

Analyse de contenus visuels complémentaires statiques et dynamiques

Sylvie Chambon

▶ **To cite this version:**

Sylvie Chambon. Analyse de contenus visuels complémentaires statiques et dynamiques. Vision par ordinateur et reconnaissance de formes [cs.CV]. Institut National Polytechnique (Toulouse), 2020. ⟨tel-04912970⟩

HAL Id: tel-04912970

https://hal.science/tel-04912970v1

Submitted on 26 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization



Habilitation à Diriger des Recherches (HDR)

Délivrée par : *Institut National Polytechnique de Toulouse (Toulouse INP)*
Unité de Recherche : *Institut de Recherche en Informatique de Toulouse, IRIT (UMR 5505)*

Analyse de contenus visuels complémentaires statiques et dynamiques : Détection de saillance et de similarités en 2D et en 3D pour reconnaitre des objets en identifiant forme et apparence

Sylvie Chambon

Soutenue le *12 octobre 2020* **devant le jury composé de :**

Rapporteure	Samia Bouchafa	Professeure, Université d'Evry Val d'Essonne
Rapporteur	David Fofi	Professeur, Université de Bourgogne
Rapporteur	Vincent Lepetit	Directeur de recherche, École des Ponts ParisTech
Examinateur	Ferran Marqués	Professeur, Universitat Politècnica de Catalunya
Examinateur	Peter Sturm	Directeur de recherche, INRIA Grenoble
Examinatrice	Yihong Wu	Professeure, Chinese Academy of Sciences in Beijing
Correspondante	Géraldine Morin	Professeure, Institut National Polytechnique de Toulouse

À mes parents

≪ Méfies-toi du cynisme
L'avenir appartient aux idéalistes. ≫
[Faye 13]

À mes enfants

Qu'il est beau de te voir courir
Qu'il est beau de te voir sourire
Et qu'il est bleu ce ciel, qui devant toi veut s'ouvrir
[Myrtille 05]

À Ivan

On a pris les lacets
À travers la montagne
On se promène dans les sommets
On s'accompagne
[les Crayons 12]

Remerciements

Ce manuscrit est le résultat de 15 années de cheminement, de belles épreuves et de belles rencontres et j'aimerais manifester toute ma gratitude pour toutes les personnes que j'ai pu croisées et qui m'ont donné l'envie, la motivation et la satisfaction de terminer ce travail.

Tout d'abord, je souhaite remercier les trois personnes qui ont accepté avec beaucoup d'enthousiasme et de bienveillance de relire et de rapporter sur ce manuscrit : David Fofi, Samia Bouchafa et Vincent Lepetit. J'ai énormément apprécié vos retours détaillés et positifs dans les rapports mais également vos nombreuses questions passionnantes pendant la présentation orale. Je remercie les deux examinateurs Yihong Wu et Ferran Marqués pour leur retour si chaleureux et positifs. En particulier, merci pour les questions qui m'ont permis de mettre en avant des aspects qui me tiennent à coeur, notamment sur l'investissement dans les pédagogies actives dans mes enseignements et la manière d'allier méthodes classiques de vision et méthodes plus récentes de *deep learning*. Enfin, en ce qui concerne le jury, un grand merci à Peter Sturm qui a accepté d'être président de cette habilitation. Sa présence lumineuse, chaleureuse et courageuse en cette période de crise sanitaire a apporté une touche définitivement humaine au déroulé de toute la soutenance.

Il y a deux personnes qui ont joué un rôle décisif dans la réalisation de cette habilitation et je les remercie du fond du cœur. En premier, Géraldine Morin qui a accepté d'être la garante de cette habilitation. Géraldine est un véritable modèle de courage, de motivation, de détermination et d'intelligence humaine, une collègue adorable qui m'a accompagnée et soutenue depuis mon arrivée. Quel bonheur de travailler avec elle ! En second, Luce Morin qui a accepté de m'aider dans toutes les étapes (conseils sur la rédaction, choix du jury, retours sur la présentation) sans autre retribution que ces remerciements dans ce manuscrit. Mon seul regret est donc de ne pas avoir pu l'inviter dans ce jury puisqu'il n'est pas encore possible d'assumer la présence de deux grandes chercheuses soeurs dans un même jury.

Au cours de ma carrière, certaines personnes, sans s'en rendre compte, ont été décisives pour me donner l'envie de continuer et je tiens à les remercier pour cela. Il s'agit d'Alain Crouzil qui a accepté d'être mon directeur de thèse et dont les qualités en terme de chercheur et de directeur de thèse continuent à me guider encore actuellement, Isabelle Bloch qui m'a accueillie en post-doctorat et qui fut la première figure scientifique féminine inspirante pour moi, Jérôme Idier qui m'a permis de continuer à avoir des échanges constructifs et stimulants dans mon domaine de recherche pendant ma période à l'IFSTTAR et Vincent Charvillat qui m'a permis de rejoindre l'équipe REVA, l'IRIT et surtout ma ville natale, au moment le plus inattendu pour moi. Je le remercie de m'avoir fait confiance et de m'avoir intégrée dans son équipe de recherche sans hésiter.

J'ai une pensée émue pour tous mes anciens collègues de l'IFSTTAR qui m'ont joyeusement et gentiment accompagnée pendant 4 ans, je les remercie d'avoir rendu le quotidien agréable et sympathique : Jean-Marc Moliard, Christian Gouraud, Philippe Nicolle qui ne m'ont jamais laissé tomber, mais également Vincent Baltazart et Xavier Dérobert qui nous ont accueillis dans leur équipe. Merci à Fabien Treyssède et Sandra Escoffier à qui j'ai pu me confier tant de fois !

Merci à tous mes collègues de l'équipe REVA qui m'ont accompagnée et aidée : Jean-Denis et ses conseils avisés, Pierre sans qui je n'aurais jamais postulé, Simone et sa gentillesse incroyable, Axel qui a même accepté de relire quelques passages. Un énorme merci à Thierry qui a relu attentivement une grande partie du manuscrit, qui m'a aidé sur les figures et qui a accepté que je ne relise pas une

dernière fois son manuscript de thèse avant de l'envoyer. Un énorme merci à Thomas qui m'a beaucoup aidée sur les aspects LaTeXde ce manuscrit et bien sûr sur la réalisation de la diffusion en direct (sans toi, je n'y serais pas arrivée).

Tout ce travail n'aurait pas pu être réalisé sans les doctorants qui m'ont acceptée comme co-directrice ou co-encadrante. Merci à Guillaume, Rabih, Bastien, Marie-Anne, Damien et Thierry. C'était beaucoup de finesse, d'élégance et de travail bien sûr, mais également, de l'humour, de grandes émotions, et beaucoup d'aide de votre part ! Je remercie également les deux post-doctorants avec qui j'ai eu la joie de travailler : Hatem et Chafik. C'était un vrai bonheur de partager nos idées, de bénéficier de votre incroyable force de travail, de votre optimisme et de votre gentillesse. Je remercie tous les stagiaires de master 2 avec qui j'ai eu la chance de travailler : Roberta, Noémie, Hanan, Maya et Emilie.

Je remercie aussi tous les collègues du département Sciences du Numérique qui ont manifesté leur soutien et leur intérêt. En particulier, merci Aurélie, Philippe et Xavier pour les conseils sur la rédaction, merci à Katia pour les conseils pour la présentation, merci à Ronan pour l'aide précieuse pour le masque spécial conférence (je ne sais pas comment j'aurais pu m'en sortir sans).

J'ai l'incroyable chance d'avoir des amis fidèles et généreux, je les remercie ici pour toute leur bienveillance au quotidien : Charlotte, Jean-Baptiste, Carine, Marie, Yolaine, Marie-Claude, Philippe, Caroline, Antonio, Adrien, Florence, Anne, Karine, Sébastien, Solenn, Laure, Gaël, Laetitia, Manu, Claire, Eric, Thomas, Laurence, Mélanie, Marie, Audrey, Clarisse. Un remerciement tout particulier pour Eric et Carine qui m'ont donnée un toit pour travailler au calme et pour les relectures bienveillantes de Claire et Laetitia.

Ma famille a été particulièrement soutenante et bienveillante dans cette période et je les remercie du fond du coeur. Encore une fois, vous etiez tous là pour me soutenir et m'encourager. Un grand merci à me parents qui ont relu le manuscrit avec tant de patience ! Merci pour tous vos messages de tendresse pendant la rédaction et même pendant la soutenance !

Il n'y aurait pas de joie à terminer ce travail sans l'amour, la joie et l'humour au quotidien dans ma maison. Merci à mes deux petits loulous qui déménagent, merci à celui qui amène la douceur et le soleil.

Table des matières

Table des figures xvii

Liste des tableaux xix

Liste des algorithmes xxi

Notations xxiii

Avant propos : comment lire ce manuscrit ? xxv

1. Introduction 1
- 1.1. Positionnement par rapport aux domaines de recherche . . . 2
- 1.2. Apprentissage profond : une introduction . . . 3
  - 1.2.1. Apprentissage supervisé . . . 3
  - 1.2.2. Émergence de l'apprentissage profond . . . 4
- 1.3. Présentation de l'organisation du manuscrit . . . 4

I. Détection, caractérisation et appariement de primitives 2D et 3D 7

Introduction à la détection de primitives dans différentes modalités 9

Contributions scientifiques structurantes en détection et caractérisation 2D et 3D 11

2. Primitives en 2D 13
- 2.1. Introduction sur les primitives en 2D . . . 13
- 2.2. Approches de détection et de caractérisation des points d'intérêt en 2D . . . 17
  - 2.2.1. Approches basées premier ordre . . . 18
  - 2.2.2. Approches basées région . . . 18
  - 2.2.3. Approches basées second ordre . . . 19
  - 2.2.4. Introduction des notions d'analyse multi-échelle et multi-résolution . . . 19
  - 2.2.5. Classement des publications significatives et invariance des détecteurs . . . 21
- 2.3. Étude de la complémentarité des détecteurs . . . 23
  - 2.3.1. Répétabilité et gain en répétabilité . . . 23
  - 2.3.2. Répartition spatiale et gain en répartition spatiale . . . 24
  - 2.3.3. Complémentarité . . . 24
  - 2.3.4. Protocole de tests mis en œuvre . . . 25
  - 2.3.5. Résultats sur l'étude de complémentarité des détecteurs en 2D . . . 25
  - 2.3.6. Conclusion sur l'évaluation et les complémentarités des détecteurs en 2D . . . 26
- 2.4. Conclusion sur les primitives 2D . . . 27

3. Primitives en 3D 29
- 3.1. Introduction sur les primitives en 3D . . . 29

3.2. Approches de détection et de caractérisation des points d'intérêt en 3D 31
3.2.1. Introduction sur la recherche d'éléments caractéristiques 31
3.2.2. Détecter des points saillants dans un maillage 32
3.2.3. Décrire des points saillants dans un maillage 34
3.2.4. Conclusions sur la recherche de points d'intérêt dans des maillages 3D 34
3.3. Identification de points d'intérêt dans des images 3D médicales 35
3.3.1. Modèle de respiration et synthèse d'images TDM intermédiaires 36
3.3.2. Modèle physique 3D de la surface du poumon 37
3.3.3. Sélection de points de repère 38
3.3.4. Conclusion sur l'identification de points d'intérêt 3D en imagerie médicale . . . 39
3.4. Conclusion sur les primitives 3D 39

4. Appariement 2D/3D **41**
4.1. Introduction sur la mise en correspondance 2D/3D 41
4.2. Mise en correspondance 2D/3D directe 43
4.3. Choix d'une représentation commune des données 2D et 3D 44
4.4. Appariement de données 2D d'origine différente 44
4.5. Résumé de l'approche proposée 46
4.6. Extraction de primitives d'intérêt en 3D 47
4.6.1. Représentation de la surface d'un objet 47
4.6.2. Gradient d'ombrage moyen, *Average Shading Gradient*, ASG [Plötz 17] 47
4.6.3. Primitives de saillance curviligne, *Curvilinear Saliency*, CS 48
4.7. Représentation image 51
4.7.1. Primitives curvilignes proposées pour les images 51
4.8. Robustesse à la texture et au fond 53
4.8.1. Saillance curviligne multi-échelle, *Multi-Curvilinear Saliency*, MCS 53
4.8.2. Extraction de la texture : méthodes existantes 53
4.8.3. Suppression du fond avec des courbes de *focus* : approches existantes 54
4.8.4. Courbes de *focus* multi-échelle ou *Multi-Focus Curves*, MFC 55
4.9. Résultats de détection de primitives 57
4.9.1. Comparaison avec les approches existantes en détection de primitives 57
4.9.2. Critères d'évaluation des primitives détectées 57
4.9.3. Bases de données testées pour la détection de primitives 58
4.9.4. Analyses des résultats de détection de primitives 58
4.10. Recalage entre images 2D et modèles 3D 60
4.11. Résultats obtenus pour l'estimation de la pose 62
4.12. Conclusions et perspectives au recalage 2D/3D 65

Conclusion sur la détection de primitives en 2D et en 3D et leur appariement **67**

II. Appariements multiples **69**

Introduction sur la mise en correspondance multiple **71**

Contributions scientifiques structurantes en mise en correspondance **73**

5. Mise en correspondance **75**
5.1. Introduction sur la mise en correspondance de pixels 75
5.2. Coût global de mise en correspondance 76
5.2.1. Coût de correspondance 76

5.2.2. Coût des contraintes . . . 76
5.2.3. Coût global développé . . . 77
5.3. Coût local de mise en correspondance . . . 77
5.4. Appariement en utilisant une segmentation en régions . . . 78
5.5. Évaluation et comparaison des approches de mise en correspondance . . . 80
5.5.1. Images de tests utilisées et proposées pour l'évaluation de l'appariement . . . 80
5.5.2. Hypothèse sur les images stéréoscopiques utilisées . . . 80
5.5.3. Critères d'évaluation utilisés pour mesurer la qualité d'une carte de disparités . 81
5.6. Conclusion sur la mise en correspondance de pixels . . . 81

6. Contributions à la mise en correspondance stéréoscopique 83
6.1. Étude de l'utilisation de la couleur dans des approches globales . . . 83
6.1.1. Coûts globaux d'appariement faisant intervenir la couleur . . . 84
6.1.2. Optimisation du coût de mise en correspondance . . . 86
6.1.3. Étude des zones de distorsions radiométriques . . . 86
6.1.4. Résultats de mise en correspondance globale utilisant la couleur . . . 87
6.1.5. Conclusion sur l'utilisation de la couleur dans des approches globales . . . 89
6.2. Combinaisons de mesures de corrélation . . . 90
6.2.1. Étude de complémentarité . . . 91
6.2.2. Algorithme de mise en correspondance par fusion de N_m mesures . . . 93
6.2.3. Résultats de mise en correspondance par combinaison de mesures . . . 94
6.2.4. Conclusion sur l'appariement par combinaison de mesures de corrélation . . . 95
6.3. Appariement par propagation de germes . . . 96
6.3.1. Appariement par propagation de germes avec mesure de confiance . . . 97
6.3.2. Résultats expérimentaux sur l'appariement par propagation de germes . . . 99
6.3.3. Conclusion sur la mise en correspondance par propagation . . . 100
6.4. Appariement par vote . . . 101
6.4.1. Algorithme proposé pour la mise en correspondance par vote . . . 102
6.4.2. Résultats de la mise en correspondance par sondage de régions . . . 104
6.4.3. Conclusion sur la mise en correspondance par sondage des régions . . . 105
6.5. Conclusion sur les contributions en mise en correspondance stéréoscopique . . . 107

7. Analyse de scènes par vidéos multiples avec recouvrement 109
7.1. Introduction sur l'utilisation de vidéos multiples avec recouvrement . . . 109
7.2. Reformulation de trajectoire requête . . . 112
7.2.1. Algorithme de reformulation de trajectoire requête . . . 112
7.2.2. Corrélation de fonctions d'activité . . . 112
7.2.3. Mise en correspondance de cellules dans des vidéos . . . 113
7.2.4. Classement des vidéos par intérêt lié à la complémentarité des vues . . . 113
7.2.5. Résultats de reformulation de requête . . . 113
7.2.6. Conclusion sur la reformulation de trajectoires . . . 116
7.3. Construction du graphe de relations entre vidéos . . . 116
7.3.1. Définition de la notion d'histoire . . . 116
7.3.2. Notion de distance entre histoires . . . 117
7.3.3. Estimation du graphe de relations . . . 118
7.3.4. Résultats de construction de graphes . . . 119
7.3.5. Conclusion sur l'analyse de scène par vidéos multiples avec recouvrement . . . 121

Conclusion sur la mise en correspondance conjointe 123

III. Segmentation photométrique et géométrique **125**

Introduction à la segmentation multi-critères **127**

Contributions scientifiques structurantes en segmentation **129**

8. Méthodes de segmentation et de sur-segmentation d'images **131**
8.1. Introduction sur la notion de segmentation . . . 131
8.2. Approches de segmentation existantes . . . 133
8.3. Notion de sur-segmentation et de superpixels . . . 134
8.3.1. Introduction à la sur-segmentation . . . 134
8.3.2. Domaines d'application des superpixels . . . 134
8.4. Définition des superpixels . . . 135
8.4.1. Propriétés . . . 135
8.4.2. Construction/Modélisation . . . 137
8.4.3. Classement et comparaison des différents constructeurs de superpixels . . . 139
8.5. Conclusion sur les superpixels . . . 139

9. Superpixels adaptés aux scènes planaires **141**
9.1. Objectif applicatif : la segmentation de scènes urbaines pour une ville confortable . . . 141
9.2. Problématique de la segmentation de scènes urbaines . . . 142
9.3. Approches existantes en segmentation de scènes urbaines . . . 142
9.4. Méthode proposée pour l'analyse de scènes urbaines . . . 143
9.5. Définition des superpixels géométriques . . . 143
9.6. Segmentation à partir de superpixels géométriques . . . 146
9.7. Conclusions sur la segmentation de scènes urbaines . . . 147

10. Détection d'éléments fins par chemins minimaux **151**
10.1. Objectif applicatif : maintenance des chaussées pour la sécurité routière . . . 151
10.2. Problématique de la détection de fissures de la chaussée . . . 151
10.3. Résumé des contributions en segmentation de fissures . . . 152
10.4. Approches existantes en segmentation de fissures . . . 153
10.4.1. Approches photométriques . . . 153
10.4.2. Approches photométriques et géométriques . . . 153
10.4.3. Approches introduisant la notion de topologie . . . 153
10.4.4. Approches de segmentation de fissures par chemins minimaux . . . 153
10.5. Chemins minimaux adaptés aux images de fissures . . . 155
10.6. Sélection de chemins minimaux pour la détection de fissures . . . 156
10.7. Protocole d'évaluation pour la détection de fissures . . . 158
10.7.1. Images de référence testées . . . 158
10.7.2. Choix des méthodes comparées . . . 159
10.7.3. Critère d'évaluation . . . 160
10.8. Étude des performances . . . 160
10.8.1. Influence des paramètres . . . 160
10.8.2. Évaluations et comparaisons sur les données synthétiques . . . 162
10.8.3. Comparaison sur les images réelles . . . 162
10.9. Conclusions sur la segmentation de fissures de chaussée . . . 164

11. Segmentation combinée avec la saillance curviligne **167**
11.1. Objectif applicatif : traitement éco-responsable de la vigne . . . 167
11.2. Problématique de segmentation des images d'insectes . . . 167

11.3. Approches existantes pour segmenter des images d'insectes . . . 168
11.4. Reconnaissance s'appuyant sur SVM et des descripteurs HCS . . . 169
11.4.1. Présentation générale de l'algorithme proposé . . . 169
11.4.2. Étape (1) : détection robuste des contours des insectes . . . 169
11.4.3. Étape (2) : affinement de la segmentation pour supprimer les artefacts liés aux insectes collés entre eux . . . 169
11.4.4. Étape (3) : classification des insectes . . . 170
11.5. Évaluation de l'approche de reconnaissance de papillons . . . 171
11.5.1. Images testées pour la reconnaissance de papillons . . . 171
11.5.2. Étude des paramètres de l'approche de reconnaissance de papillons . . . 171
11.5.3. Résultats de segmentation de papillons . . . 171
11.6. Conclusion sur la détection de papillons . . . 172

Conclusion sur la segmentation multi-critères **173**

IV. Apprentissage profond pour l'analyse de contenus visuels complémentaires **175**

Détection, mise en correspondance et segmentation pour identifier des objets **177**

12.Éléments indispensables pour l'apprentissage profond **179**
12.1. Base de données pour l'apprentissage profond . . . 179
12.2. Différents réseaux de neurones . . . 180
12.2.1. Réseaux de neurones convolutifs, *Convolutional Neural Network*, CNN . . . 180
12.2.2. Réseaux de neurones récurrents, *Recurrent Neural Network*, RNN . . . 180
12.2.3. Réseaux antagonistes génératifs, *Generative Adversarial Networks*, GAN . . . 181
12.3. Aspects à maîtriser pour utiliser un réseau de neurones . . . 182

13.Perspectives **183**
13.1. Vers une estimation robuste de la pose par détection de marqueurs et de points d'intérêt 183
13.2. Vers des squelettes texturés et appariés à base de points d'intérêt . . . 185
13.3. Vers une combinaison de multiples vidéos désynchronisées avec ou sans recouvrement . 187
13.4. Combinaison et corrélation d'information multimédia . . . 190
13.4.1. Corrélation de données au service du suivi médical post-cancer . . . 190
13.4.2. Combinaison de vidéos au service d'une ville intelligente, confortable et sécurisée 192
13.5. Vers la reconstruction d'un modèle 3D complet de qualité garantie . . . 193
13.6. Vers une utilisation conjointe et uniformisée de points saillants en 2D et en 3D . . . 195
13.7. Vers une mise en correspondance et une segmentation conjointe . . . 196

Bibliographie **200**

Annexes **233**

A. Annexe relative à la détection de primitives en 2D et en 3D **235**
A.1. Détecteurs de points d'intérêt en 2D : les détails . . . 235
A.1.1. Construction d'une pyramide d'images . . . 235
A.1.2. Détecteurs du premier ordre . . . 236
A.1.3. Vers le multi-échelle : Harris-Laplace [Mikolajczyk 04] . . . 238
A.1.4. Détecteurs de régions . . . 238

A.1.5. Détecteurs du second ordre . . . 243
A.1.6. Kitchen et Rosenfeld [Kitchen 82] . . . 243
A.2. Descripteurs de points d'intérêt en 2D . . . 248
A.3. Rappels de géométrie différentielle . . . 252
A.3.1. Notations . . . 252
A.3.2. Définitions: Carte différentielle – Plan tangent – Carte de Gauss . . . 252
A.3.3. Courbures : formes fondamentales . . . 254
A.3.4. Analyses de l'indicatrice de Dupin . . . 256
A.3.5. K, courbure gaussienne de $\mathcal{D}$: preuve . . . 256
A.3.6. Preuve de la proposition 2 . . . 257
A.4. Résultats de détection de primitives avec MCS et MFC . . . 257

B. Éléments nécessaires à la mise en correspondance 259
B.1. Mesures de corrélation . . . 259
B.1.1. Corrélation croisée . . . 259
B.1.2. Mesures classiques . . . 259
B.1.3. Corrélation des champs de gradients . . . 260
B.1.4. CENSUS, mesure de recensement . . . 260
B.1.5. *Smooth Median Powered Deviation* . . . 261
B.1.6. Information mutuelle . . . 262
B.2. Représentation de la couleur . . . 262
B.2.1. Systèmes de primaires . . . 262
B.2.2. Systèmes luminance-chrominance . . . 264
B.2.3. Systèmes perceptuels . . . 265
B.2.4. Systèmes d'axes indépendants . . . 265
B.2.5. Systèmes hybrides . . . 266
B.2.6. Synthèse des systèmes de représentation de la couleur . . . 266
B.3. Résultats liés à la mise en correspondance globale utilisant la couleur . . . 266

C. Méthodes de segmentation existantes 271
C.1. Approches par contours . . . 271
C.2. Approches par régions . . . 273
C.2.1. Approches par histogramme . . . 273
C.2.2. Croissance de régions . . . 274
C.2.3. Ligne de partage des eaux . . . 275
C.3. Approches inspirées des approches de classification . . . 276
C.3.1. k-moyennes . . . 276
C.3.2. *Mean-Shift* [Comaniciu 02] . . . 277
C.3.3. *Support Vector Machine* . . . 279

D. Présentation résumée de l'ensemble de la carrière 281
D.1. Période présentée . . . 281
D.2. Intéractions dans les différentes équipes de recherche . . . 281
D.3. Encadrements de recherche . . . 282
D.3.1. Thèses soutenues . . . 282
D.3.2. Thèse en cours . . . 283
D.3.3. Encadrements de post-doctorants . . . 283
D.3.4. Encadrements de stagiaires de master 2 . . . 283
D.4. Collaborations . . . 284
D.4.1. Collaborations internationales . . . 284

D.4.2. Collaborations nationales 285
D.4.3. Collaborations au sein du laboratoire IRIT 286
D.5. Responsabilités scientifiques 286
D.5.1. Projets européens 286
D.5.2. Implication internationale 286
D.5.3. Projets nationaux 287
D.5.4. Implication nationale 287
D.5.5. Vulgarisation scientifique 288
D.6. Enseignement 288
D.7. Investissement pédagogique et administratif 289

Table des figures

1.1. Domaines abordés en traitement d'images et vision par ordinateur. 2
1.2. Reconnaître un objet en 2D à partir d'un modèle 3D. 9

2.1. Exemples de travaux de recherche sur les primitives d'intérêt en 2D. 13
2.2. Exemples de projets d'enseignement sur les primitives d'intérêt en 2D. 14
2.3. Étapes classiques du fonctionnement d'un détecteur. 15
2.4. Classement des approches de détection de points d'intérêt en 2D. 17
2.5. Illustration de l'introduction d'une analyse multi-échelle dans une pyramide d'images (multi-résolution). 19
2.6. Résultats obtenus avec quatre détecteurs de points d'intérêt en 2D. 21
2.7. Résultats de la répétabilité des détecteurs en faisant varier ϵ. 26

3.1. Exemples de primitives d'intérêt en 3D. 29
3.2. Détection de points saillants dans des maillages 3D par étude des normales à la surface. 32
3.3. Extraction de points d'intérêt dans des modèles 3D. 33
3.4. Extraction de plans de symétrie et de translation dans des modèles 3D. 34
3.5. Problème de recalage d'images 3D médicales. 35
3.6. Modèle physique de la respiration utilisé pour le recalage multi-modal basé points d'intérêt. 37
3.7. Sélection de points d'intérêt en 3D (vue sur une coupe transversale d'un seul poumon). 39
3.8. Résultat de recalage par points d'intérêt entre une image TEP et des images TDM. . 40

4.1. Reconnaître un objet en 2D à partir d'une collection de modèles 3D. 41
4.2. Apparier une image avec des modèles 3D. 46
4.3. Coniques réelles associées à l'indicatrice de Dupin. 50
4.4. Illustration de la saillance curviligne. 51
4.5. Modèle optique pris en compte pour définir le flou de mise au point [Zhuo 11]. 55
4.6. Modèle de flou dans [Zhuo 11]. 56
4.7. Cartes de détection obtenues avec CS ainsi que MCS et MFC sur 5 échelles. 59
4.8. Pourcentage de répétabilité pour 2 exemples de modèle 3D. 60
4.9. Exemple de scores de répétabilité obtenus entre une image et 40 cartes de profondeur. 63
4.10. Taux de précision du recalage 2D/3D en fonction du rang de similarité. 64
4.11. Exemples de recalage 2D/3D corrects avec CS/MFC sur la base *PASCAL3D+*. 65

5.1. Étapes des méthodes de mise en correspondance fondées sur les régions. 79
5.2. Exemples de données testées pour la mise en correspondance stéréoscopique. 80

6.1. Exemple d'intérêt de la couleur pour la mise en correspondance. 83
6.2. Estimation des zones de distorsions radiométriques. 86
6.3. Évaluation quantitative réalisée pour 8 systèmes de représentation de la couleur. . . . 87
6.4. Coûts obtenus avec l'intensité seule et une des composantes chromatiques seule. . . . 88
6.5. Erreurs d'appariement avec ou sans couleur, avec méthode globale ou locale. 89

6.6. Étude de la complémentarité. 92
6.7. Erreurs d'appariement en fonction du nombre de mesures théoriquement combinées. . 93
6.8. Cartes de disparités obtenues avec la combinaison automatique de 2 mesures. 95
6.9. Comparaison mise en correspondance par corrélation et par propagation de germes. . 96
6.10. Évaluation de la de mise en correspondance par propagation (densité $50\% < D < 75\%$). 99
6.11. Évaluation de la mise en correspondance par propagation (densité de $D \geq 75\%$). . . 100
6.12. Cartes de disparités obtenues par propagation de germes. 101
6.13. Étapes de la mise en correspondance par sondage de régions. 102
6.14. Principe du sondage de régions pour la mise en correspondance par vote. 103
6.15. Carte des zones difficiles à apparier. 104
6.16. Visualisation des valeurs de confiance pour l'approche d'appariement par vote. . . . 105
6.17. Cartes de disparités obtenues par sondage des régions. 106

7.1. Illustration des dispositifs de caméras que nous souhaitons étudier. 109
7.2. Illustration des ensembles de vidéos que nous souhaitons étudier. 110
7.3. Première problématique abordée dans le cadre de l'étude de réseaux de caméras. . . . 110
7.4. Deuxième problématique abordée dans le cadre de l'étude de réseaux de caméras. . . 111
7.5. Algorithme de reformulation de trajectoires. 112
7.6. Classement des vidéos par degré d'intérêt. 114
7.7. Évaluation de la reformulation de trajectoires. 115
7.8. Approche de construction d'un graphe de liens dans une collection de vidéos. 116
7.9. Notion d'histoire pour rechercher des similarités dans des vidéos. 117
7.10. Construction de l'histoire d'une cellule. 117
7.11. Analyse multi-échelle pour construire un graphe des liens dans une collection de vidéos. 118
7.12. Résultat de l'estimation du graphe de relations des vidéos. 120

8.1. Exemples de segmentation. 132
8.2. Différences de résultats avec différents constructeurs de superpixels. 140

9.1. Système mobile d'acquisition `imajbox`. 141
9.2. Triplet d'images successives acquises par l'`imajbox`. 142
9.3. Approche de segmentation de scènes urbaines par superpixels géométriques. 144
9.4. Intégration de l'information géométrique dans la construction de superpixels. 145
9.5. Analyse du comportement des termes utilisés dans la distance d'agrégation proposée. 146
9.6. Comparaison visuelle des approches SLIC et GEOM. 147
9.7. Rappel et erreur de sous-segmentation obtenus pour SLIC et GEOM. 148
9.8. Segmentation sémantique en plan sur des données synthétiques et *Merton College III*. 149

10.1. Système d'acquisition utilisé pour détecter des fissures. 152
10.2. Résultats intermédiaires de segmentation de fissures par chemins minimaux. 157
10.3. Influence du paramètre P sur les chemins détectés et les temps d'exécution. 160
10.4. Évolution des vrais positifs parmi les points extrémités en fonction du paramètre k_e. . 161
10.5. Précision, sensibilité et valeur DSC pour la détection des fissures réelles. 161
10.6. Résultats de segmentation de fissures synthétiques. 162
10.7. Précision, sensibilité et valeur DSC pour la détection des fissures synthétiques. . . . 162
10.8. Résultats de segmentation de fissures réelles. 163
10.9. Précision, sensibilité et valeur DSC moyennes sur les 36 images réelles. 164
10.10Résultats de détection de fissures avec MPS. 164

11.1. Exemple d'image de papillons et de résultat obtenus. 168
11.2. Approche de segmentation de papillons Eudemis proposée. 170

11.3. Résultats de détection de papillons dans des images réelles. 172

13.1. Recherche de la pose associée à l'image requête à l'aide d'images de référence. 184
13.2. Reconstruction 3D à partir des squelettes perspectifs des formes. 185
13.3. Analyse de réseaux de caméras complexes avec prise en compte du décor. 189
13.4. Traitement des données relatives aux patients suivis après un cancer. 191
13.5. Reconstruction 3D du Capitole de Toulouse par MVS [AliceVision 18]. 193
13.6. Illustrations de la compression spatiale et temporelle d'images. 195
13.7. Exemples de superpixels calculés sur un couple stéréoscopique. 197
13.8. Exemple de superpixels utilisés pour effectuer la mise en correspondance. 198
13.9. Résultat d'appariement en combinant approche par corrélation et sur-segmentation. . 198

A.1. Comportement de l'auto-corrélation suivant le type de points. 237
A.2. Masque circulaire utilisé par le détecteur SUSAN. 239
A.3. SUSAN : Comportement de la fonction de similarité. 239
A.4. Octaves et échelles pour le détecteur SIFT. 245
A.5. Algorithme de détection des points SIFT. 246
A.6. Utilisation de l'image intégrale. 248
A.7. Illustration du *Box filtering*. 249
A.8. Histogramme des orientations pour le descripteur SIFT. 250
A.9. Illustration d'une carte gaussienne associée à une surface. 253

B.1. Mesure CENSUS. 261
B.2. Tests pour les coûts globaux d'appariement. 267
B.3. Cartes des erreurs de disparités obtenues pour l'image Moebius. 268

C.1. Approche classique par détection de contours puis fermeture. 271
C.2. Approche par ligne de partage des eaux. 275
C.3. Approches par k-moyennes avec différents critères. 276
C.4. Application de l'algorithme de *Mean-Shift*. 278

D.1. Carrière résumée en une frise. 291

Liste des tableaux

2.1. Publications significatives sur la détection de points d'intérêt en 2D. 21
2.2. Invariance des différents détecteurs présentés. 22
2.3. Détecteurs les plus complémentaires en 2D. 26

4.1. Pourcentage moyen d'intersection des détecteurs de points d'intérêt pour *web collection*. 58
4.2. Pourcentage moyen d'intersection des détecteurs de points d'intérêt pour *PASCAL3D+*. 60
4.3. Précision de la pose estimée avec CS, ASG, AR, IG et nos propositions MFC et MCS. 63
4.4. Erreur moyenne sur l'estimation de la pose. 64

6.1. Pourcentage d'appariements erronés en fonction des mesures combinées. 94
6.2. Choix des paramètres pour l'approche par propagation de germes. 100
6.3. Résultats de la mise en correspondance par sondage de régions. 105

8.1. Classement des constructeurs de superpixels. 140

11.1. Évaluation et comparaison d'approches de segmentation de papillons. 172

A.1. Distance de Hausdorff moyenne des détecteurs de points d'intérêt pour *web collection*. 257
A.2. Distance de Hausdorff moyenne des détecteurs de points d'intérêt pour *PASCAL3D+*. 258

B.1. Systèmes représentatifs. 269
B.2. Systèmes représentatifs (suite). 270

D.1. Enseignements réalisés de 2002 à 2011. 289
D.2. Enseignements réalisés depuis 2011. 289

Liste des Algorithmes

1. Algorithme de mise en correspondance par propagation séquentielle de germes. 97
2. Estimation du lien entre deux vidéos. 119
3. Algorithme de segmentation de fissures par sélection de chemins minimaux. 158
4. Post-traitement des fissures détectées par sélection de chemins minimaux. 159
5. Segmentation par contours actifs. 273
6. Algorithme *mean-shift*. 278

Notations

Règles et conventions

x, N	Les scalaires sont toujours notés en caractères normaux.
$\tilde{x}, \tilde{y}$	Ce sont les coordonnées cartésiennes locales.
$\mathbf{v}$	Les vecteurs ou fonctions vectorielles sont noté(e)s en minuscules et en gras.
$\mathbf{A}$	Les matrices ou fonctions matricielles sont notées en majuscules et en gras.
$\mathcal{E}$	Les ensembles sont notés en lettres calligraphiées.

Notations communes à tous les chapitres

$\mathbf{v}^T, \mathbf{A}^T$	Vecteur, respectivement matrice, transposé(e) de $\mathbf{v}$, respectivement $\mathbf{A}$.
$\|\mathbf{v}\|$	Norme euclidienne du vecteur $\mathbf{v}$.
$\mathbf{P}$	Point 3D de la scène.
$\mathbf{p}$	Point 2D/pixel d'une image.
I	Image qui à chaque point/pixel associe un niveau de gris.
I_l	Cela correspond à la $l^{\text{ième}}$ image ou, si $l \in \{g, d\}$ respectivement à l'image gauche et l'image droite.
$I(\mathbf{p})$, $I(x, y)$	Niveau de gris de $\mathbf{p}$, aux coordonnées (x, y), dans l'image I.
$I_x, I_y, I_{xx}, I_{yy}, I_{xy}$	Images des dérivées premières et secondes de l'image I.
$I_x(\mathbf{p}), \ldots, I_{xy}(i, j))$	Valeurs des dérivées premières et secondes au point $\mathbf{p}$ aux coordonnées (i, j). Pour simplifier, parfois nous noterons, par exemple, seulement I_x pour $I_x(i, j)$ ou $I_x(\mathbf{p})$.
$\nabla I(\mathbf{p})$, $\nabla I(x, y)$	Vecteur gradient au point $\mathbf{p}$, aux coordonnées (x, y), dans l'image I.
∇f	Vecteur gradient associé à la fonction (scalaire) f.
$\#(\mathcal{E})$	Nombre d'éléments de l'ensemble $\mathcal{E}$.
$\bigotimes$	Opérateur de convolution.
$\mathbf{F}_x$	Dérivée partielle du premier ordre de la fonction vectorielle $\mathbf{F}$ en x, $\frac{\partial \mathbf{F}}{\partial x}$.
$\mathbf{F}_{xy}$	Dérivée partielle du second ordre de la fonction vectorielle $\mathbf{F}$ en x et y, $\frac{\partial^2 \mathbf{F}}{\partial x \partial y}$.
$\mathbf{p}^{i,j}$	Pixel de I de coordonnées $(i\ j)^T$.
E_{global}	Coût global.
$E_{\text{correspondance}}$	Coût d'attache aux données.
$E_{\text{contrainte}}$	Coût des contraintes.
E_{local}	Coût local.
$E_{\text{dissimilarité}}$	Coût de dissimilarité.
$E_{\text{voisinage}}$	Coût du voisinage.
E_{lissage}	Coût de lissage.
$\mathcal{S}$	Support du coût global.
$\text{ZA}(\mathbf{p}_g^{i,j})$	Zone d'agrégation associée à $\mathbf{p}_g^{i,j}$ pour le coût local.
$\text{ZAC}(\mathbf{p}_g^{i,j})$	Zone d'agrégation associée à $\mathbf{p}_g^{i,j}$ pour le coût des contraintes.

NOTATIONS COMMUNES À TOUS LES CHAPITRES (suite)

$\mathrm{Z}_d(\mathbf{p}_g^{i,j})$	Zone de recherche associée à $\mathbf{p}_g^{i,j}$.
$f_{\text{dissimilarité}}$	Fonction de la différence des niveaux de gris.
$\mathbf{f}$	Vecteur contenant les niveaux de gris des pixels des fenêtres de corrélation dans I avec $\mathbf{f} = (\cdots I^{i+p,j+q} \cdots)^T = (\cdots I^k \cdots)^T$ où $p \in [-N_v; N_v]$, $q \in [-N_h; N_h]$.
$\|\mathbf{f}\|_P$	P-normes.
$\mathbf{f}_1 \cdot \mathbf{f}_2$	Produit scalaire.
$\overline{\mathbf{f}}$	Moyenne des valeurs de $\mathbf{f}$.
sgn	Fonction signe.
D_{Ham}	Distance de Hamming.

Avant propos : comment lire ce manuscrit ?

L'objectif de ce manuscrit est de produire un document complet en présentant, d'une part, les **contributions de recherche**, et, d'autre part, un **état de l'art des domaines abordés** afin de mettre en avant des perspectives qui s'inscrivent à la fois dans la continuité des travaux que nous avons réalisés et dans la réalité du contexte bibliographique actuel.

Plus précisément, nous avons souhaité rédiger des synthèses bibliographiques les plus complètes et exhaustives possibles des domaines dans lesquels s'inscrivent nos recherches. Le but est de donner une vision globale pour permettre au lecteur d'avoir le recul nécessaire pour situer toutes les contributions présentées, qui sont toutes des **contributions postérieures à la thèse soutenue en décembre 2005**.

De plus, une partie des descriptions bibliographiques est étroitement liée à des **enseignements de spécialité** dispensés à l'ENSEEIHT et nous le signalerons dès que possible. Enfin, certains des travaux exposés ont fait l'objet d'une recherche plus appliquée que théorique et la distinction entre les deux types de contributions est indiquée dans le manuscrit.

À chaque début de partie, dans un chapitre dédié dont le nom débute par **contributions scientifiques structurantes**, plus précisément aux pages 11, 73 et 129, toutes les contributions relatives à cette partie sont présentées de manière très synthétique par un texte encadré, comme celui-ci.

De plus, pour faciliter la lecture, et le positionnement de chaque chapitre de ce manuscrit, un code couleur, sur le bord des pages, indique la catégorie, de la manière suivante :

	Il s'agit d'un chapitre qui correspond à une contribution liée à l'**enseignement/une synthèse bibliographique**.
	Il s'agit d'un chapitre qui correspond à une contribution liée à mes **recherches, généralement validées par des publications, citées dans le chapitre**.
	Il s'agit d'un chapitre qui correspond à une contribution liée à mes **recherches appliquées, généralement complétées par une production de code, et également validées par des publications**.
	Il s'agit d'un chapitre qui correspond à des **perspectives**.
	Les introductions, transitions, et conclusions permettant de faire le lien ont cette couleur. Les synthèses des **contributions scientifiques structurantes** portent aussi cette couleur.

Enfin, vous pouvez trouver un CV détaillé sur toute la carrière en annexe D, page 281.

Chapitre 1.

Introduction

Un enfant voit avec la même acuité qu'un adulte à partir d'un an mais il lui faudra toute sa petite enfance pour comprendre visuellement son environnement [Grégoire 15] !

Dans le domaine de la vision par ordinateur, la reconnaissance des éléments qui composent une scène (objets, personnes, décors) est un domaine très actif qui nécessite de nombreuses compétences en détection d'objets ou de primitives d'intérêt, caractérisation de ces éléments d'intérêt, appariement, en particulier entre acquisitions de nature différente, segmentation de scènes ou d'objets particuliers, et enfin en classification. Tous ces thèmes peuvent être abordés en utilisant ou non des données d'apprentissage, c'est-à-dire, avec ou sans connaissance *a priori*, avec ou sans données d'exemple, avec ou sans base de données annotées.

Les besoins sont de plus en plus importants, aussi bien pour des applications tout public, comme la reconnaissance automatique de visages dans des photographies ou l'ajout d'information sur des images dans les réseaux sociaux, que pour des applications plus ciblées, comme dans l'industrie avec la reconnaissance de défauts sur des pièces manufacturées, ou encore des applications plus confidentielles, comme l'aide aux enquêtes à partir de vidéos de scènes de crime. En parallèle, les données disponibles sont de plus en plus variées car de plus en plus faciles à acquérir *via* les webcams couramment disponibles sur les ordinateurs personnels, les smartphones, les caméras de surveillance, les capteurs de profondeur, en ce qui concerne le grand public, mais également *via* des systèmes privés plus élaborés ou plus coûteux, comme les scanners 3D ou les capteurs LIDAR, *LIght Detection And Ranging*, ou encore les modèles de CAO, Conception Assistée par Ordinateur.

Un système de vision complet devrait pouvoir exploiter l'intérêt de tous les systèmes de capture (optiques ou non) mis à sa disposition. Et, en effet, de nombreux travaux tentent de combiner astucieusement toutes ces sources d'information. Dans ce manuscrit, nous verrons que les efforts de recherche se sont concentrés sur la manière de combiner toutes les sources d'information accessibles : images optiques, images médicales, modèles 3D mais également tous les types de résultats intermédiaires disponibles : points d'intérêt, sur-segmentation, segmentation, afin d'apparier et de combiner au mieux toutes les informations disponibles. Ainsi, la première partie a pour objectif de synthétiser les travaux concernant la détection et la caractérisation des objets en combinant la variété des données disponibles et la variabilité des informations qu'il est possible d'extraire. La deuxième partie développe les travaux concernant l'appariement, encore une fois, en prenant en compte plusieurs sources d'information, comme, par exemple, de multiples résultats intermédiaires de mise en correspondance. Enfin, la troisième partie aborde la segmentation en utilisant plusieurs critères (géométriques et photométriques) pour caractériser les formes à reconnaître, à différentes échelles. Cette partie permettra d'introduire les différentes applications investies au cours des recherches menées.

Tous les éléments vus dans ce manuscrit portent sur les outils de vision par ordinateur pour la reconnaissance. Actuellement, ces outils sont de plus en plus étudiés et, surtout, adaptés pour introduire les outils d'apprentissage par réseaux de neurones. En effet, les avancées scientifiques en terme

d'apprentissage profond, débutées en 2010, révolutionnent tous les concepts associés à ces outils. Il paraît donc important de prendre en compte ces évolutions pour comprendre les futures directions de recherche dans ces domaines et les perspectives présentées prennent en compte cet aspect.

Ainsi, dans cette introduction, dans un premier temps, nous essayons de définir précisément les domaines abordés par nos travaux, et, dans un deuxième temps, nous présentons une courte introduction sur l'apprentissage profond.

1.1. Positionnement par rapport aux domaines de recherche

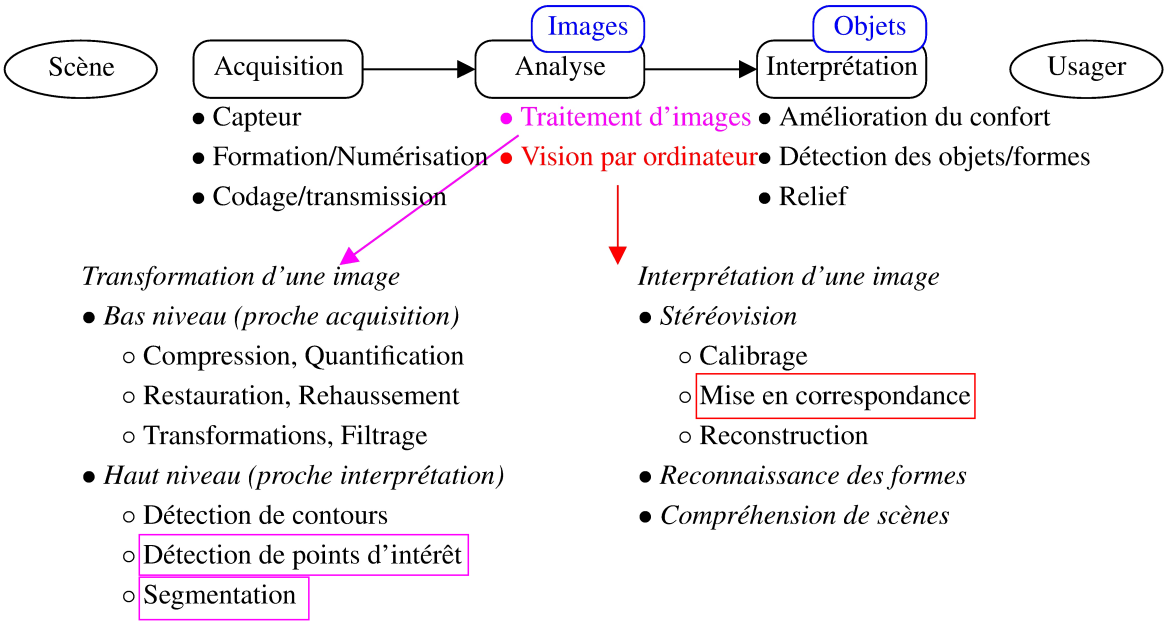


FIGURE 1.1. – Domaines abordés en traitement d'images et vision par ordinateur – Nous indiquons les domaines étudiés dans ce manuscrit, encadrés en magenta pour le traitement d'images et en rouge pour la vision par ordinateur, parmi tous les domaines que nous avons considérés et classés pour mettre en œuvre une approche complète de la scène à l'usager, en passant par les trois étapes fondamentales : acquisition, analyse et interprétation.

Les aspects abordés en traitement d'images et en vision par ordinateur couvrent de très nombreux domaines : de l'acquisition d'une scène à sa compréhension, comme nous pouvons le voir dans la figure 1.1. Plus précisément, le **traitement d'images** permet de modifier les images manipulées afin, soit d'en améliorer la visualisation, soit d'en extraire une autre information, comme par exemple, rehausser le contraste de l'image pour mieux faire apparaître les contours, ou encore augmenter la dynamique des niveaux de gris d'un objet spécifique de la scène que nous souhaitons mettre en évidence. Par ailleurs, nous pouvons citer la compression et la quantification (réduction du nombre de couleurs utilisées dans l'image) qui permettent de diminuer la quantité de données contenue dans l'image, tout en conservant l'information utile, afin d'en faciliter la transmission. En complément, la **vision par ordinateur** concerne l'extraction d'une information sur la scène, comme le relief, la reconnaissance d'un objet particulier ou une mesure relative aux objets de la scène.

La nuance entre ces deux domaines est assez délicate et nous comprenons bien que certaines disciplines peuvent aussi bien être étiquetées « traitement d'images » que « vision par ordinateur ». La segmentation ou la détection de points d'intérêt sont des exemples assez clairs sur cette difficulté à

établir un classement, puisque dans de nombreuses conférences en vision par ordinateur, des approches de segmentation ou de détection de points d'intérêt sont introduites. Cependant, dans le découpage présenté dans la figure 1.1, nous avons choisi de le classer dans les approches de traitement d'images. En effet, une segmentation seule ou une extraction de points d'intérêt seule, apporte peu d'information directe sur la scène. C'est l'utilisation de la segmentation combinée à une analyse sémantique qui va nous permettre de conclure sur la composition de la scène. C'est l'appariement et la triangulation des points d'intérêt qui va nous apporter une information de profondeur nécessaire pour la compréhension de la scène.

Ainsi, dans le cadre des travaux que nous avons menés, les domaines suivants ont été abordés :

- la **détection de points d'intérêt**, partie I ;
- la **mise en correspondance**, partie II ;
- la **segmentation**, partie III.

Les différentes méthodes que nous avons étudiées et introduites nous ont permis de manipuler aussi bien des **images**, des **vidéos**, des **images 3D** et des **images 3D temporelles** que des **modèles 3D** (surfaciques). fig:error Tous les outils utilisés, étudiés ou introduits, appartiennent majoritairement au domaine de la vision par ordinateur classique mais face à l'essor de l'apprentissage profond dans ce domaine, actuellement, nous nous sommes intéressés à ces outils qui nous semblent incontournables. Notre objectif n'est pas de proposer de nouvelles approches d'apprentissage profond, mais d'être en capacité d'identifier les outils majeurs utilisés dans les domaines visés (détection, appariement et segmentation pour la reconnaissance d'objets) pour être capable de les comprendre, de les utiliser, et de proposer des perspectives s'appuyant sur ces nouveaux outils.

Ainsi, nous proposons, dans chaque partie, d'une part, de définir l'impact de l'introduction de l'apprentissage profond sur l'avenir de la recherche dans les différents domaines abordés, et, d'autre part, de positionner les futurs travaux de recherche dans ce contexte, en particulier dans les perspectives de recherche traitées dans la partie IV. C'est pourquoi, dans un premier temps, nous effectuons une brève introduction à l'apprentissage profond et aux réseaux de neurones.

1.2. Apprentissage profond : une introduction

1.2.1. Apprentissage supervisé

Pour commencer, il semble important de définir ce que nous appelons l'apprentissage supervisé puisque l'apprentissage profond en est un sous-domaine. En effet, il s'agit d'un type d'apprentissage supervisé faisant intervenir des réseaux de neurones. Pour rédiger cette section, nous nous sommes inspirés du cours de [Carlier 19].

L'**apprentissage supervisé** vise à produire des programmes capables d'effectuer une tâche sans l'avoir explicitement codée [Samuel 59]. Il est dit que le programme apprend de son expérience pour répondre à la tâche, seulement s'il possède une mesure de performance qui augmente avec son expérience [Mitchell 97]. L'apprentissage supervisé nécessite donc deux éléments :

(1) un ensemble d'apprentissage, c'est-à-dire la plupart du temps des données annotées ;

(2) la construction d'un prédicteur qui va minimiser la différence entre les étiquettes/les valeurs réelles et les étiquettes à prédire [1].

Pour construire ce prédicteur de nombreux algorithmes ont été introduits : les plus connus sont les arbres de décision [Quinlan 86] et les forêts aléatoires [Breiman 01]. Actuellement, les plus utilisés sont les approches s'appuyant sur des réseaux de neurones.

En informatique, le concept de **neurone** formel, étroitement lié au concept des neurones en biologie n'est pas récent. Il a été introduit par [McCulloch 43] et a ensuite été utilisé dans un système

1. Nous parlons de classification lorsqu'il faut donner une étiquette et nous parlons de régression lorsqu'il faut donner une valeur.

d'apprentissage, avec la notion de **perceptron**, dans [Rosenblatt 58]. Un neurone possède plusieurs entrées qui permettent une réponse en sortie. En détail, la sortie d'un neurone correspond à un poids calculé à partir des entrées pondérées. Plus formellement, si nous notons ω_i les poids d'entrée, avec $\mathbf{w}$ le vecteur contenant les poids, nous obtenons la sortie y à partir des entrées x_i, stockées dans le vecteur $\mathbf{x}$, de la manière suivante :

$$y = f(\mathbf{w}^T\mathbf{x} + b), \tag{1.1}$$

où b correspond à un biais et f est la fonction d'activation. Malheureusement, ce modèle de perceptron ne permet pas la résolution de problèmes non-linéaires. C'est ce qui a été présenté dans [Marvin 69] et qui a, entre autres, entraîné un désintérêt pour les réseaux de neurones. Heureusement, dans [Rumelhart 86], ces travaux sont étendus en prenant en compte plusieurs couches et les premiers réseaux opérationnels multicouches apparaissent. Dans un réseau de neurones multicouches, nous distinguons : la couche d'entrée (les données), la couche de sortie (le résultat) et toutes les couches intermédiaires qui sont dites cachées. À l'origine, le terme **apprentissage profond** correspond aux réseaux avec au minimum deux couches cachées ou intermédiaires. Ainsi, le réseau LeNet [Lecun 98], qualifié alors de profond, possède 5 couches. Depuis 2012, avec l'arrivée du réseau AlexNet [Krizhevsky 12], le terme d'apprentissage profond correspond plutôt à plusieurs dizaines de couches.

1.2.2. Émergence de l'apprentissage profond

Dans la littérature, les réseaux de neurones sont régulièrement utilisés en vision par ordinateur, comme par exemple pour de la segmentation [Cheng 01, Bray 06a], mais cette utilisation reste anecdotique par comparaison avec l'essor à partir de 2010. En effet, comme toute méthode d'apprentissage, les approches à base de neurones ont été peu employées car elles nécessitent, entre autres, des calculs importants en terme d'apprentissage. De plus, dans de nombreuses applications, il n'existait pas de base de données annotées, de taille suffisamment importante, pour effectuer correctement la phase d'apprentissage. C'était le cas, par exemple, au début de nos travaux sur la segmentation de fissures de chaussées [Chambon 09]. Récemment, ces deux verrous ont été dépassés grâce à la mise à disposition de moyens d'acquisition de données de moins en moins coûteux, de plus en plus faciles à manipuler, mais également par la circulation des données de plus en plus simple (aussi bien en termes de partage de données qu'en termes de transfert de données). De plus, l'annotation de données devient de plus en plus facile grâce à l'arrivée de plateformes de *crowdsourcing*. Enfin, la puissance de calcul bien supérieure aux puissances possibles à la naissance des premiers réseaux de neurones a également contribué à cet essor. Une autre avancée majeure qui a permis l'utilisation efficace des réseaux de neurones a été une meilleure compréhension et une meilleure utilisation des **fonctions d'activation**, *learning activation function* [Carlier 19]. Initialement, la fonction d'activation utilisée était une sigmoïde responsable, en partie, des premiers résultats peu performants. La fonction reLU, *rectified Linear Unit*, introduite par [Nair 10], a permis de pallier les défauts de la fonction sigmoïde.

Ainsi l'amélioration de la fonction d'activation, la capacité à gérer de grandes bases de données et l'augmentation de la puissance de calcul, a permis le succès de l'apprentissage profond tel que nous le connaissons actuellement [LeCun 15, Goodfellow 16]. Cette brève introduction nous permet ainsi d'utiliser les différents termes liés aux réseaux de neurones dans la suite du manuscrit.

1.3. Présentation de l'organisation du manuscrit

Ce manuscrit présente trois parties relatives aux domaines que nous avons énumérés au § 1.1. Plus précisément, nous présentons nos contributions suivant ces trois articulations :

(1) **Détection et caractérisation de primitives en 2D et en 3D pour un appariement multimodal** – Malgré un état de l'art conséquent en détection de primitives, aussi bien

en 2D, cf. chapitre 2, qu'en 3D, cf. chapitre 3, il y a peu d'études abordant l'évaluation de la complémentarité de détecteurs 2D et 3D. Nos évaluations et nos propositions sont donc présentées de manière séparée, avant de présenter notre contribution principale qui est l'objet du chapitre 4, sur le recalage 2D/3D faisant intervenir la notion de saillance curviligne.

(2) **Appariements multiples** – Après une présentation de la mise en correspondance par corrélation, résumé du travail réalisé pendant la thèse, cf. chapitre 5, nous abordons toutes les contributions réalisées en mise en correspondance, cf. chapitre 6, sur les aspects suivants : prise en compte de la couleur pour la mise en correspondance globale, combinaisons de mesures de corrélation pour la mise en correspondance locale, appariement par propagation de germes et appariement par vote. Le dernier chapitre de cette partie, cf. chapitre 7, nous permet d'introduire nos travaux les plus récents et les plus prometteurs en abordant l'analyse de scène par vidéos multiples avec recouvrement.

(3) **Segmentation photométrique et géométrique** – Toute la carrière a été jalonnée de participations à des recherches plus appliquées et cette dernière partie regroupe les principales avancées ayant donné lieu à des collaborations industrielles et, ainsi à des productions de codes utilisés à des fins applicatives. Plus précisément, après une présentation des approches majeures en segmentation et en sur-segmentation, cf. chapitre 8, nous présentons une stratégie de segmentation de scènes urbaines, cf. chapitre 9, une méthode de détection de fissures de chaussée, cf. chapitre 10 et, enfin une approche de segmentation d'insectes, cf. chapitre 11.

Enfin, dans une dernière partie nous présentons les perspectives de recherche envisagées. Tout d'abord dans la continuité des travaux de recherche qui n'ont pas fait l'objet d'une présentation détaillée dans ce manuscrit, sur l'utilisation de marqueurs artificiels pour estimer la pose d'un objet ainsi que des travaux de recherche sur l'estimation du squelette 2D et 3D de formes, à partir de multiples images. Nous nous concentrerons ensuite sur la suite à donner aux travaux amorcés pour l'analyse de vidéos multiples et les liens potentiels avec des développements applicatifs. Puis, en lien avec la première partie, nous introduisons les travaux que nous envisageons, dans la continuité et la généralisation de la mise en correspondance 2D/3D, notamment, en prenant en compte des données de nature plus variée et en suivant une analogie à la bande dessinée. Ensuite, en lien avec la seconde partie, nous présentons les pistes de recherche liées à la reconstruction 3D en abordant la notion de prochaine vue complémentaire, et enfin, en développant diverses pistes de recherches au sujet de la mise en correspondance multiple que nous souhaitons orienter vers des outils plus modernes (superpixels et apprentissage profond).

Première partie

Détection, caractérisation et appariement de primitives 2D et 3D

Introduction à la détection de primitives dans différentes modalités

La détection de primitives en 2D, comme par exemple, dans des images ou des cartes de profondeur, ainsi que la détection de primitives en 3D, dans des images 3D ou des modèles surfaciques, est essentielle, dans de nombreux domaines, afin de détecter, reconnaître, suivre et segmenter. Du fait de cette multitude d'utilisations, la littérature est très abondante sur ce sujet et il est important d'en distinguer les différentes tendances afin d'être en mesure, d'une part, de faire les choix adaptés à la situation, lorsque cette détection est une simple étape dans un processus plus complexe, et, d'autre part, d'en identifier les limites, pour déterminer les pistes de recherche.

L'objectif de cette partie est, dans un premier temps, de faire un état des lieux de ce qui existe en 2D et en 3D, en comparant et en évaluant les différents détecteurs existants. Cet état des lieux nous permettra d'introduire une première contribution bibliographique sur l'évaluation de la complémentarité des détecteurs que nous avons réalisé en 2D. Nous présenterons, ensuite, la proposition d'un détecteur de points d'intérêt dans des images 3D médicales.



FIGURE 1.2. – Reconnaître un objet en 2D à partir d'un modèle 3D – Nous supposons qu'un utilisateur fait l'acquisition d'une image d'un objet qui pose problème. Nous obtenons donc une image 2D d'un objet connu en 3D (nous faisons l'hypothèse que nous possédons une collection de modèles 3D). Notre question : comment reconnaître l'objet en 2D inconnu et photographié dans des conditions très différentes de l'objet connu en 3D ? Dans cette problématique, nous devons faire face aux difficultés suivantes : présence d'une texture, d'un éclairage et d'un environnement perturbé. Nous pouvons également noter qu'il peut y avoir une dégradation de l'objet photographié par rapport à son état d'origine.

L'intérêt de cet état des lieux est de mettre en évidence le fait qu'il y a peu d'outils de détection qui ont été développés dans le but de mettre en correspondance des données 2D avec des données 3D. Pourtant, de tels outils semblent indispensables au vu des données 2D et 3D qu'il est possible d'acquérir aujourd'hui. Pour illustrer ces propos, nous reprenons l'exemple de l'application que nous avons envisagée pour le projet Mobville dont les différents éléments sont présentés dans la figure 1.2. Nous souhaitons être capable de reconnaître automatiquement des objets afin de qualifier leur état (niveau de dégradation). Pour cela, nous supposons que nous avons deux types de données disponibles :

d'une part, une image 2D de l'objet à reconnaître et à qualifier et, d'autre part, une base de données de modèles 3D de divers objets, en lien avec le type de scène acquise. Dans l'exemple, il s'agit d'une scène urbaine et nous supposons avoir une base de données de modèles 3D de mobiliers urbains.

Ainsi, le dernier chapitre de cette partie nous permettra d'introduire notre contribution dans ce domaine de recalage 2D/3D : l'utilisation d'un détecteur 3D s'appuyant sur la notion de saillance curviligne et d'un détecteur 2D utilisant la même notion mais avec une approche multi-échelle et l'introduction du principe des cartes de *focus*, lié au plan de mise au point ou plan focal. De plus, nous avons mis en place un protocole d'évaluation permettant de qualifier les performances d'un détecteur appliqué en 2D et en 3D. Nous avons validé l'utilisation conjointe des deux détecteurs 2D et 3D en les utilisant pour estimer la pose d'objets.

Contributions scientifiques structurantes en détection et caractérisation 2D et 3D

Dans le chapitre 2 de cette partie, toute la bibliographie présentée constitue une partie de l'Unité d'enseignement Vision, Réalité Augmentée et Applications, proposée en troisième année aux étudiants de l'ENSEEIHT, dans le parcours Multimédia [Chambon 15b]. Plus précisément, cela fait partie des deux classes inversées, des 4 travaux pratiques et de l'apprentissage par projet transversal réalisés en 18 séances.
À la fin de ce premier chapitre nous introduisons les contributions présentées dans la publication [Gales 10b], liées à la thèse de Guillaume Galès [Gales 11], co-encadrement avec Alain Crouzil, IRIT-TCI, sous la direction de Patrice Dalle. Enfin, nous précisons que la contribution dans le cadre de l'extraction de points d'intérêt pour la détection de fissures [Chambon 11a] correspond à ce chapitre, mais, nous ne détaillerons pas ce travail.
Le chapitre 3 permet d'introduire les contributions en imagerie médicale avec les publications suivantes : [Chambon 07a, Chambon 07b, Chambon 11e] et au travers du projet ANR MARIO – Modélisation de l'Anatomie normale et pathologique pour le Recalage non linéaire entre Images CT et TEP en Oncologie, pendant mon post-doctorat à Télécom Paris, de septembre 2006 à novembre 2007, sous la supervision d'Isabelle Bloch et en collaboration avec un doctorant Antonio Moreno Ingelmo. Ce travail m'a également permis d'être en collaboration avec l'Université d'Orlando aux États-Unis, notamment avec Anand Santhanam et Jannick Rolland.
Enfin, le chapitre 4 présente la contribution principale de cette partie sur la mise en correspondance 2D/3D et qui a été publiée dans : [Rashwan 16, Rashwan 17, Rashwan 19a]. Ce travail a été réalisé en colloboration avec le post-doctorant Hatem Rashwan, financé par le projet régional « Mob-Ville – Application mobile pour la signalisation des alertes urbaines » porté par la société `Imajing`. Actuellement, Hatem Rashwan est chercheur à l'Université de Tarragone, Universitat Virgili in Rovira, et nous continuons à travailler ensemble, ce qui a permis une publication en lien avec ces travaux [Abdulwahab 19]. Ce travail a été réalisé avec la collaboration de mes collègues au sein de l'équipe IRIT-REVA : Pierre Gurdjos, Géraldine Morin et Vincent Charvillat.

Chapitre 2.

Primitives en 2D

2.1. Introduction sur les primitives en 2D

Qu'est ce que c'est ?

Nous appelons une **primitive** un élément pouvant être extrait d'une image afin d'obtenir une information caractéristique de la scène comme la présence d'un objet, une distance, un relief (c'est-à-dire une altitude ou une profondeur). Ces primitives peuvent être des pixels, des régions, des contours, des polygones, tout ensemble de points permettant d'obtenir l'information recherchée. Dans la suite, nous étudierons particulièrement la notion de **primitive d'intérêt**, pour ensuite se consacrer à la notion de **points d'intérêt**.

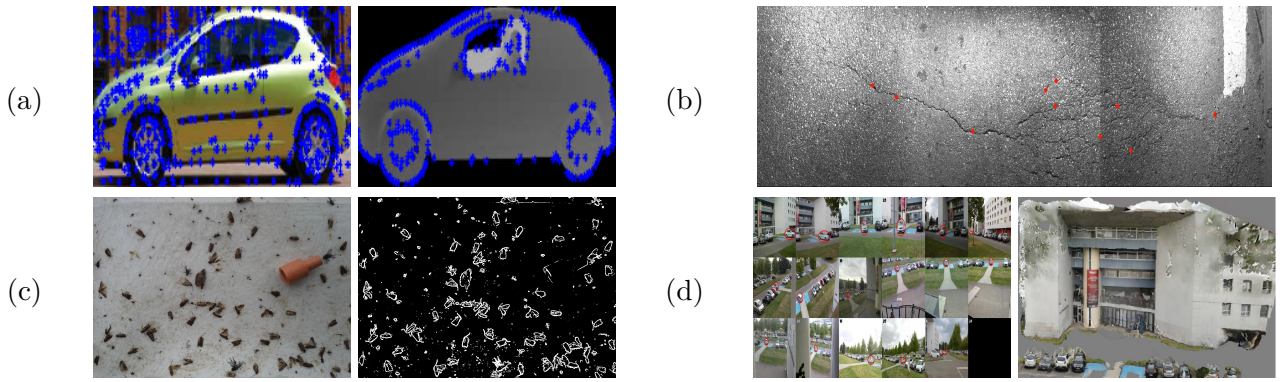


FIGURE 2.1. – Exemples de travaux de recherche sur les primitives d'intérêt en 2D – En (a), nous cherchons à obtenir des points d'intérêt répétables entre une image et une carte de profondeur. Pour cela nous exploitons les maximums de courbure dans de multiples échelles [Rashwan 16]. Ce travail sera détaillé dans le chapitre 4. En (b), nous recherchons des points qui possèdent une corrélation forte dans une direction donnée afin de mettre en évidence des points d'intérêt (en rouge) appartenant à une fissure de la chaussée [Chambon 11a]. Ce travail ne sera pas développé dans ce manuscrit mais, en revanche, le travail le plus saillant sur cette thématique, la détection de fissures de chaussée, sera repris dans le chapitre 10. En (c), nous caractérisons des points d'intérêt sur des contours (en blanc) afin de détecter des insectes [Bakkay 18a]. Ce travail sera détaillé dans le chapitre 11. En (d), nous avons étudié la possibilité d'effectuer une reconstruction 3D à partir de multiples caméras sur un jeu de données publié dans [Malon 18]. Nous reprenons ces travaux dans le chapitre 7.

Une **primitive d'intérêt**, cf. exemples des figures 2.1 et 2.2, est donc un ensemble de points sur une image qui correspond à la projection de points/zones particulières d'une scène, d'un objet étudié : coins, jonctions, variation d'intensité, de texture, forme particulière, objet particulier. Il s'agit donc de

primitives ayant des caractéristiques particulières : fort contraste, texture forte, couleurs non corrélées au reste de la scène. Ces caractéristiques particulières doivent permettre la reconnaissance, et donc le suivi, non ambigu par comparaison avec les autres entités de la scène.

(e) (f) (g)

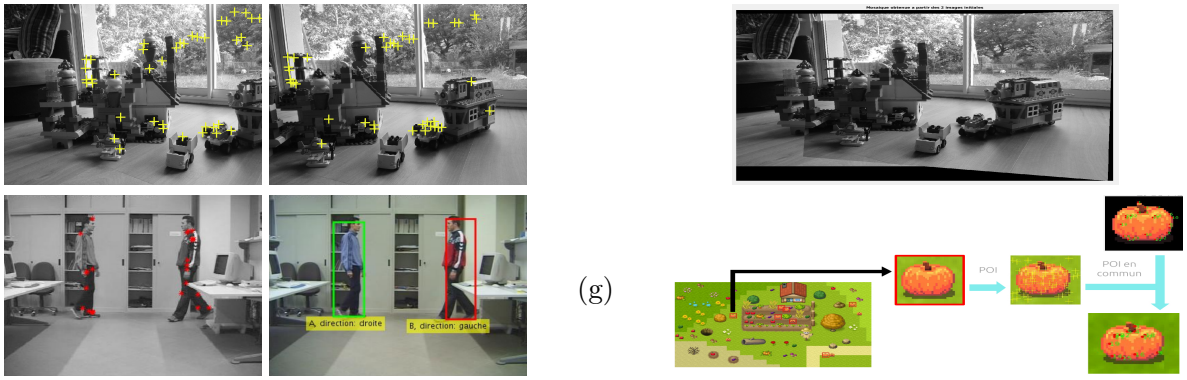


FIGURE 2.2. – Exemples de projets d'enseignement sur les primitives d'intérêt en 2D – En (e), dans le cadre de travaux pratiques proposés à l'ENSEEIHT, nous détectons des points d'intérêt dans un couple d'images (croix jaunes) afin de construire un panorama de la scène (à droite). En (f), pour un bureau d'étude proposé à l'ENSEEIHT, une solution consiste à détecter des points d'intérêt (en rouge) afin de suivre les personnes en mouvement. Le résultat est présenté sous la forme de boîtes englobantes indiquant la direction du déplacement (à droite). En (g), au cours d'un Apprentissage par Problème et Projet (APP), les étudiants sont amenés à détecter des points d'intérêt afin de reconnaître des éléments dans un jeu vidéo.

Pourquoi ?

Détecter des points d'intérêt puis caractériser des points, d'intérêt ou non, est une tâche que nous retrouvons dans de très nombreuses applications en vision par ordinateur. En effet, utiliser des points caractéristiques d'un objet que nous visualisons permet de le reconnaître dans un premier temps, mais également, de caractériser, dans un second temps, ce qui le distingue des autres points de la scène. Par exemple, il peut être nécessaire de comprendre comment cet objet est différent des objets de la même catégorie ou quelles sont les qualités intéressantes qu'il possède pour extraire des informations, ou au contraire, de quelle manière il est dégradé par rapport à d'autres objets de la même classe. Ces analyses nécessitent la plupart du temps une étape de caractérisation suivie d'une étape de mise en correspondance. Nous pouvons citer de nombreuses applications existantes, illustrées dans les figures 2.1 et 2.2, comme :

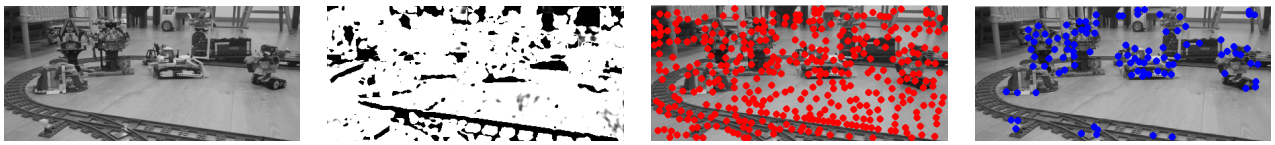
- La reconnaissance de formes (reconnaître des points de fortes courbures pour détecter des objets particuliers, comme un visage [Wang 10], ou de manière plus générale, une personne, ou encore des objets [Rashwan 17]), cf. figure (a) ;
- La reconnaissance de défauts sur des objets (reconnaître des points particuliers pour identifier un défaut [Amhaz 16]), cf. figure (b) ;
- La segmentation (reconnaître des points contours pour délimiter des régions, comme dans [Bakkay 18a]), cf. figure (c) ;
- La mise en correspondance [Gales 10b] et plus généralement la reconstruction 3D (estimer un vecteur de déplacement grâce à la connaissance des correspondances de points, sous certaines hypothèses et calculs, et être capable de reconstruire le relief d'une scène) [Agarwal 11], cf. figure (d) ;
- La construction de mosaïque d'images ou plus simplement de panorama (mise en correspondance de points caractéristiques qui possèdent des caractéristiques discriminantes par rapport aux

autres points de la scène et sont donc plus faciles et fiables à mettre en correspondance) [Shum 98, Capel 04], cf. figure (e) ;
- Le suivi d'objets (reconnaître des points caractéristiques d'un objet particulier, comme un piéton, pour le suivre) [Gauglitz 11], cf. figure (f) ;
- L'indexation (reconnaître des régions avec des caractéristiques particulières de couleur et de forme pour indexer) [Bres 99], cf. figure (g).

Dans l'énoncé même de ces applications, nous comprenons que la détection ne peut se passer de la caractérisation, afin d'être par la suite exploitée, par exemple pour de la reconnaissance. C'est la raison pour laquelle, ce chapitre regroupe ces deux aspects.

Caractériser une primitive, c'est trouver des caractéristiques qui permettront de la reconnaître, de l'apparier, de la suivre et cette étape est aussi cruciale que la détection. Traditionnellement, détection et caractérisation sont distinguées l'une de l'autre, c'est le cas pour le détecteur de Harris [Harris 88], très largement présent dans la littérature. Mais, depuis ces deux dernières décennies, et notamment depuis la célèbre publication de [Lowe 04], la littérature abonde de méthodes permettant à la fois de détecter et de caractériser.

Comment ?



(a) (b) (c) (d)

FIGURE 2.3. – Étapes classiques du fonctionnement d'un détecteur – À partir de l'image originale, en (a), nous calculons une réponse, en (b). Plus le pixel est sombre, plus la réponse est élevée. En (c), nous montrons la réponse obtenue (pastilles rouges) après suppression des non-maxima locaux. Ici, nous avons choisi un voisinage de 9×9. Enfin, en (d), il s'agit des points sélectionnés (pastilles bleues) après un seuillage. Dans cet exemple, nous avons choisi de prendre les 100 premiers points.

En général, un détecteur de points d'intérêt suit les étapes suivantes, cf. figure 2.3, pour chaque point $\mathbf{p}$ de coordonnées (x, y) :

(1) **Calcul de la réponse R d'un détecteur, cf. figure 2.3.(b)** – Plus précisément, R possède le profil suivant :

$$R : \mathbb{N}^2 \to \mathbb{R}^n.$$

Nous utilisons n car certains détecteurs donnent une réponse en dimension n, notamment les détecteurs multi-échelle.

(2) **Suppression des non-maxima locaux, cf. figure 2.3.(c)** – Cette étape est réalisée en prenant en compte un certain voisinage $\mathcal{V}$ à déterminer. La fonction réponse est modifiée de la manière suivante :

$$R_{\text{modifie}}(x, y) = \begin{cases} R_{\max}(x, y) & \text{si } R(x, y) = R_{\max}(x, y) \\ 0 & \text{sinon.} \end{cases} \tag{2.1}$$

La réponse $R_{\max}(x, y)$ est la réponse maximale, suivant un critère à définir, dans le voisinage $\mathcal{V}(\mathbf{p})$. Plus de détails sur les possibilités de choix sur ce critère sont donnés dans [Harzallah 11, page 18]. Cette étape est importante pour tenter d'éviter les amas de points d'intérêt mais elle est optionnelle et n'est donc pas toujours réalisée.

(3) **Sélection des `Nb` points maximaux, cf. figure 2.3.(d)** – Les points **p** pour lesquels $T(R(x,y))$ est vraie sont conservés. La plupart du temps $T(R(x,y))$ correspond au test suivant : $R(x,y) > S$ avec S, un seuil à fixer. Fixer ce seuil, quelle que soit l'approche choisie, reste toujours délicat. Plutôt que d'utiliser un seuil sur la réponse qui n'est pas toujours évident à fixer intuitivement, dans la littérature, un seuil sur le nombre de points souhaités est habituellement utilisé.

(4) **Affinement de la détection** – Il peut s'agir d'une détection plus fine (sous-pixellique, c'est le cas pour [Lowe 04]) ou d'une suppression des points erronés (des points qui ne sont plus considérés comme des points d'intérêt). Cette étape est également optionnelle.

Les étapes (2) et (3) ne sont pas obligatoirement réalisées toutes les deux. En effet, nous pouvons imaginer les stratégies suivantes :

- Nous appliquons uniquement l'étape (2) :
  - L'avantage est qu'il y aura des points bien répartis sur toute l'image.
  - L'inconvénient est que certains points n'auront pas des caractéristiques suffisamment saillantes pour être suivis correctement et il y a un risque élevé d'erreurs de suivi.
- Nous appliquons uniquement l'étape (3) :
  - L'avantage est qu'il y aura des points plutôt faciles à suivre car avec des caractéristiques très discriminantes.
  - L'inconvénient est que certaines zones de l'image ne posséderont aucun point suivi.

Qualité du résultat ?

Dans la littérature, de nombreuses approches ont été mises en œuvre et il est évident qu'il est nécessaire d'avoir des critères de comparaison pour évaluer la qualité des résultats obtenus. Ainsi, un premier aspect important lorsque il est nécessaire d'apparier des primitives, et en particulier des primitives d'intérêt, est la propriété de *répétabilité*. La définition que nous donnons ici, similaire à celle de [Szeliski 10], est la suivante :

> La répétabilité d'une primitive correspond à sa capacité à être détectée quelle que soit l'image ou la donnée dans laquelle elle apparaît. Plus précisément, si une primitive est détectée dans une représentation donnée, alors, elle doit également être détectée dans une autre représentation.

Une définition formelle de cette propriété sera donnée dans notre étude de la complémentarité des détecteurs en 2D, cf. § 2.3.1. Il est également souhaité que la détection soit exacte, précise. En effet, souvent, la localisation est imprécise, cela est dû aux différents opérateurs de lissage ou d'approximation utilisés. C'est pour cela que de nombreuses méthodes proposent une étape de **relocalisation**.

Il peut être nécessaire d'être robuste à certaines difficultés (cette nécessité est souvent liée à la notion de suivi). La plupart du temps ces propriétés de robustesse sont à rapprocher de l'étape d'appariement abordée dans la partie II. Dans [Tuytelaars 06], les auteurs résument ainsi parfaitement les propriétés attendues pour un détecteur « idéal »:

- Sa définition doit être locale afin de lui permettre d'être robuste aux occultations.
- Il doit être invariant, ou co-variant, aux transformations d'images, ce qui signifie qu'une transformation appliquée à l'image n'affectera pas le résultat du détecteur, ou l'affectera de manière consistante avec la transformation appliquée.
- Il doit être robuste à tous types de dégradations de l'image : bruits, flou, discrétisation ou encore compression.
- Une grande quantité de primitives d'intérêt doit être trouvée quelque soit l'objet, même s'il est de petite taille.
- La détection doit être rapide, c'est-à-dire proche du traitement en temps réel.

Bien évidemment, les auteurs ajoutent à ces caractéristiques celles que nous avons déjà énoncées : précision et répétabilité. Cette dernière notion est plutôt définie comme la capacité à être discriminant,

c'est-à-dire la capacité du détecteur à proposer des primitives qui peuvent être appariées avec le moins d'ambiguïté possible à travers tout type de représentation de cette primitive dans une large base de données. Il est intéressant de souligner la contradiction entre le fait d'être répétable et discriminant, tout en étant robuste à des transformations, cela nous permet de mettre en évidence la difficulté de cette tâche de détection.

Pour terminer sur cet aspect, des publications ont tenté d'évaluer et comparer plus finement les performances des différents détecteurs, comme les articles de référence suivants : [Mikolajczyk 04, Aanæs 12]. Nous reviendrons sur ces publications par la suite, dans notre étude de la complémentarité des détecteurs en 2D, § 2.3, page 23.

Et le descripteur ?

Entre 1980 et 2000, ce sont essentiellement des détecteurs qui ont été proposés et très peu de descripteurs. Le plus connu est le descripteur associé au détecteur SIFT, *Scale Invariant Feature Transform*, mais nous effectuons une présentation des descripteurs principaux en annexe, cf. A.2. De plus, la plupart du temps, la manière de calculer la différence entre deux descripteurs est une simple distance euclidienne ou la corrélation croisée centrée normalisée ou encore une distance de Mahalanobis. Il est possible d'utiliser bien d'autres distances, cf. partie II, mais cet aspect n'a pas été beaucoup investi dans la littérature.

2.2. Approches de détection et de caractérisation des points d'intérêt en 2D

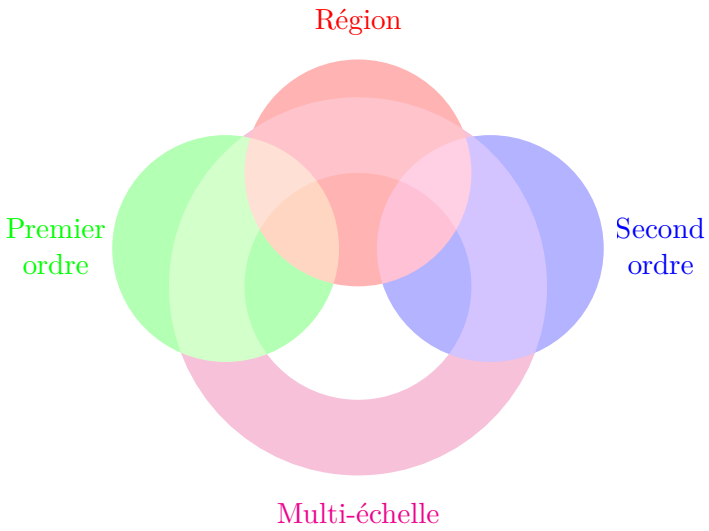


FIGURE 2.4. – Classement des approches de détection de points d'intérêt en 2D – Ce graphique nous permet d'illustrer les interactions entre les différents éléments que nous avons identifiés. Ainsi, nous choisissons de distinguer trois familles (les trois cercles) mais ce qui relie ces familles c'est la généralisation de certains opérateurs à la multirésolution (l'anneau). Les intersections mettent en évidence qu'il existe bien des approches qui peuvent être classées dans plusieurs familles et que, dans la table 2.1, nous avons fait un choix. À notre connaissance, aucune approche ne mélange approche du premier et du second ordre.

La littérature est très abondante sur le sujet de la détection de primitives d'intérêt en 2D et il est délicat d'établir une classification des approches. En effet, parmi toutes les approches existantes, il

est possible d'effectuer un tri en distinguant les approches locales (prise en compte d'un voisinage), des approches globales (comme les approches utilisant les outils de transformation de Fourier). Nous pouvons également distinguer les primitives photométriques, qui s'appuient sur l'étude de l'intensité ou de la texture, des primitives géométriques, qui prennent en compte des notions plus proches de la géométrie, comme les notions de segment ou d'intersection. Enfin, un critère important pour caractériser les différentes approches existantes est de considérer l'aspect multi-résolution ou multi-échelle, c'est-à-dire, est-ce que l'approche proposée fait intervenir différentes résolutions des images ou des objets manipulés, ou est-ce que l'approche fait intervenir différentes échelles dans les outils utilisés, comme différentes tailles de filtrage. Enfin, les acquisitions actuelles, aussi bien en 2D qu'en 3D permettent la collecte de nombreux paramètres liés aux conditions d'acquisition, ainsi que de nombreuses informations extraites en pré-traitement, c'est ce que nous appelons de manière générale des métadonnées. Par exemple, il est possible d'obtenir la date, l'heure, les dimensions, la focale, la vitesse d'obturation, les coordonnées GPS ou encore le modèle de l'appareil. Toutes ces informations sont surtout utiles pour positionner et calibrer. Nous pouvons également citer toutes les applications ajoutées qui permettent, par exemple, la détection automatique des visages, des yeux, des zones de flou, une première indexation de l'image (comme intérieur, extérieur, jour, nuit) [Hu 15]. Parmi toutes ces possibilités, nous avons choisi d'extraire les aspects que nous avons jugés importants, c'est-à-dire les plus exprimés et les plus utilisés dans la littérature. C'est ce que nous détaillons dans les paragraphes suivants et que nous résumons dans la figure 2.4.

2.2.1. Approches basées premier ordre

Les approches s'appuyant sur une détection de contours [Canny 86], c'est-à-dire exploitant les dérivées premières de l'image, sont les approches fondatrices de cette discipline. Nous les appellerons les approches de premier ordre. L'approche la plus célèbre consiste à exploiter le détecteur de Harris [Harris 88], cf. page 236, qui permet de détecter plus particulièrement des coins. Le principe est d'estimer à quel point un point étudié est corrélé à ses voisins, en s'appuyant sur la publication initiale de [Moravec 80], cf. page 236.

2.2.2. Approches basées région

Ce genre de technique propose de définir une région caractéristique autour du point étudié et, ensuite, d'étudier le comportement de cette région, afin de déterminer le caractère unique du point étudié. Un point d'intérêt ainsi obtenu est très différent car il correspond au centre d'une région d'intérêt et non à un point caractéristique sur le contour d'une forme donnée. Dans [Tuytelaars 06], les auteurs insistent sur le fait que ce type de détecteur correspond à une nouvelle manière de décrire un point d'intérêt : il ne s'agit plus de juste sélectionner des lieux caractéristiques dans une image pour en extraire l'information pertinente pour faire des calculs ou tout simplement pour réduire les temps de calcul mais il s'agit de trouver une nouvelle façon de représenter l'image permettant d'en décrire les objets, sans pour autant s'appuyer sur une segmentation. Ainsi, de nombreuses versions de détecteurs locaux s'appuyant sur l'étude de région d'intérêt autour d'un point, parfois appelées *blob*, ont été introduites. L'approche la plus connue est celle de SUSAN, *Smallest Univalue Segment Assimilating Nucleus* [Smith 97], cf. § A.1.4, qui étudie la proportion de pixels photométriquement proches du pixel étudié dans une région circulaire autour de lui. Une approche similaire mais plus simple à implémenter et permettant une accélération des calculs est FAST [Rosten 06], cf. § A.1.4.

Dans [Matas 02], les auteurs considèrent simplement les composantes connexes d'une image seuillée. Ils appelent ce détecteur MSER, *Maximally Stable Extremal Regions*, cf. § A.1.4 pour plus de détails. Pour le détecteur IBR, *Intensity Based Regions* [Tuytelaars 04], les auteurs introduisent la notion de blob. Les régions d'intérêt détectées autour des points prennent en compte les maxima le long de droites partant du point étudié. L'approche se termine en calculant une approximation par une ellipse de la région formée par tous ces maxima locaux, cf. § A.1.4. Enfin, une dernière approche est celle

de [Deng 07], PCBR, *Principal Curvature-Based Regions*. Elle est introduite comme une approche permettant d'étudier des régions où des contours sont détectés par ligne de partage des eaux. Cette approche s'appuie également sur le calcul de la courbure principale, c'est pourquoi, pour nous, elle appartient également à la dernière catégorie d'approches que nous distinguons.

2.2.3. Approches basées second ordre

Les approches par détection de la courbure sont les approches de second ordre puisqu'elles exploitent les dérivées secondes. Un des détecteurs les plus anciens est celui de Beaudet, appelé aussi le détecteur Hessien [Beaudet 78], décrit dans le § A.1.5. Ce détecteur estime la courbure de la surface alors que le détecteur de Kitchen et Rosenfeld [Kitchen 82], cf. § A.1.6, utilise la notion de courbure du contour passant par le point considéré. De plus, chaque point utilisé est pondéré par la norme du gradient car les auteurs souhaitent prendre en compte les cas où la norme du gradient est faible (donc un contour peu contrasté, qui peut correspondre à un bruit). Ce détecteur s'appuie sur les dérivées secondes et donc sur la notion de courbure de la surface. Plus récemment, les auteurs de [Fischer 14] ont proposé un détecteur s'appuyant sur la courbure, notée κ. Cette courbure est exprimée comme le changement, noté q, de gradient image le long de la tangente. Ce terme q est donc une approximation de κ. Dans la publication de [Deng 07], les auteurs s'appuient sur la valeur propre minimale (qui met en valeur une frontière sombre sur un fond clair) ou maximale (qui met en valeur une frontière claire sur un fond sombre) de la matrice Hessienne afin d'estimer la courbure principale dans un contexte de multi-résolution, cf. § A.1.4.

2.2.4. Introduction des notions d'analyse multi-échelle et multi-résolution



Figure 2.5. – Illustration de l'introduction d'une analyse multi-échelle dans une pyramide d'images (multi-résolution) – Ces illustrations sont extraites de la thèse de Guillaume Galès [Gales 11]. De haut en bas, la résolution des images est diminuée, formant ainsi une pyramide d'images. De gauche à droite un filtrage gaussien est appliqué avec un filtre de plus en plus grand, pour fournir une analyse multi-échelle.

Plus récemment, de nombreuses approches ont été introduites afin d'avoir un détecteur multi-

échelle ou multi-résolution. Ici, nous faisons le choix d'appeler l'échelle le niveau de détails utilisé pour observer/analyser une image, une scène, un ensemble de données, alors que la résolution est liée à la taille de l'image utilisée, cf. figure 2.5. Suivant l'échelle d'analyse choisie, les difficultés suivantes peuvent apparaître : avec une échelle trop faible, des détails non significatifs (un motif de surface) peuvent être pris en compte et produire des détections non pertinentes alors qu'avec une échelle trop importante, des détails pertinents de la scène mais de petite taille peuvent être ignorés. De plus, dans le cas du suivi de primitives, nous pouvons observer des changements de résolution entre images/objets comparés. Pour toutes ces raisons, le célèbre détecteur SIFT, Scale Invariant Features Transform [Lowe 04], a été introduit. Par la suite, le raisonnement multi-échelle a été généralisé à d'autres détecteurs connus, comme celui de Harris [Mikolajczyk 04]. Dans de nombreux cas, cette capacité à s'adapter aux différentes échelles va se traduire par le fait de faire varier un (ou plusieurs) paramètre(s) utilisés par le détecteur. Par exemple, dans le cas du détecteur SIFT, c'est la taille du filtre gaussien utilisé pour le lissage et pour le calcul du laplacien qui va varier. Cette variation permet de mettre en évidence les détails plus ou moins fins des données étudiées. L'inconvénient principal de ces approches étant le temps de calcul, des approches mettant en œuvre des astuces algorithmiques ont été introduites, comme celle de [Bay 08] avec le détecteur SURF, *Speeded Up Robust Features*. Encore une fois, ces approches s'appuient essentiellement sur la détection des variations de texture et d'intensité.

Les détecteurs mono-échelle utilisent une réponse associée à un lissage dépendant d'un paramètre : l'échelle. Ils détectent les points ou régions à une échelle unique. Cette échelle étant *a priori* inconnue, les approches multi-échelles cherchent à s'affranchir de cette difficulté en effectuant une analyse sur de multiples échelles et à fusionner les résultats obtenus pour fournir une réponse multi-échelle. Dans la littérature, il a été démontré que le noyau gaussien est le seul lissage linéaire permettant une analyse multi-échelle correcte [Yu 11]. Ainsi, les détecteurs actuels construisent une pyramide gaussienne, calculent une mesure d'intérêt spatiale normalisée pour chaque pixel et pour chaque niveau de la pyramide, puis détectent les extrema localisés en x, y, σ dans cet espace-échelle 3D. Le lecteur intéressé pourra trouver les manières les plus classiques de construire une pyramide d'images, dans l'annexe A.1.1.

Ainsi, l'opérateur de premier ordre que nous avons déjà introduit a été généralisé au multi-échelle par [Mikolajczyk 04] pour donner l'approche Harris-Laplace, cf. § A.1.3.

De même, en ce qui concerne les détecteurs basés régions, [Tuytelaars 04] a proposé une version multi-échelle avec l'opérateur EBR, *Edge-Based Regions*, qui étudie les familles de parallélogrammes définis à partir du point étudié et des points de contours passant par le point étudié, cf. § A.1.4.

Encore une fois, les auteurs de [Mikolajczyk 04] ont généralisé l'opérateur Hessien, du second ordre, au multi-échelle, Hessien-Laplace, cf. § A.1.5.

La méthode SIFT, *Scale Invariant Feature Transform* [Lowe 04], cf. § A.1.6, s'appuie sur la détection des extrema en échelle et en espace du laplacien. Les auteurs utilisent des différences de gaussiennes pour approximer le laplacien et ils effectuent les calculs avec différentes tailles (échelles) de filtre dans différentes résolutions (utilisation d'une pyramide d'images). La particularité de ce détecteur est que les auteurs ont proposé l'utilisation d'un descripteur associé qui considère l'orientation principale qui correspond au gradient dominant dans le voisinage du point considéré. Le descripteur associé à ce détecteur prend ainsi en compte un histogramme local de ces orientations. Il existe une variante permettant de diminuer les temps de calcul : SURF, *Speeded Up Robust Features* [Bay 08], cf. § A.1.6. Enfin, le dernier détecteur de second ordre multi-échelle est CSS, *Curvature Scale Space* [Mokhtarian 98], cf. § A.1.6, qui recherche les extrema de la courbure le long des contours en utilisant une approche multi-échelle.

2.2.5. Classement des publications significatives et invariance des détecteurs

La table 2.1 nous permet de reprendre les différents détecteurs considérés en fonction de leur catégorie. De plus, la table 2.2 permet de résumer toutes les invariances que nous avons répertoriées dans la littérature sur tous les détecteurs étudiés. Toutes ces approches sont détaillées dans l'annexe A, sachant que c'est dans la dernière catégorie : second ordre avec multi-échelle, que nous présenterons des contributions.

	Premier ordre	**Région**	**Second ordre**
Mono-échelle	HA [Moravec 80, Harris 88]	SUSAN [Smith 97] FAST [Rosten 06] MSER [Matas 02] IBR [Tuytelaars 04]	BE [Beaudet 78] KR [Kitchen 82]
Multi-échelle	Harris-laplace [Mikolajczyk 04]	KA [Kadir 04] EBR [Tuytelaars 04]	Hessien-Laplace [Mikolajczyk 04] SIFT [Lowe 04] SURF [Bay 08] CSS [Mokhtarian 98] *MFC* [Rashwan 17]
		PCBR [Deng 07]	

Table 2.1. – Publications significatives sur la détection de points d'intérêt en 2D.

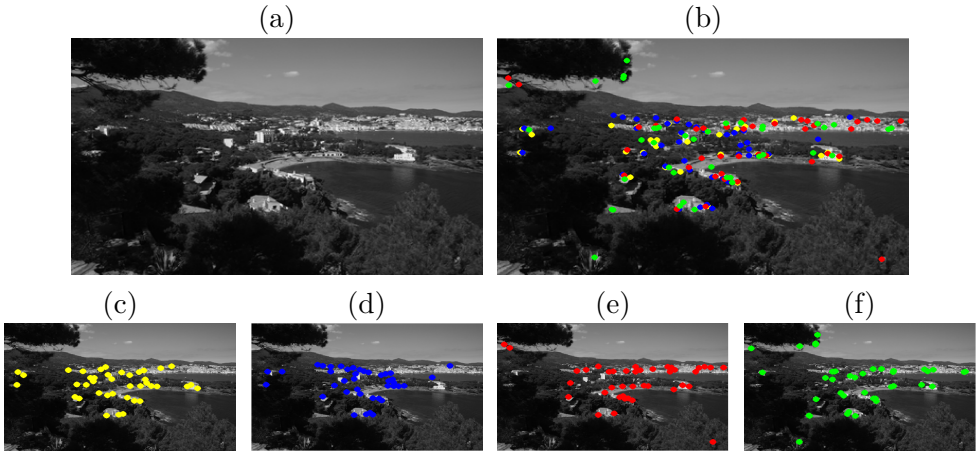


Figure 2.6. – Résultats obtenus avec quatre détecteurs de points d'intérêt en 2D – En (a), nous présentons l'image originale étudiée et en (b), l'ensemble des points d'intérêt détectés. Nous pouvons observer plus distinctement les différences avec les résultats présentés séparément, en (c) avec Harris (pastilles jaunes), en (d), avec le détecteur de Beaudet (pastilles bleues), en (e) avec le détecteur de Kitchen et Rosenfeld (pastilles rouges) et en (f), le détecteur SURF (pastilles vertes). Il s'agit du même code couleur utilisé en (b). Il faut noter que les résultats présentés sont ceux obtenus pendant les travaux pratiques et le projet réalisés aux cours des enseignements de spécialité à l'ENSEEIHT (les étudiants codent les 3 premiers détecteurs et réutilisent celui codé par `Matlab`, pour le dernier).

Dans la figure 2.6, nous présentons le résultat de détection de points d'intérêt avec quatre détecteurs différents. Nous pouvons observer que les points obtenus sont assez complémentaires les uns des autres. Nous comprenons ainsi la difficulté pour choisir un détecteur et également l'intérêt qu'il peut y avoir

à combiner différentes catégories de détecteurs.

Ainsi, nous avons présenté notre manière de classer et de comparer les détecteurs en fonction de leur définition et de leurs invariances. Le lecteur intéressé pourra trouver encore plus de détails sur les détecteurs présentés, dans l'annexe A et également dans la thèse de Guillaume Galès, en particulier une page dédiée à la comparaison des paramètres utilisés [Gales 11, pages 100-101]. Dans les travaux que nous avons menés, nous avions besoin de points d'intérêt et en particulier d'appariements de points d'intérêt. Nous pouvons tenter de déterminer quel détecteur est le plus adapté aux différents problèmes que nous nous posons, mais, nous avons pensé que nous aurions plus de recul sur l'existant si nous nous attachions à étudier la complémentarité des différents détecteurs existants. C'est la raison pour laquelle, une première étude proposée a consisté à qualifier la complémentarité des détecteurs existants.

Détecteur	Affines photométriques	Géométriques non affine	Affines géométriques	Changement d'échelle
[Moravec 80, Harris 88]				
Harris-laplace [Mikolajczyk 04]				×
SUSAN [Smith 97]				
FAST [Rosten 06]				
MSER [Matas 02]	×	×		
IBR [Tuytelaars 04]	×	×	×	
KA [Kadir 04]	×	×	×	×
EBR [Tuytelaars 04]	×	×	×	×
PCBR [Deng 07]				×
[Beaudet 78] [Kitchen 82]				
Hessien-Laplace [Mikolajczyk 04]				×
SIFT [Lowe 04]		×		×
SURF [Bay 08]				×
CSS [Mokhtarian 98]	×	×	×	×
MFC [Rashwan 17]				×

Table 2.2. – Invariance des différents détecteurs présentés – Dans ce tableau, nous répertorions les invariances annoncées par les auteurs des différentes publications. Le détecteur MCF [Rashwan 17] est la contribution apportée au domaine et il sera détaillé dans le chapitre 4.

2.3. Étude de la complémentarité des détecteurs

Il existe quelques travaux qui ont proposé des études comparatives de détecteurs et un travail de référence sur la comparaison des descripteurs [Mikolajczyk 05]. Nous citons ici les travaux qui nous semblent les plus complets et les plus proches de ce que nous présentons dans ce manuscrit : la comparaison des détections. Nous nous intéressons essentiellement aux critères utilisés.

Dans la littérature, le critère le plus utilisé pour évaluer et comparer des détecteurs est le critère de répétabilité, cf. définition en page 16. Pour qualifier la robustesse d'un détecteur face à une transformation d'une image à une autre (rotation, changement d'échelle, transformation perspective, changement d'illumination), il est également possible d'utiliser ce critère, comme dans [Schmid 00, Mikolajczyk 04]. Toutefois, il existe des travaux plus complets qui évaluent et comparent les détecteurs en examinant les critères suivants : le respect de la contrainte épipolaire, la fiabilité de la reconstruction 3D, une détection fiable de l'échelle [Aanæs 12].

Nous nous sommes particulièrement intéressés à ce que nous appelons la complémentarité, comme définie dans [Mikolajczyk 05]. Dans ces travaux, la complémentarité entre différents détecteurs est évaluée en effectuant au préalable une segmentation de l'image et en évaluant le comportement des détecteurs dans chaque région. Plus précisément, le nombre de points détectés par un même détecteur est évalué dans chaque région. Idéalement, des détecteurs complémentaires doivent détecter des points dans des régions distinctes. Une autre manière de calculer la complémentarité [Aanæs 12] consiste à calculer D_A, la moyenne des distances minimales entre les points détectés avec le premier détecteur A et ceux donnés par la vérité terrain, ainsi que D_B, la même chose avec le deuxième détecteur B, et, enfin, D_{AB} la même chose en prenant le minimum entre les distances D_A et D_B. La mesure de complémentarité est un simple rapport entre ces 3 quantités, sachant que chacune est pondérée par le nombre de points d'intérêt concernés.

Les travaux de [Aanæs 12] présentent une étude comparative de dix détecteurs différents (Harris, Harris-Laplace, Hessien-Laplace et deux variantes, MESR, IBR, EBR, FAST et SIFT) sur une soixantaines d'images (Il s'agit de scènes d'intérieur avec un éclairage contrôlé et présentant plusieurs types de scènes complexes, notamment avec des occultations). Les auteurs obtiennent une mise en correspondance de référence en utilisant une méthode par lumière structurée sur le même principe que [Scharstein 03]. L'approche de ce papier est d'effectuer une détection dans chaque image puis d'effectuer une mise en correspondance de points d'intérêt vers points d'intérêt.

À présent, nous présentons l'étude de complémentarité que nous avons proposée [Gales 10b] et qui est complémentaire aux travaux que nous venons de décrire. Cette étude nous a servi pour proposer des approches de mises en correspondance par propagation de germes, dont nous parlons dans la partie II, chapitre 6.3. Dans un premier temps, nous donnons les détails des critères définis et utilisés avant de présenter les éléments qui constituent notre protocole d'évaluation et une vue synthétique des résultats obtenus. Il est important de distinguer le critère qui fait intervenir les résultats dans les deux images, donné au § 2.3.1, de celui qui considère le résultat dans une seule image, donné au § 2.3.2.

2.3.1. Répétabilité et gain en répétabilité

Nous avons abordé la notion de répétabilité et, à présent, voici une manière précise de la définir. Si nous notons $\mathbf{p}$, un point dans une image I_1, $\mathbf{p}^{'}$ son correspondant théorique [1] dans une autre image, I_2 alors ce point est répétable s'il existe un point $\mathbf{q}$, détecté dans I_2 et tel que :

$$\|\mathbf{p}^{'} - \mathbf{q}\| \leq \epsilon. \tag{2.2}$$

1. Cela suppose de connaître le déplacement du point $\mathbf{p}$ et donc d'avoir une vérité terrain.

Nous notons $\mathcal{P}_{D,i}$ l'ensemble des points détectés dans l'image I_i par le détecteur D. Ainsi le taux de répétabilité entre les deux images, noté $R(I_1, I_2)$, est donné par :

$$R(\mathcal{P}_{D,1}, \mathcal{P}_{D,2}) = \frac{R(\mathcal{P}_{D,1} \to \mathcal{P}_{D,2}) + R(\mathcal{P}_{D,2} \to \mathcal{P}_{D,1})}{2}. \tag{2.3}$$

avec

$$R(\mathcal{P}_{D,i} \to \mathcal{P}_{D,j}) = \frac{\#\{\text{Points répétables de } \mathcal{P}_{D,i} \text{ vers } \mathcal{P}_{D,j}\}}{\#\{\mathcal{P}_{D,i}\}}. \tag{2.4}$$

En ce qui concerne la notion de répétabilité, nous calculons pour chaque détecteur, notés D_1 et D_2, leurs répétabilités respectives, notées R_1 et R_2. Nous obtenons le **gain en répétabilité** ainsi :

$$G_R(\mathcal{P}_{D_1} \cup \mathcal{P}_{D_2}) = R(\mathcal{P}_{D_1} \cup \mathcal{P}_{D_2}) - \max(R_1, R_2). \tag{2.5}$$

Nous avons simplifié l'expression $(\mathcal{P}_{D_1,i} \cup \mathcal{P}_{D_2,i}, \mathcal{P}_{D_1,j} \cup \mathcal{P}_{D_2,j})$ par $(\mathcal{P}_{D_1} \cup \mathcal{P}_{D_2})$.

2.3.2. Répartition spatiale et gain en répartition spatiale

Le taux de répartition permet d'évaluer la répartition des points. Pour évaluer cette répartition, comme dans [Aanæs 12] [2], nous segmentons l'image en régions de couleurs homogènes (à l'aide de l'algorithme *mean-shift*, que nous présentons au § C.3.2) en faisant l'hypothèse que chaque région ne présente pas de discontinuité de profondeur et donc que cette segmentation est une sous-segmentation de la carte des profondeurs. La répartition est le rapport entre le nombre de régions qui contiennent au moins un point et le nombre total de régions, soit :

$$S(\mathcal{P}_{D_1}) = \frac{\#\{\text{Région avec au moins 1 point détecté}\}}{\#\{\text{Régions}\}}, \tag{2.6}$$

où $\mathcal{P}_{D_i}$ correspond à l'ensemble des points détectés par le détecteur D_i. Ainsi, le gain en répartition spatiale lorsque nous combinons deux détecteurs est le suivant :

$$G_S(\mathcal{P}_{D_1} \cup \mathcal{P}_{D_2}) = S(\mathcal{P}_{D_1} \cup \mathcal{P}_{D_2}) - \max(S(\mathcal{P}_{D_1}), S(\mathcal{P}_{D_2})). \tag{2.7}$$

Il est important de souligner que nous pourrions utiliser n'importe quelle autre technique de segmentation et que ce calcul est dépendant de cette segmentation. Le résultat obtenu serait donc différent mais nous supposons que le comportement observé restera cohérent d'une segmentation à l'autre. Ce critère est donc un indicateur du comportement pour une segmentation donnée.

2.3.3. Complémentarité

Pour la complémentarité, nous considérons trois aspects :

(1) La contribution d'un détecteur D_1 par rapport à un détecteur D_2 ;

(2) Le gain de répétabilité obtenu en combinant deux détecteurs ;

(3) Le gain en terme de répartition spatiale.

Plus précisément, la contribution, $C(D_1, D_2)$ d'un détecteur par rapport à un autre est donnée par :

$$C(D_1 \to D_2) = \frac{\#\{\mathcal{P}_{D_2}\} - \#\{\mathcal{P}_{D_1} \cap \mathcal{P}_{D_2}\}}{\#\{\mathcal{P}_{D_1}\}}, \tag{2.8}$$

Pour déterminer l'intersection, nous considérons que toute paire de points $(\mathbf{p}_1, \mathbf{p}_2)$ détectés respectivement par D_1 et D_2 est confondue si $\|\mathbf{p}_1 - \mathbf{p}_2\| < \epsilon$. Cette mesure n'est pas symétrique puisque nous

2. Ces travaux ont été publiés après la thèse de Guillaume Galès.

ne pouvons pas obtenir strictement le même nombre de points avec chaque détecteur. C'est la raison pour laquelle nous définissons :

$$C(D_1, D_2) = C(D_2, D_1) = \min(C(D_1 \to D_2), C(D_2 \to D_1)). \tag{2.9}$$

2.3.4. Protocole de tests mis en œuvre

Pour effectuer nos tests, nous avons utilisé une partie des images que nous décrivons au § 5.5. Plus précisément, nous avons testé onze images de la base de données de *Middlebury*, introduite dans [Scharstein 02b]. Il est important de noter que pour cette étude préliminaire, les images utilisées présentent des déformations et des changements de points de vue assez faibles, puisqu'il s'agit de couples d'images stéréoscopiques. Nous avons pu étudier les détecteurs suivants :

Parmi les approches du premier ordre :

(1) Moravec, noté MO, cf. § A.1.2, page 236 ;

(2) Harris, noté HA, cf. § A.1.2, page 236 ;

(3) *Ainsi qu'une version multi-échelle*, cf. § A.1.3 : Harris-Laplace, noté HAL ;

Parmi les approches basées région :

(4) SUSAN, noté SU, cf. § A.1.4 ;

(5) FAST, noté FA, cf. § A.1.4 ;

Ainsi qu'un détecteur multi-échelle :

(6) Kadir, noté KA, cf. § A.1.4 ;

Parmi les approches du second ordre :

(7) Beaudet, noté BE, cf. § A.1.5 ;

(8) Kitchen-Rosenfeld, noté KR, cf. § A.1.6

Ainsi que des détecteurs multi-échelle :

(9) Hessian-Laplace, noté HEL, cf. § A.1.5 ;

(10) SIFT, noté SI, cf. § A.1.6 ;

(11) SURF, noté SR, cf. § A.1.6.

2.3.5. Résultats sur l'étude de complémentarité des détecteurs en 2D

La répétabilité calculée pour chaque détecteur correspond à la répétabilité moyenne obtenue sur toutes les images testées. Pour mesurer la répétabilité, nous avons fait varier ϵ, cf. équation (2.2), avec les valeurs suivantes : $\{0, 1.5, 3\}$. Les résultats sont exposés dans la figure 2.7 et mettent en avant les trois détecteurs les plus répétables : Harris, FAST et SIFT. Il est important de remarquer que pour $\epsilon = 0$, la répétabilité est très faible, ce qui implique que pour obtenir une précision d'appariement élevée, effectuer une mise en correspondance de points d'intérêt vers points d'intérêt n'est pas adapté. Il est préférable de faire une recherche de correspondants dans l'image entière (réduite grâce à la prise en compte de contraintes).

En terme de distribution spatiale, nous avons considéré deux segmentations différentes, en considérant une sous-segmentation (une centaine de régions environ) et une sur-segmentation (500 régions environ). Il est évident que les résultats exposés sont fortement dépendants du type de régions obtenues et plus généralement du type de segmentation utilisée et il convient donc de considérer les résultats obtenus avec prudence. De plus, pour modérer les résultats obtenus, il est important de prendre en considération le nombre de points détectés (la cardinalité) car il est possible d'avoir une bonne répartition mais avec peu de points. La conclusion des évaluations que nous avons menées,

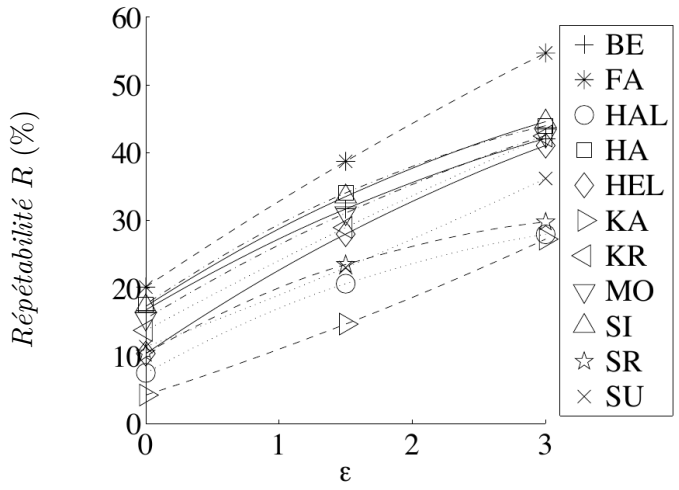


FIGURE 2.7. – Résultats de la répétabilité des détecteurs en faisant varier ϵ, la distance acceptable entre le point détecté dans l'image correspondante et le point correspondant exact.

cf. [Gales 10b], est que le meilleur compromis, c'est-à-dire cardinalité relativement élevée et distribution spatiale régulière, est toujours obtenu avec les mêmes détecteurs : Harris, FAST et SIFT.

En terme de complémentarité, nous avons donc étudié toutes les paires/combinaisons possibles entre deux détecteurs parmi les onze étudiés, sur les trois critères que nous avons proposés : la contribution d'un détecteur par rapport à un autre, D, le gain en répétabilité, G_R, et le gain en distribution, G_S. Dans la table 2.3, nous mettons en lumière les détecteurs les plus complémentaires par rapport aux trois qui ont obtenus les meilleurs performances : Harris, FAST, SIFT. De plus, cette étude a mis en évidence les couples de détecteurs les plus complémentaires entre eux, c'est-à-dire ceux qui présentent les meilleurs gains en les combinant, sans dépasser les résultats obtenus pour les trois détecteurs les plus performants : Kadir et SUSAN, Kitchen-Rosenfeld et Hessien-Laplace, Moravec et Kadir.

Détecteur	C	G_R	G_S
Harris	Hessien-Laplace	Hessien-Laplace	Hessien-Laplace
FAST	Hessien-Laplace	Beaudet	Kitchen-Rosenfeld
SIFT	Hessien-Laplace	Hessien-Laplace	Kadir

TABLE 2.3. – Détecteurs les plus complémentaires en 2D – Nous résumons les détecteurs les plus complémentaires aux détecteurs Harris, FAST, SIFT suivant les trois critères définis : contribution d'un détecteur par rapport à un autre, D, gain en répétabilité, G_R, et gain en distribution, G_S.

2.3.6. Conclusion sur l'évaluation et les complémentarités des détecteurs en 2D

Ainsi, cette étude nous permet de mettre en évidence les détecteurs les plus performants, en cohérence avec les résultats de la littérature, et également les plus complémentaires entre eux. Cette étude préliminaire nous a permis de développer des approches de mise en correspondance par propagation de germes (les germes étant des appariements de points d'intérêt) que nous détaillerons dans la partie II, chapitre 6.3. Ces premiers résultats mettent en évidence la difficulté à obtenir des points répétables d'une image à l'autre et nous compléterons ces résultats en étudiant la complémentarité des détections en 2D et en 3D, cf. chapitre 4.

2.4. Conclusion sur les primitives 2D

Cet état de l'art sur les primitives en 2D, ainsi que cette première étude de complémentarité, nous a permis d'identifier les limites des approches en 2D, ainsi que leur complémentarité, en terme de définition, de performances et d'invariance. Nous avons mis également en avant le fait que le questionnement actuel ne se porte plus sur la détection en elle-même, mais bien sur la construction d'un descripteur ou du moins d'un détecteur complet, dans le sens où un descripteur est systématiquement associé au détecteur. En particulier, des études récentes portent sur la manière de construire des descripteurs qui contiennent l'information la plus discriminante possible et la plus répétable possible d'un point de vue à l'autre, d'une acquisition à l'autre. Ainsi, des approches par réseaux de neurones convolutifs, cf. § 12.2.1, ont été introduits, dès 2015, avec notamment les travaux de [Simo-Serra 15, Tian 17]. Utiliser un réseau convolutif permet d'extraire des primitives d'intérêt en suivant les mêmes principes que ceux utilisés pour un détecteur classique de type Harris [Harris 88] ou à base de différences de gaussiennes [Lowe 04]. En effet, les opérateurs utilisés pour construire les couches d'un réseau de neurones convolutifs sont assez similaires de ceux utilisés par ces détecteurs. Dans les travaux de [Laguna 19], un réseau convolutif est donc proposé dans ce sens et il permet d'extraire le même type de primitives, en multi-résolution et à partir d'un apprentissage profond, ce qui permet de dépasser les performances des détecteurs classiques, sans apprentissage. Dans [Simo-Serra 15, Tian 17], les auteurs utilisent des réseaux siamois afin d'apprendre aussi bien les similarités que les dissimilarités. Ils mettent ainsi en œuvre un descripteur de 128 éléments qu'ils comparent à SIFT et montrent ainsi le gain de précision et de rappel obtenu. Dans [Yi 16], une approche plus élégante consiste à construire les trois éléments nécessaires à cette tâche : détection, calcul de l'orientation et construction du descripteur. Chaque tâche est représentée par un réseau de neurones. Cette approche a également l'avantage de prendre en considération les variations liées aux changements de saisons et d'échelle. Enfin, nous pouvons citer des travaux qui vont au delà d'une analyse par paires d'images, en faisant intervenir des triplets d'images [Balntas 16].

Dans tous ces travaux, il a été mis en évidence que les données actuelles d'apprentissage sont encore assez limitées et qu'un gain accru de performance pourrait être obtenu si ces bases de données étaient complétées [Luo 18]. Pour pallier cette difficulté, les auteurs proposent de prendre en considération des aspects géométriques, et, en particulier, la notion d'échantillonnage des données liées à ce type d'approches.

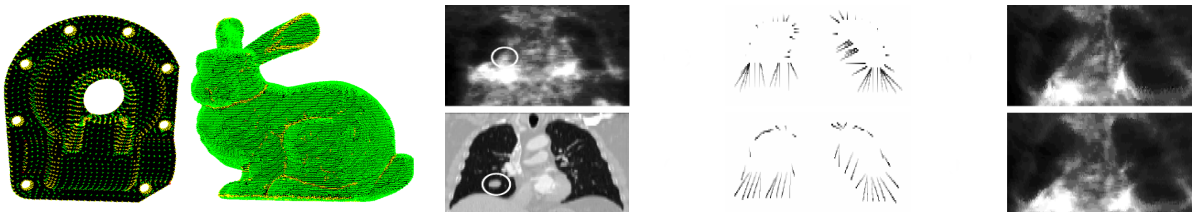
Tout le travail présenté dans ce chapitre correspond à une prise en compte d'images 2D et, à présent, nous allons aborder le travail réalisé en 3D. Nous allons mettre en évidence le lien entre les techniques utilisées dans ces deux modalités avant d'introduire nos travaux sur la détection de points saillants combinant des données d'entrée en 2D et en 3D.

Chapitre 3.

Primitives en 3D

3.1. Introduction sur les primitives en 3D

Lorsque nous abordons la notion de primitives en 3D, il est important de comprendre que la variabilité des données est très importante et que, par conséquent, les approches développées vont être très différentes d'une modalité à une autre. En effet, nous pouvons travailler sur des modèles 3D crées par des artistes avec de nombreux détails et de nombreux attributs liés à la photométrie ou des nuages de points en 3D générés par des acquisitions avec des lasers ou par une reconstruction éparse obtenue grâce à des images 2D prises de différents points de vue. Suivant les applications visées, les recherches portent ainsi sur la manipulation de maillages plus ou moins fins, de modèles surfaciques, de modèles filaires, ou de nuages de points, sans aucune topologie. Les données manipulées sont également plus ou moins régulières (par régularité, nous parlons de la densité de points par unité de volume ou de surface). Nous comprenons bien, avec l'énumération de ces configurations possibles que, par exemple, estimer un gradient en utilisant un voisinage peut s'avérer inapplicable dans certaines situations. Enfin, la dimension ou la taille des données manipulées est également différente du contexte 2D, avec souvent, des dimensions beaucoup plus élevées. Il est donc plus fréquent que les approches développées possèdent une étape de réduction de la taille des données. Un autre aspect qui diffère de la littérature en 2D concerne le type des primitives d'intérêt recherchées. En effet, ces primitives sont plus souvent des régions, des zones que des points [Léon 17]. Tous ces aspects donnent une littérature abondante et très variée. Le but de ce chapitre est d'en présenter les principales tendances sans viser à être exhaustif.



(a) (b)

FIGURE 3.1. – Exemples de primitives d'intérêt en 3D – En (a) il s'agit d'une détection de points caractéristiques sur deux exemples de modèles 3D : *Casting* et *Bunny*. Nous avons distingué les points plats (en vert), les points d'arêtes (en jaune) et les points coins (en rouge). Ces résultats sont extraits du travail de stage de master 1 de Thierry MALON. En (b), sur une coupe longitudinale d'une image 3D, nous cherchons à suivre des points à la courbure particulière, c'est-à-dire, dont la courbure de la surface (ici celle du poumon) est importante. Ces points correspondent aux extrémités reliées par des segments en noirs, dans les images du milieu. Ces résultats sont extraits de [Chambon 11e].

Qu'est ce que c'est ?

Comme en 2D, une primitive 3D va correspondre à un élément saillant, un élément caractéristique 3D. La définition d'une primitive 3D est d'autant plus délicate que la nature des données 3D peut être très différente. En effet, cela peut être une carte de profondeur, un nuage de point 3D (sans notion de topologie), un maillage. Nous parlons de données structurées (volumes/voxels, maillage) ou de données nons structurées (nuages de points). L'information disponible est donc fondamentalement différente et les calculs que nous pouvons utiliser sont donc très différents [Bustos 09].

Toutefois, suivant l'application visée, comme par exemple la compression, un point caractéristique peut s'avérer utile sur tous les types de zones du modèle, c'est-à-dire aussi bien sur une zone de forte courbure que sur une zone de faible courbure. C'est ce que nous pouvons voir sur l'exemple (a) de la figure 3.1. Dans ce contexte, nous souhaitions obtenir toutes les informations nécessaires pour compresser un objet et nous avons donc détecté des points dans ces deux types de zones.

Pourquoi ?

En 3D, l'intérêt de la détection de points d'intérêt est assez différent de ce que nous pouvons trouver en 2D. La problématique de l'utilisation de primitives d'intérêt est plutôt liée aux aspects suivants :

(1) L'amélioration des formes [Kim 06] ;

(2) La sélection de point de vue [Feixas 09], en particulier pour guider les points visuels d'attention [Kim 08] ;

(3) La segmentation de modèles [Katz 05] ;

(4) La reconnaissance ou l'identification de modèles ou de parties de modèles [Gal 06] ;

(5) Le recalage de maillage [Tam 13] ;

(6) La reconnaissance des éléments clés ou des symétries dans les objets pour en simplifier la représentation. Par exemple, identifier ces éléments est utile à la compression, à des fins d'optimisation de stockage et de transmission. De plus, la plupart des objets naturels ou manufacturés possèdent des symétries ou des motifs répétés. Nous pouvons voir un exemple dans la figure 3.1.(a). Un autre exemple est également fourni en imagerie médicale, cf. figure 3.1.(b). Dans des images de cette nature, la reconnaissance de certains points particuliers doit aider à identifier les caractéristiques d'un organe.

Comment ?

Il est important de noter que, dans la littérature, les efforts se sont beaucoup concentrés sur l'utilisation de la courbure et que de nombreux outils sont en réalité une généralisation des détecteurs en 2D à la 3D, comme le détecteur SIFT [Lowe 04]. De plus, les approches suivent le même schéma que celui présenté dans le § **Comment ?** en 2D, page 15.

Ce qui change entre la 2D et la 3D est le besoin d'ajouter une étape de réduction des données. Et parallèlement, les approches par multi-résolution sont utilisées de manière moins fréquente, car décimer ou subdiviser un maillage est une étape coûteuse. Nous pouvons toutefois citer, comme exemple, les travaux de [Nader 14] qui aborde cet aspect multi-résolution.

Un autre aspect important est que dans une image 2D, nous avons, la plupart du temps, une structure, une topologie régulière, contrairement aux nuages de points 3D ou aux maillages qui ne sont pas toujours réguliers. En effet, un pixel (ou un voxel) a toujours un nombre régulier de voisins (4 ou 8 suivant le type de connexité envisagée et moins sur les bords) alors qu'un nuage de points 3D peut avoir un nombre variable de points voisins et un maillage peut avoir un nombre variable de faces adjacentes.

De plus, nous pouvons distinguer deux manières différentes de réaliser cette tâche : utilisation directe de la 3D (le nuage de points ou le maillage), nous parlons alors d'approche basée points, ou projection

des données 3D utilisées dans un espace 2D afin d'appliquer directement un détecteur défini en 2D, nous parlons alors d'approches basées images [Pang 15]. De nombreuses approches basées images s'appuient sur des outils du premier ordre, cf. § 2.2.1, comme [Lee 07], ou du second ordre, cf. § 2.2.3, comme [Lee 13a]. Les approches basées points peuvent également utiliser les notions de points et de frontières [Mineo 19] et également des opérateurs du second ordre comme SIFT en 3D [Ohtake 04, Lee 05, Scovanner 07, Flitton 10, Sattler 11] ou la notion de courbure [Judd 07]. Dans les travaux de [Lei 17], la courbure en multi-résolution est utilisée dans des nuages de points 3D.

En conclusion, en 3D, nous retrouvons toutes les familles distinguées dans la figure 2.4. Il existe ainsi des approches du premier ordre [Hafiz 15], dont l'approche de [Ahmed 18] faisant intervenir un algorithme s'appuyant sur le *mean-shift*, cf. § C.3.2 pour détecter des coins et des frontières. Nous pouvons distinguer ces principales caractéristiques liées à la 3D : les approches du second ordre sont nettement favorisées par rapport aux autres et il y a peu d'adaptation d'approches basées régions en 2D, comme présentées dans le § 2.2.2.

Qualité du résultat ?

Une qualité attendue est liée à la répétabilité, comme nous l'avons déjà défini en 2D, cf. page 2.1. Cette propriété est d'autant plus difficile à obtenir que nous avons des données beaucoup moins régulières (nuage de points épars, maillage avec une résolution ou une régularité différente).

Et le descripteur ?

Les descripteurs les plus simples s'appuient sur la couleur (lorsqu'une texture est disponible) [Alexandre 12]. De nombreuses approches utilisent les *spin images* [Johnson 99] qui correspondent à une image 2D représentant les points de la surface appartenant au voisinage du point d'intérêt considéré [Darom 12]. En particulier, ce concept a été largement repris et étendu dans les travaux de [Mitra 06]. Dans [Castellani 08], le descripteur utilisé, pour des points d'intérêt dans des maillages, s'appuie sur l'apprentissage d'un champ de Markov caché. Plus précisément, les auteurs prennent en compte les sommets pour chaque voisinage séparément (1-voisinage, 2-voisinage ...). Dans [Zaharescu 09], c'est le descripteur HoG, cf. § A.2, qui est adapté aux maillages. De la même façon que pour la détection, une adaptation du descripteur SIFT a été réalisée pour des maillages 3D [Maes 10]. Dans les travaux de [Darom 12], il s'agit de variantes proposées pour les approches basées *spin image* et l'approche SIFT en 3D.

3.2. Approches de détection et de caractérisation des points d'intérêt en 3D

En introduction, nous avons conclu que nous pouvions classer les approches en 3D de la même façon qu'en 2D. Ce que nous proposons dans cette section n'est pas un équivalent du travail réalisé dans le § 2.2 mais c'est une illustration de ce qui est possible en 3D, notamment avec des maillages. Le travail présenté dans cette section n'a pas été publié mais il résulte du suivi de deux étudiants en stage de master 1 : Thierry Malon, en 2014, et Martin Jacquet, en 2015, co-encadrés avec Géraldine Morin, dans mon équipe (REVA), et Sandrine Mouysset, équipe APO, Algorithmes Parallèles et Optimisation. Nous avons étudié et combiné des outils existants. C'est pour cette raison que ce que nous présentons dans cette section fait partie de l'analyse de l'état de l'art que nous proposons.

3.2.1. Introduction sur la recherche d'éléments caractéristiques

Un intérêt à définir et identifier des points saillants dans des modèles 3D est d'être en capacité d'identifier des symétries ou des motifs répétés [Mitra 06, Mitra 12]. En effet, de nombreux objets

manufacturés présentent des symétries ou des motifs répétés et l'intérêt applicatif premier est d'être en capacité de simplifier le stockage du modèle, c'est-à-dire de faire de la compression de modèles. Dans un premier temps, nous avons commencé à étudier des maillages. La difficulté à manipuler des maillages est que le volume de données à traiter est plus important qu'une simple image 2D et les temps de calcul peuvent être importants. De plus, un même objet 3D peut être représenté par différents maillages, réguliers ou non et il est souhaitable d'obtenir les mêmes points répétables d'un maillage à l'autre, que celui-ci soit régulier ou non. De la même manière qu'en 2D, nous pouvons utiliser des opérateurs de courbure comme ceux présentés au § A.1.5 et comme dans [Cohen-Steiner 03], en utilisant la courbure moyenne, définie comme la moyenne des courbures principales, et la courbure gaussienne, définie comme le produit des courbures principales. La notion de courbure principale sera introduite en détails à la page 49. Nous pouvons également citer les approches spectrales, comme celle de [Wang 12c]. Cette approche a l'intérêt de distinguer clairement ces trois catégories de points : les points qui correspondent à des régions plates, les points qui correspondent à des arêtes et enfin les points qui correspondent à des sommets. Enfin, il est possible d'utiliser des opérateurs de premier ordre, notamment en utilisant des informations de normales à la surface [Weber 10].

Ainsi, nous avons choisi d'étudier une approche du premier ordre en nous appuyant sur [Weber 10] et une approche du second ordre, notamment en nous appuyant sur le détecteur/descripteur SIFT, cf. § A.1.6. Dans la suite, nous allons donner les détails des étapes étudiées :

(1) La détection des points saillants : cette étape a pour but de réduire le nombre de points à traiter en extrayant des points caractéristiques, au sens de points d'intérêt comme déjà introduit suivant les deux stratégies étudiées : premier ordre et second ordre ;

(2) La description des points détectés et ainsi l'estimation des isométries grâce aux descripteurs construits.

Dans les paragraphes qui suivent, nous avons utilisé des modèles classiques de la littérature [1].

3.2.2. Détecter des points saillants dans un maillage

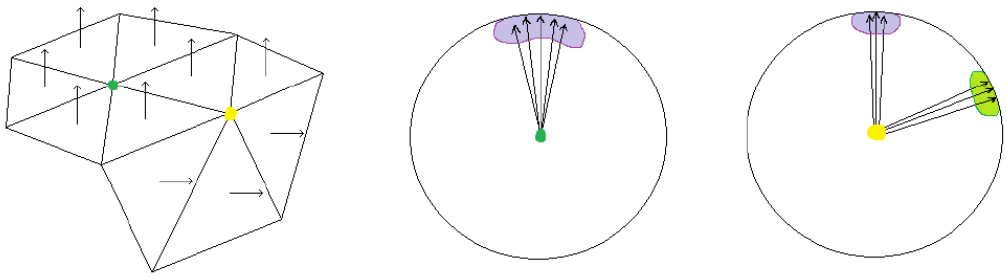


FIGURE 3.2. – Détection de points saillants dans des maillages 3D par étude des normales à la surface – Nous présentons, à gauche, un exemple de topologie locale d'un maillage : nous distinguons le 1-voisinage d'un point d'une zone plate (point vert) et le 1-voisinage d'un point sur une arête vive (point jaune). Nous visualisons les projections selon les normales aux faces adjacentes, pour la zone plane, puis pour la zone avec une arête vive. En effectuant une classification suivant l'écart angulaire, nous obtenons une seule classe dans le premier cas, et deux classes distinctes dans le second.

Approche [Weber 10] – Initialement, cette approche a été élaborée pour des nuages de points et non pour des maillages. Nous avons donc adapté cette méthode à des maillages. Cette technique permet d'obtenir des points saillants en dehors des zones planes. Pour chaque point du maillage étudié,

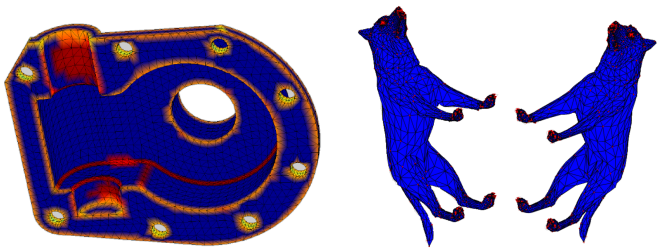
1. http://graphics.stanford.edu/data/3Dscanrep/

le principe est le suivant : plus les normales entre deux faces adjacentes, avec ce sommet en commun, sont différentes, c'est-à-dire plus elles présentent un écart angulaire important, et plus le sommet joignant ces faces a une forte probabilité d'être un point saillant du maillage, cf. figure 3.2. Pour chaque point du maillage, nous considérons les vecteurs normaux du 1-voisinage (l'ensemble des faces adjacentes au point) ou du 2-voisinage (l'ensemble des faces adjacentes aux points des faces adjacentes au point). Nous projetons ensuite le centre de la sphère sur sa surface selon tous ces vecteurs normaux. Enfin, nous classons ces points en regroupant ceux qui se trouvent à une distance inférieure à un seuil T à choisir [2]. Une fois ces calculs réalisés, il est possible de classer les points en trois catégories :

- Les points plats : quand il n'y a qu'une classe.
- Les points appartenant à une arête vive : Lorsqu'il y a deux classes.
- Les points appartenant à un sommet : Lorsqu'il y a trois classes.

Ce sont ces deux dernières catégories que nous utiliserons pour déterminer les isométries. Cette approche assez simple nous a permis d'obtenir des résultats encourageants et elle a l'avantage de n'avoir qu'un seul paramètre à choisir, le seuil T. En revanche, elle est coûteuse en temps de calcul [3].

Approche [Darom 12] – Dans la littérature, de nombreux papiers ont tenté de reproduire les travaux de [Lowe 04], cf. A.1.6, en suivant chacun des éléments de cette approche : calcul d'une réponse multi-échelle et multi-espace, puis construction d'un descripteur en s'appuyant sur des histogrammes locaux [Castellani 08, Zaharescu 09, Maes 10, Darom 12]. Ces méthodes ont toutes cherché à traduire chaque étape de l'algorithme initial au contexte 3D. Logiquement, pour respecter le principe de cet algorithme, il est nécessaire de construire des maillages à différentes résolutions, pour être cohérent avec les pyramides d'images 2D utilisées. Mais la plupart du temps, les approches en 3D n'effectuent pas cette étape, jugée trop coûteuse, et travaillent uniquement sur une octave. Seul [Castellani 08] a pris en compte cet aspect, en mettant en évidence un gain de robustesse de son approche par rapport aux autres. Toutes les approches citées effectuent un filtrage gaussien à différentes échelles, tout comme l'approche SIFT. Elles construisent donc une octave, à partir du maillage original. Ce qui diffère est la manière d'appliquer le filtre gaussien. Une première façon de faire est d'effectuer des différences de normes [Castellani 08, Zaharescu 09, Darom 12] mais [Maes 10] s'appuie sur la notion de courbure. Dans l'article de [Darom 12], le maillage est « transformé » en carte de profondeur, se ramenant ainsi à un cas 2D. Nous abordons ici des notions qui seront à nouveau abordées dans le chapitre 4.



(a) (b)

FIGURE 3.3. – Extraction de points d'intérêt dans des modèles 3D – En (a), nous montrons, les points obtenus pour *Cranck* avec la première approche étudiée s'appuyant sur les normales. Plus le point est rouge, plus il est saillant. En (b), nous illustrons les résultats obtenus pour *wolf* avec SIFT adapté au 3D. Pour apprécier les symétries que nous pouvons déjà reconnaître, les points détectés sont montrés suivant deux points de vue symétrique.

Nous avons choisi d'expérimenter l'approche de [Darom 12]. Afin d'être robuste à la variabilité du

2. En pratique, nous avons choisi $T = 0.25$ pour $T \in [0; 1]$.
3. Par exemple, elle prend plusieurs minutes sur un modèle de 25000 points.

maillage étudié qui peut posséder plus ou moins de sommets, les auteurs proposent de calculer en premier lieu la densité locale de points pour chaque sommet et ce sont ces densités qui serviront pour calculer les σ des filtres gaussiens utilisés. Ensuite, les différences de gaussiennes sont calculées, de la même façon qu'en 2D, et les extrema locaux (en prenant en compte le 1-voisinage et les échelles directement inférieure et supérieure) sont extraits. La figure 3.3 nous permet de présenter les types de résultats que nous obtenons. Les deux résultats sont exploitables et présentent l'avantage d'être indépendants du maillage.

3.2.3. Décrire des points saillants dans un maillage

Pour l'approche [Weber 10], nous avons choisi d'utiliser comme descripteur les éléments suivants :

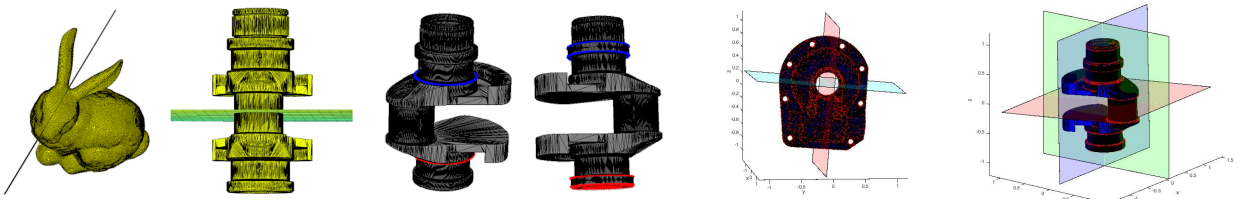
- La norme de la somme des normales aux faces adjacentes ;
- L'écart-type des angles entre les normales aux faces et le vecteur normal moyen.

Le premier élément est assez classique mais des configurations de projetés différentes peuvent donner la même norme. C'est pourquoi nous introduisons le second élément. Ensuite, plus un point possède de faces dans son 1-voisinage, plus il possédera aussi de normales, et la somme de ses normales sera différente de celle en un point de même forme locale, mais avec un découpage des faces différent. Pour obtenir un résultat répétable, quelle que soit la connectivité, dans le calcul de la somme des normales, nous ajoutons une pondération de chaque normale. Le poids est proportionnel à la portion angulaire formée par l'angle de la face qui contient le point étudié, par rapport à la somme des portions angulaires. Nous ajoutons également cette pondération dans le calcul de l'écart-type. Ce descripteur est donc peu dépendant de la topologie ou connectivité du maillage mais il est difficilement généralisable au 2-voisinage.

Pour l'approche [Darom 12], nous avons simplement appliqué SIFT en 2D à une carte de profondeur estimée localement à partir du maillage. À l'aide des points d'intérêt et de leurs descripteurs associés définis dans les sections précédentes, nous avons pu identifier les plans de symétrie et et les translations, comme nous pouvons le voir dans la figure 3.4.

3.2.4. Conclusions sur la recherche de points d'intérêt dans des maillages 3D

Nous avons étudié deux détecteurs/descripteurs différents et évalué leur intérêt pour détecter des symétries. Nous souhaitons que cette étude bibliographique préliminaire soit investie dans les travaux que nous envisageons en perspective au § 13.6. À présent, nous allons présenter les travaux réalisés sur la détection de points d'intérêt en 3D en imagerie médicale multimodale.



(a) (b) (c) (d)

FIGURE 3.4. – Extraction de plans de symétrie et de translation dans des modèles 3D – En (a), nous montrons, en vert, les différents plans de symétrie estimés pour deux modèles 3D, *bunny* et *cranck*. En (b), nous illustrons quelques unes des translations trouvées (en bleu et en rouge) pour le modèle *cranck*. Ces deux premiers cas correspondent au premier détecteur/descripteur utilisé s'appuyant sur les normales aux surfaces. En (c), avec le modèle *casting*, et en (d), avec le modèle *cranck*, nous montrons les résultats obtenus pour le détecteur/descripteur basé SIFT. Dans ces deux exemples, nous affichons les plans de symétrie les plus probables que nous ayons calculés.

3.3. Identification de points d'intérêt dans des images 3D médicales

Le recalage multi-modal est utilisé dans de nombreuses méthodes qui visent des applications diagnostiques et thérapeutiques. En particulier, le recalage d'images tomodensitométriques (TDM) et d'images de tomographie par émission de positons (TEP), en imagerie thoracique, permet d'associer des informations complémentaires sur la nature et la localisation des tumeurs du poumon [Vogel 06]. Un recalage linéaire ne permet pas de prendre en compte les déformations dues aux mouvements respiratoires. Il existe des machines combinées TEP/TDM qui permettent de limiter les différences d'orientation du patient entre les différentes acquisitions et de supposer que les images sont recalées linéairement. Malheureusement, ces techniques ne permettent pas de résoudre les problèmes liés à la respiration. Il est donc nécessaire de mettre en œuvre un recalage non linéaire [Shekhar 05].

Nous pouvons distinguer deux types de méthodes de recalage non linéaire : les approches utilisant une mesure de similarité entre les images à recaler et les approches s'appuyant sur une mise en correspondance de points d'intérêt [Maintz 98, Pluim 03, Zitovà 03]. Dans la plupart de ces méthodes, les connaissances physiologiques sur le corps humain sont ignorées. Utiliser une mise en correspondance de points d'intérêt anatomiques permet déjà une première approche pour tenter d'intégrer des caractéristiques physiologiques dans le processus de recalage. Une grande partie des méthodes à base de points d'intérêt utilisent l'information de courbure. Cependant, peu d'articles exposent la manière de détecter ces points d'intérêt [Betke 03, Rohr 01]. Une deuxième approche consiste à introduire un modèle de respiration au cours de l'estimation du recalage. Alors qu'il existe de nombreux travaux prenant en compte un modèle de respiration pour la visualisation, pour corriger des artefacts dans les images acquises ou pour estimer les déformations d'un poumon lors d'une radiothérapie, peu de méthodes prennent en compte ce type de modèle pour réaliser un recalage multi-modal.

Nous proposons une approche pour introduire un modèle 3D surfacique du poumon au cours du cycle respiratoire dans une méthode de recalage afin de garantir des déformations physiologiquement réalistes. Nous nous appuyons sur l'utilisation de deux volumes [4] TDM et d'un volume TEP, illustrés sur la figure 3.5 (sur une coupe). Les deux volumes TDM correspondent à deux instants du cycle respiratoire, alors que l'image 3D TEP est moyennée au cours de la respiration. De plus, nous introduisons une méthode de détection automatique de points d'intérêt s'appuyant sur la courbure.

Après avoir résumé les travaux existants sur les modèles de respiration, cf. § 3.3.1, ainsi que les méthodes de recalage associées, nous présenterons le modèle de respiration que nous proposons, cf. § 3.3.2. Ensuite, nous décrirons la manière de détecter automatiquement des points d'intérêt sur la surface du poumon, en utilisant la courbure, car il s'agit de notre contribution à la méthode complète de recalage, cf. § 3.3.3.



(a) (b) (c)

FIGURE 3.5. – Problème de recalage d'images 3D médicales – Vue coronale des images TDM à deux instants du cycle respiratoire (a) et (b) et l'image TEP correspondante (c) pour un même patient (moyennée sur l'intervalle de temps nécessaire à l'acquisition).

4. Nous choisissons, pour simplifier la lecture de remplacer le terme *image 3D* par *volume*.

3.3.1. Modèle de respiration et synthèse d'images TDM intermédiaires

Les techniques de radiothérapie avec asservissement respiratoire se sont développées afin d'améliorer l'irradiation des tumeurs abdominales ou pulmonaires. La prise en compte des mouvements dus à la respiration peut être réalisée à deux niveaux : au cours de la reconstruction des modèles volumiques ou au cours du traitement. Dans le cas de la reconstruction des modèles volumiques, les méthodes proposées sont dépendantes du matériel [Crawford 96, Rit 06]. Dans le cas des méthodes qui effectuent une prise en compte au cours du traitement, nous avons distingué trois catégories de méthodes :

(1) *Actives* – Le patient est acteur de son traitement. À l'aide d'un appareil mesurant le volume et le débit de sa respiration (un spiromètre), il visualise en temps réel son cycle respiratoire et bloque sa respiration à des instants donnés où l'irradiation de la tumeur sera la plus efficace.

(2) *Passives* – Une méthode d'asservissement au mouvement, *gating*, est mise en œuvre, c'est-à-dire que le système est capable de détecter les mouvements et de couper automatiquement la respiration du patient. Nous pouvons différencier les méthodes par suivi de marqueurs [Nehmeh 04, Schweikard 00] (ce sont des méthodes invasives, la plupart du temps) des méthodes empiriques qui s'appuient sur des acquisitions préliminaires [McClelland 06, Rohlfing 01].

(3) *Basées modèle* – Il s'agit des méthodes les plus difficiles à mettre en œuvre mais elles permettent un meilleur suivi de la surface du poumon en s'appuyant sur un modèle de respiration [Sarrut 05].

Nous nous intéressons aux dernières méthodes décrites car elles permettent une plus grande précision pour suivre la surface du poumon au cours du cycle respiratoire et, à l'opposé des méthodes passives, elles ne nécessitent pas une mise en œuvre délicate au niveau du traitement radiothérapique.

Les travaux qui intègrent un modèle de respiration dans leur technique de recalage s'appuient sur des segmentations d'IRM pour simuler des images TEP à différents instants du cycle respiratoire et ces différents TEP sont utilisés pour affiner le recalage entre images IRM et TEP [Pollari 04]. D'autres utilisent également des IRM déjà recalées pour estimer un modèle de respiration [Rohlfing 04, Sundaram 05]. Dans le domaine du recalage d'images du foie, une approche variationnelle qui combine la segmentation et le recalage d'images TDM a été également proposée pour estimer l'influence des mouvements respiratoires [Ehrhardt 07]. Ces travaux utilisent ou estiment un modèle de respiration pour visualiser, simuler ou faire des études médicales, mais aucun de ces modèles n'a réellement été proposé pour affiner un recalage multi-modal, dans le cadre des applications en radiothérapie.

De nombreux modèles de respiration ont déjà été développés. Nous pouvons distinguer deux catégories de modèles : géométriques ou physiques. La méthode géométrique la plus populaire s'appuie sur les NURBS, *Non-Uniform Rational B-Spline*. Une surface NURBS, exprimée par un rapport entre B-Splines, est une représentation paramétrique bidirectionnelle d'un objet. Le modèle NCAT, *NURBS-based CArdiac-Torso*, a été proposé pour modéliser les variations cardiaques sur des images SPECT/TEMP, *Single photon computed tomography*, tomographie à émission monophotonique [Segars 02]. Une approche multi-résolution est proposée dans [Rohlfing 01] pour des images de résonance magnétique, IRM, 4D. Dans [Guerrero 05], le modèle NCAT est utilisé pour générer, à partir d'une image TDM originale, des images TDM à d'autres instants de la respiration pour effectuer un recalage élastique.

Les modèles physiques permettent de décrire l'influence du flux d'air à l'intérieur du poumon. Dans de nombreux cas, l'estimation de ce modèle nécessite l'utilisation d'un coordinateur actif de la respiration ou ABC, *Active Breathing Coordinator* [Sarrut 05]. Ce système permet d'acquérir des images du poumon à différents instants du cycle respiratoire avec le volume d'air réel présent dans les poumons. D'autres méthodes s'appuient sur une relation entre la pression et le volume, PV [Zordan 06]. Dans [Narusawa 01, Venegas 98], les auteurs supposent que la relation entre la pression et le volume est une fonction sigmoïdale et proposent une méthode pour en estimer les paramètres. Dans [Promayon 97], le modèle distingue les zones déformables et non déformables dans le poumon.

Nous avons choisi d'utiliser un modèle physique car les modèles de ce type sont facilement adaptables pour simuler les mouvements du poumon spécifiques à un patient, sans utiliser un équipement de

contrôle de la respiration.

3.3.2. Modèle physique 3D de la surface du poumon

Cette partie de notre travail a été réalisée en collaboration avec l'université d'Orlando, en particulier avec Anand Santhanam et Jannick Rollanc. Le modèle de respiration que nous utilisons a été initialement présenté par eux dans [Santhanam 07] et sa construction, à partir d'un ensemble de volumes TDM acquis pour un patient de référence, comporte les deux étapes suivantes, cf. Figure 3.6 :

(1) Estimation des paramètres de la relation PV, qui joue le rôle d'ABC ;

(2) Estimation des déformations à partir des images TDM 4D.



(a) Relation PV

(b) Fin d'expiration

(c) Fin d'inspiration

Figure 3.6. – Modèle physique de la respiration utilisé pour le recalage multi-modal basé points d'intérêt – La relation Pression–Volume (a) et deux maillages (b-c) du modèle de respiration obtenus avec les TDM 4D de référence. Il s'agit du modèle initial de respiration qui s'appuie sur des images de référence, avant l'adaptation au patient.

Ce modèle va nous permettre de générer des maillages dont les nœuds correspondent aux points de la surface du poumon, à différents instants du cycle respiratoire. Dans l'étape (1), à partir d'une relation PV calculée sur un patient de référence (cas normal), le volume et la forme du poumon 3D sont simulés à différents instants du cycle respiratoire. Pour l'étape (2), les vecteurs de déplacement des nœuds estimés à partir des images TDM 4D du patient de référence sont utilisés comme initialisation. La direction et l'amplitude des déplacements des points sur la surface du poumon sont estimées en respectant la contrainte d'évolution linéaire du volume, c'est-à-dire le fait que l'étirement des tissus pulmonaires est linéairement dépendant de l'augmentation du volume du poumon. La totalité des forces appliquées à chaque nœud (qui est relative au flux d'air dans le poumon, c'est-à-dire à la croissance des régions alvéolaires) est estimée à partir de la relation PV, étape (1), et de l'orientation du poumon (pour prendre en compte l'effet de la gravité). Une déformation physique s'appuyant sur l'estimation d'une fonction de Green est ensuite exploitée pour déformer les maillages 3D. Ces fonctions comportent un facteur physiologique (qui représente l'élasticité des tissus et la croissance des régions alvéolaires) et un facteur structurel (la distance entre les différents nœuds du maillage). Pour estimer ces différents facteurs, une approche itérative est utilisée : à chaque étape, la force associée à chaque nœud est localement mise à jour en prenant en compte les nœuds voisins (la croissance alvéolaire est localement normalisée). L'algorithme s'arrête lorsque les forces sont équilibrées. Cette approche a été évaluée à partir d'un ensemble d'images TDM obtenues pour un patient au cours du cycle respiratoire [Santhanam 07]. Le but de ce travail consiste à proposer une méthodologie permettant d'intégrer ce modèle de respiration dans une méthode de recalage non linéaire afin d'obtenir des déformations du poumon physiologiquement plausibles. De plus, le niveau de développement de ce modèle est suffisant pour répondre à l'objectif que nous nous fixons : intégrer des contraintes physiologiques réalistes dans une méthode de recalage.

3.3.3. Sélection de points de repère

Pour estimer l'ensemble des paramètres de la transformation présentée au paragraphe précédent, il est nécessaire de détecter et mettre en correspondance des points d'intérêt sur les deux volumes à recaler : l'image 3D TDM et l'image 3D TEP. Nous détectons donc des voxels d'intérêt. La qualité du recalage dépend de cette sélection et de la mise en correspondance associée. La sélection de ces voxels d'intérêt peut être manuelle (c'est le cas dans de nombreuses méthodes) [West 05], semi-automatique [Hartkens 02], ou automatique [Rohr 01]. La sélection manuelle est fastidieuse et longue. Dans [Hartkens 02], les auteurs mettent en évidence l'intérêt d'utiliser une méthode semi-automatique qui permet d'intégrer les connaissances d'experts dans un processus automatique. La sélection automatique permet de réduire considérablement les temps de traitement et d'obtenir une bonne précision. La plupart des méthodes automatiques utilisent la courbure. De manière générale, pour définir des points saillants dans des modèles volumiques, nous avons déjà indiqué les approches utilisant la courbure moyenne ou la courbure gaussienne [Cohen-Steiner 03] mais il existe des méthodes hybrides, comme celle de [Betke 03] qui combine auto-corrélation (premier ordre) et courbure (second ordre).

Nous supposons que les points anatomiques caractéristiques, c'est-à-dire qui fournissent une information significative pour estimer la déformation du poumon, correspondent à des points de forte courbure sur la surface du poumon. Ainsi, nous avons choisi d'effectuer une sélection automatique s'appuyant sur la courbure gaussienne et la courbure moyenne. De plus, afin de prendre en compte la forme de toute la surface du poumon, nous proposons d'obtenir une répartition « pseudo-uniforme ». Notre algorithme s'appuie sur les étapes suivantes :

(1) Calcul de la courbure pour chaque voxel de la surface du poumon (des volumes TDM et TEP) ;

(2) Classement des voxels dans l'ordre décroissant des valeurs absolues des courbures ;

(3) Sélection des voxels en utilisant la courbure et un critère de distance ;

(4) Si une sélection « pseudo-uniforme » est souhaitée, ajout de voxels dans les zones où la courbure est nulle.

Les courbures moyennes et gaussiennes ont été estimées sur les volumes, c'est-à-dire les images tridimensionnelles binaires résultant de la segmentation des poumons, en nous appuyant sur les formules analytiques de la courbure [Beil 97]. Les dérivées ont été calculées en utilisant les différences finies. Dans l'étape 3, nous considérons $\mathcal{V} = \{\mathbf{v}_i\}_{i=0..N_\mathcal{S}}$, l'ensemble des voxels par ordre décroissant des valeurs absolues de la courbure, avec $N_\mathcal{S}$ le nombre de voxels sur la surface, et $\mathcal{V}_\mathcal{L} = \{\mathbf{v}_{\mathcal{L}i}\}_{i=0..N_\mathcal{L}}$, l'ensemble des points d'intérêt, avec $N_\mathcal{L}$ le nombre de points d'intérêt. Pour chaque voxel $\mathbf{v}_i \in \mathcal{V}$ (pour $i = 0$ à $N_\mathcal{S}$) avec une courbure non nulle, nous ajoutons $\mathbf{v}_i$ dans $\mathcal{V}_\mathcal{L}$, si $\forall\ \mathbf{v}_j \in \mathcal{V}_\mathcal{L}$, $d_g(\mathbf{v}_i, \mathbf{v}_j) > T$ où d_g est la distance géodésique sur la surface du poumon et T est un seuil à choisir. Avec ce type de sélection, certaines zones (les zones plates, à courbure nulle) ne contiendront aucun point d'intérêt, c'est pourquoi nous avons ajouté l'étape 4 : pour chaque voxel sur la surface du poumon $\mathbf{v}_i \in \mathcal{V}$ avec une courbure nulle, s'il n'y a aucun voxel tel que $\mathbf{v}_j \in \mathcal{V}_\mathcal{L}$ avec $d_g(\mathbf{v}_i, \mathbf{v}_j) < T$, alors nous ajoutons $\mathbf{v}_i$ à $\mathcal{V}_\mathcal{L}$.

La distance géodésique sur la surface est calculée par une méthode de propagation de type Chanfrein, conditionnelle à la surface. La méthode utilisée est proche de l'algorithme de Dijkstra [Dijkstra 59] qui permet de trouver le chemin le plus court entre deux points [Klette 04, p. 129]. Nous ne prenons en compte que les voxels de la surface du poumon. La connexité utilisée pour la propagation comprend les voxels dans un voisinage $3 \times 3 \times 3$ du voxel courant. Quatre variantes ont été testées :

(1) Moy – Courbure moyenne sans répartition uniforme ;

(2) Gau – Courbure gaussienne sans répartition uniforme ;

(3) NonUni – Courbures moyenne et gaussienne sans répartition uniforme ;

(4) Uni – Courbures moyenne et gaussienne avec répartition « pseudo-uniforme ».

Lorsque les deux types de courbure sont utilisés (méthodes NonUni et Uni), l'ensemble $\mathcal{V}$ correspond à une fusion des voxels par ordre décroissant de courbure moyenne et gaussienne, en prenant alternativement un voxel de chaque ensemble de courbures. Ces stratégies sont comparées dans la figure 3.7. Les résultats donnés par Moy et Gau sont très différents et l'intérêt de les combiner est illustré par les résultats obtenus avec NonUni. La méthode Uni permet d'ajouter des points dans les zones localement plates comme l'illustre la figure 3.7.

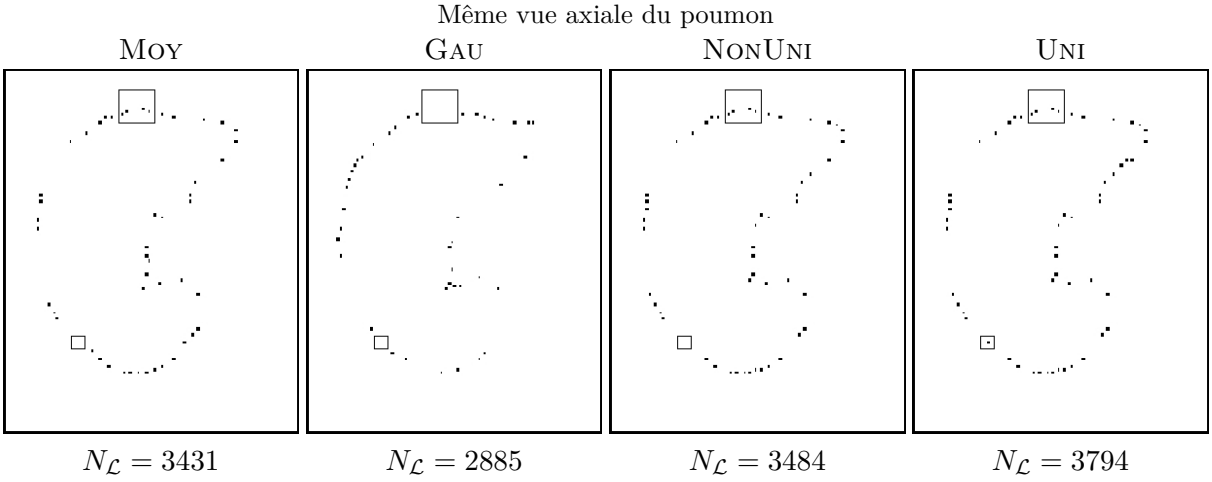


Figure 3.7. – Sélection de points d'intérêt en 3D (vue sur une coupe transversale d'un seul poumon) – Dans chaque image, deux régions sont identifiées par des rectangles. Dans le plus grand rectangle, il n'y a pas de points détectés par la méthode gau alors qu'il y en a avec la méthode moy. Avec la méthode de fusion, moy-gau, ces points ont été conservés. Dans le plus petit rectangle, aucun point n'est sélectionné, quelle que soit la courbure utilisée. La méthode moy-gau-uni permet bien d'obtenir un point dans cette région.

3.3.4. Conclusion sur l'identification de points d'intérêt 3D en imagerie médicale

Ce travail a été développé dans le cadre d'un projet ANR nommé MARIO, dont tout le processus a été publié dans [Chambon 11d]. Ce projet a permis de développer une méthodologie complète de recalage TEP/TDM [Chambon 07a]. Une des contributions importante est la proposition d'une méthodologie pour introduire un modèle de respiration dans une méthode de recalage TEP/TDM garantissant des déformations physiologiquement plausibles de la surface du poumon. Ainsi, le recalage proposé s'appuie sur la proposition d'une méthode de détection automatique de points d'intérêt [Chambon 07b] qui permet de combiner les courbures gaussienne et moyenne tout en fournissant une répartition « pseudo-uniforme », et sur l'introduction de contraintes de rigidité sur les tumeurs et sur des organes à risque tels que le cœur. Des exemples de résultats obtenus sont présentés dans la figure 3.8. Ils illustrent à quel point l'utilisation d'un modèle de respiration et de contraintes peut améliorer la qualité du recalage.

Nous n'avons pas développé l'approche de recalage que nous avons mise en œuvre, mais tous les détails sont fournis dans [Moreno-Ingelmo 08].

3.4. Conclusion sur les primitives 3D

Nous avons présenté un état de l'art des approches classiques en détection de primitives 3D. Actuellement, des approches s'appuient sur des réseaux de neurones convolutifs classiques, cf.§ 12.2.1, [Zhou 18, Ghadai 19]. Dans le travail de [Shu 19], deux réseaux sont combinés : le premier permet de prendre en

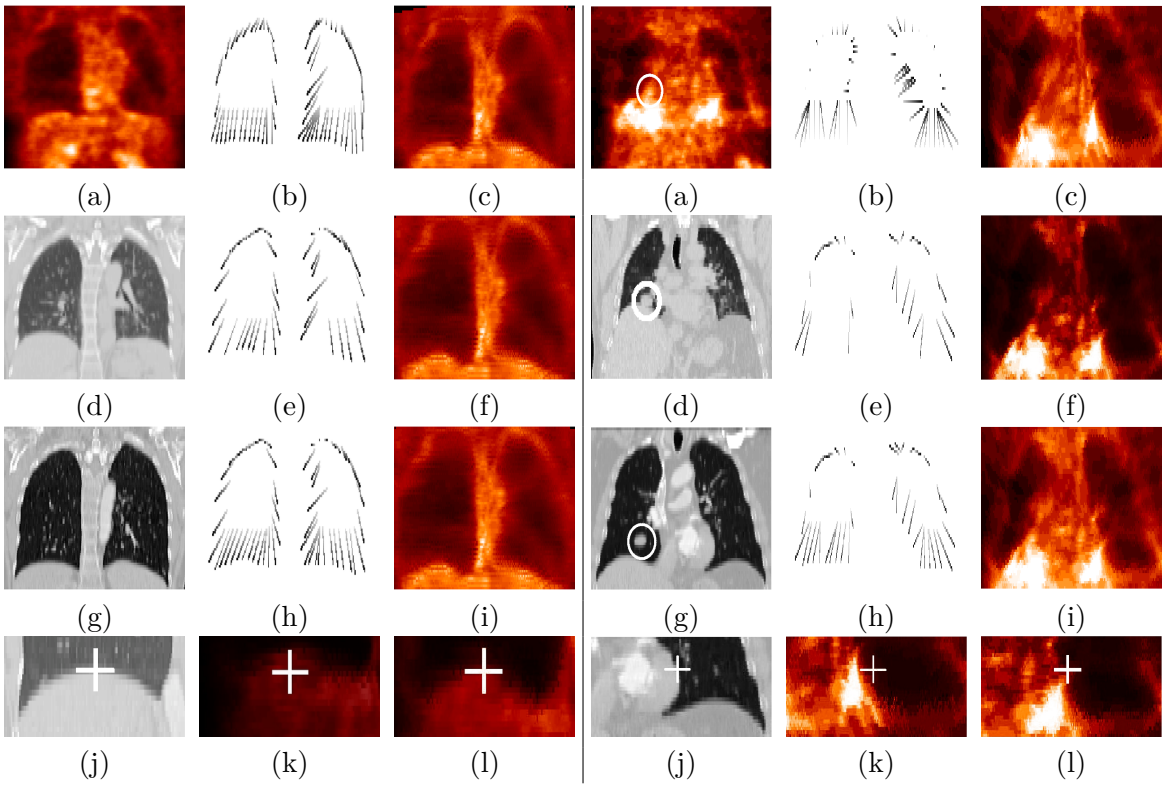


FIGURE 3.8. – Résultat de recalage par points d'intérêt entre une image TEP (a) et des images TDM originales (d, g), dans un cas normal, à gauche, et dans un cas pathologique (la tumeur est mise en évidence par le cercle blanc), à droite. Le recalage obtenu en (c) est sans modèle de respiration, alors qu'en (f) il est obtenu en utilisant un modèle de respiration et une détection de points d'intérêt non uniforme et, enfin, en (i) le modèle de respiration et une détection de points d'intérêt « pseudo-uniforme » ont été utilisés. De même, les correspondances entre les points d'intérêt sur les images TEP et TDM en respiration maximale sont illustrées en (b) pour la méthode sans modèle de respiration, en (e) avec le modèle de respiration et une détection de points d'intérêt non uniforme et en (h) avec le modèle de respiration et une détection de points d'intérêt « pseudo-uniforme ». Sur la quatrième ligne, il s'agit d'un détail des images. La croix permet de visualiser le même point sur l'image TDM originale (j), l'image TEP recalée sans modèle (k) et avec modèle de respiration (l). L'utilisation du modèle permet d'obtenir une meilleure correspondance entre les surfaces des poumons.

considération des caractéristiques géométriques en s'appuyant sur trois descripteurs (les plus utilisés dans la communauté) alors que le second permet d'obtenir une carte des probabilités que les points de la surface soient des points d'intérêt (ce dernier réseau s'appuie sur un auto-encodeur, cf. § 12.2.3). Une dernière classification classique est appliquée pour répartir des points d'intérêt sur la surface.

Tout ce chapitre a mis en évidence le fait que l'étude de la courbure de la surface reste le critère le plus étudié en 3D. Étant donné qu'il existe également de nombreuses approches en 2D, il paraît assez naturel d'utiliser cette information de manière conjointe entre 2D et 3D. De plus, nous avons séparé primitives en 2D et primitives en 3D car il existe peu de travaux qui confrontent des approches de détection aussi bien en 2D qu'en 3D. Il existe bien évidemment des approches qui combinent ces deux types d'information, et nous allons en décrire quelques unes au chapitre suivant, § 4.2, mais peu évaluent la répétabilité des résultats obtenus en 2D et en 3D, et peu proposent une méthodologie adaptée à cette différence de type des données manipulées. Il s'agit donc des objectifs que nous visons dans le chapitre suivant.

Chapitre 4.

Appariement 2D/3D

4.1. Introduction sur la mise en correspondance 2D/3D

Qu'est ce que c'est ?

Il s'agit de retrouver des primitives en 2D et en 3D, comme celles définies aux § 2.1 et 3.1, qui correspondent aux mêmes entités physiques. La difficulté principale réside dans le fait que les définitions utilisées en 2D et en 3D doivent être cohérentes entre elles pour que ces primitives puissent être appariées. Bien souvent, les informations utilisées prennent peu en compte la texture ou la couleur et s'appuient sur des aspects liés à la forme des éléments à reconnaître, comme les contours occultants ou les silhouettes.

Dans les travaux que nous présentons dans cette partie, notre objectif est de caractériser les changements ou plutôt les dégradations sur des éléments de scènes urbaines en comparant un modèle 3D sans défaut et sans texture, avec une image 2D d'un objet avec ses propres caractéristiques photométriques, comme illustré dans la figure 4.1.

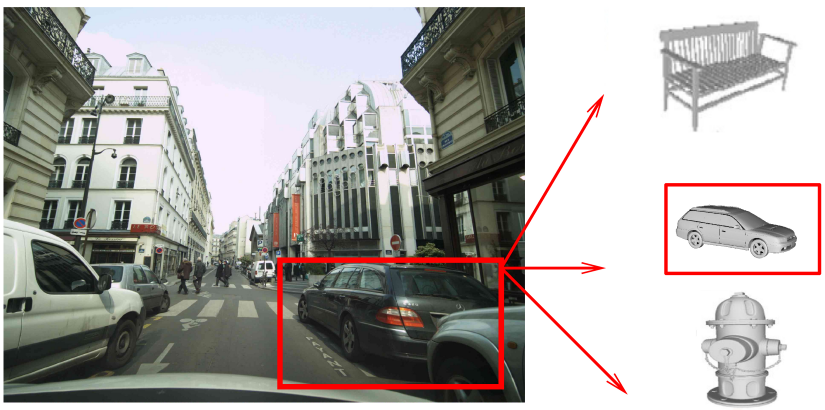


FIGURE 4.1. – Reconnaître un objet en 2D à partir d'une collection de modèles 3D – Nous avons déjà exposé le contexte applicatif dans la figure 1.2. Nous supposons que l'objet à étudier a été délimité par une boîte englobante. Notre objectif est de reconnaître cet objet et sa pose parmi une collection d'objets dont nous possédons un modèle 3D non texturé. Il est important de noter qu'en dehors de la texture, la forme de l'objet en 3D peut également être différente suivant le modèle que nous avons dans notre base de données et la photographie acquise.

Pourquoi ?

Dans le cas de la mise en correspondance de données de nature différente, un des domaines les plus concernés est celui de l'analyse et de la compréhension d'images médicales où l'appariement des données de source différente est crucial pour une meilleure prise en charge de la pathologie (détection précise des zones à soigner pour un traitement le plus adapté possible) [Nam 11]. Pour un état de l'art conséquent dans ce domaine, nous recommandons la publication suivante [Markelj 12]. En robotique, la nécessité de mettre en correspondance des données de nature différente et, notamment des données 2D avec des données 3D, est également importante [Wong 92]. Cette publication [Pomerleau 15] permet d'avoir une présentation de l'état de l'art dans ce domaine. En reconnaissance biométrique, nous pouvons également citer des travaux en reconnaissance de visages [Yang 08b]. Apparier directement une image 2D avec des données 3D peut être utile pour déterminer la localisation, c'est-à-dire le lieu exact, d'une photographie, en supposant qu'un nuage de points a été préalablement reconstruit [Irschara 09].

Dans le travail présenté, nous souhaitons reconnaître des objets à partir d'une image 2D couleur texturée et un modèle 3D sans texture.

Comment ?

Avant de détailler les différentes manières d'effectuer ce type d'appariement, nous citons un ensemble de travaux remarquables mais dont les données d'entrée sont différentes de celles que nous utilisons. Nous pensons qu'il est important de citer ces méthodes car malgré la différence soulignée, les éléments présentés et utilisés pour effectuer l'appariement se rapprochent de ce qu'il est possible de faire dans le contexte de notre étude. Certains systèmes sont capables de générer simultanément des modèles 3D ou des cartes de profondeur et des images 2D, simplifiant en partie le problème du recalage des données comme avec des capteurs RGB-D [Cao 17a]. Toujours en utilisant des capteurs RGB-D, les auteurs de [Song 14] identifient des objets en utilisant des modèles 3D mais ils ne s'appuient que sur les cartes de profondeur, ce qui permet de s'affranchir de la couleur, de la texture et, éventuellement du fond texturé. Dans [Toshev 09], des objets sont reconnus dans des vidéos, à partir de modèles 3D connus. Une segmentation du mouvement permet d'extraire chaque objet. Ensuite, la silhouette de l'objet est utilisée pour réaliser l'appariement. L'avantage de cette approche est que les auteurs possèdent de nombreux points de vue sur l'objet à reconnaître. L'opérateur proposé dans [Su 13] est générique et peut s'appliquer aussi bien en 2D qu'en 3D mais c'est la reconnaissance de formes qui est étudiée, et non directement l'appariement de ces éléments en 2D et 3D. Les auteurs font l'hypothèse d'une détection (suivie d'une phase d'échantillonnage de la forme) qui n'est pas entachée d'erreur. Une fois de plus, seul le contour extérieur de la forme est pris en compte. Dans [Hinterstoisser 11], les auteurs cherchent à reconnaître des objets non texturés. En soit, dans le contexte de la mise en correspondance classique s'appuyant sur la photométrie et la texture, cette tâche est délicate, mais ces travaux s'affranchissent de cette difficulté en proposant d'utiliser des cartes de gradients. Nous pouvons également citer une approche qui effectue en parallèle l'appariement de multiples images avec de multiples modèles 3D pour estimer la pose d'un objet en 2D et en proposer une reconstruction 3D [Huang 15]. L'originalité de l'approche vient, entre autres, du fait d'imposer une contrainte d'unicité. Enfin, certaines méthodes s'appuient sur des *a priori* sur les formes étudiées, comme des objets articulés [Ion 11].

Le problème de recalage d'objets 2D et 3D que nous nous posons est différent car nous n'avons qu'une seule image et aucune contrainte sur les objets manipulés. Ce problème est difficile car les données acquises dans les deux modalités présentent des caractéristiques très différentes. En effet, dans des photographies, l'apparence de l'objet dépend, d'une part, des caractéristiques intrinsèques de l'objet, comme la texture, la couleur, l'albédo, et d'autre part, des conditions d'acquisition, comme la pose et l'éclairage. De plus, dans une image prise dans un environnement non contrôlé, le fond (le reste de la scène qui contient l'objet) peut également posséder une texture ou des formes qui ajoutent des difficultés supplémentaires pour comparer le modèle 3D et l'image. À l'inverse, pour un modèle 3D, dans la plupart des cas, la texture de l'objet n'est pas modélisée et il n'y a pas de fond.

Recaler des données 2D et 3D correspond à résoudre à la fois un problème d'estimation de pose [Crombez 18] et un problème de correspondance. De manière très classique, pour effectuer cette tâche, des points d'intérêt peuvent être détectés, caractérisés puis appariés [Wu 08, Agarwal 11], comme des points SIFT [Crombez 18]. Dans le § 4.2, nous allons détailler les différentes approches existantes en appariement direct 2D/3D mais une autre façon de résoudre ce problème d'appariement est de considérer deux sous-problèmes [Shaiek 13] :

(1) Choisir une représentation commune à ces deux types de modalités ;

(2) Apparier les objets en utilisant cette représentation commune.

Ainsi, le problème de mise en correspondance 2D/3D est transformé en problème de mise en correspondance 2D/2D. Nous avons choisi de présenter les travaux de recherche dans ce cadre car diviser le problème en deux sous-problèmes permet de simplifier la compréhension et le traitement.

Pour conclure, la littérature fournit de nombreuses méthodes permettant de combiner des outils différents mais pas réellement d'outils permettant une analyse commune. De plus, à notre connaissance, peu d'approches tentent de qualifier quantitativement la répétabilité des outils entre diverses sources de données. Ce que nous présentons par la suite permet d'aborder ces deux aspects.

Qualité du résultat ?

Nous utilisons la même définition de la répétabilité que celle présentée dans la formulation (2.2), page 23. De plus, nous évaluerons l'intérêt des outils proposés en évaluant la qualité du recalage obtenu ainsi que la qualité de la pose estimée.

Et le descripteur ?

Comme pour la détection, les descripteurs utilisés s'appuient sur ceux existants en 2D et en 3D, comme par exemple le descripteur SIFT [Wu 08], le descripteur HoG [Aubry 14, Lim 14] ou une version moins coûteuse en terme de calculs, dans [Choy 15]. Nous pouvons également citer les travaux de [Ricard 05a, Ricard 05b] qui introduit un descripteur applicable aussi bien en 2D qu'en 3D et s'appuyant sur la transformation radiale angulaire, *Angular Radial Transform* (ART).

4.2. Mise en correspondance 2D/3D directe

Nous rappelons qu'il y a deux manières distinctes de réaliser cette tâche : effectuer une mise en correspondance directe entre 2D et 3D, c'est ce que nous décrivons dans ce paragraphe ou utiliser une représentation 2D commune, que nous détaillons au § 4.3, avant d'effectuer l'appariement, cf. § 4.4.

C'est au début des années 2000 que le problème de recalage entre une photographie 2D et un modèle 3D d'une scène a connu un essor. Dans de nombreux articles, l'objectif visé est d'obtenir la pose 3D d'un objet. De manière classique, nous pouvons classer les approches existantes en deux familles : les méthodes indirectes et les méthodes directes [Paudel 14]. Nous parlons de **recalage indirect**, lorsque l'approche proposée s'inspire des approches de recalage 3D-3D ou lorsque l'objectif est d'estimer des paramètres pour effectuer un recalage global, comme, par exemple, dans l'approche standard : *Iterative Closest Point*, ICP [Besl 92]. Ces approches peuvent donc être assimilées à des approches globales, elles ne fournissent pas un appariement point à point.

Pour les approches par **recalage direct**, bien évidemment, des approches se sont limitées à définir des correspondances de manière manuelle [Dellepiane 07] mais cela n'a pas d'intérêt pour réaliser une mise en correspondance à grande échelle dans une base de données conséquente. Nous pouvons citer des approches assez naturelles et proches de ce qui est fait en 2D comme les approches de [Wu 08, Meierhold 10, Sattler 11, Agarwal 11, Lee 13b] où les correspondances sont obtenues en exploitant directement les points SIFT obtenus en 2D et en 3D. Il est également possible d'utiliser une analyse multi-spectrale [Ricard 05a]. Dans les travaux de [Lee 13c], il est simplement supposé

que les objets étudiés possèdent peu de texture, ce qui est peu réaliste. Une autre façon de traiter le problème des difficultés d'appariement et de choisir des primitives plus structurées comme des lignes [Kamgar-Parsi 11, Xu 17], des plans [Tamaazousti 11] ou des boîtes englobantes [Liu 05]. C'est le cas en imagerie médicale où les auteurs de [Cyr 00] utilisent essentiellement les contours des objets. Toutes ces primitives sont bien adaptées dans un contexte de scènes dites *Manhattan World*[1]. D'autres techniques utilisent des hypothèses fortes sur la scène ou les objets observés comme la méthode de la ligne d'horizon de [Ramalingam 09] ou une approche exploitant un modèle 3D prédéfini, comme dans [Clarkson 01]. Toutes les méthodes citées sont dédiées au contexte d'une application bien précise. D'autres approches proposent de s'appuyer sur les courbes qui présentent des propriétés particulières en terme de géométrie différentielle de la surface. Par exemple, dans [Ohtake 04], les courbes sont supposées correspondre à des paraboles, *paraboloic lines*, qui doivent se situer sur des points de courbure extrême. Ceci permet de prendre en compte des propriétés caractéristiques de la surface de l'objet mais ces critères ne sont pas invariants lorsque le point de vue change. Des approches s'attachent à observer le contexte de la forme en 2D et en 3D [Belongie 02] mais les données utilisées sont simples : les objets en 2D sont binaires et il n'y a pas de fond. Nous pouvons noter que dans tous ces travaux, la répétabilité des primitives obtenues n'est pas étudiée.

4.3. Choix d'une représentation commune des données 2D et 3D

Représentations communes possibles – Nous pouvons envisager deux types de rendu : sous la forme d'une image de synthèse [Campbell 01, Rosenhahn 04] ou sous la forme d'une carte de profondeur [Bo 11, Darom 12, Choy 15]. Enfin, plus récemment, [Plötz 17] a proposé l'utilisation d'un gradient moyenné, *average shading gradients*. Cette technique consiste à calculer la moyenne des normales des gradients en supposant un ensemble d'éclairages possibles. En effet, pour tenter de prendre en compte l'inconnue sur la source d'éclairage, les auteurs envisagent un ensemble d'éclairages couvrant l'ensemble (discrétisé) des directions d'éclairage envisageables. Cette technique astucieuse pour gérer l'inconnue sur l'éclairage ne prend pas directement en compte la variabilité liée à la couleur et à la texture. Elle ne permet pas non plus de prendre en compte la présence d'un fond texturé, et donc de diminuer l'impact de ce fond sur la caractérisation de la forme à reconnaître.

Représentation commune choisie – Ainsi, parmi ces 3 manières de représenter un objet 3D en 2D, nous avons choisi les cartes de profondeur. En effet, nous pensons que ce type de représentation permet de mieux capturer les formes significatives de l'objet, sans chercher à effectuer un quelconque rendu synthétique d'informations de couleur ou de texture (que nous ne possédons pas dans les modèles manipulés). L'aspect délicat de cette représentation réside dans le choix de l'ensemble des points de vue pour générer les cartes de profondeur. Cet aspect sera abordé au § 4.11.

4.4. Appariement de données 2D d'origine différente

Après avoir choisi une représentation commune, il est nécessaire de **mettre en correspondance des primitives entre deux modalités différentes dans une représentation commune**. De manière générale, un appariement peut être réalisé de manière partielle ou dense [Crombez 18]. De plus, la mise en correspondance peut être locale ou globale [Scharstein 02a]. Les primitives appariées peuvent être plus ou moins élaborées : points d'intérêt, cf. chapitres 2 et 3, contours [Lee 07], segments, courbes [Judd 07]. Certaines approches combinent différents types de primitives, comme dans [Bo 11].

1. Terme apparu en vision par ordinateur dans [Coughlan 99], il suppose que deux directions quelconques de la scène sont soit parallèles soit perpendiculaires.

Méthodes existantes – Dans le cadre de la mise en correspondances locale de données 2D de nature différente, nous pouvons citer l'approche de [Rosenhahn 04] qui utilise un appariement dense par corrélation, cf. chapitre 5. Nous pouvons également citer l'approche de [Campbell 01] où les silhouettes sont extraites dans les images de synthèse et les images réelles afin d'être appariées. Cette approche d'appariement local et partiel ne permet pas de prendre en compte les contours occultants ou les contours significatifs à l'intérieur de la forme, ce qui nous semble indispensable pour caractériser au mieux un objet. Dans [Godil 11], les auteurs introduisent la notion de saillance dans des maillages. Cette saillance est estimée en utilisant la courbure moyenne à différentes résolutions. Dans [Judd 07], la notion de lignes de crêtes apparentes, *apparent ridges*, AR, est introduite. Il s'agit de l'ensemble de courbes dont les points correspondent à des maxima locaux de la surface. Ces travaux sont proches de ce que nous utilisons pour détecter des contours d'intérêt dans les données que nous manipulons, c'est pourquoi, par la suite, nous comparerons nos résultats avec cette approche. Dans tous ces travaux, la variabilité des données manipulées n'est pas prise en compte.

Finalement, dans [Plötz 17], déjà introduit au § 4.3, les coins sont détectés en 3D en utilisant le détecteur de Harris en 3D. Pour caractériser les points, les auteurs utilisent les images de gradients moyens obtenus sur les images de rendu. Ainsi, pour chaque image requête, de manière similaire, ils détectent des coins, en multi-échelle, et caractérisent le point en prenant les gradients calculés dans le voisinage. Ainsi, chaque point 2D est apparié avec tous les points de la base de données, en utilisant le descripteur HoG. Malheureusement, tous les calculs de gradient sont sensibles à la présence de texture, et en particulier d'un fond, dans les images optiques. En conséquence, les auteurs introduisent une étape de raffinage s'appuyant sur un processus de *RANdom SAmple Consensus*, RANSAC [Fischler 81] pour améliorer la pose ainsi estimée. Cette approche reste la plus proche de ce que nous introduisons par la suite, et nous nous comparerons à elle après avoir donné tous les détails significatifs.

Proposition – Une propriété attendue est que **la détection, ainsi que la description, soient répétables entre l'image et la carte de profondeur de l'objet obtenu suivant le même point de vue**. Intuitivement, nous pensons qu'entre ces deux types de représentation, l'élément commun est la forme de l'objet que nous avons associée à la notion de courbure. Ainsi, nous arrivons à notre proposition : étudier ce que nous appelons les formes curvilignes ou plus précisément, la saillance curviligne. La notion de primitives liées à de régions curvilignes est largement détaillée dans la thèse de [Lemaitre 08]. L'intérêt est de considérer à la fois les contours des silhouettes des objets, mais également, les contours significatifs à l'intérieur de l'objet d'intérêt, également appelés dans la littérature les contours externes et internes, *outer and inner contours*, ou occultant et auto-occultants, *occluding and self-occluding*. Nous souhaitons utiliser ces primitives car, intuitivement, nous estimons que ce sont les plus robustes en terme de variations d'intensité, de couleur et de pose. De plus, nous allons exploiter à la fois la notion de saillance curviligne et d'orientation de la courbure. Pour fusionner ces deux informations, nous proposons une nouvelle formulation du descripteur HoG, cf. § A.2. Nous introduisons également un score de répétabilité dans le descripteur utilisé.

Plus précisément, nous souhaitons utiliser des primitives qui utilisent l'information liée aux gradients de l'image car nous supposons qu'ils sont significatifs sur les contours de l'objet, en proposant une méthode pour rendre l'exploitation de ces gradients robuste aux changements d'illumination et à la présence de texture, dans les photographies. Pour cela, nous faisons deux hypothèses :

(1) Lorsque les contours, liés à la notion de courbure, sont générés par la forme, et non par la texture ou les variations d'éclairage, ils peuvent être détectés dans différentes échelles. Nous allons exploiter cette redondance de détection pour distinguer les contours liés à la forme de ceux liés à la texture ou à la variation d'éclairage.

(2) Lorsque nous photographions un objet d'intérêt, la mise au point est réalisée sur cet objet. Nous pouvons supposer que les contours seront les plus nets au niveau de la forme. Nous allons donc exploiter le même principe que les algorithmes qui quantifie le degré de flou en chaque pixel en supposant que plus un contour est net, plus il a de chance d'être un contour significatif.

Pour la suite, nous présentons l'ensemble des éléments qui composent la méthode proposée, cf. § 4.5. Nous donnerons ainsi les éléments permettant d'analyser la surface du modèle 3D, cf. § 4.6, avant de présenter l'analyse similaire en 2D, cf. 4.7. Pour compléter l'analyse en 2D, nous allons présenter les traitements supplémentaires effectués, cf. § 4.8. Ils prennent en compte, d'une part des aspects multi-échelle, § 4.8.1, et, d'autre part des aspects liés à la notion de *focus*, § 4.8.4. Enfin, nous présenterons les résultats obtenus sur la détection de primitives, § 4.9, le recalage 2D/3D, § 4.10 et l'estimation de la pose, § 4.11.

4.5. Résumé de l'approche proposée

L'approche est présentée dans la figure 4.2. Cela nous permet de visualiser la manière dont les différents éléments proposés sont combinés pour donner le résultat final. Nous avons décidé de détecter les éléments saillants du point de vue de la courbure, aussi bien en 2D, qu'en 3D. C'est ce que nous appelons la saillance curviligne, *Curvilinear Saliency*, CS. Toutefois, nous cherchons à proposer une approche robuste au fond et, de manière générale, à la texture présente dans les images. Ainsi, nous introduisons l'exploitation de cette saillance à différentes échelles, d'où le terme de saillance curviligne multi-échelle, *Multiscale Curvilinear Saliency*, MCS. Nous utilisons également la notion de *focus* afin d'introduire ce que nous appelons les courbes de *focus* multi-échelle, *Multi-Focus Curves*, MFC.

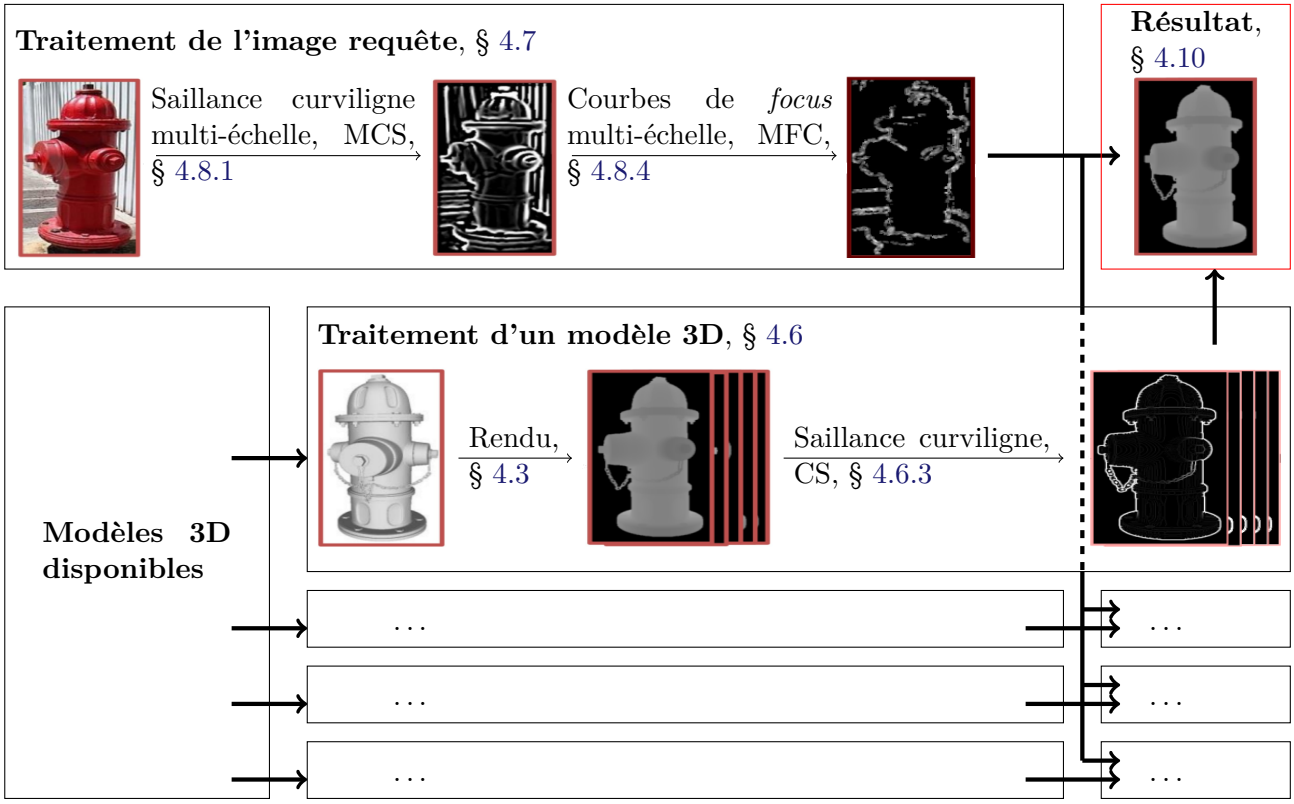


FIGURE 4.2. – Apparier une image avec des modèles 3D – Afin d'être en mesure de comparer des images avec des modèles 3D, nous utilisons une collection de cartes de profondeur obtenues suivant différents points de vue. Ensuite, nous détectons des primitives d'intérêt dans les deux modalités. Puis, chaque carte de profondeur de chaque modèle est comparée avec l'image requête afin de déterminer la carte de profondeur la plus proche. Nous obtenons donc pour chaque modèle 3D, la pose correspondante à l'image en 2D. Enfin, l'objet est reconnu en sélectionnant la carte de profondeur la plus proche de l'image 2D parmi toutes les poses identifiées (le cadre rouge). La façon de choisir l'objet reconnu a été détaillée dans notre publication [Rashwan 17], cf. § 4.10.

4.6. Extraction de primitives d'intérêt en 3D

Nous allons introduire la façon de représenter les surfaces en 3D en mettant en évidence les différences significatives avec le modèle introduit par [Plötz 17]. Pour le lecteur intéressé, des rappels de géométrie différentielle sont donnés en annexe A.3.

4.6.1. Représentation de la surface d'un objet

Posons $\mathcal{M}$ la surface d'un objet paramétré[2] par $\mathbf{M}(\mathbf{x}) \triangleq [X(\mathbf{x}), Y(\mathbf{x}), Z(\mathbf{x})]^\top$, où $\mathbf{x} = [x, y]^\top$ parcourt le domaine restreint de l'image produite par une caméra donnée et délimitée par les contours occultants d'un objet. Plus précisément, nous supposons que $\mathcal{M}$ est tel que chacun des points qui lui appartient, $\mathbf{M}(\mathbf{x})$, c'est-à-dire visible suivant le point de vue de la caméra, est en correspondance point à point de manière perspective avec un point de l'image $\mathbf{x} = [x, y]^\top$, de la manière suivante : $x = X(\mathbf{x})/Z(\mathbf{x})$ et $y = Y(\mathbf{x})/Z(\mathbf{x})$. En conséquence, nous obtenons :

$$\mathbf{M}(\mathbf{x}) = Z(\mathbf{x})[\mathbf{x}^\top, 1]^\top. \tag{4.1}$$

Soit $\mathbf{n}$ la normale unitaire de $\mathcal{M}$ en $\mathbf{P} = \mathbf{M}(\mathbf{x})$. La carte gaussienne $\mathbf{N} : \mathcal{M} \to \Sigma$ de $\mathcal{M}$ en $\mathbf{P}$ est la carte qui assigne à $\mathbf{P}$ le vecteur $\mathbf{N}(\mathbf{P}) = \pm\mathbf{n}$ sur la sphère unitaire Σ telle que $\mathbf{N}$ soit différentiable. En utilisant la notation $\bar{\mathbf{N}}(\mathbf{P}) = \mathbf{M}_x(\mathbf{x}) \times \mathbf{M}_y(\mathbf{x})$ pour $\mathbf{x} = \mathbf{M}^{-1}(\mathbf{P})$, cette carte peut être estimée avec :

$$\mathbf{N}(\mathbf{P}) = \frac{\bar{\mathbf{N}}(\mathbf{P})}{\left\|\bar{\mathbf{N}}(\mathbf{P})\right\|}, \tag{4.2}$$

où :

$$\bar{\mathbf{N}} = \mathbf{M}_x \times \mathbf{M}_y = Z\left[-Z_x, -Z_y, xZ_x + yZ_y + Z\right]^\top. \tag{4.3}$$

Nous pouvons montrer que la matrice jacobienne 3×2 de $\mathbf{N}$ s'écrit :

$$\mathtt{J}_{\mathbf{N}} = \begin{bmatrix}\mathbf{N}_x & \mathbf{N}_y\end{bmatrix} = \left(\mathtt{I} - \mathbf{N}\mathbf{N}^\top\right)\mathtt{J}_{\bar{\mathbf{N}}}, \tag{4.4}$$

où les colonnes de $\mathtt{J}_{\bar{\mathbf{N}}} = \begin{bmatrix}\bar{\mathbf{N}}_x & \bar{\mathbf{N}}_y\end{bmatrix}$ ont cette forme :

$$\bar{\mathbf{N}}_\star = \begin{bmatrix} Z_x Z_\star - Z_{x\star} Z \\ Z_\star Z_y - Z_{\star y} Z \\ xZ_{x\star}Z + yZ_{\star y}Z + Z_\star(xZ_x + yZ_y + 3Z)\end{bmatrix} \tag{4.5}$$

et où $\star$ représente x ou y.

4.6.2. Gradient d'ombrage moyen, *Average Shading Gradient*, ASG [Plötz 17]

[Plötz 17] assume que la fonction d'intensité d'une image, dans le cas d'une source de lumière parallèle $\mathbf{s}$ dans $\mathbb{R}^3$, est donnée par :

$$I(x, y) \propto \max(0, -\mathbf{N}(x, y) \cdot \mathbf{s}) \text{ avec } \mathbf{s} \in \mathbb{R}^3 \tag{4.6}$$

L'équation (4.6) signifie que la réflectance qui décrit le matériau de l'objet est supposée lambertienne avec un albédo constant[3]. De plus, une hypothèse importante est faite : **le fond de l'image doit**

2. Attention, dans cette section, nous utilisons les lettres calligraphiées pour désigner une surface régulière, ce qui ne suit pas les notations annoncées en page xxiii.

3. Une fonction d'ombrage générale est définie par $I(x, y) = \rho(\mathbf{M}(x, y)) \max(0, -\mathbf{N}(x, y) \cdot \mathbf{s})$, où $\rho(\mathbf{M}(x, y))$ est l'albédo du point de l'objet $\mathbf{M}(x, y)$.

être constant, comme c'est le cas d'un plan. Ainsi, les auteurs proposent que l'amplitude du gradient de la fonction d'ombrage soit la primitive utilisée dans les images 2D. Ensuite, pour réaliser le recalage avec les données 3D, représentées par un modèle 3D non texturé, le principe est de générer des vues virtuelles de l'objet de différents points de vue. Or, il est clairement impossible de générer de telles vues virtuelles qui suivent les fonctions d'ombrage données par (4.6) sans avoir d'information préalable sur la direction de l'éclairage et donc, plus généralement, sans information sur $\mathbf{s}$. C'est la raison pour laquelle les auteurs proposent de remplacer l'amplitude du gradient dans les images virtuelles par le calcul de la primitive correspondante, à savoir la valeur moyenne de l'amplitude du gradient calculée sur toutes les orientations possibles de l'éclairage. C'est cela que les auteurs appellent : *Average Shading Gradient magnitude*. Si nous notons $\|\nabla I\|$ l'amplitude du gradient de la fonction d'ombrage (4.6) alors l'amplitude du gradient d'ombrage moyen est donnée par :

$$\overline{\|\nabla I\|} = \int_{\mathcal{S}} \|\nabla I\| \, \mathrm{d}\mathbf{s} \tag{4.7}$$

avec $\|\nabla I\|^2 = I_x^2 + I_y^2$ et où le vecteur $\mathbf{s}$, cf. (4.6), varie suivant la sphère unitaire $\mathcal{S}$ dans $\mathbb{R}^3$, et $\mathrm{d}\mathbf{s}$ est l'élément de volume. En appliquant les inégalités de Jensen pour dériver la forme fermée suivante sur $\overline{\|\nabla I\|}$, les auteurs obtiennent :

$$\begin{aligned} \overline{\|\nabla I\|} &\leq \sqrt{\int_{\mathcal{S}} \|\nabla I\|^2 \, \mathrm{d}\mathbf{s}} \\ &= \gamma \sqrt{\left(\|\mathbf{N}_x\|^2 + \|\mathbf{N}_y\|^2\right)} \text{ avec } \gamma = \sqrt{\frac{\pi}{3}} \end{aligned} \tag{4.8}$$

Les auteurs remarquent que cette expression permet une très bonne approximation de $\overline{\|\nabla I\|}$. Il s'agit de la contribution remarquable de [Plötz 17] pour prendre en compte l'inconnue sur la direction de l'éclairage $\mathbf{s}$.

4.6.3. Primitives de saillance curviligne, *Curvilinear Saliency*, CS

Comme nous l'avons déjà annoncé, notre but est de trouver une représentation commune entre les données 2D, les images optiques, et les données 3D, un modèle 3D *a priori* sans information de texture. Nous avons besoin de cette représentation commune pour mettre en correspondance ces données, dans le but de reconnaître des objets. Ainsi, dans un premier temps, nous allons détailler la représentation 3D utilisée, comment cette représentation permet de représenter un modèle 3D de différents points de vue et comment cette représentation est adaptée pour comparer aux données 2D.

Premièrement, les données 3D sont représentées par un ensemble de cartes de profondeur générées pour un ensemble donné de positions de la caméra. Une carte de profondeur notée $Z(x, y)$ associe à chaque point de l'image (x, y) la coordonnée Z, relative à la caméra, du point de l'objet 3D, cf. expression (4.1), qui se projette en (x, y). Les différentes positions sont choisies sur un ensemble de sphères concentriques et en échantillonnant les angles d'élévation et d'azimut ainsi que la distance entre la caméra et l'objet. Notons $\mathcal{D}$ la profondeur de la surface qui correspond à la surface 3D dont le paramétrage est donné par :

$$\mathbf{D}(x, y) = [x, y, Z(x, y)]^\top . \tag{4.9}$$

Il est important de noter la différence avec (4.1). De plus, nous devons remarquer que deux surfaces de profondeur, c'est-à-dire deux cartes de profondeur, résultat de la projection suivant deux points de vue différents, ne sont pas égales au sens d'une transformation euclidienne.

À présent, que pouvons nous dire sur les points remarquables que nous souhaitons obtenir sur une carte de profondeur ?

C'est ici que nous introduisons la notion de courbures principales. Nous notons κ_1 et κ_2 les cour-

bures principales de la surface. Nous souhaitons détecter les discontinuités de profondeur en cherchant les points de $\mathcal{D}$ qui ont une courbure principale élevée dans une direction et une courbure principale faible dans la direction orthogonale. De manière intuitive, nous cherchons à trouver des primitives le long d'éléments de surface « allongés ». Nous appelons cette notion la saillance curviligne ou *curviness saliency*, CS. Ces points correspondent à ce qui est appelé couramment dans la littérature des crêtes, *ridges* et des vallées, *valleys*, de la surface [Crowley 84]. Ces courbures principales peuvent être combinées de façons différentes mais nous avons choisi de les combiner en utilisant la différence, soit : $\kappa_1 - \kappa_2$. Nous expliquons nos motivations dans les paragraphes suivants.

Courbures et directions principales

Nous considérons un point $\mathbf{P} = \mathbf{D}(x, y)$ et $\mathbf{N}'(x, y)$ la carte gaussienne associée à $\mathcal{D}$. Cette carte assigne à $\mathbf{P}$ la normale unitaire de $\mathcal{D}$:

$$\mathbf{N}' = \frac{\bar{\mathbf{N}}'}{\left\|\bar{\mathbf{N}}'\right\|} \text{ où } \bar{\mathbf{N}}' = \mathbf{D}_x \times \mathbf{D}_y = \alpha \begin{bmatrix} -\nabla Z \\ 1 \end{bmatrix} \tag{4.10}$$

avec $\nabla Z = [Z_x, Z_y]^\top$ et $\alpha = 1/\sqrt{1 + \|\nabla Z\|^2}$.

Nous avons les deux colonnes de la matrice jacobienne $\mathtt{J}_\mathbf{D}$ de $\mathcal{D}$ définie par : $\mathbf{D}_x = [1, 0, Z_x]^\top$ et $\mathbf{D}_y = [0, 1, Z_y]^\top$. Ainsi, la **première forme fondamentale de** $\mathcal{D}$ peut être calculée de cette façon :

$$\mathtt{I}_\mathbf{P} = \mathsf{I}_3 + \nabla Z \nabla Z^\top \tag{4.11}$$

et la **seconde forme fondamentale de** $\mathcal{D}$ par :

$$\mathtt{II}_\mathbf{P} = \alpha \mathtt{H}_Z \tag{4.12}$$

où $\mathtt{H}_Z$ est la matrice hessienne de Z, c'est-à-dire la matrice formée des dérivées secondes partielles de Z en x et y.

Les **courbures principales** de $\mathcal{D}$ en $\mathbf{P}$ correspondent aux valeurs propres κ_α ($\alpha = 1, 2$) de $\mathtt{I}_\mathbf{P}^{-1}\mathtt{II}_\mathbf{P}$. Ces valeurs propres sont toujours réelles. Dans le plan tangent $T_\mathbf{P}(\mathcal{D})$, les coordonnées locales des **directions principales** de $\mathcal{D}$ en $\mathbf{P}$ sont données par les vecteurs propres $\mathbf{e}_\alpha$ de $\mathtt{I}_\mathbf{p}^{-1}\mathtt{II}_\mathbf{p}$, soient les directions 3D principales qui s'écrivent $\mathtt{J}_\mathbf{D}\mathbf{e}_\alpha$. [Koenderink 92] a écrit : "*it is perhaps not superfluous to remark here that the simple (eigen-)interpretation in terms*[4] *of* $\mathtt{II}_\mathbf{P} = \alpha \mathtt{H}_Z$ *is only valid in representations where* $\nabla Z = \mathbf{0}$", ce qui est la condition nécessaire pour que le point concerné soit considéré comme un extremum local.

Grâce à la proposition 1, présentée à la page 255 de l'annexe A.3, nous savons que la courbure principale κ_α en $\mathbf{P}$ associée à la direction 3D principale donnée par $\mathbf{T}_\alpha = \mathtt{J}_\mathbf{D}\mathbf{e}_\alpha$ est égale à l'amplitude en absolu du changement sur la normale, soit :

$$|\kappa_\alpha| = \left\|\mathrm{d}\mathbf{N}'_\mathbf{P}(\mathbf{T}_\alpha)\right\| \tag{4.13}$$

où $\mathrm{d}\mathbf{N}'_\mathbf{P}(\mathbf{T})$ est la différentielle de $\mathbf{N}'$ en $\mathbf{P}$ dans la direction $\mathbf{T}$. C'est ce résultat que nous allons utiliser pour la représentation 2D utilisée par la suite, cf. §4.7. À présent, nous allons justifier le choix $\kappa_1 - \kappa_2$, avec $\kappa_1 \geq \kappa_2$, comme critère pour détecter des primitives d'intérêt.

Primitive curviligne

Sans perte de généralité, nous posons à présent que κ_1 et κ_2 sont les courbures principales ordonnées obtenues par calcul des valeurs propres de $\mathtt{I}_\mathbf{p}^{-1}\mathtt{II}_\mathbf{p}$ avec $\kappa_1 \geq \kappa_2$. Nous rappelons que nous souhaitons

4. i.e., en négligeant $\mathtt{I}_\mathbf{P}$.

détecter les points qui se trouvent sur des éléments de surface « allongés ». Ainsi, nous détectons les points tels que cette différence soit élevée :

$$CS(x,y) = \kappa_1(x,y) - \kappa_2(x,y). \tag{4.14}$$

L'expression (4.14) est ce que nous appelons la saillance curviligne, *Curvilinear Saliency*, CS. Curviligne signifie qui appartient à une ligne courbe. La justification de ce choix est donnée par la suite.

Soit $\mathbf{P}$ un point sur $\mathcal{D}$, nous posons $(\tilde{x},\tilde{y})$ les coordonnées cartésiennes dans le plan tangent $T_{\mathbf{P}}(\mathcal{D})$. Cela signifie que nous prenons $\mathbf{P}$ comme origine et nous considérons la base orthonormée formée par les directions principales $\{\mathbf{e}_1,\mathbf{e}_2\}$. En conséquence, $\mathcal{D}$ peut être localement assimilée à la nouvelle paramétrisation $\mathbf{F}(\tilde{x},\tilde{y}) = [\tilde{x},\tilde{y},F(\tilde{x},\tilde{y})]^\top$, pour une fonction d'élévation F. Dans ce cas précis, nous pouvons montrer que $\mathtt{I}_{\mathbf{P}}$ est la matrice identité et, donc, $\mathtt{I}_{\mathbf{P}}^{-1}\mathtt{II}_{\mathbf{P}} = \mathtt{II}_{\mathbf{P}} = \mathrm{diag}(\kappa_1,\kappa_2)$ est exactement la matrice hessienne de F. Pour un ϵ strictement positif et suffisamment petit, nous considérons, sur les deux plans parallèles à $T_{\mathbf{P}}(\mathcal{D})$ aux distances $\pm\epsilon$ de $T_{\mathbf{P}}(\mathcal{D})$, les courbes $\mathscr{C}_\pm = \{(\tilde{x},\tilde{y}), \mathbf{F}(\tilde{x},\tilde{y}) \in T_{\mathbf{P}}(\mathcal{D}) \mid F(\tilde{x},\tilde{y}) = \pm\epsilon\}$. Il a été montré [Gallier 01, p500] que l'approximation de premier ordre des intersections de $\mathcal{D}$ avec ces deux plans parallèles est l'union des deux coniques (une réelle et une virtuelle) d'équations $\mathtt{II}_{\mathbf{P}}(\tilde{x},\tilde{y}) = \pm 2\epsilon$.. Cette union est connue comme étant l'*indicatrice de Dupin* quand celle-ci est écrite sous la forme canonique (c'est-à-dire en remplaçant 2ϵ par 1).

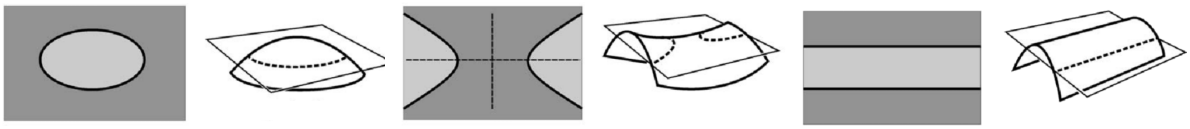


FIGURE 4.3. – Coniques réelles associées à l'indicatrice de Dupin.

La conique réelle de Dupin caractérise la forme locale de $\mathcal{D}$ et fournit donc une information locale sur la géométrie de premier ordre de la surface, au moins pour les points où la conique n'est pas dégénérée. Comme illustré dans la figure 4.3, cette conique se spécialise en :

- **Parabole** si la courbure gaussienne est nulle, soit $\kappa_1\kappa_2 = 0$;
- **Ellipse** si $\kappa_1\kappa_2 > 0$ et
- **Hyperbole** si $\kappa_1\kappa_2 < 0$.

Ainsi, les points sont elliptiques, hyperboliques ou paraboliques. Des détails sont donnés dans l'annexe A.3.4. La saillance curviligne CS est significative quand $\kappa_1 \gg \kappa_2$, ce qui veut dire que nous sommes en présence de deux foyers éloignés l'un de l'autre et donc d'une ellipse très allongée ou d'une hyperbole très « écrasée », cf. figure 4.3. Ces cas correspondent au cas où le point se trouve sur une discontinuité de profondeur. À l'inverse, lorsque $\kappa_1 \simeq \kappa_2$, la conique se rapproche d'un cercle et la distance entre les deux foyers est très petite.

Une façon simple de calculer une primitive curviligne

Après quelques manipulations algébriques, nous pouvons montrer que : $\mathtt{I}_{\mathbf{P}}^{-1}\mathtt{II}_{\mathbf{P}} = \frac{1}{\alpha}\mathtt{M}$ où

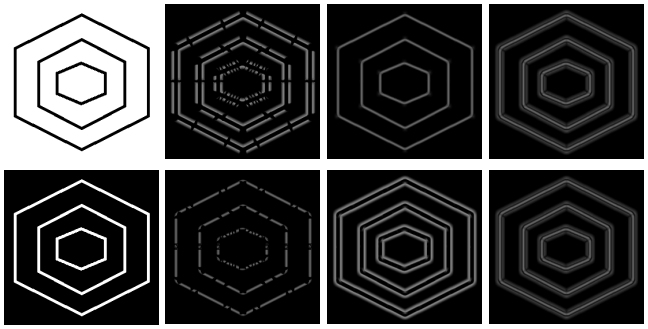
$$\mathtt{M} \triangleq \begin{bmatrix} (Z_y^2+1)Z_{xx} - Z_xZ_yZ_{xy} & (Z_y^2+1)Z_{xy} - Z_xZ_yZ_{yy} \\ (Z_x^2+1)Z_{xy} - Z_xZ_yZ_{xx} & (Z_x^2+1)Z_{yy} - Z_xZ_yZ_{xy} \end{bmatrix}$$

Proposition 1 *Le carré de la saillance curviligne, telle que nous l'avons définie, est donné par :*

$$\begin{aligned} \mathrm{CS}^2 &\triangleq \|\nabla Z\|^2 \left((\mathrm{trace}\,\mathtt{M})^2 - 4\det\mathtt{M} \right) \\ &= 4\|\nabla Z\|^2 (\bar{\kappa}^2 - K) \end{aligned} \tag{4.15}$$

où $\bar{\kappa}$ est la courbure moyenne de $\mathcal{D}$, et K est la courbure gaussienne (une preuve est fournie en annexe A.3.5).

Nous pouvons noter que l'utilisation de la courbure principale la plus élevée ou la plus haute, seule, n'est pas adaptée pour définir de manière fiable la notion de crête [Rashwan 16]. Sur la figure 4.4, nous montrons différentes détections obtenues en utilisant seulement la courbure principale minimale ou seulement la courbure principale maximale, comme proposé par [Deng 07]. Nous voyons que la courbure maximale fournit une réponse élevée seulement en présence d'une ligne sombre sur un fond clair alors que la courbure minimale, à l'inverse, fournit une réponse maximale en présence d'une ligne claire sur un fond sombre. La différence entre courbures maximale et minimale, $\kappa_1 - \kappa_2$, permet bien évidemment d'obtenir une réponse élevée dans les deux cas.



(a) (b) (c) (d)

FIGURE 4.4. – Illustration de la saillance curviligne de deux formes – En (a), nous observons la forme à étudier. Le résultat de détection de primitives d'intérêt à partir du minimum parmi les deux valeurs propres obtenues est donné en (b), alors qu'en qu'en (c), il s'agit du maximum, et qu'en (d), il s'agit de la différence.

4.7. Représentation image

4.7.1. Primitives curvilignes proposées pour les images

De manière similaire au travail de [Plötz 17], nous supposons un modèle d'éclairage lambertien suivant l'expression (4.6) c'est-à-dire $I(x,y) \propto \max(0, -\mathbf{N}(x,y) \cdot \mathbf{S})$. Nous rapellons que la normale unitaire est donnée par $\mathbf{N}(x,y) = \bar{\mathbf{N}}(x,y) / \left\| \bar{\mathbf{N}}(x,y) \right\|$ où $\bar{\mathbf{N}}$ est définie par l'expression (4.3) et dépend donc uniquement de la profondeur $Z(x,y)$ et de ses dérivées à l'ordre 1.

Notre objectif est de détecter des points d'intérêt sur la surface d'intensité de $\mathcal{I}$. Ces primitives curvilignes sont extraites sur la surface de profondeur $\mathcal{D}$ relative à la position de la caméra. De plus, nous souhaitons vérifier que ce sont de bons candidats pour être mis en correspondance. Plus précisément, les primitives détectées dans $\mathcal{I}$ doivent être similaires aux primitives détectées sur $\mathcal{D}$. Cette condition peut être assurée si dans les deux cas les calculs sont réalisés avec le même ordre de dérivation sur $Z(x,y)$. En effet, c'est ce qui peut nous garantir d'avoir des caractéristiques « compatibles » pour réaliser l'appariement. Assurer cette forme de compatibilité doit permettre d'assurer une certaine propriété de répétabilité. Le fait que I dépende de $Z(x,y)$ et de ses dérivées à l'ordre 1 implique que la détection des primitives dans $\mathcal{I}$ doit être relative à des variations de la surface $\mathbf{I}(x,y)$ d'ordre 1, c'est-à-dire relative à sa différentielle le long de la direction appropriée. Nous considérons un point $\mathbf{Q} = \mathbf{I}(x,y)$ sur la surface image et $\mathrm{d}\mathbf{I}_\mathbf{Q} : U \to \mathbb{R}^3$ comme la différentielle de $\mathbf{I}$ en $\mathbf{Q}$. De plus, nous nous donnons une direction unitaire $\mathbf{v} = [a,b]^\top$ dans le plan image xy. Enfin, nous avons $\mathrm{d}\mathbf{I}_\mathbf{Q}(\mathbf{v}) = a\mathbf{I}_x + b\mathbf{I}_y = \mathrm{J}_\mathbf{I}\mathbf{v}$

comme étant la matrice jacobienne de $\mathbf{I}$, $\mathbf{I}_x = [1, 0, I_x]^\top$ et $\mathbf{I}_y = [0, 1, I_y]^\top$, avec

$$I_\star = \frac{1}{2}\left(\operatorname{sgn}\mathbf{N}\cdot\mathbf{s} - 1\right)\left(\mathbf{N}_\star\cdot\mathbf{s}\right) \tag{4.16}$$

où $\star$ représente soit x, soit y. Il s'agit bien d'un calcul à l'ordre 1 de la variation de la surface image en $\mathbf{Q}$ et ce calcul est bien compatible avec le calcul curviligne que nous avons proposé d'utiliser pour la surface de profondeur, c'est-à-dire que nous avons bien le même ordre de dérivation sur Z.

Afin d'obtenir un calcul scalaire, nous définissons également les vecteurs unitaires suivants : $\mathbf{T}_1 = \mathtt{J}_\mathbf{I}\frac{\nabla I}{\|\nabla I\|}$ ainsi que $\mathbf{T}_2$ en effectuant une rotation de $\mathbf{T}_1$ par $\frac{\pi}{2}$. Pour $\alpha = 1, 2$, nous définissons :

$$|\mu_\alpha| = \|\mathrm{d}\mathbf{I}_\mathbf{Q}\left(\mathbf{T}_\alpha\right)\| \tag{4.17}$$

qui est la différentielle de $\mathbf{I}$ le long de la direction unitaire $\mathbf{T}_\alpha$ dans le plan image. Il est possible de montrer aisément que $\nabla I/\|\nabla I\|$ est le vecteur propre de

$$\begin{aligned}\mathtt{J}_\mathbf{I}^\top\mathtt{J}_\mathbf{I} = \begin{bmatrix}\mathbf{I}_x\cdot\mathbf{I}_x & \mathbf{I}_x\cdot\mathbf{I}_y\\ \mathbf{I}_x\cdot\mathbf{I}_y & \mathbf{I}_y\cdot\mathbf{I}_y\end{bmatrix} &= \begin{bmatrix}1+(I_x)^2 & I_xI_y\\ I_xI_y & 1+(I_y)^2\end{bmatrix}\\ &= \mathtt{I} + \nabla I\nabla I^\top\end{aligned} \tag{4.18}$$

associé à la plus grande valeur propre μ_α. Nous pouvons alors remarquer la similarité avec l'expression des courbures principales de la surface de profondeur (4.17). De plus, la matrice (4.18) est la première forme fondamentale de $\mathcal{I}$. Clairement, les valeurs maximale et minimale de la forme quadratique suivante $\|\mathrm{d}\mathbf{I}_\mathbf{Q}(\mathbf{v})\|^2$ correspondent aux deux valeurs propres de la matrice de la première forme fondamentale donnée par (4.18) [5]. Par une approche similaire à celle décrite dans le § 4.6.3, nous pouvons proposer comme opérateur de détection $\mu_1 - \mu_2$, avec $\mu_1 \geq \mu_2$.

Proposition 2 *Soit μ_1 et μ_2 les deux valeurs propres de la matrice de la première forme fondamentale $\mathtt{J}_\mathbf{I}^\top\mathtt{J}_\mathbf{I}$ de $\mathcal{I}$, dans l'ordre descendant. Alors, nous avons :*

$$\mu_1 - \mu_2 = \|\nabla I\|^2 . \tag{4.19}$$

Cette proposition est prouvée en annexe, page 257.

Encore une fois, comme dans le § 4.6.3, la forme locale de $\mathcal{I}$ en $\mathbf{Q}$ peut être décrite par l'excentricité de la conique, donnée ici par la forme quadratique $\mathbf{v}^\top\mathtt{J}_\mathbf{I}^\top\mathtt{J}_\mathbf{I}\mathbf{v} = \pm 1$. À présent, nous pouvons nous demander comment interpréter cette conique. Le développement de Taylor au premier ordre pour des changements infinitésimaux $(\mathrm{d}x, \mathrm{d}y)$ dans le voisinage de $\mathbf{Q} = \mathbf{I}(x, y)$ donne :

$$\mathbf{I}(x + \mathrm{d}x, y + \mathrm{d}y) - \mathbf{I}(x, y) \approx \mathtt{J}_\mathbf{I}[\mathrm{d}x, \mathrm{d}y]^\top \tag{4.20}$$

Pour toute direction unitaire $\mathbf{v} = [a, b]^\top$ dans le plan xy, la forme quadratique $\mathbf{v}^\top\mathtt{J}_\mathbf{I}^\top\mathtt{J}_\mathbf{I}\mathbf{v}$ correspond à la partie linéaire g de la croissance de la longueur de l'arc de $\mathbf{I}(x, y)$ en $\mathbf{I}(x + a, y + b)$. De plus :

$$g^2 = \|\mathrm{d}\mathbf{I}_\mathbf{Q}((\mathrm{d}x, \mathrm{d}y)\|^2 = \mathbf{v}^\top\mathtt{J}_\mathbf{I}^\top\mathtt{J}_\mathbf{I}\mathbf{v}. \tag{4.21}$$

La remarque suivante est assez importante et n'a pas été mentionnée dans les travaux de [Plötz 17] : la primitive définie par l'expression (4.8) est tout simplement la norme de Frobenius de la matrice jacobienne $\mathtt{J}_\mathbf{N}$ de la carte $\mathbf{N}(x, y)$, cf. expression (4.4), à une constante près γ. Cela signifie que cette expression décrit le comportement au second ordre de la surface $\mathcal{M}$ relative à la norme en un point

5. Si $\mathbf{A}^\top\mathbf{A}$ est de rang plein, alors le maximum, respectivement le minimum, de $\|\mathbf{A}\mathbf{x}\|^2 = \mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x}$ sous les contraintes $\|\mathbf{x}\|^2 = 1$ est donné par la plus grande, respectivement la plus petite, valeur propre de $\mathbf{A}^\top\mathbf{A}$. Ici , $\mathbf{A} = \mathtt{J}_\mathbf{I}^\top\mathtt{J}_\mathbf{I}$ est de taille 2×2 et généralement de rang plein.

dans son voisinage immédiat. Ainsi, en utilisant les résultats montrés en (4.3), (4.4) et (4.5), nous pouvons affirmer que la primitive extraite dans la carte de profondeur dépend uniquement de X, Y, Z et de leurs dérivées d'ordre 2. Ceci est cohérent, en terme d'ordre des dérivées de X, Y, Z, avec les primitives $\|\nabla I\| = \sqrt{I_x^2 + I_y^2}$ détectées dans les images d'intensité où $I_\star$, avec $\star \in \{1, 2\}$, est donnée dans (4.16).

Pour conclure ces deux dernières sections, nous avons présenté la notion de saillance curviligne, et comment les primitives associées à cette notion peuvent être extraites en 3D, puis, en 2D. Dans la section suivante, nous montrons comment, en 2D, ce calcul peut encore être amélioré afin de le rendre plus robuste à la présence d'une texture sur l'objet et à la présence d'un fond dans l'image.

4.8. Robustesse à la texture et au fond

4.8.1. Saillance curviligne multi-échelle, *Multi-Curvilinear Saliency*, MCS

Contrairement aux images de profondeur qui représentent des formes 3D sans texture, les images d'intensité contiennent une forme texturée. En conséquence, la saillance curviligne, CS, estimée dans les images d'intensité est affectée par la présence de ces textures. Notre idée est d'utiliser l'hypothèse suivante : une analyse multi-échelle doit permettre de discriminer les points d'intérêt qui appartiennent à la forme de ceux qui appartiennent à une texture. Pour calculer la pyramide d'images, nous utilisons une technique de lissage qui préserve les contours, appelée diffusion anisotrope, *anisotropic filter* [Paris 09] [6]. Pour les plus basses résolutions de la pyramide, les contours détectés sont considérés comme fiables mais avec une localisation peu précise. De plus, les détails de la scène ne sont pas pris en compte. Pour les plus hautes résolutions de la pyramide, les détails sont préservés, mais en conséquence, les détections présentent de nombreux faux positifs dans la texture. Les scores CS des détails et des textures sont élevés dans les niveaux de plus haute résolution mais ces scores diminuent dès que la résolution diminue. Afin de combiner les informations présentes à chaque échelle, nous utilisons la valeur de CS pour chaque pixel à travers les N_c échelles. Plus précisément, si la valeur CS est plus grande qu'un seuil T_{CS}, à toutes les échelles, alors la saillance curviligne maximale de ce pixel est conservé. Cette valeur maximale à travers les différentes échelles est ce que nous appelons la saillance curviligne multi-échelle, *Multi-Curvilinear Saliency*, notée MCS dans la suite. Le seuil utilisé est une fonction qui dépend du nombre d'images de la pyramide n. En effet, ce seuil est défini par : $T_{CS} = e^{-n}$ afin qu'il ait le comportement suivant : quand n est petit, alors T est élevé et *vice versa*. De plus, ce qu'il faut remarquer et qui est important c'est que si la valeur CS est inférieure à ce seuil T_{CS}, seulement sur un niveau de la pyramide, alors nous considérons que ce point appartient à une texture ou à un détail non significatif de la forme et ainsi, il est supprimé de l'image finale des scores de saillance curviligne multi-échelle. Ainsi, nous espérons que l'ajout de cette analyse multi-échelle va permettre de réduire significativement l'impact de la texture sur la détection des points d'intérêt. Enfin, dans la prochaine section, nous allons présenter la dernière amélioration proposée pour augmenter la robustesse de l'approche en cas de présence de texture sur l'objet et/ou d'un fond texturé dans l'image d'intensité de l'objet. Pour cela, nous présentons un bref état de l'art sur les travaux existants en matière de détection de la texture dans une image et en particulier les approches faisant intervenir l'estimation de courbes de *focus*.

4.8.2. Extraction de la texture : méthodes existantes

De nombreuses approches ont été introduites pour permettre d'extraire les textures présentes dans une image naturelle [Aujol 06, Karacan 13, Zhang 14, Jeon 14]. Dans ces approches, l'image est vue comme la composition de deux éléments : la texture et le reste. L'idée est donc de supprimer la texture

6. Cette approche tente de séparer les composantes basse-fréquence (les frontières dues aux contours d'objets) des composantes basses fréquences (les textures) en préservant les contours les plus importants dans l'image.

tout en conservant les frontières importantes de l'image. La technique favorisée est celle du lissage de l'image dans un premier temps, d'où la difficulté à ne pas détruire les frontières des objets, pour, dans un deuxième temps, extraire la structure en s'appuyant sur des connaissances *a priori*. Ces méthodes de décomposition en structure et texture sont analogues au filtrage passe bas - passe haut en traitement classique du signal et de l'image. Cependant, même si cela peut paraître justifié de considérer que la structure contient des frontières bien contrastées, ceci n'est pas toujours le cas suivant les conditions d'éclairage et la texture elle-même peut contenir des frontières suffisamment bien contrastées pour « résister » à ce filtrage.

Une autre manière de procéder consiste à utiliser la notion de *focus* ou *focusness*. Plus précisément, une carte de *focus* est définie comme étant inversement proportionnelle au « degré » de flou de mise au point, *blurriness*, en chaque point de l'image [Jiang 13]. C'est un outil fréquemment utilisé pour estimer la profondeur d'une scène [Zhuo 11] mais également pour corriger la présence de flou ou tout simplement améliorer la qualité de l'image. La présence de flou dans l'image peut être due à de nombreuses raisons : ce qui explique la difficulté de la tâche et l'existence de nombreuses publications dans ce domaine [Tai 09]. Le lecteur intéressé peut trouver de nombreux détails techniques sur des approches prenant en compte les effets dûs au flou généré par la pénombre ou l'ombre [Zhuo 11], en particulier, en multi-résolution [Jiang 13]. Le flou d'une image est habituellement mesuré dans les zones qui contiennent des contours car c'est au niveau des contours qu'il apparaît ou du moins qu'il est mesurable [Elder 98]. La plupart des algorithmes existants [Jiang 13, Zhuo 11] évaluent ce flou en prenant en compte le rapport entre les frontières détectées à deux niveaux de lissage différents, c'est-à-dire que nous comparons l'image originale avec une version lissée d'elle-même. Ainsi, nous proposons simplement d'appliquer le même principe sur nos images, en calculant le rapport entre deux niveaux d'images de saillance curviligne afin d'extraire le taux de flou en s'appuyant précisant sur la technique développée dans [Zhuo 11]. En revanche, pour les aspects multi-résolution, nous nous inspirons des principes expliqués dans [Jiang 13].

4.8.3. Suppression du fond avec des courbes de *focus* : approches existantes

Les approches permettant de retrouver la forme 3D à partir des cartes de *focus* sont appelées les approches de reconstruction de forme à partir du *defocus* ou *Shape From Defocus*, SFD [Pentland 87]. Le *defocus* est ce que nous appelons en français les points en dehors du plan de *focus*, c'est-à-dire du plan de mise au point. Nous proposons un traitement à la manière de ces approches pour réaliser des cartes représentant les « courbes de *focus* », *focus curves*. Plus précisément, nous voulons que le niveau de flou soit estimé à partir des cartes de saillance curviligne des images 2D.

Un flou lié à la mise au point apparaît lorsqu'un point est en dehors du plan focal ou champ de mise au point, comme nous le présentons dans la figure 4.5. Quand le point est à une distance focale d_f, tous les rayons passant par lui convergent vers un seul point sur le capteur. En revanche, quand $d \neq d_f$, les rayons forment une région floue sur le capteur. Plus précisément, cette région est appelée le cercle de confusion ou *circle of confusion*, CoC. Nous notons c le diamètre de ce cercle.

Dans [Jiang 13, Zhuo 11], le flou dû au *defocus* a été modélisé par une fonction de l'étalement du point ou *Point Spread Function*, PSF. Cette PSF est habituellement approchée par une fonction gaussienne $g(x, \sigma)$ où la déviation standard est relative au degré de flou, c'est-à-dire qu'elle est définie comme étant proportionnelle au diamètre du cercle de confusion, $\sigma \propto c$. Plus précisément, nous avons :

$$c = \frac{|d - d_f|}{d} \frac{f}{d - f}, \tag{4.22}$$

où d, d_f, f correspondent respectivement à la distance, la distance focale et la longueur de la focale.

Dans [Zhuo 11], une méthode pour estimer le flou a été introduite pour le cas 1D, cf. figure 4.6. Plus précisément, les auteurs utilisent la notion de frontière floutée. L'intensité d'un point sur une frontière

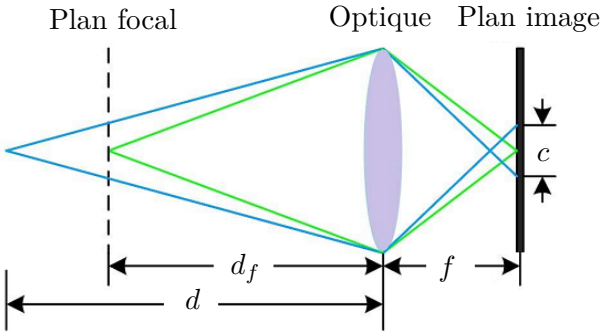


FIGURE 4.5. – Modèle optique pris en compte pour définir le flou de mise au point [Zhuo 11].

lissée, pour simuler le flou, $i(x)$ est définie par :

$$i(x) = f(x) \otimes g(x, \sigma), \tag{4.23}$$

où $f(x) = Au(x) + B$ est la frontière idéale, c'est-à-dire telle qu'elle doit être représentée sans flou et $u(x)$ est une fonction en escalier. Les termes A et B correspondent respectivement à l'amplitude et au biais par rapport à la localisation $x = 0$. Le gradient de la frontière lissée, c'est-à-dire floutée, est donné par :

$$\begin{aligned} \nabla i_1(x) =& \nabla(i(x) \otimes g(x, \sigma_0)) \\ =& \nabla((Au(x) + B) \otimes g(x, \sigma) \otimes g(x, \sigma_0)) \\ =& \frac{A}{\sqrt{2\pi(\sigma + \sigma_0)}} e^{\left(-\frac{x^2}{2(\sigma^2+\sigma_0^2)}\right)}, \end{aligned} \tag{4.24}$$

où σ_0 est la déviation standard associée au noyau gaussien de lissage. Ainsi, l'amplitude du gradient du rapport entre la frontière originale et celle floutée artificiellement est donnée par :

$$\begin{aligned} R =& \frac{|\nabla i_1(x)|}{|\nabla i(x)|} \\ =& \sqrt{\frac{\sigma^2 + \sigma_0^2}{\sigma^2}} e^{-\left(\frac{x^2}{2(\sigma^2)} - \frac{x^2}{2(\sigma^2+\sigma_0^2)}\right)}. \end{aligned} \tag{4.25}$$

Le rapport est maximal quand $x = 0$ et ce rapport maximal, noté R, est donné par :

$$R = \sqrt{\frac{\sigma^2 + \sigma_0^2}{\sigma^2}}. \tag{4.26}$$

Ainsi, la proportion de flou, notre inconnue, notée σ, peut être estimée par :

$$\sigma = \frac{\sigma_0}{\sqrt{R^2 - 1}}. \tag{4.27}$$

4.8.4. Courbes de *focus* multi-échelle ou *Multi-Focus Curves*, MFC

Une manière assez intuitive d'utiliser les cartes de saillance curviligne est de remplacer la détection des frontières par la détection des saillances curvilignes. Nous estimons ainsi non plus des cartes de saillance curviligne mais des cartes de courbes de *focus* de la saillance curviligne. De plus, nous

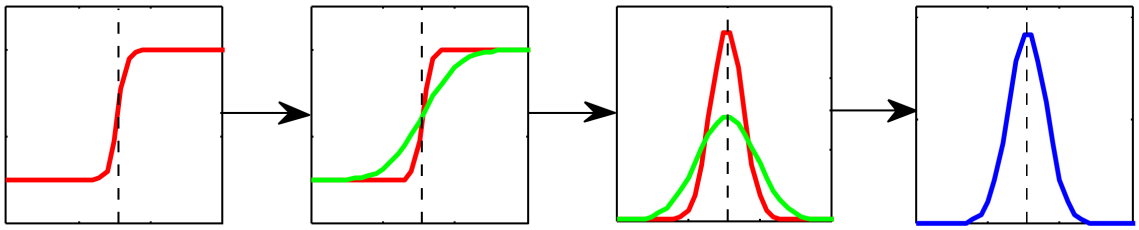


FIGURE 4.6. – Modèle de flou dans [Zhuo 11] – La ligne noire en pointillé représente la position théorique d'une frontière. Les termes $\bigotimes$ et ∇ correspondent respectivement à l'opérateur de convolution et au calcul du gradient. De plus, σ_1 est l'écart type associé à la fonction de lissage gaussienne.

estimons ces courbes de *focus* sur plusieurs échelles, contrairement au travail mono-échelle réalisé dans la publication de [Zhuo 11]. Pour cela, toutes les informations obtenues à chacune des échelles testées sont combinées pour fournir le résultat final. Le but de cet étape est de s'affranchir le plus possible de la texture de l'objet et du fond, afin d'obtenir un résultat de détection de saillance le plus proche possible de celui obtenu pour la carte de profondeur.

En suivant le même principe que celui exposé dans le § 4.8.3, la saillance curviligne est donnée par :

$$CS = \alpha(I_x^2 + I_y^2) \otimes g(x, y, \sigma). \tag{4.28}$$

L'image de saillance curviligne artificiellement floutée, à l'échelle i, notée CS_i peut être définie par :

$$CS_i = \alpha(I_x^2 + I_y^2) \otimes g(x, y, \sigma) \otimes g(x, y, \sigma_i), \tag{4.29}$$

avec $i = 1, 2, \ldots, n$ et n le nombre d'échelles considérées. Finalement, le rapport entre l'image originale de saillance curviligne et celle floutée artificiellement est donné par :

$$\begin{aligned} R_i &= \frac{CS_i}{CS} \\ &= \frac{\sigma^2 + \sigma_i^2}{\sigma^2} e^{-\left(\frac{x^2+y^2}{2(\sigma^2)} - \frac{x^2+y^2}{2(\sigma^2+\sigma_i^2)}\right)}. \end{aligned} \tag{4.30}$$

Dans le voisinage d'un pixel, la réponse atteint son maximum quand $x = 0$ et $y = 0$, ainsi, nous avons :

$$\begin{aligned} R_i &= \frac{CS_i}{CS} \\ &= \frac{\sigma^2 + \sigma_i^2}{\sigma^2} \\ &= 1 + \frac{\sigma_i^2}{\sigma^2}. \end{aligned} \tag{4.31}$$

Pour terminer, le degré de flou que nous cherchons à estimer σ_i est donné par :

$$\sigma_i = \frac{\sigma_i}{\sqrt{R_i - 1}}. \tag{4.32}$$

Pour n échelles, nous obtenons donc $n - 1$ échelles pour les courbes de focus en utilisant le rapport entre la saillance curviligne de plus bas niveau, c'est-à-dire le niveau de l'image et chacun des $n - 1$ niveaux de la pyramide restants. En suivant les mêmes principes que ceux énoncés au § 4.8.1, nous définissons donc la notion de courbes de *focus* multi-échelle ou *Multi-Focus Curves* , MFC, qui

correspond à la fusion de toutes les courbes de *focus* en une seule carte, en conservant uniquement les pixels qui ont un focus élevé sur les $n-1$ niveaux. Nous utilisons un seuillage similaire en prenant $T_{MFC} = e^{-n}$, comme pour le § 4.8.1. Ainsi, si un pixel a une valeur élevée à chaque échelle, nous conservons la valeur maximale obtenue parmi toutes les échelles, pour construire la carte de courbes multi-focus. En conclusion, les plus grandes valeurs de MFC sont révélatrices des frontières qui sont peu floutées et donc des frontières d'importance pour décrire la forme. À l'inverse, les faibles valeurs indiquent les frontières les plus floutées. Nous espérons donc que les frontières mises en évidence dans les images de saillance curviligne soient similaires à celles mises en évidence dans les cartes de profondeur.

4.9. Résultats de détection de primitives

4.9.1. Comparaison avec les approches existantes en détection de primitives

Nous rappelons que notre objectif le plus important dans ce travail est d'introduire un opérateur de détection qui soit le plus répétable possible entre des images d'intensité (2D) texturées et des modèles 3D sans texture. Et, en particulier, nous avons introduit un détecteur que nous souhaitons plus répétable que les détecteurs classiques de la littérature. En conséquence, nous comparons les primitives détectées en 2D avec :

(1) Le gradient image ou *Image Gradient*, noté IG ;

(2) La première approche proposée : la saillance curviligne multi-échelle ou *Multi-Scale Curvilinear Saliency*, MCS ;

(3) La seconde approche introduite : les courbes de *focus* multi-échelle, *Multi-Scale Focus Curves*, MFC.

De plus, nous mesurons la répétabilité avec deux autres détecteurs adaptés aux modèles 3D :

(1) La gradient d'ombrage moyen ou *Average Shading Gradient*, ASG [Plötz 17] ;

(2) La norme de Frobenius de la hessienne ou *Hessian Frobenius Norm*, HFN ;

(3) Le détecteur s'appuyant sur la saillance curviligne, CS.

Les opérateurs MFC et MCS ont également été comparés à 9 détecteurs 2D classiques de la littérature [Rashwan 19a]. Ici, nous avons choisi de présenter la comparaison avec ceux que nous avons mis en avant au chapitre 2 : Harris [Harris 88] (premier ordre), SIFT [Lowe 04], SURF [Bay 08] et PCBR [Deng 07] (second ordre et multi-échelle).

4.9.2. Critères d'évaluation des primitives détectées

Pour évaluer et comparer ces détecteurs 2D, nous avons choisi d'utiliser ces deux mesures :

(1) Le **Pourcentage d'intersection ou *Intersection percentage*, IP** : Il s'agit de la probabilité qu'un point d'intérêt en 2D soit trouvé proche d'un point d'intérêt détecté en 3D, dans la carte de profondeur [Rashwan 16].

(2) La **Distance de Haussdorf ou *Hausdorff distance* (HD)** : C'est une mesure classique pour comparer deux ensembles de points A et B :

$$HD(A,B) = \max\left(h(A,B), h(B,A)\right),$$

où $h(A,B) = \max_{a\in A} \min_{b\in B} \parallel a-b \parallel$. Plus cette distance est faible, plus les ensembles comparés sont similaires.

4.9.3. Bases de données testées pour la détection de primitives

Il est difficile de trouver une base de données qui corresponde parfaitement au problème que nous nous posons : évaluer la répétabilité entre primitives extraites en 2D et en 3D. C'est pourquoi, nous avons choisi d'évaluer deux ensembles de données : un que nous avons crée à partir d'images et de modèles collectés séparément sur internet et un autre issu de la littérature, soit :

(1) *Web collection* : nous avons collecté des données pour 10 objets différents. Plus précisément, nous avons récupéré 15 images réelles de chacun de ces objets en choisissant des vues les plus proches possibles des vues que nous générons pour les cartes de profondeur. De plus, pour tenter d'évaluer la robustesse aux différentes conditions d'acquisition, pour un même objet, nous avons choisi des images très différentes en terme de couleur, texture et modèle.

(2) *PASCAL3D+* [Y. Xiang 14] : Il nous permet de montrer la robustesse à un passage à grande échelle. Il contient 12 catégories d'objets rigides. Nous avons sélectionné les images sans occulations, soit approximativement 1000 images réelles et 3 modèles 3D au minimum pour chaque catégorie. Les images réelles sont acquises dans des conditions variables : éclairage, complexité du fond, niveau de contraste. Nous avons choisi des modèles 3D sans texture [7].

Toutes les cartes de profondeur ont été générées en utilisant `Matlab 3D Model Renderer` [8].

4.9.4. Analyses des résultats de détection de primitives

Méthodes	MFC	MCS	PCBR [Deng 07]	Harris [Harris 88]	SIFT [Lowe 04]	SURF [Bay 08]
Car	59	50	46	04	03	03
Shoe	38	31	31	03	10	01
Plane	58	55	38	04	10	03
T-Rex	66	64	59	06	02	05
Elephant	37	32	32	03	05	03
Fhydrant	56	51	42	04	02	09
Jeep	69	62	58	05	05	06
Mug	57	54	50	03	04	03
Teddy	44	39	32	05	09	04
Pistol	69	67	61	09	09	04

TABLE 4.1. – Pourcentage moyen d'intersection des détecteurs de points d'intérêt pour la base de test *web collection* – Plus cette valeur est élevée, plus le détecteur est répétable. Il s'agit de la moyenne calculée entre toutes les images réelles avec différentes textures et différents éclairages et toutes les images de profondeur associées. Ce calcul est réalisé pour tous les points de vue envisagés.

Comme nous pouvons le constater dans les tableaux de résultats 4.1 et 4.2, l'approche proposée MFC est capable de fournir le nombre le plus élevé de primitives répétables, malgré la présence de textures variables et de changements d'illumination. Plus précisément, MFC obtient un pourcentage d'intersection moyen, IP moyen, plus grand que 56%, alors que MCS et PCBR, obtiennent un taux respectivement plus grand que 50%, et plus grand que 44%, pour *web collection*. Pour *PASCAL3D+*, MFC obtient toujours le meilleur IP moyen, soit 46%. De plus, comme nous pouvons le voir dans les tableaux de résultats A.1 et 4.3, présentés en annexe A.4, les deux approches proposées donnent toujours les deux plus faibles distances de Haussdorf moyenne. Plus précisément, MFC est inférieure à 35 et, MCS est inférieure à 52. Pour compléter, nous présentons les résultats obtenus sur deux

7. Les données sont accessibles en ligne ici : http://tf3dm.com/.
8. Les données sont accessibles en ligne ici : http://www.openu.ac.il/home/hassner/projects/poses/

exemples avec la base de données la plus importante : PASCAL3D+ [Y. Xiang 14], cf. figure 4.8. Les pourcentages de répétabilité obtenus illustrent le fait que le détecteur IG est affecté par la texture alors que le couple MFC, pour les cartes de profondeur, et CS, pour les images réelles, permet d'obtenir la meilleure répétabilité, parmi toutes les combinaisons testées.

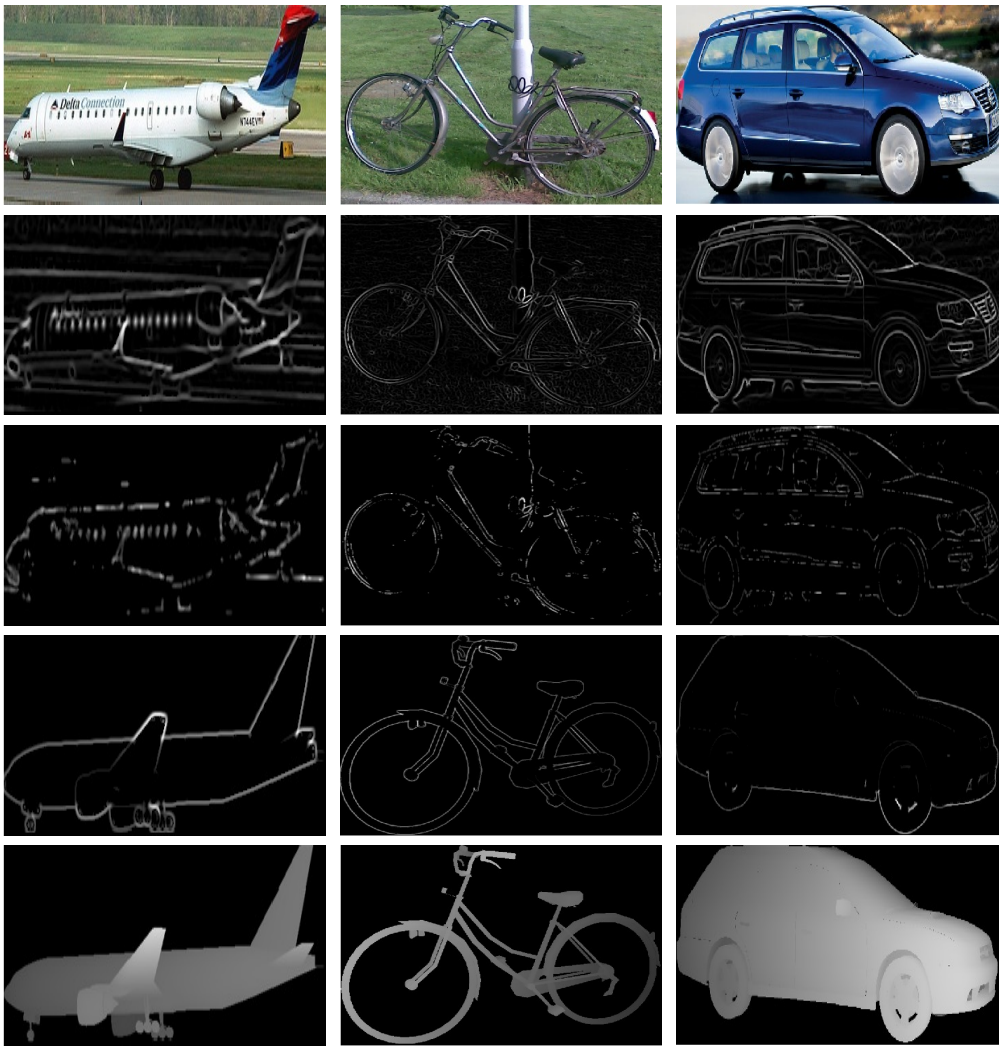


FIGURE 4.7. – Cartes de détection obtenues avec CS ainsi que MCS et MFC sur 5 échelles – La première ligne présente l'image 2D alors que la dernière correspond à la carte de profondeur avec le même point de vue. Ensuite, nous présentons dans l'ordre les résultats obtenus en combinant 5 échelles, avec MCS (ligne 2) puis MFC (ligne 3). Enfin, nous présentons l'utilisation de CS sur la carte de profondeur (ligne 4).

Robustesse du détecteur proposé aux changements d'illumination et temps d'exécution – Dans la publication [Rashwan 19a], nous avons également montré que l'approche proposée est robuste aux changements d'illumination. De plus, nous avons étudié l'impact de ces détecteurs en terme de temps d'exécution et les résultats obtenus montrent que les temps de calcul ne sont pas affectés de manière significative. En effet, les opérateurs proposés sont moins coûteux que les approches SURF ou SIFT. De plus, MFC et MCS ont des temps de calcul deux fois moins importants que PCBR.

Méthodes	MFC	MCS	PCBR [Deng 07]	Harris [Harris 88]	SIFT [Lowe 04]	SURF [Bay 08]
Plane	55	50	37	09	08	13
Bicycle	69	61	57	08	16	24
Boat	42	36	28	10	06	10
Bus	31	24	17	06	02	04
Car	44	41	24	08	03	06
Diningtable	40	38	19	05	04	08
Motorbike	59	53	48	09	06	14
Sofa	67	66	60	11	18	20
Train	31	28	14	07	03	05
Tvmonitor	44	40	37	06	12	11
Chair	58	54	42	10	09	18
Bottle	56	54	51	14	18	21

TABLE 4.2. – Pourcentage moyen d'intersection des détecteurs de points d'intérêt pour la base de test *PASCAL3D+* – Plus cette valeur est élevée, plus le détecteur est répétable. Il s'agit de la moyenne calculée entre toutes les images réelles avec différentes textures et différents éclairages et toutes les images de profondeur associées. Ce calcul est réalisé pour tous les points de vue envisagés.

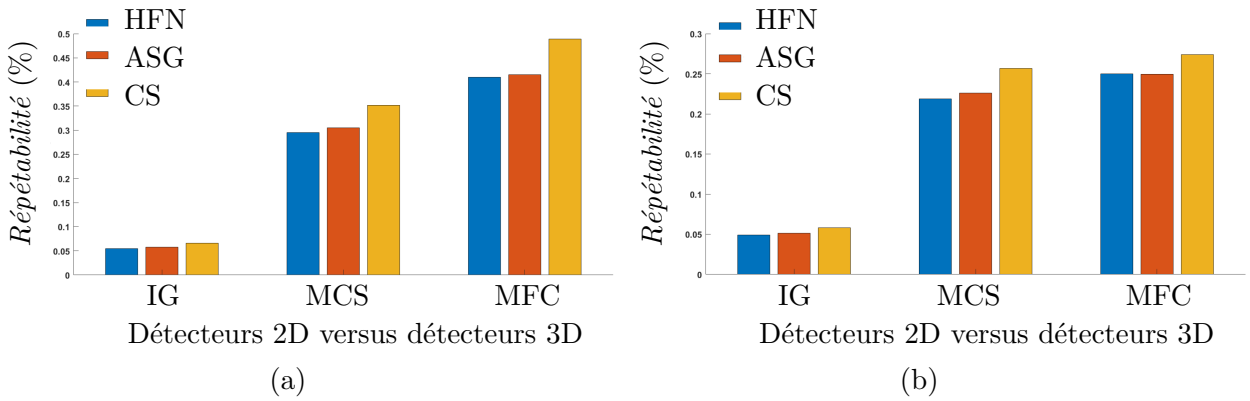


FIGURE 4.8. – Pourcentage de répétabilité pour 2 exemples de modèle 3D – le modèle *car* en (a) et le modèle *sofa* en (b).

4.10. Recalage entre images 2D et modèles 3D

L'objectif de notre étude est de proposer une approche de recalage entre données 2D et données 3D à partir de correspondances de points d'intérêt. Nous présentons donc maintenant la manière d'effectuer ce recalage et les résultats obtenus. Ainsi, à présent, nous allons aborder la définition d'un descripteur adapté. Nous souhaitons trouver la vue d la plus proche du point de vue d'acquisition d'une image 2D. Cette vue est sélectionnée à partir d'un ensemble d'images, plus précisément de cartes de profondeur, obtenues à partir du modèle 3D. Nous notons d_k, $k = 1 \dots N$, l'ensemble des vues avec N le nombre total de vues, c'est-à-dire le nombre de cartes de profondeur. Nous supposons que l'objet à reconnaître est contenu dans une boîte englobante et que nous souhaitons obtenir sa pose 3D. Estimer la pose consiste à estimer les angles d'élévation et d'azimuth, notés respectivement h et a ainsi que la distance entre l'objet et la caméra, notée v. Pour chaque modèle 3D, les cartes de profondeur ont été générées de manière uniforme autour de la sphère en faisant évoluer h, a et v de manière à obtenir N vues par modèle. Nous expliquons en détails les choix réalisés dans le § 4.11.

De manière à décrire l'objet dans toutes les images réelles et dans toutes les cartes de profondeur,

nous proposons une adaptation du descripteur HoG [Dalal 05a], décrit dans le § A.2 et déjà utilisé dans la publication la plus proche de notre travail [Plötz 17]. Plus précisément, nous introduisons la notion d'histogramme de saillance curviligne ou *Histogram of Curvilinear Saliency*, HCS.

Tout d'abord, pour exploiter les différents opérateurs proposés pour la détection, nous effectuons une normalisation. Ainsi, dans le cas des cartes de profondeur, CS est multiplié par le vecteur propre e_{H_1} qui correspond à la plus grande valeur propre de la matrice M dans l'expression (4.15) :

$$\overrightarrow{CS} = CS.\overrightarrow{e_{H_1}}. \tag{4.33}$$

De la même façon, pour les images réelles, les valeurs MCS sont multipliées par le vecteur propre e_{S_1} qui correspond à la saillance curviligne $\lambda_1 - \lambda_2$:

$$\overrightarrow{MCS} = MCS.\overrightarrow{e_{H_1}}. \tag{4.34}$$

Enfin, les valeurs de MFC sont également multipliées par e_{S_1} :

$$\overrightarrow{MFC} = MFC.\overrightarrow{e_{S_1}}. \tag{4.35}$$

Pour le détecteur HoG, une fenêtre glissante permet de générer un descripteur qui correspond à un histogramme quantifié des orientations des gradients dans une région d'intérêt. De plus, l'orientation et l'amplitude du gradient au point étudié sont utilisées pour construire le descripteur. Suivant le même principe, et en nous inspirant du travail préliminaire présenté dans [Fischer 14], nous proposons un descripteur qui contient l'orientation de la courbure ainsi que l'amplitude de CS, MCS et MFC, combinée aux histogrammes quantifiés de ces valeurs.

Ainsi, nous notons le descripteur HCS dans une image 2D $\mathbf{D}_I$, et les descripteurs HCS des différentes cartes de profondeur $\mathbf{D}_{d_k}$. Pour comparer D_q à chaque D_{d_k}, nous utilisons le même score de similarité que celui utilisé dans [Aubry 14] :

$$\mathbf{S}_{hcs}(k,h,a,v) = (\mathbf{D}_{d_k} - \boldsymbol{\mu}_s)^T \Sigma^{-1} \mathbf{D}_I, \tag{4.36}$$

avec $k = 1...N$ et où Σ et $\boldsymbol{\mu}_s$ sont, respectivement, la matrice de covariance et la moyenne de tous les descripteurs sur toutes les cartes de profondeur.

De plus, pour valider la correspondance avec une vue, une similarité globale est évaluée en mesurant combien chaque point de l'image individuellement peut être apparié avec un point détecté dans la carte de profondeur, c'est-à-dire, en évaluant combien chaque point de l'image vue est répétable. Plus précisément, ce score de répétabilité normalisé entre 0 et 1 est la probabilité que les points d'intérêt dans l'image d'intensité soient trouvés à proximité d'un point d'intérêt extrait dans la carte de profondeur. Il est noté $Rep_{d_k \dashrightarrow I}$. Comme la vue la plus proche doit obtenir la répétabilité la plus élevée, par comparaison avec toutes les autres vues, la dissimilarité s'appuyant sur ces scores de répétabilité est donnée par $R_{d_k} = 1 - Rep_{d_k \dashrightarrow I}$. Ainsi, la similarité globale est définie par :

$$\mathbf{S}_{rep}(k,h,a,v) = e^{\left(\frac{-(\mathbf{R}_{d_k} - \mu_r)^2}{2\,\sigma_r^2}\right)}, \tag{4.37}$$

avec μ_r la valeur moyenne de $\mathbf{R}_{d_k}$ et σ_r la déviation standard [9]. Finalement, en combinant toutes les similarités des descripteurs HCS et la similarité s'appuyant sur la répétabilité, la probabilité finale est donnée par :

$$\mathbf{S}(k,h,a,v) = \mathbf{S}_{hcs}(k,h,a,v) \odot \mathbf{S}_{rep}(k,h,a,v). \tag{4.38}$$

9. Nous avons choisi $\sigma_r = 0.1$.

où $\odot$ est le produit d'Hadamard. Pour estimer la pose, afin de tenter de corriger certaines erreurs, nous avons choisi de sélectionner, dans un premier temps, les 3 meilleures vues suivant le score $\mathbf{S}(k, h, a, v)$. Ensuite, parmi ces 3 vues, nous conservons uniquement celles qui sont cohérentes entre elles, c'est-à-dire celles qui ne sont pas trop éloignées d'une position moyenne définie par les deux autres. Enfin, la pose est affinée en générant de nouvelles vues dans les intervalles définis par les maxima et minima de h, a et v parmi les vues conservées et en répétant le processus de sélection des meilleures vues telles que nous venons de l'introduire. Les itérations s'arrêtent lorsque la différence entre les valeurs minimales et maximales sont inférieures à un seuil [10].

4.11. Résultats obtenus pour l'estimation de la pose

Alignement et représentation des modèles 3D – Un aspect délicat sur la génération des différentes vues concerne le choix du nombre de vues générées. Plus ce nombre est important et plus l'estimation de la pose pourra être précise mais plus les temps de calcul seront importants. En conséquence, nous avons choisi une approche incrémentale, déjà introduite au § 4.10. Nous générons environ $N = 700$ vues au départ car cela permet d'avoir un temps de calcul raisonnable. Ce qui donne h qui évolue par pas de 50^o, a par pas de 20^o et v par pas de 30 cm. Ce choix donne une répartition à peu près uniforme.

Pour comparer les cartes de profondeur et les images 2D, nous avons également ajouté un prétraitement pour placer les données dans des échelles similaires et les placer dans le même repère ayant pour origine le centre de l'objet ou du modèle, cf. les détails fournis dans [Rashwan 19a].

Enfin, de la même manière que pour HoG, nous avons utilisé une quantification en 9 classes [Dalal 05a] pour le descripteur HCS. De plus, chaque image et chaque carte de profondeur est divisée en une grille régulière de cellules de taille 8×8 [11]. Le descripteur final est donc une concaténation des histogrammes obtenus sur chaque cellule et pondérés par l'amplitude.

Analyse préliminaire – Pour reconnaître un objet, et surtout identifier sans ambiguïté sa pose, nous souhaitons que la probabilité que des points d'intérêt dans une image 2D se trouvent proches de ceux de la carte de profondeur associée au même point de vue soit élevée. En revanche, nous voulons que cette probabilité soit faible quand il s'agit de la carte de profondeur associée à un autre point de vue. C'est bien le type de comportement que nous observons dans la figure 4.9. Nous considérons trois type de détecteurs : un s'appuyant sur la détection des contours par Sobel, SIFT et MCS. Nous obtenons un score plus élevé pour la bonne vue dans les 3 cas mais pour Sobel, cette différence est assez faible. Ce qui est remarquable dans le résultat que nous montrons est que c'est bien MCS qui obtient la différence la plus significative entre la bonne vue et les autres, ce qui nous permet d'identifier de la manière la plus fiable possible la pose. SIFT a été proposé pour être robuste à de nombreuses transformations donc ce résultat, où la variation est moins importante, est cohérent.

Analyses des résultats – Les expérimentations suivantes ont été réalisées sur la base *PASCAL3D+*. Nous avons calculé le taux de précision de la pose. Pour les modèles 3D, nous comparons les 3 approches déjà présentées : CS, ASG et les lignes de crêtes apparentes, *apparent ridges*, AR [Judd 07], et pour les images d'intensité, nous utilisons les 3 techniques déjà introduites : IG [Plötz 17], MCS et MFC, les deux approches proposées. Comme nous pouvons le constater dans la table 4.3, effectuer le recalage en utilisant CS combinée à MFC, permet d'obtenir les meilleures résultats par comparaison avec tout autre combinaison d'opérateurs. De plus, le taux de précision est réduit de 25% par rapport aux approches ASG ou IG. Il faut noter que AR obtient les plus mauvais résultats, quelque soit le

10. Nous avons choisi $|dh| = 5, |da| = 5$ et $|dv| = 1$.

11. Ce choix empirique a été réalisé à partir de tests avec des grilles de 4×4, 8×8 et 16×16. La grille 8×8 est celle qui a permis d'obtenir les meilleurs résultats.

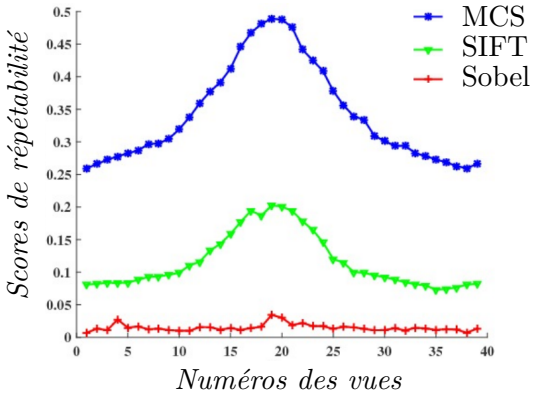


FIGURE 4.9. – Exemple de scores de répétabilité obtenus entre une image et 40 cartes de profondeur – Parmi les 40 vues, celles à la position du milieu (coordonnée 20 sur l'axe des abscisses). Cet exemple nous permet de comparer les comportements des détecteurs de Sobel, SIFT et celui proposé. Comme attendu, SIFT est assez robuste aux changements de points de vue alors que MCS y est sensible.

type de représentation utilisée pour les images d'intensité. Comme attendu, l'utilisation de ASG est adaptée dans les cas de modèles non texturés et pour reconnaître des objets non texturés. Toutes ces analyses et conclusions sont confirmées par les résultats présentés dans la table 4.4 où les détails sur la qualité de l'estimation de la pose sont fournis.

Méthodes	3D	CS			ASG			AR		
	2D	MFC	MCS	IG	MFC	MCS	IG	MFC	MCS	IG
Plane		0.85	0.83	0.62	0.84	0.80	0.59	0.78	0.70	0.50
Bicycle		0.81	0.76	0.60	0.80	0.78	0.61	0.74	0.73	0.49
Boat		0.78	0.71	0.58	0.75	0.70	0.57	0.71	0.68	0.52
Bus		0.87	0.82	0.56	0.82	0.80	0.52	0.75	0.74	0.51
Car		0.86	0.85	0.58	0.86	0.83	0.51	0.76	0.72	0.47
Diningtable		0.86	0.83	0.61	0.81	0.81	0.60	0.79	0.77	0.54
Motorbike		0.79	0.78	0.60	0.78	0.75	0.58	0.69	0.62	0.52
Sofa		0.85	0.81	0.64	0.80	0.72	0.61	0.68	0.61	0.53
Train		0.87	0.86	0.70	0.81	0.82	0.71	0.74	0.67	0.58
Tvmonitor		0.83	0.80	0.55	0.80	0.79	0.54	0.66	0.64	0.52
Chair		0.80	0.76	0.62	0.78	0.73	0.56	0.68	0.66	0.54
Bottle		0.77	0.74	0.61	0.79	0.71	0.59	0.62	0.64	0.50

TABLE 4.3. – Précision de la pose estimée avec CS, ASG [Plötz 17], AR [Judd 07], IG et nos propositions MFC et MCS.

Pour approfondir la présentation de ces résultats, nous avons choisi de montrer le pourcentage de cas où la vue similaire se trouve parmi les premières vues sélectionnées. Plus précisément, nous avons d'abord trié toutes les vues suivant les scores de similarité et ensuite nous avons analysé les r meilleurs scores, avec $r = 1, 3, 5, 10, 20$. Nous cherchons enfin si la pose similaire est présente parmi ces r premières réponses. Les résultats obtenus sont présentés dans la figure 4.10. Dans ces expérimentations, la précision représente le pourcentage de cas où la vue similaire se trouve dans les r

réponses considérées. Bien évidemment, plus r augmente et plus le taux de précision est élevé. Mais le résultat important à retenir dans ces graphiques est que le taux de précision est toujours le plus élevé en combinant CS avec n'importe quelle représentation 3D, et parmi les 3 représentations 3D, MFC est celle qui obtient les meilleurs résultats. Finalement, nous présentons quelques résultats visuels dans la figure 4.11 afin d'illustrer la variété d'exemples différents (textures variées, changements de points de vues et d'éclairage) où la méthode retenue obtient des résultats intéressants.

Méthodes	(a)		(b)		(c)		(d)	
	Est.	Clo.	Est.	Clo.	Est.	Clo.	Est.	Clo.
CS/MFC	16.5^o	4.8^o	08.8^o	1.2^o	5.6^o	0.8^o	18	7
CS/MCS	16.0^o	5.2^o	11.4^o	1.5^o	6.1^o	1.1^o	21	8
ASG/MFC	19.2^o	5.3^o	10.1^o	1.3^o	5.1^o	0.8^o	22	9
ASG/MCS	19.6^o	5.9^o	13.6^o	1.9^o	6.2^o	1.2^o	23	11
AR/MFC	28.7^o	7.1^o	16.5^o	2.5^o	8.5^o	1.8^o	36	13
AR/MCS	29.5^o	8.0^o	17.3^o	3.1^o	9.2^o	2.0^o	39	17

TABLE 4.4. – Erreur moyenne sur l'estimation de la pose (Est.) – Nous présentons les erreurs sur la pose de la caméra en termes d'élévation, en (a), d'azimuth, en (b), sur les angles, en (c) et enfin en distance (en centimètres). Le terme Clo. correspond à la vue la plus proche de la pose similaire. Ces résultats montrent que la meilleure combinaison est MFC/CS.

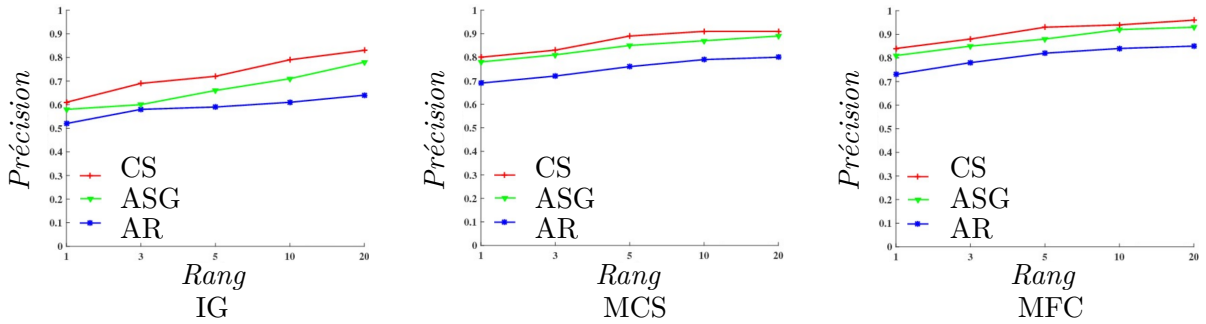


FIGURE 4.10. – Taux de précision du recalage 2D/3D en fonction du rang de similarité – La précision représente le pourcentage de cas où la vue similaire se trouve dans les r premiers rangs des réponses ordonnées suivant le degré de similarité.

Temps d'éxécution pour effectuer le recalage – Bien évidemment, les temps d'exécution peuvent être améliorés car nous n'avons pas optimisé l'implantation, mais ce qui nous semble important est de noter que 53% du temps d'exécution dépend du rendu et de l'extraction des points d'intérêt dans la carte de profondeur, alors que l'extraction des points d'intérêt dans l'image représente seulement 6%. Finalement, le recalage prend 40% du temps.

Comparaison avec une approche utilisant un réseau de neurones convolutifs – La méthode que nous avons proposée a été comparée avec deux approches d'estimation de la pose faisant intervenir des réseaux de neurones convolutifs, cf. § 12.2.1. Les résultats exposés dans [Rashwan 19a] sur le jeu de données, *PASCAL3D+* montrent que la qualité des résultats que nous obtenons est comparable à la qualité des résultats obtenus avec [Su 15, Tulsiani 15]. Ces approches ont dû effectuer le rendu de millions d'images synthétiques pour entraîner leurs modèles et obtenir ces résultats, ce que nous n'avons pas eu besoin de faire pour notre approche.



FIGURE 4.11. – Exemples de recalage 2D/3D corrects avec CS/MFC sur la base *PASCAL3D+* – Nous présentons dans la première colonne l'image requête, puis, dans la seconde colonne, la vue correspondante dans le modèle 3D et enfin la superposition de l'image requête (en vert) avec la vue estimée (en violet). Même si les modèles n'ont pas exactement les mêmes formes que l'image requête, le recalage est réalisé de manière satisfaisante.

4.12. Conclusions et perspectives au recalage 2D/3D

Ainsi, nous avons proposé une analyse de l'existant en terme de recalage 2D/3D et nous avons introduit une approche en tentant de prendre en compte au mieux la variabilité entre une image d'intensité et un modèle 3D non texturé. En particulier, nous avons mis en évidence les raisons pour lesquelles l'approche proposée prend en compte les spécificités de cette tâche. Plus précisément, nous avons introduit un protocole d'évaluation pour évaluer la répétabilité des détections entre une image d'intensité et une carte de profondeur. La méthode introduite a permis d'étudier de manière distincte deux sujets : comment représenter les données en 2D et en 3D pour être ensuite en capacité de les apparier. Ainsi, nous avons introduit un détecteur 3D s'appuyant sur la saillance curviligne et un détecteur 2D utilisant le même principe mais en ajoutant une analyse multi-échelle et en utilisant la notion de *focus*. L'intérêt de l'approche proposée a été validé par une évaluation quantitative de la pose estimée à partir du recalage 2D/3D.

Nous pouvons envisager des améliorations à ce travail, notamment en prenant en compte des approches récentes en lien avec ce travail. Par exemple, en ce qui concerne les représentations communes entre données en 2D et en 3D, nous pouvons citer l'approche de [Grabner 19b] qui définit des cartes de localisation ou *location fields* en 2D et en 3D. Pour les modèles 3D, ces cartes sont calculées très simplement en effectuant le rendu. En revanche, en 2D, les auteurs ont besoin d'effectuer une estimation en effectuant un apprentissage préalable par un réseaux de neurones convolutifs. Nous pourrions également combiner le travail de détection et d'appariement avec une information de segmentation,

comme dans le travail de [Huang 15] ou en suivant les principes que nous avons étudiés et que nous présentons dans le chapitre 9 ou dans nos perspectives, cf. § 13.7.

Enfin, la suite envisagée pour ce travail est de reconnaître les défauts des objets. Pour cela, dans un premier temps, il est impératif d'effectuer une étude de la robustesse de cette approche aux occultations et aux changements de l'objet, afin de déterminer quelles sont les adaptations à réaliser pour intégrer cette qualification des défauts.

Conclusion sur la détection de primitives en 2D et en 3D et leur appariement

Nous avons présenté un état de l'art le plus exhaustif possible sur la détection de primitives, et plus précisément de points d'intérêt, en 2D et en 3D afin de comprendre, d'une part, les grands principes et, d'autre part, les liens entre détecteurs en 2D et en 3D. Cela correspond à ce que nous avons présenté aux chapitres 2 et 3, mais également à toutes les informations ajoutées en annexe A et qui constituent en grande partie les cours réalisés à l'ENSEEIHT. L'objectif de cette présentation était également de montrer nos études sur la complémentarité des détecteurs en 2D, cf. § 2.3, ainsi que l'intérêt de combiner des données multimodales en 3D en imagerie médicale, cf. § 3.3. Enfin, le chapitre 4 a mis en avant un outil commun pour analyser des images en 2D et des modèles 3D, en utilisant des cartes de profondeur : la courbure. Nous avons également proposé un détecteur de points d'intérêt en 2D faisant intervenir les notions de saillance curviligne et de courbes de *focus*, avec une analyse multi-échelle, pour prendre en compte, le mieux possible, les différences entre images 2D et modèles 3D afin de s'en affranchir. Les évaluations réalisées montrent l'intérêt de l'approche proposée que nous avons confirmé, grâce à l'utilisation applicative dans [Abdulwahab 19] mais également dans les travaux de segmentation que nous présentons au chapitre 11.

Nous pourrions approfondir ce travail, notamment en étudiant encore plus les aspects liés à la détection du flou de *focus*. En effet, des travaux récents sur l'analyse de scènes d'intérieur permettent d'estimer la focale avec un apprentissage profond [Grabner 19a]. Nous pourrions étudier la possibilité d'exploiter ce type d'approche ainsi que l'utilisation de cette estimation de la focale dans notre calcul du flou de *focus*. Dans la partie IV relative aux perspectives, nous développons également des travaux incluant ce travail, aux § 13.1 et 13.2. Il s'agit de perspectives dans la continuité de thèses co-encadrées. C'est également le cas au § 13.6 qui correspond à un projet d'encadrement de thèse. Plus particulièrement, nous aimerions prendre en compte les aspects temporels mais également l'association avec une segmentation sémantique. En effet, cela nous semble pertinent dans le contexte de recherche actuel, cf. les travaux réalisés, dans un contexte de segmentation urbaine, par [Armagan 17].

Bien que cette partie se concentre sur l'aspect détection et caractérisation, nous avons commencé à aborder les notions d'appariement et nous allons approfondir cet aspect dans toute la partie suivante. Encore une fois, notre objectif est d'étudier la complémentarité des approches existantes afin de proposer des méthodes permettant de combiner les intérêts de chaque approche. Nous nous concentrons en grande partie sur la mise en correspondance stéréoscopique mais nous ouvrons également cette thématique à l'appariement de vidéos multiples.

Deuxième partie

Appariements multiples

Introduction sur la mise en correspondance multiple

De nombreux travaux ont été proposés pour détecter des primitives spécifiques, faiblement corrélées à leur environnement, afin de les apparier plus facilement, c'est-à-dire sans ambiguïté. Du fait de cette définition, notamment du respect du critère de répétabilité, la phase d'appariement se limite souvent à une simple corrélation croisée, voire une simple différence, comme nous l'avons vu dans la partie précédente. Par exemple, dans le cas de SIFT, cf. § A.1.6, une simple utilisation de la corrélation centrée normalisée suffit. La première partie de ce manuscrit a porté sur l'étude de ces primitives et de leur combinaison afin de réaliser un appariement multimodal. Cette nouvelle partie se concentre sur la manière de réaliser l'appariement.

Dans un premier temps, nous allons étudier la mise en correspondance dense. Un appariement dense ou *a minima* moins épars qu'un appariement de points d'intérêt, permet une reconstruction 3D plus fine. En effet, dans le cas d'une scène d'intérieur simple comportant de nombreux objets plans, à partir d'une reconstruction partielle de la scène, nous pouvons facilement interpoler les surfaces estimées pour calculer une reconstruction complète sans commettre d'erreurs importantes. Lorsque la scène est plus complexe, avec des objets naturels, aux surfaces plus chaotiques, effectuer une mise en correspondance dense permettra de reconstruire des détails plus fins. En contrepartie, cette étape d'appariement est plus délicate, entraînant des erreurs de correspondance qu'il faut être en capacité de traiter. C'est pourquoi, comme pour la première partie, nous avons porté une attention particulière aux évaluations précises des erreurs d'appariement des approches proposées.

Ainsi, une grande partie de notre travail se place dans le contexte de l'utilisation de couples stéréoscopiques, ou de séquences d'images où des couples stéréoscopiques peuvent être extraits et nous nous préoccupons essentiellement de mise en correspondance de pixels. Nous étudions en particulier la manière de rendre cette étape plus robuste à certaines difficultés. Plus précisément, nous montrons comment différentes informations peuvent être fusionnées, comme la couleur, de multiples mesures, des points d'intérêt appariés de manière fiable, des approches de segmentation voire de sursegmentation en superpixels. Dans ce contexte, nous appelons mise en correspondance conjointe un algorithme d'appariement qui va combiner les intérêts de multiples approches d'appariement ou de multiples informations extraites. À la fin de cette partie, nous introduisons un nouveau contexte de mise en correspondance intégrant des aspects temporels en manipulant des vidéos multiples.

Dans la littérature, l'habitude est de distinguer, parmi les nombreuses méthodes de mise en correspondance, les approches locales des approches globales [Brown 03]. Une autre façon de procéder est de décomposer la description en éléments constituants [Chambon 05a, chapitre 1]. Cette description permet d'affiner la séparation en quatre catégories : les méthodes locales, les méthodes globales ne faisant pas intervenir de critère local, les méthodes mixtes (avec critère local qui intervient dans un processus global) et les méthodes à passages multiples (qui font intervenir de manière séquentielle et complémentaire différentes approches locales et/ou globales). Nous nous intéressons essentiellement à des approches locales ou à passages multiples. Les difficultés de la mise en correspondance sont nombreuses : bruits, manque de texture, occultations, discontinuités de profondeur. Nous abordons en particulier la gestion des zones faiblement texturées, la prise en compte des occultations et des problèmes liés aux discontinuités de profondeur.

Contributions scientifiques structurantes en mise en correspondance

Le début de cette partie s'inscrit dans la continuité des perspectives énoncées à la fin de la thèse [Chambon 05a, page 185]. Le premier chapitre 5 introduit la bibliographie relative à l'Unité d'Enseignement Vision, Réalité Augmentée et Applications, proposée en troisième année aux étudiants de l'ENSEEIHT, dans le parcours Multimédia [Chambon 15b]. Les contributions correspondant aux perspectives de la thèse qui ont été concrétisées sont présentées dans le chapitre 6. Plus précisément, nous présentons :

(1) Une étude approfondie de l'impact de la couleur – Dans le § 6.1, nous caractérisons de manière plus détaillée l'intérêt de la couleur dans des approches d'appariement [Bleyer 08a, Bleyer 10a] notamment en déterminant les zones de distorsions radiométriques et en étudiant l'impact de ces zones sur les performances des méthodes globales. Ce travail a été réalisé en collaboration avec Michael Bleyer, à l'université de Vienne, Autriche.

(2) Une étude précise de l'apport de chaque mesure de corrélation par rapport aux autres et de l'intérêt de les combiner – Dans le § 6.2, nous présentons l'étude de l'intérêt de combiner plus de deux mesures de corrélation pour les approches locales. Notamment nous avons caractérisé le gain de performances que nous pouvons espérer de cette combinaison et nous avons introduit un nouvel algorithme de combinaison. Ce travail a été publié dans [Chambon 10b, Chambon 11b].

(3) Une prise en compte des difficultés liées au manque de texture en étudiant et en proposant des approches par croissance de germes combinées aux mesures robustes proposées – Nous étudions l'impact de l'utilisation de points d'intérêt, dans le § 6.3. À partir de correspondances fiables, nous souhaitons propager les résultats dans leur voisinage jusqu'à obtenir une mise en correspondance dense [Gales 10c, Gales 12, Prévost 13a, Prévost 13b].

(4) Une étude de la gestion des occultations en combinant les mesures proposées à une segmentation ou une sur-segmentation – Nous avons introduit une nouvelle approche s'appuyant sur la notion de vote dans le § 6.4. Ces travaux ont été publiés dans [Gales 10a].

Ces deux derniers points ont été réalisés au cours de la thèse de Guillaume Galès [Gales 11], co-encadrement avec Alain Crouzil, IRIT-TCI, sous la direction de Patrice Dalle.
La fin de cette partie traite du cas plus général de la mise en correspondance de vidéos en exploitant les éléments redondants, les activités concomitantes entre vidéos. Ces travaux ont été publiés dans [Malon 18, Guyot 19, Malon 19a, Malon 19b] et sont réalisés dans le cadre de la thèse de Thierry Malon (thèse en cours), co-encadrement avec Alain Crouzil, IRIT-MINDS, sous la direction de Vincent Charvillat, IRIT-REVA.

Chapitre 5.

Mise en correspondance

5.1. Introduction sur la mise en correspondance de pixels

Nous pouvons mettre en correspondance des primitives plus compliquées que des pixels [Jawahar 02], comme des lignes [Bouchafa 06] ou des points d'intérêt, cf. partie I, mais dans cette partie nous nous concentrons sur les pixels. Une manière très répandue de classer les méthodes de mise en correspondance est de distinguer les méthodes locales des méthodes globales, c'est ce qui est proposé dans [Scharstein 02b, Brown 03]. Dans ce paragraphe, nous reprenons les éléments présentés dans [Chambon 05a, chapitre 1], et en particulier, les familles de méthodes considérées : les approches locales, les approches globales ne faisant pas intervenir de mesure de corrélation, les approches globales faisant intervenir une mesure de corrélation et enfin, les approches mixtes qui combinent différentes approches d'appariement ou différents outils de vision, comme les approches faisant intervenir une segmentation en régions.

Ainsi, nous nous plaçons dans le cas où la mise en correspondance est effectuée de l'image gauche vers l'image droite mais toutes les explications et les formules peuvent être utilisées dans le cas d'une mise en correspondance de la droite vers la gauche. Une manière de représenter le résultat d'une mise en correspondance consiste à associer à chaque pixel $\mathbf{p}_g^{i,j} = (i\ j)^T$ de l'image gauche un vecteur appelé disparité, défini par ses composantes $(u-i\ v-j)^T$ où $(u\ v)^T$ sont les coordonnées dans l'image droite du point correspondant au pixel $\mathbf{p}_g^{i,j}$.

Nous pouvons assimiler le problème de la mise en correspondance à la recherche d'une fonction de disparité d qui attribue une disparité à chaque pixel $\mathbf{p}_g^{i,j}$. Dans le cas général, nous avons :

$$\begin{aligned} d\ :\ \mathbb{N}^2 &\longrightarrow\ \mathbb{R}^2 \\ \mathbf{p}_g^{i,j} &\longmapsto\ d(\mathbf{p}_g^{i,j}) = (u-i\ \ v-j)^T. \end{aligned} \tag{5.1}$$

Lorsqu'une disparité réelle est calculée, il s'agit d'un appariement au sous-pixel. La plupart des méthodes effectuent une mise en correspondance de pixel à pixel, c'est-à-dire qu'elles calculent des disparités entières, et dans ce cas, nous avons :

$$d : \mathbb{N}^2 \rightarrow \mathbb{Z}^2.$$

Nous commençons donc par donner des rappels sur le calcul des coûts utilisés pour une approche globale, cf. § 5.2, puis pour une approche locale, cf. § 5.3. Nous présentons ensuite une catégorie de méthodes mixtes faisant intervenir une segmentation en régions, cf. § 5.4, avant de présenter les protocoles d'évaluation que nous utilisons, cf. § 5.5, et qui nous permettent de valider tous les résultats présentés au chapitre 6.

5.2. Coût global de mise en correspondance

Le coût global de mise en correspondance qui évalue à quel point un ensemble de correspondances est valide a pour forme générale :

$$E_{\text{global}}(d) = (1-\lambda)E_{\text{correspondance}}(d) + \lambda E_{\text{contrainte}}(d), \tag{5.2}$$

où $\lambda \in [0;1]$ permet d'ajuster l'influence des poids entre les deux termes :

- $E_{\text{correspondance}}$ qui est le coût de correspondance ou terme d'attache aux données.
- $E_{\text{contrainte}}$ qui est le coût des contraintes modélisant les interactions entre les pixels considérés.

Résoudre le problème de la mise en correspondance, formulé de cette manière, consiste à trouver la fonction de disparité d qui minimise ce coût E_{global}.

5.2.1. Coût de correspondance

Ce terme évalue à quel point un ensemble de correspondances est fiable. Il est déterminé par la somme sur le support $\mathcal{S}$ des coûts locaux et il est de la forme :

$$\begin{aligned} E_{\text{correspondance}}(d) &= \sum_{\mathbf{p}_g^{i,j} \in \mathcal{S}} E_{\text{local}}(\mathbf{p}_g^{i,j}, \mathbf{p}_d^{u,v}) \\ \text{avec } \mathbf{p}_d^{u,v} &= \mathbf{p}_g^{i,j} + d(\mathbf{p}_g^{i,j}). \end{aligned} \tag{5.3}$$

Le coût local correspond au coût d'une correspondance et il est défini par :

$$E_{\text{local}}(\mathbf{p}_g^{i,j},\ \mathbf{p}_d^{u,v}) = \sum_{\substack{\mathbf{p}_g^{i',j'} \in \text{ZA}(\mathbf{p}_g^{i,j}) \\ \mathbf{p}_d^{u',v'} \in \text{ZA}(\mathbf{p}_d^{u,v})}} E_{\text{dissimilarité}}(\mathbf{p}_g^{i',j'}, \mathbf{p}_d^{u',v'}), \tag{5.4}$$

où $\text{ZA}(\mathbf{p}_l^{i,j})$ correspond à la zone d'agrégation considérée (pixels pris en compte pour le calcul du coût local). La zone d'agrégation d'un pixel $\mathbf{p}_l^{i,j}$ est le plus souvent un voisinage de $\mathbf{p}_l^{i,j}$, c'est-à-dire qu'elle correspond à un ensemble de pixels connexes qui contient $\mathbf{p}_l^{i,j}$. Un exemple classique de voisinage est une fenêtre carrée de taille $(2t+1) \times (2t+1)$ centrée sur $\mathbf{p}_l^{i,j}$. Le terme $E_{\text{dissimilarité}}$ est le coût de dissimilarité : il évalue à quel point deux pixels ne se ressemblent pas.

5.2.2. Coût des contraintes

Il modélise les contraintes entre les pixels du support $\mathcal{S}$ et les pixels de la zone d'agrégation. Il y a N_c contraintes et ainsi N_c zones d'agrégation associées que nous notons ZAC car cette zone est différente de la zone d'agrégation associée au coût local, ZA, et elle correspond la plupart du temps à un voisinage. Le coût des contraintes est donné par :

$$E_{contrainte}(d) = \sum_{c=0}^{N_c-1} E^c_{contrainte}(d) = \sum_{c=0}^{N_c-1} \sum_{\mathbf{p}_g^{i,j} \in \mathcal{S}} E^c_{\text{voisinage}}(\mathbf{p}_g^{i,j}), \tag{5.5}$$

Le coût du voisinage permet d'évaluer l'effet de la contrainte sur le pixel étudié et ses voisins et il est défini par :

$$E^c_{\text{voisinage}}(\mathbf{p}_g^{i,j}) = \sum_{\mathbf{p}_g^{i',j'} \in \text{ZAC}^c(\mathbf{p}_g^{i,j})} E^c_{\text{lissage}}(\mathbf{p}_g^{i,j}, \mathbf{p}_g^{i',j'}). \tag{5.6}$$

Le terme E^c_{lissage} permet de comparer les disparités associées aux deux pixels considérés et il est appelé coût de lissage. La zone d'agrégation associée à la c$^{\text{ème}}$ contrainte est notée ZAC^c.

5.2.3. Coût global développé

Nous avons présenté les deux termes du coût global de mise en correspondance, nous pouvons à présent reprendre l'équation (5.2) et la détailler :

$$E_{\text{global}}(d) = \sum_{\mathbf{p}_g^{i,j} \in \mathcal{S}} \left((1-\lambda) \underbrace{\sum_{\substack{\mathbf{p}_g^{i',j'} \in \text{ZA}(\mathbf{p}_g^{i,j}) \\ \mathbf{p}_d^{u',v'} \in \text{ZA}(\mathbf{p}_d^{u,v})}} E_{\text{dissimilarité}}(\mathbf{p}_g^{i',j'}, \mathbf{p}_d^{u',v'})}_{\text{coût local}} + \lambda \sum_{c=0}^{N_c-1} \underbrace{\sum_{\mathbf{p}_g^{i',j'} \in \text{ZAC}^c(\mathbf{p}_g^{i,j})} E^c_{\text{lissage}}(\mathbf{p}_g^{i,j}, \mathbf{p}_g^{i',j'})}_{\text{coût du voisinage}} \right) \quad (5.7)$$

$$\text{avec } \mathbf{p}_d^{u,v} = \mathbf{p}_g^{i,j} + d(\mathbf{p}_g^{i,j}).$$

Les choix à faire pour ce coût global de mise en correspondance sont :

- λ – Le plus souvent, les publications n'abordent pas cette pondération.
- $E_{\text{dissimilarité}}$ – Ce coût s'appuie souvent sur les niveaux de gris des pixels. Cet aspect a été discuté dans la thèse [Chambon 05a] et nous donnerons seulement les conclusions sur les mesures de similarité les plus remarquables que nous avons mises en avant dans [Chambon 11c].
- E^c_{lissage} – Ce coût, dans la majorité des cas, est lié à la disparité associée à chaque pixel ou au fait que le pixel est considéré comme occulté ou non.
- ZA – Nous distinguons les méthodes qui utilisent une zone d'agrégation réduite à un seul pixel (le pixel étudié), ce sont les méthodes globales, des méthodes qui utilisent une zone d'agrégation plus importante, ce sont les méthodes locales et les méthodes mixtes.
- ZAC^c – Les zones d'agrégation des contraintes ne sont, en général, pas identiques aux zones d'agrégation du coût local. Dans de nombreux cas, la zone d'agrégation pour le coût local est limitée au pixel étudié alors que la zone d'agrégation des contraintes est un 4-voisinage.
- $\mathcal{S}$ – Le support peut être le pixel considéré ou l'ensemble des N_g pixels considérés dans l'image ou les N_g^{col} pixels qui se trouvent sur la même ligne (dans le cas des images rectifiées, nous reviendrons sur cette définition au § 5.5.2).

Les méthodes locales effectuent N_g minimisations des coûts locaux et la zone d'agrégation est toujours strictement supérieure à un pixel (au pixel étudié), alors que les méthodes globales effectuent une minimisation du coût global qui est une fonction des coûts locaux et la zone d'agrégation est souvent réduite au pixel considéré.

5.3. Coût local de mise en correspondance

Les méthodes locales sont aussi appelées méthodes par corrélation. Une corrélation indique un degré de ressemblance entre deux ensembles de données. Par abus de langage, dans ce chapitre, quand nous parlons de méthode par corrélation, la mesure de corrélation utilisée évalue le degré de dissimilarité entre deux ensembles de pixels (score de corrélation). Pour ces méthodes, nous avons :

- S – Il y a N_g supports, chacun étant réduit à un singleton.
- ZA – Il y a deux possibilités : utiliser une fenêtre carrée centrée sur le pixel considéré ou encore des fenêtres adaptatives [Veksler 02, Yoon 06, Zhang 09]. Les approches les plus sophistiquées

s'appuient sur l'utilisation de poids attribués aux pixels de la zone considérée et permettent, entre autres, la gestion des déformations liées à la perspective [Hosni 13]. Nous pouvons également citer les approches prenant en compte un sous-ensemble de pixels non connexe [Muresan 15], en s'appuyant sur les mêmes principes que les *Local Binary Pattern*, LBP, cf. § A.2.

- $E_{\text{dissimilarité}}$ – Une des variantes les plus utilisées est :

$$E_{\text{dissimilarité}}(\mathbf{p}_g^{i',j'}, \mathbf{p}_d^{u',v'}) = f_{\text{dissimilarité}}(I_g^{i',j'} - I_d^{u',v'}) = (I_g^{i',j'} - I_d^{u',v'})^2, \tag{5.8}$$

où $f_{\text{dissimilarité}}$ est une fonction de la différence des deux niveaux de gris.

Les méthodes locales n'ont pas de coût des contraintes et seul le coût local, équation (5.4), est utilisé. Ainsi la fonction de disparité d est déterminée de la façon suivante :

$$\forall\ \mathbf{p}_g^{i,j} \in \mathcal{P}_g \quad d(\mathbf{p}_g^{i,j}) = (u - i\ v - j)^T \text{ avec } \mathbf{p}_d^{u,v} = \underset{\mathbf{p}_d^{u',v'} \in \mathrm{Z}_d(\mathbf{p}_g^{i,j})}{\operatorname{argmin}} E_{\text{local}}(\mathbf{p}_g^{i,j},\ \mathbf{p}_d^{u',v'}), \tag{5.9}$$

où $\mathcal{P}_g$ est l'ensemble des pixels considérés dans l'image, $\mathrm{Z}_d(\mathbf{p}_g^{i,j})$ correspond à la zone de recherche associée au pixel $\mathbf{p}_g^{i,j}$. La méthode d'optimisation employée est la méthode de recherche exhaustive, *winner take all*, notée WTA, c'est-à-dire que le pixel qui obtient le meilleur coût est sélectionné.

En mise en correspondance stéréoscopique, les approches par corrélation sont très populaires, du fait de leur implémentation simple et de leur temps de calcul relativement faible. En particulier, la publication de [Yoon 05] a mis en évidence l'efficacité de ce type d'approches en obtenant un bon classement sur l'évaluation proposée par [Scharstein 02b]. Nous considérons qu'une mesure de corrélation évalue à quel point deux pixels et leurs voisinages se ressemblent. Même si dans la littérature de nombreuses mesures de corrélation ont été évaluées et comparées (mesure classique, mesures basées dérivées, mesures avec transformation non paramétrique) [Hirschmüller 09], le choix d'une mesure reste assez difficile et dans [Chambon 11c], nous avons proposé une évaluation plus systématique et plus complète qui a permis de caractériser plus finement le comportement de chaque mesure et de proposer des recommandations de choix parmi un ensemble de cinq familles de mesures. Ces cinq familles correspondent aux mesures s'appuyant sur : un produit scalaire en niveaux de gris, une distance classique entre niveaux de gris, une distance entre champs de vecteurs gradients, une distance non-paramétrique, une mesure issue des statistiques robustes. Les deux dernières catégories sont les plus performantes en présence de bruits et/ou d'occultations alors que les deux premières sont les plus performantes dans les zones sans difficulté. Plus de détails sont fournis dans [Chambon 11c] et également en annexe, § B.1 où nous présentons les six mesures que nous utilisons par la suite.

Maintenant que nous avons rappelé les éléments essentiels des approches les plus classiques, nous présentons les approches combinées à une segmentation en régions, voire une sursegmentation.

5.4. Appariement en utilisant une segmentation en régions

Le principe des méthodes utilisant des régions repose sur l'hypothèse qu'une région de couleur homogène dans l'image suffisamment petite peut être la projection d'une même surface de la scène et que cette surface peut être approchée par un modèle dans l'espace des disparités, par exemple un plan ou une surface *B-Spline*, [Hong 04, Sun 05b, Klaus 06, Wang 08, Yang 08a, Yang 09, Taguchi 08, Bleyer 10b]. L'utilisation de ces régions peut aider la mise en correspondance de régions homogènes et permet également d'éviter des artefacts au niveau des discontinuités de profondeur. Le modèle plan est relativement simple et donne de bonnes approximations, en particulier avec les plus petites régions, mais peut se révéler imprécis lorsque les régions ne sont pas les projections de facettes planes. Dans ce dernier cas, le modèle *B-Spline* peut donner de meilleures approximations. Cependant, il est plus complexe à configurer (l'ordre et le vecteur nodal doivent être choisis avec soin) et d'importantes oscillations peuvent apparaître dans le résultat [Lin 03, Bleyer 10b].

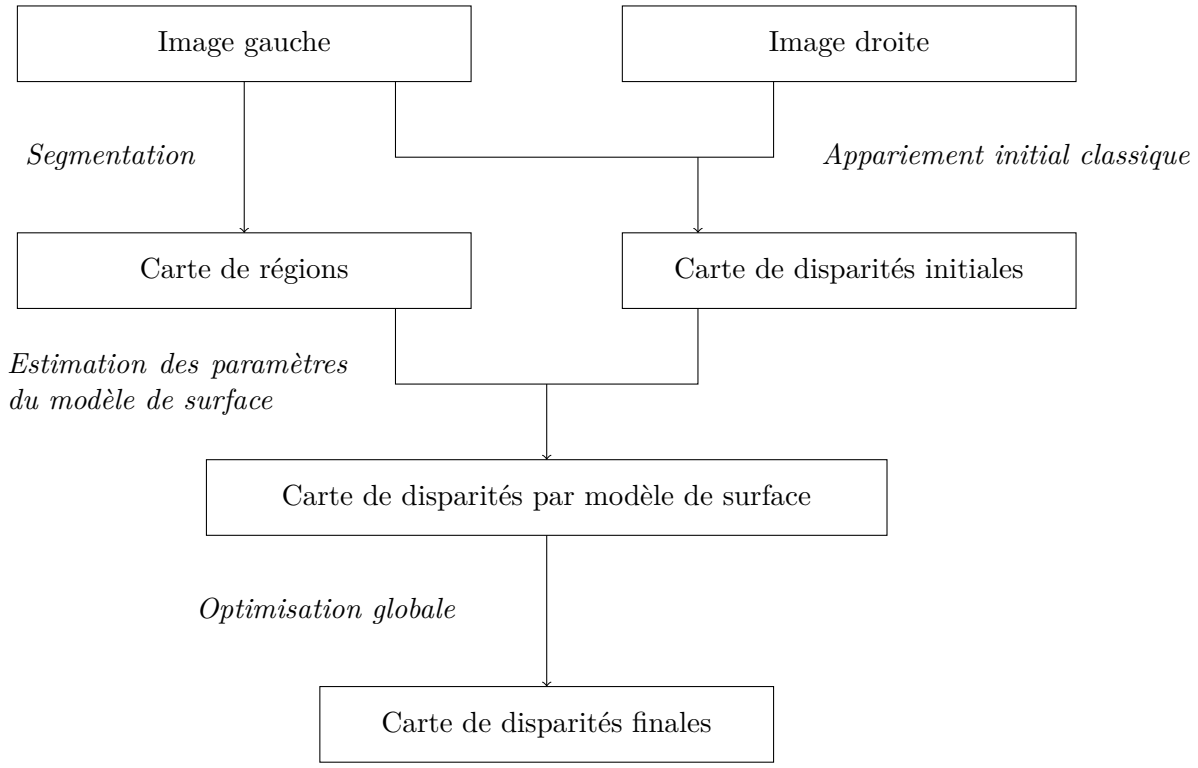


FIGURE 5.1. – Étapes des méthodes de mise en correspondance fondées sur les régions.

Ces méthodes comportent plusieurs étapes résumées par la figure 5.1 :

- **Segmentation des images en régions** – Une partition de l'image est calculée en utilisant un algorithme de segmentation en régions de couleur homogène [Gonzalez 04, Trémeau 04]. La méthode *mean-shift* est la plus utilisée dans ce contexte [Sun 02, Klaus 06, Wang 08, Yang 09, Bleyer 10b]. Cette approche est développée au § C.3.2. Il est important de noter que pour cette étape, une sur-segmentation est presque toujours calculée. Elle permet d'avoir des petites régions qui sont plus facilement approchées par un modèle plan.
- **Appariement initial classique** – Pour initialiser les disparités, une méthode de mise en correspondance locale est généralement employée comme celles décrites dans le § 5.3.
- **Estimation des paramètres du modèle choisi** – Le modèle plan est relativement simple et donne de bonnes approximations des disparités, en particulier pour les petites régions. Il s'agit du modèle le plus couramment utilisé mais le modèle *B-Spline* a également été étudié dans [Lin 03, Bleyer 10b]. Les paramètres du modèle choisi sont habituellement estimés à l'aide d'une méthode d'estimation robuste à partir des disparités initiales : moindres carrés pondérés itératifs [Hong 04], RANSAC, *random sample consensus*, [Yang 08a, Yang 09] ou décomposition des paramètres [Wang 08, Klaus 06].
- **Optimisation globale** – Les résultats sont raffinés en appliquant une optimisation globale. Nous pouvons distinguer les méthodes suivantes :
  - *Affectation d'un plan* – Les différents plans de disparité sont obtenus à partir des modèles estimés et chaque pixel est affecté à l'un de ces plans de telle sorte qu'un coût global sur l'image entière soit optimisé. Cela permet de corriger certaines erreurs notamment dues à une mauvaise segmentation. Cette solution est utilisée dans [Hong 04, Bleyer 10b] où la technique d'optimisation employée est la coupure de graphe, et également dans [Klaus 06] où la technique d'optimisation employée est la propagation de croyance.
  - *Raffinement des paramètres* – Les paramètres des modèles préalablement estimés peuvent être

affinés en minimisant un coût global. Cette solution est notamment utilisée dans [Wang 08, Yang 08a].

Toutes les contributions que nous présentons au chapitre 6 sont évaluées et comparées et pour cela, nous avons utilisé le même protocole, que nous décrivons dans la section suivante.

5.5. Évaluation et comparaison des approches de mise en correspondance

5.5.1. Images de tests utilisées et proposées pour l'évaluation de l'appariement

Dans la littérature, un travail conséquent et fondateur a permis de répertorier et comparer les méthodes les plus performantes, essentiellement parmi des approches globales [Scharstein 02b]. Les auteurs de cet article ont aussi mis un protocole d'évaluation à la disposition de la communauté [1]. Nous avons utilisé les bases de données des articles suivants : [Scharstein 02b, Scharstein 03] que nous avons complétées avec des couples de tests générés dans nos travaux [Chambon 05a, chapitre 2]. Nous obtenons donc une base de 42 paires d'images, avec une vérité terrain ou une carte de disparités associée, cf. figure 5.2 : un stéréogramme aléatoire, deux paires d'images synthétiques, une paire d'images composées d'objets plans et finalement 38 paires d'images réelles. Plus précisément, les images de Middlebury sont issues de différentes versions, de plus en plus complexes : 6 paires ont été introduites en 2001 [Scharstein 02b], puis 2 en 2003 [Scharstein 03] et 30 en 2005 et 2006 [Scharstein 07, Hirschmuller 07].

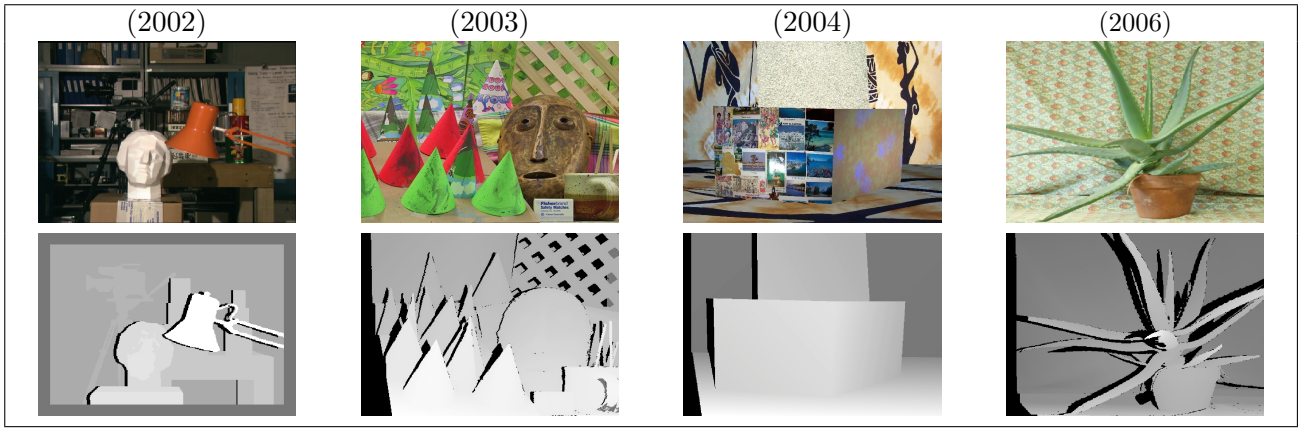


FIGURE 5.2. – Exemples de données testées pour la mise en correspondance stéréoscopique – Sur la première ligne, l'image gauche et sur la deuxième ligne, la carte de disparités associées. Tous les détails se trouvent dans les publications et le site web, cf. § 5.5 (utilisation de vision active et/ou de contraintes géométriques sur la scène).

5.5.2. Hypothèse sur les images stéréoscopiques utilisées

Jusqu'à présent, nous avons présenté les outils d'appariement dans le cas général, c'est-à-dire en supposant que les déplacements d'une image à l'autre peuvent être réalisés dans les deux dimensions.

En réalité, et c'est le cas de la base de données présentée au § 5.5.1, la plupart des couples stéréoscopiques sont rectifiés. La rectification épipolaire consiste à se ramener à une configuration particulière du capteur, appelée configuration parallèle, dans laquelle deux pixels qui se correspondent

1. http://vision.middlebury.edu/stereo/eval/

se trouvent sur la même ligne. Ainsi le déplacement entre un pixel dans l'image gauche et son correspondant dans l'image droite correspond à une simple translation. Il s'agit donc d'une transformation géométrique à appliquer aux deux images. La difficulté réside dans l'estimation des paramètres de cette transformation, ce qui n'est pas le coeur de ce travail et nous renvoyons à l'ouvrage de référence [Hartley 04]. **Dans le chapitre 6, tous les travaux présentés ont été testés sur des images en configuration parallèle, contrairement au chapitre 7.**

5.5.3. Critères d'évaluation utilisés pour mesurer la qualité d'une carte de disparités

Nous supposons que nous possédons une vérité terrain, c'est-à-dire une estimation de la disparité pour chaque pixel de l'image que nous jugeons suffisamment fiable pour servir de référence. Ainsi, si nous notons d_{ref} la fonction de disparité théorique et **occ** la disparité symbolique attribuée à un pixel occulté, il est possible de calculer l'erreur suivante pour évaluer la qualité d'une disparité estimée :

$$\text{Er}_g^{i,j} = \left\| d(\mathbf{p}_g^{i,j}) - d_{\text{ref}}(\mathbf{p}_g^{i,j}) \right\|. \tag{5.10}$$

Le calcul de $\text{Er}_g^{i,j}$ n'a de signification que si $d(\mathbf{p}_l^{i,j}) \neq \textbf{occ}$ et $d_{\text{ref}}(\mathbf{p}^{i,j}) \neq \textbf{occ}$.

Ainsi, nous pouvons définir les critères suivants :

1. *Pourcentage d'appariements corrects, noté C* – Un appariement est correct si $\text{Er}_g^{i,j} = 0$.
2. *Pourcentage d'appariements acceptés, noté A* – Un appariement est accepté si $0 < \text{Er}_g^{i,j} < t$.
3. *Pourcentage d'appariements erronés, noté* Er – L'appariement est erroné si $\text{Er}_g^{i,j} \geq t$.
4. *Densité d'une carte de disparités, notée D* – La densité du résultat correspond au rapport du nombre de pixels mis en correspondance sur le nombre total de pixels de l'image. Si nous notons R l'image des disparités colonnes associée à une image, alors, nous pouvons calculer

$$D(R) = \frac{\#\left\{\mathbf{p}_g^{i,j} | d(\mathbf{p}^{i,j}) \neq \textbf{occ}\right\}}{N_g}. \tag{5.11}$$

5. *Taux de rappel, noté T_{ra}* – Il s'agit du nombre d'appariements corrects sur le nombre d'appariements effectués.
6. *Taux d'acceptation, noté T_a* – Cela correspond à un taux de rappel avec une certaine tolérance aux erreurs. Cela correspond donc au nombre d'appariements acceptés sur le nombre d'appariements effectués.

5.6. Conclusion sur la mise en correspondance de pixels

Ce chapitre nous a permis de présenter les grandes familles d'approches en insistant sur celles qui font parties des contributions du chapitre suivant : les approches par corrélation et les approches par régions. De plus, nous avons présenté les jeux de données et les critères utilisés pour valider, évaluer et comparer. Le chapitre suivant sera donc consacré aux diverses contributions en appariement stéréoscopique de pixels.

Chapitre 6.

Contributions à la mise en correspondance stéréoscopique

6.1. Étude de l'utilisation de la couleur dans des approches globales

Intuitivement, nous pouvons espérer que l'utilisation de la couleur améliore les résultats obtenus à partir d'une image en niveaux de gris. En effet, la couleur peut permettre de lever des ambiguïtés en fournissant plus d'information que les niveaux de gris. Par exemple, si nous observons l'image de la figure 6.1, nous pouvons constater que beaucoup de formes réalisées en origami semblent être, dans l'image en niveaux de gris, dans des couleurs similaires alors qu'elles ne le sont absolument pas dans l'image en couleur.



FIGURE 6.1. – Exemple d'intérêt de la couleur pour la mise en correspondance – Nous pouvons espérer au vu de cet exemple que la couleur apporte une information plus riche que les niveaux de gris, en particulier quand nous observons la forme rose en origami sur le fond en tissu madras bleu. Cette forme semble être de la même couleur que le support sur lequel elle est. Or, si nous observons le même détail en couleur, nous constatons que le support est beige/jaune, puisque c'est un carton.

Nous avons étudié l'impact de la couleur pour des approches de mise en correspondance par corrélation [Chambon 05b] et nous avons, entre autres, montré que cette amélioration était significative uniquement dans les zones chromatiques, c'est-à-dire des zones où l'information de couleur est présente et significative par rapport aux niveaux de gris, cf. [Chambon 05a, chapitre 4]. L'ensemble des images testées comportait 14.47% à 86.34% de zones chromatiques et cela explique bien les différences modérées de comportement entre couleur et niveaux de gris. Par la suite, nous avons étudié l'utilisation de la couleur dans des **approches globales**. Malgré les premiers résultats encourageants dans la littérature [Okutomi 92, Koschan 93, Mühlmann 02] et dans les travaux que nous avons menés [Chambon 05b, Bleyer 08a], notamment, dans [Bleyer 08a] où nous avons obtenu plus de 25% d'amélioration globale dans les expérimentations menées, nous allons ici, d'une part, confirmer ces résultats préliminaires, mais, d'autre part, ajouter des détails sur la raison pour laquelle ces performances ont été obtenues et quelles sont leurs limites. En particulier, nous mettons en pa-

rallèle notre étude avec les améliorations apportées par les approches robustes aux distorsions radiométriques comme celle de [Hirschmüller 09]. Dans ce papier, les auteurs testent quasi exclusivement des opérateurs s'appuyant sur les niveaux de gris mais ils présentent également des résultats préliminaires en utilisant la couleur et ils constatent des améliorations assez mineures en utilisant la couleur avec leurs mesures robustes aux distorsions radiométriques et en concluent que la couleur n'apporte pas grand chose à la mise en correspondance. Cela contredit les résultats préliminaires obtenus. Ainsi, de manière similaire à nos travaux précédents, nous tentons de trouver les zones où les améliorations sont significatives et nous étudions, en particulier, l'impact de la couleur dans des zones de **distorsions radiométriques**. Nous appelons une distorsion radiométrique tout ensemble de pixels ayant une intensité ou une couleur différente d'une image à l'autre, ce qui ne respecte pas l'hypothèse de consistance photométrique généralement faite dans la littérature. Il y a de nombreuses causes possibles pour observer une distorsion radiométrique comme des problèmes de réglage du capteur (temps d'exposition non adapté impliquant une sur ou sous exposition ou simplement un écart), le vignettage [1], ou, plus simplement, des conditions d'éclairage légèrement différentes d'une image à l'autre. Dans notre étude, le résultat principal est le suivant : dans les cas où nous observons une différence photométrique entre pixels correspondants due à des variations de luminosité, la couleur apporte une amélioration du résultat par comparaison avec les niveaux de gris. Cela signifie que l'introduction de la couleur agit dans une méthode donnée permet de rendre cette approche insensible à des distorsions radiométriques. Cela nous permet de compléter l'analyse des résultats fournis dans [Hirschmüller 09] : effectivement, si les distorsions radiométriques ont déjà été prises en compte par un autre moyen, comme une mesure de similarité robuste aux distorsions radiométriques alors l'utilisation de la couleur est peu utile. Par la suite, nous présentons les coûts de mise en correspondance mis en place, cf. § 6.1.1, la méthode d'optimisation utilisée, cf. § 6.1.2, la manière d'estimer les zones de distortions radiométriques pour évaluer les résultats dans ces zones, cf. § 6.1.3 et enfin les résultats obtenus, cf. § 6.1.4.

6.1.1. Coûts globaux d'appariement faisant intervenir la couleur

Coût global de mise en correspondance utilisé – Nous reprenons donc l'équation (5.7), en supprimant le terme de pondération λ [2] :

$$E_{\text{global}}(d) = E_{\text{correspondance}}(d) + E_{\text{contrainte}}(d). \tag{6.1}$$

Coût des contraintes – Nous allons en premier lieu détailler les choix réalisés pour $E_{\text{contrainte}}$. De manière classiue, nous utilisons seulement la contrainte de lissage des disparités dans le voisinage et nous obtenons :

$$E_{\text{contrainte}}(d) = \sum_{\mathbf{p}_g^{i,j} \in \mathcal{S}} E_{\text{voisinage}}(\mathbf{p}_g^{i,j}) = \sum_{\mathbf{p}_g^{i,j} \in \mathcal{S}} \left(\sum_{\mathbf{p}_g^{i',j'} \in \text{ZAC}(\mathbf{p}_g^{i,j})} E_{\text{lissage}}(\mathbf{p}_g^{i,j}, \mathbf{p}_g^{i',j'}) \right). \tag{6.2}$$

Pour le voisinage, ZAC, nous choisissons simplement les 4 voisins. Habituellement, un seul voisin est considéré, mais, pour être robuste aux erreurs d'appariement, nous pensons qu'il est plus adapté de

1. Il s'agit du cas où nous observons un assombrissement des coins de l'image qui forme une sorte d'encadré, de vignette, autour de la scène acquise. Ce défaut peut être provoqué par l'objectif ou le filtre utilisé.

2. Expérimentalement, les résultats obtenus par Michael Bleyer dans le cadre de ses recherches nous ont permis de fixer ce paramètre.

considérer les 4 voisins. Ainsi, nous utilisons la fonction de lissage suivante :

$$E_{\text{lissage}}(\mathbf{p}_g^{i,j}, \mathbf{p}_g^{i',j'}) = \begin{cases} 0 & \text{Si } d(\mathbf{p}_g^{i,j}) = d(\mathbf{p}_g^{i',j'}) \\ P_1 & \text{Si } |d(\mathbf{p}_g^{i,j}) - d(\mathbf{p}_g^{i',j'})| = 1 \\ P_2 & \text{Sinon.} \end{cases} \tag{6.3}$$

Les constantes P_1 et P_2 sont à définir, sachant que P_1 représente la pénalité lorsqu'il y a une faible différence de disparité et P_2 la pénalité sur les discontinuités de profondeur [3]

Coût de correspondance – Enfin, nous définissons le terme d'attache aux données, $E_{\text{correspondance}}(d)$, en reprenant l'équation (5.3). Dans un premier temps, nous allons étudier des coûts locaux complets, comme exposé au § 5.2, en suivant l'équation (5.4), cf. § 6.1.1. Dans un second temps, nous ne considérons pas de zone d'agrégation et ainsi l'équation (5.4) devient :

$$\begin{aligned} E_{\text{correspondance}}(d) = & \sum_{\mathbf{p}_g^{i,j} \in \mathcal{S}} \rho(\mathbf{p}_g^{i,j}, \mathbf{p}_d^{u,v}) \\ & \text{avec } \mathbf{p}_d^{u,v} = \mathbf{p}_g^{i,j} + d(\mathbf{p}_g^{i,j}). \end{aligned} \tag{6.4}$$

Nous avons simplement renommé $E_{\text{local}}(\mathbf{p}, \mathbf{q})$ en $\rho(\mathbf{p}, \mathbf{q})$. Cette expression représente la fonction de dissimilarité entre deux pixels. Ainsi, elle évalue à quel point la couleur entre deux points appariés, $\mathbf{p}$ et $\mathbf{q}$, est différente. Dans la suite, au § 6.1.1, nous allons détailler les choix réalisés pour la fonction de dissimilarité ρ.

Mesures de corrélation testées – Nous avons choisi de tester deux approches présentées en page 71 : ZNCC, cf. § B.1.1, et CENSUS [Zabih 94], cf. § B.1.4. Pour généraliser à la couleur, comme proposé dans [Chambon 05b], pour ZNCC, le calcul est réalisé sur les trois composantes directement, car l'expression se généralise simplement en dimension 3, alors que pour CENSUS, les calculs sont réalisés séparément sur chaque composante et ensuite nous faisons la somme des trois résultats.

Fonction de dissimilarité – Nous utilisons la **valeur absolue de la différence**. Il s'agit simplement du coût de dissimilarité le plus utilisé en mise en correspondance globale. Bien sûr, cette mesure ne permet pas de prendre en compte les distorsions radiométriques et la comparaison à cette fonction doit nous permettre d'identifier les coûts qui donnent le meilleur comportement dans un cas de distorsions radiométriques. En utilisant la couleur, nous appelons ce coût ρ_{AD} *Absolute Difference*.

$$\rho_{AD}(\mathbf{p}_g^{i,j}, \mathbf{p}_d^{u,v}) = \sum_{k=1,\ldots,3} |c_k(\mathbf{p}_g^{i,j}) - c_k(\mathbf{p}_d^{u,v})| \tag{6.5}$$

où $c_k(\mathbf{p})$ correspond à la valeur de la ième composante couleur associée au pixel $\mathbf{p}$. Il faut noter que nous comparons nos résultats à ceux obtenus avec les images en niveaux de gris et dans ce cas $\rho_{AD}\mathbf{p}_g^{i,j}, \mathbf{p}_d^{u,v}) = |I_g(i,j) - I_d(u,v)|$. Nous utilisons également l'**information mutuelle**, les détails sont donnés au § B.1.6.

Systèmes de représentation de la couleur – Nous avons considéré les systèmes étudiés dans [Bleyer 08a, Chambon 05b] à l'exclusion de HSI qui a montré de faibles performances dans [Bleyer 08a, Chambon 05b]. Nous avons ainsi pris en compte trois types de systèmes différents :

(1) Les systèmes de primaires : RGB et XYZ ;

3. Empiriquement, nous avons choisi $P_1 = \frac{P_2}{2}$. Et nous avons choisi P_2 pour chaque coût de mise en correspondance testé dans nos expérimentations. Le but de ce choix est de permettre de tester les meilleurs résultats que nous pouvons espérer pour chaque coût et ainsi obtenir les résultats les plus favorables pour chacun.

(2) Les systèmes Luminance-Chrominance : L_{UV} , L_{AB}, AC_1C_2 et YC_1C_2 ;

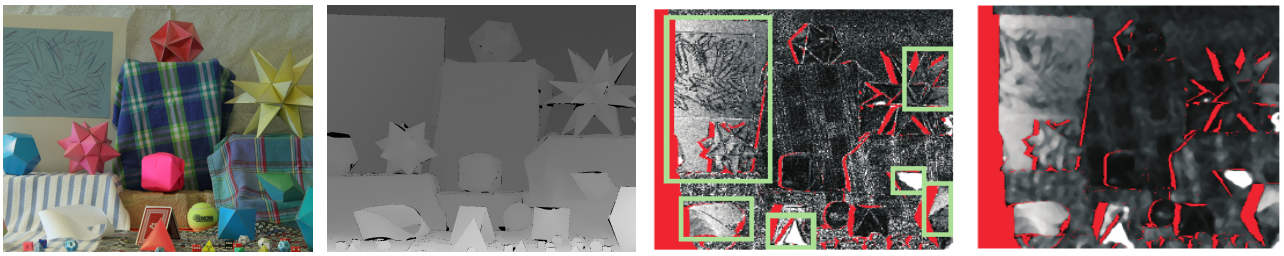
(3) Les systèmes de composantes indépendantes : $I_1I_2I_3$ et $H_1H_2H_3$.

Plus de détails sont donnés sur ces systèmes, dans l'annexe B.2. De plus, nous fournissons également le résultat obtenu en niveaux de gris. Ainsi, si nous combinons quatre coûts de mise en correspondance avec neuf systèmes de représentation, nous avons donc trente-six coûts possibles.

6.1.2. Optimisation du coût de mise en correspondance

Il existe de nombreux travaux sur l'optimisation de coûts [Veksler 05] comme celui donné par l'équation (6.1). Une méthode d'optimisation classique très utilisée est la propagation de croyance [Sun 05b] ou les approches basées coupure de graphe [Deng 05]. Étant donné le nombre important de coûts que nous souhaitons évaluer, il paraît important de favoriser des approches peu coûteuses en temps de calcul. En conséquence, nous avons choisi d'utiliser les coûts de correspondance proposés dans la méthode d'optimisation introduite par [Bleyer 08b]. Cette approche consiste à construire un graphe des 4 voisins en chaque pixel et de chercher l'optimum dans le graphe ainsi construit en utilisant la programmation dynamique. Cette approche tente également de prendre en compte le problème des occultations en effectuant une mise en correspondance bidirectionnelle pour les détecter. Cette détection préalable des occultations est ensuite intégrée dans le coût global à minimiser.

6.1.3. Étude des zones de distorsions radiométriques



(a) (b) (c) (d)

FIGURE 6.2. – Estimation des zones de distorsions radiométriques – En (a), il s'agit de l'image *Moebius*, en (b) il s'agit de la carte de disparités théoriques associée et enfin en (c), nous pouvons visualiser les coûts associés à la vérité terrain. Plus le pixel est clair et plus le coût associé est important. Les pixels rouges correspondent aux pixels occultés. En (d), nous visualisons les zones de distorsions radiométriques, obtenues en prenant le filtre médian de (c).

Dans cette étude, nous souhaitons mettre en évidence les différences de résultats obtenues dans les zones affectées par des distorsions radiométriques et dans celles non-affectées. Pour cela, il a été nécessaire de mettre en œuvre un outil afin de détecter ces régions dans les images testées, de manière suffisamment fiable, afin de s'en servir comme vérité terrain. Pour cela, nous avons utilisé les coûts associés à la vérité terrain. C'est-à-dire que pour tout pixel non occulté, nous récupérons le correspondant attribué par la vérité terrain. Ensuite, nous calculons la valeur absolue de la différence des niveaux de gris. Nous construisons ainsi une carte de ces valeurs absolues. Nous pouvons voir un exemple de résultat obtenu dans la figure 6.2 où plus le pixel est clair et plus cette différence est élevée. Intuitivement, nous pouvons comprendre que les zones étendues les plus claires, comme celles marquées en vert dans cette illustration, correspondent aux régions de distorsions radiométriques entre les deux images, c'est-à-dire les zones que nous cherchons à mettre en avant. Nous rappelons que ces distorsions radiométriques peuvent être la conséquence de changements d'illumination ou de temps d'exposition variable d'une image à l'autre. Toutefois, ces distorsions peuvent également être

la conséquence d'un bruit lié au capteur, ou d'artefacts dus à l'échantillonnage [Birchfield 98] mais également au problème de *stereo matting* [Bleyer 09, Xiong 07]. Pour tenter d'être robuste au bruit, nous appliquons un masque médian, ce qui nous permet d'obtenir le résultat en (c) [4].

6.1.4. Résultats de mise en correspondance globale utilisant la couleur

Tests visuels – Nous avons choisi de tester les images de la base de données de [Scharstein 02a], présentées au § 5.5. Pour chaque paire, nous avons estimé les cartes de distorsions géométriques, comme nous pouvons voir dans la figure B.2, annexe B.3. Dans ce jeu de données, bien que les images aient été acquises dans un environnement très contrôlé, elles contiennent d'importantes zones de distorsions géométriques. De plus, pour certaines, les plus récentes, les auteurs ont choisi de fournir des versions comportant des variations d'illuminations et des variations de temps d'exposition afin de simuler des distorsions radiométriques.

Tests quantitatifs sur l'apport de la couleur dans une approche globale – Pour réaliser ces tests, nous avons classiquement calculé le pourcentage d'erreurs parmi les pixels non occultés, noté $\mathrm{Er}_{\mathcal{V}}$, comme dans [Scharstein 02b]. Il s'agit du calcul décrit au § 5.5.3, appliqué uniquement aux pixels non occultés. Ce qui est plus spécifique à notre étude correspond à l'ajout des deux critères suivants : le pourcentage d'erreur dans les zones de distorsions radiométriques, noté $\mathrm{Er}_{\mathcal{D}}$ et à l'extérieur de ces zones, noté $\mathrm{Er}_{\mathcal{C}}$, cf. les cartes de distorsions radiométriques présentées dans la figure B.2, annexe B.3. Nous précisons que, pour ces trois critères, nous considérons qu'il y a une erreur si la différence entre le pixel correspondant calculé et celui de la vérité terrain est supérieur à 1. Pour les zones de distorsions et les zones hors distorsions, nous pondérons l'influence de chaque pixel par la valeur estimée pour calculer la carte de distorsions radiométriques. À ce stade, il est important de remarquer que la carte de distorsions radiométriques n'est pas binaire car nous avons souhaité nous affranchir du choix d'un seuil de binarisation. Par la suite, nous utilisons également les notations suivantes : $\overline{\mathrm{Er}_{\mathcal{V}}}$, $\overline{\mathrm{Er}_{\mathcal{D}}}$ et $\overline{\mathrm{Er}_{\mathcal{C}}}$ qui correspondent aux moyennes de ces erreurs sur l'ensemble des images testées.

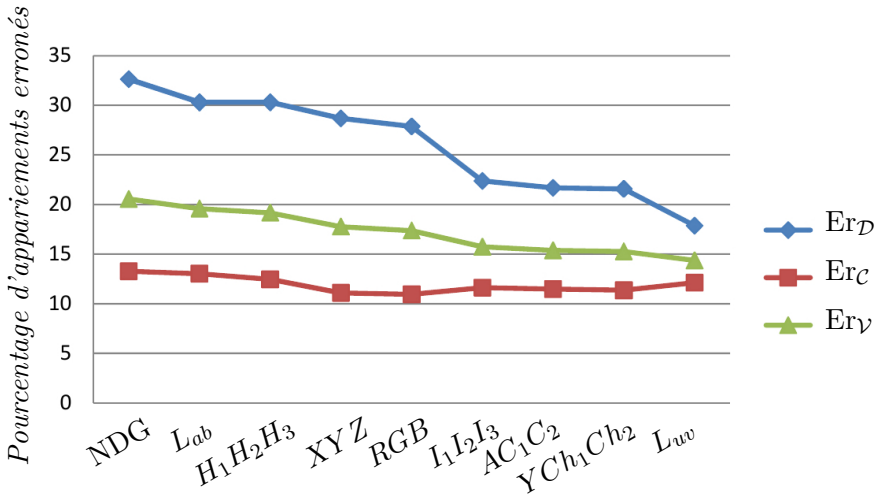


FIGURE 6.3. – Évaluation quantitative réalisée pour les niveaux de gris (NDG) et les 8 systèmes de représentation de la couleur utilisés – Pour tous les tests, la fonction de dissimilarité est une norme L_1. De manière analogue aux résultats obtenus dans notre travail précédent [Bleyer 08a], nous pouvons remarquer que l'utilisation de la couleur permet d'obtenir des résultats de meilleure qualité que lorsque les niveaux de gris seuls sont utilisés (cf. les résultats obtenus avec $\mathrm{Er}_{\mathcal{V}}$).

4. Nous avons choisi des tailles de 7 de manière empirique et en nous appuyant sur les valeurs fréquemment utilisées dans la littérature.

Les évaluations quantitatives sont présentées dans la figure 6.3 et confirment que la couleur permet d'obtenir de meilleurs résultats que ceux obtenus avec les niveaux de gris. En particulier, les systèmes luminance-chrominance semblent les plus adaptés. De plus, ces résultats illustrent le fait que les performances sont améliorées dans les zones de fortes distorsions radiométriques sans pour autant dégrader les résultats obtenus dans les zones sans distorsions. En effet, le critère $\mathrm{Er}_{\mathcal{D}}$ évolue de 32.6%, quand les niveaux de gris sont utilisés, à 17.8% quand c'est *Luv*. Lorsque nous utilisons des images en niveaux de gris, avec notre ensemble d'images testées acquises avec un éclairage contrôlé [5], nous avons mis en évidence qu'une grande partie des résultats erronés provient des zones de distorsions radiométriques. Plus précisément, 57% des erreurs proviennent de ces zones alors qu'elles ne représentent qu'une sous-partie de l'image, en moyenne 33%. Dans la figure 6.3, le système *Luv* obtient les résultats les plus intéressants. La particularité de ce type de système est de tenter de séparer l'intensité (contenue dans L), de la couleur « pure » (décrite dans les canaux u et v). Nous supposons que le gain de performance obtenu pour ce système provient de cette séparation. C'est la raison pour laquelle, nous étudions les résultats obtenus en utilisant uniquement L ou uniquement u (nous n'affichons pas v qui produit des résultats similaires à u) dans la figure 6.4. Dans cette figure, nous exposons les résultats dans les zones de distorsions radiométriques où la composante couleur obtient des résultats bien plus performants et dans les zones sans distorsions radiométriques où cette tendance s'inverse. Ainsi, la conclusion de cette étude est qu'il est nécessaire de combiner l'utilisation des deux pour obtenir des résultats exploitables dans les deux types de zones. Afin de combiner au mieux cette information, nous pensons qu'il est préférable d'utiliser des systèmes de représentation de la couleur qui permettent de distinguer ces 2 informations. Dans les systèmes $H_1H_2H_3$, XYZ, RGB et $I_1I_2I_3$, la couleur et l'intensité sont « présentes » dans chaque canal alors que dans les systèmes AC_1C_2 , YC_1C_2 et Luv, une composante est dédiée à l'intensité et les deux autres à la couleur, ce sont donc plutôt des systèmes à favoriser [6].

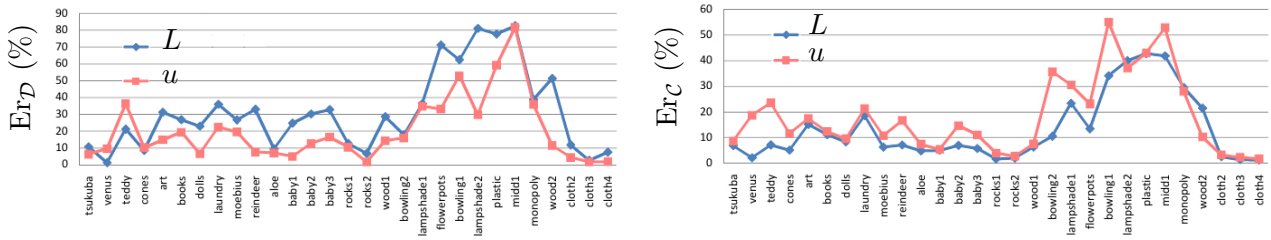


FIGURE 6.4. – Comparaison des coûts obtenus avec la composante intensité seule (composante L) et une des composantes chromatiques seule (composante U) – Nous avons comparé les résultats obtenus séparément afin d'évaluer l'impact de chaque type de canal sur la qualité du résultat obtenu. Les résultats exposés à gauche, respectivement à droite, illustrent le fait que l'utilisation du canal associé à la chromaticité permet d'obtenir de meilleurs résultats, respectivement de moins bons résultats, que ceux obtenus en utilisant seulement la composante de luminosité, respectivement dans les zones avec distorsions radiométrique et celles sans distorsions radiométriques.

Enfin, nous comparons les résultats obtenus à ceux obtenus par les approches prenant directement en compte les problèmes de distorsions radiométriques, cf. figure 6.5 : la valeur absolue, notée AD, des différences en niveaux de gris et en couleur, en utilisant le système *Luv* (meilleur résultat de la figure 6.3, notée *ADLuv*, ainsi que les mesures de corrélation suivantes : MI (*Mutual Information*), NCC et CENSUS sur des images en niveaux de gris [7]. Les résultats obtenus montrent que les mesures NCC et CENSUS qui ont la capacité de prendre en compte les difficultés dans les zones de distor-

5. Il est important de souligner que le jeu de données comporte donc très probablement moins de zones problématiques que dans des acquisitions réelles.
6. Il est nécessaire de faire une exception pour le système *Lab* pour lequel nous n'avons pas obtenu de bons résultats.
7. Les détails sur ces mesures sont donnés en annexe B.1.

sions radiométriques obtiennent des résultats aussi satisfaisants que ceux obtenus par des mesures classiques appliquées à la couleur. De plus, dans les zones sans difficultés, elle permettent également d'obtenir de meilleurs résultats. Un exemple assez significatif est l'amélioration obtenue pour $\mathrm{Er}_{\mathcal{V}}$ qui diminue de 20.5% pour AD_Gris à 6.7% pour CENSUS. En revanche, l'information mutuelle MI ne permet pas d'obtenir de meilleurs résultats par rapport à ces deux mesures et obtient le même niveau que AD_Couleur. Au vu des résultats obtenus, nous comprenons mieux pourquoi lorsque nous utilisons la couleur pour appliquer ZNCC et CENSUS, l'amélioration des résultats n'est pas aussi significative qu'avec les autres mesures, comme nous avons pu le voir dans nos résultats préliminaires dans [Chambon 05a, chapitre 4] ou encore dans les travaux de [Hirschmüller 09]. Dans ce dernier papier, les auteurs ont déclaré que la couleur n'avait pas d'intérêt et que l'information d'intensité seule était bien suffisante. Nos résultats affinent ce commentaire en ajoutant qu'effectivement, l'apport de la couleur reste minime si les images acquises ne présentent pas de régions de distorsions radiométriques. De plus, cette difficulté peut être prise en compte si le coût utilisé est ZNCC ou CENSUS. Ces coûts sont plus coûteux en termes de temps de calcul, ainsi, utiliser la couleur combinée à une mesure classique a l'intérêt de diminuer les temps de calcul, tout en préservant une certaine robustesse dans ces zones de distorsions radiométriques.

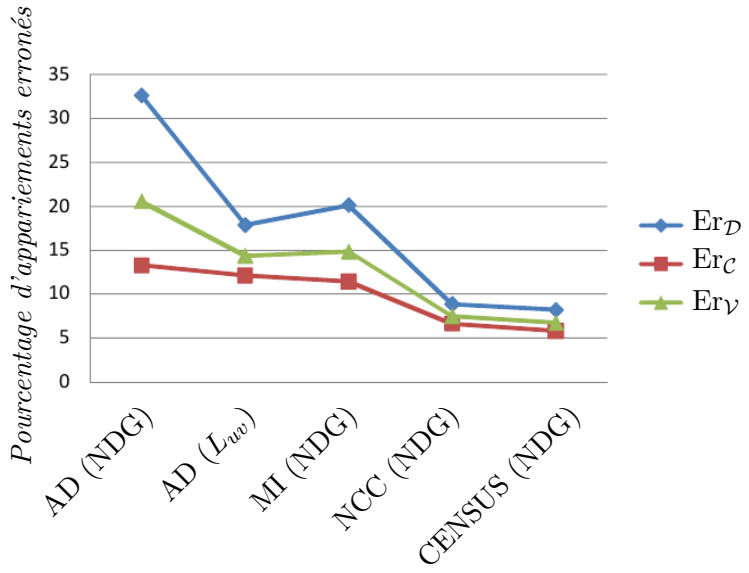


FIGURE 6.5. – Erreurs d'appariement avec ou sans couleur, avec méthode globale ou locale. – Nous montrons les résultats pour deux mesures de corrélation standard, NCC et CENSUS, ainsi que MI, *Mutual Information*. Dans les deux cas, l'utilisation de la couleur permet d'améliorer visuellement la qualité des résultats obtenus.

6.1.5. Conclusion sur l'utilisation de la couleur dans des approches globales

Ce travail nous a permis d'approfondir celui amorcé dans [Chambon 05a, chapitre 4]. D'une part, nous avons mis en lumière d'autres zones où la couleur permet d'améliorer la qualité des résultats : les zones de distorsions radiométriques. D'autre part, nous avons identifié dans quelles conditions ces améliorations sont significatives : lorsqu'un coût local faisant intervenir une distance classique, comme une valeur absolue, est utilisé. De plus, en accord avec les travaux développés dans [Hirschmüller 09], nous pensons que dans de nombreuses situations, l'information de couleur disponible n'est pas réellement exploitable, en particulier, quand nous n'avons pas d'information sur le blanc de référence [Can Karaimer 18]. C'est le cas dans le jeu de données Middlebury

que nous utilisons. Nous pourrions approfondir cet aspect, en partant des publications de Michael Brown [8] [Nguyen 17, Can Karaimer 19]. Il serait nécessaire de définir le minimum d'information nécessaire à obtenir sur la couleur pour que celle-ci soit réellement exploitable. Avec cette connaissance, nous pourrions envisager, d'une part, une adaptation des approches proposées pour intégrer cette information, et, d'autre part, des adpatations possibles dans le cas où elle n'est pas disponible. Bien entendu, les travaux que nous avons présentés ici peuvent être encore plus approfondis en ajoutant d'autres approches d'optimisation, comme celles citées dans [Chambon 05a, chapitre 1], ou encore, en testant d'autres coûts de lissage, comme ceux présentés dans [Woodford 08] et d'autres coûts de dissimilarité [Chambon 05a, Yoon 07, chapitre 3]. Enfin, dans la même idée que les travaux développés dans [Tombari 08, Wang 06], nous pensons qu'il est intéressant d'intégrer au coût de mise en correspondance d'autres aspects, comme l'utilisation d'autres mesures de corrélation, c'est ce que nous allons étudier dans le § 6.2, ou l'introduction de notion de segmentation ou de résultats de segmentation pour conditionner la propagation des correspondances, c'est ce que nous allons étudier dans les § 6.3 et 6.4.

6.2. Combinaisons de mesures de corrélation

Dans le cadre des méthodes locales, il est encore difficile de mesurer les qualités respectives de chacune des méthodes proposées, en particulier de chacune des mesures de corrélation introduites. Pourtant, ces méthodes restent actuellement les plus rapides et les plus utilisées pour des applications assez diverses, comme la robotique, l'étude de déformations de surfaces [Sutton 08]. La première étude que nous avons menée [Chambon 05b, chapitre 3 à 5] a mis en évidence l'intérêt des mesures robustes, dans les zones présentant des occultations, ainsi que l'intérêt des mesures classiques, dans les zones sans occultation. Nous avons donc proposé des approches permettant de les combiner. Les résultats obtenus sont encourageants mais ils restent encore trois questions à explorer :

(1) De manière plus générale, quelles sont les mesures de corrélation qui ont une réelle « complémentarité » ?

(2) Est-il judicieux de combiner plus de mesures de corrélation et si oui, combien ?

(3) Si nous arrivons à répondre à ces deux questions, est-il envisageable de proposer de nouveaux algorithmes plus performants que ceux initialement proposés ?

Ainsi, nous avons repris les 41 mesures étudiées et présentées dans [Chambon 11c], et nous avons évalué la qualité des résultats d'appariement sur les 42 couples d'images présentés au § 5.5 afin d'étudier la complémentarité de ces mesures. En particulier, nous étudions la complémentarité des mesures que nous avions identifiées comme les plus performantes par famille de mesures, cf. les mesures présentées au § B.1. Pour cette étude, bien qu'il soit possible de quantifier la qualité des résultats suivant de nombreux critères, cf. [Chambon 05a, chapitre 2], nous nous concentrons sur un unique critère, le pourcentage d'appariements erronés, que nous notons Er.

Nous notons que peu de travaux se sont intéressés à ce type de combinaison. L'approche la plus proche que nous ayons trouvée est celle de [Wegner 09] où les auteurs comparent et combinent quatre mesures différentes (SAD, *Sum of Absolute Differences*, somme des valeurs absolues des différences, RANK [Bhat 98], GC, cf. § B.1.3 et CENSUS, cf. § B.1.4). Pour combiner ces mesures, ils utilisent un produit deux à deux des scores de corrélation. La mesure est utilisée au sein d'un algorithme global de mise en correspondance par propagation de croyance [Sun 03]. L'étude que nous proposons approfondit ce travail car, d'une part, nous étudions la notion de complémentarité de ces mesures, et, d'autre part, nous généralisons ce travail à toutes les mesures présentées dans [Chambon 11c].

8. rencontré lors de ma visite à l'Université de Singapour en 2016.

6.2.1. Étude de complémentarité

Plus précisément, nous étudions le pourcentage d'appariements erronés pour chaque mesure seule, puis, pour chaque combinaison de mesures possible. Dans un premier temps, nous envisageons une combinaison théorique toujours juste. C'est-à-dire que sur un ensemble de candidats correspondants obtenus par différentes mesures, nous gardons systématiquement le bon correspondant, le bon candidat, si une des mesures a été en capacité de l'identifier. Le résultat de cette étude est fourni dans la figure 6.7 qui montre les résultats du pourcentage d'appariement pour une combinaison de 2 à 41 mesures. Les résultats sont présentés sous deux formes différentes : un tableau avec les valeurs de pourcentages (pour une lecture précise) et une courbe de l'évolution des pourcentages moyens d'erreurs suivant le nombre de mesures combinées (pour visualiser le comportement).

Notations – Nous utilisons les notations suivantes :

- M_i, avec $i \in \{1; \ldots; N_m\}$, correspond aux N_m mesures testées.
- $d_{th}(\mathbf{p})$ est la disparité théorique associée au pixel $\mathbf{p}$. C'est-à-dire celle donnée par la carte de disparités théoriques.
- $d_i(\mathbf{p})$ est la disparité donnée par l'application de l'algorithme de mise en correspondance locale en utilisant la mesure M_i.
- $d_{tc}^{N_m}(\mathbf{p})$ est la disparité obtenue en utilisant la combinaison théorique ou optimale de N_m mesures.

Combinaison théorique – Nous définissons la notion de combinaison théorique optimale. Elle est obtenue pour chaque pixel $\mathbf{p}$ de la manière suivante :

> Si $\exists\ i \in \{1; \ldots; N_m\}$ tel que $d_i(\mathbf{p}) = d_{th}(\mathbf{p})$
> alors $d_{tc}^{N_m}$ est correcte, c'est-à-dire que $d_{tc}^{N_m} = d_{th}(\mathbf{p})$,
> sinon $d_{tc}^{N_m}$ est erronée.

Configurations étudiées – Ainsi, nous avons étudié les configurations suivantes :

(C_1) $N_m = 2$: toutes les combinaisons de mesures, soit 41×40 combinaisons, ont été testées.

(C_2) $N_m = 41$: les 41 mesures ont été théoriquement combinées.

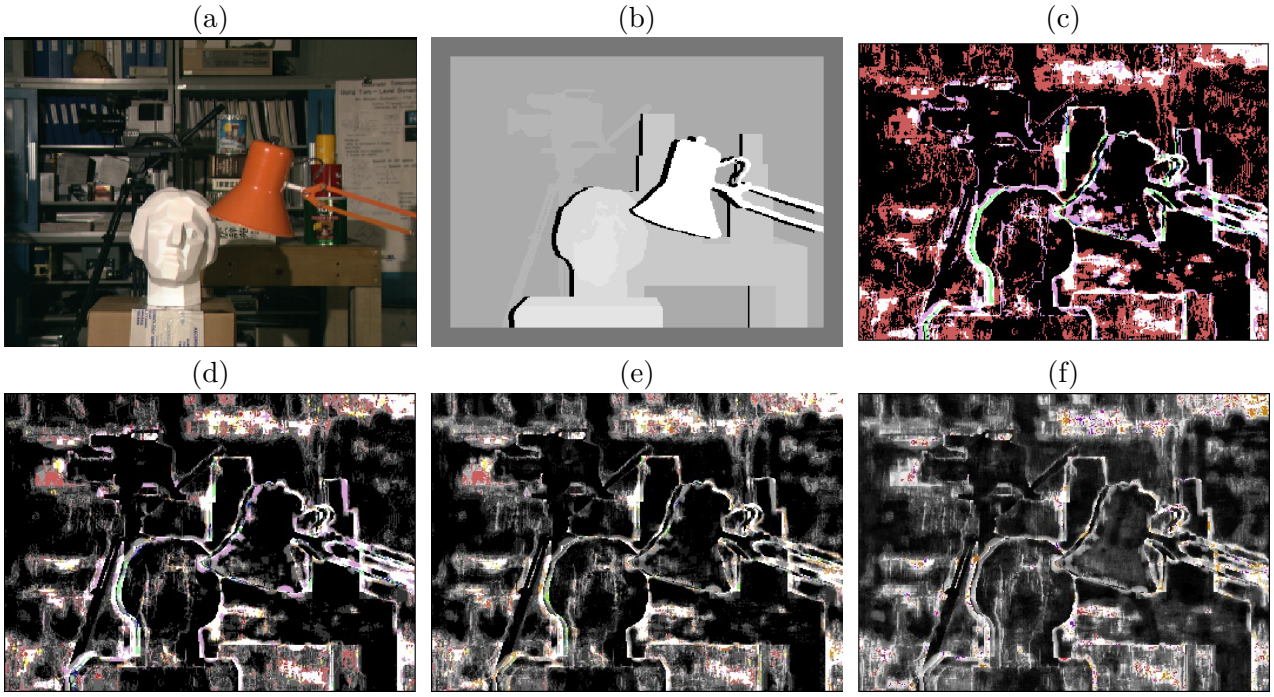
(C_3) $N_m \in \{3; \ldots; 40\}$: nous avons testé toutes les combinaisons entre 3 et 40 mesures. Nous avons donc effectué $\sum_{k=3}^{40} C_{41}^k$ combinaisons différentes, avec C_{41}^k le nombre de possibilités pour prendre k éléments parmi 41.

Dans la suite de ce paragraphe, nous faisons référence à un certain nombre de mesures de corrélation, dont les formulations sont rappelées en annexe B.1 : NCC, cf. B.1.1, LSAD, cf. § B.1.2, GC, cf. § B.1.3, CENSUS, cf. § B.1.4 et SMAD, cf. B.1.5. Elles correspondent aux mesures obtenant les meilleurs résultats, par catégorie, cf. [Chambon 11c].

Résultats obtenus – Afin de comprendre les comportements de ces combinaisons, nous présentons en premier lieu une analyse visuelle sur un exemple donné dans la figure 6.6 :

(1) *Visualisation des disparités correctes pour les deux mesures les plus complémentaires, cf. figure 6.6.(c)* – Comme attendu, SMAD et GC se complètent avantageusement : SMAD permet de compenser les erreurs de GC dans les zones proches des discontinuités et occultations alors que GC compense celles de SMAD dans les zones faiblement texturées.

(2) *Visualisation des disparités correctes en combinant 5, 10 ou 41 mesures, cf. figures 6.6.(d), (e) et (f)* – Ces résultats montrent que GC semble la mesure la plus complémentaire aux autres et SMAD est la deuxième plus complémentaire, ce qui est assez cohérent avec le premier résultat visuel exposé. Il faut également remarquer que le résultat que nous montrons pour la combinaison de 10 mesures, en (d), est assez proche de celui avec 41 mesures, en (f).



Mesure	NCC	LSAD	GC	CENSUS	SMAD
Zones sans occultation					
Zones d'occultations					

FIGURE 6.6. – Étude de la complémentarité – En (a), nous présentons une image de la paire étudiée et, en (b), il s'agit de la carte de disparité théorique. Pour les cartes présentées de (c) à (f), les pixels en niveaux de gris correspondent aux pixels où plus d'une mesure a donné le bon appariement parmi les N_m combinaisons possibles. Plus le pixel est foncé, plus le nombre de bons appariements est important. Le code couleur permet d'indiquer quand une seule mesure a fourni le bon correspondant. De plus, nous avons distingué les zones d'occultations des zones sans occultation.

Pour renforcer cette analyse visuelle, nous commentons les résultats quantitatifs présentés dans la figure 6.7 :

- Grâce à (C_1), nous avons mis en évidence la combinaison permettant la complémentarité la plus importante, c'est-à-dire celle qui permet d'obtenir les meilleures performances suivant notre protocole d'évaluation. Ainsi, la combinaison GC/SMAD permet de réduire le pourcentage moyen des erreurs à 14.13% (moyenne calculée sur les 42 images de test).
- Grâce à (C_2), nous avons mis en évidence la limite supérieure qui peut être atteinte, en combinant toutes les mesures et en supposant une fusion parfaite, à savoir une réduction du pourcentage moyen des erreurs à 7.26%.
- Grâce à (C_3), nous remarquons qu'en ajoutant une mesure à la combinaison la plus performante, i.e. GC/SMAD, quelle que soit la mesure ajoutée, le pourcentage moyen des erreurs est à 13%. Soit, 1% de mieux qu'avec GC/SMAD. De manière générale, cette erreur diminue régulièrement jusqu'à 10 mesures combinées pour atteindre des performances similaires à celles obtenues avec 42 mesures combinées.

En conclusion, combiner des mesures peut apporter une amélioration significative de la qualité des résultats puisque nous pouvons espérer réduire de 7% (en combinant 2 mesures) à 17% (avec 41 mesures) par rapport au meilleur résultat de corrélation à une seule mesure (à savoir 20% avec la

mesure CENSUS). Dans le paragraphe suivant, nous avons proposé un premier algorithme de mise en correspondance dense permettant de combiner les résultats obtenus par N_m mesures différentes et nous avons validé ce nouvel algorithme en combinant GC et SMAD.

Mesure	Er
NCC	23.2
LSAD	23.3
GC	21
CENSUS	**20.2**
SMAD	27.9
GC-SMAD	**14.13**

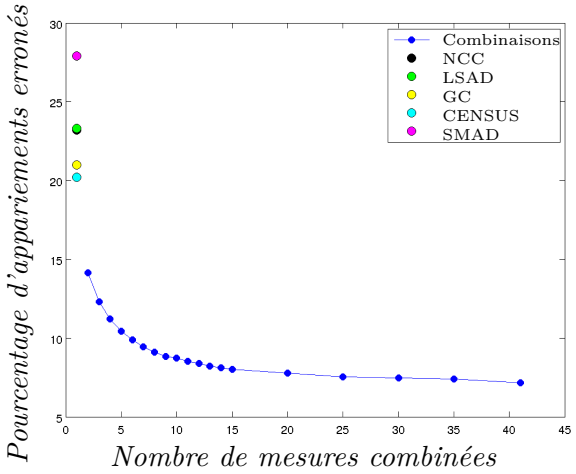


Figure 6.7. – Erreurs d'appariement en fonction du nombre de mesures théoriquement combinées – L'objectif de ce graphe et de ce tableau (qui montrent la même information) est d'illustrer le nombre maximal de mesures qu'il est nécessaire de combiner afin d'obtenir les performances optimales théoriques. Il apparaît que l'utilisation de 10 mesures permet d'obtenir le maximum théorique. Une seconde remarque importante : à partir de 2 mesures combinées, le gain est le plus important.

6.2.2. Algorithme de mise en correspondance par fusion de N_m mesures

Nous avions déjà proposé des algorithmes de fusion dans [Chambon 05a, chapitre 5], dans le but de prendre en compte les différences de résultats obtenues dans les zones d'occultations. Pour cela, nous nous appuyions sur la détection des zones d'occultations, tâche délicate. Les algorithmes proposés permettaient des améliorations mais elles étaient limitées par la qualité de la détection des occultations. De plus, la séparation du comportement de chaque mesure dans chacune des zones n'est pas aussi nette et parfois dans des zones sans occultation, les mesures robustes peuvent obtenir un meilleur résultat que certaines mesures classiques et inversement. Ainsi, le principe de cette nouvelle approche est de ne plus chercher à détecter ces zones directement mais à fusionner les résultats afin de combiner l'intérêt de chaque mesure pour tout type de zones, avec ou sans occultation, avec ou sans problème de texture, avec ou sans problème d'éclairage. Le principe de cet algorithme est de donner plus d'importance à la disparité qui obtient le plus de votes. Ainsi, nous distinguons deux étapes : l'estimation des disparités pour chaque mesure puis la fusion. Cet algorithme applique les deux règles suivantes, relative aux deux étapes :

(1) *Initialisation pour chaque pixel* $\mathbf{p}$ – Si plus de la moitié des mesures donnent le même correspondant alors cette correspondance est validée et est considérée comme fiable. Nous appelons $d_f^{N_m}$ la disparité finale conservée après la fusion des N_m mesures. De plus, nous définissons $N_d(\mathbf{p})$ qui permet de compter le nombre de mesures qui donnent la disparité d :

$$N_d(\mathbf{p}) = \#\{i | d_i(\mathbf{p}) == d\}.$$

Ainsi, l'initialisation se formalise avec :

$$\begin{array}{ll} \textbf{Si} & \exists\ d \mid \left(\left(d = \underset{d}{\operatorname{argmax}}\ N_d(\mathbf{p})\right)\ \&\&\ \left(N_d(\mathbf{p}) \geq \frac{N_m}{2}\right)\right) \\ \textbf{alors} & d_f^{N_m}(\mathbf{p}) \leftarrow d \\ \textbf{sinon} & \text{la disparité est indéterminée.} \end{array} \tag{6.6}$$

(2) *Affinement* – Pour les pixels indéterminés, pour chaque mesure utilisée, nous évaluons la fiabilité de la disparité estimée. Plus précisément, plus la disparité moyenne dans son voisinage est proche de sa propre disparité, plus cette disparité est considérée comme fiable. Plus précisément, nous estimons l'ambiguïté, notée f_A, pour chaque disparité possible $d_i(\mathbf{p})$. Nous supposons que cette fonction dépend des disparités des pixels voisins, disparités obtenues par la même mesure. Plus précisément, elle est évaluée en prenant l'écart de la disparité étudiée à la disparité moyenne dans son voisinage, soit :

$$f_A(d_i(\mathbf{p})) = |d_i(\mathbf{p}) - \frac{1}{\#\mathcal{V}(\mathbf{p})} \sum_{\mathbf{p}' \in \mathcal{V}(\mathbf{p})} d_i(\mathbf{p}')|. \tag{6.7}$$

avec $\mathcal{V}(\mathbf{p})$ le voisinage[9] de $\mathbf{p}$. Nous estimons donc ainsi la disparité la plus fiable, pour chaque pixel $\mathbf{p}$:

$$d = \underset{i \in \{1, N_m\}}{\operatorname{argmin}}\ f_A(d_i(\mathbf{p})). \tag{6.8}$$

Enfin, cette disparité d est conservée uniquement si l'ambiguïté associée n'est pas trop importante, c'est-à-dire seulement si $f_A(d(\mathbf{p}) < \epsilon$, sinon, le pixel est considéré comme occulté[10].

6.2.3. Résultats de mise en correspondance par combinaison de mesures

Méthode	**H+O**	**O**	**H**	**R**
GC seule	25.6	17.5	19.6	15.9
Fusion GC-SMAD proposée	22.1	13.5	16.9	13.5
Fusion GC-SMAD théorique	19.4	10.8	15.4	10.8

Table 6.1. – Pourcentage d'appariements erronés en fonction des mesures combinées – **H** représente les résultats obtenus pour les images avec des zones non-texturées, comme la paire *Tsukuba*, **O**, les images qui contiennent de nombreuses occultations, comme la paire *Aloe* et **R**, les images sans difficulté majeure, comme la paire *Cones* (cf. la figure 6.8 pour les exemples de ces images). Le pourcentage de correspondances erronées obtenu avec la méthode de fusion proposée permet d'améliorer les résultats obtenus avec la mesure GC seule, et en particulier, dans les cas de scènes complexes.

Nous présentons ici les résultats de fusion obtenus avec les mesures GC et SMAD. Le tableau 6.1 montre les améliorations obtenues sur le pourcentage de correspondances erronées sur les 42 images réelles de la base de données présentées au § 5.5. Cette amélioration est de 2.47, pour les images présentant les scènes les plus simples, à 4.08, avec les images présentant les scènes les plus complexes, c'est-à-dire avec de nombreuses zones d'occultations ou de zones non texturées. Cependant, nous pouvons souligner que cette amélioration n'atteint pas l'amélioration théorique présentée dans le tableau de gauche de la figure 6.7. Une autre manière d'apprécier la qualité des résultats obtenus

9. Les 8 voisins ont été pris en compte dans nos tests.
10. Nous avons choisi $\epsilon = 1$.

est d'observer les cartes de disparités obtenues, comme celles présentées dans la figure 6.8. Les cartes de disparités obtenues par fusion sont visuellement les plus appréciables car elles présentent le moins de faux négatifs (c'est-à-dire les fausses occultations qui sont notées en noir). De plus, les véritables zones d'occultations sont mieux délimitées dans ces cartes de résultats.

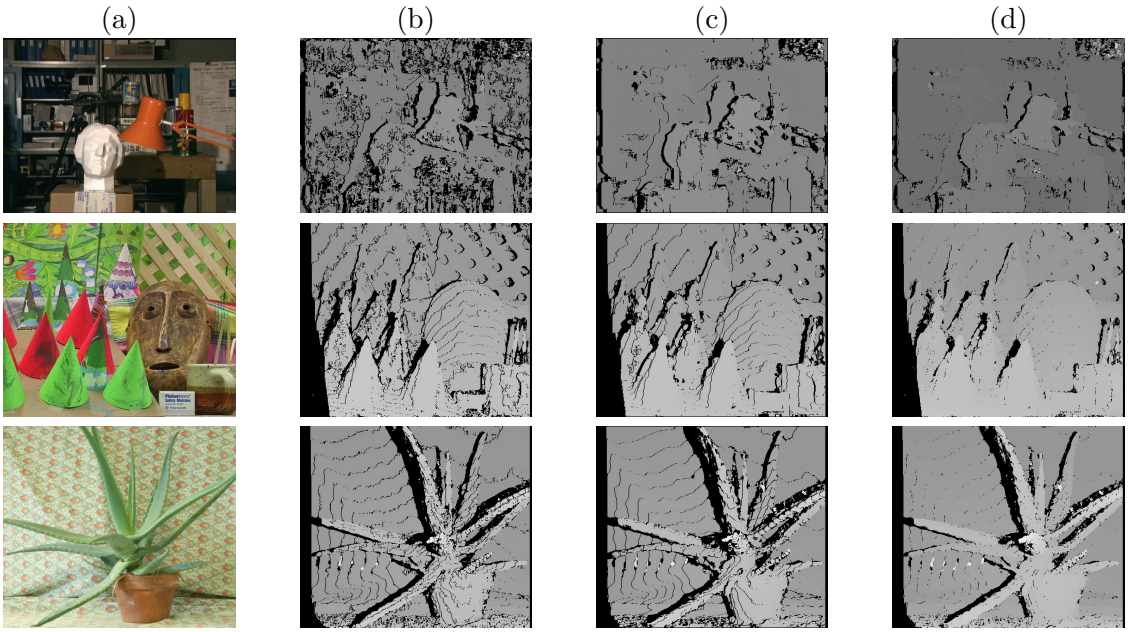


FIGURE 6.8. – Cartes de disparités obtenues avec la combinaison automatique de 2 mesures – En (a), il s'agit de l'image de gauche, puis les cartes de disparités sont présentées pour SMAD, en (b), pour GC, en (c) et enfin pour la FUSION, en (d). Comme attendu, les résultats de la fusion présentent moins de faux négatifs, en particulier avec les images *Cones* et *Aloe*. En revanche, l'exemple de *Tsukuba* illustre les limites de la méthode.

6.2.4. Conclusion sur l'appariement par combinaison de mesures de corrélation

Ce paragraphe nous a permis de présenter un étude approfondie sur l'origine des meilleurs résultats obtenus pour chaque mesure suivant son contexte d'utilisation. Pour cela, une première étude de la complémentarité des mesures a mis en évidence la combinaison la plus avantageuse : GC et SMAD. Nous avons ensuite introduit un nouvel algorithme de mise en correspondance permettant de les combiner automatiquement. Ces résultats préliminaires sont encourageants et nous permettent de mettre en avant les perspectives de ce travail de fusion, en soulevant les questions suivantes :

(1) Quels peuvent être les résultats de cet algorithme de fusion si nous utilisons plus de deux mesures ? Est il possible d'atteindre les performances théoriques en combinant plus de mesures ?

(2) Ces résultats montrent qu'il est possible de trouver des points ou des régions dans l'image qui constituent des éléments fiables à apparier (ici, certains points seront toujours bien appariés quelle que soit la mesure). Pouvons nous imaginer un algorithme qui propagerait les points fiables à apparier pour obtenir les correspondances dans les régions les plus difficiles ?

(3) Si nous imaginons être en capacité de déterminer les différents types de régions présentes dans l'image (avec ou sans occultation, de même caractéristique photométrique), est-il possible de combiner ces améliorations à une sur-segmentation ?

Ce dernier questionnement a été exploré dans les travaux que nous présentons dans les deux sections suivantes, cf. § 6.3 et 6.4.

6.3. Appariement par propagation de germes

Nous avons pu constater dans les paragraphes précédents que les approches locales étudiées et pour lesquelles nous avons proposé des améliorations ont obtenu des résultats perfectibles dans les zones de difficultés telles que les zones proches des occultations ou des discontinuités de profondeur. Une autre manière de prendre en compte ces difficultés est d'utiliser une approche globale, mais, ces approches sont coûteuses et dans nos travaux, une fois de plus, nous allons chercher à atteindre ces performances tout en conservant un processus local. Comme présenté précédemment, la littérature est abondante sur le sujet mais peu de publications ont exploré des méthodes par propagation de germes.

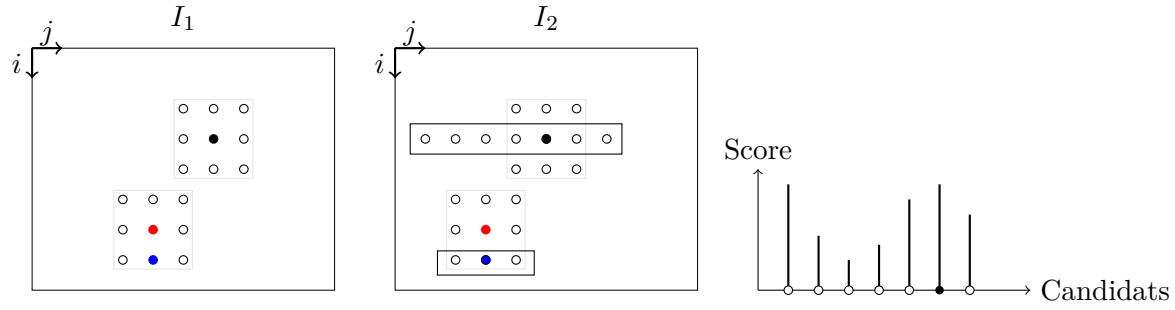


FIGURE 6.9. – Comparaison mise en correspondance par corrélation et par propagation de germes – Dans cet exemple, nous montrons que la zone de recherche dans une approche classique de mise en correspondance par corrélation est bien plus importante que dans une approche par propagation de germes. En effet, dans le premier cas, pour le point en noir, sans connaissance *a priori*, la recherche s'effectue sur toute la ligne correspondante alors que dans le second cas, pour le point bleu, la zone de recherche se réduit à un voisinage du point germe, le point rouge, pour lequel le correspondant est connu. Ainsi, dans cet exemple, la zone de recherche de 7 pixels est réduite à 3 pixels.

Le principe d'une approche par propagation de germes est de supposer qu'il est possible d'avoir un ensemble de correspondances très fiables et de propager ces correspondances pour obtenir une carte de disparités dense. Ces germes sont la plupart du temps des points d'intérêt appariés. Le premier intérêt de ce type d'approche est de diminuer les erreurs de mise en correspondance. En effet, à chaque itération, pour chercher une nouvelle correspondance, la zone de recherche est réduite par rapport à une approche classique, cf. figure 6.9. Dans cet exemple, nous observons la distribution des scores dans toute la zone de recherche prise en compte pour une approche classique par corrélation et nous constatons qu'il existe plusieurs candidats potentiels pour apparier un point (en noir) : plus précisément le candidat 1 et le candidat 6. Cette remarque nous permet d'introduire les notions d'ambiguïté et d'imprécision telles que décrites dans [De Joinville 01]. L'ambiguïté correspond à la distance entre un candidat et le second candidat correspondant. L'ambiguïté est donc de 5 pixels dans cet exemple. L'imprécision correspond à la largeur du pic principal de cette distribution. Dans cet exemple, le cinquième et le septième candidat obtiennent des scores proches. Ainsi, l'imprécision est de plus ou moins 1 pixel. Dans le cas de la propagation de germes (un germe en bleu apparaît dans cet exemple), la zone de recherche du correspondant d'un point (comme celui en rouge) du voisinage du germe est réduite au voisinage du point germe (le point bleu). Nous espérons que cela permet de réduire l'ambiguïté et l'imprécision que nous venons d'introduire. Le deuxième avantage, également lié à la zone de recherche réduite, est le gain en temps de calcul.

De manière générale, les étapes de la mise en correspondance par propagation de germes sont :

1. **Détection et mise en correspondance de germes** – Cela peut être un ensemble de points d'intérêt appariés par mise en correspondance classique par corrélation, comme dans [Otto 89, Lhuillier 02, Kannala 07, Schmid 00, Tuytelaars 08], ou une sélection réalisée directement à partir d'une carte de disparités initiales [Chen 99b], en appliquant un ensemble

de contraintes.

2. **Propagation des germes** – Deux approches sont généralement distinguées :
   - *Approche simultanée* : à chaque itération, chaque germe est propagé parmi ses 8 voisins simultanément, c'est-à-dire que nous recherchons l'ensemble des correspondants pour tous les points dans un 8-voisinage d'un germe [Chen 99b].
   - *Approche séquentielle* : à chaque itération, un seul germe est propagé. Cela signifie qu'il faut établir une stratégie pour choisir le « meilleur » germe à propager, selon un critère à définir [Otto 89, Lhuillier 02, Ma 05, Kannala 07, Čech 09]. Dans ce type de propagation, il reste à déterminer un critère pour stopper la propagation. Bien souvent, il s'agit d'un seuil sur le score de corrélation. Il est également possible d'ajouter un critère pour stopper la propagation dans les zones homogènes, comme dans [Lhuillier 02].

Dans la partie I, nous avons largement abordé la notion de détection de points d'intérêt, et, dans la suite de ce chapitre, nous allons nous consacrer à la technique de propagation. Nous avons choisi d'explorer la méthode de propagation séquentielle afin de tenter de favoriser les correspondances les plus fiables (critère que nous allons définir). Classiquement, pour la propagation, une approche par corrélation simple est utilisée mais, comme nous avons pu le constater dans les chapitres précédents, cette mesure peut contenir des ambiguïtés ou des imprécisions, cf. la distribution des scores exposée dans la figure 6.9. Pour limiter cet effet, nous proposons de combiner une mesure de corrélation avec d'autres aspects afin d'introduire la notion de mesure de confiance. Plus précisément, cette mesure combine l'utilisation d'une distance photométrique et géométrique au germe, ainsi qu'une mesure de corrélation que nous espérons la moins ambiguë possible. Cette nouvelle approche a été validée sur le jeu de données présentés au § 5.5.

6.3.1. Appariement par propagation de germes avec mesure de confiance

Nous rappelons que nous nous intéressons aux approches de propagation séquentielle comme décrit dans l'algorithme 1. Voici les notations spécifiques utilisées dans ce paragraphe :

- Nous notons $\mathcal{C}_g^{l,l'}$ la paire $(\mathbf{p}_g^{k,l}, \mathbf{p}_d^{k,l'})$ qui correspond à un germe et son point correspondant dans l'autre image.
- Nous notons $\mathcal{C}_p^{j,j'}$ la paire $(\mathbf{p}_g^{i,j}, \mathbf{p}_d^{i,j'})$ qui correspond à un point (quelconque) et son correspondant dans l'autre image.

Algorithme 1 : Algorithme de mise en correspondance par propagation séquentielle de germes – Les contraintes sont détaillées dans le § 6.3.1.

```
Fonction Mise_en_correspondance_par_propagation_séquentielle (I_g, I_d, S)
  Entrées :
  I_g, I_d : images manipulées
  S : ensemble des points germes, c'est-à-dire l'ensemble des paires C_g^{l,l'} = (p_g^{k,l}, p_d^{k,l'})
  Sorties : Carte de disparités quasi-dense
répéter
    Sélection du « meilleur » germe dans S
    pour chaque chaque pixel p_g^{i,j} dans le 8 voisinages de p_g^{k,l} faire
        Trouver le meilleur candidat p_d^{i,j'} dans la zone de recherche réduite induite par p_d^{k,l'}
        si C_p^{j,j'} = (p_g^{i,j}, p_d^{i,j'}) satisfait un ensemble de contraintes alors
            S ← S ∪ C_p^{j,j'}
jusqu'à S = ∅ ;
```

Peu de publications dans le domaine de la mise en correspondance par propagation de germes portent sur l'étude du critère utilisé pour la propagation, qui classiquement, est une mesure de corrélation classique. Nous proposons donc d'étudier cet aspect plus en détails et, en particulier, nous introduisons un nouveau critère de sélection de meilleur germe et de propagation. Plus précisément, nous nous appuyons sur un critère de confiance qui comporte trois aspects différents :

1. **La distance au germe** – Nous supposons que plus le point étudié est proche d'un germe, plus la fiabilité de la correspondance trouvée est importante. Ainsi, en suivant une loi normale, nous calculons la probabilité suivante :

$$P_s(\mathcal{C}_g^{l,l'}, \mathcal{C}_p^{j,j'}) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(|j'-l'|)^2}{2\sigma_s^2}} \tag{6.9}$$

   avec $\mathcal{C}_p^{j,j'}$ une correspondance étudiée et $\mathcal{C}_g^{l,l'}$ la correspondance germe considérée. Une valeur faible pour σ donne plus de poids à une correspondance proche du germe alors qu'une valeur élevée permettra plus de flexibilité et doit donc empêcher la propagation de correspondances erronées.

2. **La différence colorimétrique vis à vis du germe** – Nous faisons l'hypothèse que si la couleur du point étudié est proche du germe alors il y a plus de chances que les correspondants se trouvent dans le même voisinage. Plus précisément, dans les résultats présentés par le protocole d'évaluation de Middlebury [Scharstein 02b], les approches par régions ont donné des résultats compétitifs. Ces approches s'appuient sur l'hypothèse que les zones de couleurs homogènes doivent correspondre aux mêmes surfaces d'objets. Nous souhaitons exploiter la même hypothèse tout en nous affranchissant de l'étape coûteuse de segmentation. Ainsi, nous intégrons cette hypothèse en ajoutant ce calcul, qui suit, une fois de plus, une loi normale :

$$P_c(\mathcal{C}_g^{l,l'}, \mathcal{C}_p^{j,j'}) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(\|c(\mathbf{p}_d^{i,j'}) - c(\mathbf{p}_g^{k,l})\|)^2}{2\sigma_c^2}} \tag{6.10}$$

   où $c(\mathbf{p})$ correspond à la couleur (un vecteur) du point $\mathbf{p}$. Nous avons choisi l'espace L_{ab}.

3. **Le score de corrélation avec prise en compte de l'ambiguïté** – Nous voulons réduire l'ambiguïté ou à minima évaluer l'ambiguïté du score calculé pour réduire son influence. Afin de propager au mieux les germes dans les zones de petits détails, nous avons donc choisi une fenêtre de corrélation de taille 3×3. Toutefois, plus la taille de la fenêtre de corrélation est petite, plus l'ambiguïté et l'imprécision sont importantes. Nous avons choisi la mesure GC, cf. § B.1.3, car nous avons mis en évidence qu'elle est la moins ambiguë avec une fenêtre de corrélation de petite taille [Chambon 11c]. De plus, nous évaluons la moyenne de l'ambiguïté de ce score pour toute la zone de recherche et cela nous permet d'effectuer une normalisation :

$$P_{GC}(\mathcal{C}_g^{l,l'}, \mathcal{C}_p^{j,j'}) = \frac{\mathrm{GC}(\mathbf{f}_g^{i,j}, \mathbf{f}_d^{i,j'})}{\sum_{\mathbf{p}_d^{i,j''} \in \mathrm{Z}_d^{\text{réduite}}(\mathbf{p}_g^{i,j})} \mathrm{GC}(\mathbf{f}_g^{i,j}, \mathbf{f}_d^{i,j''})} \tag{6.11}$$

   où $\mathrm{Z}_d^{\text{réduite}}(\mathbf{p}_g^{i,j})$ est la zone de recherche du point correspondant, cette zone étant réduite grâce à la propagation induite par le germe $\mathcal{C}_g^{l,l'}$.

Ainsi, la probabilité d'une correspondance, étant donné un germe dans son voisinage, est donnée par :

$$P(\mathcal{C}) = P_s(\mathcal{C}) P_c(\mathcal{C}) P_{GC}(\mathcal{C}), \tag{6.12}$$

avec $\mathcal{C} = (\mathcal{C}_g^{l,l'}, \mathcal{C}_p^{j,j'})$. Ce critère est utilisé, à la fois pour sélectionner le germe à propager et pour calculer le score de corrélation pendant la propagation. De plus, pour accepter ou non une nouvelle

correspondance, nous avons utilisé une contrainte de seuil sur le score de corrélation. En dessous d'un seuil t à déterminer, une correspondance est invalidée.

6.3.2. Résultats expérimentaux sur l'appariement par propagation de germes

Choix des éléments de test – Comme pour tous les chapitres de cette partie, nous avons choisi de tester les images présentées au § 5.5. Pour comparer les approches, nous avons choisi les critères suivants : le pourcentage d'appariements corrects, noté C, et le pourcentage de densité, noté D. De plus, nous avons comparé les approches suivantes :

- Méthode notée P_{RM} – Propagation séquentielle avec la mesure présentée au paragraphe précédent en faisaint varier $\sigma_s \in [0.75; 30]$ et $\sigma_c \in [0; 10^5]$. Les germes utilisés sont ceux s'appuyant sur les détecteurs de points d'intérêt les plus complémentaires, comme nous l'avons défini au § 2.3 et publié dans [Gales 10b].
- Méthode notée P_{corr} – Propagation séquentielle avec une mesure de corrélation classique. Nous avons choisi GC.
- Méthode notée $B.L.A$, *Basic Local algorithm* – Algorithme de mise en correspondance classique en faisant intervenir la mesure GC avec une fenêtre de taille 3×3 pour être en cohérence avec la propagation et une fenêtre 9×9 qui est la fenêtre optimale suivant les résultats obtenus dans [Chambon 11c].

Pour ces trois algorithmes testés, nous avons choisi de faire varier le seuil sur le score de corrélation en prenant $t \in [0; 1]$.

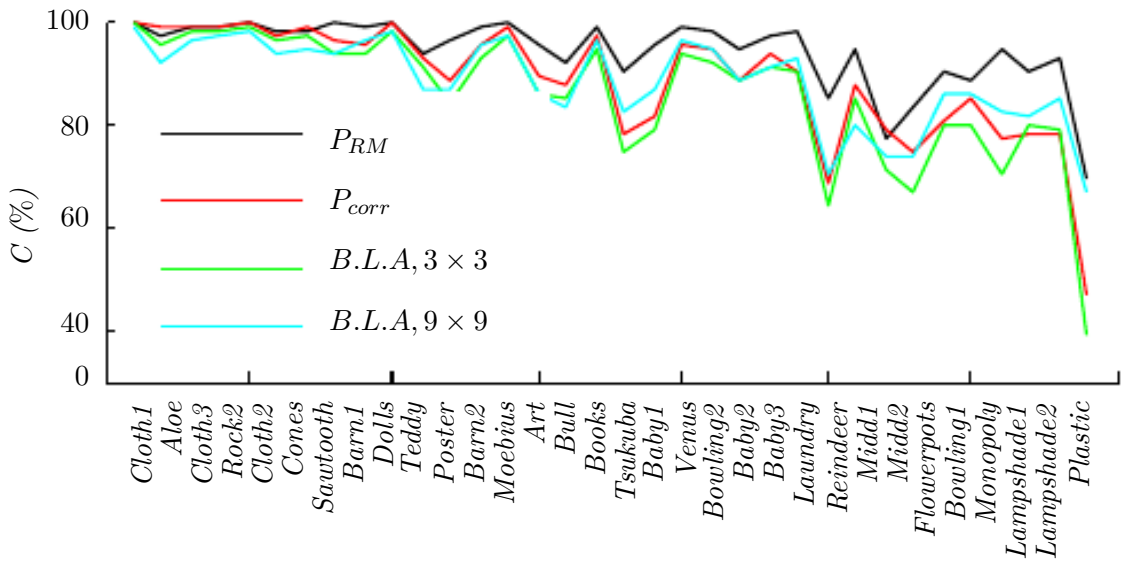


FIGURE 6.10. – Évaluation de l'approche de mise en correspondance par propagation – Nous montrons les pourcentages d'appariements corrects, pour une densité de $50\% < D < 75\%$.

Analyse des résultats de mise en correspondance par propagation de germes – Nous présentons les résultats obtenus en termes de pourcentage d'appariements corrects sur les figures 6.10 et 6.11, pour les 34 images ordonnées des images les plus texturées aux images les moins texturées. Pour chaque approche, nous affichons le résultat avec le jeu de paramètres qui maximisent ce pourcentage. Nous précisons dans la table 6.2, la configuration qui permet les meilleurs résultats sur la plupart des images, pour chaque approche étudiée. Lorsque nous visons une densité $50\% \leq D < 75\%$, P_{RM} permet d'obtenir les meilleurs résultats pour la plupart des paires étudiées, excepté les images 1, 3,

5 et 7 qui sont bien texturées et présentent peu de difficultés. Avec une densité telle que $D \geq 75\%$, P_{RM} permet d'obtenir les meilleurs résultats, en dehors des images bien texturées 2 et 7. Dans les deux cas de densité, l'approche P_{corr} apporte de meilleurs résultats que l'approche par corrélation classique, $B.L.A.$. Comme attendu, nous notons que l'approche $B.L.A._{3\times3}$ se comporte mieux que $B.L.A._{9\times9}$ dans les cas d'images bien texturées et *vice versa*. La figure 6.12 permet de présenter des cartes de disparités obtenues avec les différentes approches testées et pour une densité $D \geq 75$. Nous avons choisi ces exemples car il s'agit d'images difficiles, contenant de nombreuses occultations et des textures répétitives. Les résultats obtenus sont visuellement cohérents avec les résultats quantitatifs.

Méthode	Paramètres	$50\% < D < 75\%$	$D \geq 75\%$
P_{RM}	(σ_s, σ_c, t)	$(1, 750, 4.7 \times 10^{-6})$	$(1, 1000, 2.8 \times 10^{-6})$
P_{corr}	(s, t)	$(2, 0.8)$	$(2, 0.65)$
$B.L.A.$	$(3 \times 3, 9 \times 9)$	$(0.825; 0.825)$	$(0.725; 0.65)$

Table 6.2. – Choix des paramètres pour l'approche par propagation de germes.

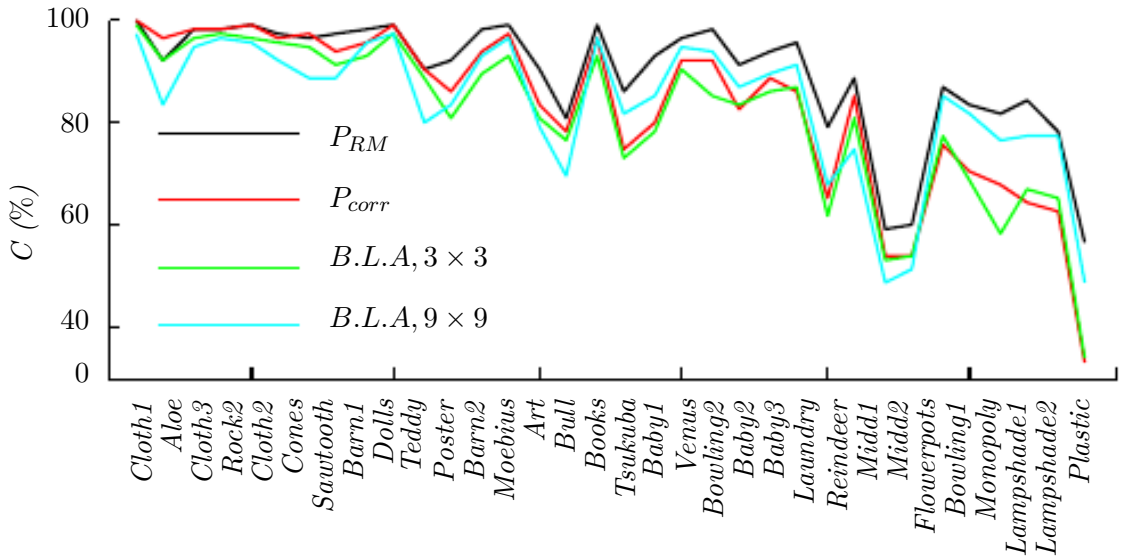


Figure 6.11. – Évaluation de l'approche de mise en correspondance par propagation – Nous montrons les pourcentages d'appariements corrects, pour une densité de $D \geq 75\%$.

6.3.3. Conclusion sur la mise en correspondance par propagation

Nous avons ainsi introduit une approche dont la contribution principale porte sur une mesure de confiance dans un processus de mise en correspondance par propagation de germes. De plus, les résultats présentés ont bénéficié des deux études précédentes pour : le choix du critère de corrélation [Chambon 11c] ainsi que le choix des détecteurs de points d'intérêt complémentaires [Gales 10b]. Ce travail a été enrichi en utilisant de multiples mesures [Gales 11, chapitre 5] et nous pourrions aller encore plus loin en reprenant les principes utilisés dans le § 6.2. Dans l'algorithme que nous avons proposé, nous avons contraint la propagation avec une mesure de confiance sur les appariements mais il paraît assez intuitif de prendre en compte une segmentation pour contraindre la propagation. C'est pourquoi nous nous sommes également intéressés à l'utilisation de cartes de segmentation en régions.

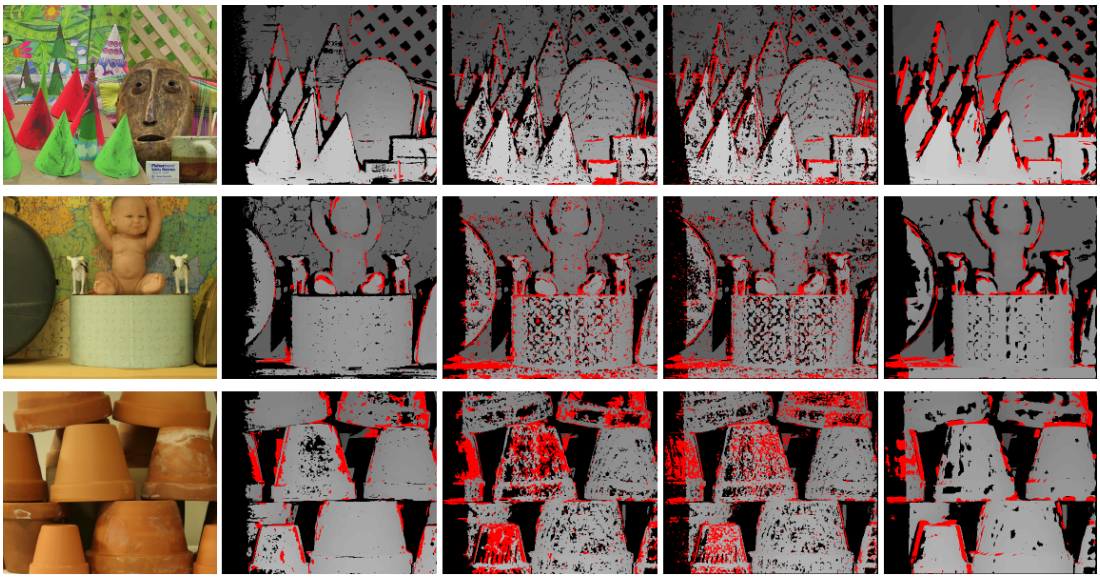


FIGURE 6.12. – Cartes de disparités obtenues par propagation de germes – Il s'agit des exemples d'images Cones, Baby3 et Flowerpots, soit respectivement les image 8, 24 et 29 dans les graphes présentés dans les figures 6.10 et 6.11. Les erreurs sont marquées en rouge. De gauche à droite, il s'agit des images gauches de des paires stéréoscopiques testées puis des résultats obtenus avec les approches : P_{RM}, P_{corr}, $B.L.A.$ avec une fenêtre de corrélation 3×3 puis 9×9.

6.4. Appariement par vote

La mise en correspondance par propagation présentée au § 6.3 donne de bons résultats quand nous cherchons un compromis entre le taux d'acceptation et la densité. Plus le taux d'acceptation est élevé, moins denses sont les résultats. L'objectif de ce chapitre est de proposer une méthode qui permette d'obtenir des taux d'acceptation élevés mais qui donne des résultats denses. Selon le protocole d'évaluation de Middlebury [11] qui compare les méthodes de mise en correspondance dense pour la stéréovision binoculaire, la plupart des méthodes qui sont bien classées sont fondées sur les régions. Leur principe a déjà été présenté au § 5.4. Ces méthodes consistent à calculer une carte de régions de l'image en faisant l'hypothèse que chaque région de couleur homogène est la projection d'une même surface de la scène. Puis, les paramètres d'un modèle de surface sont estimés pour chaque région. Enfin, une étape finale impliquant une optimisation globale est réalisée pour raffiner les résultats. Une telle approche doit permettre d'être robuste dans les zones homogènes et d'estimer une disparité dans les zones occultées. Elle a donc l'intérêt de combiner les avantages d'une approche locale (rapidité et simplicité des calculs) avec ceux d'une approche globale (gestion différente et complémentaire des difficultés liées aux occultations et aux zones homogènes). Toutefois, les disparités calculées dépendent du modèle de surface choisi, le plus courant étant le modèle plan. Ce choix ne permet pas toujours une bonne approximation, notamment si un plan est choisi pour représenter une surface conique. De plus, l'étape d'optimisation est délicate car le coût utilisé peut être relativement complexe (prise en compte de nombreuses contraintes pour modéliser au mieux le problème de la mise en correspondance). C'est pourquoi, nous proposons une méthode s'appuyant sur les régions dont le principe repose sur un processus aléatoire et itératif. Pour chaque pixel, à partir de plusieurs segmentations, d'une carte initiale de disparités et d'un modèle de surface, nous calculons différentes approximations de sa valeur de disparité. Chaque approximation constitue un vote et la valeur qui obtient le plus de suffrages est sélectionnée comme disparité finale. Nous faisons l'hypothèse que chaque région de couleur homogène

11. `http://vision.middlebury.edu/stereo/`

est la projection d'une même surface et donc que les pixels d'une même région ont des disparités proches. Par cette approche, nous souhaitons donc obtenir les performances d'une approche globale tout en évitant l'étape d'optimisation globale. Nous présentons donc la méthode introduite au § 6.4.1, ainsi que les résultats obtenus, cf. § 6.4.2.

6.4.1. Algorithme proposé pour la mise en correspondance par vote

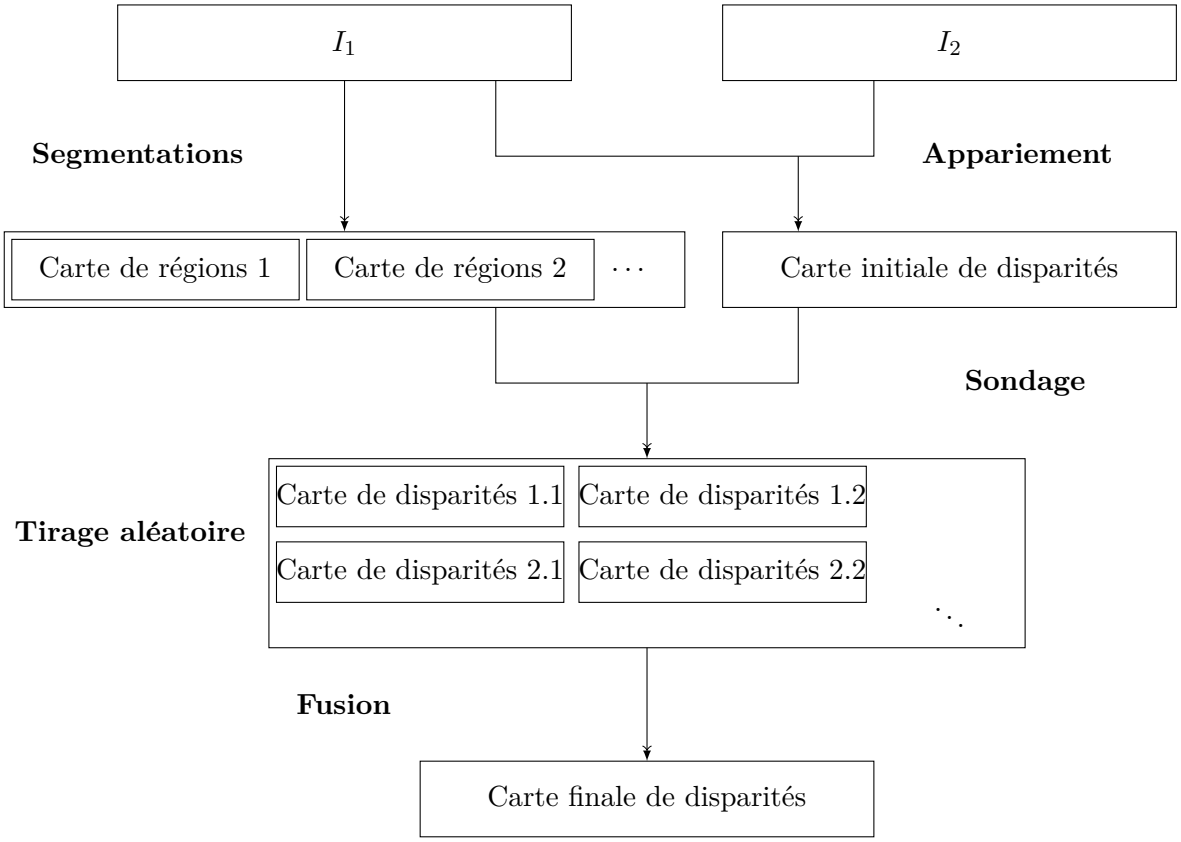


FIGURE 6.13. – Étapes de la mise en correspondance par sondage de régions.

L'approche proposée est résumée dans la figure 6.13. Les différentes étapes sont :

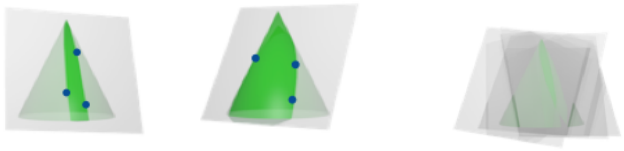
- **Calcul de différentes segmentations** –Les méthodes fondées sur les régions utilisent généralement une sur-segmentation, c'est-à-dire de petites régions. En effet, plus les régions sont petites, plus elles peuvent être approchées par un plan sans commettre une grande erreur. Cependant, plus les régions sont petites, plus il est difficile d'obtenir des disparités initiales correctes dans les zones proches des occultations et des discontinuités de profondeur. Cela pose un problème pour l'approche proposée qui s'appuie sur ces disparités. Plus une région est petite, plus le calcul de la disparité a des chances d'être précis si nous supposons une disparité initiale bien estimée. Plus une région est grande et plus nous augmentons les chances de compléter des zones de difficultés, comme des zones faiblement texturées ou proches d'une occultation. En revanche, nous augmentons aussi le risque de regrouper des zones de disparités différentes. Le fait de combiner ces différentes segmentations doit nous permettre de limiter les inconvénients de chacune, c'est-à-dire les erreurs d'appariement induites par chacune.
- **Estimation d'un appariement initial** – Il s'agit d'une première carte de disparités estimée sans utiliser les régions. Cela permet de calculer les paramètres des différents plans pour chaque région. Cette carte doit *a priori* présenter un bon compromis entre densité et taux d'acceptation. La densité élevée permet d'avoir davantage de disparités à sonder pour obtenir un nombre plus

important de votes. Le taux d'acceptation élevé permet de réduire le nombre de votes menant à des résultats erronés.

- **Calcul des disparités induites par chaque plan tiré aléatoirement pour chaque région** – Nous simplifions $d(\mathbf{p}^{i,j})$ par d. Pour chaque région, de chaque niveau de segmentation, un triplet de points $\left\{[i_1 \quad j_1 \quad d_1]^\top, [i_2 \quad j_2 \quad d_2]^\top, [i_3 \quad j_3 \quad d_3]^\top\right\}$ est tiré au sort dans l'espace des disparités initiales (les pixels sans disparité initiale ne sont pas pris en compte dans le tirage au sort). Ainsi, les paramètres $[p_1 \quad p_2 \quad p_3]^\top$ du plan sont donnés par :

$$\begin{bmatrix} i_1 & j_1 & 1 \\ i_2 & j_2 & 1 \\ i_3 & j_3 & 1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}. \tag{6.13}$$

Ce principe est illustré sur la figure 6.14 pour l'exemple d'une région représentant un cône. À chaque tirage, trois points marqués en bleu sont tirés au sort et le plan passant par ces trois points permet de calculer une disparité approchée pour chaque pixel de la région.



Premier tirage Deuxième tirage Tous les plans induits

FIGURE 6.14. – Principe du sondage de régions pour la mise en correspondance par vote – Nous montrons les plans induits (en gris) par tous les triplets (des exemples, en bleu, sur les deux premières illustrations) tirés au sort pour une région représentant un cône.

- **Attribution d'une disparité finale par sondage** – Ensuite, chaque pixel de chaque région reçoit un vote pour la disparité donnée par :

$$d_n = p_1 k + p_2 l + p_3 \tag{6.14}$$

où d_n est la valeur de disparité du $n^{\text{ième}}$ vote pour le pixel de coordonnées (k, l). Le processus est répété m fois pour obtenir m votes. Le tirage au sort peut être réalisé de façon :
  - *Uniforme* – Tous les triplets ont la même probabilité d'être tirés au sort.
  - *Biaisée* – Une valeur de confiance est calculée pour chaque triplet. Cette valeur détermine la probabilité d'acceptation de chaque point. Cela permet de limiter le nombre de sélections de triplets contenant un point dont la disparité initiale est erronée. La mesure de confiance, notée C, de chaque pixel dépend de la similarité de chaque pixel et de ses voisins avec leurs correspondants estimés :

$$\begin{aligned} C &: \mathbb{N}^2 \to \mathbb{R} \\ C(\mathbf{p}^{i,j}) &\mapsto \left(1 - \frac{\displaystyle\sum_{k=-r}^{r} \sum_{l=-r}^{r} \left| I_g^{i+k,j+l} - I_d^{i+k,j+l+d(\mathbf{p}^{i+k,j+l})} \right|}{z} \right)^p \end{aligned} \tag{6.15}$$

où r détermine les dimensions du voisinage considéré, z est un terme de normalisation et p est un paramètre de pondération. Plus p est élevé, plus le tirage est biaisé en faveur des pixels qui ont les valeurs de confiance les plus élevées. Néanmoins, si p est trop élevé, nous

risquons de toujours tirer au sort les mêmes triplets. De plus, nous proposons pour chaque région de ne conserver pour le tirage au sort que les $h\%$ de pixels les plus fiables selon la mesure de confiance (les choix de n, p et h seront discutés à la page 104).

- **Fusion** – Pour chaque pixel, la valeur de disparité qui obtient le plus de votes lui est attribuée en affinant cette valeur au sous-pixel suivant la technique de [Chen 02].

6.4.2. Résultats de la mise en correspondance par sondage de régions

Protocole expérimental – Les images testées sont les mêmes que celles présentées au § 5.5. Nous avons considéré des **zones difficiles**, notées ZD. Ces zones difficiles sont déterminées à partir de la vérité terrain en considérant qu'un pixel appartient à une zone difficile s'il n'est pas occulté et si, dans son voisinage, il existe une discontinuité de profondeur ou un pixel occulté. Les dimensions de ce voisinage sont données par les dimensions de la fenêtre de corrélation utilisée. La figure 6.15 montre en blanc les zones difficiles de l'image *Cones* pour une fenêtre de taille 11×11. Pour évaluer les résultats du sondage, nous utilisons les critères suivants, déjà définis au § 5.5.3 : la densité du résultat, D, le pourcentage d'appariements erronés, Er, le taux de rappel, T_{ra} avec $t = 0.5$ et le taux d'acceptation, T_a, avec $t = 1$.

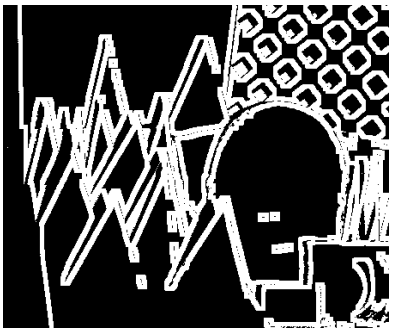


FIGURE 6.15. – Carte des zones difficiles à apparier (en blanc) pour une image de la paire *Cones* pour une taille de fenêtre de corrélation 11×11.

Nous avons besoin de choisir un certain nombre de paramètres liés à la segmentation *mean-shift*, cf. § C.3.2. Nous fixons les différentes segmentations utilisées au nombre de 4. Nous choisissons une segmentation rudimentaire ($h_s = 2, h_r = 12$), c'est-à-dire que nous supposons que les régions principales de la scène sont déterminées, 2 segmentations intermédiaires ($h_s = 12, h_r = 6$ et $h_s = 8, h_r = 4$), où nous espérons obtenir les principaux objets, et une segmentation des détails ($h_s = 12, h_r = 2$). Ces segmentations ont donné de bons résultats sur les images testées lors de nos tests préliminaires. En outre, nous avons pu constater que le choix des différentes segmentations ne fait pas varier les résultats de manière significative. Toutefois, si des erreurs importantes de segmentation sont présentes sur une carte (en particulier pour la segmentation rudimentaire), il est impératif que celles-ci ne réapparaissent pas sur les autres segmentations. Nous testons la méthode avec une carte initiale calculée par mise en correspondance par corrélation avec CENSUS, cf. § B.1.4, en ajoutant une contrainte de seuil sur le score de corrélation (fixé à 0.7). Pour le sondage, nous testons un tirage au sort uniforme et un tirage au sort biaisé. Nous prenons $r = 1$, $p = 10$ et nous testons $h \in \{35\%, 50\%, 75\%\}$. Nous faisons également varier le nombre de tirages pour chaque région $n \in \{5, 25, 75\}$. La figure 6.16 montre la carte de confiance pour le couple *Cones* où, plus le niveau de gris d'un pixel est clair, plus il a de chances d'être tiré au sort. Cette figure présente également les histogrammes des valeurs de confiance pour $p = 1$ et $p = 10$. Avec $p = 1$, la plupart des pixels ont une valeur de confiance très élevée et donc quasiment la même probabilité d'être tirés au sort. Avec $p = 10$, les valeurs de confiance sont beaucoup mieux réparties dans l'intervalle $[0; 1]$.

Analyse des résultats obtenus – La table 6.3 présente les 10 meilleurs résultats moyens obtenus sur les images testées. Plus le nombre de tirages par région est important, meilleurs sont les résultats. Nous observons que de bons résultats peuvent être obtenus à partir de 25 tirages. Nous constatons également que le tirage au sort biaisé permet d'obtenir les meilleurs résultats. Les meilleurs taux d'acceptation obtenus sont de l'ordre de 94% sur toute l'image et de 90% dans les zones difficiles. En 2011, la méthode proposée a été comparée avec celles classées sur le site de Middlebury et elle a obtenu des résultats comparables. En effet, la méthode par sondage se positionnait dans le premier quart du classement lorsque les résultats sont évalués au sous-pixel ($t \leq 0.75$) et dans la première moitié dans les autres cas. De plus, sur la figure 6.17, nous montrons les meilleurs résultats où les disparités erronées sont marquées en rouge pour différentes valeurs du seuil t.

n	p	h	T_a	T_{aZD}
75	10	35%	94.3 ± 3.3	90.1 ± 5.2
25	10	50%	94.2 ± 3.4	90.0 ± 5.5
75	10	50%	94.2 ± 3.4	89.9 ± 5.3
25	10	35%	94.2 ± 3.4	89.9 ± 5.2
75	10	75%	93.9 ± 3.6	89.4 ± 5.5
25	10	75%	93.8 ± 3.6	89.2 ± 5.6

TABLE 6.3. – Résultats de la mise en correspondance par sondage de régions – Tri par valeur de taux d'acceptation T_a. Nous affichons également le taux d'acceptation dans les zones difficiles, T_{aZD}.

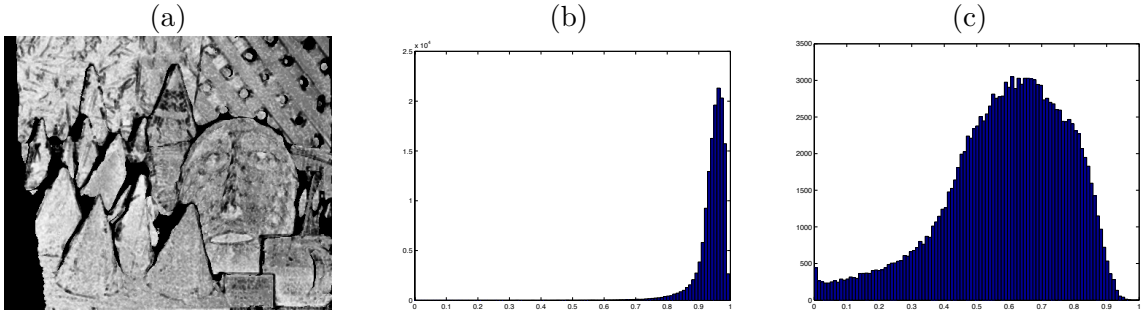


FIGURE 6.16. – Visualisation des valeurs de confiance pour l'approche d'appariement par vote – Carte (a) et histogrammes des valeurs de confiance pour $p = 1$ (b) et $p = 10$ (c).

6.4.3. Conclusion sur la mise en correspondance par sondage des régions

Dans cette section, nous avons proposé une méthode permettant d'obtenir des cartes de disparités denses en s'appuyant sur différents niveaux de segmentation en régions de couleur homogène. Une carte initiale de disparités calculées avec une méthode locale est sondée pour chaque région par tirage au sort et, pour chaque pixel, la valeur de disparité qui a reçu le plus de votes est attribuée. Nous pensons que la mise en correspondance par sondage de régions constitue une bonne stratégie de mise en correspondance dense en stéréovision binoculaire. Nous avons montré qu'en moyenne, les meilleurs résultats sont obtenus à partir de la carte de disparités calculées par la mesure de corrélation CENSUS.

Cependant, un inconvénient de cette méthode, est le tirage aléatoire qui fait que nous ne pouvons pas obtenir deux fois exactement le même résultat en lançant deux fois l'algorithme. Toutefois, bien que les tirages aléatoires ne garantissent pas l'obtention des mêmes résultats quand l'algorithme est relancé plusieurs fois, lorsque nous utilisons un nombre de tirages supérieur à 75, nous obtenons des résultats très proches. Il existe des approches s'appuyant sur une stratégie de vote [Bouchafa 06] différente, et une perspective serait d'étudier d'autres stratégies pour les adapter à nos données d'entrée.

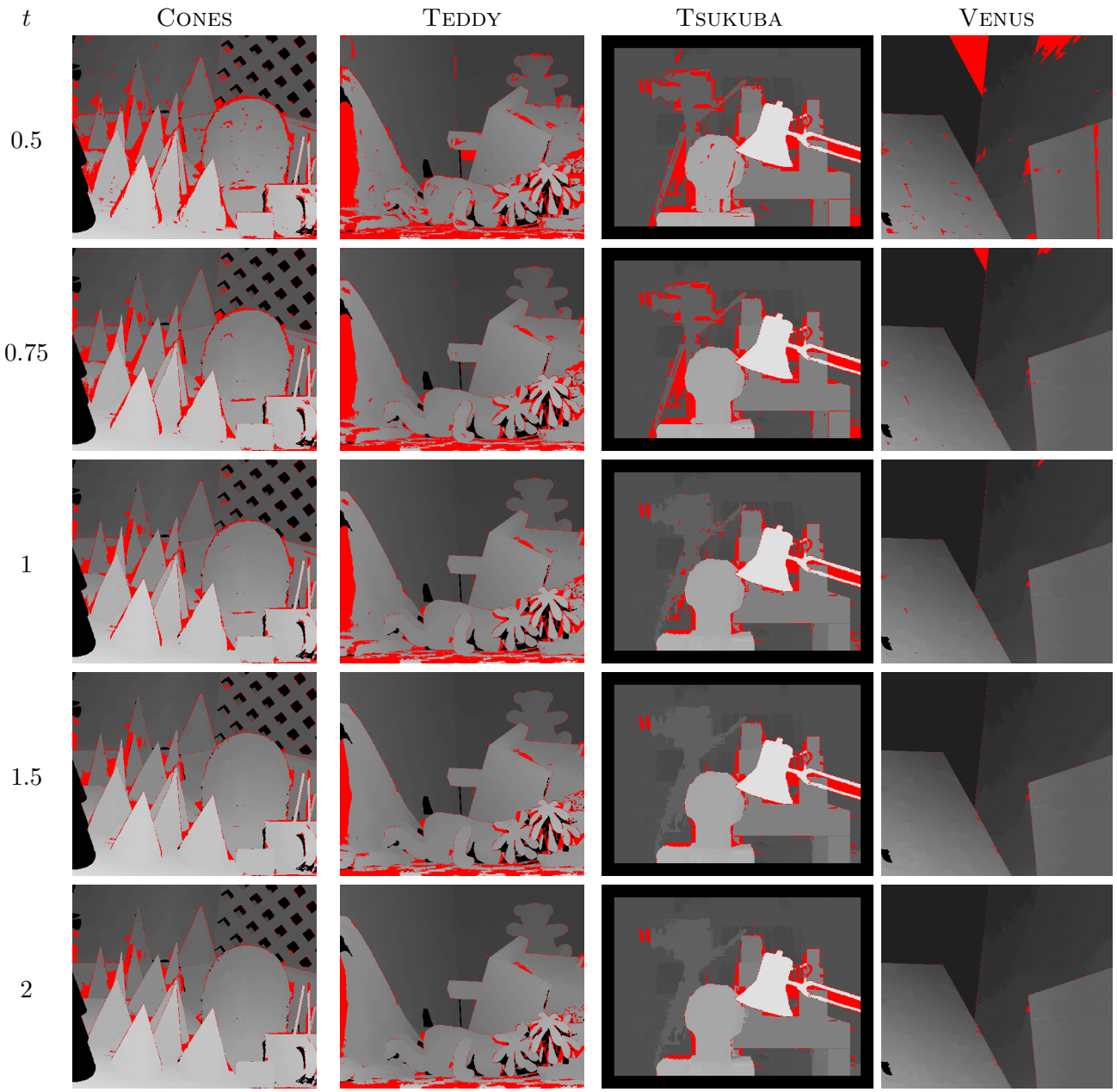


Figure 6.17. – Cartes de disparités obtenues par sondage des régions – En rouge, il s'agit des disparités erronées pour différentes valeurs de t, seuil sur l'erreur acceptée pour la disparité.

6.5. Conclusion sur les contributions en mise en correspondance stéréoscopique

Nous avons présenté dans ce chapitre différentes techniques de mise en correspondance de pixels pour la stéréovision binoculaire. Après avoir approfondi notre étude sur l'intérêt d'utiliser la couleur et la complémentarité des mesures de corrélation, nous avons traité plus particulièrement deux approches : la mise en correspondance par propagation de germes et la mise en correspondance fondée sur les régions. La première est une alternative à la mise en correspondance locale « classique ». Elle s'appuie sur un ensemble d'appariements fiables, les germes, qui permet de réduire certaines ambiguïtés. Mais ces germes doivent être sélectionnés avec soin pour éviter de propager des erreurs. Bien que de nombreux travaux traitent du problème de la mise en correspondance fiable (en particulier pour les points d'intérêt), peu d'entre eux s'intéressent à ce problème pour la sélection de germes dans le cadre de la mise en correspondance pour la stéréovision binoculaire. C'est la raison pour laquelle nous avons proposé une étude détaillée de cet aspect. La seconde approche permet d'obtenir les meilleurs résultats, parmi toutes nos contributions, selon le protocole d'évaluation de Middlebury [12].

Au final, voici les éléments que nous avons envisagés pour une utilisation conjointe [13] de ces deux approches [Gales 11, chapitre 6] :

- *Calcul de la carte initiale avec l'approche par propagation de germes* – Pour la sélection de germes, le détecteur FAST, associé aux mesures CENSUS ou GC, est celui qui a donné les meilleurs résultats pour la sélection des germes. La meilleure stratégie consiste à détecter et mettre en correspondance les points d'intérêt de telle sorte que les germes obtenus soient très fiables, même si leur répartition n'est pas maximale. Il est alors possible d'améliorer cet ensemble de germes, en le fusionnant à d'autres ensembles issus d'autres couples détecteur-corrélation complémentaires. Pour pouvoir être utilisée comme carte initiale pour la méthode de sondage de régions, la carte obtenue doit présenter un bon compromis entre la densité et le pourcentage d'appariements acceptés.
- *Mise en correspondance dense par sondage de régions* – Enfin, la mise en correspondance dense peut être obtenue par sondage de régions en biaisant le tirage au sort des disparités initiales afin de favoriser les pixels qui ont les disparités initiales les plus fiables.

Par la suite, nous avons continué à explorer des approches faisant intervenir une segmentation, notamment en utilisant la notion de superpixels, cf. 8.3 qui introduit cette notion. Les travaux préliminaires que nous avons réalisés dans le cadre de la mise en correspondance s'appuyant sur l'utilisation des superpixels, ont mis en évidence qu'il convient de mesurer l'impact de la taille des superpixels, ainsi que la régularité de ces superpixels. Il est également important d'évaluer l'influence de différentes approches de superpixels pour compléter les travaux initiés par [Xie 10]. Cela fait partie des travaux futurs que nous développons à la section 13.7 de la partie IV sur les perspectives.

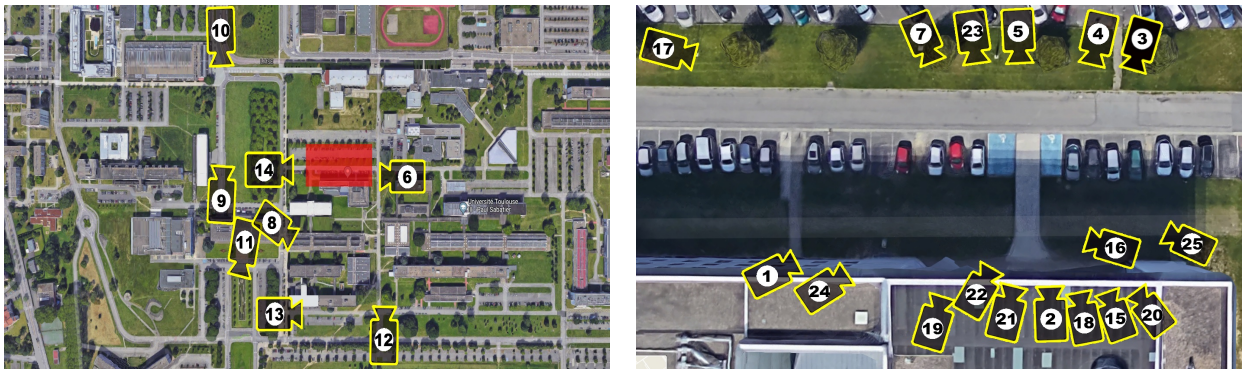
Nous avons jusqu'ici traité des couples d'images stéréoscopiques. Les travaux suivants permettent la prise en compte de vidéos multiples.

12. `http://vision.middlebury.edu/stereo`
13. Cette partie n'a pas été directement développée dans ce manuscrit.

Chapitre 7.

Analyse de scènes par vidéos multiples avec recouvrement

7.1. Introduction sur l'utilisation de vidéos multiples avec recouvrement



(a) (b)

FIGURE 7.1. – Illustration des dispositifs de caméras que nous souhaitons étudier – Voici un dispositif de caméras que nous avons mis en place et nous avons publié les vidéos extraites dans [Malon 18]. Il s'agit d'un ensemble de 25 vidéos synchronisées avec un scénario et une annotation manuelle par boîtes englobantes. Nous avons nommé ce jeu de données *Toulouse Campus Dataset (ToCaDa)*. Nous montrons, en (a), la vue globale du dispositif et, en (b), un zoom de la vue globale au niveau du rectangle rouge. La position de chaque caméra est présentée : nous pouvons remarquer que certaines caméras ont peu (comme 14 et 6) ou pas (comme 8, 9, 11, 12, 13) de recouvrement avec les autres. En revanche, certains groupes de caméras possèdent une large zone commune : toutes celles présentées dans la vue (b).

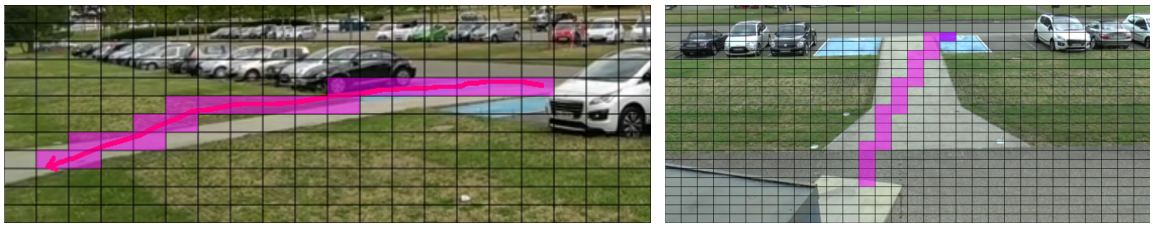
Certains des travaux présentés dans ce manuscrit traitent, dans la partie I, de l'utilisation d'images multimodales, cf. chapitre 3, ou de combinaison d'images et de cartes de profondeur, cf. chapitre 4. Au début de cette partie II, nous avons introduit des traitements sur des couples stéréoscopiques, cf. chapitre 6. De plus, nous avons introduit, au début de ce manuscrit, le fait qu'il y a de plus en plus de données multimédia disponibles pour effectuer de la reconnaissance et de l'apprentissage. C'est également le cas des vidéos, et, en particulier, dans le cas d'un évènement particulier : compétition, phénomènes climatiques, manifestations ou attentats, un très grand nombre de vidéos vont être acquises, par des systèmes de surveillance ou plus simplement par des outils d'acquisition grand public, comme les caméras de smartphone. Dans le cas d'un évènement important, voire d'un évènement lié



(a) (b)

FIGURE 7.2. – Illustration des ensembles de vidéos que nous souhaitons étudier – Toutes les vues présentées sont au même instant dans toutes les vidéos acquises. Les vues présentées en (a) correspondent à des vidéos qui possèdent un recouvrement entre elles et elles sont acquises par l'ensemble des caméras visualisées dans la figure 7.1.(b). La voiture entourée en rouge est visible dans toutes les caméras. En revanche, en (b), nous présentons un exemple de la vue au même instant pour toutes les caméras sans recouvrement.

à une enquête, cette source d'information est capitale mais elle est également difficile à exploiter. C'est la raison pour laquelle, dans nos travaux actuels [Malon 19b], nous nous intéressons aux vidéos avec recouvrement. Plus précisément, nous cherchons à recouper les informations contenues dans les différentes vidéos afin d'identifier les groupes de vidéos qui observent une même partie de la scène, cf. la figure 7.1 montrant le type de dispositif que nous souhaitons exploiter et la figure 7.2 qui illustre le type de données que nous souhaitons manipuler.



(a) (b)

FIGURE 7.3. – Première problématique abordée dans le cadre de l'étude de réseaux de caméras – En (a), nous montrons une trajectoire tracée par un utilisateur. C'est ce que nous appelons une trajectoire requête. À partir de cette trajectoire, nous proposons, en (b), de la retrouver, ou plus précisément, de la reformuler dans une autre vidéo permettant de visualiser la même scène depuis un autre point de vue.

Déterminer manuellement les groupes de vidéos qui observent une même partie de la scène est une tâche très coûteuse en temps, et cela peut avoir un impact considérable dans un contexte applicatif donné, en particulier dans le cas d'une enquête policière. Nous avons donc besoin d'outils les plus automatiques possible pour identifier les vidéos en correspondance ainsi que les éléments saillants à étudier dans les vidéos. Nous souhaitons que ces outils répondent aux questions suivantes :

(1) Quelles sont les vidéos qui permettent de visualiser les mêmes éléments mais avec des angles de vue différents ?

(2) Quelle vidéo offre la meilleure vue d'un élément étudié ?

Plus précisément, nous avons commencé à répondre aux problématiques suivantes, illustrées dans les figures 7.3 et 7.4 :

(1) Quelles vidéos d'une collection permettent de mieux visualiser une trajectoire requête, cf. figure 7.3.(a) ? Avec les deux sous-problèmes suivants :
   (a) Nous voulons obtenir la trajectoire reformulée pour chaque vidéo, cf. figure 7.3 où nous pouvons observer la trajectoire fournie en (a) reformulée dans une autre caméra en (b).
   (b) Nous souhaitons obtenir la liste des vidéos triées par ordre décroissant de pertinence.

(2) De manière générale, comment construire un graphe de relations, cf. figure 7.4 ? C'est-à-dire, connaissant un ensemble de vidéos, cf. figure 7.4.(a), est-il possible de déterminer celles qui permettent de visualiser une même partie de la scène (ce sont les vidéos avec un lien dans la figure 7.4.(b)) ? De plus, comment les caméras sont-elles orientées les unes par rapport aux autres ?

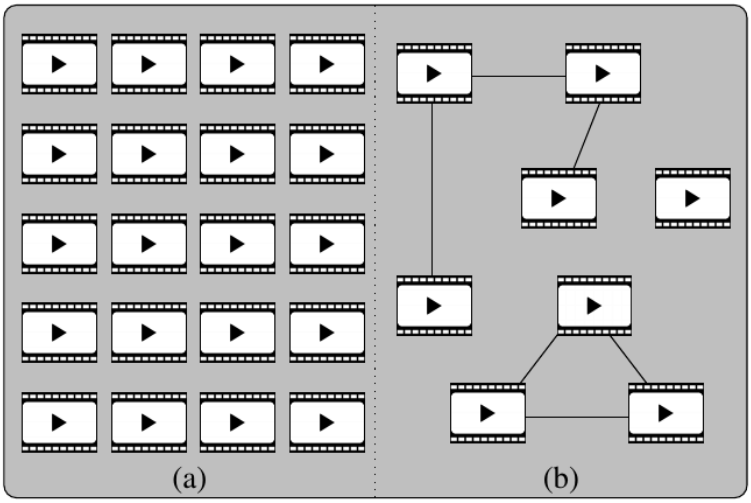


FIGURE 7.4. – Deuxième problématique abordée dans le cadre de l'étude de réseaux de caméras – Nous souhaitons obtenir à partir de l'ensemble de vidéos, en (a), un graphe de liens entre vidéos, comme celui présenté en (b).

Dans ce contexte, nous pourrions envisager l'utilisation de nombreux outils de la vision, comme par exemple : des outils de reconstruction 3D ou des méthodes de ré-identification. Nous faisons l'hypothèse de points de vue très différents, ainsi que de systèmes d'acquisition variables et de qualité plus ou moins bonne. Toutes ces techniques semblent donc utiles mais peu exploitables dans ce contexte très peu contraint. Dans ces travaux, nous exploitons donc surtout l'intuition suivante : si nous observons une activité similaire simultanément dans deux vidéos, il est probable qu'elles observent la même scène. Nous supposons donc, au préalable, que :

(1) Les vidéos manipulées sont synchronisées.

(2) Les objets en mouvement ont été préalablement détectés. Cette détection est réalisée image par image, sous la forme de boîtes englobantes, et nous n'avons pas de notion de suivi. Cela nous semble être un résultat que nous pouvons obtenir avec des outils de la littérature, comme avec l'outil YOLO, *You Only Look Once* [Redmon 18], ou *Mask-RCNN, Regional Convolutional Neural Network* [He 20].

Ainsi, plus formellement, voici le principe général utilisé :

> Deux régions extraites de deux caméras différentes systématiquement et simultanément occupées ou non occupées sont des régions en correspondance.

Dans ce domaine de recherche, notre première difficulté a été d'obtenir des données afin de valider les concepts proposés. Dans un premier temps, la littérature étant assez faible sur cet aspect, nous avons constitué notre propre base de données [Malon 18], cf. figure 7.1 et 7.2. L'originalité de ce jeu de données réside dans le fait de proposer aussi bien des annotations visuelles que des annotations auditives [Guyot 19].

7.2. Reformulation de trajectoire requête

7.2.1. Algorithme de reformulation de trajectoire requête

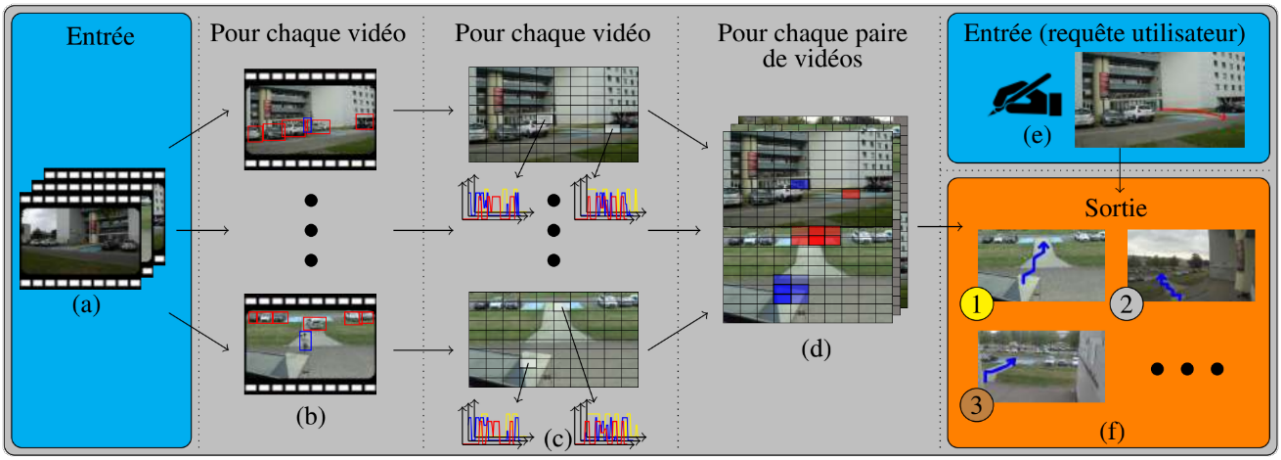


FIGURE 7.5. – Algorithme de reformulation de trajectoires – À partir des vidéos disponibles, en (a), nous supposons également que nous avons une détection image par image d'éléments dans différentes catégories (chaque catégorie est représentée par une même couleur), en (b). En découpant les vidéos en cellules selon une grille régulière, nous construisons alors les fonctions d'activité de chaque cellule pour chaque catégorie d'éléments en (c) : il s'agit du taux de recouvrement de la cellule au cours du temps (i.e. à mesure que défilent les images) par des éléments de la catégorie en question. Ces fonctions d'activité nous permettent d'établir des correspondances entre cellules de différentes images (et ainsi des différentes vidéos), cf. (d). Grâce à ces correspondances, nous sommes capables de reformuler une trajectoire exprimée par un utilisateur dans une caméra donnée, en (e), dans toutes les autres caméras, en (f). Suivant les réponses obtenues, nous effectuons en sortie, en (f), un classement des vidéos qui apportent le plus d'information par rapport à la caméra initiale.

Pour répondre à cette première problématique, nous avons introduit l'algorithme présenté dans la figure 7.5, [Malon 19b, Malon 19a]. L'originalité principale de ce travail repose sur le fait que nous ne faisons pas un suivi précis ou une ré-identification des personnes et des véhicules mais que nous utilisons de manière très simple les corrélations d'activité dans les images.

7.2.2. Corrélation de fonctions d'activité

Nous supposons que quel que soit le couple de caméras possédant un recouvrement, une partie seulement de la caméra sera concernée. Dans ce contexte, il paraît pertinent de faire une étude dans des sous-régions de l'image que nous nommerons cellules dans la suite de ce chapitre. Ainsi, si nous notons V_1 et V_2 les deux vidéos que nous souhaitons comparer, pour une catégorie donnée, notée ω, pour chaque paire de cellules $(c_i^{V_1}, c_{i'}^{V_2})$, nous utilisons un simple score de corrélation (qui est relatif aux mesures definies au chapitre 5). Plus précisément, nous évaluons une corrélation croisée centrée

normalisée, cf. équation (B.2) :

$$M_\omega\left(c_i^{V_1}, c_{i'}^{V_2}\right) = \frac{\sum_{t=1}^{T}\left(a_i^{V_1,\omega}(t) - \overline{a_i^{V_1,\omega}}\right)\left(a_{i'}^{V_2,\omega}(t) - \overline{a_{i'}^{V_2,\omega}}\right)}{\sqrt{\sum_{t=1}^{T}\left(a_i^{V_1,\omega}(t) - \overline{a_i^{V_1,\omega}}\right)^2 \sum_{t=1}^{T}\left(a_{i'}^{V_2,\omega}(t) - \overline{a_{i'}^{V_2,\omega}}\right)^2}}, \tag{7.1}$$

avec les notations suivantes :

- $a^{V_i,\omega}(t)$ la valeur de la fonction d'activité, dans la vidéo V_i, liée à la catégorie ω, à l'instant t ;
- $\overline{a^{V_i,\omega}}$ la moyenne sur tous les instants de la fonction d'activité, dans la vidéo V_i, liée à la catégorie ω.

7.2.3. Mise en correspondance de cellules dans des vidéos

Grâce à cette mesure permettant d'estimer les correspondances entre cellules de différentes vidéos, nous pouvons reformuler une trajectoire dans une autre image/vidéo en utilisant le critère suivant :

$$\underset{(i'_1,\ldots,i'_M)}{\operatorname{argmax}} \frac{\frac{1}{M}\sum_{k=1}^{M} M_\omega(c_{i_k}^{V_1}, c_{i'_k}^{V_2})}{1 + \sum_{k=1}^{M-1} \max(0, ||i'_k - i'_{k+1}|| - 1)}, \tag{7.2}$$

avec les notations suivantes :

- $(i_1, \ldots, i_M)$ la séquence de cellules requête ;
- $(i'_1, \ldots, i'_M)$ la séquence de cellules reformulée ;
- $M_\omega(c_{i_k}^{V_1}, c_{i'_k}^{V_2})$ le score de correspondance défini par (7.1).

Cette expression implique que nous allons, d'une part, avec le numérateur, favoriser les paires de cellules qui se correspondent le plus, et, d'autre part, avec le dénominateur, favoriser la séquence de cellules qui maximise le nombre de cellules adjacentes deux à deux.

7.2.4. Classement des vidéos par intérêt lié à la complémentarité des vues

Enfin, dans ce travail, afin de classer les vidéos par ordre d'intérêt par rapport à la vidéo d'origine, nous avons pris en compte la longueur de la trajectoire reformulée, car nous faisons l'hypothèse que plus la trajectoire est longue, plus elle a de chances de révéler des détails intéressants, ou plus précisément, des détails nouveaux, par rapport à la vidéo d'origine.

7.2.5. Résultats de reformulation de requête

La figure 7.6 nous permet d'illustrer le type de résultat que nous obtenons. Nous pouvons constater que les résultats obtenus sont assez satisfaisants dans le sens où nous obtenons des trajectoires reformulées correctement ainsi que des vidéos qui permettent d'obtenir un point de vue assez différent. La vidéo la moins bien classée, pour chaque cas, reste une vidéo qui contient bien un recouvrement avec la vidéo initiale étudiée. Pour aller plus loin dans notre évaluation, nous avons évalué la qualité des cartes de correspondances, des reformulations de trajectoire ainsi que la pertinence du classement des vues offrant une meilleure visualisation.

Rappel du jeu de données utilisé – Nous avons utilisé le jeu de données ToCaDa [Malon 18]. Il contient 25 vidéos dans lesquelles une trentaine d'objets de trois catégories (piéton, moto et voiture) sont présents. Parmi toutes les vues, 15 présentent de larges recouvrements et sont situées autour d'un



FIGURE 7.6. – Classement des vidéos par degré d'intérêt – La première colonne permet de présenter 3 trajectoires requête exprimées dans 3 vidéos différentes (trajectoire dessinée en rouge par l'utilisateur). Ces 3 trajectoires sont exprimées pour 3 catégories différentes : personnes, motos et voitures. Dans les colonnes 2 à 4, nous montrons les 3 premières vidéos classées alors que la dernière colonne correspond à la dernière vidéo proposée.

même bâtiment, tandis que les autres vues filment des régions plus éloignées, sans recouvrement. Les vidéos sont synchronisées et durent près de 5 minutes, ce qui paraît suffisant pour estimer les cartes de correspondances. En effet, il est préférable que les vidéos durent assez longtemps pour que les cartes de correspondances puissent être apprises : plus la durée commune des vidéos est importante, moins nous risquons de corréler les actions de régions qui ne se correspondent pas. Enfin, nous nous sommes servis des étiquettes de catégorie et des détections de boîtes englobantes fournies dans le jeu de données pour calculer les cartes de correspondances des cellules.

Évaluation de la qualité des correspondances – Nous estimons qu'une carte de correspondances est de bonne qualité si, lorsqu'à un instant donné, un même objet apparaît simultanément dans deux vidéos, les cartes de correspondances des cellules couvertes par la boîte englobante de l'objet dans la première vidéo offrent un score de correspondance élevé au niveau des cellules couvrant la boîte englobante de la deuxième vidéo. Plus formellement, nous définissons le taux de correspondance entre une paire de boîtes englobantes B_1 et B_2 correspondant à un même objet vu simultanément dans V_1 et V_2 par :

$$\sum_{(\mathbf{c_i}^{V_1},\mathbf{c_j}^{V_2})} \frac{|c_{\mathbf{i}}^{V_1} \cap B_1|}{|B_1|} \frac{|c_{\mathbf{j}}^{V_2} \cap B_2|}{|B_2|} \frac{C_\omega(c_{\mathbf{i}}^{V_1}, c_{\mathbf{j}}^{V_2})}{||C_\omega(c_{\mathbf{i}}^{V_1}, V_2)||_2} \tag{7.3}$$

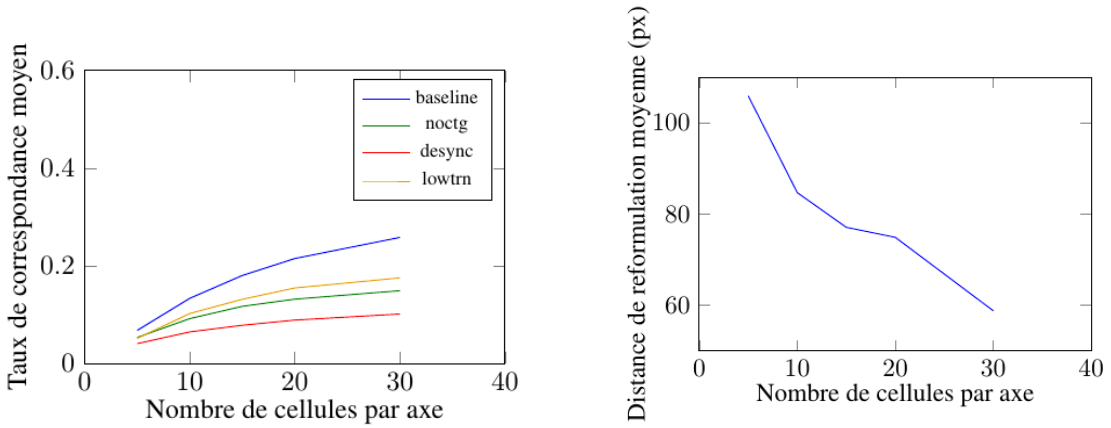
Le premier terme du produit pondère chaque cellule de la première vue par le pourcentage de la boîte englobante qu'elle couvre. Le second terme pondère de la même façon dans la deuxième vue. Le troisième terme pondère finalement par la corrélation entre paires de cellules issues des deux boîtes englobantes. Le taux de correspondance global est obtenu en moyennant l'expression (7.3) sur l'ensemble des boîtes englobantes simultanées d'un objet, puis en moyennant sur tous les objets communs à la paire de vidéos, puis, enfin, en moyennant sur toutes les paires possibles de vidéos. La figure 7.7.(a) présente les taux de correspondance pour différentes configurations et différents nombres de cellules. Nous avons évalué le comportement de la méthode :

- *Sans utilisation de la catégorie :* chaque cellule dispose alors d'une seule fonction d'activité mesurant indifféremment la présence d'objets quelle que soit leur catégorie.
- *En présence d'une désynchronisation :* nous avons crée un décalage d'une seconde entre les

paires de vidéos.

- *Avec des vidéos de durée plus courte :* nous avons utilisé des vidéos tronquées, c'est-à-dire, réduites de moitié en terme de temps.

Les résultats ont révélé que l'étape de calcul des cartes de correspondances est très sensible à ces trois types de perturbations, et en particulier, à la distinction des catégories. Ce résultat était prévisible et nous espérons simplement que ces taux de correspondance soient suffisants pour nous permettre de correctement effectuer l'étape suivante : la reformulation.



(a) (b)

FIGURE 7.7. – Évaluation de la reformulation de trajectoires – Nous présentons la distance de reformulation moyenne DTW en pixels pour différents nombres de cellules en (a) et le taux de correspondances entre vidéos en (b). Ces deux courbes sont évaluées pour la base de données ToCaDa [Malon 18]. Pour (b), nous indiquons ce taux pour différents nombres de cellules et différentes configurations : sans dégradation, *baseline*, sans distinction des catégories, *noctg*, avec une désynchronisation d'une seconde entre les vidéos, *desync* et en apprenant les cartes de correspondances sur seulement la moitié du jeu de données, *lowtrn*.

Évaluation de la qualité des reformulations – Pour mesurer la qualité de nos reformulations de trajectoire, nous avons tracé 10 trajectoires requêtes au niveau du sol dans différentes caméras et leur avons appliqué notre méthode de reformulation pour obtenir les trajectoires reformulées dans les autres caméras. Pour se comparer à une trajectoire de référence, nous avons calculé l'homographie plane entre chaque paire de caméras à partir des coins des places de parking bleues. Puis, nous avons mesuré la distance *Dynamic Time Warping* (DTW) [Zhang 06], en pixels, entre les trajectoires obtenues à partir de l'homographie et les trajectoires reformulées. La figure 7.7.(b) présente la distance DTW moyenne, en pixels, sur des vidéos de tailles 960×540 pour différents nombres de cellules. Plus le nombre de pixels augmente et plus la trajectoire reformulée est fiable. Ces résultats valident donc la méthode de reformulation de trajectoire.

Concernant le classement des vidéos, les vues qui ne présentent pas de recouvrement sont correctement filtrées en utilisant un seuil $\sigma = 0.3$. En effet, quasiment aucune correspondance ne peut être apprise du fait qu'il n'y ait pas d'objets de la même catégorie présents simultanément au cours du temps. Parmi les 5 meilleures vues obtenues sur nos 10 trajectoires requêtes, 72% d'entre elles donnent effectivement une meilleure ou aussi bonne visibilité de la trajectoire requête, cf. les exemples déjà présentés dans la figure 7.6.

7.2.6. Conclusion sur la reformulation de trajectoires

À partir d'une collection de vidéos et sans se servir des paramètres de calibrage des caméras, la méthode proposée permet de classer avec succès un sous-ensemble des vidéos qui présentent du recouvrement avec une vidéo dans laquelle une requête trajectographique est tracée en fonction de la visibilité qu'elles offrent de la trajectoire reformulée. Les vidéos qui ne sont pas liées à la scène sont également correctement filtrées. Il n'est fait usage d'aucune méthode de ré-identification, la simple présence simultanée d'objets de même catégorie dans différentes caméras permet d'apprendre les correspondances spatiales. La méthode est particulièrement adaptée à un contexte de vidéo-surveillance en extérieur du fait des heures d'enregistrement vidéo disponibles depuis divers points de vue, permettant d'obtenir des cartes de correspondances robustes. À présent, nous allons présenter la façon de construire un graphe de relations entre vidéos à partir de l'étude des correspondances.

7.3. Construction du graphe de relations entre vidéos

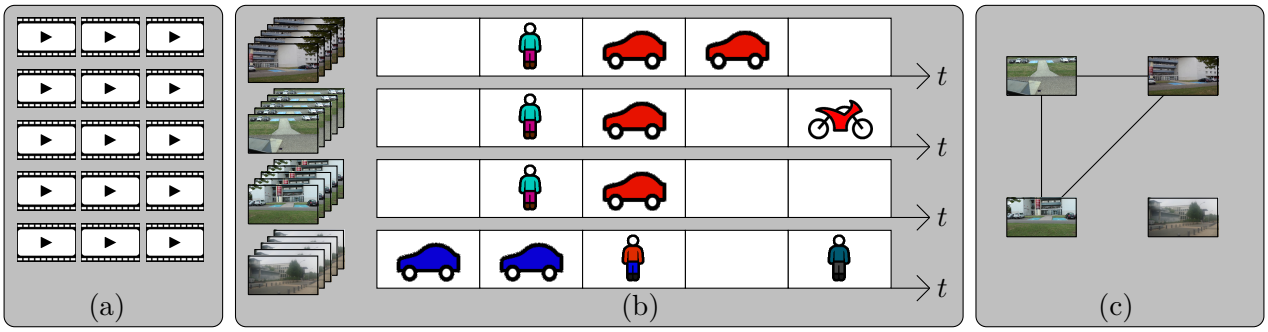


FIGURE 7.8. – Approche proposée pour la construction d'un graphe de liens dans une collection de vidéos – Pour chaque vidéo d'un ensemble de vidéos statiques (a), les objets présents sont détectés pour formuler les *histoires* des vidéos (b), puis des liens sont formés entre les vidéos aux histoires similaires pour construire un graphe de liens (c) où chaque arête indique une paire de vidéos avec des champs de vue qui se recouvrent.

Pour nous permettre de construire un graphe de relations entre vidéos, comme présenté dans la figure 7.8, nous avons introduit un concept nouveau : la notion d'**histoire**. De plus, nous avons mis en place une analyse multi-échelle temporelle et spatiale, car nous espérons ainsi rendre l'approche d'estimation des liens la plus robuste possible aux changements d'échelles et à une faible désynchronisation tout en souhaitant avoir des temps d'exécution raisonnables.

7.3.1. Définition de la notion d'histoire

Nous rappelons que chaque image de chaque vidéo est découpée en régions, appelées des cellules. Les objets sont régulièrement détectés, comme dans la reformulation de trajectoire, à chaque pas de temps $\tau_i = \tau \times i$. L'histoire d'une région correspond à une séquence temporelle de tableaux de contingence. Ces tableaux listent tous les objets détectés dans une région ainsi que leurs caractéristiques (catégorie, couleur, vitesse), cf. figure 7.9. Certains pixels peuvent être occupés par plusieurs objets en même temps (cela est relatif à la distance de l'objet à la caméra ou encore à l'imprécision de la détection). En conséquence, nous supposons qu'un élément occupe une cellule, uniquement si les points situés au niveau du sol se trouvent dans cette cellule. De manière similaire à [Khan 03], nous prenons le milieu du côté inférieur de la boîte englobante, cf. figure 7.10. Si nous notons R une région, alors l'histoire à travers les différents pas de temps fixés $\tau_0, ..., \tau_i$ est notée $S_R^{0 \to i}$. L'histoire au temps τ_j est notée S_R^j

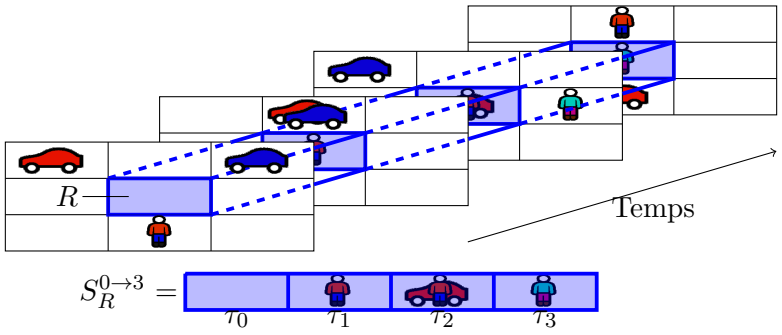


FIGURE 7.9. – Notion d'histoire pour rechercher des similarités dans des vidéos – À intervalles réguliers, les objets sont détectés dans chaque cellule de la vidéo et nous leur attribuons un vecteur de caractéristiques comme la catégorie (personne, moto et voiture) ou l'apparence (comme la couleur). L'histoire d'une région donnée, comme R dans cette figure (la case bleutée à chaque instant τ) est représentée par le vecteur comme celui de cette figure, en bas. Ainsi, pour cette cellule, nous constatons qu'il n'y a aucun objet à l'instant τ_0, puis, à partir de l'instant τ_1, une personne apparaît ... et ainsi de suite.

et l'histoire sur toute la vidéo S_R. Comme nous pouvons le voir dans la figure 7.11, nous calculons des tables de contingences à différentes échelles temporelles, ϕ_t, et spatiales, ϕ_{sp}, et cela nous permet de construire des histoires à différentes échelles spatio-temporelles en prenant les unions des tables de contingences suivant l'analyse multi-échelle souhaitée.

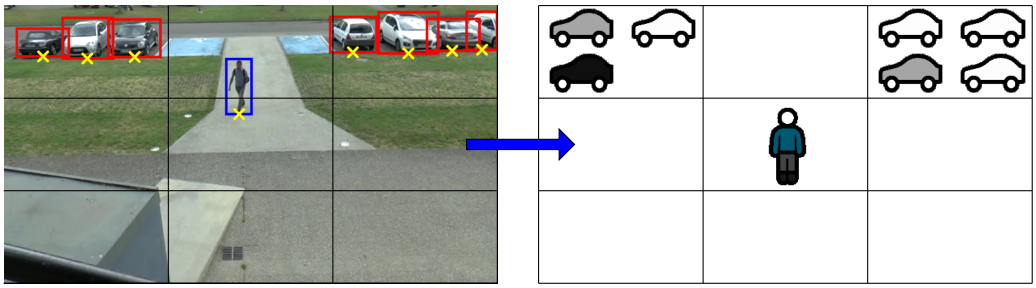


FIGURE 7.10. – Construction de l'histoire d'une cellule – À gauche, nous présentons une image extraite d'une vidéo, découpée en 9 cellules où nous pouvons visualiser les objets détectés avec leurs boîtes englobantes (rouge lorsqu'il s'agit d'une voiture, bleue lorsqu'il s'agit d'un piéton). Le point de contact au sol est matérialisé par les croix jaunes. Ainsi, à droite, nous pouvons visualiser la table de contingence qui permet de répertorier tous les éléments par cellule. La région assignée est celle où se trouve la croix jaune.

7.3.2. Notion de distance entre histoires

Afin de définir la distance entre deux histoires de même longueur S_R et $S_{R'}$, nous introduisons d'abord la notion d'objet commun aux deux histoires. Un objet O est commun aux deux histoires s'il apparaît dans la région R de la première vidéo au temps τ_i et si un objet O' de même catégorie que O et d'apparence similaire apparaît dans la région R' ou dans son voisinage spatial (par exemple, dans son 8-voisinage), au temps τ_i ou dans son voisinage temporel $\{\tau_{i-k}, ..., \tau_i, ..., \tau_{i+k}\}$. La similarité d'apparence est établie si la distance entre les descripteurs d'apparence est inférieure au seuil d'apparence σ_{app}. Nous pouvons alors dire que O a trouvé un correspondant dans R'.

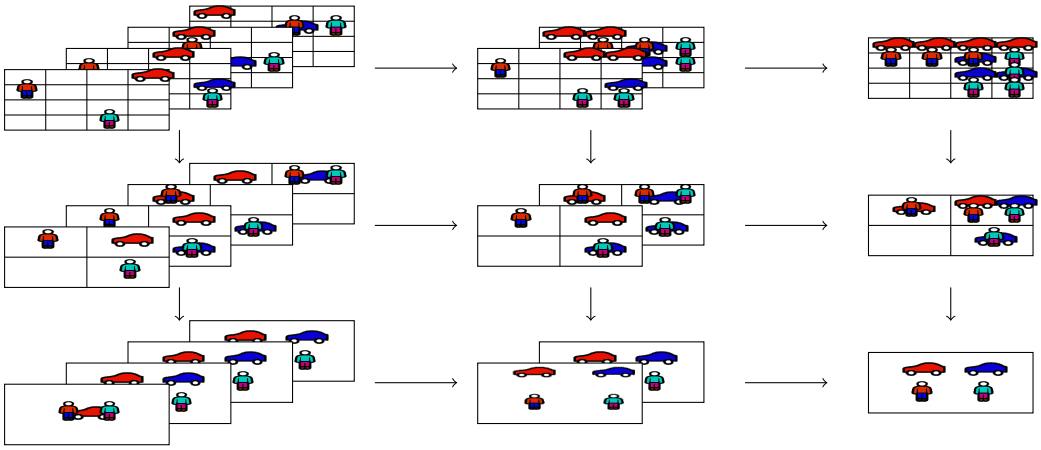


FIGURE 7.11. – Analyse multi-échelle pour construire un graphe des liens dans une collection de vidéos – Nous illustrons ici l'intérêt de l'analyse multi-échelle avec un exemple. De la gauche vers la droite, l'échelle temporelle est réduite en prenant l'union des histoires successives par pas de temps, deux à deux, alors que de bas en haut, c'est la résolution spatiale qui est réduite en prenant l'union des cellules adjacentes. Ainsi, l'histoire en haut à gauche est la plus détaillée alors que celle en bas à droite est la plus simplifiée.

Aucune ré-identification d'objets par paire n'est effectuée : plusieurs objets différents de R peuvent trouver un même correspondant dans R'. C'est en cela que notre approche cherche à repérer une cohérence globale entre objets détectés au cours du temps sans chercher à ré-identifier précisément les objets. Notons $C(S_R, S_{R'})$ la proportion d'objets apparaissant dans R qui ont trouvé un correspondant dans $R^{'}$ et $|S_R|$, respectivement $|S'_R|$, le nombre d'objets apparaissant dans l'histoire S_R, respectivement S'_R. Nous définissons la distance $d(S_R, S_{R'})$ entre deux histoires par :

$$d(S_R, S_{R'}) = 1 - \frac{C(S_R, S_{R'}) + C(S_{R'}, S_R)}{|S_R| + |S_{R'}|}. \tag{7.4}$$

Cette distance peut s'interpréter comme le pourcentage d'objets issus des deux histoires confondues qui ne trouvent pas de correspondant. Il est important de noter que nous avons pris en compte uniquement les objets mobiles dans la scène, car, avec cette notion d'histoires, les objets statiques peuvent être une source d'erreur pour associer des vidéos. Par exemple, un ensemble de voitures garées ressemblent fortement à n'importe quel autre ensemble de voitures garées [1]. Dans la section suivante, nous allons voir comment cette distance est utilisée pour construire le graphe de relations.

7.3.3. Estimation du graphe de relations

Nous présentons l'algorithme proposé pour trouver les liens dans les vidéos dans l'algorithme 2. Il s'appuie sur l'étude de toutes les paires de vidéos possibles. Les histoires sont ainsi comparées de l'échelle la plus grossière (celle du coin inférieur droit dans la figure 7.11) à l'échelle la plus fine (le coin supérieur gauche dans la même figure), si nécessaire. Plus précisément, nous commençons par comparer les histoires à l'échelle de toute l'image, soit $s = 1$, et avec le pas temporel le plus important. Si la dissimilarité est au delà d'un seuil haut, σ^s_{reject}, qui dépend de l'échelle s, nous sommes assez certains que les vidéos n'ont rien en commun. En revanche, si la dissimilarité est inférieure à un seuil

1. Pour ce pré-traitement, nous avons simplement calculé l'indice de Jaccard entre chaque objet à la dissimilarité faible, à différents pas de temps successifs, et nous avons éliminé ceux qui ont un recouvrement trop important, par simple seuillage.

bas, σ^s_{accept}, les vidéos sont très probablement prises depuis des points de vue très proches et quasiment similaires. Dans ces deux cas, l'étude de la paire de vidéo s'arrête, elles ne sont pas comparées à une échelle plus fine. Dans tous les autres cas, c'est-à-dire ceux d'une dissimilarité entre ces deux seuils, les comparaisons sont affinées en réduisant les échelles spatio-temporelles. Ainsi, les cellules sont divisées en sous-régions avec un facteur d'échelle ϕ_{sp} et les histoires de chaque sous-région sont exprimées avec une échelle temporelle plus fine ϕ_t. Comme nous effectuons des comparaisons plus précises, les seuils sont mis à jour en conséquence : plus l'échelle est fine, plus σ^s_{accept} augmente et plus σ^s_{reject} diminue [2]. Ce processus est répété pour toutes les paires de cellules. Au final, le taux de correspondance entre deux vidéos est directement lié à la proportion de cellules appariées avec cet algorithme.

Algorithme 2 : Estimation du lien entre deux vidéos – Ce que nous appelons *liens* est l'ensemble des régions qui peuvent avoir un lien, *candidats* est l'ensemble des triplets de régions mises en correspondance, avec l'échelle associée. Quand la paire $(R_1, R_2, scale)$ est rejetée, aucune correspondance n'est recherchée dans des échelles plus fines (c'est ce que nous avons voulu mettre en avant en ajoutant l'instruction *null*). La fonction Subdiv(R) retourne toute les sous-régions contenues dans R à l'échelle suivante.

```
1  Fonction estimation_liens (S^{V1}, S^{V2})
   Entrées : S^{V1} et S^{V2} : histoires de V1 et V2
   Sorties : liens : ensemble des liens entre V1 et V2
2  liens = [ ] ;                          /* Ensemble de régions où nous supposons un lien */
3  candidats = {(V1, V2, 1)} ;                /* Premier candidat étudié : toute la vidéo */
4  /* Recherche dans les échelles spatio-temporelles de plus en plus fines        */
5  tant que candidats ≠ ∅ faire
6      /* Récupération de la première correspondance de régions avec son échelle   */
7      (R1, R2, s) = candidats[0]
8      si d(S_R1, S_R2) ⩾ σ^s_accept alors
9          liens = liens ∪ {(R1, R2)}
10     sinon
11         si d(S_R1, S_R2) ⩽ σ^s_reject alors
12             /* Ajout de tous les triplets de la région subdivisée                */
13             candidats = candidats ∪ { (r1, r2, s+1), ∀ (r1, r2) ∈ Subdiv(R1) × Subdiv(R2) }
14         sinon
15             null
16     candidats = candidats \ candidats[0]
```

7.3.4. Résultats de construction de graphes

Données testées – Comme au § 7.2.5, il est difficile d'obtenir des jeux de données qui correspondent à l'application visée et nous avons dans un permier temps utilisé la base de données ToCaDa [Malon 18]. Nous rappelons qu'elle contient plus de 20 vidéos synchronisées positionnées sur le campus de l'Université de Toulouse III. Parmi ces vidéos, 16 caméras possèdent des zones de recouvrement alors que les autres ne possèdent pas de recouvrement. Nous rappelons que ces vidéos ont été acquises dans un contexte contrôlé et elles filment une scène que nous avons scénarisé au préalable. Ainsi, pour enrichir notre jeu de données, nous avons ajouté 4 caméras de vidéosurveillance publique disponibles sur Youtube. Seulement 2 possèdent une zone de recouvrement. Nous avons ainsi ajouté 8 vidéos. Le résultat attendu sur ce jeu de données est un graphe avec un nœud par vidéo où les 16

2. Expérimentalement, nous avons choisi $\sigma^s_{reject} = 0.9^s$ et $\sigma^s_{accept} = 1 - 0.9^s$.

caméras avec recouvrement de ToCaDa sont liées, ainsi que les paires de vidéos liées dans le cas des caméras de surveillance sur Youtube. Toutes les autres caméras ne doivent pas présenter de liens.

Détails d'implémentation – Comme pour la reformulation de trajectoire, nous avons détecté les objets avec *Mask-RCNN, Regional Convolutional Neural Network* [He 20] et avec un pas temporel $\tau = 500$ ms. Mask-RCNN nous permet d'obtenir pour chaque objet sa catégorie ainsi qu'un masque binaire, plus précis qu'une simple boîte englobante. Les propriétés que nous avons exploitées consistent simplement en un histogramme couleur quantifié sur 16 niveaux, comme proposé dans [Yang 14].

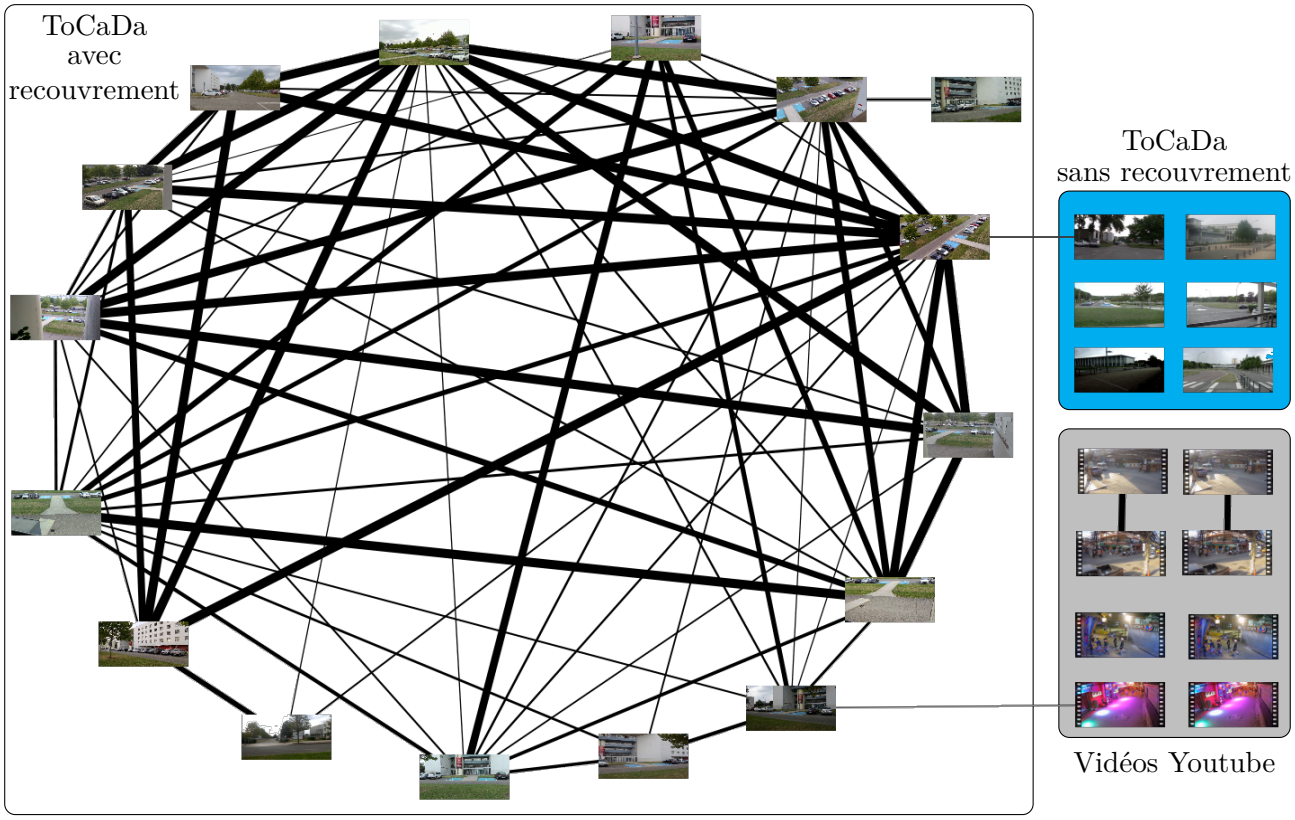


FIGURE 7.12. – Résultat de l'estimation du graphe de relations des vidéos – Voici le résultat obtenu en 3 minutes sur un simple Intel Core i7-7820HQ, avec 4 échelles différentes, soit $s = 1...4$ et $\phi_{sp} = 3$ avec un pas $\phi_t = 10$.

Résultats obtenus – La figure 7.12 nous permet de présenter visuellement les résultats obtenus. L'épaisseur du lien dans le graphe est proportionnel au taux de correspondance obtenu entre les deux vidéos. Comme attendu, les vidéos avec recouvrement sont toutes liées entre elles, elles forment une clique, et nous les avons disposées en cercle pour mettre en évidence ce résultat. Certaines des vidéos ont été acquises dans des conditions dégradées, comme cela pourrait être le cas d'une caméra de smartphone tenue à la main et en conséquence, certains objets statiques sont tout de même détectés comme mobiles, ce qui diminue le taux de correspondance. Parmi toutes les vues avec recouvrement, une seule vidéo (en haut à droite) ne possède qu'un seul lien avec une autre vidéo. Il se trouve que dans cette vidéo la première minute n'a pas été acquise et donc, elle est comparée aux autres qui ont une minute de plus ce qui la pénalise. En effet, aucun objet ne peut être trouvé en commun, donc tout objet détecté dans une autre caméra pendant cette minute fera partie de la proportion d'objets qui ne trouvent pas de correspondant. Enfin, deux liens faux ont été estimés entre deux vues disjointes et mettent en évidence les limites de l'approche : sans ré-identification, et avec peu de propriétés

exploitées, dans certains cas, des objets différents sont appariés et entraînent ces erreurs. Toutefois, dans ces deux cas, le taux de correspondance est assez faible.

7.3.5. Conclusion sur l'analyse de scène par vidéos multiples avec recouvrement

Cette approche a été proposée pour permettre une navigation facilitée dans une collection de vidéos afin d'en extraire des informations complémentaires sur des éléments d'une scène étudiée. Cette méthode a l'avantage de proposer une solution dans des conditions extrêmes d'acquisition, mais assez réalistes, et elle ne nécessite pas de calibrage ou de metadata. Plus précisément, nous avons introduit la notion d'histoires, de distance entre ces histoires et finalement, une méthode de construction de graphe de relations. Les résultats obtenus sont encourageants et contiennent peu d'erreurs.

Toutefois, de nombreuses améliorations sont envisagées. Nous aimerions utiliser d'autres descripteurs, une description sémantique et même des descripteurs de sons. Ensuite, pour améliorer le graphe, il nous paraît important de travailler sur la détection de liens entre caméras sans recouvrement. En effet, parfois, il existe un lien de séquence entre vidéos (un élément détecté en sortie d'une caméra sera visible à l'entrée d'une autre). Connaître ces liens est également intéressant pour faire une requête dans une large collection de vidéos.

Nous allons développer les perspectives liées à ce travail de manière plus détaillée, dans la partie IV, § 13.3.

Conclusion sur la mise en correspondance conjointe

Cette partie nous a permis, d'une part, de présenter toutes les contributions dans le domaine de la mise en correspondance stéréoscopique, et qui ont fait suite aux travaux de thèse [Chambon 05a]. Malgré tous les résultats encourageants obtenus et publiés, nous continuons à travailler sur ce sujet et nous proposons des perspectives dans ce sens au § 13.7. Le dernier chapitre aborde la notion d'appariement dans un contexte avec plus de capteurs et une dimension dynamique liée au travail de thèse en cours avec Thierry Malon. Le paragraphe 13.3 nous permettra de présenter un certains nombre de perpectives sur ce sujet.

Depuis les débuts de l'utilisation massive de l'apprentissage profond dans des approches de vision stéréoscopique que nous datons de 2012, de nombreuses publications ont été proposées sur le thème de la recherche de similarités, et en particulier, pour reconnaître des régions dans des images, comme dans [Shen 19]. Ces approches utilisent de manière classique des réseaux de neurones convolutifs [Han 15, Tolias 16, Radenović 16, Iscen 17], notamment en utilisant des réseaux siamois [Zbontar 15], introduits au § 2.4. Dans [Mayer 16], les auteurs introduisent le premier réseau qui intègre toutes les étapes classiques de mise en correspondance telles que décrites dans [Scharstein 02b] et proposent également le premier jeu de données synthétique avec vérité terrain, très utile à la communauté, pour permettre un entraînement des réseaux de neurones dans ce domaine de recherche. Par la suite, les travaux de [Liu 17, Liu 19b] permettent de perfectionner encore ces réseaux en introduisant un nouveau type de couche appelé *dilated convolutional layer* qui permet essentiellement de prendre en compte un voisinage plus grand et donc plus significatif pour effectuer la mise en correspondance. Tout aussi récemment, les auteurs de [Pang 17] ont ajouté au réseau DispNetC [Mayer 16] une étape d'affinement des disparités en formant ainsi un réseau à deux étapes. Le réseau GC-Net [Kendall 17] permet d'exploiter l'utilisation des volumes de coûts, *cost volume*[3]. La notion de multi-échelle a également été abordée dans [Zhang 14] en introduisant le concept de pyramide spatiale dans la phase de *pooling*, expliquée au § 12.2.1. Ainsi, dans ses travaux, [Chang 18] combinent cette notion d'analyse multi-échelle avec la convolution dilatée introduite dans [Liu 17].

Dans [Yao 18], un réseau est proposé pour prendre en considération des vues multiples et [Yao 19] améliore ce réseau en utilisant des réseaux récurrents. Dans ce travail, le principal intérêt à utiliser un réseau récurrent est de réduire les coûts en terme de calcul. Ces nouvelles techniques apportent un éclairage nouveau pour résoudre les problèmes de mise en correspondance de pixels. En particulier, ces approches mettent encore plus l'accent sur le traitement des données aberrantes [Zhang 19] ou le respect de contraintes sur les résultats obtenus [Zhou 15]. Ces approches permettent d'obtenir des résultats performants qui peuvent surpasser les approches classiques, c'est-à-dire sans apprentissage profond, mais elles entraînent également d'autres difficultés comme la gestion de la mémoire. De nombreux papiers traitent ce problème [Zhang 19]. Les travaux de [Zhou 15] abordent la mise en correspondance multi-images et s'attachent à généraliser le concept de vérification bidirectionnelle, ou contrainte de symétrie, à de multiples images. C'est ce qu'ils appellent la consistance circulaire. Pour prendre en considération ce type de contrainte, le nombre de vérifications possibles et nécessaires à

3. Il s'agit de construire un volume contenant tous les coûts possibles pour chaque disparité testée.

faire dans une collection d'images est lié de manière exponentielle au nombre d'images. Cela engendre également des problèmes de mémoire à gérer au mieux. Ces travaux consistent d'ailleurs à travailler sur une mise en correspondance de points d'intérêt et non une mise en correspondance dense. Dans ce papier, les auteurs ont aussi souligné la difficulté à construire une base d'apprentissage. En effet, mettre en correspondance des points d'intérêt, de manière massive, pour construire une base de donnée s'avère être une tâche délicate.

Une autre façon de traiter le problème de mise en correspondance est de l'aborder comme un problème de classification [Luo 16]. Cela a l'avantage de permettre l'utilisation d'un réseau de neurones avec moins de couches, moins de paramètres et donc moins coûteux en temps de calculs. Dans la suite des travaux de [Zhou 15], la mise en correspondance est réalisée en utilisant également une segmentation sémantique [Wang 18]. Des travaux comme ceux de [Wu 19] permettent d'exploiter tous les aspects récents introduits et utilisés dans des réseaux de neurones : prendre en compte des volumes de disparités, dans une approche pyramidale, tout en considérant une segmentation sémantique. Ces travaux sont les plus proches de ce que nous avons introduits dans cette partie sur la mise en correspondance conjointe et nous souhaitons donc, d'une part, identifier les éléments réellement comparables à ceux qui sont spécifiques à notre travail, afin de proposer des approches permettant d'introduire ces outils d'apprentissage profond tout en ajoutant les éléments sur lesquels nous avons travaillés. Nous aborderons plus de détails dans la partie IV, au § 13.7.

Troisième partie

Segmentation photométrique et géométrique

Introduction à la segmentation multi-critères

La segmentation d'images est un outil indispensable dans de nombreuses applications, comme illustré dans [Szeliski 10, chapitre 5]. Classiquement, nous pouvons distinguer les approches par contours des approches par régions, comme exposé dans les travaux de [Brun 96, partie 2]. Toutefois, dans la littérature, les différents outils de segmentation et de vision ont été combinés pour tenter d'améliorer les résultats obtenus, en utilisant de multiples segmentations [Mathevet 99], en fusionnant les résultats de manière spatio-temporelle [Chen 98, Chen 99a, Barnes 09], et en combinant les approches régions et les approches contours [Kermad 97, Sebari 07, El Merabet 13]. Une autre manière d'exploiter différents outils est de fusionner des informations de nature différente, en particulier des données 2D et 3D [Rosenhahn 07, Bray 06b, Kholi 08, Dambreville 08]. D'autres travaux permettent également l'utilisation d'attributs d'origines différentes ou encore la prise en compte de recouvrement spatial entre différentes images pour effectuer une segmentation. Par exemple, les publications de [Heitz 08, Tu 10] permettent d'utiliser une information contextuelle (les objets à segmenter ont des positions relatives particulières ou connues, les uns par rapport aux autres) alors que les travaux de [Zhuge 04, Weippl 07] font intervenir la sémantique (les régions sont associées à un mot clé, une définition). Nous pouvons citer les approches hiérarchiques qui font intervenir la notion de superpixels [Ren 03, Fulkerson 09, Achanta 12]. Enfin, la notion d'apprentissage apparaît depuis longtemps dans ce type de méthodes [Collins 02] et l'utilisation de l'apprentissage profond a connu un essor très important depuis 2010, avec des approches célèbres comme celle de [Krizhevsky 12]. Face à l'ampleur des applications visées et des approches développées pour y répondre, il est difficile d'évaluer, de comparer les résultats et, ainsi, de choisir une approche adaptée. En conséquence, de nombreux jeux de données ont été créés et mis à disposition, comme, par exemple, dans la publication de [Martin 01]. Avec l'arrivée de l'apprentissage profond, de nombreuses bases de données conséquentes sont apparues avec notamment la base ImageNet [Russakovsky 15].

Dans la littérature, face à un problème réel de segmentation, les étapes suivies sont la plupart du temps les suivantes : pré-traitement des données, binarisation ou classification et enfin post-traitement. Lorsque la segmentation fait appel à un apprentissage, le pré-traitement consiste, en partie, à sélectionner et traiter une base de données, et, dans le cas particulier de l'apprentissage profond, ce pré-traitement inclut une préparation des données qui permettra de les rendre exploitables dans le réseau de neurones mis en œuvre. Tous ces pré-traitements et étapes de préparation peuvent avoir un impact important sur la qualité du résultat obtenu. Le post-traitement permet souvent d'extraire des résultats plus interprétables pour réaliser la reconnaissance, l'identification de l'objet d'intérêt ou des objets de la scène. Il est aussi souvent proposé pour tenter de corriger les erreurs de segmentation et, encore une fois, cela peut avoir un impact important sur la qualité du résultat obtenu. Le but de cette analyse est de mettre en avant les différents éléments qui peuvent amener au succès d'un algorithme afin d'en comprendre les causes : quelle étape permet d'atteindre ce résultat ? Comment chaque partie contribue au résultat obtenu ? Dans cette partie, la démarche que nous avons suivie dans chacune des propositions faites est d'identifier le plus clairement possible tous ces éléments et, surtout, d'évaluer leur influence sur la qualité du résultat obtenu.

Nous avons eu à gérer des données dynamiques différentes : séquence d'images avec recouvrement

pour la segmentation de scènes urbaines, cf. chapitre 9, vidéo d'images de chaussée, cf. chapitre 10, et séquences d'images de pièges pour insectes 11. Pour le premier exemple, nous avons pu exploiter la redondance d'information entre images alors que pour les deux cas suivants, ce n'est pas le cas. En effet, les vidéos de chaussée ne permettent pas d'acquisition redondante et pour les pièges pour insectes, nous n'avons pas obtenu les données espérées dans les temps. Nous avons dû pour ces deux derniers cas faire face à un contexte de segmentation assez difficile et il est apparu important d'exploiter de manière conjointe des informations photométriques et des informations géométriques.

Contributions scientifiques structurantes en segmentation

Dans le chapitre 8 de cette partie, toute la bibliographie présentée constitue une partie de l'Unité d'Enseignement Image, Modélisation et Rendu, proposée en deuxième année aux étudiants de l'ENSEEIHT, dans le parcours Multimédia. Plus précisément, cela fait partie des deux classes inversées et des 8 travaux pratiques dont les résultats obtenus illustrent le premier chapitre [Chambon 15a].
Le chapitre 9 permet d'introduire la contribution en analyse de scènes urbaines sur la notion de segmentation en superpixels, en collaboration avec l'entreprise `imajing`. Ces travaux ont été publiés dans : [Bauda 15b, Bauda 15a] et ils ont été réalisés au cours de la thèse de Marie-Anne Bauda [Bauda 16], co-encadrée avec Pierre Gurdjos, sous la direction de Vincent Charvillat, IRIT-REVA.
Le chapitre 10 présente la contribution principale en segmentation de fissures de chaussées en combinant des éléments photométriques et géométriques. Ce travail a été publié dans [Amhaz 14, Amhaz 16] et a été réalisé pour le projet projet européen TRIMM – *Tomorrow's Road Infrastructure Monitoring and Management* porté par l'IFSTTAR (2013–2014). De plus, il est lié, en partie, à la thèse de Rabih Amhaz [Amhaz 15], co-encadrée avec Vincent Baltazart, IFSTTAR, sous la direction de Jérôme Idier, IRCCyN, Nantes. Ce travail a permis la publication d'une base de données annotée [a]. Nous avons également réalisé un dépôt de logiciel [b] élaboré pendant la thèse de Rabih Ahmaz sur la détection automatique de fissures dans des images par la SATT [c] de l'Ouest et TTT [d]
Ce sujet m'a permis une collaboration avec une post-doctorante, Argyro Karatanou et un doctorant, Nicolas Courdray. Nous avons publié une approche s'appuyant sur un algorithme de ligne de partage des eaux multi-échelle [Coudray 10]. Nous avons également publié avec Aurélien Cord, collaboration interne à l'IFSTTAR, une approche de classification de sous-images extraites d'images de chaussées, en utilisant un opérateur d'appuyant sur Adaboost [Cord 11]. En ce qui concerne ce chapitre, nous ne présenterons que le travail issu de notre dernière publication sur le sujet [Amhaz 16].
Récemment, la société `Kawantech` a montré son intérêt dans ce domaine et a financé une expertise sur la détection de chaussées par des caméras installées sur des mâts d'éclairage public.
Le chapitre 11 fait le lien entre les travaux présentés au chapitre 4 (détection de points d'intérêt) et la segmentation. Nous avons proposé une approche combinant l'utilisation des opérateurs de courbure et une segmentation par régions [Bakkay 18a, Bakkay 18b]. Ces travaux ont été réalisés avec un post-doctorant, Chafik Bakkay, financé par le projet FUI VITI OPTIMUM 2.0, en tant que prestataire. L'algorithme a été transféré à l'entreprise `SiConsult`.

a. Sur la page web dédiée, https://www.irit.fr/ Sylvie.Chambon/Crack_Detection_Database.html, il y a : les 5 catégories d'images en noir et blanc avec des textures différentes annotées (segmentations de référence que l'utilisateur peut exploiter pour effectuer une évaluation quantitative) et les résultats de 4 approches (3 approches proposées et publiées, notamment [Chambon 09, Chambon 11d, Chambon 11a] et 1 approche de l'état de l'art qui a été implémentée [Nguyen 11], au cours du stage de master 1 de Gaël Michelin).

b. Le certificat a été délivré par l'Agence pour la Protection des Programmes avec le numéro IDDN, *Inter Deposit Digital Number* : IDDN.FR.001.420010.000.S.P.2018.000.21000.

c. Sociétés d'Accélération du Transfert de Technologies.

d. Toulouse Tech Transfert.

Chapitre 8.

Méthodes de segmentation et de sur-segmentation d'images

8.1. Introduction sur la notion de segmentation

Qu'est ce que c'est ?

Nous souhaitons définir la segmentation ainsi :

> La segmentation a pour but de définir une partition de l'image sachant que chaque partition représente une information homogène et cohérente de la scène et qui a du sens.

Il y a autant de définitions de segmentation que de manières de définir qu'une information est homogène et cohérente. C'est pourquoi, souvent, on parle de problème de segmentation mal posé. Il n'existe pas de segmentation idéale et le choix d'une méthode est lié à, entre autres :

- La nature des images (éclairage, texture, ...) ;
- La forme des primitives à extraire ;
- Des contraintes de temps.

Dans la littérature, les segmentations fond/forme (deux classes seulement) sont distinguées des segmentations à plusieurs objets. De plus, nous considérons également ces trois catégories d'approches :

- La **segmentation supervisée**, qui utilise une base d'apprentissage pour avoir des connaissances *a priori* sur les régions recherchées.
- La **segmentation semi-supervisée**, où une seule connaissance *a priori* est utilisée et généralement, il s'agit de la connaissance du nombre de régions recherchées.
- La **segmentation non supervisée**, où aucune connaissance *a priori* n'est utilisée.

Plus formellement, voilà une manière de définir une segmentation :

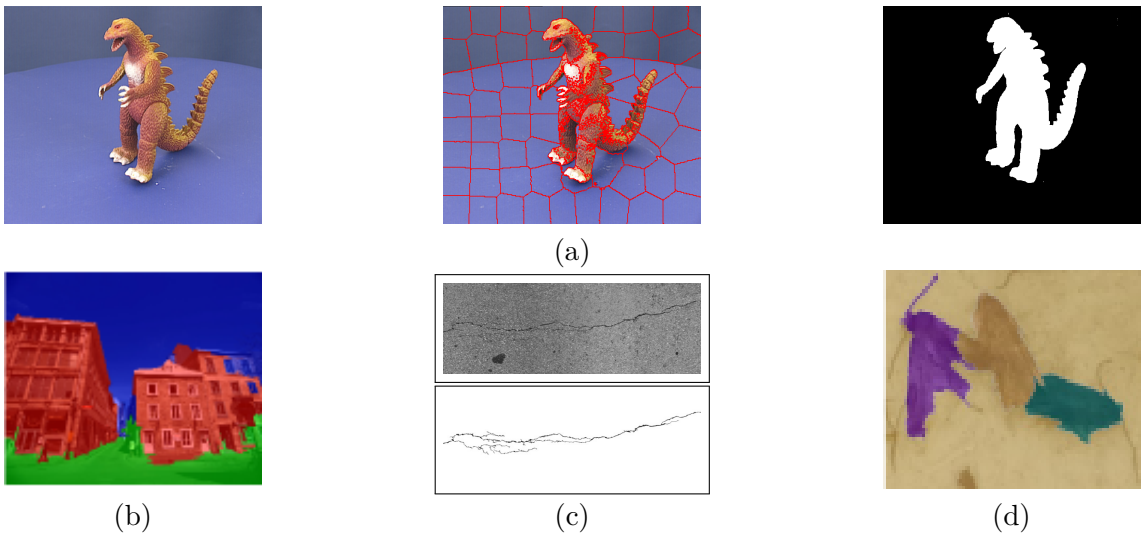
> Segmenter consiste à créer une partition de l'image I en sous-ensembles R_i appelés régions tels que :
> $\forall\ i\ R_i \neq \emptyset$
> $\forall\ i, j$ avec $i \neq jR_i \cap R_j = \emptyset$
> $I = \cup_i R_i$

Une région est donc un ensemble de pixels ayant des propriétés communes qui les différencient des régions voisines : niveaux de gris, couleurs, textures, formes. Une région peut être connexe ou non.

Pourquoi ?

Ils existent de nombreuses applications qui s'appuient sur une segmentation d'images, en particulier en imagerie médicale pour délimiter différents organes, les travaux que nous avons présentés dans le § 3.3 s'appuient notamment sur une segmentation préalable du poumon. Nous pouvons distinguer les cas de segmentations de type fond/forme, avec une seule forme à reconnaître comme dans l'exemple

8.1.(a). Nous pouvons aussi envisager des cas de segmentations d'une même forme sur un fond mais cette forme peut être présente en plusieurs exemplaires ou parties, comme la détection d'une fissure (la forme à segmenter) dans une chaussée (le fond texturé), ou des insectes à identifier sur une image de piège, cf. figure 8.1.(c) et (d). Enfin, nous pouvons avoir des cas de segmentation de scènes qui comportent plus d'éléments distincts, comme des scènes urbaines, cf. figure 8.1.(d).



(a) (b) (c) (d)

FIGURE 8.1. – Exemples de segmentation – En (a), dans le cadre de travaux pratiques proposés à l'ENSEEIHT, nous sur-segmentons une image d'origine (à gauche) afin d'en extraire une segmentation fond/forme (à droite). En (b), nous présentons un résultat de segmentation de scènes urbaines, chapitre § 9 (il s'agit des différentes couleurs appliquées sur la scène). En (c), il s'agit d'un résultat présenté au chapitre 10 de segmentation de fissures de la chaussée (indiquée en noire en bas). Enfin, en (d), nous montrons un résultat de détection de papillons sur une image de piège (il s'agit d'un extrait et la segmentation obtenue est représentée par les couleurs ajoutées sur l'image d'origine), cf. chapitre 11.

Comment ?

Contrairement à la détection de points d'intérêt que nous avons abordée dans la partie I, il est difficile de définir un algorithme générique. Nous proposons donc de présenter les principales familles au § 8.2. Des détails sont également fournis en annexe C.

Qualité du résultat ?

La littérature est abondante sur ce domaine de l'évaluation de la qualité de la segmentation, comme dans [Zhang 96, Y. J. Zhang 01], et les travaux [Pont-Tuset 15] donnent une bonne synthèse de cet aspect, pour le lecteur intéressé. De la même façon que pour la mise en correspondance, en supposant que nous possédons une segmentation de référence faisant office de vérité terrain, en segmentation, nous pouvons évaluer les pourcentages de : vrais positifs, *True Positives* (TP), faux positifs, *False Positives* (FP), faux négatifs, *False negatives* (FN) et vrais négatifs, *True Negatives* (TN). Les valeurs trouvées pour chacune de ces quatre catégories de pixels peuvent être combinées afin de déterminer une performance globale. Plus précisément, voici les critères qui peuvent être évalués :

- La précision (P) permet de mettre en évidence la proportion de vrais positifs parmi les positifs

(la segmentation estimée) :

$$P = \frac{TP}{TP + FP} \tag{8.1}$$

- La sensibilité (S) met en évidence la proportion de bonnes détections par rapport à la forme à segmenter. Cela nous donne une indication indirecte sur la proportion de manques :

$$S = \frac{TP}{TP + FN} \tag{8.2}$$

- Le coefficient de similarité ou *DICE Similarity Coefficient*[1] (DSC) est la moyenne harmonique entre la précision et la sensibilité :

$$DSC = \frac{2\,TP}{2\,TP + FP + FN} \tag{8.3}$$

Dans le cas d'une segmentation parfaite, $DSC = 1$ alors que $DSC = 0$, quand la segmentation est entièrement fausse. Comme ce coefficient permet de combiner précision et sensibilité, il est très utilisé.

8.2. Approches de segmentation existantes

Il est difficile d'établir une classification exhaustive des méthodes de segmentation, mais nous avons fait le choix de distinguer les trois méthodes suivantes :

(1) Les approches **contours**, où nous cherchons les contours entre objets peu semblables, c'est-à-dire les contours entre zones hétérogènes. La technique la plus utilisée s'appuient sur la notion de contours actifs (*snakes*) [Kass 88].

(2) Les approches **régions**, où nous cherchons à regrouper les pixels similaires pour former des régions homogènes, comme :

  (a) La fusion/division de régions (*split and merge*) [Horowitz 76].

  (b) La croissance de régions, comme la ligne de partage des eaux, *watershed* [Vincent 91].

(3) Les approches s'appuyant sur le principe de la **classification** [Duda 01], les plus connues et les plus utilisées étant :

  (a) Les approches bayésiennes [Domingos 97] ;

  (b) Les approches par Espérance-Maximisation (EM) [Dempster 77] ;

  (c) Les k-moyennes (k-*means*) [MacQueen 67] ;

  (d) Le *mean-shift* [Comaniciu 02] ;

  (e) Les approches par *Support Vector Machine* (SVM) [Boser 92].

Il existe également des approches hybrides ou duales, qui mélangent les deux premières approches[2]. Enfin, quelle que soit la méthode choisie, les critères utilisés (couleur, position, etc.) sont déterminants. Nous présentons des détails sur ces trois types d'approches en annexe C.

Actuellement, il existe des approches plus avancées permettant de combiner différentes approches notamment :

- en utilisant des segmentations multiples [Mathevet 99];
- en s'appuyant sur la notion de sur-segmentation par superpixels permettant une première estimation fiable de la segmentation finale [Moore 08] ;

1. Dans la littérature, on parle aussi de mesure/score f_1.

2. Il est intéressant de noter que le système visuel humain est plus apte à détecter des contours que des zones homogènes.

- en combinant des informations de nature différente plus riches que le niveau de gris ou la couleur comme des informations 2D (contours, points d'intérêt, régions) [Ibenthal 00], des informations 3D [Kholi 08] ou des informations contextuelles [Hoiem 05b].

Ainsi, nous allons présenter les approches par superpixels, cf. § 8.3, pour permettre l'introduction d'une première contribution au sujet de la segmentation de scènes planaires dans le chapitre 9.

8.3. Notion de sur-segmentation et de superpixels

8.3.1. Introduction à la sur-segmentation

Dans la littérature, de plus en plus de papiers traitent de la construction et/ou de l'utilisation des superpixels [Achanta 12]. En vision par ordinateur, l'élément élémentaire étudié pour réaliser une segmentation, un suivi ou un appariement est très souvent le pixel. Un certain nombre d'approches rapides et efficaces ont ainsi été mises en œuvre. Toutefois, l'information relative à un seul pixel (intensité, couleur, position) peut être insuffisante dans certains cas. En effet, les caractéristiques d'un pixel seul fournissent bien souvent une information ambiguë et incomplète. De plus, le pixel, seul, ne représente pas une entité physique naturel qui a du sens. Ainsi, très rapidement la notion d'exploitation du voisinage a été introduite, cf. l'utilisation de mesure de corrélation, de régions d'intérêt, voir de *patches*. Ces primitives, appelées primitives intermédiaires, permettent d'enrichir l'information exploitée en ajoutant des caractéristiques photométriques et/ou géométriques plus riches que celles associées à un pixel isolé. Cette notion de région d'intérêt peut encore être perfectionnée avec la notion de points et objets d'intérêt, cf. partie I. Ces détecteurs sont devenus de plus en plus précis, de plus en plus efficaces et ont surtout contribué à améliorer les performances dans le cadre d'applications précises comme la classification, la reconnaissance d'objets spécifiques et le suivi. Toutefois, si nous reprenons l'exemple de la détection d'objets particuliers, les modèles proposés permettent l'amélioration des approches de classification d'objets mais, du fait de la spécialisation de ces opérateurs, dans un but applicatif précis (cf. segmentation fond/forme, reconnaissance de classe d'objets), ceux-ci sont difficiles à utiliser en dehors du contexte ou de l'application pour laquelle ils ont été proposés. De la même façon, l'utilisation de *patches* permet une grande stabilité utile au cadre du suivi, par exemple, mais ceux-ci ne possèdent pas de propriétés photométriques et/ou géométriques suffisamment spécifiques pour servir, par exemple, de base à une segmentation de l'image. Toutes ces raisons ont amené à exploiter une information différente, certes moins discriminante qu'un objet ou un point d'intérêt mais plus riche en information photométriques et/ou géométriques, moins stable en terme de suivi mais avec une forme permettant une meilleure adaptation à la réalité de la scène, à savoir des sur-segmentations de l'image, qui ont finalement été appelées des superpixels depuis [Ren 03].

8.3.2. Domaines d'application des superpixels

Les superpixels ont été utilisés dans les domaines suivants :

- Bien évidemment, comme étape de pré-segmentation ou plus exactement de sur-segmentation avant une segmentation, comme dans [Hoiem 05b]. Dans cet article, les superpixels sont fusionnés en prenant en compte la catégorie associée à chaque superpixel et en vérifiant la cohérence des catégories des superpixels adjacents. Il s'agit donc de combiner segmentation sémantique et utilisation des superpixels.
- Des approches d'indexation d'images exploitent ces outils. Par exemple, dans l'article de [Zhang 10], une image requête est mise en correspondance avec les images indexées en exploitant les superpixels.
- En détection et reconnaissance d'objets, nous pouvons citer des approches comme celle de [Mori 05] où les superpixels sont utilisés pour rechercher un modèle de corps humain dans une image. L'auteur utilise les superpixels pour deux raisons : réduire l'espace de recherche du

modèle et avoir un support spatial du masque le plus précis possible, par rapport à une simple boîte englobante.

- Le suivi d'objets rencontre de nombreuses difficultés comme le changement d'échelle, d'orientation ou même de luminosité de l'objet. De plus, l'objet peut, au cours de la séquence être occulté partiellement ou même totalement le temps de quelques images. La difficulté augmente encore lorsque l'objet suivi est non rigide et qu'il se déforme donc d'une image à l'autre. L'hypothèse souvent faite est que deux images consécutives sont suffisamment proches temporellement pour que les modifications de l'objet soient faibles (aussi bien du point de vue de ses modifications internes comme la rotation ou le changement d'échelle que du point de vue de ses modifications externes comme les occultations). Ainsi, les auteurs de [Ren 07] cherchent à effectuer le suivi en s'appuyant sur des superpixels.
- La construction de carte de saillance [3] peut également s'appuyer sur l'utilisation de superpixels. Les premiers algorithmes s'appuyaient sur des études psycho-visuelles, ainsi que sur des dispositifs d'*eye-tracking*. Si ces méthodes de cartes de saillance sont très utilisées, elles ont néanmoins le défaut de représenter l'information qu'elles apportent sous la forme de tâches de lumineuses : une carte de saillance apporte en soi peu d'information sur la structure des objets intéressants. C'est une des raisons pour laquelle [Lu 12] propose une méthode de calcul de carte de saillance basée sur les superpixels. L'approche des auteurs consiste à rechercher des propriétés particulières dans l'image, comme la symétrie, au niveau du superpixel, et donc à construire des cartes de saillance qui possèdent une cohérence liée à la structure décrite par les superpixels. Ainsi les structures d'objet apparaissent sur les cartes de saillance.

8.4. Définition des superpixels

Bien évidemment, il n'existe pas de définition classique des superpixels plébiscitée par tous les chercheurs car chacun a sa manière propre de les définir. Toutefois, nous pouvons tenter de donner une définition suffisamment générique qui puisse être reprise par chacune des approches proposées.

> Un superpixel est une structure locale et cohérente qui représente une zone ou une partie d'un même objet.

Les variations de définitions d'un superpixel viennent en fait de l'interprétation mathématique que nous pouvons faire des termes employés dans cette définition, à savoir : structure, locale et cohérente. Plus précisément, le sens de ces termes dépendra de l'ensemble des propriétés souhaitées. Il est évident que cette définition est très proche de la définition de la segmentation que nous avons fournie au § 8.1. La seule différence est que chaque région doit se restreindre à un seul objet (ou une partie de l'objet). Les différents constructeurs de superpixels que nous présentons ont été publiés entre 2000 [Shi 00] et 2012 [Achanta 12]. Toutefois, le mot superpixels apparaît pour la première fois dans [Ren 03].

8.4.1. Propriétés

Dans un premier temps, nous décrivons les propriétés couramment utilisées dans la littérature pour la construction de superpixels, avant d'introduire les principaux algorithmes de construction que nous pouvons trouver dans la littérature. Nous présentons d'abord les propriétés héritées de la collection de pixels qui composent le superpixel, à savoir les propriétés d'apparence et les propriétés spatiales. Nous nous intéressons ensuite aux propriétés d'échelle, relative à leur support spatial, à la résolution de l'image ou encore au nombre de superpixels souhaité.

Par la suite, nous utilisons les notations suivantes :

3. La carte de saillance d'une image permet de modéliser l'attention visuelle humaine sur une image donnée. Plus précisément, une carte de saillance associée à une image se présente sous la forme d'une image en niveaux de gris pour laquelle les régions les plus brillantes correspondent aux points qui attirent le plus l'attention d'un utilisateur, et donc qui révèlent les objets intéressants de l'image.

- N_S : le nombre de superpixels recherché ;
- $\mathcal{S}$: un ensemble de superpixels, contenant $\#(\mathcal{S})$ superpixels, et S_i un superpixel d'indice i de cet ensemble ;
- $\mathbf{p}_i$: le pixel d'indice i du superpixel S ;
- σ_S: l'écart-type de la couleur au sein du superpixel S ;
- $\mathcal{V}_S$: l'ensemble des voisins d'un superpixel S donné.

Propriétés d'apparence – Nous distinguons les aspects suivants :

- *Statistiques sur la colorimétrie ou l'intensité :* les propriétés d'apparence d'un superpixel donné résument les propriétés des pixels qui le composent en appliquant des statistiques de premier ordre (moyenne/écart-type) sur l'intensité ou la couleur des pixels.
- *Différence intra/inter région :* les méthodes proposées dans [Shi 00, Felzenszwalb 04] minimisent la dissimilarité entre deux régions tout en maximisant la similarité entre pixels au sein d'une même région. Cette notion est appelée *différence intra/inter région*. Plus formellement, dans [Shi 00], voici le critère C_{DIIR} qui est minimisé :

$$C_{DIIR}(\mathcal{S}) = \frac{1}{\#(\mathcal{S})} \sum_{S_i \in \mathcal{S}} \frac{\frac{1}{\#(\mathcal{V}_{S_i})} \sum_{S_l \in \mathcal{V}_{S_i}} \sigma_{S_i \cup S_l}}{\sigma_{S_i}} \tag{8.4}$$

  Avec la méthode décrite dans [Felzenszwalb 04] , le critère utilisé pour évaluer la différence intra/inter région entre deux superpixels $\mathcal{S}_1$ et $\mathcal{S}_2$ est donné par :

$$C_D(S_1, S_2) = \frac{\min\left(\max_{(\mathbf{p}_i, \mathbf{p}_j) \in S_1} \delta(\mathbf{p}_i, \mathbf{p}_j) + \frac{c_s}{\#(S_1)}, \max_{(\mathbf{p}_k, \mathbf{p}_l) \in S_2} \delta(\mathbf{p}_k, \mathbf{p}_l) + \frac{c_s}{\#(S_2)}\right)}{\min_{\mathbf{p}_i \in S_1, \mathbf{p}_k \in S_2} \delta(\mathbf{p}_i, \mathbf{p}_k)} \tag{8.5}$$

  où c_s est une constante permettant de conditionner l'échelle des superpixels, c'est-à-dire que plus c_s est grand plus la taille des superpixels obtenus sera importante. Le terme δ définit une distance (par exemple, la différence entre les niveaux de gris de deux pixels).
- *Texture :* seul [Shi 00], à notre connaissance, introduit la notion de texture dans la distance utilisée. Dans ce travail, la texture est caractérisée en appliquant des différences de gaussiennes à différentes échelles.

Propriétés spatiales – Nous pouvons considérer les éléments suivants :

- *Position :* lorsqu'elle est prise en compte pour la construction des superpixels dans un critère tel qu'une distance [Shi 00, Achanta 12], la position peut contraindre les superpixels à être connexes et renforcer leur compacité. Ces deux aspects sont définis dans les deux points suivants.
- *Connexité :* un superpixel S est connexe si et seulement si pour toutes paires de pixels $(\mathbf{p}^i, \mathbf{p}^j) \in S$, il existe un chemin les reliant et qui passe uniquement pas des pixels de S. Différentes approches sont proposées pour respecter cette propriété de connexité. Dans [Achanta 12], il s'agit d'une étape de post-traitement qui traitent les cas de superpixels non connexes [4]. Dans les méthodes de construction basées graphe comme [Felzenszwalb 04], des relations spatiales spécifiques sont définies par la notion de voisinage entre les pixels (les n plus proches voisins ou les 8-voisins) et ainsi la construction même du graphe respecte cette connexité.
- *Compacité :* ce terme est souvent employé dans la description des superpixels mais il est clairement défini seulement dans [Levinshtein 09] où les auteurs la caractérise comme le rapport entre

4. Les parties non connexes les plus petites du superpixel étudié sont rattachées au superpixel voisin ayant la plus grande taille.

le carré du périmètre d'un superpixel et son aire :

$$C_{Comp}(\mathcal{S}) = \frac{1}{\#(\mathcal{S})} \sum_{S_i \in \mathcal{S}} \frac{P_{S_i}^2}{A_{S_i}}, \tag{8.6}$$

où P_S le périmètre du superpixel et A_S son aire. Certains auteurs proposent une construction des superpixels la plus compacte possible en exploitant la distance au centre du superpixel [Achanta 12] alors que d'autres s'appuient sur l'utilisation d'une grille pour initialiser les superpixels et limitent la déformation de cette grille [Moore 09] en exploitant les contours de l'image, c'est-à-dire que les contours des superpixels doivent correspondre aux contours de l'image. Nous pouvons également citer les approches qui imposent une contrainte sur le nombre de pixels par superpixels de manière à réguler la taille des superpixels, comme [Levinshtein 09].

- *Régularité topologique :* une autre propriété souvent favorisée est la régularité topologique [Moore 09], c'est-à-dire, le fait que chaque superpixel possède toujours le même nombre de voisins (sauf dans le cas particulier des bords). Cette propriété est intéressante si une modélisation markovienne des interactions entre primitives est envisagée. La plupart du temps pour respecter cette régularité topologique, une configuration initiale respectant cette propriété (une grille, une disposition régulière de *patches* est utilisée et le modèle d'évolution proposée s'appuie uniquement sur des modifications de la topologie de départ qui n'ont pas d'incidence sur cette contrainte de régularité.
- *Contours de l'image :* enfin, comme précédemment mentionné, de nombreux articles souhaitent favoriser le fait que les superpixels suivent au mieux les contours de l'image. Il s'agit donc de détecter les contours et de faire en sorte que les frontières des superpixels suivent au mieux ces contours [Moore 09, Levinshtein 09, Veksler 10].

Propriété d'invariance temporelle – Des approches permettent d'intégrer la notion d'invariance temporelle ou de cohérence temporelle des superpixels le long d'un flux vidéo ou d'un flux d'images [Grundmann 10]. Dans le cas d'un flux vidéo, la contrainte la plus utilisée s'appuie sur le fait que la position d'un objet entre deux images successives varie très peu et se retrouve dans un petit voisinage de l'image précédente.

Conclusion sur les propriétés – En conclusion, il y a peu de propriétés différentes réellement abordées dans la littérature parmi les propositions de constructeurs, pourtant, les applications visées requièrent souvent plus de propriétés comme la robustesse aux diverses transformations et la cohérence spatio-temporelle.

8.4.2. Construction/Modélisation

Plusieurs techniques de construction de superpixels existent et chaque auteur prend en compte ses propres contraintes qui induisent un certain nombre de propriétés comme celles listées au § 8.4.1. La priorité reste souvent la réduction du temps de calcul en réduisant le nombre de primitives, dans le sens de régions, à étudier.

Approches s'appuyant sur la théorie des graphes – Dans ce contexte, une représentation par graphe se décrit ainsi : les nœuds sont les superpixels et les liens, les relations de voisinage entre superpixels. Et, en conséquence, la méthode d'optimisation utilisée pour construire le graphe est la minimisation d'un critère (qui prend en compte les propriétés attendues) par coupure de graphes, *graph-cut*. Un algorithme générique est présenté dans [Shi 00]. Nous pouvons distinguer deux types de structure :

- *Régulière :* celle de [Moore 09] commence par prendre en considération une grille régulière dans l'image en fonction du nombre de superpixels souhaités. Puis chacune des droites de cette grille va évoluer en courbes à l'intérieur d'une bande autour de la droite. Cette évolution est réalisée afin de maximiser la somme des gradients des pixels situés sur la courbe dans le but de forcer les contours des superpixels à être sur les contours de l'image. Les auteurs souhaitent que cet algorithme fournisse plus de superpixels dans les zones de l'image où la densité de contours est plus élevée. Les bandes où évoluent les courbes séparant les superpixels sont donc construits de manière à avoir chacune la même probabilité d'obtenir un contour.
- *Non régulière :* la plupart des approches proposées fournissent des graphes réguliers de superpixels et seul [Felzenszwalb 04] propose une version de son algorithme qui ne respecte pas la propriété de régularité topologique pour donner la priorité à la cohérence photométrique des superpixels.

Approches statistiques – Les algorithmes de coupure de graphe décrits précédemment sont considérés comme coûteux en temps de calcul. Ainsi, d'autres approches s'appuient sur des méthodes de classification classique comme les k-moyennes, *k-means*, cf. § C.3.1, [Vedaldi 08, Achanta 12]. L'approche SLIC, *Simple Linear Iterative Clustering* proposée par [Achanta 12] est la plus connue et la plus utilisée. Dans l'algorithme introduit, une étape d'initialisation distribue de manière régulière les N_c centres des superpixels souhaités. Plus précisément, les centres, notés $\mathbf{c}_k$, sont les centres des cases d'une grille dont chaque case à une largeur et une hauteur de taille V qui dépend du nombre de superpixels souhaités. Puis, ces centres sont ajustés localement, dans un voisinage 3×3, pour éviter qu'ils ne se trouvent sur un contour (déplacement vers le gradient local le plus faible). L'algorithme utilisé par la suite est un *k-means* et pour attribuer chaque pixel à un superpixel (une classe), la distance D_s, combinaison pondérée des distances dans l'espace colorimétrique et dans l'espace géométrique, est utilisée :

$$D_s(\mathbf{p}, \mathbf{c}_k) = \sqrt{d_a(\mathbf{p}, \mathbf{c}_k)^2 + \left(\frac{m}{V}\right)^2 d_p(\mathbf{p}, \mathbf{c}_k)^2} \tag{8.7}$$

où d_a, d_p correspondent aux distances euclidiennes entre $\mathbf{c}_k$ et $\mathbf{p}$, respectivement sur le critère d'apparence, la couleur, et le critère de position. Le poids m permet de régler l'influence de chacune des distances. Pour affecter un superpixel d'appartenance à chaque pixel de l'image, ce calcul est réalisé pour chaque centre qui se trouve dans un voisinage $2V \times 2V$. De nombreuses variantes ont été proposés pour SLIC, comme, par exemple, celle de [Wang 12b] qui modifie simplement la forme des régions initiales en prenant des hexagones. Les auteurs de [Achanta 12] eux-mêmes ont amélioré cet algorithme en proposant une version non itérative, SNIC, *Simple Non-Iterative Clustering* [Achanta 17].

Approches par croissance de germes – Un dernier groupe de méthodes s'appuient sur l'initialisation et la croissance de superpixels soumis à des contraintes. Parmi ces algorithmes, les TurboPixels de [Levinshtein 09] sont les plus connus. Tout d'abord, pour le placement des germes, les auteurs utilisent exactement la même façon de faire que SLIC, c'est la suite qui est différente. Il est important de noter que les auteurs utilisent des techniques qui s'apparentent à la notion de *Level Set* ou de courbe de niveaux [Chan 02]. Ces approches sont elles mêmes assez proches des approches par contours actifs [Kass 88], cf. § C.1. Le principe est de faire évoluer des courbes dans l'image qui permettront de définir les frontières des superpixels. Par cette définition, nous pouvons constater que nous nous approchons du concept des contours actifs. Nous notons $\mathbf{c}$ un vecteur contenant les coordonnées de la courbe paramétrée par p, le paramètre qui parcourt la courbe, et t le paramètre d'évolution du temps. Le terme $\mathbf{n}$ correspond à la normal externe associée et chaque point se déplace à la vitesse v avec l'équation d'évolution de la courbe suivante :

$$\frac{\partial \mathbf{c}}{\partial t} = v\mathbf{n}.$$

L'évolution de la courbe est implémentée en représentant $\mathbf{c}$ par une courbe de niveau d'une fonction lisse Ψ telle que :

$$\Psi : \mathbb{R}^2 \times [0, \tau) \rightarrow \mathbb{R}^2 \tag{8.8}$$

$$\Psi_t = -v||\nabla\Psi|| \tag{8.9}$$

En pratique cette fonction Ψ correspond à la distance euclidienne signée entre le point et la frontière. Cette distance est positive lorsque le point se trouve assigné à la région et négative sinon. Ensuite, dans cet algorithme, l'état Ψ^{n+1} des superpixels à l'itération $n+1$ est défini en fonction de l'état des superpixels à l'itération précédente Ψ^n de la manière suivante, en suivant la discrétisation, au premier ordre en fonction de t de l'équation 8.9 :

$$\Psi^{n+1} = \Psi^n - v_I v_B \|\nabla\Psi^n\|\Delta t. \tag{8.10}$$

Chaque application de cette équation correspond à une évolution de la courbe Ψ à chaque pas de temps. Cette équation est appliquée jusqu'à ce qu'il n'y ait plus d'évolutions possibles. Le terme le plus important est le produit $v_I v_B$. Le scalaire v_I dépend de la structure locale de l'image et de la géométrie du superpixel en chaque point de la frontière alors que le scalaire v_B dépend de la proximité des points de la frontière avec les points de la frontière des superpixels dans son voisinage. Ces contraintes sont liées par exemple à la courbure des superpixels, à la position de la frontière actuelle des superpixels par rapport à des contours de l'image ou encore à la position d'un superpixel par rapport à un autre (pour empêcher deux superpixels de se recouvrir). Ainsi, v_I et v_B décrivent des vélocités et recouvrent des termes qui accélèrent ou freinent l'expansion des superpixels en fonction des contraintes extérieures. Ce modèle exploite donc simultanément des propriétés d'apparence et des propriétés spatiales.

Dans cette famille, nous pouvons également citer le constructeur SEEDS, *Super-pixels Extracted via Energy-Driven Sampling* [Van den Bergh 12]. À l'aide d'un processus itératif et d'une initialisation sur une grille régulière, seuls les pixels frontaliers à deux superpixels voisins peuvent changer de centre de rattachement en optimisant un terme prenant en compte le caractère homogène du superpixel ainsi que la régularité de sa forme. Plus précisément, l'homogénéité est relative à la fonction de densité de l'histogramme couleur au sein de la région et la régularité de la forme dépend du nombre de superpixels voisins localement (en chaque pixel frontalier).

8.4.3. Classement et comparaison des différents constructeurs de superpixels

Pour conclure, le tableau 8.1 permet de comparer tous les constructeurs présentés en prenant en compte les divers aspects abordés : les critères d'apparence utilisés et le modèle de construction utilisé. Le comportement d'une partie des constructeurs est illustré dans la figure 8.2 sur des images de vignes qui correspondent à des travaux réalisés dans l'équipe REVA [5].

8.5. Conclusion sur les superpixels

L'analyse de l'état de l'art s'appuient essentiellement sur les publications avant 2012, mais ce domaine de recherche est toujours très actif. Ainsi, certains travaux n'ont pas été abordés mais nous donnons ici les différentes références pour obtenir plus de détails. Des travaux ont été développés sur l'utilisation de polygones pour approximer les formes [Duan 15, Forsythe 16, Bauchet 18]. D'autres approches mettent en avant l'intérêt d'utiliser une analyse globale de l'image, plutôt qu'une analyse locale, comme dans l'approche SLIC, en exploitant des outils d'analyse spectrale [Li 15]. Enfin, en s'ap-

5. Il s'agit de la thèse de Jérôme Guénard [Guénard 13], co-encadrée par Pierre Gurdjos et Géraldine Morin, sous la direction de Vincent Charvillat.

Catégorie	Référence	Apparence			Spatiales					Temp.
		S	*D*	*T*	*Pos.*	*Comp.*	*Conn.*	*Rég.*	*Cont.*	
Graphe	[Shi 00]	–			–					
	[Felzenszwalb 04]	–	–		–					
	[Moore 09]	–					–	–	–	
	[Grundmann 10]	–	–		–					–
	[Veksler 10]	–	–			–				–
Statistiques	[Vedaldi 08]	–			–					
	[Achanta 12]	–		–	–	–	–		–	–
	[Wang 12b]	–				–	–	–		
Germes	[Levinshtein 09]	–				–	–	–	–	
	[Van den Bergh 12]	–			–	–	–		–	

Table 8.1. – Classement des constructeurs de superpixels – L'abbréviation Temp. correspond à la prise en compte de l'aspect temporel. Pour les propriétés d'apparence et spatiales, nous utilisons les notations suivantes : *S*, Statistiques sur la colorimétrie, *D*, Différence intra/inter région, *T*, Texture, *Pos.*, position, *Comp.*, compacité, *Conn.*, connexité, *Rég.*, pour le critère de régularité, *Cont.*, pour l'utilisation des contours de l'image.

puyant sur le concept de la ligne de partage des eaux, *watershed*, une nouvelle forme de superpixels a été introduite : les *waterpixels* [Machairas 15]. Nous avons réalisé cette analyse des l'état de l'art des superpixels car ils seront utilisés dans une partie des travaux que nous présentons dans cette partie, plus précisément le chapitre suivant. Nous reviendrons également sur cette notion de superpixels dans les perspectives, partie IV, § 13.7.

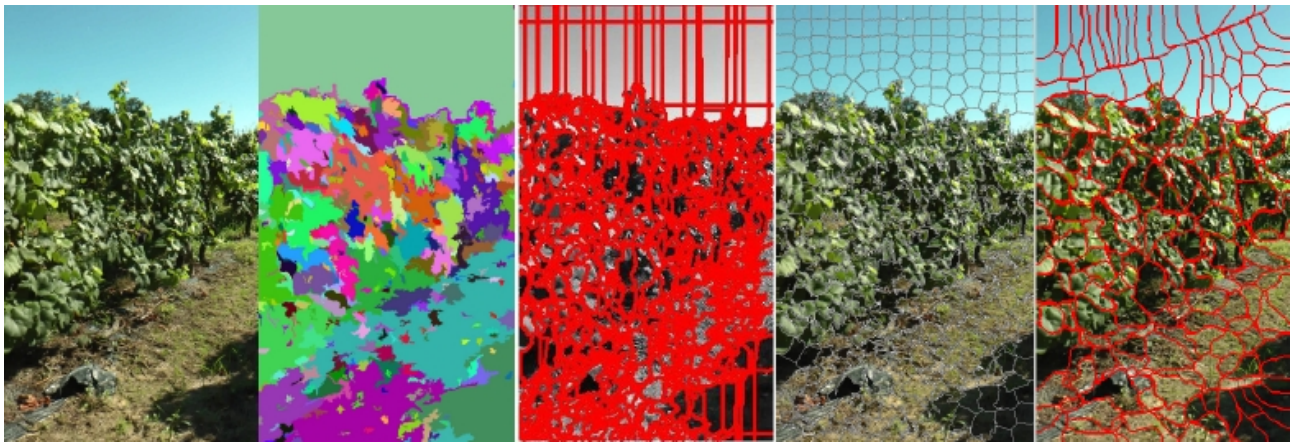


Figure 8.2. – Différences de résultats avec différents constructeurs de superpixels – De la gauche vers la droite, il s'agit de l'image originale, un résultat de superpixel donné par [Felzenszwalb 04], par [Moore 09], par [Achanta 12] et enfin par [Levinshtein 09].

Chapitre 9.

Superpixels adaptés aux scènes planaires

9.1. Objectif applicatif : la segmentation de scènes urbaines pour une ville confortable

Nous nous plaçons dans un contexte de cartographie mobile ou instantanée. Différentes technologies existent actuellement pour accéder aux données multimédia associées aux endroits physiques. Un des outils cartographiques les plus connus actuellement est Google Maps [1]. L'information est distribuée par couche, chacune contenant des données spécifiques comme la couleur des pixels, le nom de la rue ou d'un magasin. Les images géo-référencées des rues sont également proposées comme nouvelle source d'information liée à la cartographie. Actuellement, en terme de recherche, de nombreux travaux tentent d'exploiter cette source riche d'informations variées pour en extraire automatiquement des éléments d'intérêt, de la reconnaissance d'objets spécifique, comme le mobilier urbain, à la segmentation sémantique des scènes disponibles.

Pour la production de données cartographiques, la société `imajing` a développé l'`imajbox`, présentée dans la figure 9.1, qui permet d'acquérir des séquences d'images géo-référencées et orientées, au cours de la trajectoire d'un véhicule donné. Le type d'images acquises est illustrées dans la figure 9.2. Elle est équipée d'un capteur CCD (*Charge-Coupled Device*) d'une fréquence maximale d'acquisition de $7Hz$ et de taille 5 Mpx (2050×2448), et d'un système de positionnement GNSS, *Global Navigation Satellite System*, composé d'une antenne GPS, *Global Positionning System* et d'une centrale inertielle IMU, *Inertial Measurement Unit* composée d'un gyroscope, d'un accéléromètre et d'une boussole. Les données produites sont ensuite traitées par des algorithmes de navigation et de traitement d'images afin d'être visualisées, analysées et annotées par les utilisateurs avec le logiciel `imajview`. Le service `imajnet` permet d'accéder aux données sur internet.

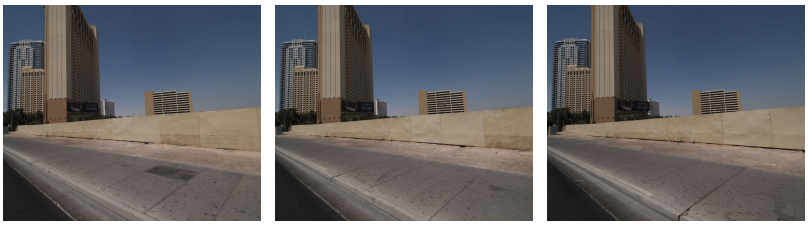


`imajbox 1` `imajbox 2`

FIGURE 9.1. – Système mobile d'acquisition `imajbox` – L'`imajbox 1` a été distribuée de 2009-2014. L'`imajbox 2` est disponible en 2015 en deux versions : compacte ou renforcée.

1. https://www.google.fr/maps/

Un des principaux usages de ce système est la réalisation d'inventaires et d'outils de gestion des équipements : signalisation horizontale et verticale, candélabres et points lumineux, glissières, obstacles latéraux, géométrie de l'environnement. La technologie portable d'acquisition `imajbox` permet de produire rapidement un très grand volume d'images traitées, en grande partie, manuellement. Ainsi, les travaux présentés dans ce chapitre s'inscrivent dans la proposition d'outils d'automatisation de traitement.



(a) (b) (c)

FIGURE 9.2. – Triplet d'images successives acquises par l'`imajbox` – En (b), il s'agit de l'image que nous appellerons centrale ou de référence, en (a) et (c) il s'agit des deux images dites adjacentes.

9.2. Problématique de la segmentation de scènes urbaines

L'objectif est de proposer une segmentation de la scène afin d'identifier les différents éléments qui la composent pour détecter les changements entre deux séquences d'images acquises à des instants différents. Le sujet principal ne porte pas sur la proposition de structures sémantiques, toutefois, nous avons introduit une segmentation sémantique adaptée aux données traitées par `imajing`. La segmentation des environnements urbains offrent des spécificités permettant d'en simplifier l'analyse (environnement presque plan par morceaux) et nous introduisons une méthodologie qui les intègre. L'intérêt de ce travail repose sur l'exploitation de données qui permettent une grande redondance (au sein de la même séquence acquise, mais également, une redondance temporelle). De plus, nous avons accès à des données de nature différente : points 3D, images géo-référencées et description sémantique de la scène.

9.3. Approches existantes en segmentation de scènes urbaines

La segmentation d'images est un problème délicat que nous avons déjà abordés au chapitre 8 et qui est complété par l'annexe C. Ici, nous nous intéressons particulièrement aux approches introduisant des connaissances contextuelles, comme celles de [Hoiem 05b]. Dans ces approches, les connaissances forment usuellement un « contexte » au sein duquel les inférences visuelles sont simplifiées. De nombreuses tâches de vision par ordinateur réputées délicates sont facilitées lorsqu'elles s'opèrent de manière contextuelle [Marques 11]. En particulier, de nombreux travaux ont mis en évidence l'intérêt d'aller au delà du contexte simplement basé sur l'apparence [Tu 10] et ont montré qu'une compréhension holistique de l'image [Hoiem 08, Heitz 08] permettait d'améliorer le processus d'analyse de la scène. Ainsi, les auteurs de [Gould 09] présentent une approche qui s'appuient simultanément sur les associations entre les pixels et les régions auxquelles ils appartiennent et les caractéristiques sémantiques et géométriques qui leur sont associées. Toutefois, ces techniques s'avèrent très coûteuses. D'un côté, elles travaillent au niveau du pixel pour chaque image et, d'un autre, elles ont besoin d'un grand nombre de données d'apprentissage. Actuellement, il est possible d'obtenir des reconstructions de scènes 3D à partir de plusieurs vues. Les travaux les plus connus à ce sujet sont ceux de [Furukawa 09]. Cette possibilité d'obtenir rapidement une reconstruction 3D d'une scène, même

partielle, ouvrent de nombreuses possibilités d'applications comme par exemple la génération d'effets spéciaux, cf. le projet AliceVision [2] ou la manipulation de modèles 3D pour l'édition et la manipulation [Morin 14, Durix 16a]. Dans la suite de ce travail, nous supposons qu'il est possible d'utiliser ce type d'outils pour obtenir des données d'entrées exploitables pour le reste de l'application.

9.4. Méthode proposée pour l'analyse de scènes urbaines

Nous nous intéressons à l'analyse de scènes urbaines, comme dans l'exemple fourni dans la figure 9.3, en (4). Chaque scène est représentée par une collection d'images acquises depuis un véhicule en mouvement et nous avons choisi de combiner des informations géométriques et photométriques afin de proposer une segmentation en plans, comme en (4). Nous faisons l'hypothèse que les objets d'intérêt sont associées à six classes principales d'objets : sol, ciel, végétation, constructions urbaines, mobiliers urbains et autres objets. Il s'agit du résultat de segmentation attendu. Nous supposons que les données de départ correspondent aux sorties d'un module de calcul de la structure et du mouvement, *Structure-and/from-Motion*, SfM [Cui 15] [3], voire d'un module de stéréoscopie multi-vues, *Multi-View Stereo*, MVS [Furukawa 10].

L'approche proposée s'appuie sur une représentation intermédiaire des images par des superpixels, point (3). Ce pré-traitement est essentiel car il nous permet d'aborder l'aspect de segmentation sémantique pour obtenir le résultat attendu. Les travaux existants que nous avons déjà présenté [Mori 04, Hoiem 05a, Gould 08] montrent l'importance et l'intérêt d'une telle approche, notamment pour réduire les temps de calcul par rapport à des méthodes classiques de segmentation. Dans notre cas, les constructeurs de superpixels doivent préserver certains contours occultants des objets considérés car ce critère a une influence directe sur les performances finales [Hanbury 08]. Nous nous intéressons en particulier à l'intégration d'une information photométrique s'appuyant sur un *a priori* géométrique de planéité des surfaces dans la construction de superpixels. Une vue d'ensemble de l'approche proposée est représentée dans la figure 9.3. La section suivante nous permet de détailler l'approche de sur-segmentation proposée.

9.5. Définition des superpixels géométriques

Notre objectif est d'obtenir une décomposition de l'image I en un nombre inconnu K d'entités géométriquement et sémantiquement cohérentes à partir de plusieurs vues calibrées d'une scène qui peut être considérée plane par morceaux.

Calculs préliminaires – En plus des données d'entrées déjà présentées au § 9.4 et dans la figure 9.3, nous supposons que nous sommes capables d'estimer les plans dominants de la scène avec, par exemple, une approche comme celle décrite dans [Bouchafa 09, Yang 10]. Pour chaque région planaire, notée z_i connaissant les paramètres du plan et la géométrie épipolaire, nous pouvons estimer l'homographie induite par le plan de support, comme dans [Bauda 13]. Cette homographie permet de calculer la région recalée z_j, alignée sur la région de référence. Les deux zones z_i et z_j sont alors comparées avec une mesure de similarité s'appuyant sur un produit scalaire et introduite dans [Wang 02] car c'est celle qui donne les meilleures résultats dans nos expérimentations [Bauda 15b] et [Bauda 16, Chapitre 3]. Plus précisément, la mesure UQI, *Universal Quality Index*, est donnée par :

$$\mathrm{UQI}(z_i, z_j) = \frac{4\sigma_{z_i z_j}\mu_{z_i}\mu_{z_j}}{(\sigma_{z_i}^2 + \sigma_{z_j}^2)(\mu_{z_i}^2 + \mu_{z_j}^2)}. \tag{9.1}$$

2. https://alicevision.org/

3. Un exemple de code disponible : VisualSFM: A Visual Structure from Motion System : http://ccwu.me/vsfm/, 2011.

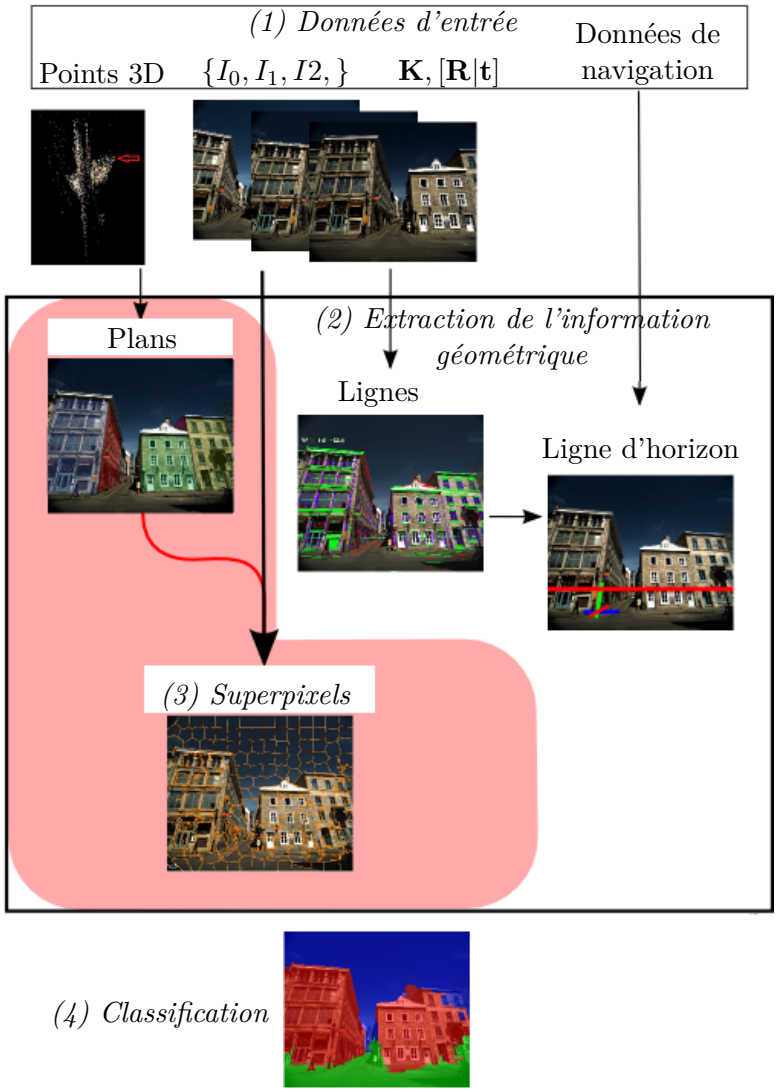


FIGURE 9.3. – Approche de segmentation de scènes urbaines par superpixels géométriques – À partir des données d'entrée, des informations géométriques sont extraites (plans, lignes et horizon) afin de réaliser une sur-segmentation guidée par la géométrie, afin de proposer un résultat de classification en trois classes (ciel, sol, objets verticaux). Nos contributions (mise en avant par le fond rose) portent sur l'intégration de l'hypothèse de planéité dans la construction des superpixels.

Extraction et intégration de l'information géométrique – Nous décrivons ici l'information géométrique que nous proposons d'extraire afin de l'intégrer au constructeur de superpixels [Achanta 12]. Plus précisément, nous proposons de construire deux cartes : une carte de planéité, qui correspond au calcul de la normale $\vec{n}$ à la surface 3D plane à laquelle appartient chaque pixel p, et une carte de similarité, qui correspond au calcul de la similarité comme décrit à la page 143. Ces cartes sont illustrées dans la figure 9.4. Quelle que soit la carte calculée, il est important de tenter d'éliminer les cas où le résultat obtenu ne semble pas fiable, c'est-à-dire, tous les cas où la valeur obtenue est inférieure à un seuil $< \epsilon$, à fixer. Ainsi, la mesure UQI binarisée est notée d_{UQI} et ce terme est utilisé dans la nouvelle distance proposée pour construire les superpixels.

Les deux cartes présentées précédemment sont intégrées dans la construction de superpixels géométriques, et plus précisément, la distance d'agrégation que nous proposons. Cette nouvelle énergie

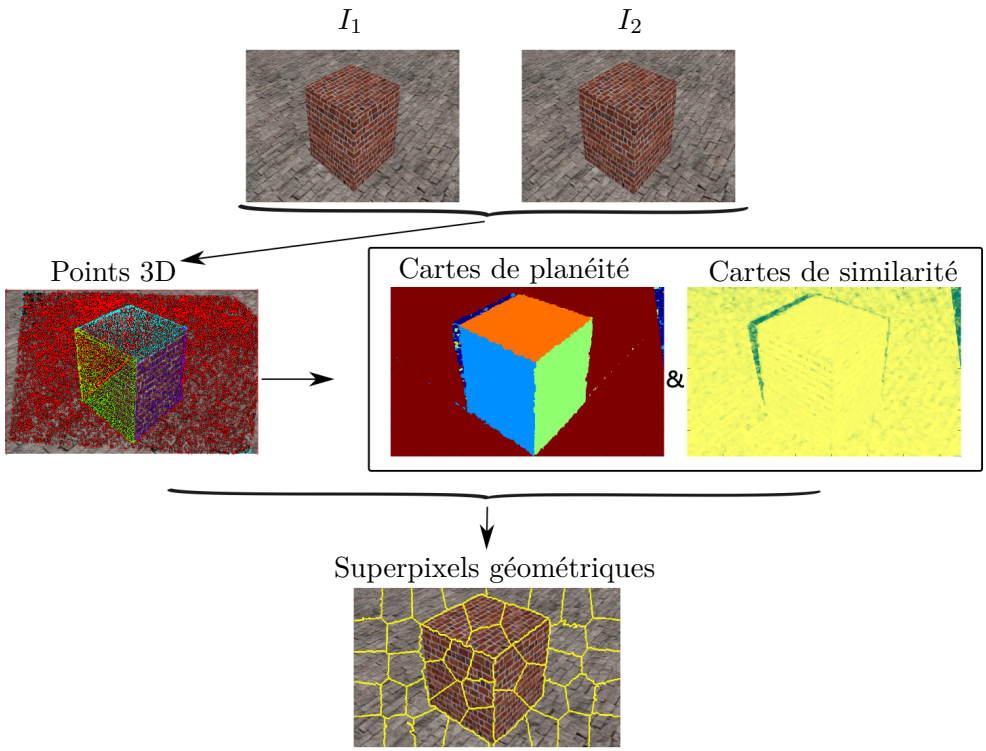


FIGURE 9.4. – Intégration de l'information géométrique dans la construction de superpixels – La première ligne présente les deux images calibrées utilisées (ici, il s'agit de simples images de synthèse). La seconde ligne présente la projection du nuage de points 3D dans l'image 2D puis la carte de planéité qui indique l'appartenance de chaque pixel à un plan et, enfin, la carte de similarité inter-image qui correspond à la valeur de la mesure UQI, cf. expression 9.1. Le calcul de ces deux dernières cartes est présenté au § 9.5. La dernière ligne expose enfin le résultat de la méthode de sur-segmentation géométrique cohérente avec les surfaces de la scène.

à minimiser, notée D_{SP}, intègre à la distance D_s, proposée par [Achanta 12] et définie par l'expression (8.7), une information géométrique, d_g :

$$\begin{aligned} D_{SP} &= \sqrt{\lambda D_s + (1-\lambda) d_g} \\ &= \sqrt{\lambda (d_{c_0} + d_{s_0}) + (1-\lambda) d_g}. \end{aligned} \tag{9.2}$$

Ce terme d_g prend en compte la géométrie de la scène en fusionnant l'information donnée par les cartes de planéité et de similarité. Il est défini par :

$$d_g(\mathbf{p}_j, \mathbf{p}_i) = 1 - d_{\mathbf{n}}(\mathbf{p}_j, \mathbf{p}_i) d_{UQI}(z_j, z_i). \tag{9.3}$$

La distance entre normales est notée $d_{\mathbf{n}}$ et indique l'appartenance d'un germe au même plan qu'un pixel voisin. Nous proposons deux manières différentes de définir $d_{\mathbf{n}}$: par calcul exact et par binarisation. Si nous supposons que l'orientation des surfaces est connue en chaque pixel de l'image, comme avec les données de synthèse présentées dans [Bauda 15b], alors nous savons que la variation d'orientation est donnée par le cosinus de l'angle que forme les normales en ces deux points :

$$d_{\mathbf{n}}(\mathbf{p}_j, \mathbf{p}_i) = \frac{1 + cos(\mathbf{n}_j, \mathbf{n}_i)}{2} \tag{9.4}$$

Dans le cas simplifié, les orientations des surfaces reconstruites sont obtenues par approximation. Nous ne considérons que deux états : soit deux pixels appartiennent au même plan, soit ils n'appartiennent pas au même plan. Nous ne considérons pas les variations d'orientation mais seulement l'appartenance des pixels à un même plan :

$$d_{\mathbf{n}}(\mathbf{p}_j, \mathbf{p}_i) = \begin{cases} 1 & \text{si } \mathbf{n}_i = \mathbf{n}_j \\ 0 & \text{sinon} \end{cases}$$

En conclusion, ce constructeur nécessite donc le choix de trois paramètres : le nombre de superpixels attendus, ainsi que λ et ϵ [4]. Le comportement des trois termes d_{s_0}, d_{c_0} et d_g de la distance proposée D_{SP} est illustré dans la figure 9.5. Nous observons bien que, comme attendu, c'est le terme d_g qui met en évidence les deux plans présents dans la zone étudiée.

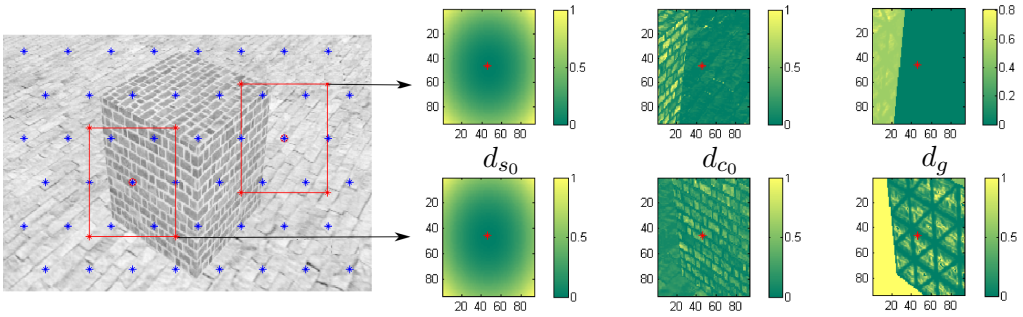


FIGURE 9.5. – Analyse du comportement des termes utilisés dans la distance d'agrégation proposée – L'image à gauche présente une image de synthèse où nous indiquons en bleu des germes pour estimer des superpixels. Nous étudions le comportement des termes pour deux cas de germes indiqués en rouge. L'interprétation des cartes de distances présentées est la suivante : plus la couleur du pixel est proche du vert, plus la distance est petite. Nous visualisons les valeurs obtenues, de la gauche vers la droite, pour d_{s_0}, d_{c_0} et d_g dans deux cas où les informations spatiale et colorimétrique ne sont pas suffisantes pour discriminer les deux surfaces adjacentes. Seule, la distance géométrique permet de distinguer les deux surfaces. Sur la première ligne : le germe se trouve sur le plan support (sol) alors que sur la deuxième, le germe appartient à une face de la boîte.

Analyse du comportement des superpixels géométriques proposés – Nous notons GEOM l'approche que nous avons proposée. Une comparaison visuelle des résultats entre SLIC et GEOM est donnée dans la figure 9.6. Une comparaison quantitative a également été réalisée afin d'évaluer la sur-segmentation à l'aide de l'erreur de sous-segmentation et de la valeur de rappel. Représentées sur les graphiques de la figure 9.7, ces mesures montrent que le rappel est le plus élevé et l'erreur de sous-segmentation la plus faible avec GEOM. Nous pouvons noter que pour un nombre de superpixels donné, par exemple entre 100 et 200, l'erreur relative vaut entre $\frac{0{,}5-0{,}43}{0{,}43} \simeq 0,16$ et $\frac{0{,}57-0{,}52}{0{,}52} \simeq 0,10$, c'est-à-dire que notre approche améliore de 10 à 16% le rappel de la sur-segmentation.

9.6. Segmentation à partir de superpixels géométriques

Dans cette partie nous exposons deux exemples de résultats de segmentation que nous avons obtenus en appliquant l'algorithme présenté dans la figure 9.3, en utilisant les superpixels géométriques que nous avons définis au § 9.5. Nous avons mis en place une vérité terrain en utilisant un outil de segmentation manuelle comme celui développé dans [Giro-i-Nieto 10]. Avec l'exemple d'*Oxford Merton*

4. Dans les expérimentations qui suivent, nous avons choisi de manière empirique $\lambda = 0.4$ et $\epsilon = 0.4$.

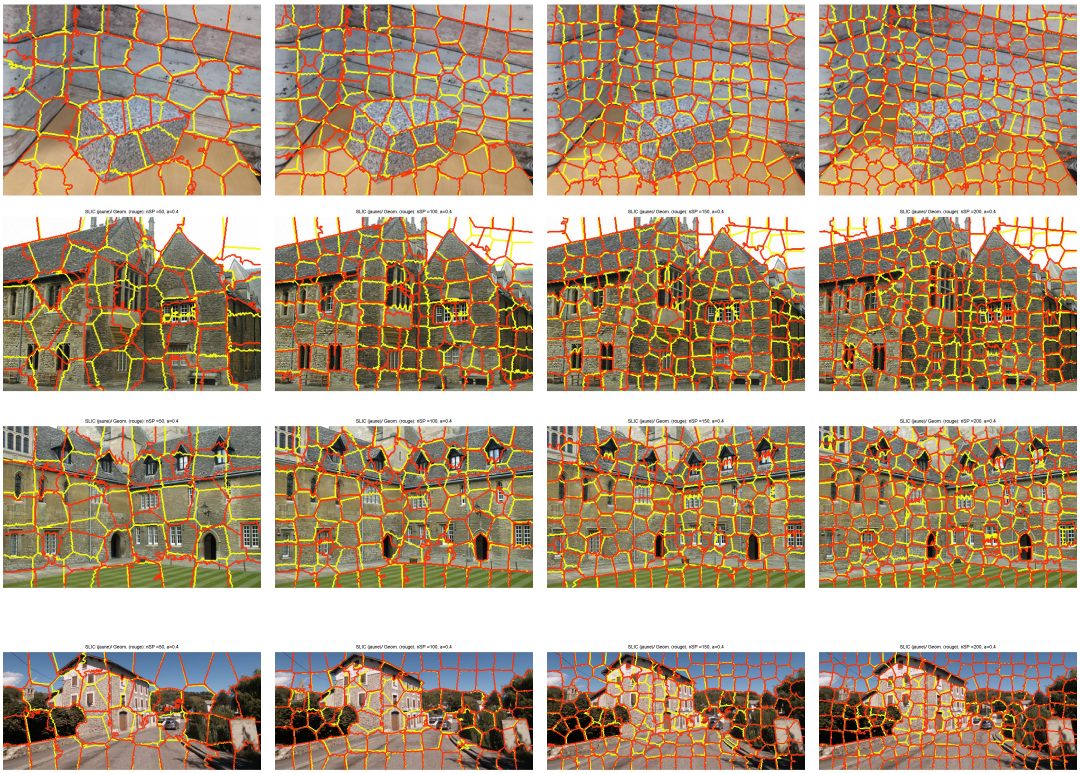


FIGURE 9.6. – Comparaison visuelle des approches SLIC et GEOM – Sur 4 images de notre corpus, nous visualisons les contours des superpixels obtenus avec l'approche SLIC (en jaune) et notre approche GEOM (en rouge), pour 50, 100, 150 et 200 superpixels.

College III, figure 9.8, l'évaluation au niveau de l'ensemble des bâtiments permet de constater l'apport de l'utilisation des superpixels. En effet, les taux de faux positifs obtenus par l'approche utilisant des superpixels géométriques sont plus élevées que pour SLIC. De plus, les erreurs sont plus faibles. Nous pouvons noter la difficulté à obtenir une annotation correcte sur les plans correspondant au toit car il y a peu de points 3D associés et ils sont généralement entâchés d'erreurs. Ces exemples ne sont qu'un échantillon de l'ensemble des scènes possibles. Ils nous permettent de démontrer l'utilité d'intégrer de l'information géométrique dans la construction de primitives intermédiaires : les superpixels géométriques, afin d'obtenir une meilleure segmentation sémantique en plan. En effet, l'approche de superpixels intégrant de l'information géométrique, permet de gérer des entités cohérentes (compacité, connexité) et de densifier l'étiquetage, tout en respectant la géométrie de la scène, en particulier lors du changement d'orientation des surfaces texturées. Cependant, les surfaces trop petites dans l'image, non ou mal représentées dans le nuage de point ne peuvent pas être traitées.

9.7. Conclusions sur la segmentation de scènes urbaines

La contribution principale de cette partie repose sur la mise au point d'un algorithme de construction de superpixels exploitant la planéité des régions de l'image. En effet, nous avons adapté l'approche développée dans [Achanta 12] au cas des scènes planaires. Cette contribution a été intégrée dans une chaîne de traitement complète permettant à partir de données redondantes (temporellement et spatialement) de proposer une segmentation sémantique de la scène.

Dans ce travail, nous avons extrait les différents éléments géométriques de la scène à partir des

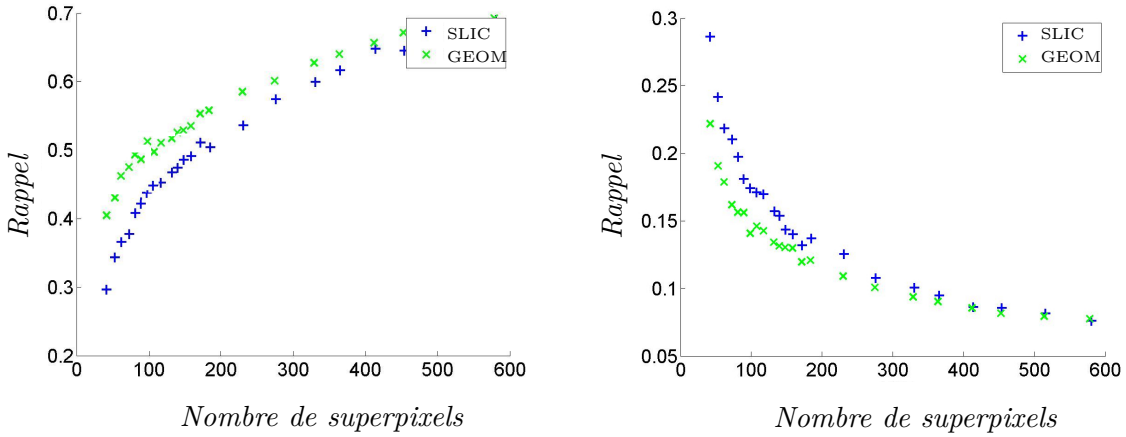
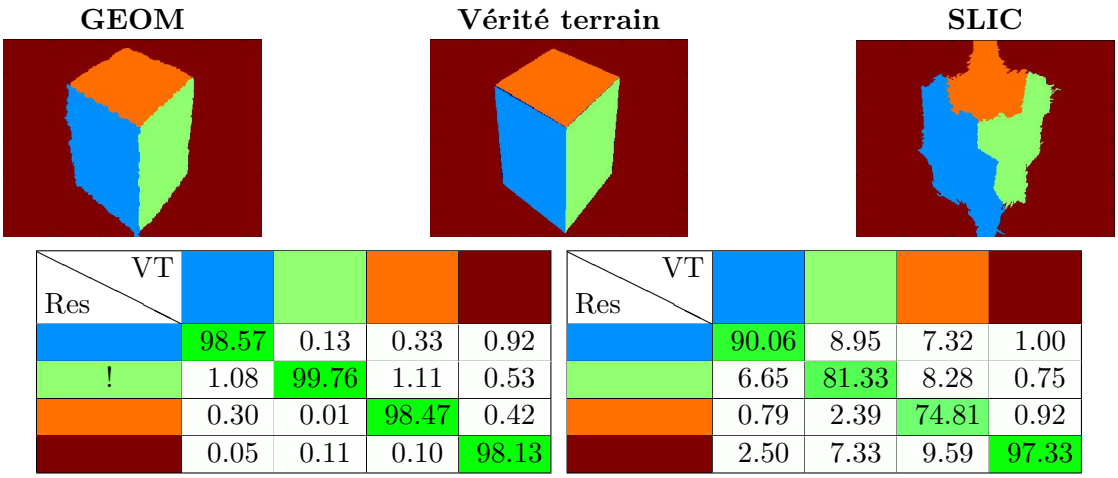


FIGURE 9.7. – Rappel et erreur de sous-segmentation obtenus pour SLIC et GEOM – Les courbes représentent le rappel (à gauche) et l'erreur de sous-segmentation (à droite).

techniques utilisées au sein de l'entreprise. Nous pourrions à présent envisager d'autres solutions, notamment en utilisant des outils de segmentation sémantique, comme cela a déjà été étudié dans un contexte de segmentation urbaine dans [Armagan 17]. Le contexte de notre étude comporte des similitudes avec celui exposé dans [Stekovic 20], mais ce travail a été réalisé pour des scènes d'intérieur. Ce qui est intéressant dans cette approche c'est qu'il est possible d'extraire la structure 3D de la scène avec une seule image. Nous pourrions également intégré ces résultats dans nos calculs préliminaires.

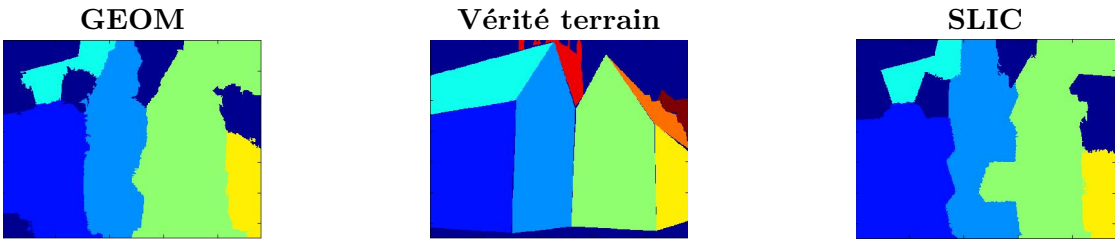
Dans le cadre de ce travail, nous avons aussi étudié la possibilité d'utiliser des points d'intérêt, comme ceux étudiés au chapitre 2. Nous pourrions enrichir l'approche de segmentation proposée en utilisant des points d'intérêt pour distinguer les différents éléments qui composent la scène et guider la sur-segmentation préliminaires.

Dans le travail présenté, nous avons pris en compte uniquement la redondance entre acquisitions d'une même séquence. Nous n'avons pas exploité la redondance des acquisitions dans le temps. Ainsi, nous souhaitons généraliser ce travail en exploitant mieux les aspects temporels, d'une part, et en exploitant cette redondance pour détecter les changements.



Res \ VT				
	98.57	0.13	0.33	0.92
!	1.08	99.76	1.11	0.53
	0.30	0.01	98.47	0.42
	0.05	0.11	0.10	98.13

Res \ VT				
	90.06	8.95	7.32	1.00
	6.65	81.33	8.28	0.75
	0.79	2.39	74.81	0.92
	2.50	7.33	9.59	97.33



Res \ VT					
	94.71	1.21	3.79	0.00	0.00
	0.84	93.69	6.55	0.02	0.00
	0.00	0.00	44.15	0.00	0.00
	0.00	3.97	0.00	99.54	2.32
	0.00	0.00	0.00	0.37	96.26
	4.45	1.13	45.50	0.07	1.42

Res \ VT					
	95.74	13.76	2.34	0.00	0.00
	0.00	78.04	5.57	11.86	0.00
	0.00	0.04	47.00	0.00	0.00
	0.00	6.56	0.00	87.90	1.900
	0.00	0.00	0.00	0.11	83.96
	4.26	1.60	45.09	0.13	14.14

FIGURE 9.8. – Segmentation sémantique en plan sur des données synthétiques et *Merton College III* – Nous comparons deux approches : GEOM et SLIC. Nous présentons la visualisation des cartes de résultats ainsi que des matrices de confusion associées aux classes considérées. Nous considérons deux niveaux de segmentation. Le premier correspond à la segmentation entre zones planes et zones non planes. Le deuxième niveau de segmentation considère les différents plans de la scène.

Chapitre 10.

Détection d'éléments fins par chemins minimaux

10.1. Objectif applicatif : maintenance des chaussées pour la sécurité routière

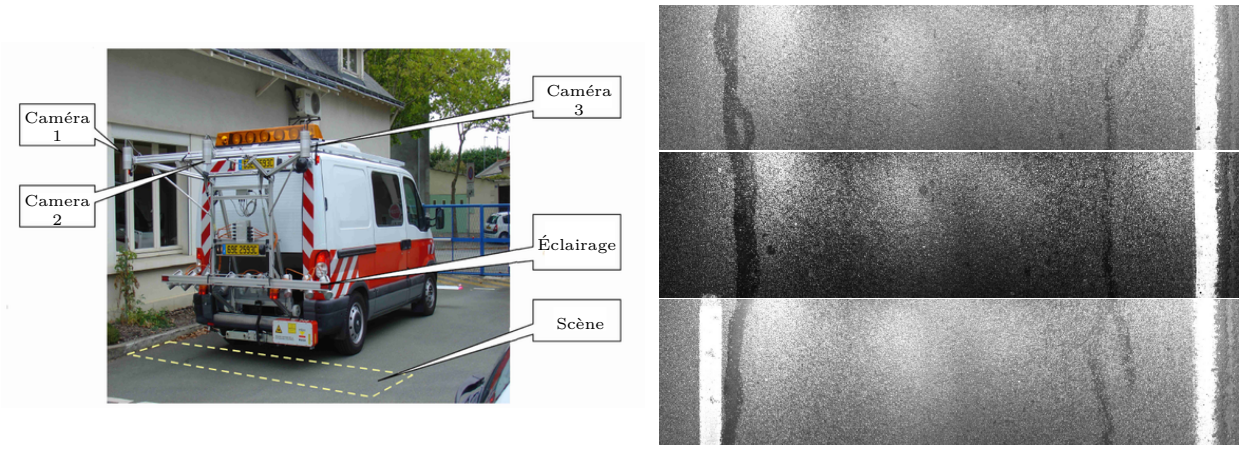
La surveillance automatique des routes est primordiale pour assurer la sécurité des usagers. Dans de nombreux pays, la détection des défauts de la chaussée est une étape incontournable pour inspecter l'état des routes. En effet, l'objectif est de détecter des défauts comme les arrachements ou encore les fissures, afin de planifier les tâches de maintenance et de réparation de la route. Le défaut le plus commun est celui des fissures. L'inspection visuelle humaine a été progressivement remplacée par l'acquisition automatique de données, si possible à la vitesse de circulation aussi bien avec des caméras CCD, *Charge-Coupled Device*, que des caméras CMOS, *Complementary Metal-Oxide-Semiconductor* [Ahmed 09].

Ainsi, de nombreuses techniques de détection automatique ont été développées afin de réaliser cette inspection, éventuellement en proposant une supervision humaine des résultats obtenus. Il est important de noter qu'il n'est pas nécessaire d'avoir un traitement en temps réel mais, en revanche, il est important qu'une base de données effectuée dans la journée puisse être traitée en un temps raisonnable pour fournir un résultat au gestionnaire le lendemain des acquisitions. En ce qui concerne les systèmes d'acquisition, nous pouvons citer les travaux de [Gavilán 11] qui fournissent un historique des systèmes développés.

Il est important de distinguer les techniques pour classer des images (avec ou sans défaut) des approches pour détecter la forme exacte d'un défaut. En France, détecté la localisation exacte de la fissure ainsi que sa forme est important pour évaluer la gravité du défaut. Ainsi, dans nos travaux de recherche, nous avons essentiellement traité le cas de la détection précise du défaut (à l'exception du travail publié dans [Cord 11] où nous nous sommes concentrés sur la classification d'images de défauts). Dans ce chapitre, nous présentons donc la manière de traiter ce cas de segmentation particulier : un objet fin et faiblement représenté (la fissure) dans un environnement fortement texturé (la chaussée), comme illustré dans la figure 10.1, en (b), avec le système d'acquisition présenté en (a).

10.2. Problématique de la détection de fissures de la chaussée

En imagerie 2D, le plus souvent avec un capteur CCD, le défaut de fissure est associé à des pixels d'intensité plus faibles que les autres pixels de la chaussée alors qu'en imagerie 3D, le défaut est bien évidemment lié à une variation d'élévation. L'imagerie 3D, dans ce contexte, est une technique plus délicate, c'est-à-dire qui nécessite souvent une vitesse de circulation plus faible et qui a un coût plus élevé. Ce ne sera donc pas le propos de ce chapitre mais nous pouvons penser qu'avec la démocratisation des lasers ou des LIDAR, *LIght Detection And Ranging* [Urano 19] cette technique



(a) (b)

FIGURE 10.1. – Système d'acquisition utilisé pour détecter des fissures – En (a), nous illustrons que le système d'acquisition utilisent 3 capteurs positionnés à 1 mètre du sol et en (b) nous donnons l'exemple de 3 triplets d'image obtenus. La surface présentée contient 2 réparations de fissures verticales. Dans certaines conditions d'acquisition, les 3 capteurs n'ont pas toujours les mêmes réglages et l'illumination globale peut être assez différente. Cette variabilité peut induire des artefacts assimilables à des défauts de chaussée. Ces problèmes sont partiellement résolus par une étape de pré-traitement. Cette étape de pré-traitement corrige les défauts qui peuvent apparaître aux jonctions des images issues des 3 capteurs. Nous n'avons aucune maîtrise/emprise sur ce pré-traitement.

devrait se développer. Ainsi, nous nous concentrons sur la détection des défauts dans des images en niveau de gris mais nous n'avons pas de doute que ces techniques peuvent se généraliser au 3D. Détecter automatiquement ce type de défaut est un véritable défi car dans la plupart des cas, le défaut est fin, irrégulier et assez proche des détails de la texture du revêtement [Oliveira 13, Li 11], cf. figure 10.1.(b).

10.3. Résumé des contributions en segmentation de fissures

De nombreuses approches ont été proposées dans la littérature, des plus simples s'appuyant uniquement sur une information photométrique [1] aux plus complexes prenant en compte des aspects géométriques de la fissure. Nous avons présenté dans [Chambon 11d] un état de l'art des systèmes d'acquisition et des méthodes de détection. Parmi les différentes familles d'approches que nous avons distinguées, nous avons introduit les contributions suivantes :

- une méthode par modélisation markovienne [Chambon 09, Chambon 10c] ;
- une approche s'appuyant sur des lignes de partage des eaux multi-échelle [Coudray 10] ;
- une méthode en utilisant les outils de stéréovision [Chambon 10a] ;
- une méthode par détection de points d'intérêt et contours actifs [Chambon 11a] ;
- une méthode s'appuyant sur une approche de classification de type *Adaboost* [Cord 11] (dans ce cas, nous ne détectons pas la forme de la fissure mais nous classons des sous-images en fonction de la présence d'un défaut, en prenant toutes les catégories de défauts et non juste les fissures).

Toutes ces méthodes nous ont permis d'obtenir des résultats surpassant l'état de l'art, mais, dans ce chapitre, nous détaillerons uniquement l'approche qui nous a permis d'obtenir les meilleurs résultats, à savoir celle s'appuyant sur les chemins minimaux [Amhaz 14, Amhaz 16].

1. L'hypothèse utilisée est celle qu'un pixel de fissure a une intensité lumineuse plus faible que les autres pixels de la scène.

10.4. Approches existantes en segmentation de fissures

10.4.1. Approches photométriques

Parmi les approches photométriques, largement employées dans la littérature, un simple seuillage a souvent été employé. Ce seuil peut être global [Acosta 92], local adaptatif [Salari 11, Tang 13] ou par hystérésis [Hu 10]. Pour déterminer ce seuil de nombreuses approches d'optimisation ont été utilisées, des plus classiques, comme l'algorithme d'Otsu [Otsu 79] aux plus récentes, comme celles faisant intervenir des outils de la logique floue [Cheng 99] ou des algorithmes génétiques [Tomikawa 99]. Toutes ses approches montrent des performances limitées car *a priori*, dans ce contexte, un simple seuil ne permet pas de distinguer les deux populations en présence dans l'image. Ainsi, des approches plus complexes ont introduit la notion d'apprentissage supervisé en supposant un mélange de gaussiennes et en estimant les paramètres de ces gaussiennes [Koutsopoulos 93], en utilisant un réseau de neurones [Cheng 01, Bray 06a] ou une classification bayésienne [Oliveira 08]. Pour être performantes, toutes ces approches doivent faire appel à une étape d'apprentissage importante, assez difficile à réaliser encore récemment, car il y avait peu d'images disponibles.

10.4.2. Approches photométriques et géométriques

D'autres approches ajoutent un critère géométrique pour prendre en considération des ensembles de pixels, localement. Nous pouvons citer les approches s'appuyant sur des opérateurs de morphologie mathématique comme une dilation ou une érosion. L'objectif de ces méthodes est de tenter de réduire les discontinuités de la détection des fissures, d'une part, et de réduire les fausses détections, d'autre part [Tanaka 98, Na 12, Tang 13]. Ce type d'approches nécessite le choix de nombreux paramètres qui sont délicats à déterminer automatiquement. Les fissures peuvent avoir des longueurs et des largeurs très variables, c'est pourquoi dans la littérature, il existe également des approches multi-échelles. Nous pouvons citer, entre autres, les approches par décomposition en ondelettes [Zhou 05, Subirats 06]. La difficulté majeure de toutes ces approches concerne la façon de fusionner les résultats à différentes échelles.

10.4.3. Approches introduisant la notion de topologie

Les approches les plus avancées introduisent des **critères géométriques sur la topologie des fissures pour compléter le critère photométrique utilisé**. Plus précisément, dans ces approches les fissures sont considérées comme des ensemble de pixels contigus et de faible intensité lumineuse. Par exemple, les approches s'appuyant sur des champs de Markov [Delagnes 95] favorisent la continuité entre pixels fissures en utilisant une approche statistique. Le problème de ces approches est qu'elles sont coûteuses en terme de temps de calcul et elles nécessitent de fixer un certain nombre de paramètres qui doivent être appris. À ce stade, il est important de remarquer que ce problème de détection de fissures dans des images de chaussée peut être assimilé à d'autres problèmes connus en recherche comme, par exemple : les problèmes de suivi dans des vaisseaux, avec des images IRM, ou encore la recherche des contours des objets pour effectuer une segmentation. Ainsi, dans la littérature, certaines de ces approches ont simplement été adaptées au problème de la détection de fissures, comme l'approche tenseur [Zou 12], ou les approches s'appuyant sur des contours actifs ou des contours géodésiques [Kaul 12, Tang 13]. Dans ce dernier cas, une étape importante est la recherche de chemins minimaux ou de chemins les plus courts, comme dans [Nguyen 11]. Il s'agit des méthodes que nous avons choisies de d'étudier et d'améliorer.

10.4.4. Approches de segmentation de fissures par chemins minimaux

De manière générale, le problème de recherche de chemin minimal correspond à trouver le meilleur chemin entre pairs de nœuds dans un graphe en minimisant un coût qui généralement est la somme

des coûts de toutes les arêtes qui composent le chemin. En traitement d'images, le coût comprend en général deux termes. Le premier terme, appelé énergie externe ou terme d'attache aux données, prend en considération l'intensité des pixels le long du chemin alors que le second, appelé énergie interne ou terme de régularisation, évalue localement la régularité du chemin [Kaul 12, Mille 12].

Un algorithme efficace et très célèbre pour résoudre le problème de recherche de chemins minimaux est l'algorithme de [Dijkstra 59]. Nous pouvons également citer l'approche de [Sethian 95], l'algorithme de *fast marching*, qui permet de minimiser l'énergie dans le cas où l'énergie interne s'appuie sur la longueur du chemin. C'est grâce à l'algorithme de Dijkstra qu'il est possible d'obtenir une information géométrique de haut niveau tout en conservant des temps de calcul raisonnables et un coût en mémoire acceptable. Toutefois, pour appliquer la recherche de chemins minimaux, nous avons besoin de connaître au préalable les positions des deux points à relier. Malheureusement, ce genre de connaissance n'est pas disponible dans un contexte entièrement non supervisé. De plus, les fissures ne peuvent pas être considérées comme de simples lignes. Il s'agit d'un réseau complexe à géométrie inconnue. Afin de pallier les difficultés que nous venons de décrire, la plupart des approches s'appuyant sur une recherche de chemins minimaux reposent sur ces étapes [Chambon 11a, Nguyen 11, Zou 12, Amhaz 16] :

- Sélection des points de l'image les plus sombres en supposant qu'ils appartiennent très probablement à une fissure ;
- Calculs de tous les chemins minimaux possibles entre deux pairs de points voisins.

Pour la première étape, dans [Nguyen 11], les auteurs considèrent que chaque point de l'image peut potentiellement être un point fissure. Pour la seconde étape, [Chambon 11d, Zou 12] utilisent un terme de régularisation pour favoriser une trajectoire lisse alors que [Nguyen 11, Amhaz 14] utilisent un coût qui est la somme des intensités des pixels qui composent le chemin, sans aucune énergie interne. Nous pouvons considérer que la fissure est trop irrégulière pour envisager l'utilisation d'un terme de régularisation et donc, nous nous concentrons sur le deuxième type d'approche.

La contribution la plus significative de notre travail est de considérer que la fissure peut être caractérisée uniquement sur sa photométrie et sur une hypothèse assez minimaliste : les pixels qui appartiennent à une fissure forment un chemin continu avec une forme arbitraire. Les travaux les plus proches de ce travail sont ceux de [Nguyen 11] et la différence majeure entre ces travaux et les nôtres réside dans les hypothèses faites sur la forme de la fissure. En effet, dans [Nguyen 11], les auteurs supposent qu'une fissure a obligatoirement une direction privilégiée. Cette hypothèse permet de réduire fortement la recherche de chemins minimaux. Au contraire, dans nos travaux, nous ne faisons aucune restriction sur la forme et l'orientation des fissures mais nous ajoutons des contraintes sur les pixels sélectionnés au départ pour calculer les chemins minimaux. De plus, nous avons ajouté des étapes de post-traitement pour fusionner les chemins et calculer la largeur des défauts, en ajoutant des pixels, en fonction de leur position et de leur intensité. Un intérêt important de l'approche proposée est qu'elle est totalement non-supervisée, à chaque étape de l'algorithme. Tous les paramètres de cette approche seront présentés et nous pouvons déjà mettre en avant le fait que pour tous les tests de performance que nous avons réalisés, ces paramètres ont des valeurs par défaut et fixes, c'est-à-dire, non manuellement optimisés pour chaque image.

Ainsi, nous allons présenter les détails sur le principe de calcul des chemins minimaux et notamment leur utilisation dans ce contexte applicatif précis, cf. § 10.5. Cette présentation nous permettra d'introduire l'algorithme proposé, dont le principe repose sur les chemins minimaux, et qui fait intervenir trois étapes, cf. § 10.6. Dans cette section, nous présenterons également les deux étapes de post-traitement permettant d'affiner la détection. Cette approche a été validée expérimentalement sur un ensemble de données que nous avons construits et partagés avec la communauté. L'intérêt de cette validation est d'avoir permis la comparaison avec cinq algorithmes différents : quatre algorithmes que nous avons proposés mais également l'approche la plus proche de nos travaux [Nguyen 11], cf. § 10.7.

10.5. Chemins minimaux adaptés aux images de fissures

En théorie des graphes, la recherche de chemin le plus court consiste à estimer un chemin entre deux sommets tel qu'il minimise un coût. Généralement, ce coût est la somme des poids des arcs qui composent le chemin. Dans le cas des fissures, nous ne cherchons pas le chemin le plus court, c'est-à-dire la fissure la plus courte, mais nous cherchons la fissure qui minimise un coût qui ne prend pas en compte une distance géométrique. C'est pourquoi, nous avons choisi de parler de chemins minimaux plutôt que de chemins les plus courts et c'est le terme que nous utiliserons pour la suite.

En traitement d'images, nous pouvons considérer une image comme un graphe connecté de pixels et nous associons aux sommets des poids qui correspondent aux niveaux de gris des pixels. L'extraction de chemins minimaux permet de traiter de nombreux problèmes en imagerie [Sun 05a], comme la segmentation d'images médicales [Deschamps 01], l'extraction de réseaux routiers dans des images satellites [Cohen 96] ou encore la reconnaissance d'objets [Yan 06].

Pour mettre en œuvre un algorithme d'estimation de chemins minimaux, il est nécessaire de choisir ces deux éléments :

1. Le **coût** utilisé – Ce coût doit nous permettre de faire le tri entre les solutions optimales pour résoudre le problème et les solutions les plus défavorables. En traitement d'images, beaucoup d'algorithmes considèrent uniquement les niveaux de gris ou les couleurs. De plus, de nombreuses méthodes introduisent un terme de régularisation afin de minimiser à la fois une distance, un terme de régularité et un terme de longueur [Kaul 12, Mille 12]. Utiliser un tel coût est totalement approprié pour des applications comme la segmentation de vaisseaux sanguins en imagerie médicale car les objets à détecter ou à segmenter ont une forme régulière avec une longueur constante. Dans le cas des fissures de chaussée, il n'y a aucune régularité et bien évidemment, la longueur est très variable. En effet, nous pouvons affirmer que la fissure a une forme chaotique.
2. La méthode d'**optimisation** utilisée pour estimer le chemin qui minimise le coût – Les approches les plus connues sont celles de [Dijkstra 59] ou le *fast marching* [Sethian 95].

La façon de construire le coût influence le choix de la méthode d'optimisation et *vice versa*. Dans le cas des contours actifs, l'objectif est de trouver une courbe qui correspond au contour d'un objet d'intérêt en minimisant de manière itérative un coût ou une énergie. Les contours actifs sont détaillés au § C.1 et nous reprenons donc l'expression de l'énergie E définie pour la courbe $\mathcal{C}$ [Kass 88] :

$$E(\mathcal{C}) = \int_0^L \Big(w_1 \|\mathcal{C}'(v)\|^2 + w_2 \|\mathcal{C}''(v)\|^2 + \mathcal{P}(\mathcal{C}(v)) \Big)\, dv, \tag{10.1}$$

où $c(v)$ est la courbe dans l'image, en 2D, $[0, L]$ est son domaine de définition et L, la longueur de la courbe. Nous rappelons que les deux premiers termes correspondent aux *forces internes*, liées à la régularité de la courbe alors que le troisième terme correspond aux forces externes $\mathcal{P}$. Afin de refléter au mieux les caractéristiques d'une fissure, il paraît naturel d'utiliser une force externe faisant intervenir uniquement les niveaux de gris des pixels. De plus, il n'y a pas de contrainte de régularité ou de lissage, puisque, par définition, une fissure à une forme chaotique. En conséquence, dans ces travaux nous utilisons la fonction de coût c donnée par :

$$c(p_{ij}) = \frac{1}{\#(p_{ij})} \sum_{\mathbf{p}=i}^{j} I(\mathbf{p}), \tag{10.2}$$

où p_{ij} est l'ensemble des points avec $\mathbf{p}_i$ la source et $\mathbf{p}_j$ la destination. Le terme $\#(p_{ij})$ correspond à la longueur de ce chemin. Nous avons envisagé l'utilisation d'un terme de régularisation pour limiter les changements topologiques, comme dans [Chambon 11d]. Cependant, le coût défini par (10.2) est celui que nous avons retenu car cette formulation permet d'envisager tous les chemins possibles, chaotiques ou non. Pour minimiser ce coût, deux techniques sont majoritairement utilisées : l'algorithme de

Dijkstra [Dijkstra 59] ou une approche par *fast marching* [Cohen 96]. Elles sont utilisées indifféremment dans la littérature [Chambon 11d, Nguyen 11].

L'aspect important à retenir sur ces approches est que, pour chacun des algorithmes proposés, un ensemble de chemins minimaux est estimé et, parmi tous ces chemins, seul un sous-ensemble sera conservé. Il y a donc ainsi deux aspects importants à considérer :

(1) Quels sont les chemins estimés au départ ? Ou plus précisément, quels points de l'images sont utilisés comme début et fin de chemin ? Plus le nombre de points germes est important et plus les calculs seront importants.

(2) Quel sous-ensemble de chemins est conservé ?

Sur le premier point, certaines approches considèrent tous les points de l'image, comme dans [Nguyen 11]. Dans ce cas, pour réduire les calculs, les auteurs imposent une contrainte de direction (seules 4 directions sont autorisées) et de taille de chemin (ils ne peuvent pas dépasser 30 pixels de longueur). Ces contraintes nous semblent inappropriées étant donné que la fissure a une forme chaotique et est de longueur variable. C'est pourquoi nous avons proposé une approche s'appuyant sur une détection de points d'intérêt [Chambon 11a]. Ceci peut s'avérer coûteux en terme de temps de calcul et les points fissures ne correspondent pas toujours à la notion de points d'intérêt classique, comme nous avons pu étudier dans la partie I. En ce qui concerne le second point, dans [Nguyen 11], les chemins minimaux estimés en chaque pixel de l'image donne un score de probabilité d'être un pixel appartenant à une fissure. Ce score est ensuite utilisé pour réaliser la segmentation. Cette façon de procéder ne permet plus d'exploiter la notion de continuité dans le tracé de la fissure. C'est la raison pour laquelle, nous avons orienté nos travaux sur l'utilisation des coûts des chemins entre points germes [Amhaz 14]. Plus précisément, nous conservons des chemins, en fonction de leur coût, et non des pixels pris un par un.

10.6. Sélection de chemins minimaux pour la détection de fissures

L'approche que nous proposons est appelée MPS, *Minimal Path Selection*. Nous faisons l'hypothèse simple qu'un chemin avec un coût faible est probablement un chemin appartenant à une fissure. De plus, nous analysons l'image de deux façons : de manière locale (cela correspond à la détection des points germes) et de manière globale (cela correspond à la sélection de l'ensemble de chemins minimaux appartenant à une fissure). Dans la suite, nous supposons que l'image n'a pas de défauts d'éclairage (c'est-à-dire que nous supposons que tous les défauts d'éclairage ont été pré-traités si besoin). Notre approche comporte également deux post-traitements, le premier pour réduire certains artefacts, comme les barbules et les boucles, et le second pour estimer l'épaisseur de la fissure. La figure 10.2 nous permet de présenter les résultats intermédiaires obtenus pour cette approche.

Les trois étapes principales de l'algorithme MPS sont présentées dans l'algorithme 3. Pour l'étape (C_1) de sélection de l'ensemble des points amorces $\mathcal{E}$, notre objectif est d'avoir le plus de points possibles à l'intérieur des fissures à détecter. Pour cela, nous avons simplement pris les minima locaux dans une grille régulière de l'image où chaque cellule a une taille $P \times P$. Toutefois, certains de ces candidats sont ensuite supprimés si leurs niveaux de gris est au dessus d'un seuil $T_e = \mu - k_e\sigma$, avec μ et σ la moyenne et l'écart-type des niveaux de gris de toute l'image. Le choix de k_e sera expliqué au § 10.8.1.

Pour l'étape (C_2) d'estimation des chemins minimaux notés p_{ij}, nous avons simplement estimé tous les chemins possibles entre pairs d'amorces (e_i, e_j) en utilisant l'algorithme de [Dijkstra 59]. Nous savons que chaque amorce, e_i, détectée dans une cellule de la grille est entourée d'au plus 8 cellules dans son voisinage. Nous notons $\mathcal{V}(e_i)$ l'ensemble des amorces contenues dans ces cellules adjacentes et, pour estimer les chemins minimaux, nous utiliserons donc une sous-image de taille $3P \times 3P$. Rappelons nous qu'en utilisant le coût défini en (10.2), nous n'imposons aucune contrainte sur la forme et la longueur des chemins.

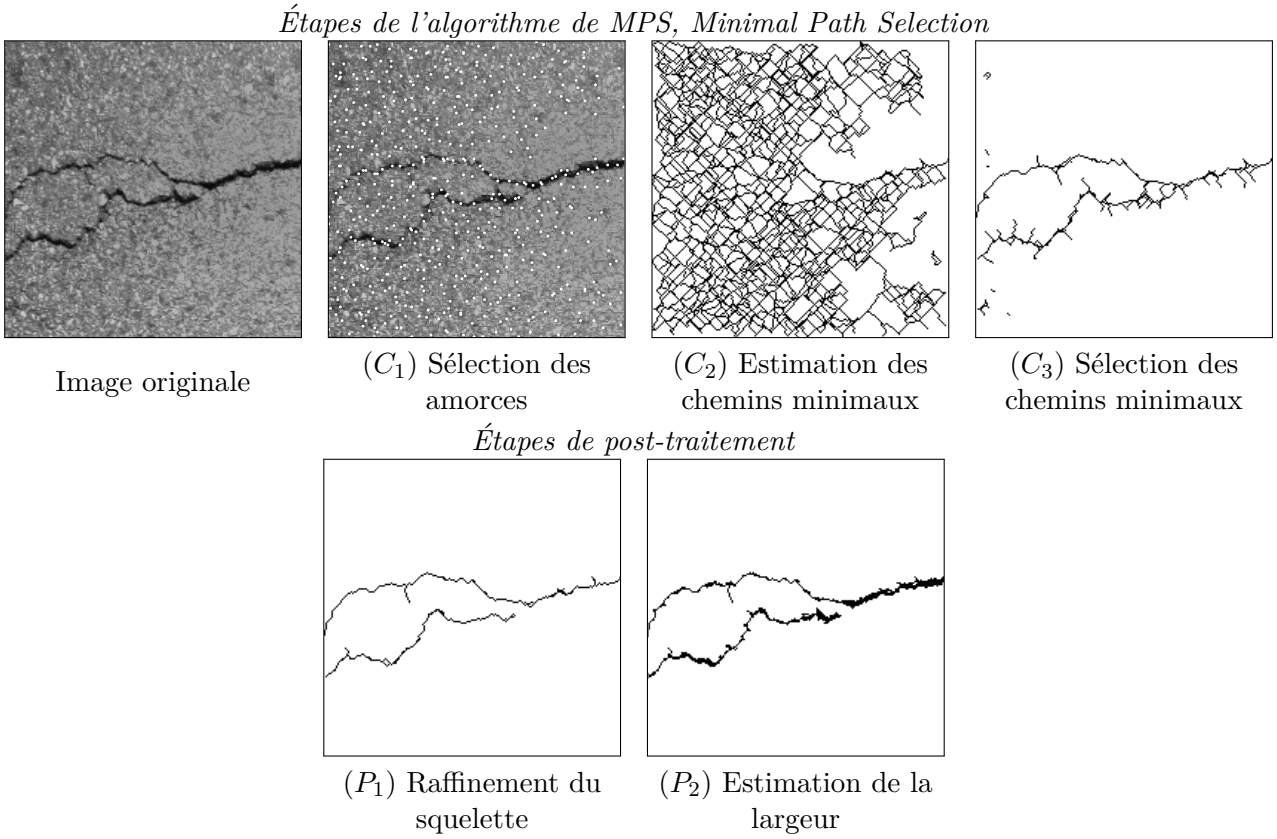


FIGURE 10.2. – Résultats intermédiaires de l'algorithme de segmentation de fissures par recherche de chemins minimaux – Les étapes (C_1) à (C_3) sont données dans l'algorithme 3 alors que les post-traitements (P_1) et (P_2) sont présentés dans l'algorithme 4.

Nous savons que parmi tous les chemins minimaux estimés une faible proportion se trouve réellement à l'intérieur d'une fissure. Nous savons également que les chemins qui doivent être gardés doivent *a priori* correspondre aux chemins les plus sombres. C'est la raison pour laquelle, pour l'étape (C_3) de sélection des chemins minimaux, nous avons décidé d'appliquer un seuil sur la fonction de coût (10.2). Nous avons choisi ce seuil ainsi : $T_c = \mu_c - k_c\sigma_c$, où μ_c est la moyenne et σ_c, l'écart type. Le choix de k_c sera expliqué au § 10.8.1. Ainsi, comme nous pouvons le constater dans la figure 10.2(C_3), cette étape permet de faire converger le résultat vers le trace de la fissure. Toutefois, il reste des artefacts que nous tendons de prendre en compte dans les post-traitements.

Les deux types de post-traitement proposés concernent : la suppression des chemins erronés et le calcul de la largeur de la fissure. Ces étapes sont présentées dans l'algorithme 4.

L'objectif de l'étape (P_1) est d'éliminer tous les défauts potentiels de la détection initiale, c'est-à-dire : les petits chemins isolés ou les chemins en boucles. Pour le premier point nous appliquons un simple seuillage sur la longueur de la fissure. Le choix du seuil T_s est expliqué au § 10.8.1. Pour les boucles, il est important de comprendre que leur origine est liée à la sélection d'un point de départ ou d'arrivée qui n'appartient pas à la fissure. Ainsi, pour traiter ce problème, nous avons décidé de distinguer les différentes parties qui composent une boucle en identifiant les points de la fissure où il y a une jonction. Par exemple, une fissure en forme de A, H ou P, sera re-découpée en 4, 5 et 2 segments respectivement. Ensuite, un nouveau seuillage des coûts de chacun de ces nouveaux chemins est appliqué en utilisant le même seuil T_c qu'à l'étape (C_3).

Une fois la topologie de la fissure estimée, nous souhaitons, dans une dernière étape (P_2), calculer l'épaisseur de cette fissure. Pour cela nous effectuons une propagation de la fissure en utilisant

Algorithme 3 : Algorithme de segmentation de fissures par sélection de chemins minimaux – Il s'agit du pseudo-code des 3 étapes principales de l'algorithme *Minimal Path Selection* (MPS).
I_P^i : Sous-image (image) numéro i de taille $P \times P$
$\mathcal{V}(\mathbf{a})$: Liste des amorces dans le voisinage de $\mathbf{a}$ (maximum 8)
p_{ij} : Chemin minimal entre $\mathbf{p}_i$ et $\mathbf{p}_j$ avec le coût $c(p_{ij})$

Fonction Détection_par_sélection_de_chemins_minimaux (I)
Entrées : I : image pour laquelle les fissures doivent être détectées
Sorties : I_f : image binaire indiquant les points fissures
$\mathcal{S}$: ensemble des chemins correspondant à une fissure
`/*` (C_1) `- Sélection des amorces` $\mathcal{E}$ `*/`
`/* Sélection locale des points avec les niveaux de gris les plus faibles */`
$\mathcal{M} = \{\mathbf{p}_i \mid \mathbf{p}_i = \text{argmin}_{\mathbf{p} \in I_P^i}(I(\mathbf{p}))\}$
Estimation du seuil T_e
Sélection de $\mathcal{E} = \{\mathbf{e}_i \in \mathcal{M} \mid I(\mathbf{e}_i) \leq T_e\}$
`/*` (C_2) `- Estimation des chemins minimaux` p_{ij} `*/`
pour chaque *paire d'amorces* $(\mathbf{e}_i, \mathbf{e}_j)$ *avec* $\mathbf{e}_i \in \mathcal{E}$, $\mathbf{e}_j \in \mathcal{E}$ *et* $\mathbf{e}_j \in \mathcal{V}(\mathbf{e}_i)$ **faire**
  Estimation du chemin minimal p_{ij} et de son coût $c(p_{ij})$
`/*` (C_3) `- Sélection des chemins minimaux` $\mathcal{S}$ `*/`
Estimation du seuil T_c
Estimation du squelette de la fissure $\mathcal{S} = \{p_{ij} \mid c(p_{ij}) \leq T_c\}$

également un seuillage. Celui-ci est adaptatif est le calcul du seuil est réalisé ainsi : $T_w = \mu_w + k_w \sigma_w$, où μ_w et σ_w sont la moyenne et l'écart-type des points couramment détectés dans la fissure. Le choix pour k_w sera présenté au § 10.8.1. La propagation est un processus itératif qui s'arrête lorsqu'il n'y a plus de point ajouté.

10.7. Protocole d'évaluation pour la détection de fissures

10.7.1. Images de référence testées

L'approche introduite a été validée sur une image de synthèse (texture de chaussée aléatoire en suivant la distribution des niveaux de gris d'une image réelles et ajout d'une fissure avec une forme chaotique, avec ou sans ramifications et d'épaisseur variant entre 1 et 2 pixels), cf. figure 10.6, pour tester les performances et limites dans un contexte entièrement contrôlé. Ensuite, nous avons complété les tests avec 66 images réelles réparties en 5 groupes qui correspondent à des acquisitions sur des systèmes différents (un système statique et 4 systèmes dynamiques) et avec des chaussées de différents types. Pour les systèmes d'acquisition dynamiques comme celui utilisé à l'IFSTTAR et présenté dans la figure 10.1, les défauts liés aux conditions d'éclairage non-uniformes ont été pré-traités. Estimer manuellement le tracé de la fissure est une tâche délicate et très variable d'un utilisateur à un autre. C'est la raison pour laquelle, dans [Chambon 11d], nous avons défini des segmentations de référence en utilisant 4 opérateurs et en fusionnant les 4 résultats, sur la première base de données que nous avions obtenues. Par la suite, lorsque nous avons pu obtenir plus d'images, nous avons opté pour une segmentation semi-automatique en cliquant sur des points d'intérêt dans la fissure et en appliquant une recherche du chemin minimal, suivant l'approche de [Dijkstra 59].

Algorithme 4 : Post-traitement des fissures détectées par sélection de chemins minimaux – Il s'agit du pseudo code des raffinements réalisés après la détection afin d'éliminer les défauts de ramifications erronées, d'une part, et de calculer la largeur des fissures, d'autre part. Nous avons repris les notations utilisées pour l'algorithme 3. Nous précisons ici que $\mathcal{V}(\mathcal{S})$ correspond au voisinage du squelette $\mathcal{S}$ (8-connectivité).

Fonction Post_Traitement $(I_f, \mathcal{S})$
Entrées : I_f : image binaire indiquant les points fissures
$\mathcal{S}$: ensemble des chemins correspondant à une fissure
Sorties : I_f et $\mathcal{S}$ post-traités
`/* (`P_1`) - Affinement du squelette` $\mathcal{S}$ `*/`
`/* Suppression des chemins courts isolés */`
Estimation du seuil T_s
$\mathcal{S} = \{p_{ij} \in \mathcal{S} \mid \#(p_{ij}) \leq T_s\}$
`/* Suppression des ramifications et des boucles */`
Détection des extrémités et des jonctions $\mathbf{e}_i^m$
Ajustement des chemins p_{ij} entre $(\mathbf{e}_i^m, \mathbf{e}_j^m)$ avec $\mathbf{e}_j^m \in \mathcal{V}(\mathbf{e}_i^m)$
Mise à jour des coûts $c(p_{ij})$
Amélioration de $\mathcal{S} = \{p_{ij} \in \mathcal{S} \mid c(p_{ij}) \leq T_c\}$
`/* (`P_2`) - Détection de la largeur */`
Estimation du seuil T_w
répéter
  Sélection de $\mathcal{S}_v = \{\mathbf{v}_i \in \mathcal{V}(\mathcal{S}) \mid I(\mathbf{v}_i) \leq T_w\}$
  $\mathcal{S} = \mathcal{S} \cup \mathcal{S}_v$
jusqu'à $\mathcal{S}$ *inchangé* ;

10.7.2. Choix des méthodes comparées

Il est important de constater qu'il y a peu de publications qui fournissent une étude comparative des différentes approches. Cela est essentiellement dû au fait que beaucoup de techniques développées sont propriétaires et donc il n'y a pas de description précise ni de code disponible. Dans ces conditions, nous avons essentiellement comparé les différents types d'approches proposées. Toutefois, nous avons implémenté l'approche de [Nguyen 11] car c'est la plus proche de nos travaux. Ainsi, nous exposons les résultats pour :

- **Des méthodes basées modèle de Markov** où un critère photométrique est utilisé (nous supposons simplement que les pixels de fissures sont plus sombres que les autres) ainsi qu'un critère géométrique locale (localement, la fissure est un segment). Plus précisément, les résultats sont exposés pour :
  - M_1 [Chambon 09] où, d'une part, un filtrage multi-échelle 2D adaptatif dans quatre directions différentes permet de renforcer le contraste, et, d'autre part, une segmentation Markovienne est réalisée. Le modèle utilisé consiste à étudier des blocs de taille 3×3 et à envisager 4 configurations possibles (4 segments dans 4 directions). Les résultats obtenus à toutes les échelles sont ensuite fusionnés pour donner le résultat global.
  - M_2 [Chambon 11d] est une amélioration de [Chambon 09] où 16 configurations sont envisagées.
- **Des approches par recherche de chemins minimaux** qui correspondent aux approches les plus proches de MPS. Ainsi, nous présentons les résultats pour :
  - GC, *Geodesic Contour* [Chambon 11a] qui utilise les contours géodésiques entre points d'intérêt automatiquement sélectionnés. L'originalité de cette approche est de considérer des points d'intérêt suivant un critère d'auto-corrélation directionnel.

- FFA, *Free Form Anisotropy* [Nguyen 11] qui s'appuie sur l'estimation de tous les chemins minimaux possibles de 30 pixels de longueur et dans 4 directions différentes. Ensuite, les pixels appartenant à une fissure sont déterminés en effectuant un seuillage des coûts des chemins estimés.
- MPS0 [Amhaz 14] est une version préliminaire à la version MPS présentée dans ce chapitre.

10.7.3. Critère d'évaluation

Nous avons utilisé le critère *DSC*, cf. § 8.1, équation (8.3). Ce critère a la particularité d'être particulièrement adapté dans le cas d'une population faiblement représentée par rapport au reste de l'image à segmenter. Ce qui est le cas avec les fissures que nous souhaitons segmenter et qui représentent la plupart du temps moins de 5% de l'image. En pratique un critère *DSC* au delà de 0.7 indiques un résultat entièrement satisfaisant [Zijdenbos 94]. Enfin, sachant que la construction de la vérité terrain est très délicate, nous avons autorisé une erreur de 2 pixels dans le voisinage de la région segmentée dans la vérité terrain.

10.8. Étude des performances

10.8.1. Influence des paramètres

Ainsi, nous avons besoin de fixer 5 paramètres : La taille des sous-images de l'image à considérer, P, pour l'analyse locale, étape (C_1), les seuils T_e, étape (C_1), T_c, étape (C_2), T_s, étape (P_1), et T_w, étape (P_2).

La valeur de P ne doit pas être trop grande pour éviter des temps d'exécution trop longs, cela vient de l'algorithme de [Dijkstra 59] à appliquer. Cependant, si la taille de P est trop petite, nous perdons l'intérêt d'une analyse par chemins minimaux. Nous pouvons voir ces effets dans la figure 10.3. De plus, les temps d'exécution ont été calculés avec différentes valeurs de P et ainsi, nous avons choisi, en pratique, de prendre $P = 8$. Ce choix est assez cohérent avec les travaux proposés dans [Huang 06, Li 12] qui effectuent une analyse de l'image par blocs en prenant des tailles entre 7×7 et 9×9.

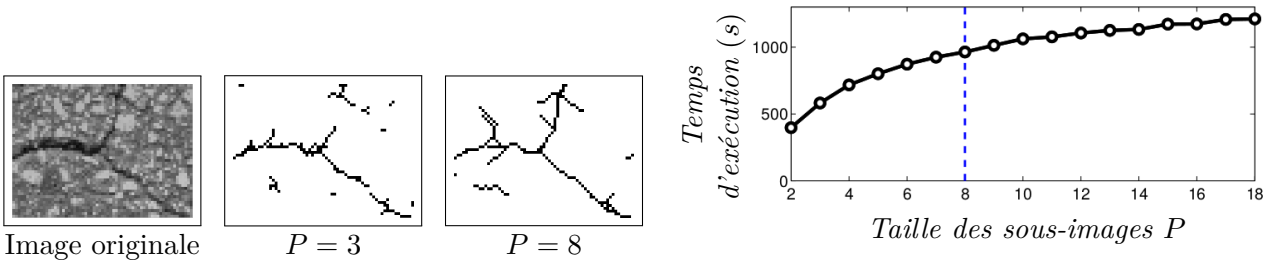


FIGURE 10.3. – Influence du paramètre P sur les chemins détectés et les temps d'exécution – Lorsque P est faible, comme par exemple $P = 3$, des parties de la fissure sont manquantes et nous obtenons donc une forme discontinue. Pour les temps d'exécution, il s'agit du temps moyen pour toutes les images de la base de données.

Le choix de k_e permet de déterminer le seuil $T_e = \mu - k_e\sigma$ qui permet de ne retenir que les points d'extrémité les plus sombres, pour augmenter la probabilité d'avoir un pixel appartenant à une fissure, comme point d'extrémité. Les valeurs les plus faibles de k_e vont avoir tendance à éliminer peu de points et ainsi engendrer une approche coûteuse en terme de temps de calcul. En revanche, une valeur élevée élimine trop de points au risque de perdre toute chance de détecter certaines parties d'une fissure. Nous

illustrons ce comportement dans la figure 10.4. En pratique, nous avons donc choisi $k_e = 1$. Toutefois, il est important de noter que le nombre de vrais positifs est assez faible. C'est le fait d'utiliser les chemins minimaux qui va nous permettre d'obtenir une bonne détection des fissures. Cette utilisation est au cœur de l'approche proposée.

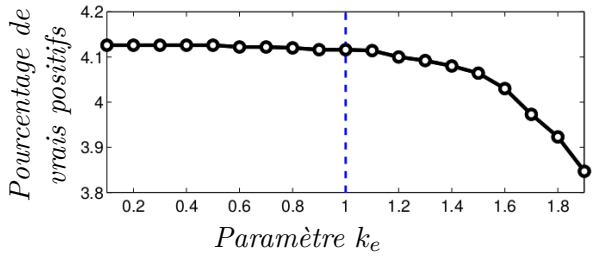


FIGURE 10.4. – Évolution du nombre de vrais positifs parmi les points extrémités en fonction du paramètre k_e – Pour la valeur choisie $k_e = 1$, 6% des points extrémités sont éliminés.

De même, le choix de k_c détermine la valeur de $T_c = \mu_c - k_c\sigma_c$ qui permet de sélectionner les chemins les plus sombres détectés à l'étape (C_2). Expérimentalement, nous avons choisi $k_c = 1$ qui permet d'éliminer la plupart des faux positifs en gardant la majorité des vrais positifs. La figure 10.5.(a) permet de donner le taux de performance en fonction de k_c et confirme que $k_c = 1$ est le meilleur compromis en terme de segmentation de fissures. Pour T_s, cf. figure 10.5.(b), le meilleur choix est $T_s = 60$ car empiriquement cette valeur permet le meilleur compromis entre précision et sensibilité. Enfin, le choix de k_w détermine $T_w = \mu_w + k_w\sigma_w$ utilisé pour propager la fissure et calculer ainsi son épaisseur. Une valeur adaptée pour k_w doit améliorer la sensibilité sans perdre en précision. Ainsi, la valeur $k_w = 0.6$ semble correspondre à la valeur qui maximise DSC, cf. figure 10.5.(c).

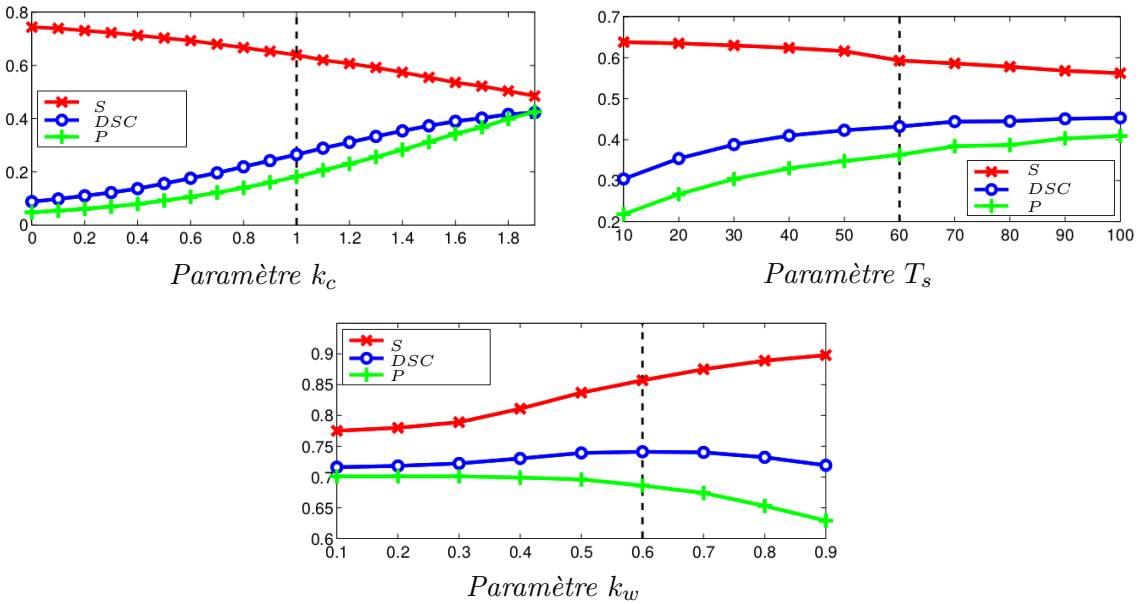


FIGURE 10.5. – Précision (P), sensibilité (S) et valeur DSC pour la détection des fissures réelles – Nous avons fait varier les paramètres k_c, T_s et k_w. Ces résultats sont obtenus sur toute la base de données réelles.

10.8.2. Évaluations et comparaisons sur les données synthétiques

Les figures 10.6 et 10.7 illustrent la sensibilité à la texture des méthodes à base de modèle markovien M_1 et M_2. L'approche GC dépend très fortement de l'initialisation des points d'intérêt. En effet, tout point détecté en dehors de la fissure va fortement perturber la qualité du résultat obtenu et en particulier, le pourcentage de faux positifs. En revanche, l'approche FFA permet de présenter un résultat continu mais les contraintes de direction et de longueur ne permettent pas à la méthode de détecter les fissures avec une forme chaotique. la méthode MPS0 permet d'améliorer encore la qualité des résultats mais elle présente encore des artefacts qui diminuent ses performances comme la présence de boucle dans la détection. Ainsi, la valeur DSC moyenne obtenue vaut 0.75. Comme expérée, l'approche MPS atteint un score de 0.83 et les résultats visuels exposés dans la figure 10.6 mette en évidence que MPS peut détecter des fissures de forme et d'épaisseur variables. En résumé, sur ces images synthétique GC obtient les plus mauvais résultats alors que FFA, M_1 et M_2 obtient des résultats intermédiaires par rapport aux meilleurs performances obtenues par MPS0 et MPS.

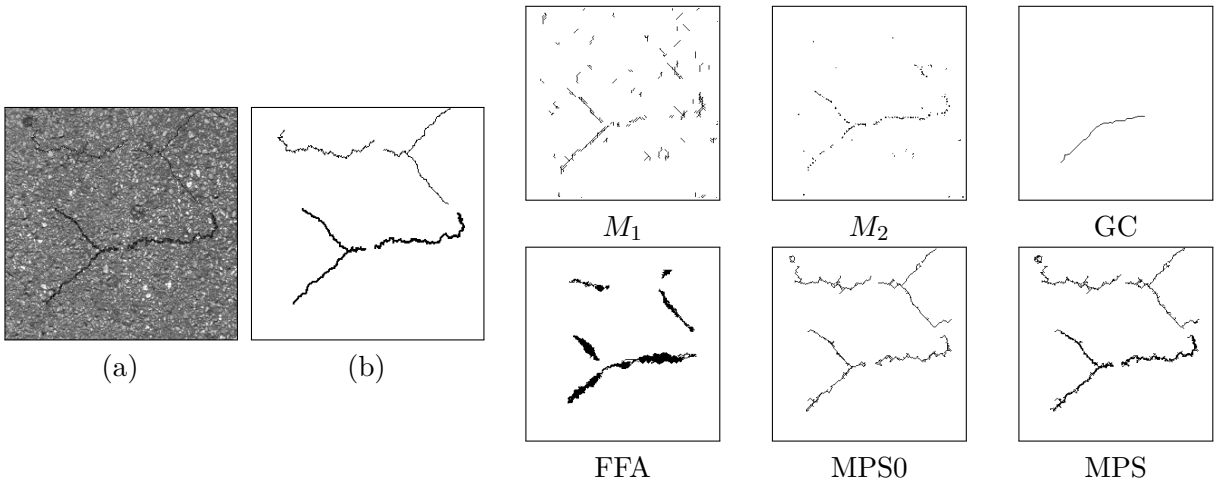


FIGURE 10.6. – Résultats de segmentation de fissures synthétiques – En (a), il s'agit de l'image originale, avec la segmentation de référence utilisée en (b). Nous illustrons tous les résultats obtenus avec toutes les méthodes testées : M_1, M_2, GC, FFA, MPS0 et MPS sur une image réelle. Nous indiquons à chaque fois la valeur obtenue pour la précision, P, la sensibilité, S et le critère DSC.

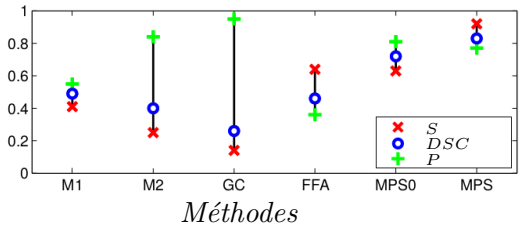


FIGURE 10.7. – Précision, sensibilité et valeur DSC pour la détection des fissures synthétiques, cf. figure 10.6.

10.8.3. Comparaison sur les images réelles

Toutes les images de la base présentée au § 10.7.1 ont été testées et évaluées quantitativement. Pour illustrer les comportements, nous montrons également des exemples de segmentation, cf. figure 10.9. Pour les approches M_1 et M_2, les trois critères utilisés (p, S et DSC) sont inférieurs à 0.5 car les

taux de faux positifs et faux négatifs sont trop élevés. L'approche GC est la plus précise avec très peu de faux positifs mais en contre-partie, le taux de faux négatifs est très élevé. L'approche FFA a le comportement opposé puisqu'il y a très peu de faux négatifs mais un nombre important de faux positifs. Ainsi, tous ces résultats confirment les résultats préliminaires obtenus avec les images de synthèse : MPS est l'approche plus satisfaisante, suivie par MPS0, FFA, M_1, M_2 et, enfin, GC, par ordre décroissant de performance.

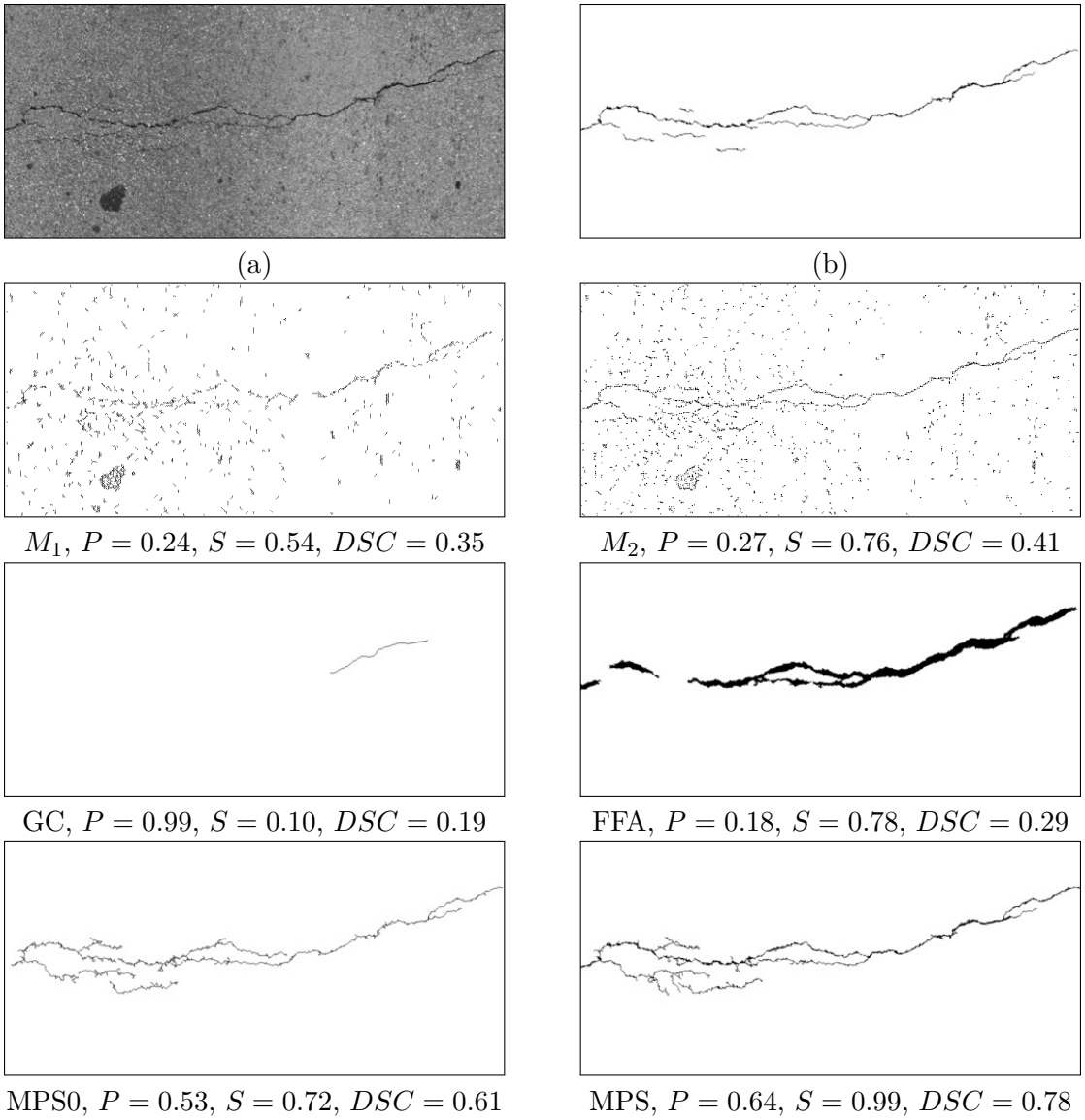


FIGURE 10.8. – Résultats de segmentation de fissures réelles – En (a), il s'agit de l'image originale, avec la segmentation de référence utilisée en (b). Nous illustrons tous les résultats obtenus avec toutes les méthodes testées : M_1, M_2, GC, FFA, MPS0 et MPS sur une image réelle. Nous indiquons à chaque fois la valeur obtenue pour la précision, P, la sensibilité, S et le critère DSC.

Nous avons affiné nos tests en étudiant en particulier la sensibilité aux changements de texture de l'approche MPS. En effet, les matériaux utilisés pour construire la chaussée sont variables et nous souhaitons illustrer le comportement de la méthode en cas de variation de la texture. Ainsi, nous montrons les résultats obtenus pour trois images A, B et C, par ordre de difficulté croissante, cf. figure 10.10. Alors que l'approche MPS obtient de bons résultats pour A et B, ce qui est confirmé par la valeur de DSC obtenue, cf. la table 10.10, les résultats obtenus en C sont moins satisfaisants,

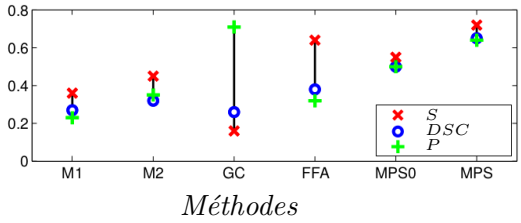
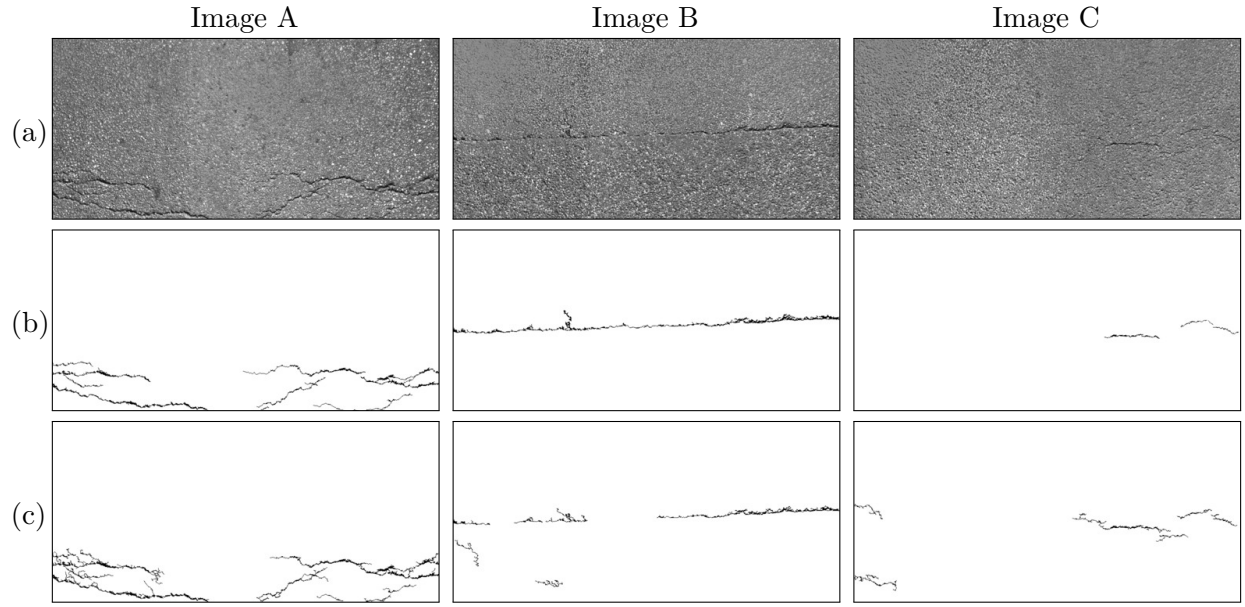


FIGURE 10.9. – Précision, sensibilité et valeurs DSC moyennes sur les 36 images réelles.

illustrant ainsi les cas limites d'application de notre approche. En 10.10.(d) nous présentons également les résultats des 5 autres approches sur ces trois mêmes images et nous observons le même type de comportement.



Valeurs du critère DSC associé						
	M_1	M_2	GC	FFA	MPS0	MPS
Image A	0.35	0.37	0.24	0.44	0.58	0.77
Image B	0.18	0.28	0.42	0.37	0.37	0.75
Image C	0.06	0.18	0.00	0.24	0.18	0.32

FIGURE 10.10. – Résultats de détection de fissures avec MPS – Nous présentons 3 images avec des revêtements de route différents, et donc, 3 textures différentes. Pour chacune de ces images, nous indiquons l'image originale, en (a) avec sa segmentation de référence en (b), et le résultat que nous obtenons en (c).

10.9. Conclusions sur la segmentation de fissures de chaussée

Dans ce chapitre, nos contributions ont porté sur la mise en œuvre d'une approche originale pour traiter un sujet délicat : la segmentation d'un élément fin et faiblement contrasté dans un contexte fortement texturé. De plus, nous avons mis en place une base de données de tests qui a été partagée. Par la suite, ce travail a été repris dans le projet européen TRIMM, *Tomorrows Road Infrastructure Monitoring and Management* [Karlsson 14, Wright 14]. Après la thèse de Rabih AMHAZ,soit 4 ans

après mon départ de l'IFSTTAR, ce travail a été poursuivi pour être adapté aux systèmes d'imagerie 3D, de type LIDAR, dans le cadre d'une thèse dirigée par Vincent Baltazart [2] et a donné lieu, entre autres, à cette publication : [Kaddah 19].

Nous avons noté que, par la suite, des publications ont été proposées pour étendre ce concept de continuité mais avec des outils différents, comme la notion de segment de fissure introduite dans [Weng 19] ou l'utilisation de superpixels dans [Steckenrider 20]. Comme pour les autres domaines que nous avons abordés, la classification de fissures s'est modernisée avec l'introduction d'approches s'appuyant sur des réseaux convolutifs [Zhang 16, Zou 18], mais également des réseaux adversaires [Liu 19a]. Une publication de [Eisenbach 17] présente la manière de préparer les données dans ce contexte applicatif pour être en capacité d'utiliser une méthode par réseaux de neurones. Enfin, le jeu de données que nous avons proposé a été repris dans [Fan 18] et, depuis, de nombreuses bases de données ont été partagées. Ainsi, [Majidifard 19] présente la base de données la plus complète, à notre connaissance. Enfin, il est possible de récupérer du code pour tester ce type d'approche, avec le travail de [Drouyer 19] [3]. Les approches que nous avons citées ont pour but de classer les images en deux classes : avec ou sans fissure. Dans ce chapitre nous avons surtout abordé la segmentation au niveau du pixel d'une fissure. Les travaux de [Fei 19] introduisent une architecture pour obtenir ce type de résultat. Ainsi, le travail réalisé dans ce chapitre pourrait être amélioré en utilisant des outils d'apprentissage profond, aussi bien en s'appuyant sur une classification de sous-images pour délimiter des zones à étudier, qu'en adaptant un réseau existant en prenant en compte la notion de chemins telle que nous l'avons utilisée.

2. Je n'ai pas souhaité participer à cet encadrement de thèse car ayant quitté l'IFSTTAR, je ne souhaitais plus continuer dans ce domaine applicatif.

3. *via* le lien : http://www.ipol.im/pub/pre/282/

Chapitre 11.

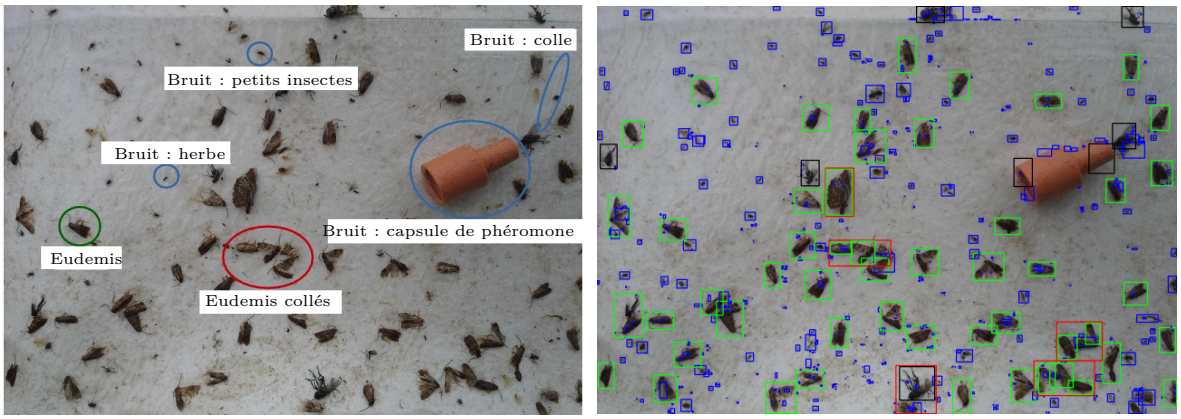
Segmentation combinée avec la saillance curviligne

11.1. Objectif applicatif : traitement éco-responsable de la vigne

Avec la prise de conscience face aux dangers que représentent les insecticides, pesticides et ovicides, pour l'environnement et l'homme, de plus en plus, les états, et par suite, les grandes entreprises, sont en quête de solutions permettant la réduction de leur utilisation, voire leur disparition. C'est le cas de l'industrie du vin et des appels à projets visant à étudier des solutions pour réduire leur utilisation. Ces travaux s'inscrivent donc dans le cadre d'un projet de type FUI, nommé VITI, qui concerne la viticulture dans le Sud de la France, et en particulier le traitement d'une larve particulière, la larve du papillon *Lobesia Botrana*, plus communément appelé Eudemis. De manière traditionnelle, les viticulteurs pulvérisent un ovicide toute la saison afin d'assurer la non-prolifération de cette larve. Toutefois, les études ont montré qu'il n'est nécessaire d'utiliser cet ovicide que 2, voire 3, fois dans la saison. Pour cela, les viticulteurs placent des pièges à Eudemis. Ils étudient l'évolution de la quantité d'Eudemis et savent déterminer les moments de ponte qui correspondent aux moments où il est nécessaire de pulvériser l'ovicide. Sur de grands domaines étendus sur des centaines de kilomètres, il est difficile pour les employés viticoles de surveiller correctement ces pièges et il apparaît primordial d'automatiser au maximum cette tâche. Le système envisagé dans le cadre de ce projet est d'ajouter un appareil photographique permettant d'acquérir des photos à intervalles réguliers (quelques photos par jour), de les traiter automatiquement, afin de transmettre le comptage des insectes. Les travaux proposés dans ce chapitre ont permis d'introduire une première approche de détection de ce papillon. Cette méthode a été transférée à l'entreprise partenaire du projet, `SiConsult`.

11.2. Problématique de segmentation des images d'insectes

Détecter des insectes et les classer, dans ce contexte applicatif, est assez délicat car l'objet à reconnaître est assez petit (l'insecte est plus petit qu'une mouche), possède peu de couleurs (il possède des tons marrons, peu contrastés), a des caractéristiques de forme assez peu discriminantes (c'est un papillon, souvent piégé avec les ailes refermées). De plus, le fond de l'image contient de nombreuses difficultés : des traces de colles, la capsule de phéromone (c'est ce qui attire les papillons), des leurres (autres insectes), des poussières, des ombres portées. Il faut également noter que, comme les acquisitions sont réalisées en extérieur, il y a aussi des changements d'illumination [Schurischuster 16]. Enfin, au cours du temps, le risque d'avoir des papillons collés les uns aux autres augmentent, ce qui constitue la dernière difficulté. Toutes ces difficultés sont illustrées dans la figure 11.1 où nous montrons également le type de résultats que nous avons obtenus.



(a) (b)

FIGURE 11.1. – Exemple d'image de papillons et de résultat obtenus – En (a), nous pouvons voir un exemple d'images étudiés et des éléments que nous souhaitons segmenter, dont le papillon d'intérêt : *Lobesia Botrana* ou Eudemis, un papillon européen, parasite des vignes (cercle vert). Malheureusement, ces images contiennent des difficultés que nous appelons de manière générale, du bruit, comme, par exemple, la capsule de phéromone, de l'herbe, des insectes plus petits (tous les cercles bleus). De plus, certains papillons sont trop proches les uns des autres pour être séparés (il s'agit des cercles rouges). En effet, ces papillons se touchent et il est important de le prendre en considération. En (b), nous pouvons visualiser le type de résultat obtenus : les bruits détectés (rectangles bleus), les papillons collés ou se touchant (rectangles rouges), les papillons Eudemis (rectangles verts) et enfin les autres insectes (rectangles noirs).

11.3. Approches existantes pour segmenter des images d'insectes

De nombreuses approches ont été introduites afin de proposer une classification d'images de papillons [Kang 14, Kaya 14], bien souvent à des fins de reconnaissance et classification biologique. Dans ce cas, les images sont de bonnes résolutions, sans flou et avec un fond uniforme [Kang 14, Wang 12a]. Ainsi, des outils classiques de classification, sans adaptation spécifique, peuvent directement être appliqués [Ding 16]. Peu de travaux concernent les papillons abordés dans les problématiques liées à l'agriculture, dans ce contexte très peu contrôlé et très bruités. Dans ce contexte plus délicat, il est habituel de distinguer deux phases : une première permettant de segmenter et détourer l'insecte afin d'éliminer les bruits liés au fond, et ensuite, une seconde phase de classification. De nombreuses caractéristiques peuvent être exploitées pour effectuer la segmentation, comme la couleur, la forme et la texture [Yalcin 15, Qing 12]. Plus précisément, [Yalcin 15] exploite une approche par contour actif, alors que [Wen 15] s'appuie sur des outils de morphologie mathématique. La limite de toutes ces approches est qu'elles ne prennent pas en compte la présence d'insectes collés les uns aux autres (qui peuvent entraînés des contours occultés). C'est la raison pour laquelle, pour l'approche proposée, nous avons utilisé les descripteurs introduits au § 4.6.3.

Pour l'étape de classification, [Yalcin 15] ont simplement utilisé des k-moyennes alors que [Qing 12] ont utilisé une approche par *Support Vector Machine*, SVM [Cortes 95]. Dans [Wen 15], les auteurs ont introduit une approche par apprentissage profond. Dans un premier temps, nous n'avons pas envisagé d'apprentissage profond car nous ne possédions pas une base de données suffisamment conséquente pour effectuer un apprentissage. Pour cet aspect, nous avons proposé d'utiliser l'approche SVM [Cortes 95] combinée au descripteur multi-échelle *Histogram Of Curviness Saliency*, HCS, introduit au § 4.10.

11.4. Reconnaissance s'appuyant sur SVM et des descripteurs HCS

11.4.1. Présentation générale de l'algorithme proposé

L'approche proposée, présentée dans la figure 11.2, permet de détecter des papillons Eudemis dans des images de pièges qui contiennent de nombreuses difficultés : bruits, insectes collés les uns aux autres, leurres, comme nous avons présentés dans la figure 11.1.

La méthode commence par détecter tous les insectes (par opposition à tous les autres objets de la scène qui ne sont pas des insectes), grâce aux étapes (1) et (2), nous avons publié ces résultats préliminaires dans [Bakkay 18a]. Ensuite, nous réalisons la classification, étape (3), que nous avons publiée dans [Bakkay 18b]. Nous détaillons ces trois étapes dans les paragraphes qui suivent.

11.4.2. Étape (1) : détection robuste des contours des insectes

Pour cela, nous exploitons l'approche de détection de contours présentée au § 4.7, publiée dans [Rashwan 16]. Nous supposons que cette première détection est suffisamment robuste aux faibles contrastes et aux problèmes de texture pour fournir tous les contours de l'image, plutôt fermés. Ensuite, sur la base des contours détectés et donc des régions associées à chaque contour fermé, nous utilisons l'algorithme des k-moyennes pour classer en k classes (k étant inconnu), de la manière suivante : les contours dus à des bruits (classe 1), les contours dus à des insectes de taille équivalente au papillon recherché (classe 2) et enfin des contours dus à des insectes de taille plus importante ou des amas de papillons que nous cherchons à détecter (classe 3 à k). L'originalité de l'approche proposée repose sur un critère adapté au problème posé. En effet, pour réaliser la classification avec l'algorithme des k-moyennes, nous prenons en compte la forme de la surface comprise à l'intérieur du contour fermé. Nous avons choisi ce critère car la forme de l'insecte nous semble la caractéristique la plus discriminante pour les distinguer des autres objets qui composent la scène. De plus, nous avons rendu automatique le choix du paramètre k, le nombre de classes, en utilisant la méthode d' [Tibshirani 01].

11.4.3. Étape (2) : affinement de la segmentation pour supprimer les artefacts liés aux insectes collés entre eux

Ensuite, nous affinons la segmentation réalisée pour les classes à partir de 3, en appliquant une segmentation basée région. Le but de ce traitement est d'identifier les amas d'insectes, afin de traiter cette difficulté. Plus précisément, nous reprenons tous les contours classées entre 3 et k afin de déterminer si ces contours correspondent à un seul objet ou à plusieurs. Pour cela, nous suivons les trois étapes suivantes : dilation du contour, pour supprimer les discontinuités dans les contours, fusion des régions pour éviter des objets sur-segmentés et enfin, application de l'algorithme de segmentation par ligne de partage des eaux [Meyer 92]. À la fin de ce processus, nous avons deux cas possibles que nous détaillons. L'algorithme de partage des eaux détecte :

1. Deux ou trois régions et, dans ce cas, nous redivisons la région en fonction de ce résultat de segmentation.
2. Une seule grande région et, dans ces cas, nous récupérons ce résultat afin d'affiner la détection de l'insecte.

Après ces deux premières étapes, nous avons obtenus 87% des bruits ou autres leurres bien détectés comme non insectes et 70% des insectes sont détectés comme étant des insectes. Enfin, 82% des insectes collés ont bien été identifiés et bien redécoupés. De plus, seulement 13% des insectes sont détectés comme bruits, 17% des insectes sont identifiés comme amas d'insectes et 18% des insectes collés sont détectés comme simples insectes. Aucune zone d'insectes collés n'est classée comme bruit.

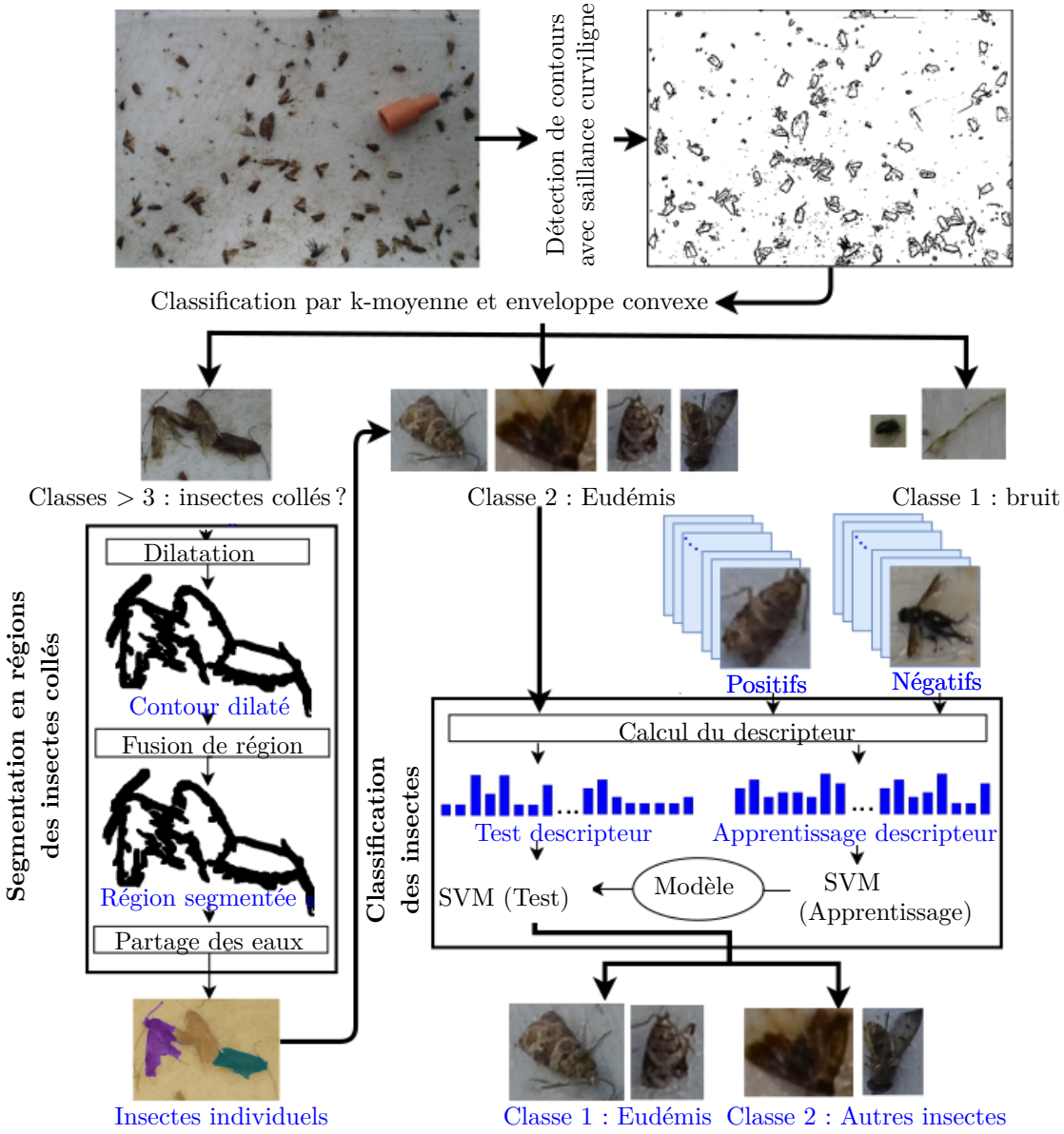


FIGURE 11.2. – Approche de segmentation de papillons Eudemis proposée.

11.4.4. Étape (3) : classification des insectes

Enfin, nous classons ces insectes en distinguant l'insecte d'intérêt, *Lobesia Botrana* ou Eudemis, des autres insectes, comme, par exemple les araignées ou les mouches. Nous utilisons une approche par *Support Vector Machine*, SVM [Cortes 95], en 2 classes : les papillons (classe 1) et les autres (classe 2). Nous utilisons également les histogrammes de saillance curviligne ou *Histogram of Curvilinear Saliency*, HCS, introduits au § 4.10. Plus précisément, nous construisons l'hyperplan qui sépare les données d'apprentissage en deux classes, en s'appuyant sur les descripteurs HCS. Cela implique que nous supposons que les deux classes ne sont pas linéairement séparables. Ainsi, nous utilisons ce coût :

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{p} \alpha_i^* y_i k(\mathbf{x}_i, \mathbf{x}) + w_0\right), \tag{11.1}$$

où $(\vec{x_i}, y_i)$ correspond à la donnée d'entraînement, $\vec{x_i}$ le descripteur de dimension p, $\vec{y_i} \in \{-1, 1\}$ les étiquettes des classes, k la fonction noyau, α_i^* les multiplicateurs de Lagrange et w_0 le biais. Ils existent de nombreuses fonctions noyaux [Cortes 95], mais, ici, nous avons expérimentalement validé le choix de celui-ci :

$$k(\vec{x_i}, \vec{x_j}) = (\vec{x_i}.\vec{x_j} + 1)^6. \tag{11.2}$$

11.5. Évaluation de l'approche de reconnaissance de papillons

11.5.1. Images testées pour la reconnaissance de papillons

Nous avons récupéré et annoté des sous-images : 2865 sous-images négatives qui présentent des leurres et 746 sous-images positives. Les images dans lesquelles nous avons extrait ces sous-images ont été acquises dans différentes conditions d'illumination. Elles possèdent de nombreux types d'insectes de taille variable et de nombreux éléments qui peuvent induire des erreurs : herbes, lignes de colle, capsule de phéromones, comme nous pouvons voir dans la figure 11.3. Étant donné que le nombre de cas positifs est beaucoup moins important que le nombre de cas négatifs, nous effectuons une augmentation de la base de données afin d'obtenir 3577 sous-images positives. Ceci nous permet également d'apprendre à l'algorithme à être robuste aux transformations géométriques basiques et aux bruits puisque la technique que nous utilisons effectuent des rotations et des translations ainsi qu'un lissage des images par filtres gaussiens [Ding 16]. Pour la validation croisée, nous avons choisi de découper aléatoirement cette base de données en 5 ensembles et le processus est répété 5 fois en prenant successivement chacun des ensembles comme ensemble de test alors que les 4 autres sont utilisés pour l'apprentissage. Cela correspond à un algorithme standard en classification [Kohavi 95].

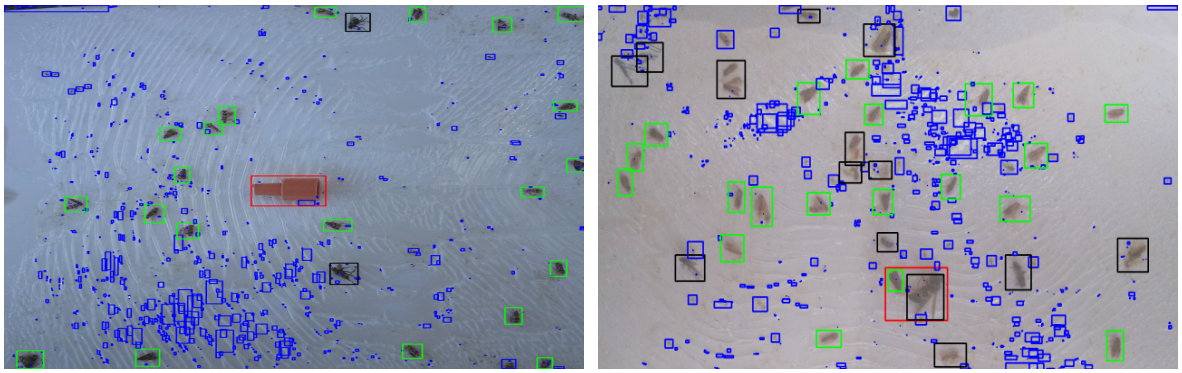
11.5.2. Étude des paramètres de l'approche de reconnaissance de papillons

Pour le calcul du descripteur, différents paramètres sont nécessaires (la taille de la fenêtre, la taille des sous-images, la taille du pas pour parcourir l'image, la taille des cellules prises en compte dans chaque sous-image, et le nombre d'éléments dans l'histogramme utilisé. Tous ces choix ont été faits en prenant en considération nos travaux précédents [Rashwan 16] (en particulier dans [Rashwan 19a], nous avons publié tous les détails nécessaires pour cela). De plus, expérimentalement, nous avons choisi de travailler avec des sous-images de taille (64×48), la taille des blocs considérés pour la construction des histogrammes est (8×8) avec un pas de (4×4) et une taille de cellule à (4×4). L'histogramme est quantifié en 9 valeurs, comme dans [Dalal 05b]. Ainsi, le descripteur contient 5940 éléments. Le seul paramètre à déterminer pour SVM est la marge d'erreur C. Nous avons choisi de tester des valeurs entre 0.1 et 0.5 et les meilleurs résultats ont été obtenus pour 0.1.

11.5.3. Résultats de segmentation de papillons

La figure 11.3 permet de présenter des résultats visuels. Nous observons que la plupart des insectes (rectangles verts et violets) et des bruits (rectangles bleus) sont bien identifiés grâce à l'algorithme de segmentation. De plus, les insectes collés sont bien séparés dans la plupart des cas (rectangles rouges). Enfin, l'étape de reconnaissance permet de bien séparer les papillons Eudemis (rectangles verts) des autres insectes (rectangles violets).

Quantitativement, nous avons comparé nos résultats à ceux obtenus avec certaines des approches citées au § 11.1. La table 11.1 montrent les résultats quantitatifs et nous pouvons constater que l'approche proposée obtient un taux de précisions de 95%.



(a) (b)

FIGURE 11.3. – Résultats de détection de papillons dans des images réelles – En (a) l'image présente des difficultés liées à l'éclairage alors qu'en (b) l'image présente une faible résolution.

TABLE 11.1. – Évaluation et comparaison d'approches de segmentation de papillons – La première méthode correspond à la méthode introduite dans ce chapitre.

Méthode	**Précision(%)**
[Bakkay 18b]	95.8 %
[Ding 16]	93.2 %
[Qing 12]	90.1 %
[Yalcin 15]	87.6 %

11.6. Conclusion sur la détection de papillons

Ce travail nous a permis de valider l'intérêt de ce que nous avons développé dans la partie I pour des applications dédiées. Plus précisément, nous avons dû adapter cette approche de détection à notre problème de segmentation. Les résultats sont encourageants et depuis, cet algorithme a été repris et implanté dans la chaîne de traitement des images acquises sur les pièges par la société `SiConsult`. Ensuite, de nombreux stagiaires, issus de l'ENSEEIHT, ont contribué à l'améliorer, au sein de l'entreprise. Et, en particulier, la phase d'apprentissage a été remplacée par un apprentissage à base de réseau de neurones convolutifs. Une perspective de ce travail est de généraliser la méthode à tous types d'insectes d'intérêt. Et, à cet effet, depuis 2019, deux stagiaires ont travaillé sur le suivi d'abeilles dans des ruches.

Conclusion sur la segmentation multi-critères

Cette dernière partie nous a permis de présenter nos travaux en segmentation dans trois domaines applicatifs très différents. En dehors des différentes perspectives et améliorations déjà introduites dans chacun des domaines, nous souhaitons à présent souligner les travaux de recherche faisant intervenir un apprentissage profond. Les travaux les plus connus qui ont permis d'améliorer les approches de classification en introduisant des réseaux de neurones sont ceux de [Krizhevsky 12], *AlexNet*, et [He 16], *ResNet*. Le succès de la première approche vient du fait qu'elle a remporté la compétition ILSVRC, *ImageNet Large Scale Visual Recognition Challenge*[1], en 2012. En particulier, la méthode s'est démarquée des autres avec une marge très importante, contrairement aux compétitions précédentes. Nous savons qu'un réseau de neurones possède plusieurs couches intermédiaires et plus le nombre de couches augmente, plus le nombre de paramètres augmente et plus le réseau sera coûteux, en terme de ressources et de temps de calculs, à entraîner. La particularité du réseau proposé par [He 16] est de proposer des réseaux simplifiés qui permettent l'introduction d'un nombre conséquent de couches intermédiaires. Plus précisément, l'architecture proposée permet d'éviter certaines couches en effectuant des connections permettant d'effectuer des raccourcis. Il faut noter que ce réseau a également remporté la compétition ILSRC, en 2015.

En détection d'objets, les opérateurs célèbres sont *Mask R-CNN*, introduit dans [He 20], et YOLO, *You Only Look Once* [Redmon 18]. Nous avons d'ailleurs utilisé cette technique dans la partie II, chapitre 7. La première approche s'appuie sur la notion de superpixels, également abordée dans le chapitre III, § 8.3 et au chapitre 9, et utilise un réseau de neurones convolutif. Deux variantes plus performantes et plus rapides ont été successivement introduites par les auteurs [Girshick 15, Shaoqing 15]. La seconde approche [Redmon 18] consiste à prédire les coordonnées des boîtes englobantes des objets en s'appuyant sur une grille d'images (découpage en cellules).

En segmentation, les approches les plus connues sont celles de [Ronneberger 15], réseau *U-Net*, en segmentation d'images médicales, et [Badrinarayanan 17], réseau *Seg-Net*, en segmentation de scènes urbaines. Ces deux réseaux utilisent des auto-encodeurs, cf. § 12.2.3.

Ces différentes techniques pourraient à présent être introduites dans les différentes applications visées. En particulier, comme information de contexte dans les travaux du chapitre 9 sur la segmentation de scènes urbaines. Mais également, comme classification des zones à traiter ou non pour la détection de fissures, cf. chapitre 10. Et, enfin, pour moderniser la phase d'apprentissage par SVM des papillons dans des images de piège, cf. chapitre 11.

1. http://www.image-net.org/challenges/LSVRC/

Quatrième partie

Apprentissage profond pour l'analyse de contenus visuels complémentaires

Détection, mise en correspondance et segmentation pour identifier des objets

Détecter, décrire, suivre, segmenter, tous ces domaines de la vision par ordinateur et du traitement d'images produisent chaque année de nombreuses publications et dans ce manuscrit, nous avons présenté les apports réalisés dans ces domaines. De plus, nous avons cherché à avoir une prise de position avec le plus de recul possible, et cela ne peut se faire sans la proposition d'un état de l'art détaillé. Il est impossible d'être exhaustif mais, pour chaque domaine, nous avons tenté d'être complet, en terme de catégories d'approches existantes, tout en étant à jour sur les techniques récentes, notamment celles liées à l'apprentissage profond. chaque algorithme proposé a également été validé par une évaluation la plus aboutie possible ainsi que le partage de jeux de données avec la communauté scientifique, c'est le cas des chapitre 7 [2] et 10 [3]. Enfin, pour valider les approches proposées, nous avons pu, pour certaines, les appliquer à des cas réels et la dernière partie de ce manuscrit a ainsi mis en évidence l'intérêt des approches développées pour des applications concrètes, en particulier en segmentation, aussi bien pour des scènes urbaines, des objets naturels (insectes) ou des objets fins (détection de défauts).

Toutes les contributions associées à chaque partie ont été listées en détails dans les chapitres commençant par **Contributions scientifiques structurantes**, page 11, pour la première partie sur la détection de primitives en 2D et en 3D, page 73, pour la seconde partie sur la mise en correspondance et l'analyse de vidéos multiples avec recouvrement, et, enfin, page 129 sur les nombreuses applications en segmentation. Dans le reste de cette partie, nous détaillons les différentes perspectives que nous envisageons.

Comme nous allons utiliser le vocabulaire associé à l'apprentissage profond, nous commençons par un bref état de l'art des différentes approches existantes.

2. http://ubee.enseeiht.fr/dokuwiki/doku.php?id=public:tocada
3. https://www.irit.fr/~Sylvie.Chambon/Crack_Detection_Database.html

Chapitre 12.

Éléments indispensables pour l'apprentissage profond

Ce chapitre a pour objectif, d'une part, de résumer les éléments nécessaires pour comprendre et utiliser un réseau de neurones pour effectuer un apprentissage profond, et, d'autre part, d'introduire le vocabulaire dédié. Afin d'étudier et d'utiliser des réseaux de neurones profonds, dans un contexte donné (détection, appariement, segmentation), il est important de distinguer les aspects majeurs pour construire un modèle permettant d'obtenir des résultats de qualité, à savoir :

(1) la base de données utilisée ainsi que

(2) le type de réseaux utilisés.

Nous détaillons ces deux aspects dans les deux paragraphes suivants.

12.1. Base de données pour l'apprentissage profond

Encore actuellement, l'utilisation des réseaux de neurones profonds peut être entravée par la base de données d'apprentissage utilisée. En effet, bien qu'il existe des bases de données très riches, certains domaines exploitent encore peu les réseaux de neurones profonds. Nous commençons ici par établir un historique des bases de données utilisées en traitement d'images ou en vision par ordinateur.

L'utilisation de base de données pour réaliser une tâche n'est pas nouvelle. Cela fait partie du processus classique de toutes les approches dites supervisées. Toutefois, depuis le début de l'utilisation massive des réseaux de neurones profonds, cette étape est devenue de plus en plus cruciale et prend de plus en plus de place, voire constitue une contribution dans de nombreux travaux de recherche. Dans la communauté, il est admis que, classiquement, des millions d'annotations sont nécessaires pour entraîner un modèle. Ici, nous présentons les bases de données les plus connues, dans les domaines de la détection et de la segmentation, car, ce sont deux des domaines que nous avons abordés dans ce manuscrit. Il est important de noter que dans le contexte de la mise en correspondance, ce qui constitue une base d'apprentissage est encore plus délicat à réaliser que pour ces deux domaines.

En détection d'objets, la base de données PASCAL VOC, *Visual Object Classes*, a été introduite dès 2007 et son historique est donné dans [Everingham 15] [1]. Elle est liée à la base de données SUN2012 [2] [Xiao 10]. PASCAL VOC a été étendue aux modèles 3D par la suite [Y. Xiang 14] et nous avons utilisé cette base dans la partie I, chapitre 4. Nous pouvons aussi citer la base ImageNet [Russakovsky 15] [3] à laquelle toutes les architectures convolutives profondes se comparent, notamment AlexNet [Krizhevsky 12] que nous avons introduit à la page 173. Nous pouvons citer la base de données MS COCO, *MicroSoft Common Objects in COntext* [4] [Lin 14] qui contient à la fois des

1. http://host.robots.ox.ac.uk/pascal/VOC/.
2. https://groups.csail.mit.edu/vision/SUN/
3. http://image-net.org/index.
4. http://cocodataset.org/#home

boîtes englobantes et des segmentations des objets d'intérêt, ce qui fait son originalité par rapport aux autres. Mais sa supériorité par rapport aux autres bases de données réside surtout dans le fait qu'elle propose de nombreuses instances d'objets (identiques ou différents) et la difficulté majeure de cette base de données est de désentrelacer les différents objets et de les reconnaître. De plus, cette base fournit plus d'annotations, sur une même classe, que PASCAL VOC et ImageNet, mais, en contre partie, le nombre de classes est moins important qu'Imagenet. Actuellement, la base de données la plus importante, avec plus de 15 millions de boîtes englobantes sur 600 classes différentes appliquées sur 2 millions d'images est Open Images [Kuznetsova 18] [5]. Cette base de donnée propose également beaucoup de segmentations avec $2,7$ millions d'instances sur 1 million d'images.

12.2. Différents réseaux de neurones

Afin de rendre les explications de cette partie les plus intuitives possibles et de faire le lien avec les objets manipulés dans ce manuscrit, nous donnerons des exemples uniquement appliqués à des images.

12.2.1. Réseaux de neurones convolutifs, *Convolutional Neural Network*, CNN

Il s'agit des réseaux les plus populaires et les plus anciens [Goodfellow 16, chapitre 9]. Ce qui les caractérise, en premier lieu, est d'effectuer des convolutions successives, qui permettent, suivant le filtre appliqué, de mettre en évidence différents aspects (comme par exemple, en traitement d'images, un filtre de lissage ou filtre de détection de contours) et qui permettent également de réduire progressivement la taille de l'image. En plus de la taille de ces convolutions, il y a un autre éléments à définir : le *padding*, la stratégie à appliquer pour filtrer sur les bords de l'image. En deuxième lieu, un réseau convolutif fait intervenir des étapes de *pooling*, le fait de découper l'image en une grille et chaque région de cette grille est représentée par une valeur (classiquement le maximum, d'où la notion de *max-pooling* mais il existe d'autres techniques). Enfin, il faut choisir le *stride*, le pas de convolution à appliquer (il ne s'agit plus d'utiliser une fenêtre glissante comme lorsque nous lissons une image). Ces deux éléments, *pooling* et *stride*, correspondent à un sous-échantillonnage que nous pouvons également retrouver pour d'autres types de réseaux de neurones. Ainsi, un réseau de neurones convolutif fait intervenir deux types de couches : les couches de convolution qui sont toujours associées à une couche de *pooling* (et il s'agit des couches les plus profondes) et des couches entièrement connectées, *fully connected*, qui correspondent aux couches les moins profondes et permettent de connecter tous les neurones entre eux.

12.2.2. Réseaux de neurones récurrents, *Recurrent Neural Network*, RNN

La particularité de ces réseaux, introduits dans [Rumelhart 86], est d'être en capacité de prendre en considération des données séquentielles. Plus précisément, une séquence d'entrées notée $[x_1, \ldots, x_t]$ va permettre de fournir une séquence de sortie notée $[y_1, \ldots, y_t]$, tout en maintenant un état interne $[h_1, \ldots, h_t]$. Typiquement, ce type de réseau permet de prendre en compte des séquences d'images, des vidéos, c'est-à-dire, plus généralement, de prendre en compte des données temporelles, comme c'est le cas en segmentation, avec les travaux de [Ventura 19a]. Un autre exemple d'utilisation dans des images est la construction d'une légende de l'image qui permet de décrire la scène suivant les différents éléments qui la composent en indiquant les liens de dépendances et d'interaction entre eux [Xu 15]. La formulation simple de ce réseau est donnée par :

$$h_t = g(Ux_t) + Wh_{t-1}y_t = f(Vh_t) \tag{12.1}$$

5. https://storage.googleapis.com/openimages/web/index.html

où U, V et W sont les paramètres du réseau. Il est dit qu'ils sont partagés dans le temps. La version la plus simple de la fonction g est de prendre la fonction tangente hyperbolique. Les réseaux récurrents les plus fréquemment utilisés sont les réseaux *LSTM, Long Short Term Memory*, introduits dans [Sepp 97]. Du fait de sa complexité accrue par rapport à un réseau convolutif, de nombreuses variantes ont été introduites pour estimer et optimiser les paramètres des réseaux récurrents. De plus, ce type de réseau implique un coût en mémoire et en temps de calcul. Toutes ces variantes et toutes le difficultés associées sont décrites dans l'ouvrage de référence suivant [Goodfellow 16, chapitre 10].

12.2.3. Réseaux antagonistes génératifs, *Generative Adversarial Networks*, GAN

Ils ont été introduits par [Goodfellow 14]. Dans cette publication, les auteurs ont appliqué ce type de réseaux à la génération de nouvelles images avec un niveau de réalisme élevé. Le principe est de mettre en compétition deux réseaux différents dans un contexte proche de celui de la théorie des jeux : le premier réseau est appelé générateur, *generator*, et va générer des échantillons ou des données de synthèse, comme une image de synthèse, alors que le second réseau, son adversaire, le discriminateur, **discriminator** va tenter de faire la distinction entre une donnée de synthèse et une donnée réelle. Plus le générateur s'améliore, plus il produit des données réalistes et, en contre-partie, le discriminateur a plus de difficultés pour distinguer données réelles des données de synthèse. À l'inverse, plus le discriminateur s'améliore, plus il distingue correctement les données réelles des données de synthèse et, en contre-partie, le générateur fournit des données de synthèse qui s'éloignent des données réelles. Cette situation est assimilable à celle d'un jeu à deux joueurs à somme nulle, d'où le lien avec la théorie des jeux. Il faut donc mettre en place une stratégie où chacun minimise le gain maximal de l'adversaire. Cette définition met en avant la difficulté d'entraîner ce type de réseaux. Les travaux de [Radford 15, Radford 16] donnent un certain nombre de recommandations sur l'utilisation de ce type de réseaux de neurones.

Les auto-encodeurs, *auto-encoders*, servent à générer une représentation latente, *latent representation*. En effet, ce réseau est formé d'un encodeur, *encoder*, qui doit permettre d'extraire les caractéristiques principales des données d'entrée et ainsi construire un espace latent. Plus précisément, l'encodeur doit permettre de réduire la dimension en conservant suffisamment d'information discriminante pour qu'un décodeur, *decoder*, puisse retrouver la donnée d'entrée. De part cette définition, l'encodeur peut être assimilé à une ACP, Analyse en Composante Principale. Enfin, le décodeur doit être capable de reconstruire la donnée d'entrée à partir de l'espace latent. Ces réseaux permettent, notamment, de mettre en place une méthode d'apprentissage sans annotation des données d'apprentissage [Bengio 09].

Il est important de noter que des tâches en intelligence artificielle peuvent être effectuées en utilisant une méthode d'apprentissage classique, sans faire intervenir de réseaux de neurones profonds. En effet, toutes les tâches où il est évident que nous pouvons identifier explicitement les éléments caractéristiques à connaître et à apprendre n'ont pas besoin de réseaux de neurones profonds. Nous pouvons penser à la reconnaissance de motifs discriminants comme des marqueurs dans des scènes, comme nous utilisons dans [Mariyanayagam 18]. Mais, justement, dans de nombreuses tâches, il est difficile de déterminer précisément les éléments caractéristiques pour effectuer la tâche. Par exemple, lorsque nous souhaitons reconnaître des voitures, les modèles, les formes, les couleurs, sont tellement variés qu'il est difficile de définir les caractéristiques précises à exploiter. C'est à ce stade que l'utilisation des réseaux de neurones devient intéressante et ce n'est pas un hasard si cet objet, la voiture, est, actuellement, majoritairement reconnu par des approches à base de réseaux de neurones, comme l'outil que nous avons utilisé dans les travaux de [Malon 19b].

12.3. Aspects à maîtriser pour utiliser un réseau de neurones

Une fois la base de données maîtrisées et le choix d'un réseau réalisé, il reste toutefois de nombreux éléments à déterminer afin de mettre en place le réseau proposé, adapté au problème que nous cherchons à résoudre :

- La méthode d'optimisation et en particulier le nombre d'*epoch*[6] et le terme d'inertie ajouté, *momentum*, pour éviter la stabilisation dans un minimum local ;
- Le nombre de couches ;
- Le nombre de neurones.

Plus nous augmentons le nombre de couches et le nombre de neurones, plus nous augmentons le pouvoir de séparation des classes du réseau de neurones. Toutefois, passé un certain nombre de couches et/ou de neurones, le réseau peut être sur-appris[7]. De même, si le modèle est trop entraîné, cela engendre un sur-apprentissage. Il existe des techniques pour éviter le sur-apprentissage (arrêt anticipé, régularisation, connections raccourcies). Le lecteur intéressé peut trouver ces détails dans [Goodfellow 16]. Actuellement avec les architectures plus avancées, les choix sont plus souvent le nombre de répétitions d'un certain bloc comme, par exemple, un bloc résiduel dans ResNet [He 16], ou les *inception blocks* dans [Szegedy 15].

À présent que nous avons effectué cette brève introduction, nous pouvons aborder les perspectives que nous envisageons.

6. Une epoch correspond à un apprentissage sur toutes les données, c'est-à-dire que l'ensemble d'apprentissage a été visité entièrement pour le calcul des gradients.

7. Nous parlons de sous-apprentissage lorsque le modèle appris explique mal l'ensemble d'apprentissage, c'est-à-dire qu'il y a beaucoup d'erreurs. Nous parlons de sur-apprentissage lorsque le modèle appris explique tellement finement les données d'apprentissage qu'il ne peut pas se généraliser à d'autres données.

Chapitre 13.

Perspectives

Les perspectives présentées dans ce chapitre sont de trois niveaux différents :

(1) *Dans la continuité des travaux de thèses encadrées :* C'est le cas des trois premières sections, pour les § 13.1 et 13.2, les thèses ont été soutenues, alors que le § 13.3 correspond à une thèse en cours.

(2) *Dans la continuité des travaux de recherche applicative menés durant toute la carrière :* la section 13.4 présentera les différentes perspectives liées à des soumissions à des appels à projet, pour des applications médicales et des applications de villes intelligentes.

(3) *Dans la continuité des travaux de recherche menés durant toute la carrière :* alors que le § 13.5 nous permettra de proposer une perspective liée à la reconstruction de modèles 3D, le § 13.6 abordera la simplification des modèles 3D pour la compression. Ces deux propositions concernent deux thèses, avec un finacement MESR, qui commencent en septembre 2020. La section 13.7 aborde les perspectives envisagées dans la lignée de la partie II sur la mise en correspondance conjointe.

13.1. Vers une estimation robuste de la pose par détection de marqueurs et de points d'intérêt

Nous pouvons distinguer deux catégories d'approches d'estimation de la pose : les approches par SLAM, *Simultaneous Localization And Mapping*, s'appuyant sur un environnement contrôlé ou dont une partie de la scène est connue, et les approches par SFM, *Structure From Motion*, n'ayant pas d'*a priori* sur la scène [Wu 18]. Dans la partie I, nous avons commencé à aborder le problème de l'estimation de la pose, en utilisant des points d'intérêt, cf. chapitre 4. Dans le cadre de la thèse de Damien Mariyanayagam [1], nous avons également étudié le problème d'estimation de la pose, dans un contexte particulier puisque nous utilisons des marqueurs artificiels dans la scène [Mariyanayagam 20], cf. figure 13.1.(a).

Les travaux développés dans [Mariyanayagam 18] se situent dans la catégorie des approches par SLAM, puisque nous avons utilisé des marqueurs fiduciaires, des marqueurs ajoutés artificiellement dans la scène, et en particulier, des marqueurs circulaires. Dans ces travaux, nous avons proposé une première approche permettant d'obtenir des résultats encourageants en utilisant un seul cercle et en supposant que nous possédons des paramètres de calibrage par défaut. Cependant, nous avons mis en évidence les limites de cette méthode, liées à la distance de prise de vue. En effet, lorsque la résolution du marqueur est faible, c'est-à-dire lorsque la taille de la région du marqueur dans l'image acquise est inférieure à un seuil [2], il n'est plus possible de calculer la pose avec un unique cercle. Nous avons ainsi développé une méthode hybride combinant les marqueurs circulaires avec l'utilisation de points

1. Thèse co-encadrée avec Pierre Gurdjos, sous la direction de Vincent Charvillat.
2. Nous considérons qu'en dessous d'une cinquantaine de pixels, la résolution est faible.



FIGURE 13.1. – Recherche de la pose associée à l'image requête à l'aide d'images de référence – Un des marqueurs manipulés dans ce travail est présenté dans la figure (a) alors qu'en (b), nous présentons l'utilisation originale envisagée. Nous supposons que nous disposons d'images de référence des marqueurs, comme dans le cercle noir à droite. Puis, nous souhaitons reconnaître des marqueurs dans une image requête, comme l'image en bas. Pour ce travail, nous combinons l'utilisation des marqueurs et de points d'intérêt dans la scène.

d'intérêt dans l'environnement [Mariyanayagam 19], cf. figure 13.1.(b). Cette étude a porté sur la façon d'exploiter ces points pour l'estimation de la pose.

Ainsi, l'extension naturelle de ces travaux portera sur les points suivants :

(1) **Étudier la robustesse de la détection des ellipses et ainsi, l'impact des erreurs de détection des ellipses sur le calcul de la pose.** Nous avons commencé à étudier la propagation des incertitudes sur l'estimation des paramètres de l'ellipse, en utilisant l'équation de l'image du centre. Et ainsi, nous avons déjà pu observer comment une erreur sur la détection et sur le calibrage introduit une erreur sur le calcul de l'image du centre. L'image du centre est liée à la position du repère du plan du support par rapport à celui de la caméra. Il faudrait donc enrichir ce travail en propageant cette étude aux paramètres de la pose, notamment à la position relative du repère du plan de support par rapport à la caméra. Pour cela il faudrait simplifier les équations utilisées, ce qui n'est pas trivial. L'apport que nous souhaitons apporter est avant tout théorique.

(2) **Rendre plus robuste la détection des ellipses.** Le marqueur utilisé, présenté en (a), est composé de deux cercles. Or, nous nous sommes concentrés sur la détection et l'utilisation d'un seul cercle, le cercle externe. Pour la suite de ce travail, il s'agit donc de détecter deux cercles conjointement, sachant qu'ils sont dans une configuration particulière. Plus précisément, ce sont des cercles concentriques et, en conséquence, les ellipses (images des cercles) sont confocales. Nous souhaitons donc résoudre un problème, assez délicat, d'une double estimation avec contrainte. L'intérêt est d'obtenir une détection plus fiable de la position du marqueur.

(3) **Étudier l'impact du type de points d'intérêt utilisé sur les résultats obtenus.** Notamment, nous voulons mettre en avant le ou les détecteurs permettant d'obtenir les points d'intérêt les plus fiables et les plus répétables. En particulier, nous voulons reprendre le travail proposé dans [Rashwan 19a] sur la détection en 2D.

(4) **Étudier les approches de détection existantes faisant intervenir de l'apprentissage**

profond. De nombreuses approches s'appuyant sur un apprentissage profond ont été introduites, comme [Kendall 16, Duong 18, Bui 18b]. En particulier, pour des scènes d'intérieur comme celles que nous étudions [Acharya 19]. L'approche de [Bui 18a] est un exemple de réseau où l'étape de mise en correspondance est apprise. De plus, dans [Mondéjar-Guerra 18], un réseau de neurones est exploité pour reconnaître des marqueurs, ici rectangulaires, dans des conditions extrêmes. Nous envisageons d'étudier l'utilisation de ce type de réseau, pour détecter des marqueurs circulaires comme ceux que nous avons étudiés. Ainsi, nous souhaitons nous inspirer de ces approches pour, d'une part, comparer les performances, et, d'autre part, construire une nouvelle approche hybride combinant marqueurs circulaires et utilisation de points d'intérêt, à la manière de ceux introduits dans le chapitre 4, et faisant intervenir un apprentissage profond.

(5) **Faire le lien avec les approches d'estimation de la pose s'appuyant sur de l'apprentissage profond.** Avec Hatem Rashwan, nous avons également commencé à étudier l'estimation de la pose en utilisant des réseaux adversaires génératifs [Abdulwahab 20]. Nous aimerions enrichir ce travail en faisant le lien avec le travail de Damien Mariyanayagam.

13.2. Vers des squelettes texturés et appariés à base de points d'intérêt

Dans ce manuscrit, nous avons présenté de nombreux outils classiques de vision qui permettent d'obtenir une reconstruction 3D d'un objet ou d'une scène. Notamment, dans la partie I où nous avons introduit des opérateurs de détection de points d'intérêt, en 2D et en 3D, et des outils de recalage en 2D et en 3D mais également, dans la partie II, où nous avons présenté de nombreuses approches d'appariement, majoritairement en stéréovision binoculaire.

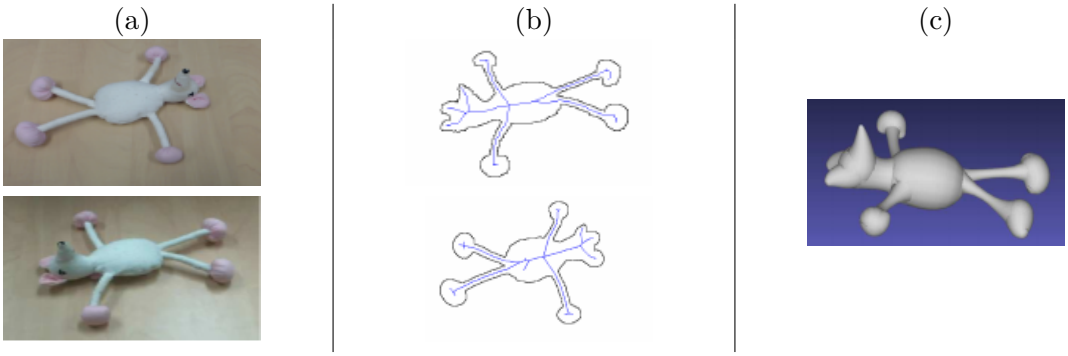


Figure 13.2. – Reconstruction 3D à partir des squelettes perspectifs des formes – À partir de plusieurs images acquises depuis différents points de vue, en (a), nous proposons d'effectuer la reconstruction de l'objet en (c). Cette reconstruction passe par une étape de segmentation fond/forme, puis un calcul du squelette, en 2D, comme en (b). À partir de ces squelettes 2D, nous sommes capables d'estimer le squelette en 3D et ainsi de reconstruire l'objet.

Au cours de ce manuscrit, nous avons identifié à plusieurs reprises la difficulté face au manque de texture. En particulier, dans la partie II, nous avons proposé des approches combinant des mesures de corrélation complémentaires, chapitre 6.2, ou des approches d'appariement par propagation, chapitre 6.3, ou par vote 6.4. Dans le cadre de la thèse de Bastien Durix [3], nous avons pu travailler sur une approche assez différente, cf. figure 13.2. En effet, nous avons mis en correspondance des squelettes 2D, figure (b), pour effectuer une reconstruction 3D [Durix 17], figure (c). L'avantage de cette nouvelle famille d'approches de reconstruction en utilisant le squelette est de s'affranchir de cette

3. Thèse co-dirigée avec Géraldine Morin.

difficulté puisqu'il n'y a plus d'appariement de points d'intérêt ou de pixels. De plus, d'un point de vue applicatif, le modèle 3D obtenu est facilement éditable, grâce à la structure du squelette et 3 à 5 images sont suffisantes pour réaliser la reconstruction. Enfin, un autre avantage de cette approche est d'être capable de construire un modèle virtuel 3D complet, c'est-à-dire fermé. Plus précisément, voici le contexte applicatif que nous visons : un utilisateur peut prendre 3 à 5 photos d'un même objet, à partir de points de vue différents. Cet objet doit être essentiellement tubulaire et fermé. Nous supposons que nous sommes capables de segmenter cet objet par rapport au fond de l'image et nous travaillons donc sur une image binaire de la forme de l'objet. Ainsi, en supposant ces données en entrée, les contributions de ce travail de thèse ont été :

(1) L'étude de la formation de la silhouette d'un objet à partir de son squelette et les propriétés de sa projection perspective. Cette étude du squelette perspectif nous a permis de développer un algorithme d'estimation de la projection perspective d'un squelette 3D curviligne, constitué d'un ensemble de courbes [Durix 15a].

(2) L'estimation du squelette 3D d'un objet à partir de ses projections associées à un ensemble d'images d'un objet [Durix 16a].

(3) La reconstruction des squelettes 3D surfaciques afin de prendre en compte des objets dont les squelettes 3D ne sont pas constitués d'un ensemble de courbes mais peuvent également contenir des parties surfaciques. Pour ce dernier algorithme, le principe est de faire grandir le squelette 3D, sous les contraintes données par les images de l'objet.

De manière générale, une difficulté des approches d'estimation de squelettes est liée à la discrétisation du contour de la forme 2D. En effet, cela induit des artefacts, et, en particulier, des branches peu informatives. Ainsi, le travail publié dans [Durix 15a] a été amélioré en prenant en compte la discrétisation du contour de la forme 2D. Pour cela, nous avons mis en place une approche par propagation d'un cercle à l'intérieur de la forme en maintenant le contact avec les bords et en modélisant explicitement le bruit associé au contour de la forme [Durix 19a, Durix 19b].

La méthode proposée permet de reconstruire l'objet en 3D mais sans aucune information de texture. C'est la raison pour laquelle, elle a ensuite été enrichie en proposant d'ajouter une texture au modèle en exploitant la texture présente dans les images acquises [Fayer 18]. La difficulté majeure pour effectuer ce travail est d'être capable d'appliquer une texture sur les parties du modèle qui ne sont pas visibles dans les diverses acquisitions. C'est pourquoi l'originalité du travail proposé est de s'appuyer sur une approche d'*inpainting*.

Comme tous les domaines abordés dans ce manuscrit (détection, appariement et segmentation), l'estimation du squelette, en 2D ou en 3D, d'un objet ou d'une personne, a également été modernisée en introduisant progressivement des outils d'apprentissage profond. Cela a commencé par l'estimation de squelette de personnes, comme dans la célèbre approche *OpenPose* [Cao 17b], mais, à présent, cela concerne toutes les approches d'estimation du squelette et un *workshop* récent, *Deep Learning for Geometric Shape Understanding*[4], sur la base d'un jeu de données commun [Demir 19], a permis la mise en lumière de nombreuses approches prometteuses [Atienza 19, Panichev 19, Nathan 19]. Dans toutes ces méthodes, une technique de réseaux de neurones adversaires génératifs, de type encodeur/décodeur, est utilisée. Plus précisément, chacun de ces papiers reprend un réseau U-Net [Ronneberger 15] que nous avons présenté dans la partie relative à la segmentation, partie III, page 173.

Afin de poursuivre notre objectif d'allier photométrie et géométrie pour mieux apparier et mieux segmenter, et, dans le contexte bibliographique que nous venons d'exposer, voici les perspectives que nous envisageons sur ce thème :

(1) L'approche de reconstruction n'est pas totalement automatique, notamment la segmentation fond/forme initiale s'appuie sur une approche semi-automatique où l'utilisateur indique les régions à l'intérieur et à l'extérieur de la forme, ces régions étant obtenues en appliquant une

4. Organisé dans le cadre de la conférence CVPR, *IEEE Conference on Computer Vision and Pattern Recognition*, en 2019.

approche de segmentation par superpixels, cf. § 8.3. Nous pourrions utiliser une approche semi-automatique encore plus avancée, par détourage, comme proposée dans [Pizenberg 17] mais il parait assez naturel de tenter d'**automatiser totalement la segmentation de la forme** pour rendre l'outil de reconstruction encore plus accessible au grand public. Pour cela, il est nécessaire d'évaluer l'impact d'une segmentation erronée sur la qualité du résultat de reconstruction obtenu. cette étude apporterait un éclairage nouveau sur les performances ainsi que des pistes d'amélioration possible. Nous envisageons d'automatiser la segmentation en continuant à utiliser une approche par superpixels couplée à l'utilisation d'une classification par réseaux de neurones, comme ceux présentés à la page 11.6, afin, grâce à la sémantique, de faire la différence entre le fond et la forme. De plus, nous pourrions imaginer une approche conjointe permettant de combiner estimation du contour de la forme et estimation du squelette.

(2) Dans les travaux réalisés, des difficultés persistent dans certains cas, comme par exemple, lorsque des formes se répètent dans un objet, et peuvent donc être confondues ou mélangées, c'est le cas, par exemple, avec les différents membres (jambes, bras) d'une personne. D'autre part, certains détails fins sont considérés comme un bruit et ne sont pas pris en compte, et, à l'inverse, un bruit sur la forme peut être assimilé à un détail à conserver. Ainsi, nous envisageons de **combiner l'approche par squelettes avec une détection de points saillants** ou d'intérêt. En effet, nous pensons réaliser une détection automatique de germes pour démarrer et/ou contrôler la propagation du squelette. Ces germes peuvent correspondre à des zones saillantes, à la manière de ce qui est détecté dans [Cao 17b] ou des points d'intérêt sur la forme, en 2D ou en 3D, à la manière de ce que nous avons réalisé dans [Abdulwahab 20]. Nous espérons que la caractérisation à la fois photométrique et géométrique permettra de distinguer les éléments ayant une forme proche les uns des autres et également de faire le tri entre détails importants de la forme et bruit sur la forme.

(3) Il est évident qu'avec 3 à 5 acquisitions par un utilisateur non expert, les clichés acquis peuvent ne pas être suffisamment complémentaires : d'une part, pour que l'approche proposée soit applicable, et, d'autre part, pour reconstruire la forme dans son intégralité. Nous aimerions donc ajouter un processus permettant d'indiquer à un utilisateur la **manière de prendre de nouveaux clichés pour compléter la reconstruction réalisée**. Cette notion est liée à la recherche de la meilleure vue suivante, dont il est question dans le § 13.5.

(4) Enfin, nous voulons, d'une part, comparer les résultats que nous obtenons avec ceux obtenus dans [Atienza 19, Panichev 19, Nathan 19] afin d'en comprendre plus finement les intérêts et désavantages les unes par rapport aux autres, pour, d'autre part, proposer de nouvelles approches permettant de prendre les avantages de chacune de ces approches. À plus long terme, nous pensons **exploiter des informations sémantiques pour affiner le traitement des différentes parties du squelette**.

13.3. Vers une combinaison de multiples vidéos désynchronisées avec ou sans recouvrement

Les travaux que nous avons présentés dans le chapitre 7 portent sur l'analyse de réseaux de caméras afin de déterminer les caméras qui ont des vues qui se recoupent. L'originalité des travaux que nous avons menés repose sur l'étude de la corrélation des activités entre vidéos, sans utiliser d'outils de suivi ou de ré-identification. Nous avons fait ce choix initial car nous avons estimé que nous n'avions pas les conditions idéales pour exploiter ces outils. Les objets présents dans les vidéos étaient de trop basses résolutions et surtout, les positions des caméras et leurs points de vue étaient trop différents. La détection d'objets est donc délicate et il est d'autant plus délicat d'en extraire un suivi fiable. Les résultats que nous avons présentés avec cette approche sans suivi et sans ré-identifcation sont encourageants mais nous souhaitons les enrichir.

Dans [Smeulders 03], les auteurs mettent en lumière les tendances pour effectuer un suivi dans de multiples vidéos. Ce problème est vu comme un problème de ré-identification en 3D, et dans ce contexte, les auteurs distinguent les trois aspects suivants qui peuvent être pris en considération :

(1) Faire de la reconnaissance et du suivi de primitives, c'est tout le sujet de la partie I ;

(2) Prendre en considération l'aspect temporel et l'accumulation d'informations dans le temps ;

(3) Effectuer une reconstruction 3D pour réaliser cette tâche en 3D.

À l'issue des travaux présentés, il semble opportun et original d'ajouter une information 3D à notre travail et, dans le cas de la reconnaissance de personnes, de distinguer les différentes parties du corps, c'est d'ailleurs ce que nous avons commencé à étudier, dans le cadre du projet européen VICTORIA, par une simple utilisation de l'algorithme existant OpenPose [Cao 17b], en 2D et en 3D. De plus, les nouveaux défis à venir portent sur la gestion des lieux très occupés, notamment avec la gestion de comportement des personnes observant la même scène [Mahmood 19], ou de foule en déplacement, comme des piétons [Gupta 18] pour anticiper des collisions. De manière classique, en vision par ordinateur, et plus particulièrement en stéréovision, pour effectuer une reconstruction 3D, il est nécessaire d'avoir des images de bonnes résolutions avec un recouvrement suffisant et la possibilité d'obtenir ou d'estimer les paramètres de calibrage de manière suffisamment fiable. Nous n'avions pas cette qualité de données en entrée dans le cadre des travaux que nous avons présentés au chapitre 7, c'est la raison pour laquelle, nous n'avons pas exploité la reconstruction 3D que nous pouvions estimer à partir des différentes caméras avec recouvrement. Dans ces travaux, nous avons également considéré des caméras statiques et faiblement désynchronisées. Nous avons également considéré qu'il n'y avait pas ou très peu d'interactions de la part des utilisateurs et enfin, nous n'avons pas établi les liens entre caméras dont les champs de vue ne se recoupent pas, mais qui ont un lien de dépendance (la sortie d'une caméra correspond à l'entrée dans une autre caméra, par exemple).

En conséquence, voici toutes les perspectives que nous envisageons pour ce travail et qui nous permettront une ouverture vers des travaux plus tournés vers l'utilisation de l'apprentissage profond et vers l'introduction des interactions utilisateurs :

(1) La **prise en compte du décor** doit nous permettre d'estimer les différences de points de vue entre caméras. Notamment, afin d'être capable d'estimer l'intérêt d'une vidéo par rapport à une autre, nous aimerions déterminer si leurs caméras sont orientées dans la même direction (côte à côte) ou au contraire, si elles sont face à face. Pour effectuer cette tâche, nous souhaitons étudier le décor, en faisant intervenir une segmentation sémantique de la scène. La contribution porterait sur l'exploitation de cette segmentation et non sur l'approche de segmentation en elle-même. Nous nous attendons à ce que des caméras placées côte à côté (et orientées dans la même direction) aient des segmentations sémantiques semblables, là où des caméras aux orientations différentes, voire qui se font face, risquent d'avoir moins d'éléments en commun. En terme de segmentation sémantique, dans la figure 13.3.(a), nous pouvons voir les vidéos, du jeu de données ToCaDa, se placer les unes par rapport aux autres en fonction de la similarité entre leurs segmentations sémantiques (il s'agit d'un résultat préliminaires que nous avons obtenus). Idéalement, à partir de ces segmentations sémantiques, nous aimerions établir un modèle de décor. Nous pourrions également envisager d'utiliser des techniques qui exploitent les vidéos multiples pour stabiliser la segmentation du décors, comme dans [Ventura 19b]. Cela implique que nous ayons un processus sélectif, c'est-à-dire que, uniquement dans les vidéos qui sont identifiées avec un recouvrement, nous affinons la segmentation du décor avec ce type d'approche. Enfin, une autre piste consiste à étudier les mouvements afin de déterminer la position des caméras et les travaux que nous avons proposés pour la reformulation de trajectoire peuvent contribuer à cette tâche.

(2) La **gestion des vidéos mobiles**, comme illustré dans la figure 13.3.(b), implique une philosophie de traitement assez différente de celle que nous avons mise en place. Il paraît évident que la première tâche consistera à trier les vidéos entre vidéos mobiles et vidéos statiques (cette tâche

peut être réalisée simplement en étudiant le mouvement global dans la vidéo). Le fait que la vidéo soit mobile implique que le recouvrement entre vidéos ne sera pas constant et stable, comme pour le cas des vidéos statiques. Cela implique également que le temps d'accumulation des évidences de recouvrement sera beaucoup plus court. Toutes ces difficultés accumulées indiquent qu'il sera nécessaire de donner plus d'importance à la reconnaissance et à la ré-identification d'objets, d'une part, et qu'il faudra s'appuyer sur un apprentissage plus conséquent, d'autre part. Enfin, il s'agit de cas limite où l'utilisateur devra jouer une part plus importante en indiquant éventuellement un objet d'intérêt repéré dans une vidéo. Ces vidéos mobiles pourront donc être utilisées en complément des vidéos statiques dans lesquelles un utilisateur indiquera un objet d'intérêt.

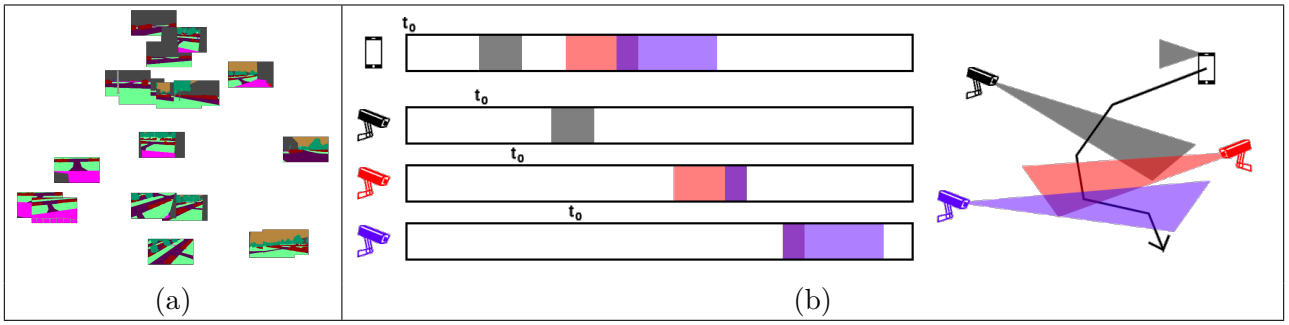


FIGURE 13.3. – Analyse de réseaux de caméras complexes avec prise en compte du décor – En (a), nous présentons des résultats préliminaires sur l'utilisation d'une segmentation sémantique de la scène afin de mettre en évidence les caméras qui ont une forte ressemblance par rapport au décor. En (b), nous illustrons les types de configuration qui n'ont pas été abordés dans nos travaux et que nous aimerions aborder. Cet exemple s'appuie sur un cas avec 3 caméras fixes et 1 caméra mobile, liée à un smartphone. Nous supposons ainsi que le champ de vue de la caméra du smartphone va croiser celui des autres caméras sur le chemin qu'il va emprunter, matérialisé par la flèche noire. Nous n'avons pas illustré le champ de vue de la caméra du smartphone car il est en mouvement et donc évolue. Cet exemple illustre également le cas des caméras qui n'ont pas des champs de vue qui se recoupent mais où il y a un lien car il est fort probable que les objets vus par la caméra noire soient vus par la caméra bleue après un certain délai. Enfin, les frises de temps à gauche mettent en lumière les problèmes de désynchronisation que nous aimerions prendre en compte.

(3) La **gestion des vidéos non-synchronisées** a pour conséquence d'identifier les problèmes de décalage entre vidéos. Les frises de temps, à gauche dans la figure 13.3.(b), illustrent ce cas. De la même façon que pour la gestion des vidéos mobiles, nous devons envisager dans un premier temps de trier les vidéos en groupe de vidéos synchronisées ou faciles à synchroniser (ce sont celles pour lesquelles nous avons l'information d'horodatage fiable) entre elles pour ensuite appliquer une stratégie de traitement de ces désynchronisations. Pour les vidéos où il est nécessaire de faire un travail de synchronisation avec les autres, nous pouvons imaginer utiliser la méthode que nous avons déjà mis en œuvre, mais, cette fois, il faut effectuer une recherche d'une part sur un sous-ensemble de la vidéo et, d'autre part, en supposant un décalage temporel bien plus important. Cela signifie que nous devons relever les défis suivants : être capable de déterminer un intervalle de temps suffisamment pertinent pour établir une comparaison, être capable de déterminer une stratégie de recherche rapide, car, si nous appliquons telle quelle la méthode actuelle, les temps de calcul seront très élevés.

(4) L'**ajout de la notion de profondeur (reconstruction 3D)** doit permettre d'améliorer les résultats. En effet, actuellement, il est possible d'estimer la profondeur, même dans des conditions non idéales, en utilisant un réseau de neurones génératifs adversaires, comme nous avons

introduit dans [Abdulwahab 20]. Le travail que nous aimerions aborder consiste à définir comment combiner cette information 3D approximative avec ce que nous avons déjà mis en place. Nous pouvons imaginer que cette information de profondeur donne une information sur les objets mais également les décors afin d'estimer la corrélation des personnes et des objets. Cette information pourrait participer au calcul de distance que nous utilisons. Nous pouvons d'ailleurs citer les travaux de [Ruiz-Hidalgo 12] qui permettent de combiner couleur et profondeur de caméras multiples pour réaliser une segmentation de la scène. De manière générale, être en capacité d'analyser les différents éléments statiques et dynamiques de la scène, en s'appuyant sur une reconstruction partielle de la scène, devrait nous permettre d'améliorer l'approche proposée. Dans ce sens, les travaux présentés dans [Jiang 17] constituent une approche intéressante pour obtenir cette analyse préliminaire.

(5) Il est nécessaire d'**ajouter les liens entre vidéos sans recouvrement mais avec un lien de dépendance**, comme c'est le cas entre les caméras noire et bleue de la figure 13.3. Pour effectuer cette tâche, nous pouvons à nouveau travailler sur les fonctions d'activité que nous avons définies. En effet, jusqu'à présent, nous avons travaillé sur la corrélation de ces fonctions tout au long de la vidéo et aux mêmes instants. À présent, il faut explorer cette concomitance d'activité mais avec un intervalle de temps suffisamment significatif pour dire qu'il y a bien une succession et non une simultanéité des activités. Cette tâche est délicate car elle augmente considérablement les calculs à réaliser. Toutefois, il paraît évident que ce travail doit être fait après avoir établi les liens directs et donc, cela réduira les temps de calcul. Cette description nous permet de mettre en évidence que ce traitement pourrait être assez lié ou similaire à celui que nous envisageons pour la gestion des vidéos non synchronisées.

(6) La **prise en compte de la détection des évènements** doit nous permettre d'augmenter la robustesse et de réduire les temps de calcul. En effet, jusqu'à présent, nous avons étudié toutes les activités, sans chercher à qualifier ces activités. Nous savons qu'en traitement de la vidéo, il y a de nombreux travaux qui portent sur l'identification des vues clés, *key frames*, pour extraire les moments les plus significatifs dans une vidéo [Liu 03] ou encore de recherche des éléments clés d'une scène, les éléments qui contiennent l'information la plus pertinente [Bellver Bueno 17]. Notre approche pourrait être adaptée en donnant plus d'importance à ces évènements pour identifier les concomitances d'évènements plutôt que de toutes les activités. Cette prise en compte des évènements devrait également nous permettre de trouver plus rapidement les liens entre vidéos sans recouvrement.

13.4. Combinaison et corrélation d'information multimédia

13.4.1. Corrélation de données au service du suivi médical post-cancer

Actuellement, de nombreuses informations sont disponibles pour suivre l'état d'un patient, en particulier, après un traitement pour un cancer. En effet, à présent, il est possible de collecter des informations sur les habitudes quotidiennes d'un patient : nourriture, activité sportive, activité sociale, qui sont des indicateurs des risques de rechute, et qui viennent compléter les informations cliniques d'autre part (imagerie, analyse biométrique, analyse génétique). Les liens entre ces indicateurs de vie quotidienne (hygiène alimentaire et état psychique) et les risques de rechute commencent à être mis en avant mais il est nécessaire d'avoir une validation statistique importante d'une part, et de lier de manière plus précise ces différents indicateurs avec les risques de rechute afin d'être en mesure de prédire les risques et également d'informer le patient et de l'aider à faire évoluer son comportement pour minimiser ses risques de rechute. Ainsi, nous souhaitons analyser et trouver des corrélations entre tous les types de données sur le patient (état physique, médical, social, comportemental, cognitif et émotionnel) afin de fournir, cf. schéma de la figure 13.4 :

(1) Une mesure du risque de rechute ;

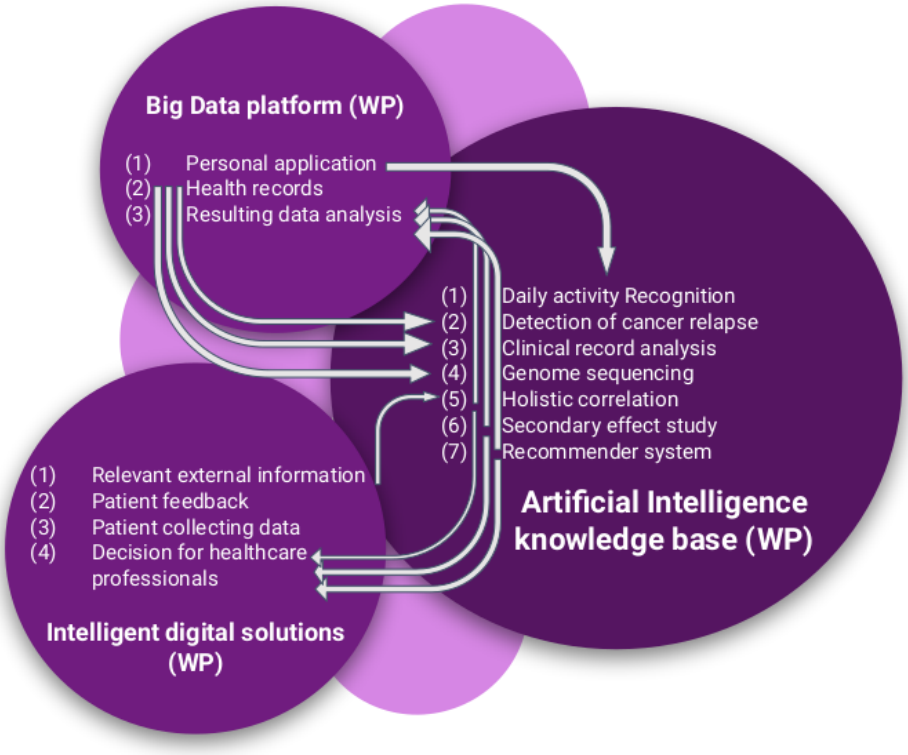


FIGURE 13.4. – Traitement des données relatives aux patients suivis après un cancer.

(2) Des recommandations personnelles et personnalisées aux patients sur son style de vie ;

(3) Une aide à la décision pour les médecins (traitement et suggestions aux patients) ;

(4) Une évaluation de la toxicité des traitements réalisés, et surtout de l'impact de ces recommandations sur la toxicité des traitements proposés.

Plus précisément, en terme de recherche, voici les points qui vont être étudiés :

- Classification automatique d'activités à partir de vidéos prises depuis des caméras "embarquées" sur le patient en utilisant de l'apprentissage profond ;
- Estimation des corrélations entre différents paramètres et mesures sur les patients et les rechutes ;
- Estimation du risque de rechute après traitement.

En terme de détection et de classification des activités quotidiennes, ce travail a déjà été initié et nous avons publié des résultat préliminaires dans [Sarker 19]. Ces travaux reposent sur l'utilisation d'une caméra de type *lifelogging*, c'est-à-dire une caméra portée par l'utilisateur et qui filme son quotidien. Cette définition même implique de nombreuses difficultés pour traiter ces données : une masse de données considérable à gérer, des images de qualité assez faible et avec des points de vue pas toujours avantageux en terme d'acquisition et/ou en terme de reconnaissance d'images (problèmes d'éclairage, manque de données, données inutiles du décor, variabilité des scènes acquises). Ainsi, la tâche de reconnaissance automatique des lieux visités par un patient est très délicate et naturellement, un traitement par apprentissage profond semble adapté à cette situation. Plus précisément, à partir des vidéos acquises, nous aimerions reconnaître les lieux visités et le type de nourriture consommée par un patient afin de déterminer sa qualité de vie.

En ce qui concerne l'estimation des corrélations entre différents paramètres et mesures, nous souhaitons explorer ces aspects en collaborant avec Sandrine MOUYSSET, équipe IRIT-APO.

13.4.2. Combinaison de vidéos au service d'une ville intelligente, confortable et sécurisée

Cette perspective applicative s'inscrit dans le cadre d'une collaboration avec l'entreprise `Kawantech`. La particularité de cette société est qu'elle a pour objectif d'exploiter des infrastructures déjà en place, à savoir les lampadaires publics, pour proposer des services d'aide à la gestion des éclairages, de manière à réduire l'impact écologique, économique des éclairages publiques. De manière plus générale, cette société souhaite proposer des services étendus, comme par exemple, la recherche de places de parking disponibles ou encore la détection de défauts de la chaussée pour la maintenance mais également l'aide à la navigation autonome de véhicules, pour avant tout améliorer la sécurité des vulnérables (piétons, cyclistes ou autres déplacements doux comme les trottinettes) et le confort des usagers (en évitant les défauts). Pour effectuer cette tâche, la société utilise des outils classiques de la vision par ordinateur : détection de mouvement, de changements, suivi d'objets d'intérêt, segmentation de la chaussée. Dans ce contexte, la société a déjà fait une demande d'expertise. Un aspect très important des travaux menés dans cette société est le respect de la vie privée et notamment du Règlement Général sur la Protection des Données (RGPD) et donc, pour cela, il n'y a pas de connexion entre lampadaire pour communiquer et il n'y a pas de systèmes de reconnaissance de personnes, de ré-identification.

Une priorité des recherches envisagées concerne l'étude des risques de collisions entre véhicules et vulnérables (piétons, vélos et autres véhicules non motorisés). Pour réaliser cette tâche, il est primordial d'avoir une analyse des mouvements de ces éléments dans la scène. Il s'agit donc, en terme de vision, de reconnaître une forme d'intérêt puis de la suivre. Ce domaine est déjà très bien étudié en vision par ordinateur [Veeraraghavan 05, Aggarwal 11]. Nous avons d'ailleurs déjà amorcé des travaux en lien avec la détection de piétons [Rashwan 19b]. Les risques de collisions entre piétons correspondent également à un domaine actif de recherche [Sayed 13]. Récemment, des techniques s'appuyant sur l'introduction de réseaux de neurones ont été mises en œuvre en utilisant des réseaux de neurones récurrents [Alahi 16, Xu 18] ou encore des réseaux adversaires génératifs [Gupta 18]. Ces méthodes sont des références dans le domaine et permettent déjà de donner des réponses satisfaisantes pour les piétons. Elles ont été validées sur de nombreux jeux de données publiques. Ces approches prennent en compte les mouvements imprévisibles liés aux interactions entre groupe de personnes mais elles ne traitent pas les comportements d'autres types de véhicules non motorisés ou des interactions et possibles collisions avec des véhicules. En ce qui concerne la collision entre vulnérables et véhicules, des études ont été menées mais essentiellement en prenant en compte une seule caméra placée sur le véhicule [Yoshitake 18]. La thèse de [Perollaz 08] met cependant en évidence l'intérêt des systèmes coopératifs, notamment la communication entre véhicules pour signaler un obstacle.

Nous souhaitons étudier des mouvements/déplacements peu rapides (en dessous de $30km/h$), d'éléments de taille raisonnable (taille inférieure à celle d'un véhicule motorisé) en prenant en compte les différents capteurs disponibles. Ainsi, de nombreuses difficultés se posent :

(1) Les formes et les tailles des objets sont variables d'une catégorie à l'autre (piétons, vélos, autres).

(2) La forme et la taille d'un même objet peut varier dans le temps et suivant le point de vue.

(3) Les différents éléments peuvent être occultés partiellement ou totalement. Ce genre de situation engendre souvent des confusions (des échanges) entre les différents éléments suivis.

(4) Les conditions d'acquisition, en particulier l'éclairage, évoluent au cours du temps générant de nombreuses difficultés : image saturée, image sous exposée, présence de flou.

L'énumération de ces contraintes et de ces difficultés mettent en avant les verrous scientifiques de ce travail : il nous faut une approche permettant d'intégrer la redondance d'information des multiples capteurs tout en prenant en compte ces problèmes. La premier intérêt du travail proposé réside dans le fait de combiner des données de nature différente : vidéos/images acquises par des caméras de surveillance ou des caméras liées à la gestion de l'espace urbain, comme l'éclairage, qui possèdent des points de vue très différents. Le second intérêt est de procéder de manière différente par rapport à

la littérature, en commençant par identifier dans le réseau de capteurs les liens entre les différents capteurs, comme dans [Malon 19b] : recouvrement ou non, délai ou lien entre capteurs et d'utiliser cette information pour effectuer le suivi des objets. Cette façon de procéder est complémentaire à la tâche de ré-identifciation et nous souhaitons étudier ces deux approches afin de les comparer puis de les utiliser de manière complémentaire.

Ainsi, en ce qui concerne la construction de descripteurs pour réaliser le suivi classique, une piste envisageable concerne l'utilisation d'une base de données d'objets 3D et de descripteurs de forme exploitant la courbure en 2D et en 3D comme dans nos travaux précédents [Rashwan 19a]. Une autre piste consisterait à extraire les silhouettes de ces vulnérables selon une méthode basée sur des modèles spatio-colorimétriques de l'arrière-plan et de l'avant-plan. Plusieurs descripteurs couleur peuvent être développés afin de caractériser les silhouettes extraites des images couleur avec un double critère utilisant à la fois des informations spatiales et colorimétriques. En sortie de cette partie, nous aurons donc les résultats des détections des vulnérables, leurs silhouettes et une signature compacte et globale caractérisant leur passage sous une caméra. Ces informations peuvent alors coopérer avec des informations issues d'une autre caméra à des fins de ré-identification. Enfin, nous souhaitons construire des graphes de lien entre vidéos/capteurs en exploitant les détections d'objets dans les images, en exploitant des détections mono-images, comme nous l'avons réalisé dans les travaux précédents [Malon 19b]. L'intérêt de calculer ce graphe est de permettre, *a priori*, de limiter la recherche pour la ré-identification, et/ou *a posteriori*, de corriger les erreurs de ré-identification. Cette partie suppose qu'une synchronisation des acquisitions d'images a été préalablement réalisée.

13.5. Vers la reconstruction d'un modèle 3D complet de qualité garantie

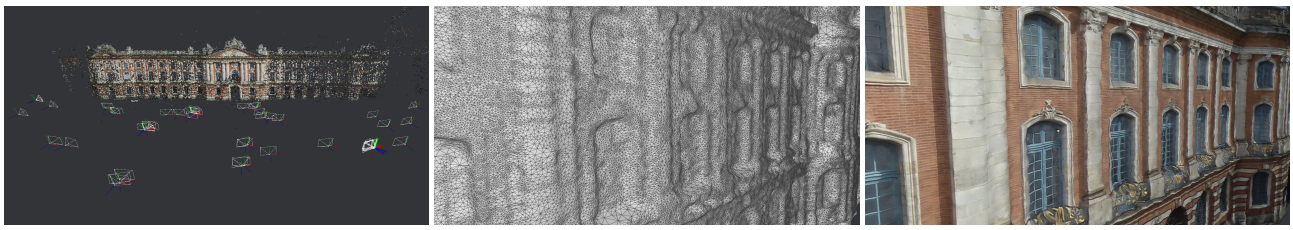


FIGURE 13.5. – Exemple de reconstruction 3D du Capitole de Toulouse par la technique MVS [AliceVision 18] – Dans l'image de gauche, nous pouvons remarquer une densité plus faible de points reconstruits sur la partie droite du bâtiment. Cette différence est due au nombre insuffisant de prises de vue. Cela ne permet pas de reconstruire les détails géométriques (image au centre) et photométriques (à droite).

Dans ce manuscrit, d'une part, nous avons manipulé des modèles 3D, cf. chapitres 3 et 4, et, d'autre part, nous avons travaillé sur l'étape de mise en correspondance, indispensable à la reconstruction, aussi bien au chapitre 4, que dans la partie I. Les perspectives présentées dans les § 13.1 et 13.2, nous ont également permis d'aborder les notions d'estimation de la pose, d'une part, et de reconstruction d'objets, d'autre part, mais dans des cas particuliers, marqueurs fiduciaires pour l'un et objets tubulaires pour l'autre. À présent, nous nous intéressons à la reconstruction de modèle 3D à partir d'un ensemble d'images. Nous ne souhaitons pas limiter ce nombre d'images, contrairement aux travaux présentés au § 13.2, mais, cette fois, nous souhaitons guider l'utilisateur pour qu'il puisse choisir le point de vue qui lui permettra de compléter au mieux son modèle 3D. De plus, nous savons qu'il est possible de générer des modèles 3D aussi bien avec des images qu'avec des données 3D ou RGB-D et nous souhaitons prendre en compte tous ces types de données. Enfin, dans cette section, quand nous parlons de modèles 3D manipulés, il s'agit de modèles surfaciques représentés par un maillage.

Il existe déjà de nombreuses approches performantes [Furukawa 15] permettant même de générer des modèles à très grande échelle [Heinly 15] à partir d'un ensemble quelconque d'images [Snavely 06] ou d'un flux vidéo [Taketomi 17]. De plus, les techniques actuelles s'appuient sur des données de plus en plus multimodales [Kim 14]. Toutefois, peu de travaux abordent le problème de la qualité du modèle obtenu. En effet, toutes ces techniques font appel à une acquisition manuelle des données et le choix des prises de vue est laissé à l'expertise de l'opérateur qui suit des règles qualitatives générales [Wenzel 13] pour trouver un compromis entre le recouvrement des captures et l'acquisition des nouvelles parties de la scène. Le modèle final produit peut ainsi être incomplet, soit parce que les données acquises ne couvrent pas suffisamment le modèle réel, nous parlons de données manquantes, soit parce que certaines parties de l'objet sont plus difficiles que d'autres à reconstruire automatiquement par manque de recouvrement ou en présence d'occultations. Une autre difficulté est liée au fait d'avoir des résolutions différentes, c'est-à-dire un niveau de détails différent, de la géométrie et de la texture, dans les diverses parties de la scène. En conséquence, le modèle produit sera de qualité inégale, comme nous pouvons le constater sur l'exemple de la figure 13.5.

Ainsi, nous souhaitons aborder le problème de la qualité du modèle et plus précisément le problème du choix des prises de vue optimales pour générer un modèle complet et avec des résolutions géométrique et photométrique garanties et uniformes. Dans la littérature, ce problème correspond au choix de la meilleure prise de vue suivante ou *Next Best View* (NBV) [Connolly 85]. Il est important de distinguer les approches qui peuvent être utilisées en intérieur, comme celle abordée au § 13.2, des approches en extérieur, comme celles que nous souhaitons étudier. En intérieur, notamment dans la communauté de la robotique, ce problème a été largement étudié [Scott 03]. Dans ce contexte, il est nécessaire de planifier les positions des prises de vue de manière optimale pour garantir une numérisation complète d'un objet ou d'un environnement tout en minimisant le nombre d'acquisitions [Khalfaoui 13]. Les méthodes proposées dans ce contexte sont spécifiques au type de capteur 3D utilisé. De plus, contraintes par les limitations des manipulateurs robotiques, elles sont adaptées pour des tailles d'objets relativement réduites [Karaszewski 16]. Enfin, peu de travaux utilisent l'information photométrique, couleur ou texture, pour compléter le modèle [Banerjee 18].

Dans le contexte de la reconstruction par images, les auteurs de [Hoppe 12] introduisent une approche interactive pour améliorer la qualité du modèle 3D reconstruit en indiquant les régions du modèle de faible résolution géométrique. Toutefois, ils n'identifient pas la prise de vue optimale. Pour cette tâche, [Langguth 13] introduit une approche pour guider un utilisateur non expert en proposant des NBV limitées à la sphère autour de l'objet. L'approche de [Golodetz 18] est une méthode collaborative entre plusieurs agents, mais ne garantit pas une reconstruction finale complète. L'étude proposée dans [Haner 12] utilise l'estimation de l'incertitude des positions des caméras et des points reconstruits pour propager la covariance des erreurs et sélectionner la meilleure vue suivante qui minimise l'incertitude. Cependant, elle n'est pas adaptée pour des scènes de grandes dimensions et ne garantit pas la complétude géométrique et photométrique. Plus récemment, le problème de la qualité du maillage a été abordé dans [Morreale 19] mais cette approche n'assure pas la complétude géométrique et photométrique. Enfin, des approches récentes [Mendoza 20] utilisent l'apprentissage profond pour déterminer les meilleures prises de vue pour la reconstruction des objets mais elles sont limitées car elles supposent la connaissance *a priori* de la forme ou de la classe de l'objet à reconstruire et ne peuvent pas être facilement généralisées à des scènes quelconques. Enfin, nous pouvons citer des approches qui combinent des données de nature différente ainsi qu'une sursegmentation du maillage pour reconstruire le modèle 3D [Bódis-Szomorú 15].

Ainsi, dans ce contexte bibliographique, nous souhaitons proposer une méthode pour assurer la complétude du modèle 3D généré, en terme de géométrie et d'apparence photométrique. Nous souhaitons définir des critères permettant de déterminer des points de vue complémentaires et avec suffisamment de recouvrement avec les vues précédentes pour compléter la reconstruction. Nous voulons proposer des critères qui soient le plus possible indépendants du type de capteur et des données utilisées. En général, les données en entrée seront des images couleur ou des données 3D avec l'information

photométrique pour chaque point 3D, issues d'un capteur RGBD ou d'un LIDAR [5]. Nous supposons que nous travaillons avec un premier ensemble de prises de vue et un modèle 3D préliminaire et potentiellement incomplet sur lequel formuler les critères pour la recherche de prises de vue complémentaires.

Nous envisageons la possibilité d'utiliser une analyse sémantique de la scène [Minaee 20] pour extraire des informations utiles pour mieux guider le choix des prises de vue. Par exemple, détecter et reconnaître la classe des objets dans la scène pourra donner des informations supplémentaires à prendre en compte dans les critères pour bénéficier d'*a priori* sur la scène comme les propriétés des objets, la régularité des éléments présents, la répétition de motifs 3D, la structure de la scène, les types de texture, la topologie et les relations entre les objets, et la forme typique, appelée *template*, des objets. Ces contraintes peuvent guider la recherche de prises de vue mais aussi être utilisées pendant ou en aval de la reconstruction pour détecter les primitives géométriques caractérisant la scène et pour simplifier la représentation du modèle 3D [Monszpart 15, Morreale 19].

13.6. Vers une utilisation conjointe et uniformisée de points saillants en 2D et en 3D

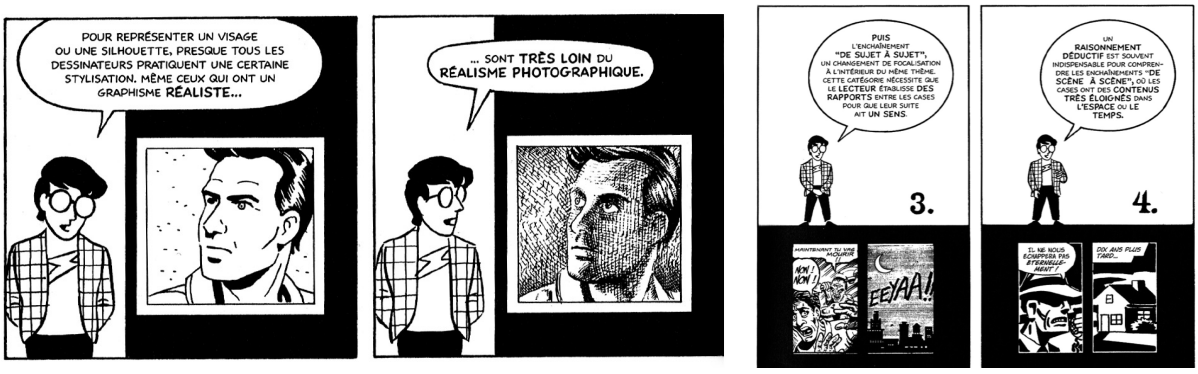


FIGURE 13.6. – Illustrations de la compression spatiale et temporelle d'images – Pour illustrer nos propos, nous utilisons des extraits de « l'art invisible » [McCloud 93]. À gauche : le dessin stylisé réduit l'information sémantique par rapport à une image complète, en gardant les éléments saillants pour la compréhension. À droite : Une information temporelle très réduite, ici avec seulement 2 vignettes et un texte très court, une action complète de l'attaque d'un personnage (couple de vignette à gauche) est racontée. Cette narration est relativement implicite.

Lorsque nous souhaitons utiliser, manipuler et partager un objet 3D potentiellement dynamique, il est possible de recueillir simplement de nombreuses données, de nature différente : images ou vidéos, séquences de modèles 3D, ou de cartes de profondeur. Nous l'avons illustré dans les travaux exposés dans la partie I. Nous savons que les modèles 3D sont intrinsèquement indépendants du point de vue alors que les images et cartes de profondeurs sont liées à un point de vue. De plus, nous avons montré que la littérature est abondante sur la caractérisation de ses données par l'extraction de points d'intérêt ou de régions saillantes : aussi bien dans des images, en 2D, cf. chapitre 2, que dans des modèles 3D, cf. chapitre 3. Nous avons déjà introduit des approches permettant de combiner des données dans différentes modalités, cf. § 3.3 et chapitre 4.

Pour manipuler et utiliser ces différentes données, il est également important de distinguer les similarités ou les répétitions dans ces données. Nous avons initié ce travail au § 3.2 et nous savons qu'il existe des approches également en 2D [Chen 13]. Nous souhaiterions donc effectuer une étude de ces

5. *LIght (ou Laser Imaging) Detection And Ranging.*

répétitions/ces similarités, en introduisant une étude et une méthode analogue à ce que nous avons réalisé en détection et recalage 2D/3D dans le chapitre 4.

Dans la partie I, nous avons abordé la notion de détection multimodale sans prendre en compte les aspects temporels. Il existe des travaux permettant de prendre en compte ces aspects pour détecter une région ou un point d'intérêt, en 2D [Laptev 04], et également en 3D [Ahmed 08].

Enfin, avec les avancées récentes en apprentissage profond, de nouvelles techniques de détection et caractérisation s'appuient sur des outils d'apprentissage, aussi bien en 2D, avec des réseaux convolutionnels siamois pour apprendre simultanément similarités et dissimilarités [Simonyan 15, Tian 17], qu'en 3D [Ghadai 19].

Le travail que nous souhaitons développer s'appuie sur une analogie à la bande dessinée (BD) et, plus particulièrement la manière très compressée, mais pour autant tout aussi chargée de sens, de représenter les données. Une BD arrive à raconter une histoire, faire vivre et évoluer des personnages, des lieux avec parfois seulement quelques traits et paroles, cf. figure 13.6. En s'inspirant des processus de la narration, en particulier celui de la BD, nous voulons extraire des données spatiales et/ou temporelles une information de saillance pour représenter le sens de ces données de façon compacte, efficace et pertinente.

Il existe de nombreuses applications qui nécessitent la détection, la caractérisation et la mise en correspondance de points saillants, comme la reconstruction 3D, l'analyse de formes ou la compression. Dans un premier temps, nous nous intéresserons particulièrement au cas de la compression où nous cherchons à déterminer à la fois les points saillants et leurs répétitions.

13.7. Vers une mise en correspondance et une segmentation conjointe

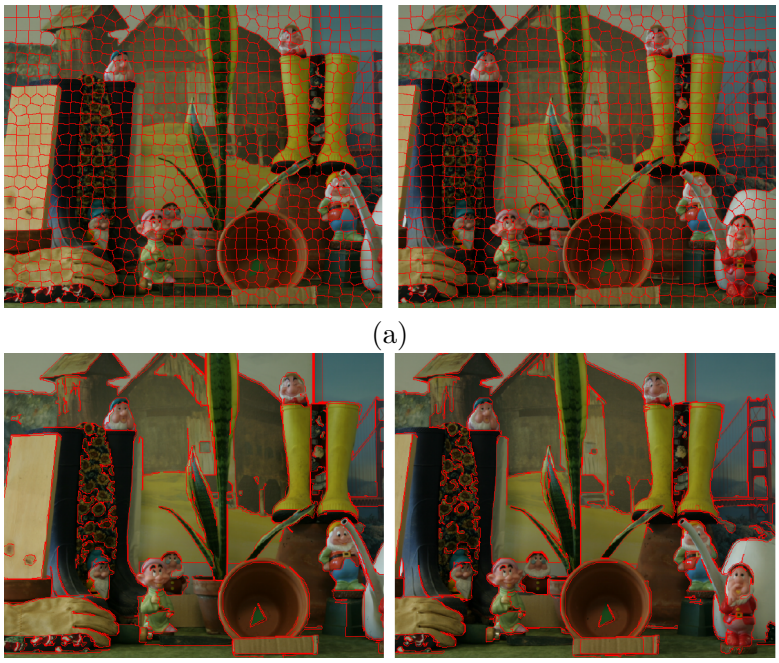
Toute une partie de ce manuscrit a été dédiée à la mise en correspondance, cf. partie II, et en particulier la mise en correspondance stéréoscopique. Nous avons montré l'intérêt de fusionner différentes mesures de corrélation, cf. § 6.2, ainsi que l'intérêt de combiner segmentation et mise en correspondance, cf. sections 6.3 et 6.4. D'autres travaux sont également en lien, notamment la segmentation de scènes urbaines, cf. chapitre 9, où nous avons utilisé conjointement une sur-segmentation et une mise en correspondance de points d'intérêt. Nous avons également commencé à étudier la possibilité de combiner mise en correspondance et superpixels [6]. Nous pouvons imaginer l'introduction des superpixels dans le processus de mise en correspondance de deux façons différentes :

1. Nous réalisons une mise en correspondance de superpixels et ces correspondances sont directement utilisées pour effectuer l'appariement au niveau du pixel.
2. Nous ajoutons des contraintes à la mise en correspondance pixellique en introduisant/propageant l'information de sur-segmentation, à la manière de ce qui a été introduit dans les sections 6.3 et 6.4. Cela signifie que nous ne cherchons pas à apparier les superpixels.

Le premier cas s'appuie sur l'hypothèse que la construction des superpixels est cohérente d'une image à l'autre dans une séquence. Notre étude des superpixels nous montre qu'il est difficile d'obtenir des superpixels qui se correspondent parfaitement, cf. les exemples fournis dans la figure 13.7 sur des images extraites de la dernière base de données mise à disposition par Middlebury [Scharstein 14].

Dans le second cas, il s'agit souvent d'approches appliquées pour améliorer les cartes de profondeur extraites par un capteur 3D ou plus précisément pour améliorer l'interpolation de cette profondeur. En effet, il est connu que l'estimation de ces profondeurs peut être assez grossière et entâchée d'erreurs [Matsuo 13, Cambra 14]. Dans ce cadre applicatif, cette segmentation est vue comme une segmentation des différentes surfaces. Dans le domaine de la mise en correspondance stéréoscopique par

6. C'était le sujet de stage de master 2 d'Emilie Jalras, en 2015. Tous les résultats préliminaires que nous présentons dans cette section sont issus de son travail.



(a)

(b)

FIGURE 13.7. – Exemples de superpixels calculés sur un couple stéréoscopique – Nous avons calculés les superpixels de l'approche [Achanta 12], en (a), et de l'approche [Felzenszwalb 04], en (b). Nous avons volontairement choisi des constructeurs très différents et nous avons conservé le même jeu de paramètres pour les deux images du couple stéréoscopique.

corrélation, nous supposons que les voisins du pixel étudié ont la même disparité ou appartiennent au même objet. Or cela n'est pas toujours le cas. C'est la raison pour laquelle dans de nombreuses publications les auteurs ont proposé d'utiliser une fenêtre adaptative prenant en compte des éléments de contours ou de segmentations. L'arrivée des superpixels, dont l'objectif est de proposer une sursegmentation de l'image prenant en compte les détails les plus fins permet d'affiner ce concept de fenêtre adaptative, puisqu'il est envisageable d'ajuster de manière encore plus fine la manière de sélectionner le voisinage impliqué. Ainsi, nous pouvons citer l'approche locale de [Tong 10] qui s'appuie sur l'utilisation de trois niveaux de segmentation différents. Le premier niveau de segmentation doit permettre de segmenter les différentes parties de la scène, en respectant les contours. Le second niveau est un affinement des régions détectées au premier niveau et doit fournir les différents objets présents dans la scène segmentée. Enfin, le dernier niveau correspond à un découpage en superpixels encore plus fins (qui doit logiquement contenir des sous-parties des objets présents). La mise en correspondance s'effectue ensuite en appliquant différents algorithmes d'appariement en séquence. Le premier algorithme d'appariement permet la recherche de superpixels correspondants dans la seconde image (l'algorithme ne s'appuie que sur les superpixels détectés dans l'image initiale) alors que le second va affiner les calculs au pixel en utilisant les régions correspondantes trouvées. Les auteurs ont également introduit une approche globale où des superpixels sont utilisés [Xie 10]. Ils ont en particulier évalué l'influence du type de sursegmentation utilisé.

Toutes ces approches nous a amené à effectuer des premiers tests afin d'améliorer les techniques de mise en correspondance robuste mises en œuvre durant la thèse, [Chambon 05a, Chapitre 3]. En effet, la mesure de corrélation proposée et mise en avant s'appuie sur une utilisation partielle de la fenêtre de corrélation car nous supposons que dans le cas d'une occultation, une partie de cette fenêtre ne doit pas être prise en compte. Pour faire le choix entre pixels à conserver et pixels à rejeter, nous avons utilisé uniquement un critère photométrique. Il nous semble que ce travail peut être enrichi en

prenant en compte un critère lié à une segmentation en régions. En effet, nous avons envisagé d'utiliser le superpixel comme support pour affiner le voisinage pris en compte dans une approche de mise en correspondance par corrélation. Plus précisément, les pixels rejetés correspondent à ceux qui ne font pas partie du même superpixel, comme nous pouvons le voir dans la figure 13.8.

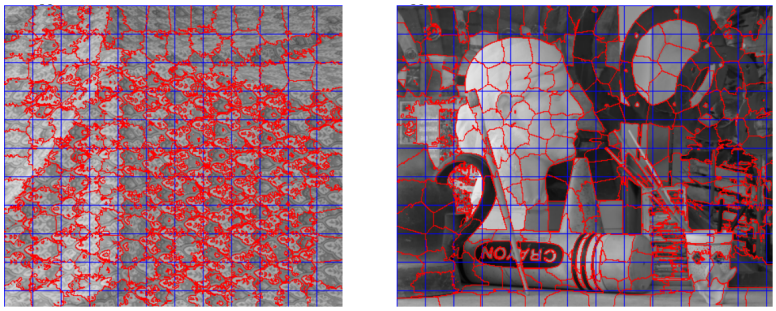
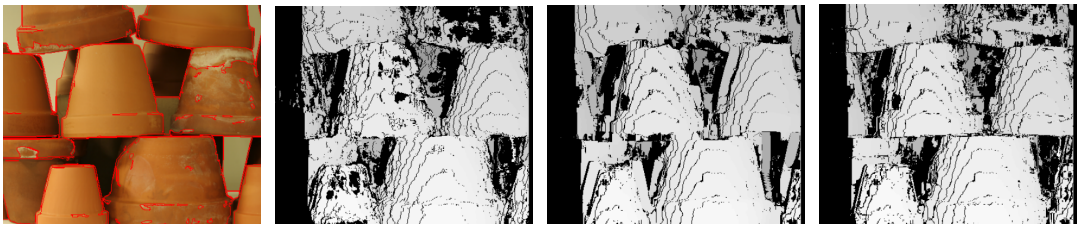


FIGURE 13.8. – Exemple de superpixels utilisés pour effectuer la mise en correspondance – Nous avons calculé les superpixels de l'approche [Achanta 12], en rouge, et nous illustrons la différence avec certaines des fenêtres de corrélation classique, notées en bleues, utilisées dans une approche par corrélation classique.

Les premiers résultats que nous avons obtenus mettent en évidence le fait qu'utiliser une sur-segmentation comme celle de [Felzenszwalb 04], qui favorise la similarité photométrique plus que la régularité des superpixels, nous permet d'obtenir les résultats les plus performants en terme de pourcentage d'appariements corrects. De plus, cela permet d'améliorer les résultats particulièrement pour des scènes complexes comme celle présentée dans la figure 13.9, c'est-à-dire, des scènes faiblement texturées et comportant des motifs répétitifs.



(a) (b) (c) (d)

FIGURE 13.9. – Résultat de mise en correspondance obtenu en combinant approche par corrélation et sur-segmentation en superpixels – En (a), nous présentons une image du couple stéréoscopique avec en rouge les superpixels calculés par [Felzenszwalb 04] que nous utilisons. En (b) et (c), il s'agit des résultats classiques obtenus avec les mesures SMAD et ZNCC respectivement, alors qu'en (d), il s'agit du résultat obtenu en combinant superpixels et mise en correspondance classique avec ZNCC. Les images utilisées font partie du jeu de données publié dans [Scharstein 14], que nous n'avons pas utilisé dans les travaux de la partie II, car, les travaux de cette partie sont antérieurs à 2014.

L'intérêt de ce type d'approche combinant superpixels et algorithmes de mise en correspondance a, par ailleurs, été démontré et renforcé par la publication de [Xue 19] car, l'approche proposée obtient des résultats qui surpassent ceux obtenus par des approches par réseaux de neurones. Dans le contexte de l'apprentissage profond, des approches faisant intervenir une segmentation sémantique ont également été développées, en s'appuyant sur une mise en correspondance partielle par points d'intérêt dans [Yang 18]. La segmentation sémantique exploitée est directement intégrée dans le coût des contraintes du coût de mise en correspondance globale.

Nous souhaitons donc poursuivre ces travaux, dans un contexte d'apprentissage profond, en exploitant un coût de mise en correspondance faisant intervenir la notion de superpixels, voire la notion de segmentation sémantique, dans la suite de ce que nous avons étudié et dans la suite de ce qui a été réalisé dans la littérature en mise en correspondance avec utilisation d'apprentissage profond par réseaux de neurones.

Bibliographie

[Aanæs 12] H. Aanæs, A. Lindbjerg Dahl et K. Steenstrup Pedersen. *Interesting Interest Points. Internationl Journal on Computer Vision, IJCV*, 97(1):18–35, 2012.

[Abdulwahab 19] S. Abdulwahab, H. Rashwan, J. Cristiano, S. Chambon et D. Puig. *Effective 2D/3D Registration using Curvilinear Saliency Features and Multi-Class SVM.* Dans International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP, pages 354–361, 2019.

[Abdulwahab 20] S. Abdulwahab, H. A. Rashwan, M. A. García, M. Jabreel, S. Chambon et D. Puig. *Adversarial Learning for Depth and Viewpoint Estimation from a Single Image. IEEE Transactions on Circuits and Systems for Video Technology, CSVT*, 2020.

[Achanta 12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua et S. Süsstrunk. *SLIC Superpixels Compared to State-of-the-art Superpixel Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 34(11):2274–2282, 2012.

[Achanta 17] R. Achanta et S. Susstrunk. *Superpixels and polygons using simple non-iterative clustering.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2017.

[Acharya 19] D. Acharya, K. Khoshelham et S. Winter. *BIM-PoseNet: Indoor camera localisation using a 3D indoor model and deep learning from synthetic images. International Society for Photogrammetry and Remote Sensing, ISPRS*, 150:245–258, 2019.

[Acosta 92] J. J. Acosta, L. Adolfo et R. L. Mullen. *Low-Cost Video Image Processing System for Evaluating Pavement Surface Distress. Journal of the Transportation Research Board, TRB Journal*, 1348:63–72, 1992.

[Agarwal 11] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. Seitz et R. Szeliski. *Building Rome in a Day. Communications of the Association for Computing Machinery, ACM*, 54(10):105–112, 2011.

[Aggarwal 11] J. K. Aggarwal et M. S. Ryoo. *Human Activity Analysis: A Review. ACM Computing Surveys*, 43(3):1896–1909, 2011.

[Ahmed 08] N. Ahmed, C. Theobalt, C. Rossl, S. Thrun et H. Seidel. *Dense correspondence finding for parametrization-free animation reconstruction from video.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2008.

[Ahmed 09] M. Ahmed et C. Haas. *The Potential of Low Cost Close Range Photogrammetry towards Unified Automatic Pavement Distress Surveying.* Dans Transportation Research Board Meeting, TRB Meeting, 2009.

[Ahmed 18] S. M. Ahmed, Y. Z. Tan, C. M. Chew, A. A. Mamun et F. S. Wong. *Edge and Corner Detection for Unorganized 3D Point Clouds with Application to Robotic Welding.* Dans IEEE International Conference on Intelligent Robotic Systems, IROS, pages 7350–7355, 2018.

[Alahi 16] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei et S. Savarese. *Social LSTM: Human trajectory prediction in crowded spaces.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 961–971, 2016.

[Alexandre 12] L . A Alexandre. *3D Descriptors for Object and Category Recognition: a Comparative Evaluation.* Dans IEEE International Conference on Intelligent Robotic Systems, IROS, Workshop on Color-Depth Camera Fusion in Robotics, 2012.

[AliceVision 18] AliceVision. Meshroom: A 3D reconstruction software. 2018.

[Amhaz 13a] R. Amhaz, S. Chambon, J. Idier et V. Baltazart. *Automatic Road Crack Detection Based on a Shortest-Path Algorithm.* Dans actes du colloque GRETSI sur le traitement du signal et des images, 2013.

[Amhaz 13b] R. Amhaz, S. Chambon, J. Idier et V. Baltazart. *Une méthode de relevé automatique de fissures de chaussée fondée sur la recherche de chemins minimaux.* Dans Colloque du club Contrôle et Mesure Optique pour l'Industrie, CMOI, 2013.

[Amhaz 14] R. Amhaz, S. Chambon, J. Idier et V. Baltazart. *A new minimal path selection algorithm for automatic crack detection on pavement images.* Dans International Conference on Image Processing, ICIP, 2014.

[Amhaz 15] R. Amhaz. *Détection automatique de fissures dans des images de chaussée par sélection de chemins minimaux.* Thèse de doctorat, École Centrale, Nantes, 2015.

[Amhaz 16] R. Amhaz, S. Chambon, J. Idier et V. Baltazart. *Automatic Crack Detection on Two-Dimensional Pavement Images: An Algorithm Based on Minimal Path Selection. IEEE Transactions on Intelligent Transportation Systems, TITS*, 17(10):2718–2729, 2016.

[Armagan 17] A. Armagan, M. Hirzer, P. M. Roth et V. Lepetit. *Accurate Camera Registration in Urban Environments Using High-Level Feature Matching.* Dans British Machine Vision Conference, BMVC, 2017.

[Aschwanden 92] P. Aschwanden et W. Guggenbül. *Experimental results from a comparative study on correlation type registration algorithms.* Dans W. Förstner et S. Ruwiedel, éditeurs, Robust computer vision: Quality of Vision Algorithms, RCVQVA, pages 268–282. Wichmann, 1992.

[Atienza 19] R. Atienza. *Pyramid U-Network for Skeleton Extraction from Shape Points.* Dans IEEE Conference on Computer Vision Pattern Recognition Workshop, CVPR Workshop, 2019.

[Aubry 14] M. Aubry, D. Maturana, A. Efros, B. Russell et J. Sivic. *Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2014.

[Aujol 06] J.-F. Aujol, G. Gilboa, T. Chan et S. Osher. *Structure-Texture Image Decomposition—Modeling, Algorithms, and Parameter Selection. Internationl Journal on Computer Vision, IJCV*, 67(1):111–136, 2006.

[Badrinarayanan 17] V. Badrinarayanan, A. Kendall et R. Cipolla. *SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 39(12):2481–2495, 2017.

[Bakkay 18a] M. C. Bakkay, S. Chambon, C. Lubat et S. N. Barsotti. *Automatic detection of individual and touching moths from trap images by combining contour-based and region-based segmentation. IET Computer Vision*, 12(2):138–145, 2018.

[Bakkay 18b] M. C. Bakkay, S. Chambon, C. Lubat et S. N. Barsotti. *Support Vector Machine (SVM) Recognition Approach adapted to Individual and Touching Moths Counting in Trap Images.* Dans Visual observation and analysis of Vertebrate And Insect Behavior Workshop, VAIB, 2018.

[Balntas 16] V. Balntas, E. Riba, D. Ponsa et K. Mikolajczyk. *Learning local feature descriptors with triplets and shallow convolutional neural networks.* Dans British Machine Vision Conference, BMVC, 2016.

[Banerjee 18] D. Banerjee, K. Yu et G. Aggarwal. *Robotic Arm Based 3D Reconstruction Test Automation. IEEE Access*, 6:7206–7213, 2018.

[Barnes 09] C. Barnes, E. Shechtman, A. Finkelstein et D. B. Goldman. *PatchMatch: a randomized correspondence algorithm for structural image editing. ACM transactions on graphics,TOG*, 28:24:1–24:11, juillet 2009.

[Bartoli 07] A. Bartoli, M. Perriollat et S. Chambon. *Generalized Thin-Plate Spline Warps.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2007.

[Bartoli 10] A. Bartoli, M. Perriollat et S. Chambon. *Generalized Thin-Plate Spline Warps. Internationl Journal on Computer Vision, IJCV*, 88(1):85–110, 2010.

[Bauchet 18] J.-P. Bauchet et F. Lafarge. *KIPPI: KInetic Polygonal Partitioning of Images.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2018.

[Bauda 13] M.-A. Bauda, S. Chambon, M. Spangenberg et V. Charvillat. *Segmentation de scènes urbaines par combinaison d'information.* Dans actes du congrès francophone de Vision par Ordinateur, ORASIS, 2013.

[Bauda 15a] M.-A. Bauda, S. Chambon, P. Gurdjos et V. Charvillat. *Geometry-based Superpixel Segmentation - Introduction of Planar Hypothesis for Superpixel Construction.* Dans International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP, pages 227–232, 2015.

[Bauda 15b] M.-A. Bauda, S. Chambon, P. Gurdjos et V. Charvillat. *Image Quality Assessment for Photo-consistency Evaluation on Planar Classification in Urban Scenes.* Dans International Conference on Pattern Recognition Applications and Methods, ICPRAM, pages 328–333, 2015.

[Bauda 16] M.-A. Bauda. *Compréhension de scènes urbaines par combinaison d'information 2D/3D.* Thèse de doctorat, Institut National Polytechnique de Toulouse, INPT, 2016.

[Bay 08] H. Bay, A. Ess, T. Tuytelaars et L. Van Gool. *Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding, CVIU*, 110(3):346–359, 2008.

[Beaudet 78] P. R. Beaudet. *Rotationnally invariant image operators.* Dans International Conference on Pattern Recognition, ICPR, pages 579–583, 1978.

[Beil 97] W. Beil, K. Rohr et H. S. Stiehl. *Investigation of Approaches for the Localization of Anatomical Landmarks in 3D Medical Images.* Dans Computer Assisted Radiology and Surgery, CARS, pages 265–270, 1997.

[Bellver Bueno 17] M. Bellver Bueno, X. Giró-i-Nieto, F. Marqués et J. Torres. *Hierarchical object detection with deep reinforcement learning. Deep Learning for Image Processing Applications*, 31(164):3, 2017.

[Belongie 02] S. Belongie, J. Malik et J. Puzicha. *Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 24(4):509–522, 2002.

[Bengio 09] Y. Bengio. *Learning deep architectures for AI. Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[Besl 92] P. J. Besl et N. D. McKay. *A Method for Registration of 3-D Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 14(2):239–256, 1992.

[Betke 03] M. Betke, H. Hong, D. Thomasa, C. Princea et J. P. Kob. *Landmark detection in the chest and registration of lung surfaces with an application to nodule registration. Medical Image Analysis, MICCAI Society*, 7(3):265–281, 2003.

[Bhat 98] D. N. Bhat et S. K. Nayar. *Ordinal Measures for Image Correspondence. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 20(4):415–423, 1998.

[Birchfield 98] S. Birchfield et C. Tomasi. *A pixel dissimilarity measure that is insensitive to image sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 20:401–406, 1998.

[Bleyer 08a] M. Bleyer, S. Chambon, U. Poppe et M. Gelautz. *Evaluation of different methods for using colour information in global stereo matching approaches.* Dans International Society for Photogrammetry and Remote Sensing, ISPRS, pages 63–68, 2008.

[Bleyer 08b] M. Bleyer et M. Gelautz. *Simple but effective tree structures for dynamic programming-based stereo matching.* Dans International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP, pages 415–422, 2008.

[Bleyer 09] M. Bleyer, M. Gelautz, C. Rother et C. Rhemann. *A stereo approach that handles the matting problem via image warping.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 501–508, 2009.

[Bleyer 10a] M. Bleyer et S. Chambon. *Does Color Really Help in Dense Stereo Matching?* Dans International Symposium of Three-Dimensional Data Processing, Visualization, and Transmission, 3DPVT, 2010.

[Bleyer 10b] M. Bleyer, C. Rother et P. Kohli. *Surface Stereo with Soft Segmentation.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 1570–1577, juin 2010.

[Bo 11] L. Bo, X. Ren et D. Fox. *Depth kernel descriptors for object recognition.* Dans IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 821–826, 2011.

[Bódis-Szomorú 15] A. Bódis-Szomorú, H. Riemenschneider et L. Van Gool. *Superpixel meshes for fast edge-preserving surface reconstruction.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 2011–2020, 2015.

[Boser 92] B. E. Boser, I. M. Guyon et V. N. Vapnik. *A Training Algorithm for Optimal Margin Classifiers.* Dans Annual Workshop on Computational Learning Theory (COLT), pages 144–152, 1992.

[Bouchafa 06] S. Bouchafa et B. Zavidovique. *Efficient cumulative matching for image registration. Image and Vision Computing, IVC*, 24(1):70–79, 2006.

[Bouchafa 09] S. Bouchafa, A. Patri et B. Zavidovique. *Efficient plane detection from a single moving camera.* Dans International Conference on Image Processing, ICIP, pages 3493–3496, 2009.

[Braquelaire 97] J. P. Braquelaire et L. Brun. *Comparison and optimization of methods of color image quantization. IEEE Transactions on Image Processing, TIP*, 6(7):1048–1052, 1997.

[Bray 06a] J. Bray, B. Verma, X. Li et W. He. *A Neural Network based Technique for Automatic Classification of Road Cracks.* Dans International Joint Conference on Neural Networks, JCNN, pages 907–912, 2006.

[Bray 06b] M. Bray, P. Kohli et P. H. S. Torr. *PoseCut: Simultaneous Segmentation and 3D Pose Estimation of Humans Using Dynamic Graph-Cuts.* Dans European Conference on Computer Vision, ECCV, volume Lecture Notes in Computer Science 3952, mai 2006.

[Breiman 01] L. Breiman. *Random forests. Machine Learning, ML*, 45:5–31, 2001.

[Bres 99] S. Bres et J.-M. Jolion. *Detection of Interest Points for Image Indexation.* Dans Visual Information and Information Systems, pages 427–435. Springer Berlin Heidelberg, 1999.

[Brigger 99] P. Brigger, F. Müller, K. Illgner et M. Unser. *Centered Pyramids. IEEE Transactions on Image Processing, TIP*, 8(9):1254–1264, 1999.

[Brown 03] M. Z. Brown, D. Burschka et G. D. Hager. *Advances in Computational Stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 25(8):993–1008, 2003.

[Brun 96] L. Brun. *Segmentation d'images à base topologique.* Thèse de doctorat, Université de Bordeaux I, décembre 1996.

[Bui 18a] M. Bui, S. Albarqouni, S. Ilic et N. Navab. *Scene coordinate and correspondence learning for image-based localization. CoRR (arXiv)*, abs/1805.08443, 2018.

[Bui 18b] M. Bui, C. Baur, N. Navab, S. Ilic et S. Albarqouni. *Adversarial Joint Image and Pose Distribution Learning for Camera Pose Regression and Refinement. CoRR (arXiv)*, abs/1903.06646, 2018.

[Burt 83] P. J. Burt et E. H. Adelson. *The Laplacian Pyramid as a Compact Image Code. IEEE Transactions on Communications, TC*, 31(4):532–540, 1983.

[Bustos 09] B. Bustos et T. Schreck. Feature-Based 3D Object Retrieval, pages 1125–1128. Springer, Boston, 2009.

[Calonder 10] M. Calonder, V. Lepetit, C. Strecha et P. Fua. *BRIEF: Binary Robust Independent Elementary Features.* Dans European Conference on Computer Vision, ECCV, volume 6314, 2010.

[Cambra 14] A. B. Cambra, A. Muñoz, A. C. Murillo, J. J. Guerrero et D. Gutierrez. *Depth Interpolation via Smooth Surface Segmentation Using Tangent Planes Based on the Superpixels of a Color Image.* Dans Spanish Computer Graphics Conference, pages 29–36, 2014.

[Campbell 01] R. J. Campbell et P. J. Flynn. *A Survey Of Free-Form Object Representation and Recognition Techniques. Computer Vision and Image Understanding, CVIU*, 81(2):166–210, 2001.

[Can Karaimer 18] H. Can Karaimer et M. S. Brown. *Improving Color Reproduction Accuracy on Cameras.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2018.

[Can Karaimer 19] H. Can Karaimer et M. S. Brown. *Beyond raw-RGB and sRGB: Advocating Access to a Colorimetric Image State.* Dans Color and Imaging Conference, pages 86–90, 2019.

[Canny 86] J. Canny. *A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 8(6):679–698, 1986.

[Cao 17a] Y.-P. Cao, T. Ju, J. Xu et S.-M. Hu. *Extracting Sharp Features from RGB-D Images. Computer Graphics Forum*, 36(8):138–152, 2017.

[Cao 17b] Z. Cao, T. Simon, S. Wei et Y. Sheikh. *Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2017.

[Capel 04] D. Capel. Image mosaicing, Chapitre Image Mosaicing and Super-resolution, pages 47–79. Springer, London, 2004.

[Carlier 19] A. Carlier. Apprentissage et réseaux de neurones. Cours de l'Unité d'Enseignement IA et Multimédia, Semestre de Septembre à Janvier, Troisième année, parcours Multimédia 2019. Département Sciences du Numérique – École Nationale Supérieure en Électronique, Électrotechnique, Informatique, Hydraulique et Télécommunications, ENSEEIHT, Toulouse.

[Carron 95] T. Carron. *Segmentation d'images couleur dans la base Teinte Luminance Saturation : approche numérique et symbolique.* Thèse de doctorat, Université de Savoie, 1995.

[Castellani 08] U. Castellani, M. Cristani, S. Fantoni et V. Murino. *Sparse points matching by combining 3-D mesh saliency with statistical descriptors.* Dans Annual Conference of the European Association for Computer Graphics, Eurographics, 2008.

[Chambah 01] M. Chambah. *Analyse et traitement de données chromatiques d'images numérisées à haute résolution. Application à la restauration numérique des couleurs des films cinématographiques.* Thèse de doctorat, Université de La Rochelle, 2001.

[Chambon 05a] S. Chambon. *Mise en correspondance stéréoscopique d'images couleur en présence d'occultations.* Thèse de doctorat, Université Paul Sabatier – Toulouse III, 2005.

[Chambon 05b] S. Chambon et A. Crouzil. *Colour correlation-based matching. International Journal of Robotics and Automation, IJRA*, 20(2):78–85, 2005.

[Chambon 07a] A. Moreno-Ingelmo S. Chambon, , A. P. Santhanam, P. Kupelian, J. P. Rolland, E. D. Angelini et I. Bloch. *Thoracic CT-PET Registration Using a 3D Breathing Model.* Dans International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI, pages 626–633, 2007.

[Chambon 07b] S. Chambon, A. Moreno-Ingelmo, A. P. Santhanam, J. P. Rolland, E. D. Angelini et I. Bloch. *CT-PET Landmark-based Lung Registration Using a Dynamic Breathing Model*. Dans International Conference on Image Analysis and Processing, ICIAP, pages 691–696, 2007.

[Chambon 08] S. Chambon, A. Moreno, A. Santhanam, R. Brocardo, J. Rolland, E. Angelini et I. Bloch. *Introduction d'un modèle de respiration dans une méthode de recalage à partir de points d'intérêt d'images TEP et TDM du poumon*. Dans Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle, RFIA, 2008.

[Chambon 09] S. Chambon, P. Subirats et J. Dumoulin. *Introduction of a wavelet transform based on 2D matched filter in a Markov Random Field for fine structure extraction: Application on road crack detection*. Dans IS&T/SPIE Electronic Imaging, Image Processing: Machine Vision Applications II, 2009.

[Chambon 10a] S. Chambon. *Road Crack Extraction with Adapted Filtering and Markov Model-based Segmentation – Introduction and Validation*. Dans International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP, pages 349–354, 2010.

[Chambon 10b] S. Chambon, A. Crouzil, M. El Miziani, G. Lemarie et P. Le Callet. *Complémentarité de mesures de corrélation pour la mise en correspondance*. Dans Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle, RFIA, 2010.

[Chambon 10c] S. Chambon, C. Gourraud, J.-M. Moliard et P. Nicolle. *Road Crack Extraction with Adapted Filtering and Markov Model-based Segmentation – Introduction and Validation*. Dans International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP, pages 81–90, 2010.

[Chambon 11a] S. Chambon. *Detection of Points of Interest for Geodesic Contours: Application on Road Images for Crack Detection*. Dans International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP, pages 210–213, 2011.

[Chambon 11b] S. Chambon et A. Crouzil. *Combination of Correlation Measures for Dense Stereo Matching*. Dans International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP, pages 598–603, 2011.

[Chambon 11c] S. Chambon et A. Crouzil. *Occlusions handling in dense stereo matching. Pattern Recognition, PR*, 44(9):2063–2075, 2011.

[Chambon 11d] S. Chambon et J.-M. Moliard. *Automatic Road Pavement Assessment with Image Processing: Review and Comparison. International Journal of Geophysics*, pageID 989354, 2011.

[Chambon 11e] S. Chambon, A. Moreno, A. Santhanam, J. Rolland et I. Bloch. *MARIO : Modélisation de l'Anatomie normale et pathologique pour le Recalage non linéaire entre Images TDM et TEP en Oncologie. Traitement du Signal*, 28(3-4):275–307, 2011.

[Chambon 15a] S. Chambon. Des transformations à l'analyse de scènes. Cours de l'Unité d'Enseignement Image Modélisation et Rendu (IMR) – Partie Modélisation et Rendu, Semestre de Février à Mai, Deuxième année, parcours Multimédia 2015. Département Sciences du Numérique – École Nationale Supérieure en Électronique, Électrotechnique, Informatique, Hydraulique et Télécommunications, ENSEEIHT, Toulouse.

[Chambon 15b] S. Chambon. Introduction au suivi d'objets visuels – Détection et mise en correspondance de points d'intérêt. Cours de l'Unité d'Enseignement Vision, Réalité Augmentée et Applications – Partie Vision, Semestre de Septembre à Janvier, Troisième année, parcours Multimédia 2015. Département Sciences du Numérique – École Nationale Supérieure en Électronique, Électrotechnique, Informatique, Hydraulique et Télécommunications, ENSEEIHT, Toulouse.

[Chan 02] T. F. Chan et L. A. Vese. *A Multiphase Level Set Framework for Image Segmentation using the Mumford and Shah model. Internationl Journal on Computer Vision, IJCV*, 50(3):271–293, 2002.

[Chang 18] J.-R. Chang et Y.-S. Chen. *Pyramid stereo matching network.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 5410–5418, 2018.

[Chen 98] C.-S. Chen, Y.-P. Hung et J.-B. Cheng. *A fast automatic method for registration of partially-overlapping range images.* Dans International Conference on Computer Vision, ICCV, pages 242–248, janvier 1998.

[Chen 99a] C.-S. Chen, Y.-P. Hung et J.-B. Cheng. *RANSAC based DARCES – A New Approach to Fast Automatic Registration of Partially Overlapping Range Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 21:1229–1234, novembre 1999.

[Chen 99b] Q. Chen et G. Medioni. *A volumetric Stereo Matching Method: Application to Image-Based Modeling.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, volume 1, pages 1029–1034, juin 1999.

[Chen 02] H. Chen et P. Meer. *Robust Computer Vision through Kernel Density Estimation.* Dans European Conference on Computer Vision, ECCV, volume 1, pages 236–250, 2002.

[Chen 13] L. Chen, W. Lu, J. Ni, W. Sun et J. Huang. *Region duplication detection based on Harris corner points and step sector statistics. Journal of Visual Communication and Image Representation, VCIR*, 24(3):244–254, 2013.

[Cheng 99] H. Cheng, J. Chen, C. Glazier et Y. Hu. *Novel Approach to Pavement Cracking Detection Based on Fuzzy Set Theory. Journal of Computing in Civil Engineering*, 13(4):270–280, 1999.

[Cheng 01] H. C. Cheng, J. L. Wang, Y. G. Hu, C. Glazier, X. J. Shi et X. W. Chen. *Novel Approach to Pavement Cracking Detection Based on Neural Network.* Dans Journal of Materials Processing Technology, JMPT, pages 1764:119–127, 2001.

[Choy 15] C. B. Choy, M. Stark, S. Corbett-Davies et S. Savarese. *Enriching object detection with 2D-3D registration and continuous viewpoint estimation.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 2512–2520, 2015.

[Cie 15.2 86] Cie 15.2. Colorimetry, Second edition. Rapport Technique, Commission Internationale de l'Éclairage, 1986.

[Clarkson 01] M. J. Clarkson, D. Rueckert, D. L. G. Hill et D. J. Hawkes. *Using photo-consistency to register 2D optical images of the human face to a 3D surface model. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 23(11):1266–1280, 2001.

[Cohen-Steiner 03] D. Cohen-Steiner et J.-M. Morvan. *Restricted Delaunay Triangulations and Normal Cycle.* Dans Symposium on Computational Geometry, SCG, pages 312–321, 2003.

[Cohen 96] L. D. Cohen et R. Kimmel. *Global Minimum for Active Contour Models: A Minimal Path Approach. Internationl Journal on Computer Vision, IJCV*, 24(1):57–78, 1996.

[Collins 02] M. Collins, R. E. Schapire et Y. Singer. *Logistic Regression, AdaBoost and Bregman Distances. Machine Learning, ML*, 48:253–285, 2002.

[Comaniciu 99] D. Comaniciu et P. Meer. *Mean shift analysis and applications.* Dans International Conference on Computer Vision, ICCV, volume 2, pages 1197–1203, 1999.

[Comaniciu 02] D. Comaniciu et P. Meer. *Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 24(5):603–619, 2002.

[Connolly 85] C. Connolly. *The determination of next best views.* Dans IEEE international conference on Robotics and Automation, ICRA, pages 432–435, 1985.

[Cord 11] A. Cord et S. Chambon. *Automatic Road Defect Detection by Textural Pattern Recognition Based on AdaBoost. Computer-Aided Civil and Infrastructure Engineering*, 27(4):244–249, 2011.

[Cortes 95] C. Cortes et V. Vapnik. *Support-vector networks. Machine Learning, ML*, 20(3):273–297, 1995.

[Coudray 10] N. Coudray, A. Karathanou et S. Chambon. *Multi-resolution Approach for Fine Structure Extraction – Application and Validation on Road Images.* Dans International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP, pages 142–147, 2010.

[Coughlan 99] J. M. Coughlan et A. L. Yuille. *Manhattan World: compass direction from a single image by Bayesian inference.* Dans International Conference on Computer Vision, ICCV, volume 2, pages 941–947, 1999.

[Crawford 96] C. R. Crawford, K. F. King, C. J. Ritchie et J. D. Godwin. *Respiratory Compensation in Projection Imaging Using a Magnification and Displacement Model. IEEE Transactions on Medical Imaging, TMI*, 15(3):327–332, 1996.

[Crombez 18] N. Crombez, R. Seulin, O. Morel, D. Fofi et C. Demonceaux. *Multimodal 2D Image to 3D Model Registration via a Mutual Alignment of Sparse and Dense Visual Features.* Dans IEEE international conference on Robotics and Automation, ICRA, pages 6316–6322, 2018.

[Crouzil 96] A. Crouzil, L. Massip-Pailhes et S. Castan. *A New Correlation Criterion Based on Gradient Fields Similarity.* Dans International Conference on Pattern Recognition, ICPR, volume 1, pages 632–636, 1996.

[Crowley 84] J. L. Crowley et A. C. Parker. *A Representation for Shape Based on Peaks and Ridges in the Difference of Low-Pass Transform. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 6(2):156–170, 1984.

[Cui 15] Zhaopeng Cui et Ping Tan. *Global Structure-From-Motion by Similarity Averaging.* Dans International Conference on Computer Vision, ICCV, 2015.

[Cyr 00] C. M. Cyr, A. F. Kamal, T. B. Sebastian et B. B. Kimia. *2D-3D Registration Based on Shape Matching.* Dans IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, MMBIA, pages 198–203, 2000.

[Dalal 05a] N. Dalal et B. Triggs. *Histograms of oriented gradients for human detection.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 886–893, 2005.

[Dalal 05b] N. Dalal et B. Triggs. *Histograms of oriented gradients for human detection.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, volume 1, pages 886–893, 2005.

[Dambreville 08] S. Dambreville, R. Sandhu, A. Yezzi et A. Tannenbaum. *Robust 3D Pose Estimation and Efficient 2D Region-Based Segmentation from a 3D Shape Prior.* Dans European Conference on Computer Vision, ECCV, volume 2, pages 169–182, octobre 2008.

[Darom 12] T. Darom et Y. Keller. *Scale-Invariant Features for 3-D Mesh Models. IEEE Transactions on Image Processing, TIP*, 21(5), 2012.

[De Joinville 01] O. De Joinville, G. Maillet, H. Maître et M. Roux. *Évaluation a priori de la qualité d'un MNS.* Dans actes du congrès francophone de Vision par Ordinateur, ORASIS, pages 67–76, 2001.

[Delagnes 95] P. Delagnes et D. Barba. *A Markov random field for rectilinear structure extraction in pavement distress image analysis.* Dans International Conference on Image Processing, ICIP, volume 1, pages 446–449, 1995.

[Dellepiane 07] M. Dellepiane, M. Callieri, F. Ponchio et R. Scopigno. *Mapping highly detailed color information on extremely dense 3D models: the case of David's restoration.* Dans Annual Conference of the European Association for Computer Graphics, Eurographics – Cultural Heritage Papers, 2007.

[Demir 19] I. Demir, C. Hahn, K. Leonard, G. Morin, D. Rahbani, A. Panotopoulou, A. Fondevilla, E. Balashova, B. Durix et A. Kortylewski. *SkelNetOn 2019: Dataset and Challenge on Deep Learning for Geometric Shape Understanding.* Dans IEEE Conference on Computer Vision Pattern Recognition Workshop, CVPR Workshop, 2019.

[Dempster 77] A. P. Dempster, N. M. Laird et D.B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[Deng 05] Y. Deng, Q. Yang, X. Lin et X. Tang. *A Symmetric Patch-Based Correspondence Model for Occlusion Handling.* Dans International Conference on Computer Vision, ICCV, 2005.

[Deng 07] H. Deng, W. Zhang, E. Mortensen, T. Dietterich et L. Shapiro. *Principal Curvature-Based Region Detector for Object Recognition.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2007.

[Deriche 87] R. Deriche. *Using Canny's criteria to derive a recursively implemented optimal edge detector. Internationl Journal on Computer Vision, IJCV*, 2(6):167–187, 1987.

[Deschamps 01] T. Deschamps et L. Cohen. *Fast extraction of minimal paths in 3D images and applications to virtual endoscopy.* Dans Medical Image Analysis, pages 281–299, 2001.

[Dijkstra 59] E. W. Dijkstra. *A Note on Two Problems in Connexion with Graphs. Numerische Mathematik*, 1(1):269–271, 1959.

[Ding 16] W. Ding et G. W. Taylor. *Automatic Moth Detection from Trap Images for Pest Management. Computers and Electronics in Agriculture, CEA*, 123(C):17–28, 2016.

[do Carmo 76] M. P. do Carmo. Differential geometry of curves and surfaces. Prentice Hall, 1976.

[Domingos 97] P. Domingos et M. Pazzani. *On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, ML*, 29(2-3):103–130, 1997.

[Drouyer 19] S. Drouyer. *An « All Terrain » Crack Detector Obtained by Deep Learning on Available Databases . Image Processing On Line, IPOL*, 2019.

[Duan 15] L. Duan et F. Lafarge. *Image partitioning into convex polygons.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 3119–3127, 2015.

[Duda 01] R. O. Duda, P. E. Hart et D. G. Stork. Pattern classification. Wiley-interscience, 2001.

[Duong 18] N. Duong, A. Kacete, C. Sodalie, P. Richard et J. Royan. *xyzNet: Towards Machine Learning Camera Relocalization by Using a Scene Coordinate Prediction Network.* Dans IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), pages 258–263, 2018.

[Durix 15a] B. Durix, G. Morin, S. Chambon, C. Roudet et L. Garnier. *Towards Skeleton Based Reconstruction: From Projective Skeletonization to Canal Surface Estimation.* Dans International Conference on 3D Vision, 3DV, pages 545–553, 2015.

[Durix 15b] B. Durix, G. Morin, S. Chambon, C. Roudet, L. Garnier et O. Cots. *Caractérisation de la projection du squelette d'une surface canal 3D : Application à la reconstruction 3D à partir de deux images.* Dans Journées du Groupe de Travail en Modélisation Géométrique, GTMG, 2015.

[Durix 16a] B. Durix, G. Morin, S. Chambon, C. Roudet et L. Garnier. *Skeleton-based multiview reconstruction.* Dans International Conference on Image Processing, ICIP, pages 4047–4051, 2016.

[Durix 16b] B. Durix, G. Morin, S. Chambon, C. Roudet, L. Garnier et O. Cots. *Reconstruction basée squelette d'un objet 3D à partir de points de vue multiples.* Dans Journées du Groupe de Travail en Modélisation Géométrique, GTMG, 2016.

[Durix 17] B. Durix. *Squelettes pour la reconstruction 3D : de l'estimation de la projection du squelette dans une image 2D à la triangulation du squelette en 3D.* Thèse de doctorat, Institut National Polytechnique de Toulouse, INPT, 2017.

[Durix 19a] B. Durix, S. Chambon, K. Leonard, J.-L. Mari et G. Morin. *The Propagated Skeleton: A Robust Detail-Preserving Approach.* Dans International Conference on Discrete Geometry for Computer Imagery, DGCI, pages 343–354, 2019.

[Durix 19b] B. Durix, G. Morin, S. Chambon, K. Leonard et J.-L. Mari. *One-step Compact Skeletonization.* Dans Annual Conference of the European Association for Computer Graphics, Eurographics, pages 21–24, 2019.

[Ehrhardt 07] J. Ehrhardt, A. Schmidt-Richberg et H. Handels. *A Variational Approach for Combined Segmentation and Estimation of Respiratory Motion in Temporal Image Sequences.* Dans International Conference on Computer Vision, ICCV, 2007.

[Eisenbach 17] M. Eisenbach, R. Stricker, D. Seichter, K. Amende, K. Debes, M. Sesselmann, D. Ebersbach, U. Stoeckert et H.-M. Gross. *How to get pavement distress detection ready for deep learning? A systematic approach.* Dans International Joint Conference on Neural Networks, JCNN, pages 2039–2047, 2017.

[El Merabet 13] Y. El Merabet. *Segmentation d'images couleur par combinaison LPE-régions/LPE-contours et fusion de régions – Application à la segmentation de toitures à partir d'orthophotoplans.* PhD thesis, Université de Technologie de Belfort-Montbeliard, 2013.

[Elder 98] J. H. Elder et S. W. Zucker. *Local scale control for edge detection and blur estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 20(7):699–716, 1998.

[Everingham 15] M. Everingham, S. M. . Eslami, L. Van Gool, C. K. I. Williams, J. Winn et A. Zisserman. *The Pascal Visual Object Classes Challenge: A Retrospective. Internationl Journal on Computer Vision, IJCV*, 111(1):98–136, 2015.

[Fan 18] Z. Fan, Y. Wu, J. Lu et W. Li. *Automatic Pavement Crack Detection Based on Structured Prediction with the Convolutional Neural Network. CoRR (arXiv)*, abs/1802.02208, 2018.

[Faugeras 79] O. Faugeras. *Digital Color Image Processing Within the Framework of a Human Visual Model. IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP*, 27(4):380–393, 1979.

[Faye 13] Gaël Faye. *Qwerty.* Dans Pili pili sur un croissant au beurre, 2013.

[Fayer 18] J. Fayer, B. Durix, S. Gasparini et G. Morin. *Texturing and inpainting a complete tubular 3D object reconstructed from partial views. Computer & Graphics*, 74:126–136, 2018.

[Fei 19] Y. Fei, K. C. P. Wang, A. Zhang, C. Chen, J. Q. Li, Y. Liu, G. Yang et B. Li. *Pixel-level cracking detection on 3D asphalt pavement images through deep-learning-based CrackNet-V. IEEE Transactions on Intelligent Transportation Systems, TITS*, 2019.

[Feixas 09] M. Feixas, M. Sbert et F. González. *A unified information-theoretic framework for viewpoint selection and mesh saliency. ACM Transactions on Applied Perception, TAP*, 6(1):1–23, 2009.

[Felzenszwalb 04] P. F. Felzenszwalb et D. P. Huttenlocher. *Efficient Graph-Based Image Segmentation. Internationl Journal on Computer Vision, IJCV*, 59(2):167–181, 2004.

[Fischer 14] P. Fischer et T. Brox. *Image Descriptors Based on Curvature Histograms.* Dans German Conference on Pattern Recognition, GCPR, pages 239–249, 2014.

[Fischler 81] M. A. Fischler et R. C. Bolles. *Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the Association for Computing Machinery, ACM*, 24(6):381–395, 1981.

[Flitton 10] G. Flitton, T. Breckon et N. Megherbi Bouallagu. *Object Recognition using 3D SIFT in Complex CT Volumes.* Dans British Machine Vision Conference, BMVC, 2010.

[Forsythe 16] J. Forsythe, V. Kurlin et A. Fitzgibbon. *Resolution-Independent Superpixels Based on Convex Constrained Meshes Without Small Angles.* Dans International Conference on Advances in Visual Computing, ISVC, pages 223–233, 2016.

[Fulkerson 09] B. Fulkerson, A. Vedaldi et S. Soatto. *Class segmentation and object localization with superpixel neighborhoods.* Dans International Conference on Computer Vision, ICCV, pages 670–677, octobre– novembre 2009.

[Furukawa 09] Y. Furukawa, B. Curless, S. M. Seitz et R. Szeliski. *Manhattan-world stereo.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 1422–1429, 2009.

[Furukawa 10] Y. Furukawa et J. Ponce. *Accurate, Dense, and Robust Multi-View Stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 32(8):1362–1376, 2010.

[Furukawa 15] Y. Furukawa et C. Hernández. *Multi-View Stereo: A Tutorial. Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

[Gal 06] R. Gal et D. Cohen-Or. *Salient geometric features for partial shape matching and similarity. ACM transactions on graphics,TOG*, 25(1):130–150, 2006.

[Gales 10a] G. Gales, A. Crouzil et S. Chambon. *A Region-Based Randomized Voting Scheme for Stereo Matching.* Dans International Conference on Advances in Visual Computing, ISVC, pages 182–191, 2010.

[Gales 10b] G. Gales, A. Crouzil et S. Chambon. *Complementarity of Feature Point Detectors.* Dans International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP, pages 334–339, 2010.

[Gales 10c] G. Gales, A. Crouzil et S. Chambon. *Détection de points d'intérêt pour la mise en correspondance par propagation.* Dans Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle, RFIA, 2010.

[Gales 11] G. Gales. *Mise en correspondance de pixels pour la stéréovision binoculaire par propagation d'appariements de points d'intérêt et sondage de régions.* Thèse de doctorat, Université Paul Sabatier – Toulouse III, 2011.

[Gales 12] G. Gales, S. Chambon, A. Crouzil et J. McDonald. *Reliability measure for propagation-based stereo matching.* Dans International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS, 2012.

[Gallier 01] J.-H. Gallier. Geometric Methods and Applications: For Computer Science and Engineering. Springer, 2001.

[Garbay 79] C. Garbay. *Modélisation de la couleur dans le cadre de l'analyse d'images et de son application à la cytologie automatique.* Thèse de doctorat, Institut National Polytechnique, INP, 1979.

[Garbay 81] C. Garbay, F. Brugal et C. Choquet. *Application of Colored Image Analysis to Bone Marrow Cell Recognition. Analytical and Quantitative Cytology and Histology, AQCH*, 3(4):272–280, 1981.

[Gauglitz 11] S. Gauglitz, T. Höllerer et M. Turk. *Evaluation of interest point detectors and feature descriptors for visual tracking. Internationl Journal on Computer Vision, IJCV*, 94(3):335–360, 2011.

[Gavilán 11] M. Gavilán, D. Balcones, O. Marcos, D. F. Llorca, M. A. Sotelo, I. Parra, M. Ocaña, P. Aliseda, P. Yarza et A. Amírola. *Adaptive Road Crack Detection System by Pavement Classification. Sensors*, pages 9628–9657, 2011.

[Gevers 98] T. Gevers, A. W. M. Smeulders et H. Stokman. *Photometric Invariant Region Detection.* Dans British Machine Vision Conference, BMVC, pages 578–589, 1998.

[Ghadai 19] S. Ghadai, X. Y. Lee, A. Balu, S. Sarkar et A. Krishnamurthy. *Multi-Level 3D CNN for Learning Multi-Scale Spatial Features.* Dans IEEE Conference on Computer Vision Pattern Recognition Workshop, CVPR Workshop, 2019.

[Giro-i-Nieto 10] X. Giro-i-Nieto, N. Camps et F. Marqués. *GAT: a Graphical Annotation Tool for semantic regions. Multimedia tools and applications, MTA*, 46(2-3):155–174, 2010.

[Girshick 15] R. Girshick. *Fast R-CNN. CoRR (arXiv)*, abs/1504.08083, 2015.

[Godil 11] A. Godil et A. I. Wagan. *Salient local 3D features for 3D shape retrieval.* Dans IS&T/SPIE Electronic Imaging, 2011.

[Golodetz 18] S. Golodetz, T. Cavallari, N. A. Lord, V. A. Prisacariu, D. W. Murray et P. H. S. Torr. *Collaborative Large-Scale Dense 3D Reconstruction with Online Inter-Agent Pose Optimisation. IEEE Transactions on Visualization and Computer Graphics, TVCG*, 24:2895–2905, 2018.

[Gonzalez 04] R. C. Gonzalez, R. E. Woods et S. L. Eddins. Digital image processing using matlab. Pierson Prentice Hall, 2004.

[Goodfellow 14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, X. Bing, D. Warde-Farley, S. Ozair, A. Courville et Y. Bengio. *Generative adversarial nets.* Dans Advances in neural information processing systems, pages 2672–2680, 2014.

[Goodfellow 16] I. Goodfellow, Y. Bengio et A. Courville. Deep learning. MIT Press, 2016. http://www.deeplearningbook.org.

[Gould 08] S. Gould, J. Rodgers, D. Cohen, G. Elidan et D. Koller. *Multi-Class Segmentation with Relative Location Prior. Internationl Journal on Computer Vision, IJCV*, pages 300–316, décembre 2008.

[Gould 09] S. Gould, R. Fulton et D. Koller. *Decomposing a scene into geometric and semantically consistent regions.* Dans International Conference on Computer Vision, ICCV, octobre 2009.

[Grabner 19a] A. Grabner, P. M. Roth et V. Lepetit. *GP2C: Geometric Projection Parameter Consensus for Joint 3D Pose and Focal Length Estimation in the Wild.* Dans International Conference on Computer Vision, ICCV, 2019.

[Grabner 19b] A. Grabner, P. M. Roth et V. Lepetit. *Location Field Descriptors: Single Image 3D Model Retrieval in the Wild.* Dans International Conference on 3D Vision, 3DV, 2019.

[Gray 06] A. Gray, E. Abbena et S. Salamon. Modern Differential Geometry of Curves and Surfaces with Mathematica. Chapman & Hall/CRC, 2006.

[Grundmann 10] M. Grundmann, V. Kwatra, M. Han et I. Essa. *Efficient hierarchical graph-based video segmentation.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 2141–2148, 2010.

[Grégoire 15] M. Grégoire. Le développement de la fonction visuelle chez l'enfant : étude bibliographique. Mémoire en vue de l'obtention du certificat de capacité d'orthoptie, Ecole d'orthoptie de Clermont Ferrand, Université d'Auvergne, 2015.

[Guerrero 05] T. Guerrero, G. Zhang, W. Segars, T. C. Huang, S. Bilton, G. Ibbott, L. Dong, K. Forster et K. P. Lin. *Elastic image mapping for 4-D dose estimation in thoracic radiotherapy. Radiation Protection Dosimetry*, 115(1–4):497–502, 2005.

[Gupta 18] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese et A. Alahi. *Social GAN: Socially acceptable trajectories with generative adversarial networks.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 2255–2264, 2018.

[Guyot 19] P. Guyot, T. Malon, G. Roman-Jimenez, S. Chambon, V. Charvillat, A. Crouzil, A. Péninou, J. Pinquier, F. Sèdes et C. Sénac. *Audiovisual Annotation Procedure for Multi-view Field Recordings.* Dans International Conference on MultiMedia Modeling, MMM, pages 399–410, 2019.

[Guénard 13] J. Guénard. *Synthèse de modèles de plantes et reconstructions de baies à partir d'images*. Thèse de doctorat, Institut National Polytechnique de Toulouse, INPT, 2013.

[Hafiz 15] D. A. Hafiz, B. A. B. Youssef, W. M. Sheta et H. A. Hassan. *Interest Point Detection in 3D Point Cloud Data Using 3D Sobel-Harris Operator. International Journal of Pattern Recognition and Artificial Intelligence, PRAI*, 29(7), 2015.

[Han 15] X. Han, T. Leung, Y. Jia, R. Sukthankar et A. C. Berg. *MatchNet: Unifying feature and metric learning for patch-based matching*. Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 3279–3286, 2015.

[Hanbury 03] A. Hanbury. *A 3D-Polar Coordinate Colour Representation Well Adapted to Image Analysis*. Dans Scandinavian Conference on Image Analysis, SCIA, pages 804–811, 2003.

[Hanbury 08] A. Hanbury. *How Do Superpixels Affect Image Segmentation?* Dans Iberoamerican Congress on Pattern Recognition, CIARP, volume 5197, pages 178–186, 2008.

[Haner 12] S. Haner et A. Heyden. *Covariance Propagation and Next Best View Planning for 3D Reconstruction*. Dans European Conference on Computer Vision, ECCV, 2012.

[Hardeberg 99] J. Y. Hardeberg. *Acquisition and reproduction of colour images: colorimetric and multispectral approaches*. Thèse de doctorat, École Nationale Supérieure des Télécommunications, ENST, 1999.

[Harris 88] C. Harris et M. Stephens. *A combined corner and edge detector*. Dans Alvey Vision Conference, pages 147–151, 1988.

[Hartkens 02] T. Hartkens, D. L. G. Hill, A. D. Castellano-Smith, D. J. Hawkes, C. R. Maurer Jr., A. J. Martin, W. A. Hall, H. Liu et C. L. Truwit. *Using Points and Surfaces to Improve Voxel-Based Non-Rigid Registration*. Dans International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI, volume 2489 de *Lecture Notes in Computer Science*, pages 565–572, 2002.

[Hartley 04] R. Hartley et A. Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, 2004.

[Harzallah 11] H. Harzallah. *Contribution à la détection et à la reconnaissance d'objets dans les images*. Thèse de doctorat, Université Grenoble Alpes, 2011.

[He 16] K. He, X. Zhang, S. Ren et J. Sun. *Deep residual learning for image recognition*. Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 770–778, 2016.

[He 20] K. He, G. Gkioxari, P. Dollár et R. Girshick. *Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 42(2):386–397, 2020.

[Heinly 15] J. Heinly, J. L. Schönberger, E. Dunn et J.-M. Frahm. *Reconstructing the World* in Six Days *(As Captured by the Yahoo 100 Million Image Dataset)*. Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2015.

[Heitz 08] G. Heitz et D. Koller. *Learning Spatial Context: Using Stuff to Find Things*. Dans European Conference on Computer Vision, ECCV, volume 1, pages 30–43, octobre 2008.

[Hinterstoisser 11] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua et V. Lepetit. *Gradient response maps for real-time detection of textureless objects. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 34(5):876–888, 2011.

[Hirschmuller 07] H. Hirschmuller et D. Scharstein. *Evaluation of cost functions for stereo matching*. Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2007.

[Hirschmüller 08] H. Hirschmüller. *Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 30(2):328–341, 2008.

[Hirschmüller 09] H. Hirschmüller et D. Scharstein. *Evaluation of stereo matching costs on images with radiometric differences. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 31(9):1582–1599, 2009.

[Hoiem 05a] D. Hoiem, A. A. Efros et M. Herbert. *Automatic photo pop-up. ACM transactions on graphics,TOG*, 24(3):577–584, 2005.

[Hoiem 05b] D. Hoiem, A. A. Efros et M. Herbert. *Geometric context from a single image.* Dans International Conference on Computer Vision, ICCV, volume 1, pages 654–661, 2005.

[Hoiem 08] D. Hoiem, A. A. Efros et M. Herbert. *Closing the loop on scene interpretation.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2008.

[Hong 04] L. Hong et G. Chen. *Segment-based Stereo Matching Using Graph Cuts.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, volume 1, pages 74–81, 2004.

[Hoppe 12] C. Hoppe, M. Klopschitz, M. Rumpler, A. Wendel, S. Kluckner, H. Bischof et G. Reitmayr. *Online Feedback for Structure-from-Motion Image Acquisition. British Machine Vision Conference, BMVC*, 2012.

[Horowitz 76] S. L Horowitz et T. Pavlidis. *Picture Segmentation by a Tree Traversal Algorithm. Journal of the Association for Computing Machinery, ACM*, 23(2):368–388, 1976.

[Hosni 13] A. Hosni, M. Bleyer et M. Gelautz. *Secrets of adaptive support weight techniques for local stereo matching. Computer Vision and Image Understanding, CVIU*, 117(6):620–632, 2013.

[Hu 10] Y. Hu et C. Zhao. *A Local Binary Pattern Based Methods For Pavement Crack Detection. Journal of Pattern Recognition Research, JPRR*, 5(1):140–147, 2010.

[Hu 15] Y. Hu, S. Gao, K. Janowicz, B. Yu, W. Li et S. Prasad. *Extracting and understanding urban areas of interest using geotagged photos. Computers, Environment and Urban Systems*, 54:240–254, 2015.

[Huang 06] Y. Huang et B. Xu. *Automatic inspection of pavement cracking distress. J. Electronic Imaging*, 2006.

[Huang 15] Q. Huang, H. Wang et V. Koltun. *Single-view reconstruction via joint analysis of image and shape collections. ACM transactions on graphics,TOG*, 34(4):1–10, 2015.

[Ibenthal 00] A. Ibenthal, S. Siggelkow et R.-R. Grigat. Method of image segmentation. Brevet américain 6044114, mars 2000. U.S. Philips Corporation.

[Ion 11] A. Ion, N. M. Artner, G. Peyré, W. G. Kropatsch et L. D. Cohen. *Matching 2D and 3D articulated shapes using the eccentricity transform. Computer Vision and Image Understanding, CVIU*, 115(6):817–83, 2011.

[Irschara 09] A. Irschara, C. Zach, J. M. Frahm et H. Bischof. *From structure-from-motion point clouds to fast location recognition.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 2599–2606, 2009.

[Iscen 17] A. Iscen, G. Tolias, Y. Avrithis, T. Furon et 0. Chum. *Efficient diffusion on region manifolds: Recovering small objects with compact CNN representations.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2017.

[Jawahar 02] C. V. Jawahar et P. J. Narayanan. *Generalised correlation for multi-feature correspondence. Pattern Recognition, PR*, 35(6):1303–1313, 2002.

[Jeon 14] J. Jeon, S. Cho, X. Tong et S. Lee. *Intrinsic Image Decomposition Using Structure-Texture Separation and Surface Normals.* Dans European Conference on Computer Vision, ECCV, pages 218–233, 2014.

[Jiang 13] P. Jiang, H. Ling, J. Yu et J. Peng. *Salient Region Detection by UFO: Uniqueness, Focusness and Objectness.* Dans International Conference on Computer Vision, ICCV, pages 1976–1983, 2013.

[Jiang 17] C. Jiang, P. D. Paudel, Y. Fougerolle, D. Fofi et C. Demonceaux. *Static and dynamic objects analysis as a 3d vector field.* Dans International Conference on 3D Vision, 3DV, pages 234–243, 2017.

[Johnson 99] A. E. Johnson. *Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 21(5):433–449, 1999.

[Jolion 03] J. M. Jolion. *Stochastic Pyramid Revisited. Pattern Recognition Letters, PRL*, 24(8):1335–1342, 2003.

[Judd 07] T. Judd, F. Durand et E. H. Adelson. *Apparent ridges for line drawing. ACM transactions on graphics,TOG*, 26(3), 2007.

[Kaddah 19] W. Kaddah, M. Elbouz, Y. Ouerhani, V. Baltazart, M. Desthieux et A. Alfalou. *Optimized minimal path selection (OMPS) method for automatic and unsupervised crack segmentation within two-dimensional pavement images. The Visual Computer*, 35(9):1293–1309, 2019.

[Kadir 04] T. Kadir, A. Zisserman et M. Brady. *An Affine Invariant Salient Region Detector.* Dans European Conference on Computer Vision, ECCV, pages 228–241, 2004.

[Kamgar-Parsi 11] B. Kamgar-Parsi et B. Kamgar-Parsi. *Matching 2D image lines to 3D models: Two improvements and a new algorithm.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 2425–2432, 2011.

[Kang 14] S.-H. Kang, J.-H. Cho et S.-H. Lee. *Identification of butterfly based on their shapes when viewed from different angles using an artificial neural network. Journal of Asia-Pacific Entomology*, 17(2):143–149, 2014.

[Kannala 07] J. Kannala et S. S. Brandt. *Quasi-Dense Wide Baseline Matching Using Match Propagation.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, juin 2007.

[Karacan 13] L. Karacan, E. Erdem et A. Erdem. *Structure-Preserving Image Smoothing via Region Covariances.* Dans ACM Special Interest Group on Computer Graphics and Interactive Techniques Conference in Asia, SIGGRAPH Asia, 2013.

[Karaszewski 16] M. Karaszewski, M. Adamczyk et R. Sitnik. *Assessment of next-best-view algorithms performance with various 3D scanners and manipulator. International Society for Photogrammetry and Remote Sensing, ISPRS*, 119:320–333, 2016.

[Karlsson 14] R. Karlsson. *Implementation of Advanced Monitoring Techniques in Road Asset Management Results from the TRIMM project.* Dans Transport Research Arena, 2014.

[Kass 88] M. Kass, A. Witkin et D. Terzopoulos. *Snakes : Active contour models. Internationl Journal on Computer Vision, IJCV*, 1(4):321–331, 1988.

[Katz 05] S. Katz, G. Leifman et A. Tal. *Mesh segmentation using feature point and core extraction. International Journal of Computer Graphics, The Visual Computer*, 21(8):649–658, 2005.

[Kaul 12] V. Kaul, A. Yezzi et Y. Tsai. *Detecting Curves with Unknown Endpoints and Arbitrary Topology Using Minimal Paths.* Dans IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI, volume 34, pages 1952–1965, 2012.

[Kaya 14] Y. Kaya et L. Kayci. *Application of artificial neural network for automatic detection of butterfly species using color and texture features. International Journal of Computer Graphics, The Visual Computer*, 30(1):71–79, 2014.

[Kendall 16] A. Kendall et R. Cipolla. *Modelling uncertainty in deep learning for camera relocalization.* Dans IEEE international conference on Robotics and Automation, ICRA, pages 4762–4769, 2016.

[Kendall 17] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach et A. Bry. *End-to-end learning of geometry and context for deep stereo regression.* Dans International Conference on Computer Vision, ICCV, pages 66–75, 2017.

[Kermad 97] C. Kermad. *Image Segmentation: toward an automatic and unsupervised system through methods cooperation.* PhD thesis, Université Rennes 1, 1997.

[Khalfaoui 13] S. Khalfaoui, R. Seulin, Y. Fougerolle et D. Fofi. *An efficient method for fully automatic 3D digitization of unknown objects. Computers in Industry*, 64:1152–1160, 2013.

[Khan 03] S. Khan et M. Shah. *Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 25(10):1355–1360, 2003.

[Kholi 08] P. Kholi, J. Rihan, M. Bray et P. Torr. *Simultaneous segmentation and 3D pose estimation of humans using dynamic graph cuts. Internationl Journal on Computer Vision, IJCV*, 79(3):285–298, 2008.

[Kim 06] Y. Kim et A. Varshney. *Saliency-guided enhancement for volume visualization. IEEE Transactions on Visualization and Computer Graphics, TVCG*, 12(5):925–932, 2006.

[Kim 08] Y. Kim et A. Varshney. *Persuading visual attention through geometry. IEEE Transactions on Visualization and Computer Graphics, TVCG*, 14(4):772–782, 2008.

[Kim 14] H. Kim et A. Hilton. *Hybrid 3D feature description and matching for multi-modal data registration.* Dans International Conference on Image Processing, ICIP, pages 3493–3497, 2014.

[Kitchen 82] L. Kitchen et A. Rosenfeld. *Gray level corner detection. Pattern Recognition Letters, PRL*, 1(2):95–102, 1982.

[Klaus 06] A. Klaus, M. Sormann et K. Karner. *Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure.* Dans International Conference on Pattern Recognition, ICPR, volume 3, pages 15–18, 2006.

[Klette 04] R. Klette et A. Rosenfeld. Digital Geometry: Geometric Methods for Digital Picture Analysis. Morgan Kaufmann, 2004.

[Koenderink 92] J. J. Koenderink et A. J. Van Doorn. *Surface shape and curvature scales. Image and Vision Computing, IVC*, 10(8):557–564, 1992.

[Kohavi 95] R. Kohavi. *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Dans International Joint Conference on Artificial intelligence, IJCAI, volume 2, pages 1137–1145, 1995.

[Koschan 93] A. Koschan. *Dense Stereo Correspondence Using Polychromatic Block Matching.* Dans International Conference on Computer Analysis of Images and Patterns, CAIP, volume 719 de *Lecture Notes in Computer Science*, pages 538–542, 1993.

[Koutsopoulos 93] H. N. Koutsopoulos et A. B. Downey. *Primitive-Based Classification of Pavement Cracking Images. Journal of Transportation Engineering, JTE*, 119(3):402–418, 1993.

[Krizhevsky 12] A. Krizhevsky, I. Sutskever et G. E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks.* Dans International Conference on Neural Information Processing Systems, NIPS, pages 1097–1105, 2012.

[Kuznetsova 18] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig et V. Ferrari. *The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. CoRR (arXiv)*, abs/1811.00982, 2018.

[Laguna 19] A. B. Laguna, E. Riba, D. Ponsa et K. Mikolajczyk. *Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters.* Dans International Conference on Computer Vision, ICCV, pages 5835–5843, 2019.

[Langguth 13] F. Langguth et M. Goesele. *Guided Capturing of Multi-view Stereo Datasets.* Dans Annual Conference of the European Association for Computer Graphics, Eurographics, pages 93–96, 2013.

[Laptev 04] I. Laptev et T. Lindeberg. *Local descriptors for spatio-temporal recognition.* Dans International Workshop on Spatial Coherence for Visual Motion Analysis, pages 91–103, 2004.

[Lecun 98] Y. Lecun, L. Bottou, Y. Bengio et P. Haffner. *Gradient-based learning applied to document recognition. Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LeCun 15] Y. LeCun, Y. Bengio et G. E. Hinton. *Deep learning. Nature*, 521(7553):436, 2015.

[Lee 05] C. H. Lee, A. Varshney et D. W. Jacobs. *Mesh Saliency. ACM transactions on graphics,TOG*, 24(3):659–666, 2005.

[Lee 07] Y. Lee, L. Markosian, S. Lee et J. F. Hughes. *Line Drawings via Abstracted Shading. ACM transactions on graphics,TOG*, 26(3), 2007.

[Lee 13a] Y. Y. Lee, K. Park, J. D. Yoo et K. H. Lee. *Multi-scale Feature Matching Between 2D Image and 3D Model.* Dans ACM Special Interest Group on Computer Graphics and Interactive Techniques Conference in Asia, SIGGRAPH Asia, 2013.

[Lee 13b] Y. Y. Lee, M. K. Park, J. D. Yoo et K. H. Lee. *Multi-Scale Feature Matching between 2D Image and 3D Model.* Dans ACM Special Interest Group on Computer Graphics and Interactive Techniques Conference in Asia, SIGGRAPH Asia, 2013.

[Lee 13c] Y. Y. Lee, M. K. Park, J. D. Yoo et K. H. Lee. *Multi-scale Feature Matching Between 2D Image and 3D Model.* Dans ACM Special Interest Group on Computer Graphics and Interactive Techniques Conference in Asia, SIGGRAPH Asia, 2013.

[Lei 17] H. Lei, G. Jiang et L. Quan. *Fast Descriptors and Correspondence Propagation for Robust Global Point Cloud Registration. IEEE Transactions on Image Processing, TIP*, 26(8):3614–3623, 2017.

[Lemaitre 08] C. Lemaitre. *Définition et étude des performances d'un détecteur de structures curvilinéaires. Application à la stéréoscopie et la détection d'objets filaires.* PhD thesis, Université de Bourgogne, France, 2008.

[les Crayons 12] Amélie les Crayons. *On n'est pas fatigués.* Dans Jusqu'à la mer, 2012.

[Leutenegger 11] S. Leutenegger, M. Chli et R. Y. Siegwart. *BRISK: Binary robust invariant scalable keypoints.* Dans International Conference on Computer Vision, ICCV, pages 2548–2555, 2011.

[Levine 85] M. D. Levine. Vision in man and machine. McGraw-Hill Book Company, 1985.

[Levinshtein 09] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson et K. Siddiqi. *TurboPixels: Fast Superpixels Using Geometric Flows. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 31(12):2290–2297, 2009.

[Levy 00] A. Levy et M. Lindenbaum. *Sequential Karhunen-Loeve Basis Extraction and its Application to Images. IEEE Transactions on Image Processing, TIP*, 9(8):1371–1374, 2000.

[Lhuillier 02] M. Lhuillier et L. Quan. *Match Propagation for Image-Based Modeling and Rendering. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 24(8):1140–1146, 2002.

[Li 11] Q. Li, Q. Zou et X. Liu. *Pavement Crack Classification via Spatial Distribution Features.* Dans EURASIP Journal on Advances in Signal Processing, 2011.

[Li 12] Z. Li et X. Zou. *Skeleton detection of crack based on multiscale geometric analysis in space domain.* Dans International Conference on Automatic Control and Artificial Intelligence, pages 1862–1865, 2012.

[Li 15] Z. Li et J. Chen. *Superpixel segmentation using linear spectral clustering.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 1356–1363, 2015.

[Lim 14] J. J. Lim, A. Khosla et A. Torralba. *FPM: Fine Pose Parts-Based Model with 3D CAD Models.* Dans European Conference on Computer Vision, ECCV, pages 478–493, 2014.

[Lin 03] M. H. Lin et C. Tomasi. *Surfaces with Occlusions from layered Stereo. IEEE Conference on Computer Vision Pattern Recognition, CVPR*, 2003.

[Lin 14] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick et P. Dollár. *Microsoft COCO: Common Objects in Context. CoRR (arXiv)*, abs/1405.0312, 2014.

[Liu 03] T. Liu, H.-J. Zhang et F. Qi. *A novel video key-frame-extraction algorithm based on perceived motion energy model. IEEE Transactions on Circuits and Systems for Video Technology, CSVT*, 13(10):1006–1013, 2003.

[Liu 05] L. Liu et I. Stamos. *Automatic 3D to 2D registration for the photorealistic rendering of urban scenes.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 137–143, 2005.

[Liu 17] X. Liu, Y. Luo, Y. Ye et J. Lu. *MC-DCNN: Dilated convolutional neural network for computing stereo matching cost.* Dans International Conference on Neural Information Processing, ICONIP, pages 249–259, 2017.

[Liu 19a] W. Liu, Y. Huang, Y. Li et Q. Chen. *FPCNet: Fast Pavement Crack Detection Network Based on Encoder-Decoder Architecture. CoRR (arXiv)*, abs/1907.02248, 2019.

[Liu 19b] X. Liu et Y. Huang. *MC-HDCNN: Computing the Stereo Matching Cost with a Hybrid Dilated Convolutional Neural Network.* Dans International Conference on Neural Information Processing, ICONIP, pages 140–147, 2019.

[Lowe 04] D. G. Lowe. *Distinctive image features from scale-invariant keypoints. Internationl Journal on Computer Vision, IJCV*, 60(2):91–110, 2004.

[Lu 12] Y. Lu, W. Zhang, C. Jin et X. Xue. *Learning attention map from images.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 1067–1074, juin 2012.

[Luo 16] W. Luo, A. G. Schwing et Raquel Urtasun. *Efficient deep learning for stereo matching.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 5695–5703, 2016.

[Luo 18] Z. Luo, T. Shen, L. Zhou, S. Zhu, R. Zhang, Y. Yao, T. Fang et L. Quan. *GEO-DESC: Learning local descriptors by integrating geometry constraints.* Dans European Conference on Computer Vision, ECCV, pages 168–183, 2018.

[Léon 17] V. Léon, V. Itier, N. Bonneel, G. Lavoué et J.-P. Vandeborre. *Semantic correspondence across 3D models for example-based modeling.* Dans Eurographics Workshop on 3D Object Retrieval, 3DOR, 2017.

[Ma 05] Y. Z. Ma et W. J. Liu. *Progressive Matching Based on Segmentation for 3D reconstruction.* Dans International Conference on Computer and Information Technology, ICCIT, pages 575–579, septembre 2005.

[Machairas 15] V. Machairas, M. Faessel, D. Cárdenas-Peña, T. Chabardes, T. Walter et E. Decencière. *Waterpixels. IEEE Transactions on Image Processing, TIP*, 24(11):3707–3716, 2015.

[MacQueen 67] J. MacQueen. *Some methods for classification and analysis of multivariate observations.* Dans Berkeley symposium on mathematical statistics and probability, volume 1, pages 281–297, 1967.

[Maes 10] C. Maes, T. Fabry, J. Keustermans, D. Smeets, P. Suetens et D. Vandermeulen. *Feature detection on 3D face surfaces for pose normalisation and recognition.* Dans IEEE International Conference on Biometrics: Theory, Applications and Systems, BTAS, 2010.

[Mahmood 19] A. Mahmood et S. Al-Maadeed. *Action recognition in poor-quality spectator crowd videos using head distribution-based person segmentation. Machine Vision and Applications, MVA*, 30:1083–1096, 2019.

[Maintz 98] J. B. A. Maintz et M. A. Viergever. *A Survey of Medical Image Registration. Medical Image Analysis, MICCAI Society*, 2(1):1–36, 1998.

[Majidifard 19] H. Majidifard, P. Jin, Y. Adu-Gyamfi et W. G. Buttlar. *PID: A New Benchmark Dataset to Classify and Densify Pavement Distresses. CoRR (arXiv)*, abs/1910.11123, 2019.

[Malon 18] T. Malon, P. Guyot, G. Roman-Jimenez, S. Chambon, V. Charvillat, A. Crouzil, A. Péninou, J. Pinquier, F. Sèdes et C. Sénac. *Toulouse campus surveillance dataset: scenarios, soundtracks, synchronized videos with overlapping and disjoint views.* Dans ACM Multimedia Systems Conference, MMSys, pages 393–398, 2018.

[Malon 19a] T. Malon, S. Chambon, V. Charvillat et A. Crouzil. *Aide à la navigation dans un ensemble de vidéos par reformulation de trajectoires.* Dans actes du congrès francophone de Vision par Ordinateur, ORASIS, 2019.

[Malon 19b] T. Malon, S. Chambon, V. Charvillat et A. Crouzil. *Estimation of Correspondent Trajectories in Multiple Overlapping Synchronized Videos Using Correlation of Activity Functions.* Dans International Conference on Image Processing, ICIP, pages 994–998, 2019.

[Malon 20] T. Malon, S. Chambon, A. Crouzil et V. Charvillat. *Story comparison for estimating field of view overlap in a video collection.* Dans International Conference on Pattern Recognition, ICPR, pageÀ paraître, 2020.

[Mariyanayagam 17] D. Mariyanayagam, P. Gurdjos, S. Chambon, V. Charvillat et F. Brunet. *Rectification métrique de l'image d'un marqueur à contour circulaire à partir d'une caméra calibrée par défaut.* Dans actes du congrès francophone de Vision par Ordinateur, ORASIS, 2017.

[Mariyanayagam 18] D. Mariyanayagam, P. Gurdjos, S. Chambon, F. Brunet et V. Charvillat. *Pose estimation of a single circle using default intrinsic calibration.* Dans Asian Conference on Computer Vision, ACCV, 2018.

[Mariyanayagam 19] D. Mariyanayagam, P. Gurdjos, S. Chambon, F. Brunet et V. Charvillat. *Camera Localization by Single View Query Using One Circular Target.* Dans Scandinavian Conference on Image Analysis, SCIA, 2019.

[Mariyanayagam 20] D. Mariyanayagam. *Localisation et reconnaissance de marqueur circulaire à partir d'une vue de leur contour.* PhD thesis, Institut National Polytechnique de Toulouse, INPT, mars 2020.

[Markelj 12] P. Markelj, D. Tomaževič, B. Likar et F. Pernuš. *A review of 3D/2D registration methods for image-guided interventions. Medical Image Analysis, MICCAI Society*, 16(3):642–661, 2012.

[Marques 11] O. Marques, E. Barenholtz et V. Charvillat. *Context modeling in computer vision: techniques, implications, and applications. Multimedia tools and applications, MTA*, 51(1):1573–7721, 2011.

[Martin 01] D. Martin, C. Fowlkes, D. Tal, J. et Malik. *A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics.* Dans International Conference on Computer Vision, ICCV, volume 2, pages 416–423, juillet 2001.

[Marvin 69] M. Marvin et A. Seymour. Perceptrons. MIT Press, 1969.

[Matas 02] J. Matas, O. Chum, U. Martin et T. Pajdla. *Robust wide baseline stereo from maximally stable extremal regions.* Dans British Machine Vision Conference, BMVC, pages 384–393, 2002.

[Mathevet 99] S. Mathevet, L. Trassoudaine, P. Checchin et J. Alizon. *Combinaison de segmentations en regions. Traitement du Signal*, 16(2):93–104, 1999.

[Matsuo 13] K. Matsuo et Y. Aoki. *Depth Interpolation via Smooth Surface Segmentation Using Tangent Planes Based on the Superpixels of a Color Image.* Dans Internationl Conference on Computer Vision Workshop, ICCV Workshop, pages 29–36, 2013.

[Mayer 16] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy et T. Brox. *A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 4040–4048, 2016.

[McClelland 06] J. R. McClelland, J. M. Blackall, S. Tarte, A. C. Chandler, S. Hughes, S. Ahmad, D. B. Landau et D. J. Hawkes. *A Continuous 4 Motion Model from Multiple Respiratory Cycles for Use in Lung Radiotherapy. The international Journal of Medical Physics Research and Practice, Medical Physics*, 33(9):3348–3358, 2006.

[McCloud 93] S. McCloud. Understanding comics: The invisible art. HarperCollins Publishers, 1993.

[McCulloch 43] W. S. McCulloch et W. Pitts. *A logical calculus of the ideas immanent in nervous activity. Bulletin of mathematical biophysics*, 5:115–133, 1943.

[Meierhold 10] N. Meierhold, M. Spehr, A. Schilling, S. Gumhold et H. G. Maas. *Automatic feature matching between digital images and 2D representations of a 3D laser scanner point cloud. International archives of the photogrammetry, remote sensing and spatial information sciences, ISPRS*, 38:446–451, 2010.

[Mendoza 20] M. Mendoza, J. I. Vasquez-Gomez, H. Taud, L. E. Sucar et C. Reta. *Supervised learning of the next-best-view for 3d object reconstruction. Pattern Recognition Letters, PRL*, 133:224–231, 2020.

[Meyer 92] F. Meyer. *Color image segmentation.* Dans IET International Conference on Image Processing and its Applications, pages 303–306, 1992.

[Mikolajczyk 04] K. Mikolajczyk et C. Schmid. *Scale & Affine Invariant Interest Point Detectors. Internationl Journal on Computer Vision, IJCV*, 60(1):63–86, 2004.

[Mikolajczyk 05] K. Mikolajczyk, B. Leibe et B. Schiele. *Local features for object classe recognition.* Dans International Conference on Computer Vision, ICCV, volume 2, pages 1792–1799, 2005.

[Mille 12] J. Mille, S. Bougleux et L. Cohen. *Minimally overlapping paths sets for closed contour extraction.* Dans International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP, 2012.

[Minaee 20] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz et D. Terzopoulos. *Image Segmentation Using Deep Learning: A Survey. CoRR (arXiv)*, abs/2001.05566, 2020.

[Mineo 19] C. Mineo, S. G. Pierce et R. Summan. *Novel algorithms for 3D surface point cloud boundary detection and edge reconstruction. Journal of Computational Design and Engineering, CDE*, 6(1):81–91, 2019.

[Mitchell 97] T. M. Mitchell. Machine learning. McGraw Hill series in computer science. McGraw-Hill, 1997.

[Mitra 06] N. J. Mitra, L. J. Guibas et M. Pauly. *Partial and Approximate Symmetry Detection for 3D Geometry. ACM transactions on graphics,TOG*, 25(3):560–568, 2006.

[Mitra 12] N. J. Mitra, M. Pauly, M. Wand et D. Ceylan. *Symmetry in 3D Geometry: Extraction and Applications.* Dans Annual Conference of the European Association for Computer Graphics, Eurographics, 2012.

[Mokhtarian 98] F. Mokhtarian et R. Suomela. *Robust Image Corner Detection Through Curvature Scale Space. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 20:1376–1381, 1998.

[Mondéjar-Guerra 18] V. Mondéjar-Guerra, S. Garrido-Jurado, R. Muñoz-Salinas, M. J. Marín-Jiménez et R. Medina-Carnicer. *Robust identification of fiducial markers in challenging conditions. Expert Systems with Applications*, 93:336–345, 2018.

[Monszpart 15] A. Monszpart, N. Mellado, G. J. Brostow et N. J. Mitra. *RAPter: Rebuilding Man-made Scenes with Regular Arrangements of Planes. ACM transactions on graphics,TOG*, 34(4):103:1–103:12, 2015.

[Montanvert 91] A. Montanvert, P. Meer et A. Rosenfeld. *Hierarchical Image Analysis Using Irregular Tessellations. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 13(4):307–316, 1991.

[Moore 08] A. P. Moore, S. Prince, J. Warrell, U. Mohammed et G. Jones. *Superpixel lattices.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, juin 2008.

[Moore 09] A.P. Moore, S. J. D. Prince, J. Warrell, U. Mohammed et G. Jones. *Scene shape priors for superpixel segmentation.* Dans International Conference on Computer Vision, ICCV, pages 771–778, 2009.

[Moravec 80] H. Moravec. *Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover.* Thèse de doctorat, Université de Carnegie Mellon, CMU, 1980.

[Moreno-Ingelmo 08] A. Moreno-Ingelmo, S. Chambon, A. P. Santhanam, J. P. Rolland, E. D. Angelini et I. Bloch. *Combining a breathing model and tumor-specific rigidity constraints for registration of CT-PET thoracic data. Computer Aided Surgery*, 13(5):281–298, 2008.

[Mori 04] G. Mori, X. Ren, A. A. Efros et J. Malik. *Recovering human body configurations: combining segmentation and recognition.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, volume 2, 2004.

[Mori 05] G. Mori. *Guiding model search using segmentation.* Dans International Conference on Computer Vision, ICCV, volume 2, pages 1417–1423, 2005.

[Morin 14] G. Morin. *3D Models for ...* Habilitation à diriger des recherches, Institut National Polytechnique de Toulouse, INPT, 2014.

[Morreale 19] L. Morreale, A. Romanoni et M. Matteucci. *Dense 3D Visual Mapping via Semantic Simplification. CoRR (arXiv)*, abs/1902.07511, 2019.

[Muresan 15] M. P. Muresan, M. Negru et S. Nedevschi. *Improving local stereo algorithms using binary shifted windows, fusion and smoothness constraint.* Dans IEEE International Conference on Intelligent Computer Communication and Processing, ICCP, pages 179–185, 2015.

[Myrtille 05] Myrtille. *Petit Homme.* Dans Murmures, 2005.

[Mühlmann 02] K. Mühlmann, D. Maier, J. Hesser et R. Männer. *Calculating dense disparity maps from color stereo images. Internationl Journal on Computer Vision, IJCV*, 47(1):79–88, 2002.

[Na 12] W. Na et W. Tao. *Proximal support vector machine based pavement image classification.* Dans International Conference on Advanced Computational Intelligence, pages 686–688, 2012.

[Nader 14] G. Nader, G. Guennebaud et N. Mellado. *Adaptive multi-scale analysis for point-based surface editing. Computer Graphics Forum*, 33(7):171–179, 2014.

[Nair 10] V. Nair et G. E. Hinton. *Rectified Linear Units Improve Restricted Boltzmann Machines.* Dans International Conference on Machine Learning, ICML, page807–814, 2010.

[Nam 11] W. H. Nam, D.-G. Kang, D. Lee, J. Y. Lee et J. B. Ra. *Automatic registration between 3D intra-operative ultrasound and pre-operative CT images of the liver based on robust edge matching. The international journal of biomedical physics and engineering, Physics in Medicine & Biology, PMB*, 57(1), 2011.

[Narusawa 01] U. Narusawa. *General characteristics of the sigmoidal model equation representing quasi-static pulmonary P-V curves. Journal of Applied Physiology, JAPPL*, 92(1):201–210, 2001.

[Nathan 19] S. Nathan et P. Kansal. *SkeletonNet: Shape Pixel to Skeleton Pixel.* Dans IEEE Conference on Computer Vision Pattern Recognition Workshop, CVPR Workshop, 2019.

[Nehmeh 04] S. A. Nehmeh, Y. E. Erdi, T. Pan, A. Pevsner, K. E. Rosenzweig, E. Yorke, G. S. Mageras, H. Schoder, P. Vernon, O. Squire, H. Mostafavi, S. M. Larson et J. L. Humm. *Four-dimensional (4D) PET/CT imaging of the thorax. The international journal of biomedical physics and engineering, Physics in Medicine & Biology, PMB*, 31(12):3179–3186, 2004.

[Nguyen 11] T. S. Nguyen, S. Begot, F. Duculty et M. Avila. *Free-Form Anisotropy: A New Method For Crack Detection On Pavement Surface Images.* Dans International Conference on Image Processing, ICIP, pages 1069–1072, 2011.

[Nguyen 17] R. M. H. Nguyen et M. S. Brown. *Why You Should Forget Luminance Conversion and Do Something Better.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2017.

[Ohta 80] Y.-I. Ohta, T. Kanade et T. Sakai. *Color Information for Region Segmentation. Computer Graphics and Image Processing, CGIP*, 13(3):222–241, juillet 1980.

[Ohtake 04] Y. Ohtake, A. Belyaev et H.-P. Seidel. *Ridge-valley Lines on Meshes via Implicit Surface Fitting. ACM transactions on graphics,TOG*, 23(3):609–612, 2004.

[Ojala 02] T. Ojala, M. Pietikainen et T. Maenpaa. *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 24(7):971–987, 2002.

[Okutomi 92] M. Okutomi et G. Tomita. *Color stereo matching and its application to 3-d measurement of optic nerve head.* Dans International Conference on Pattern Recognition, ICPR, volume 1, pages 509–513, 1992.

[Oliveira 08] H. Oliveira et P. L. Correia. *Supervised Strategies For Cracks Detection in Images of Road Pavement Flexible Surfaces.* Dans European Signal Processing Conference, EUSIPCO, 2008.

[Oliveira 13] H. Oliveira et P. L. Correia. *Automatic Road Crack Detection and Characterization. IEEE Transactions on Intelligent Transportation Systems, TITS*, 14(1):155–168, 2013.

[Otsu 79] N. Otsu. *A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man and Cybernetics, TSMC*, 9(1):62–66, 1979.

[Otto 89] G. P. Otto et T. K. W. Chau. *"Region-growing" algorithm for matching of terrain images. Image and Vision Computing, IVC*, 7(2):83–94, 1989.

[Pang 15] G. Pang et U. Neumann. *Fast and Robust Multi-view 3D Object Recognition in Point Clouds.* Dans International Conference on 3D Vision, 3DV, pages 171–179, 2015.

[Pang 17] J. Pang, W. Sun, J. S. J. Ren, C. Yang et Q. Yan. *Cascade residual learning: A two-stage convolutional neural network for stereo matching.* Dans Internationl Conference on Computer Vision Workshop, ICCV Workshop, pages 887–895, 2017.

[Panichev 19] 0. Panichev et A. Voloshyna. *U-Net Based Convolutional Neural Network for Skeleton Extraction.* Dans IEEE Conference on Computer Vision Pattern Recognition Workshop, CVPR Workshop, 2019.

[Paris 09] S. Paris et F. Durand. *A Fast Approximation of the Bilateral Filter Using a Signal Processing Approach. Internationl Journal on Computer Vision, IJCV*, 81(1):24–52, 2009.

[Parlett 98] B. Parlett. The Symmetric Eigenvalue Problem. Society for Industrial and Applied Mathematics (SIAM), 1998.

[Paudel 14] D. P. Paudel, C. Demonceaux, A. Habed et P. Vasseur. *Localization of 2D Cameras in a Known Environment Using Direct 2D-3D Registration.* Dans International Conference on Pattern Recognition, ICPR, pages 196–201, 2014.

[Pentland 87] A. P. Pentland. *A New Sense for Depth of Field. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 9(4):523–531, 1987.

[Perollaz 08] M. Perollaz. *Détection d'obstacles multi-capteurs supervisée par stéréovision.* Thèse de doctorat, Université Pierre et Marie Curie, Paris 6, 2008.

[Pizenberg 17] M. Pizenberg, A. Carlier, E. Faure et V. Charvillat. *Outlining Objects for Interactive Segmentation on Touch Devices.* Dans ACM International Conference on Multimedia, MM, pages 1734–1742, 2017.

[Plötz 17] T. Plötz et S. Roth. *Automatic Registration of Images to Untextured Geometry Using Average Shading Gradients. Internationl Journal on Computer Vision, IJCV*, 125(1):65–81, 2017.

[Pluim 03] J. P. W. Pluim et J. M. Fitzpatrick. *Image Registration. IEEE Transactions on Medical Imaging, TMI*, 22(11):1341–1343, 2003.

[Pollari 04] M. Pollari, J. Lotjonen, T. Makela, N. Pauna, A. Reilhac et P. Clarysse. *Evaluation of cardiac PET-MRI registration methods using a numerical breathing phantom.* Dans IEEE International Symposium on Biomedical Imaging, ISBI, pages 1447–1450, 2004.

[Pomerleau 15] F. Pomerleau, F. Colas et R. Siegwart. *A Review of Point Cloud Registration Algorithms for Mobile Robotics. Foundations and Trends in Robotics*, 4(1):1–104, 2015.

[Pont-Tuset 15] J. Pont-Tuset et F. Marqués. *Supervised evaluation of image segmentation and object proposal techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 38(7):1465–1478, 2015.

[Promayon 97] E. Promayon, P. Baconnier et C. Puech. *Physically-Based Model for Simulating the Human Trunk Respiration Movements.* Dans Joint Conference Computer Vision, Virtual Reality and Robotics in Medicine and Medical Robotics and Computer-Assisted Surgery, CVRMed-MRCAS, volume 1205 de *Lecture Notes in Computer Science*, pages 379–388, 1997.

[Prévost 13a] S. Prévost, C. Niquin, S. Chambon et G. Galès. 3d video: from capture to diffusion, Chapitre Multi- and Stereoscopic Matching, Depth and Disparity, pages 137–155. L. Lucas and C. Loscos and Y. Remion, Wiley, 2013.

[Prévost 13b] S. Prévost, C. Niquin, S. Chambon et G. Galès. Vidéo 3d : Capture, traitement et diffusion, Chapitre Stéréocorrélation, profondeur et disparité, pages 147–164. L. Lucas and C. Loscos and Y. Remion, Traité IC2, série signal et image, Hermès–Lavoisier, 2013.

[Qing 12] Y. Qing, Q.-J. Liu, B.-J. Yang, H.-M. Chen et J. Tang. *An insect imaging system to automate rice light-trap pest identification. Journal of Integrative Agriculture*, 11(6):978–985, 2012.

[Quinlan 86] J. R. Quinlan. *Induction of decision trees. Machine Learning, ML*, 1:81–106, 1986.

[Radenović 16] F. Radenović, G. Tolias et O. Chum. *CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples.* Dans European Conference on Computer Vision, ECCV, pages 3–20, 2016.

[Radford 15] A. Radford, L. Metz et S. Chintala. *Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR (arXiv)*, abs/1511.06434, 2015.

[Radford 16] A. Radford, L. Metz et S. Chintala. *Unsupervised representation learning with deep convolutional generative adversarial networks.* Dans International Conference on Learning Representations, ICLR, 2016.

[Ramalingam 09] S. Ramalingam, S. Bouaziz, P. Sturm et M. Brand. *Geolocalization using skylines from omni-images.* Dans Internationl Conference on Computer Vision Workshop, ICCV Workshop, pages 23–30, 2009.

[Rashwan 16] H. A. Rashwan, S. Chambon, P. Gurdjos, G. Morin et V. Charvillat. *Towards multi-scale feature detection repeatable over intensity and depth images.* Dans International Conference on Image Processing, ICIP, pages 36–40, 2016.

[Rashwan 17] H. A. Rashwan, S. Chambon, P. Gurdjos, G. Morin et V. Charvillat. *Towards Recognizing of 3D Models Using A Single Image.* Dans Eurographics Workshop on 3D Object Retrieval, 3DOR, 2017.

[Rashwan 19a] H. A. Rashwan, S. Chambon, P. Gurdjos, G. Morin et V. Charvillat. *Using Curvilinear Features in Focus for Registering a Single Image to a 3D Object. IEEE Transactions on Image Processing, TIP*, 29(8), 2019.

[Rashwan 19b] H. A. Rashwan, M. Á. García, S. Chambon et D. Puig. *Gait representation and recognition from temporal co-occurrence of flow fields. Machine Vision and Applications, MVA*, 30:139–152, 2019.

[Redmon 18] J. Redmon et A. Farhadi. *YOLOv3: An Incremental Improvement. CoRR (arXiv)*, abs/1804.02767, 2018.

[Ren 03] X. Ren et J. Malik. *Learning a classification model for segmentation.* Dans International Conference on Computer Vision, ICCV, volume 1, pages 10–17, octobre 2003.

[Ren 07] X. Ren et J. Malik. *Tracking as Repeated Figure/Ground Segmentation.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2007.

[Ricard 05a] J. Ricard. *Indexation et recherche d'objets 3D à partir de requêtes 2D et 3D.* Thèse de doctorat, Université Claude Bernard Lyon 1, décembre 2005.

[Ricard 05b] J. Ricard, D. Coeurjolly et A. Baskurt. *Generalizations of angular radial transform for 2D and 3D shape retrieval. Pattern Recognition Letters, PRL*, 26(14):2174–2186, 2005.

[Rit 06] S. Rit, D. Sarrut, V. Boldea et C. Ginestet. *Extraction du signal respiratoire à partir de projections cone-beam pour l'imagerie TDM 4D. Traitement du signal, TS*, 23(3–4):307–319, 2006.

[Rohlfing 01] T. Rohlfing, C. R. Maurer et J. Zhong. *Modeling liver motion and deformation during the respiratory cycle using intensity-based free-form registration of gated MR images.* Dans SPIE Medical Imaging Conference, volume 4319, pages 337–348, 2001.

[Rohlfing 04] T. Rohlfing, C. R. Jr. Maurer, W. G. O'Dell et J. Zhong. *Modeling Liver Motion and Deformation During the Respiratory Cycle Using Intensity-Based Free-Form Registration of Gated MR Images. The international Journal of Medical Physics Research and Practice, Medical Physics*, 31(3):427–432, 2004.

[Rohr 01] K. Rohr, H. S. Stiehl, R. Sprengel, T. M. Buzug, J. Weese et M. H. Kuhn. *Landmark-Based Elastic Registration Using Approximating Thin-Plate Splines. IEEE Transactions on Medical Imaging, TMI*, 20(6):526–534, 2001.

[Ronneberger 15] O. Ronneberger, P. Fischer et T. Brox. *U-net: Convolutional networks for biomedical image segmentation.* Dans International Conference on Medical image computing and computer-assisted intervention, pages 234–241, 2015.

[Rosenblatt 58] F. Rosenblatt. *The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review*, 65(6):386–408, 1958.

[Rosenhahn 04] B. Rosenhahn, H. Ho et R. Klette. *Block matching based 2D-3D pose estimation.* Dans Image and Vision Computing, IVC, pages 309–314, 2004.

[Rosenhahn 07] B. Rosenhahn, T. Brox et J. Weickert. *Three-Dimensional Shape Knowledge for Joint Image Segmentation and Pose Tracking. Internationl Journal on Computer Vision, IJCV*, 73:243–262, 2007.

[Rosten 06] E. Rosten et T. Drummond. *Machine Learning for High-Speed Corner Detection.* Dans European Conference on Computer Vision, ECCV, pages 2564–2571, 2006.

[Rousseeuw 92] P. J. Rousseeuw et C. Croux. *L_1-Statistical Analysis and Related Methods.* Dans Y. Dodge, éditeur, Explicit Scale Estimators with High Breakdown Point, pages 77–92. Elsevier, 1992.

[Rublee 11] E. Rublee, V. Rabaud, K. Konolige et G. Bradski. *ORB: an efficient alternative to SIFT or SURF*. Dans International Conference on Computer Vision, ICCV, pages 2564–2571, 2011.

[Ruiz-Hidalgo 12] J. Ruiz-Hidalgo, J. R. Morros, P. Aflaki, F. Calderero et R. Marqués. *Multiview depth coding based on combined color/depth segmentation. Journal of Visual Communication and Image Representation, VCIR*, 23(1):42–52, 2012.

[Rumelhart 86] D. E. Rumelhart, G. E. Hinton et R. J. Williams. *Learning representations by back-propagating errors. Nature*, 323(6088):533–536, 1986.

[Russakovsky 15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg et L. Fei-Fei. *ImageNet Large Scale Visual Recognition Challenge. Internationl Journal on Computer Vision, IJCV*, 115(3):211–252, 2015.

[Salari 11] E. Salari et X. Yu. *Pavement distress detection and classification using a Genetic Algorithm*. Dans IEEE Applied Imagery Pattern Recognition Workshop, pages 1–5, 2011.

[Samuel 59] A. L. Samuel. *Some studies in machine learning using the game of Checkers. IBM journal of Research and Development*, pages 71–105, 1959.

[Santhanam 07] A. P. Santhanam, C. Imielinska, P. Davenport, P. Kupelian et J. P. Rolland. *Modeling Simulation and Visualization of Real-Time 3D Lung Dynamics. IEEE Transactions on Information Technology in Biomedicine, TITB*, 2007. À paraître.

[Sarker 19] M. K. Sarker, S. F. Banu, H. A. Rashwan, M. Abdel-Nasser, V. K. Singh, S. Chambon, P. Radeva et D. Puig. *Food Places Classification in Egocentric Images Using Siamese Neural Networks*. Dans International Conference of the Catalan Association for Artificial Intelligence, volume 319, page145, 2019.

[Sarrut 05] D. Sarrut, V. Boldea, M. Ayadi, J.-N. Badel, C. Ginestet, S. Clippe et C. Carrie. *Non-rigid registration method to assess reproducibility of breath-holding with ABC in lung cancer. International Journal of Radiation Oncology–Biology–Physis*, 61(2):594–607, 2005.

[Sattler 11] T. Sattler, B. Leibe et L. Kobbelt. *Fast image-based localization using direct 2D-to-3D matching*. Dans International Conference on Computer Vision, ICCV, pages 667–674, 2011.

[Sayed 13] T. Sayed, M. H. Zaki et J. Autey. *Automated safety diagnosis of vehicle–bicycle interactions using computer vision analysis. Safety Science*, 59:163–172, 2013.

[Scharstein 02a] D. Scharstein et Richard R. Szeliski. *A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. Internationl Journal on Computer Vision, IJCV*, 47(1):7–42, 2002.

[Scharstein 02b] D. Scharstein et R. Szeliski. *A Taxomomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. Internationl Journal on Computer Vision, IJCV*, 47(1):7–42, avril 2002.

[Scharstein 03] D. Scharstein et R. Szeliski. *High-Accuracy Stereo Depth Maps Using Structured Light*. Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, volume 1, pages 195–202, 2003.

[Scharstein 07] D. Scharstein et C. Pal. *Learning conditional random fields for stereo*. Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2007.

[Scharstein 14] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang et P. Westling. *High-resolution stereo datasets with subpixel-accurate ground truth*. Dans German Conference on Pattern Recognition, GCPR, 2014.

[Schmid 00] C. Schmid, R. Mohr et C. Bauckhage. *Evaluation of interest point detectors. Internationl Journal on Computer Vision, IJCV*, 37(2):151–172, 2000.

[Schurischuster 16] S. Schurischuster, S. Zambanini, M. Kampel et B. Lamp. *Sensor Study for Monitoring Varroa Mites on Honey Bees (Apis mellifera).* Dans Visual observation and analysis of Vertebrate And Insect Behavior Workshop, VAIB, 2016.

[Schweikard 00] A. Schweikard, G. Glosser, M. Bodduluri, M. J. Murphy et J. R. Adler. *Robotic motion compensation for respiratory movement during radiosurgery. Computer Aided Surgery, CAS*, 5(4):263–277, 2000.

[Scott 03] W. R. Scott, G. Roth et J.-F. Rivest. *View Planning for Automated Three-Dimensional Object Reconstruction and Inspection. ACM Computing Surveys*, 35(1):64–96, 2003.

[Scovanner 07] P. Scovanner, S. Ali et M. Shah. *A 3-Dimensional Sift Descriptor and Its Application to Action Recognition.* Dans ACM International Conference on Multimedia, MM, pages 357–360, 2007.

[Sebari 07] I. Sebari et D. C. He. *Les approches de segmentation d'image par coopération régions-contours. Revue Télédétection*, 7:499–506, 2007.

[Segars 02] W. P. Segars et B. M. W. Tsui. *Study of the Efficacy of Respiratory Gating in Myocardial SPECT Using the New 4-D NCAT Phantom. IEEE Transactions on Nuclear Science, TNS*, 49(3):675–679, 2002.

[Sepp 97] H. Sepp et J. Schmidhuber. *Long Short-Term Memory. Neural Computation*, 9(8):1735–1780, novembre 1997.

[Sethian 95] J. A. Sethian. *A Fast Marching Level Set Method for Monotonically Advancing Fronts.* Dans National Academy of Science, pages 1591–1595, 1995.

[Shaiek 13] A. Shaiek. *Reconnaissance d'objets 3D par points d'intérêt.* Thèse de doctorat, École Nationale Supérieure des Mines de Paris, 2013.

[Shaoqing 15] R. Shaoqing, K. He, R. Girshick et J. Sun. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.* Dans International Conference on Neural Information Processing Systems, NIPS, pages 91–99, 2015.

[Shekhar 05] R. Shekhar, V. Walimbe, S. Raja, V. Zagrodsky, M. Kanvinde, G. Wu et B. Bybel. *Automated 3-Dimensional Elastic Registration of Whole-Body PET and CT from Separate or Combined Scanners. The Journal of Nuclear Medicine, JNM*, 46(9):1488–1496, 2005.

[Shen 19] T. Shen, Z. Luo, R. Zhang, S. Zhu, T. Fang et L. Quan. *Matchable Image Retrieval by Learning from Surface Reconstruction.* Dans Asian Conference on Computer Vision, ACCV, pages 415–431, 2019.

[Shi 00] J. Shi et J. Malik. *Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 22(8):888–905, 2000.

[Shu 19] Z. Shu, S. Xin, X. Xu, L. Liu et L. Kavan. *Detecting 3D Points of Interest Using Multiple Features and Stacked Auto-encoder. IEEE Transactions on Visualization and Computer Graphics, TVCG*, 25:2583–2596, 2019.

[Shum 98] H.-Y. Shum et R. Szeliski. *Construction and refinement of panoramic mosaics with global and local alignment.* Dans International Conference on Computer Vision, ICCV, pages 953–956, 1998.

[Simo-Serra 15] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua et F. Moreno-Noguer. *Discriminative Learning of Deep Convolutional Feature Point Descriptors.* Dans International Conference on Computer Vision, ICCV, pages 118–126, 2015.

[Simonetto 05] E. Simonetto et A. Soadane. *Colour Image Watermarking Using a Visual Sub-Band Decomposition. International Journal of Robotics and Automation, IJRA*, 20(2):101–108, 2005.

[Simonyan 15] K. Simonyan et A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition.* Dans International Conference on Learning Representations, ICLR, 2015.

[Smeulders 03] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan et M. Shah. *Visual Tracking: An Experimental Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 36(7):1355–1360, 2003.

[Smith 97] S. Smith et J. Brady. *SUSAN – A New Approach to Low Level Image Processing. Internationl Journal on Computer Vision, IJCV*, 23(1):45–78, 1997.

[Snavely 06] N. Snavely, S. M. Seitz et R. Szeliski. *Photo tourism: Exploring photo collections in 3D.* Dans ACM Special Interest Group on Computer Graphics and Interactive Techniques Conference, SIGGRAPH, pages 835–846, 2006.

[Song 14] S. Song et J. Xiao. *Sliding shapes for 3d object detection in depth images.* Dans European Conference on Computer Vision, ECCV, pages 634–651, 2014.

[Steckenrider 20] J. J. Steckenrider et T. Furukawa. *A Probabilistic Superpixel-Based Method for Road Crack Network Detection.* Dans Advances in Computer Vision, pages 303–316, 2020.

[Stekovic 20] S. Stekovic, F. Fraundorfer et V. Lepetit. *General 3D Room Layout from a Single View by Render-and-Compare. CoRR (arXiv)*, abs/2001.02149, 2020.

[Su 13] J. Su, A. Srivastava et F. W. Huffer. *Detection, classification and estimation of individual shapes in 2D and 3D point clouds. Computational Statistics & Data Analysis, CSDA*, 58:227–241, 2013. The Third Special Issue on Statistical Signal Extraction and Filtering.

[Su 15] H. Su, C. R. Qi, Y. Li et L. J. Guibas. *Render for CNN: Viewpoint Estimation in Images Using CNNs Trained With Rendered 3D Model Views.* Dans International Conference on Computer Vision, ICCV, 2015.

[Subirats 06] P. Subirats, O. Fabre, J. Dumoulin, V. Legeay et D. Barba. *Automation of pavement surface crack detection with a matched filtering to define the mother wavelet function used.* Dans European Signal Processing Conference, EUSIPCO, 2006.

[Sun 02] J. Sun, H-Y. Shum et N-N. Zheng. *Stereo Matching Using Belief Propagation.* Dans European Conference on Computer Vision, ECCV, pages 450–452, 2002.

[Sun 03] J. Sun. *Stereo Matching Using Belief Propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 25(7):787–800, 2003.

[Sun 05a] C. Sun et B. Appleton. *Multiple paths extraction in images using a constrained expanded trellis. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 27:1923–1933, 2005.

[Sun 05b] J. Sun, S. B. Kang et H. Y. Shum. *Symmetric stereo matching for occlusion handling.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, volume 2, pages 399–406, 2005.

[Sundaram 05] T. A. Sundaram et J. C. Gee. *Towards a Model of Lung Biomechanics: Pulmonary Kinematics Via Registration of Serial Lung Images. Medical Image Analysis, MICCAI Society*, 9(6):524–537, 2005.

[Sutton 08] M. A. Sutton, J. H. Yan, V. Tiwari, H. W. Schreier et J. J. Orteu. *The effect of out-of-plane motion on 2D and 3D digital image correlation measurements. Optics and Lasers in Engineering*, 46:746–757, 2008.

[Swain 91] M. J. Swain et D. H. Ballard. *Color Indexing. Internationl Journal on Computer Vision, IJCV*, 7(1):11–32, 1991.

[Szegedy 15] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke et A. Rabinovich. *Going deeper with convolutions.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2015.

[Szeliski 10] R. Szeliski. Computer Vision: Algorithms and Applications. Springer, 2010. `http://szeliski.org/Book/`.

[Taguchi 08] Y. Taguchi, B. Wiburn et C. L. Zitnick. *Stereo reconstruction with mixed pixels using adaptive over-segmentation.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2008.

[Tai 09] Y.-W. Tai et M. S. Brown. *Single image defocus map estimation using local contrast prior.* Dans International Conference on Image Processing, ICIP, pages 1797–1800, 2009.

[Taketomi 17] T. Taketomi, H. Uchiyama et S. Ikeda. *Visual SLAM algorithms: a survey from 2010 to 2016. IPSJ Transactions on Computer Vision and Applications*, 9(1):16, dec 2017.

[Tam 13] G. K. Tam, Z.-Q. Cheng, Y.-K. Lai, F. C. Langbein, Y. Liu, D. Marshall, R. R. Martin, X.-F. Sun et P. L. Rosin. *Persuading visual attention through geometry. IEEE Transactions on Visualization and Computer Graphics, TVCG*, 19(7):1199–1217, 2013.

[Tamaazousti 11] M. Tamaazousti, V. Gay-Bellile, S. N. Collette, S. Bourgeois et M. Dhome. *NonLinear refinement of structure from motion reconstruction by taking advantage of a partial knowledge of the environment.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 3073–3080, 2011.

[Tanaka 98] N. Tanaka et K. Uematsu. *A Crack Detection Method in Road Surface Images Using Morphology.* Dans Workshop on Machine Vision and Applications, MVA, pages 154–157, 1998.

[Tang 13] J. Tang et Y. Gu. *Automatic Crack Detection and Segmentation Using a Hybrid Algorithm for Road Distress Analysis.* Dans IEEE Transactions on Systems, Man and Cybernetics, TSMC, pages 3026–3030, 2013.

[Tian 17] Y. Tian, B. Fan et F. Wu. *L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2017.

[Tibshirani 01] R. Tibshirani, G. Walther et T. Hastie. *Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[Tolias 16] G. Tolias, R. Sicre et H. Jégou. *Particular object retrieval with integral max-pooling of CNN activations.* Dans International Conference on Learning Representations, ICLR, 2016.

[Tombari 08] F. Tombari, S. Mattoccia, L. D. Stefano et E. Addimanda. *Classification and evaluation of cost aggregation methods for stereo correspondence.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2008.

[Tomikawa 99] T. Tomikawa. *A study of road crack detection by the meta-genetic algorithm.* Dans IEEE Africon, volume 1, pages 543–548, 1999.

[Tong 10] H. Tong, S. Liu, N. Liu et N. Barnes. *A novel object-oriented stereo matching on multi-scale superpixels for low-resolution depth mapping.* Dans IEEE International Conference of Engineering in Medicine and Biology, EMBC, pages 5046–5049, 2010.

[Toshev 09] A. Toshev, A. Makadia et K. Daniilidis. *Shape-based object recognition in videos using 3D synthetic object models.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 288–295, 2009.

[Trémeau 04] A. Trémeau, C. Fernandez-Maloigne et P. Bonton. Image numérique couleur – de l'acquisistion au traitement. Dunod, 2004.

[Trémeau 03] A. Trémeau et P. Colantoni. *Visualisation 3D dédiée à l'analyse des images couleur.* Dans actes du congrès francophone de Vision par Ordinateur, ORASIS, pages 261–268, 2003.

[Tu 10] Z. Tu et X. Bai. *Auto-Context and Its Application to High-Level Vision Tasks and 3D Brain Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 32(10):1744–1757, 2010.

[Tulsiani 15] S. Tulsiani et J. Malik. *Viewpoints and keypoints.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 1510–1519, 2015.

[Tuytelaars 04] T. Tuytelaars et L. Van Gool. *Matching Widely Separated Views Based on Affine Invariant Regions. Internationl Journal on Computer Vision, IJCV*, 59:61–85, 2004.

[Tuytelaars 06] T. Tuytelaars. Local Invariant Features: What? Why? When? How? Tutorial in European Conference on Computer Vision, ECCV, 2006.

[Tuytelaars 08] T. Tuytelaars et K. Mikolajczyk. *Local Invariant Feature Detectors: A Survey. Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280, 2008.

[Tuzel 06] O. Tuzel, F. Porikli et P. Meer. *Region covariance: A fast descriptor for detection and classification.* Dans European Conference on Computer Vision, ECCV, pages 589–600, 2006.

[Urano 19] K. Urano, K. Hiroi, S. Kato, N. Komagata et N. Kawaguchi. *Road Surface Condition Inspection Using a Laser Scanner Mounted on an Autonomous Driving Car.* Dans International Workshop on Pervasive Computing for Vehicular Systems, PerVehicle, pages 826–831, 2019.

[Čech 09] J. Čech. *Accurate and Robust Stereoscopic Matching in Efficient Algorithms.* Thèse de doctorat, Université Technique de Prague, février 2009.

[Van den Bergh 12] M. Van den Bergh, X. Boix, G. Roig, B. de Capitani et L. J. Van Gool. *SEEDS: Superpixels Extracted via Energy-Driven Sampling.* Dans European Conference on Computer Vision, ECCV, pages 13–26, 2012.

[Vandenbroucke 00] N. Vandenbroucke. *Segmentation d'images couleur par classification de pixels dans des espaces d'attributs colorimétriques adaptés. Application à l'analyse d'images de football.* Thèse de doctorat, Université des sciences et technologies de Lille 1, 2000.

[Vedaldi 08] A. Vedaldi et S. Soatto. *Quick Shift and Kernel Methods for Mode Seeking.* Dans European Conference on Computer Vision, ECCV, pages 179–192, 2008.

[Veeraraghavan 05] A. Veeraraghavan, A. K. Roy-Chowdhury et R. Chellappa. *Matching Shape Sequences in Video with Applications in Human Movement Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 27(12):1896–1909, 2005.

[Veksler 02] O. Veksler. *Stereo correspondence with compact windows via minimum ratio cycle. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 24(12):1654–1660, 2002.

[Veksler 05] O. Veksler. *Stereo Correspondence by Dynamic Programming on a Tree.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 384–390, 2005.

[Veksler 10] O. Veksler, Y. Boykov et P. Mehrani. *Superpixels and supervoxels in an energy optimization framework.* Dans European Conference on Computer Vision, ECCV, pages 211–224, 2010.

[Venegas 98] J. G. Venegas, R. S. Harris et B. A. Simon. *A comprehensive equation for the pulmonary pressure-volume curve. Journal of Applied Physiology, JAPPL*, 84(1):389–395, 1998.

[Ventura 19a] C. Ventura, M. Bellver Bueno, A. Girbau, A. Salvador, F. Marqués et X. Giro-i-Nieto. *RVOS: End-to-end recurrent network for video object segmentation.* Dans Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5277–5286, 2019.

[Ventura 19b] C. Ventura, D. Varas, V. Vilaplana, X. Giro-i-Nieto et F. Marqués. *Multiresolution co-clustering for uncalibrated multiview segmentation. Signal Processing: Image Communication*, 76:151–166, 2019.

[Vincent 91] L. Vincent et P. Soille. *Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 13(6):583–598, 1991.

[Vogel 06] W. V. Vogel, J. A. van Dalen, D. A. Schinagl, J. H. Kaanders, H. Huisman, F. H. Corstens et W. J. Oyen. *Correction of an image size difference between positron emission tomography (PET) and computed tomography (CT) improves image fusion of dedicated PET and*

CT. The international journal of biomedical physics and engineering, Physics in Medicine & Biology, PMB, 27(6):515–519, 2006.

[Wang 02] K. C. P. Wang et W. Gong. *Automated Pavement Distress Survey: A Review and A New Direction.* Dans Pavement Evaluation Conference, PEC, pages 21–25, 2002.

[Wang 06] L. Wang, M. Gong, M. Gong et R. Yang. *How far can we go with local optimization in real-time stereo matching.* Dans International Symposium of Three-Dimensional Data Processing, Visualization, and Transmission, 3DPVT, pages 129–136, 2006.

[Wang 08] Z. F. Wang et Z. G. Zheng. *A region based stereo matching algorithm using cooperative optimization.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2008.

[Wang 10] J. Wang, J. Li, W. Yau et E. Sung. *Boosting dense SIFT descriptors and shape contexts of face images for gender recognition.* Dans IEEE Conference on Computer Vision Pattern Recognition Workshop, CVPR Workshop, pages 96–102, 2010.

[Wang 12a] J. Wang, C. Lin, L. Ji et A.-P. Liang. *A new automatic identification system of insect images at the order level. Knowledge-Based Systems*, 33:102–110, 2012.

[Wang 12b] J. Wang et X. Wang. *VCells: Simple and Efficient Superpixels Using Edge-Weighted Centroidal Voronoi Tessellations. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 34:1241–1247, 2012.

[Wang 12c] X.-C. Wang, J.-J. Cao, X.-P. Liu, B.-J. Li, X.-Q. Shi et Y.-Z. Sun. *Feature detection of triangular meshes via neighbor supporting. Journal of Zhejiang University SCIENCE C*, 13(6):440–451, 2012.

[Wang 18] Q. Wang, X. Zhou et K. Daniilidis. *Multi-image Semantic Matching by Mining Consistent Features.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 685–694, 2018.

[Weber 10] C. Weber, S. Hahmann et H. Hagen. *Sharp Feature Detection in Point Clouds.* Dans IEEE International Conference on Shape Modeling and Applications, SMI, 2010.

[Wegner 09] K. Wegner et O. Stankiewicz. *Similarity measures for depth estimation.* Dans IEEE 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 3DTV-CON, 2009.

[Weippl 07] E. R. Weippl, M. Klemen, S. Fenz, A. Ekelhart et A. M. Tjoa. *The semantic desktop: a semantic personal information management system based on RDF and topic maps.* Dans VLDB conference on Ontologies-based databases and information systems, pages 135–151, 2007.

[Wen 15] C. Wen, D. Wu et H. Hu. *Pose estimation-dependent identification method for field moth images using deep learning architecture. Biosystems Engineering*, 136:117–128, 2015.

[Weng 19] X. Weng, Y. Huang et W. Wang. *Segment-based pavement crack quantification. Automation in Construction*, 105, 2019.

[Wenzel 13] K. Wenzel, M. Rothermel, F. D. Fritsch et N. Haala. *Image acquisition and model selection for multi-view stereo. International archives of the photogrammetry, remote sensing and spatial information sciences, ISPRS*, 5:251–258, 2013.

[West 05] J. B. West, C. R. Maurer et J. R. Dooley. *Hybrid Point-and-Intensity-Based Deformable Registration for Abdominal CT Images.* Dans SPIE Medical Imaging Conference, volume 5747, pages 204–211, 2005.

[Wong 92] E. K. Wong. *Model matching in robot vision by subgraph isomorphism. Pattern Recognition, PR*, 25(3):287–303, 1992.

[Woodford 08] O. Woodford, P. Torr, I. Reid et A. Fitzgibbon. *Global stereo reconstruction under second order smoothness priors.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2008.

[Wright 14] D. Wright, V. Baltazart, N. Elsworth, R. Hamrouche, J. Krarup, M. Lurdes Antunes, S. McRobbie, V. Merecos et T. Saarenketo. TRIMM D4.3: Monitoring Structural and Surface Conditions. Rapport Technique, Collaborative project FP7-285119, European Community, 2014.

[Wu 08] C. Wu, B. Clipp, X. Li, J. M. Frahm et M. Pollefeys. *3D model matching with Viewpoint-Invariant Patches (VIP).* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2008.

[Wu 18] Y. Wu, F. Tang et H. Li. *Image-based camera localization: an overview. Visual Computing for Industry, Biomedicine, and Art*, septembre 2018.

[Wu 19] Z. Wu, X. Wu, X. Zhang, S. Wang et L. Ju. *Semantic Stereo Matching with Pyramid Cost Volumes.* Dans International Conference on Computer Vision, ICCV, pages 7484–7493, 2019.

[Xiao 10] J. Xiao, J. Hays, K. Ehinger, A. Oliva et A. Torralba. *SUN Database: Large-scale Scene Recognition from Abbey to Zoo.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2010.

[Xie 10] Y. Xie, R. Cao, H. Tong, N. Liu et S. Liu. *Evaluating Multi-scale Over-segment and Its Contribution to Real Scene Stereo Matching by High-Order MRFs.* Dans IEEE International Conference on Digital Image Computing: Techniques and Applications, DICTA, pages 235–240, 2010.

[Xiong 07] W. Xiong et J. Jia. *Stereo matching on objects with fractional boundary.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2007.

[Xu 15] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel et Y. Bengio. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.* Dans International Conference on Machine Learning, ICML, pages 2048–2057, 2015.

[Xu 17] C. Xu, L. Zhang, L. Cheng et R. Koch. *Pose Estimation from Line Correspondences: A Complete Analysis and a Series of Solutions. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 39(6):1209–1222, 2017.

[Xu 18] Y. Xu, Z. Piao et S. Gao. *Encoding crowd interaction with deep neural network for pedestrian trajectory prediction.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 5275–5284, 2018.

[Xue 19] T. Xue, A. Owens, D. Scharstein, M. Goesele et R. Szeliski. *Multi-frame stereo matching with edges, planes, and superpixels.* Dans Image and Vision Computing, IVC, volume 91, 2019.

[Y. J. Zhang 01] Y. J. Zhang. *A review of recent evaluation methods for image segmentation.* Dans International Symposium on Signal Processing and its Applications, volume 1, pages 148–151, 2001.

[Y. Xiang 14] R. Mottaghi Y. Xiang et S. Savarese. *Beyond PASCAL: A Benchmark for 3D Object Detection in the Wild.* Dans IEEE Winter Conference on Applications of Computer Vision, WACV, 2014.

[Yalcin 15] Hulya Yalcin. *Vision based automatic inspection of insects in pheromone traps.* Dans IEEE International Conference on on Agro-Geoinformatics, pages 333–338, 2015.

[Yan 06] P. Yan et A. A. Kassim. *Medical Image Segmentation Using Minimal Path Deformable Models With Implicit Shape Priors. IEEE Transactions on Information Technology in Biomedicine, TITB*, 10(4):677–684, 2006.

[Yang 08a] Q. Yang, C. Engels et A. Akbarzadeh. *Near real-time stereo for weakly-textured scenes.* Dans British Machine Vision Conference, BMVC, 2008.

[Yang 08b] W. Yang, D. Yi, Z. Lei, J. Sang et S. Z. Li. *2D-3D face matching using CCA.* Dans IEEE International Conference on Automatic Face Gesture Recognition, FG, pages 1–6, 2008.

[Yang 09] Q. Yang, L. Wang, R. Yang, H. Stewénius et D. Nistér. *Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 31(3):492–504, 2009.

[Yang 10] M. Y. Yang et W. Förstner. *Plane Detection in Point Cloud Data. International Conference on Machine Control Guidance*, pages 95–104, 2010.

[Yang 14] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi et S. Z. Li. *Salient color names for person re-identification.* Dans European Conference on Computer Vision, ECCV, 2014.

[Yang 18] G. Yang, H. Zhao, J. Shi, Z. Deng et J. Jia. *SegStereo: Exploiting semantic information for disparity estimation.* Dans European Conference on Computer Vision, ECCV, pages 636–651, 2018.

[Yao 18] Y. Yao, S. Li, T. Fang et L. Quan. *MVSNet: Depth Inference for Unstructured Multi-view Stereo.* Dans European Conference on Computer Vision, ECCV, pages 3–20, 2018.

[Yao 19] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang et L. Quan. *Recurrent MVSNet for High-resolution Multi-view Stereo Depth Inference.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, 2019.

[Yi 16] K. M. Yi, E. Trulls, V. Lepetit et P. Fua. *LIFT: Learned invariant feature transform.* Dans European Conference on Computer Vision, ECCV, 2016.

[Yoon 05] K.-J. Yoon et I.-S. Kweon. *Locally Adaptive Support-Weight Approach for Visual Correspondence Search.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, volume 2, pages 924–931, 2005.

[Yoon 06] K.-J. Yoon et I. S. Kweon. *Adaptive support-weight approach for correspondence search. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 28(4):650–656, 2006.

[Yoon 07] K. Yoon et I. Kweon. *Stereo matching with the distinctive similarity measure.* Dans International Conference on Computer Vision, ICCV, 2007.

[Yoshitake 18] H. Yoshitake et M. Shino. *Risk assessment based on driving behavior for preventing collisions with pedestrians when making across-traffic turns at intersections. IATSS research*, 42(4):240–247, 2018.

[Yu 11] G. Yu et J.-M. Morel. *ASIFT: An Algorithm for Fully Affine Invariant Comparison. Image Processing On Line*, 2011.

[Zabih 94] R. Zabih et J. Woodfill. *Non-parametric Local Transforms for Computing Visual Correspondence.* Dans European Conference on Computer Vision, ECCV, pages 151–158, 1994.

[Zaharescu 09] A. Zaharescu, E. Boyer, K. Varanasi et R. Horaud. *Surface feature detection and description with applications to mesh matching.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 373–380, 2009.

[Zbontar 15] J. Zbontar et Y. LeCun. *Computing the stereo matching cost with a convolutional neural network.* Dans IEEE Conference on Computer Vision Pattern Recognition, CVPR, pages 1592–1599, 2015.

[Zhang 96] Y.J. Zhang. *A survey on evaluation methods for image segmentation. Pattern Recognition, PR*, 29(8):1335–1346, 1996.

[Zhang 06] Z. Zhang, K. Huang et T. Tan. *Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes.* Dans International Conference on Pattern Recognition, ICPR, volume 3, pages 1135–1138, 2006.

[Zhang 09] K. Zhang, J. Lu et G. Lafruit. *Cross-Based Local Stereo Matching Using Orthogonal Integral Images. IEEE Transactions on Circuits and Systems for Video Technology, CSVT*, 19(7):1073–1079, 2009.

[Zhang 10] H. Zhang, J. Xiao et L. Quan. *Supervised label transfer for semantic segmentation of street scenes*. Dans European Conference on Computer Vision, ECCV, pages 561–574, 2010.

[Zhang 14] Q. Zhang, X. Shen, L. Xu et J. Jia. *Rolling Guidance Filter*. Dans European Conference on Computer Vision, ECCV, pages 815–830, 2014.

[Zhang 16] L. Zhang, F. Yang, Y. D. Zhang et Y. J. Zhu. *Road crack detection using deep convolutional neural network*. Dans International Conference on Image Processing, ICIP, pages 3708–3712, 2016.

[Zhang 19] J. Zhang, D. Sun, Z. Luo, A. Yao, L. Zhou, T. Shen, Y. Chen, L. Quan et H. Liao. *Learning Two-View Correspondences and Geometry Using Order-Aware Network*. Dans International Conference on Computer Vision, ICCV, 2019.

[Zhou 05] J. Zhou, P. Huang et F.-P. Chiang. *Wavelet-based pavement distress classification. Journal of the Transportation Research Board, TRB Journal*, 1940:89–98, 2005.

[Zhou 15] X. Zhou, M. Zhu et K. Daniilidis. *Multi-image matching via fast alternating minimization*. Dans International Conference on Computer Vision, ICCV, pages 4032–4040, 2015.

[Zhou 18] L. Zhou, S. Zhu, Z. Luo, T. Shen, R. Zhang, M. Zhen, T. Fang et L. Quan. *Learning and Matching Multi-View Descriptors for Registration of Point Clouds*. Dans European Conference on Computer Vision, ECCV, pages 527–544, 2018.

[Zhuge 04] H. Zhuge. *Retrieve images by understanding semantic links and clustering image fragments. Journal of Systems and Software*, 73(3):455–466, novembre 2004.

[Zhuo 11] S. Zhuo et T. Sim. *Defocus map estimation from a single image. Pattern Recognition, PR*, 44(9):1852–1858, 2011.

[Zijdenbos 94] A. Zijdenbos, B. M. Dawant, R. A. Margolin et A. C. Palmer. *Morphometric Analysis of White Matter Lesions in MR Images: Method and Validation. IEEE Transactions on Medical Imaging, TMI*, 13(4):716–724, 1994.

[Zitovà 03] B. Zitovà et J. Flusser. *Image Registration Methods: A Survey. Image and Vision Computing, IVC*, 21:977–1000, 2003.

[Zordan 06] V. B. Zordan, B. Celly, B. Chiu et P. C. DiLorenzo. *Breathe Easy: Model and Control of Human Respiration for Computer Animation. Graphical Models, GMOD*, 68(2):113–132, 2006.

[Zou 12] Q. Zou, Y. Cao, Q. Li, Q. Mao et S. Wang. *CrackTree: Automatic Crack Detection from Pavement Images. Pattern Recognition Letters, PRL*, 33(3), 2012.

[Zou 18] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang et S. Wang. *Deepcrack: Learning hierarchical convolutional features for crack detection. IEEE Transactions on Image Processing, TIP*, 28(3):1498–1512, 2018.

Annexe A.

Annexe relative à la détection de primitives en 2D et en 3D

A.1. Détecteurs de points d'intérêt en 2D : les détails

Nous avons choisi de présenter les approches en les séparant en trois catégories : opérateur du premier ordre, opérateur basé région et opérateur du second ordre. De plus, pour chaque catégorie, nous indiquons les approches multi-échelle. C'est pourquoi, dans un premier temps, nous présentons la manière la plus classique de construire une pyramide d'images.

A.1.1. Construction d'une pyramide d'images

Le problème consiste à produire à partir d'une image, un ensemble d'images de résolution de plus en plus faibles. Plusieurs techniques ont été développées pour calculer des pyramides d'images, principalement, dans le domaine de la compression d'images [Burt 83]. Il existe différents types de pyramides :

- *Les pyramides régulières* – Ce sont les plus couramment employées. Le même voisinage est considéré dans la phase de réduction de la résolution des images.
- *Les pyramides irrégulières* – Elles sont utilisées de manière beaucoup plus marginale [Montanvert 91]. Nous pouvons distinguer les pyramides stochastiques exposées dans [Jolion 03], des pyramides centrées proposées dans [Brigger 99].

Les pyramides irrégulières ne sont pas utilisées dans le cadre du suivi. Elles ont souvent été proposées pour effectuer une segmentation en régions, cf. § 8. Ces pyramides semblent, en effet, mieux respecter les différentes structures de l'image et ne pas supprimer les détails pertinents. Dans la suite, nous ne détaillerons donc que les pyramides régulières.

Si nous notons n la résolution courante à calculer (qui est plus faible que la résolution $n-1$), la construction d'une pyramide régulière d'images en chaque point (i,j) est définie par :

$$I^n(i,j) = \mathrm{red}_{(i',j')\in \mathrm{ZV}(i,j)} f_r(I^{n-1}(i',j')), \tag{A.1}$$

sachant que $\mathrm{ZV}(i,j)$ correspond au voisinage de réduction (c'est une fonction qui permet de retourner l'ensemble des pixels considérés au niveau $n-1$ pour évaluer la valeur d'un pixel au niveau n), f_r est la fonction qu'il faut appliquer aux pixels pris en compte (le plus souvent il s'agit d'une pondération) et l'opérateur red est une fusion des valeurs des pixels voisins du niveau $n+1$ pour obtenir le niveau n.

Les pyramides régulières les plus courantes et les plus utilisées dans le cadre de la mise en correspondance sont :

- *Les pyramides gaussiennes* – Il s'agit d'effectuer un lissage par un filtre gaussien, puis un sous-échantillonnage, en récupérant un pixel sur quatre.
- *Les pyramides moyennes* – Une moyenne sur un voisinage 2×2 est calculée.

A.1.2. Détecteurs du premier ordre

Il y a peu de détecteurs dans cette famille et leur principe est simple : il s'agit d'estimer à quel point **un point étudié est corrélé à ses voisins**. Cette famille est lié au détecteur le plus ancien et le plus célèbre : le détecteur de [Harris 88]. Par la suite, une variante multi-échelle a été introduite et nous la présentons également.

Auto-corrélation [Harris 88]

Principe d'un détecteur basé auto-corrélation [Moravec 80] – Si (u, v) est un déplacement quelconque alors le détecteur de Moravec revient à évaluer l'impact de ce déplacement en (x, y) de la façon suivante :

$$\mathrm{E}_{(u,v)}(x, y) = \sum_{x,y} w(x, y)(I(x + u, y + v) - I(x, y))^2 \tag{A.2}$$

où :

- w correspond à un poids (0 si le point n'est pas considéré et 1 s'il est pris en compte, ce qui signifie qu'ici le voisinage est en fait toute l'image) ;
- Ainsi, nous pouvons conclure que $\mathrm{E}_{(u,v)}(x, y)$ évalue la **moyenne du changement** si la fenêtre est déplacée.

Ainsi, en utilisant cette réponse E, nous pouvons considérer les trois cas suivants, cf. figure A.1 :

(1) Dans une zone homogène, la réponse donnée par E sera faible dans toutes les directions.

(2) Dans une zone sur un contour rectiligne, les termes de E, le long du contour seront faibles, alors que perpendiculairement au contour ce sera élevé.

(3) Dans une zone de coins, voire d'un point isolé, la plupart des termes de E seront élevés et donneront donc une réponse globale élevée.

En conséquence, le principe du détecteur de Moravec est de donner comme réponse la **valeur minimale obtenue par Mor parmi toutes les directions testées** :

$$\mathrm{MOR}(x, y) = \min_{(u,v)} \mathrm{E}_{(u,v)}(x, y)$$

Améliorations proposées par Harris et Stephen – Le détecteur de Moravec présente plusieurs défauts. En particulier, Harris et Stephen ont identifié des erreurs/des problèmes et ils ont également proposé de les corriger :

(1) **La réponse est bruitée en raison des poids binaires**.

Harris Pour cela, les auteurs proposent l'utilisation de poids Gaussiens, de moyennes nulles, donc :

$$w(x, y) = e^{(-\frac{x^2+y^2}{2\sigma^2})} \tag{A.3}$$

(2) **La réponse du détecteur est anisotrope (dépendant de la direction) du fait qu'un pas discret de 45 degrés est utilisé.**

ET

(3) **La réponse est trop forte/ambiguë au niveau des contours puisque finalement, une seule valeur est conservée**.

Harris Pour répondre à ces deux problèmes/difficultés, les auteurs ont proposé d'utiliser un **développement de Taylor à l'ordre 2** au voisinage de (x, y) :

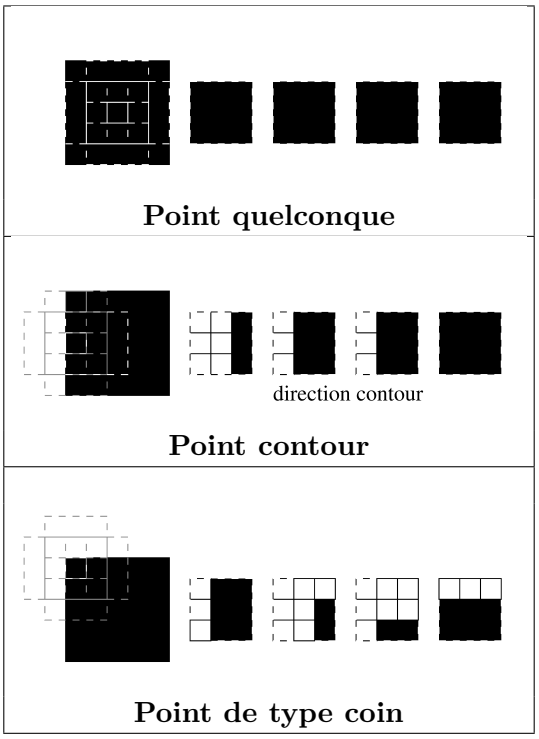


FIGURE A.1. – Comportement de l'auto-corrélation suivant le type de points – Pour chaque cas, nous présentons à gauche un point et son voisinage (en binaire), puis à droite nous développons les voisinages de ces voisins 4-connexes.

Nous pouvons écrire :

$$I(x+u, y+v) = I(x,y) + u\frac{\partial I}{\partial u} + v\frac{\partial I}{\partial v} + o(u^2, v^2). \tag{A.4}$$

D'où, si nous remplaçons dans l'équation (A.2) :

$$\mathrm{E}^{\text{modifiée}}_{(u,v)}(x,y) = \sum_{x,y} w(x,y)(u\frac{\partial I}{\partial u} + v\frac{\partial I}{\partial v} + o(u^2, v^2))^2. \tag{A.5}$$

En négligeant $o(u^2, v^2)$, ce qui est valide dans le cadre de **petits déplacements**, nous obtenons :

$$\mathrm{E}_{\text{modifiée}} = Au^2 + 2Cuv + Bv^2, \tag{A.6}$$

avec $A = \left(\frac{\partial I}{\partial u}\right)^2 \otimes w$, $B = \left(\frac{\partial I}{\partial v}\right)^2 \otimes w$ et $C = \frac{\partial I}{\partial u}\frac{\partial I}{\partial v} \otimes w$ où $\otimes$ représente la convolution.

Harris Ainsi, grâce à cette nouvelle façon d'écrire, il est possible de prendre en compte tout le comportement local à partir de E modifiée. En repartant de l'équation (A.5) ré-écrite de la manière suivante :

$$\mathrm{E}^{\text{modifiée}}_{(u,v)}(x,y) = (u,v)\dot{M}\dot{(}u,v)^T, \tag{A.7}$$

avec $M = \begin{pmatrix} A & C \\ C & B \end{pmatrix}$. **Les valeurs propres de la matrice M correspondent aux courbures principales associées à E**, ainsi :

- Si les deux courbures sont faibles, la région a une intensité constante.
- Si une des courbures est forte alors que l'autre est faible, alors la région se trouve sur un

contour.
- Si les deux courbures sont fortes alors l'intensité varie fortement dans toutes les directions et il s'agit d'un coin.

Définition de la réponse de Harris et Stephen – Ainsi, la réponse du détecteur est définie par :

$$R(x, y) = \mathrm{Det}(M) - k\mathrm{Trace}(M)^2, \tag{A.8}$$

avec $\mathrm{Det}(M) = AB - C^2$ et $\mathrm{Trace}(M) = A + B$. Avec cette formulation, **l'approche s'est affranchie du choix de la direction**. Les auteurs préconisent pour k une valeur entre 0.04 et 0.06.

Ainsi, l'analyse de R permet de conclure que les valeurs de R sont :
- positives au voisinage d'un coin,
- négatives au voisinage d'un contour et
- faibles dans une région d'intensité constante.

A.1.3. Vers le multi-échelle : Harris-Laplace [Mikolajczyk 04]

La version multi-échelle de Harris proposé dans [Mikolajczyk 04] s'appuit sur les deux étapes suivantes :

(1) **Calcul de la réponse de Harris à plusieurs échelles** – Les échelles sont définies par :

$$\sigma_k = c^k \sigma_0,$$

avec c le facteur d'échelle. La manière de fixer k dépend de l'échelle d'intégration (lissage d'un niveau d'échelle à un autre) et de l'échelle de différentiation (lissage pour calculer les dérivées). **Les extrema locaux de la réponse sont estimés en (i, j, σ). Ils vont être utilisés à l'étape suivante.**

(2) **Affinement de la sélection des points d'intérêt de coordonnées (i, j) dans l'espace échelle du laplacien** – Pour cette étape, un algorithme itératif est appliqué **pour chaque candidat retenu à la première étape**. Soient (i_0, j_0, σ_0), les coordonnées et l'échelle initiale du candidat considéré :

(a) Recherche d'un extremum pour le candidat dans l'espace échelle du laplacien pour les échelles comprises dans $[c^{-1}\sigma_k; c\sigma_k]$, avec intervalle inter-échelle déterminé par $c_2 < c$. Si un extremum n'est pas trouvé alors ce candidat est rejeté sinon σ_u, l'échelle à laquelle cet extremum a été trouvé, est conservée.

(b) Détection de la position (i_{k+1}, j_{k+1}) du maximum de la réponse de Harris le plus proche de (i_k, j_k) à l'échelle σ_u.

(c) Nouvelle itération si $\sigma_k \neq \sigma_{k+1}$ ou $(i_k, j_k) \neq (i_{k+1}, j_{k+1})$.

La sortie de cet algorithme sera donc les points d'intérêt détectés en prenant en compte l'échelle. Une version moins coûteuse est possible en étant plus sélectif à la fin de la première étape en appliquant un seuillage.

A.1.4. Détecteurs de régions

Pour ce type de détecteur, le principe commun est de considérer non plus un point d'intérêt comme un point qui se distingue des autres, mais plutôt, comme le centre d'une région qui se distingue des autres. Au final, la réponse renvoyée est un point, mais, la notion rapportée est plutôt celle d'une région. Ce n'est pas mentionné clairement dans la littérature, mais, finalement, la région caractérisant le point peut être vue comme le descripteur associé.

SUSAN, *Smallest Univalue Segment Assimilating Nucleus* [Smith 97]

Définition de la réponse du détecteur – Pour le détecteur de SUSAN, *Smallest Univalue Segment Assimilating Nucleus*, les aspects à retenir sont les suivants :

- Les auteurs appellent *nucleus*/noyau (en français) le pixel étudié, c'est-à-dire le pixel central $\mathbf{p}$.
- Une région d'intérêt autour du noyau (à rapprocher de la notion de voisinage) est utilisée, la particularité étant de considérer un **masque circulaire**, cf. figure A.2. Habituellement, la région considérée contient 36 pixels, c'est-à-dire une fenêtre 5×5 complétée par 3 pixels sur chaque côté.

$$\begin{array}{ccccccc} - & - & * & * & * & - & - \\ - & * & * & * & * & * & - \\ * & * & * & * & * & * & * \\ * & * & * & o & * & * & * \\ * & * & * & * & * & * & * \\ - & * & * & * & * & * & - \\ - & - & * & * & * & - & - \end{array}$$

FIGURE A.2. – Masque circulaire utilisé par le détecteur SUSAN – o : pixel étudié, $*$: pixel pris en compte dans le voisinage, $-$: pixel non pris en compte dans le voisinage. Il s'agit de la discrétisation d'un cercle de Bresenham de noyau 3.

La réponse du détecteur s'appuie sur l'utilisation d'une surface particulière : l'USAN. C'est une sous-région contenue dans le voisinage et qui correspond à tous les points proches d'un point de vue photométrique du noyau. Plus précisément, la surface de l'USAN centrée sur le pixel de coordonnées $\mathbf{p}$ est donnée par :

$$\mathbf{p} \mapsto USAN(\mathbf{p}) = \sum_{\mathbf{p}' \in \mathcal{V}(\mathbf{p})} e^{-\left(\frac{I(\mathbf{p}) - I(\mathbf{p}')}{T_u}\right)^6}, \tag{A.9}$$

où T_u **correspond au contraste maximal autorisé**. Ce qui signifie qu'au delà d'une différence T_u de niveau de gris, la fonction $USAN$ aura une valeur faible ce qui sera pénalisant pour cette fonction de similarité, cf. figure A.3. La détermination de la zone USAN va donc s'appuyer sur la somme des différences des niveaux de gris dans son voisinage.

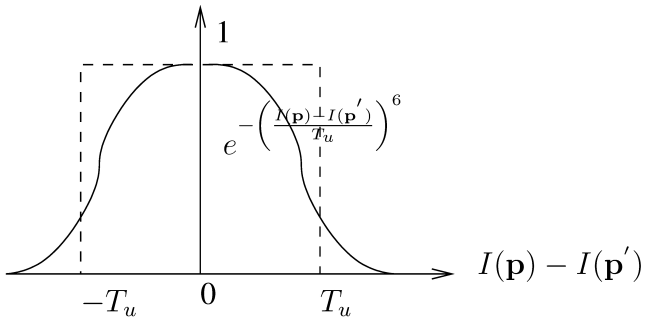


FIGURE A.3. – SUSAN : Comportement de la fonction de similarité.

Au final, la réponse de ce détecteur est donnée par :

$$SU(\mathbf{p}) = \begin{cases} T_g - USAN(\mathbf{p}) & \text{si } f(\mathbf{p}) < T_g \\ 0 & \text{sinon.} \end{cases} \tag{A.10}$$

T_g est un seuil géométrique.

Cette définition induit l'analyse suivante de la réponse :

- Si la surface de l'USAN est proche de la surface du voisinage alors le pixel est situé dans une zone homogène.
- Si la surface de l'USAN est proche de la moitié de la surface du voisinage alors le pixel est situé sur un contour.
- Si la surface de l'USAN est plus petite que la moitié de la surface du voisinage alors le pixel est un coin et c'est un point d'intérêt candidat.

Suivant les valeurs prises pour le seuil, il est possible de moduler le type de points (contours/coins) détectés. Habituellement, SUSAN est utilisé pour détecter des coins et généralement T_g est choisi tel que $T_g = \frac{s_{max}}{2}$ où s_{max} est la surface maximale de l'USAN étant donnée la taille du voisinage considéré.

Le détecteur peut également faire intervenir des post-traitements afin d'améliorer la qualité des résultats. Certains candidats peuvent être rejetés à l'aide de ces **deux contraintes** :

(1) **Contrainte du centre** – Le centre de l'USAN doit être éloigné d'une distance d (en pixels) du noyau. Les coordonnées du centre de l'USAN sont données par le barycentre de l'USAN :

$$x_c/y_c = \frac{\sum_{\mathbf{p}' \in \text{USAN}(\mathbf{p})} (x/y' - x/y) f(\mathbf{p}')}{\sum_{\mathbf{p}' \in \text{USAN}(\mathbf{p})} f(\mathbf{p}')}$$

Et il faut donc vérifier que $\sqrt{(x_c - x)^2 + (y_c - y)^2} > d$.

(2) **Contrainte de continuité** – Tous les points sur le segment reliant le noyau et le centre doivent appartenir à l'USAN.

Plus nous ajoutons de contraintes, plus nous éliminons de "mauvais" points d'intérêt mais plus nous risquons d'éliminer de "bons" points d'intérêt.

FAST, *Features from Accelerated Segment Test* [Rosten 06]

Ce détecteur possède des éléments en commun avec celui de SUSAN présenté au paragraphe précédent. En effet, le principe de ce détecteur est d'étudier une zone circulaire et d'évaluer les ensembles de pixels de même intensité qui forment une région connexe. Plus précisément, les auteurs utilisent une zone circulaire respectant l'algorithme du cercle de Bresenham de rayon 3. Parmi les 16 points sur le périmètre de ce cercle, ils vérifient si une de ces hypothèses est vraie:

1. Il existe $N_p = 12$ pixels connexes tels que $I(x', y') < I(x, y) - T$.
2. Il existe $N_p = 12$ pixels connexes tels que $I(x', y') > I(x, y) + T$.

Ce test implique le choix d'un paramètre, le seuil T. Pour simplifier l'approche et la rendre encore plus rapide, il est proposé de simplifier encore plus ce test en ne considérant que 4 des 16 pixels qui composent le contour, les pixels au position 1, 5, 9 et 13. Si 3 des 4 pixels vérifient une de ces deux hypothèses, alors, le pixel à la position (x, y) est considéré comme un point d'intérêt.

Cette approche assez simple possède les inconvénients suivants : il est difficile d'adapter cet algorithme lorsque $N_p < 12$, le choix des N_p pixels connexes implique des hypothèses fortes sur l'apparence des points d'intérêt, de nombreuses primitives sont détectées de manière adjacente. Pour toutes ces raisons, les auteurs ont proposé d'ajouter une phase d'apprentissage.

MSER, *Maximally Stable Extremal Regions* [Matas 02]

Ce détecteur repose sur l'idée que si nous seuillons progressivement une image, nous allons voir apparaître des points puis des régions qui vont ensuite se rejoindre pour fusionner ou déterminer les frontières entre régions d'intérêt. Ce principe est comparable à celui de la ligne de partage des eaux qui est d'ailleurs l'algorithme utilisé. L'ensemble de toutes les composantes connexes obtenues, à chaque itération, forme l'ensemble des régions maximales. Parmi ces régions, certaines sont considérées comme des maxima/minima car tous les points à la frontière de la région, mais n'appartenant pas à la région, ont une intensité plus grande/petite que les points appartenant à la région. De plus, parmi toutes les régions maximales, certaines vont rester stables sur plusieurs itérations. Elle constituent alors ce que les auteurs appellent les régions maximales stables. Les régions d'intérêt, dans ce contexte, sont donc, les régions maximales extrema et stables.

Ces régions sont invariantes aux transformations monotones d'intensité (et non uniquement aux transformations affines photométriques) et aux transformations géométriques homographiques ou non linéaires mais continues. L'inconvénient de ce détecteur est sa sensibilité au flou, car dans ce cas l'intensité sur les frontières est mal définie et les contours des régions sont également mal localisés.

IBR, *Intensity Based Regions* [Tuytelaars 04]

Voici les étapes suivies par ce détecteur :

1. Détection des extremums locaux d'intensité de l'image I ;
2. Analyse des profils d'intensité le longs des rayons partants de ces extremums locaux ;
3. Sélection des extremums sur chaque rayon;
4. Connexion de tous les extremums sur chaque rayon pour former une région fermée autour du point étudié ;
5. Approximation de cette région par une ellipse dont la taille est doublée pour définir la région d'intérêt finale.

Pour l'étape (2), plus précisément, les auteurs souhaitent explorer/analyser la région autour de chacun des extremums en étudiant une fonction de l'intensité le long des rayons issus de l'extremum local étudié. Pour chaque rayon [1], la fonction suivante est évaluée :

$$f_I(t) = \frac{|I(t) - I(0)|}{\max(d, \frac{1}{t}\int_0^t |I(x) - I(0)|dx}, \tag{A.11}$$

où t est l'abscisse curviligne le long du rayon et d est un nombre de valeur faible permettant d'éviter les divisions par 0. Le dénominateur permet, d'après les auteurs, d'assurer une bonne localisation des extremum.

Cet opérateur détecte des *blobs*. Dans la littérature, il s'agit d'une zone de l'image autour d'un point (d'intérêt) où les valeurs des pixels se distinguent des valeurs des pixels du voisinage. Par exemple, une zone d'une image qui représente une tache claire située dans une zone sombre constitue un *blob*. **Cette notion est assez proche de la notion de noyau abordée pour le détecteur de SUSAN, cf. § A.1.4.** Cet opérateur est invariant par transformations affines, géométriques et photométriques.

Vers le multi-échelle : Détecteur de [Kadir 04]

Cette approche s'appuie sur l'entropie de la distribution des valeurs des pixels au voisinage du point considéré, c'est-à-dire sur la « quantité » de désordre dans le voisinage du pixel considéré. Cela signifie que nous supposons que plus le voisinage est désordonné plus il est discriminant par rapport

1. À partir du point étudié, la zone circulaire autour de lui est découpée en secteur angulaire et nous prenons les rayons correspondant aux frontières de ces secteurs angulaires.

aux autres points de l'image et, au contraire, plus la distribution est uniforme, plus le point étudié est probablement situé dans une zone homogène et a donc peu d'intérêt. L'entropie est donnée par :

$$\begin{aligned} \mathcal{H} : \mathbb{N}^2 &\to \mathbb{R} \\ \mathbf{p} &\mapsto -\sum_{k,l} p\left(\mathbf{p}\right) \log\left(p\left(\mathbf{p}\right)\right) \end{aligned} \tag{A.12}$$

avec $(k,l) \in \mathbb{N}^2$ et $(k-i)^2 + (l-j)^2 < r^2$ où r est le rayon du voisinage considéré et où la probabilité $p(\mathbf{p})$ est donnée par la distribution des valeurs des pixels du voisinage. En pratique, un histogramme des valeurs des pixels composé de w intervalles est utilisé pour estimer p et la réponse suivante est calculée pour chaque candidat :

$$\mathrm{KA}\left(\mathbf{p}\right) = \mathcal{H}\left(\mathbf{p}\right) \frac{\sigma^2}{2\sigma - 1} \sum_{k,l} \left| \frac{\partial p\left(\mathbf{p}\right)}{\partial \sigma} \right|. \tag{A.13}$$

Dans [Kadir 04], il est montré comment cette formulation leur permet d'obtenir un détecteur invariant aux transformations photométriques (biais), aux transformations géométriques affines, ce qui rend ce détecteur indépendant du point de vue.

Vers le multi-échelle : EBR, *Edge-Based Regions* [Tuytelaars 04]

Les contours sont des objets assez stables par rapport aux changements de point de vue et de conditions d'éclairage. Ces détecteurs exploitent ainsi la géométrie locale : les droites et les parallélogrammes pour les *Edge-Based Regions* ou encore la courbure pour les *Curvature Scale Space*, abordé dans la partie multi-échelle, cf. § A.1.6. Voici les quatre étapes suivies pour réaliser cette détection :

1. Coins localisés grâce au détecteur de Harris ;
2. Contours détectés avec l'approche de Canny-Deriche [Canny 86, Deriche 87] ;
3. Estimation des paramètres affines invariant le long des contours avec construction d'une famille de parallélogrammes ;
4. Sélection des parallélogrammes extrema suivant une fonction.

La principale contribution réside dans les deux dernières étapes. Plus précisément, voici le principe suivi pour l'étape (3) : Si un coin p est considéré, ainsi que deux points $p_1(l_1)$ et $p_2(l_2)$ se déplaçant le long des contours partant de ce coin, alors, la distance curviligne $l_{1,2}$ de chaque point $p_{1,2}$ au coin p est définie par la surface entre le contour et la droite joignant le coin et le point $p_{1,2}$. Cette distance est invariante aux transformations isométriques. Ces deux points et le coin définissent alors une famille de parallélogrammes. Enfin, pour la dernière étape, parmi tous les parallélogrammes qui peuvent définir le coin, celui retenu est celui qui présente une valeur extremum invariante d'une fonction photométrique basée sur les moments du premier ordre de la fonction intensité. Plus précisément, les auteurs cherchent à minimiser la distance entre le centre de gravité et les diagonales du parallélogramme. Quelque part, cette façon d'agir est déjà un premier pas vers une réponse invariante à l'échelle. Cet opérateur permet de retrouver des formes parallélépipédiques, invariantes aux transformations affines géométriques et photométriques. Pour ces détecteurs, le type des primitives recherchées les destinent plutôt à des scènes structurées, comportant des éléments artificiels et bien contrastés où les frontières sont bien identifiables.

Vers le multi-échelle : PCBR, *Principal Curvature-Based Regions* [Deng 07]

Ce détecteur a été introduit afin de proposer un opérateur pour la reconnaissance de classe d'objets. Son principe est de détecter des lignes de partages des eaux, cf. section C.2.3, à partir des courbures principales estimées à différentes échelles. Nous avons placé ce détecteur à la fin de cette section sur

les détecteurs de régions, car, il est à la limite du classement que nous avons proposé, qui distingue les approches régions des approches du second ordre. Une propriété attendue par les auteurs pour ce détecteur est d'extraire des lignes de partage des eaux fiables. Pour cela, les auteurs proposent de post-traiter l'estimation des courbures principales en utilisant une fermeture morphologique combinée à un seuillage par hystérésis sur les vecteurs propres. De plus, les courbures principales ne sont conservées que si elles sont maximales sur plusieurs échelles consécutives. La façon de calculer la courbure principale, commune à de nombreuses publications, est expliquée dans la section A.1.5 et largement détaillée dans le chapitre 4 puisqu'elle est au cœur de notre contribution.

Ainsi, nous avons introduit des approches prenant en compte le voisinage afin de déterminer les points les moins corrélés à leurs voisins, cf. section A.1.2, puis, dans cette section, les approches qui à l'inverse, caractérise un point comme le centre d'une région cohérente. À présent, nous abordons la dernière famille de détecteurs qui prend en compte de opérateurs du second ordre. L'objectif est de prendre en considération la notion de courbure.

A.1.5. Détecteurs du second ordre

La plupart des détecteurs présentés dans cette dernière section s'appuie sur la notion de courbure, à l'exception de SIFT, *Scale Invariant Feature Transform*, que nous présentons en dernier. La première catégorie repose sur le fait que nous cherchons à déterminer les points saillants en terme de courbure, c'est-à-dire qui présentent un brusque changement de courbure, en supposant que cette caractéristique les rend discriminants par rapport aux autres points de l'image. Cette notion de courbure est très commune dans la littérature, notamment dans la recherche de primitives en 3D, aspects que nous abordons en détails dans le chapitre 3, partie I. L'intérêt de ce type d'approches est notamment de faire le lien avec la notion de points saillants d'une forme. Nous nous attachons donc plus ici à la notion de forme que de texture, comme c'est le cas avec les détecteurs basés auto-corrélation. La seconde catégorie, essentiellement liée au détecteur SIFT, est l'utilisation du gradient du second ordre à différentes échelles. Cela veut dire que nous nous intéressons aux points d'intérêt le long des contours (plutôt de type toit).

Beaudet ou Hessien [Beaudet 78]

Ce détecteur s'appuie sur les dérivées secondes et donc sur la notion de courbure de la surface. Le principe est de minimiser le déterminant de la matrice hessienne, soit :

$$\text{Beaudet}(x,y) = \det \begin{pmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{pmatrix} = I_{xx}I_{yy} - I_{xy}^2. \tag{A.14}$$

En effet, il représente la courbure de la surface constituée par les niveaux de gris (ou couleurs) de l'image pour chaque pixel étudié.

Vers le multi-échelle : Hessien-Laplace [Mikolajczyk 04]

La version multi-échelle du détecteur de Beaudet a également été mise en œuvre en utilisant exactement le même principe que celui exposé au § A.1.3 mais en l'appliquant au détecteur de Beaudet. Il s'agit du détecteur Hessien-Laplace.

A.1.6. Kitchen et Rosenfeld [Kitchen 82]

La réponse dépend également de la courbure, mais, cette fois, c'est la courbure du contour passant par le point considéré. De plus, chaque point utilisé est pondéré par la norme du gradient car les auteurs souhaitent prendre en compte les cas où la norme du gradient est faible (donc un contour peu

contrasté, qui peut correspondre à un bruit). Ainsi, la formulation suivante est obtenue :

$$\mathrm{KR}(x,y) = \frac{I_{xx}I_y^2 - 2I_{xy}I_xI_y + I_{yy}I_x^2}{I_x^2 + I_y^2}. \tag{A.15}$$

CSS, *Curvature Scale Space* [Mokhtarian 98]

Le détecteur en courbure et espace d'échelle, *Curvature Scale Space (CSS)* [Mokhtarian 98] recherche les extrema de la courbure le long des contours en utilisant une approche multi-échelle. Il détecte d'abord les contours en utilisant l'opérateur de Canny, fusionne les contours très proches, puis calcule la courbure $K(t,\sigma)$ de ce contour lissé par une gaussienne. Ceci définit l'espace d'échelle de courbure dont les coins sont assimilés à des extrema. Les candidats dont la réponse est trop faible ou trop proches les uns des autres ne sont pas retenus. Une variante de ce détecteur différencie le traitement des contours courts et des contours longs en déterminant automatiquement le seuil de décision. Cet opérateur est invariant par transformations affines géométriques et photométriques.

Le principal inconvénient des détecteurs basés sur les contours est la définition des paramètres utilisés par le détecteur de contours. Bien que tous les détecteurs soient paramétrés, la non-détection des contours s'avère être un handicap important.

Vers le multi-échelle : SIFT, Scale Invariant Feature Transform [Lowe 04]

Principe – La méthode SIFT, *Scale Invariant Feature Transform*, s'appuie sur la détection des extremums en échelle et en espace, notés (x,y,σ), du laplacien en (x,y,σ). Une fois ces extremums extraits, ils sont caractérisés par une orientation principale que nous définissons dans la figure A.8. Le descripteur associé à ce détecteur prend en compte un histogramme local de ces orientations, cf. section A.2.

Propriétés – Étant donné que chaque point détecté est caractérisé par l'orientation du gradient qui est invariante aux changements de contraste, cela permet une certaine **invariance aux changements de luminosité**. De plus, l'utilisation de l'histogramme local permet une certaine **robustesse aux changements de points de vue**. Toutefois, ce détecteur n'est pas invariant aux transformations affines.

Enfin, ce détecteur est complet dans le sens où les auteurs lui ont associé un descripteur permettant de décrire chacun des points et ainsi de les mettre en correspondance de manière robuste.

Étapes – Le détecteur proposé s'appuie sur les 4 grandes étapes suivantes :

(1) *Détection des extremums en échelle et en espace* : l'objectif est d'estimer un critère, ici le gradient de l'image, dans toutes les échelles et en tous points.

(2) *Localisation des points :* en s'appuyant sur un modèle, il s'agit de déterminer les points stables suivant ce modèle et ainsi d'en déduire les points d'intérêt.

(3) *Affectation d'une orientation :* une orientation principale est attribuée à chaque point.

(4) *Estimation des descripteurs :* cela est réalisé en prenant en compte un voisinage.

Les deux premières étapes correspondent à la détection des points d'intérêt alors que les deux dernières permettent le calcul d'un descripteur associé au point d'intérêt afin de réaliser la mise en correspondance ou le suivi, ils seront donc détaillés dans la section relative aux descripteurs A.2.

Détection des extremums en échelle et en espace – Il s'agit d'une approche par **filtres " en cascade "**. Les formes/points stables en échelles sont ceux qui sont détectés. Pour cela, un modèle/un noyau Gaussien est utilisé en suivant ces étapes :

(1) Les termes $L(x, y, \sigma)$ correspondent aux images filtrées à l'échelle σ et ainsi :

$$L(x, y, \sigma) = G(x, y, \sigma) \otimes I(x, y) \tag{A.16}$$

où $G(x, y, \sigma)$ est la fonction gaussienne.

(2) Pour détecter les points stables, les auteurs estiment par la suite des différences de gaussiennes de la manière suivante :

$$DoG(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma). \tag{A.17}$$

Cette différence de gaussienne n'est autre qu'une approximation du laplacien. En l'effectuant pour un ensemble de valeurs successives de σ, nous obtenons donc le laplacien normalisé en échelle. Un ensemble d'échelles correspond à une octave. Pour passer d'une octave à l'autre, la valeur de σ est généralement doublée. Ainsi, pour obtenir n échelles par octave, il faut fixer $k = 2^{1/n}$.

(3) Enfin, les maxima/minima locaux pour chaque octave sont sélectionnés en prenant en compte un voisinage 3×3 dans l'image ainsi que celui de l'échelle inférieure et de l'échelle supérieure.

Expérimentalement, les auteurs ont montré qu'il suffisait d'avoir 3 échelles par octave. En effet, il s'agit du meilleur compromis : coût/détection d'extremums locaux. De plus, il a été démontré que plus σ est élevée, plus les extremums retenus seront **répétables, c'est-à-dire que les mêmes points seront détectés quel que soit le changement entre les deux points de vue. De manière empirique, les auteurs ont choisi** $\sigma = 1.6$.

Pour résumer, pour obtenir n **échelles pour sélectionner les minima/maxima locaux, il est nécessaire de calculer** $n + 1$ **différences de gaussiennes, soit** $n + 2$ **images filtrées**. Sachant que la dernière image d'une octave et la première image de l'octave suivante ne peuvent pas être les mêmes, **il est donc nécessaire de calculer** $n + 3$ **images filtrées**, cf. figure A.4.

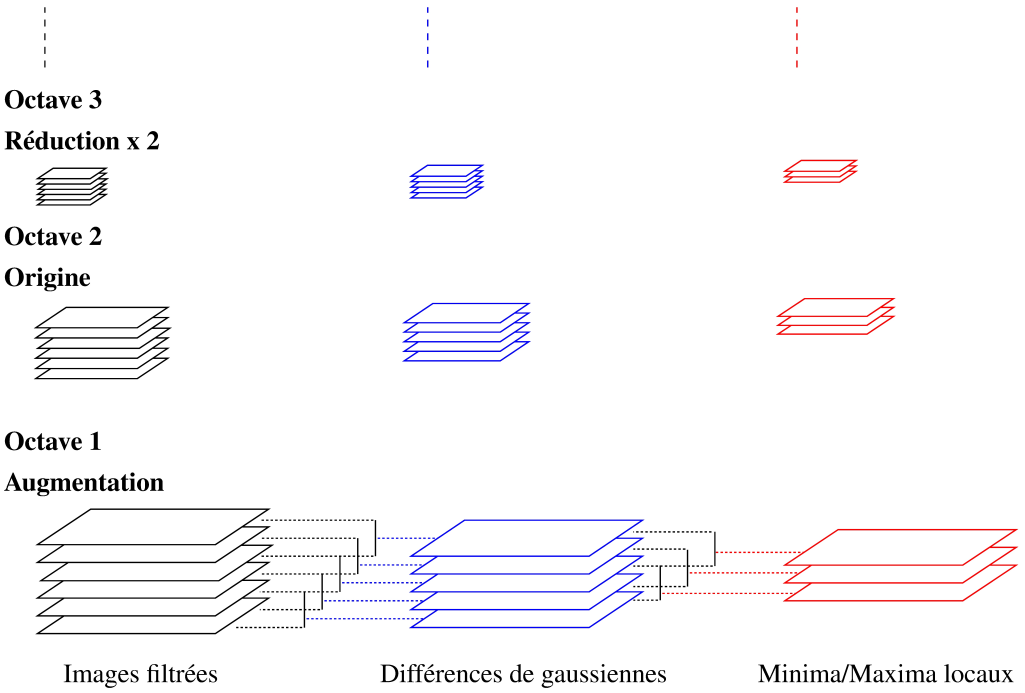


FIGURE A.4. – Octaves et échelles pour le détecteur SIFT – Pour cette exemple, nous supposons qu'il y a 3 échelles ainsi 6 images filtrées seront évaluées par octave.

Pour terminer, l'algorithme de calcul des minima/maxima locaux à travers les octaves et échelles est donc celui exposé dans la figure A.5.

```
(1) Calculer une pyramide gaussienne d'images
(2) Pour chaque octave o (chaque image de la pyramide) faire
       Pour chaque valeur de k faire
          Calculer en chaque point (x, y), G(x, y, kσ)
(3) Pour chaque octave o faire
       Pour chaque échelle s ∈ [2, n] faire
          Calculer en chaque point (x, y), DoG(x, y, σ)
(4) Pour chaque octave o faire
       Pour chaque échelle s ∈ [2, n − 1] faire
          Extraire les minima/maxima locaux
```

FIGURE A.5. – Algorithme de détection des points SIFT.

Pour le calcul de la pyramide d'images, pour contrecarrer les effets d'un éventuel bruit gaussien au départ avec un $\sigma = 0.5$ (tendance à supprimer les petits détails), les auteurs suggèrent de **doubler la résolution de l'image au départ**. Ensuite seulement, la résolution de l'image est réduite par 2 à chaque octave. Là aussi, le fait d'ajouter cette étape **permet de renforcer la répétabilité du détecteur**.

Position sub-pixellique – Pour contrecarrer les effets des lissages, par conséquent le fait que la précision sur la localisation des points peut être dégradée, les points extraits sont " relocalisés ", *i.e.,* leur position est calculée au sous-pixel. Pour cela, une approche s'appuyant sur le **développement de Taylor à l'ordre 2** est utilisée, soit :

$$D(\hat{\mathbf{p}}) = D(\mathbf{p}) + \frac{\partial D(\mathbf{p})^T}{\partial \mathbf{p}}\mathbf{p} + \frac{1}{2}\mathbf{p}^T\frac{\partial^2 D(\mathbf{p})}{\partial \mathbf{p}^2}\mathbf{p}, \tag{A.18}$$

La localisation précise de l'extremum est donnée par le passage par zéro de la dérivée suivant $\mathbf{p}$, soit :

$$\hat{\mathbf{p}} = -\frac{\partial^2 D(\mathbf{p})^{-1}}{\partial \mathbf{p}^2}\frac{\partial D(\mathbf{p})}{\partial \mathbf{p}}. \tag{A.19}$$

Si le décalage est plus grand que 0.5 **dans chaque direction, alors, il est supposé nécessaire de réaliser la relocalisation.**

Filtrage des minima/maxima locaux – Enfin, les auteurs ont ajouté deux contraintes pour la sélection des minima/maxima locaux :

(1) Une fois les points re-localisés, ceux dont le contraste est trop faible sont éliminés. Plus précisément, si l'équation (A.19) est introduite dans l'équation (A.18), nous obtenons :

$$D(\hat{\mathbf{p}}) = D(\mathbf{p}) + \frac{1}{2}\frac{\partial D(\mathbf{p})^T}{\partial \mathbf{p}}\hat{\mathbf{p}}. \tag{A.20}$$

Et dans leur expérimentations, les auteurs ont éliminé tous les points pour lesquels $|D(\hat{\mathbf{p}})| < 0.03$, en supposant une variation entre 0 et 1.

(2) Les points sur les contours sont rejetés. Pour cela, les auteurs utilisent la matrice hessienne, comme définie dans l'approche de Harris, cf. § A.1.2. La courbure principale dans la direction normale à un contour est forte et celle dans la direction du contour est faible. Les courbures principales étant proportionnelles aux valeurs propres de la matrice hessienne, seuls les points où le rapport entre la plus grande et la plus petite valeur propre est inférieur à un seuil r sont conservés. Plus formellement, il faut vérifier :

$$\frac{\text{Trace}(M)^2}{\text{Det}(M)} < \frac{(r+1)^2}{r}. \tag{A.21}$$

Dans leurs travaux, les auteurs ont pris $r = 10$.

Vers une variante accélérée de SIFT : SURF, Speeded Up Robust Features [Bay 08]

Ce détecteur utilise comme réponse une approximation du déterminant de la matrice hessienne à plusieurs échelles. L'intérêt de ce nouveau détecteur réside surtout dans le fait que les auteurs proposent une implémentation astucieuse. Plus précisément, c'est le calcul de cette approximation du Hessien qui le rend beaucoup plus rapide que les autres détecteurs de la littérature. Par exemple, SURF est moins coûteux en temps de calculs que SIFT. Le lissage gaussien est réalisé par des *box filter* et une image intégrale est alors utilisée pour calculer la convolution.

Les auteurs ont remarqué qu'en pratique, le calcul des dérivées avec un masque de convolution constitue déjà une approximation pour les autres approches de la littérature. Toutefois, ils remarquent également que cette perte de précision n'empêche pas les détecteurs basés laplacien de donner de bons résultats en terme de répétabilité (les mêmes points d'intérêt sont retrouvés d'un point de vue à l'autre) et proposent d'utiliser directement des approximations moins coûteuses en temps de calculs. Pour cela, ils introduisent l'utilisation de masques de dérivation au second ordre lissés par une gaussienne et tronqués de telle sorte que les calculs puissent être réalisés à l'aide d'images intégrales.

Image intégrale – Chaque point de l'image est la somme des niveaux de gris des points sur les lignes et colonnes précédentes. L'intérêt de ces images est de pouvoir faire par la suite des calculs très rapides. Plus formellement, ces images sont définies par :

$$I_{int}(i,j) = \sum_{k=0}^{k<i} \sum_{l=0}^{l<j} I(k,l). \tag{A.22}$$

Cette image permet de calculer facilement la somme de toutes les valeurs de I situées dans un rectangle délimité par trois opérations arithmétiques, cf. illustration dans la figure A.6 :

$$\sum_{i=a}^{b} \sum_{j=c}^{d} I(i,j) = I_{int}(b,d) - I_{int}(a,d) - I_{int}(b,c) + I_i nt(a,c). \tag{A.23}$$

Box filtering – Les masques utilisés pour calculer les dérivées sont construits de telle sorte qu'ils ne correspondent qu'à des combinaisons de sommes de valeurs de pixels situés dans différents rectangles, *i.e.*, les *box*. Ainsi, l'image intégrale peut être utilisée pour réduire le nombre d'opérations arithmétiques, cf. figure A.7.

Il existe une autre variante de ce type de détecteurs, permettant d'améliorer encore le temps de calculs, ils s'agit du détecteur BRISK, *Binary Robust Invariant Scalable Keypoints*, qui s'appuie sur des motifs binaires. Le lecteur intéressé pourra trouver les détails dans la publication suivante [Leutenegger 11].

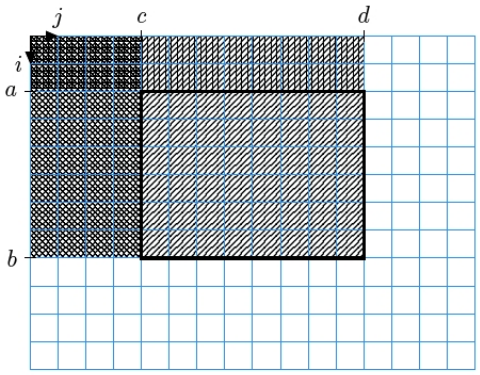


FIGURE A.6. – Utilisation de l'image intégrale.

A.2. Descripteurs de points d'intérêt en 2D

Jusqu'à présent, parmi tous les détecteurs que nous avons présentés, seulement deux correspondent à ce que nous appelons des détecteurs complets. Un détecteur complet fournit un opérateur de détection ainsi qu'un descripteur du point étudié, c'est le cas pour les détecteurs SIFT, cf. § A.1.6, et SURF, cf. § A.1.6. Dans cette section, nous présentons, d'une part, les descripteurs associés à ces détecteurs, et, d'autre part, d'autres outils qui permettent de décrire un point, sans fournir un détecteur.

Descripteur associé à SIFT et SURF

L'objectif du descripteur proposé pour effectuer la mise en correspondance est d'être suffisamment discriminant/unique pour permettre une mise en correspondance la plus fiable possible. C'est la raison pour laquelle les auteurs introduisent un descripteur de dimension 128 alors que la plupart des autres approches s'appuient sur des données 1D (niveaux de gris), 3D (couleurs) voire 5D (couleurs + position).

Affectation d'une orientation – Une fois les extremums détectés, les auteurs attribuent des orientations principales à chaque extremum en suivant les points suivants :

(1) *Estimation de la norme et de l'orientation du vecteur gradient* – Cela est calculé simplement avec :

$$\begin{aligned} m(x,y) &= \sqrt{(L(x+1,y)-L(x-1,y))^2+(L(x,y+1)-L(x,y-1))^2} \\ \theta(x,y) &= \tan^{-1}\left(\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right) \end{aligned} \tag{A.24}$$

(2) *Construction d'un histogramme* ***local*** *des orientations* θ – Pour le calcul de cet histogramme, la somme des éléments de même orientation est pondérée par la norme m et par une gaussienne, centrée sur le point étudié avec $\sigma = 1.5$. L'objectif de cette pondération est de minimiser l'influence des voisins les plus éloignés ainsi que l'influence des voisins dont le gradient est faible. En pratique, l'histogramme est discrétisé (les auteurs utilisent 36 orientations pour couvrir les 360 degrés), dans la figure A.8.(a), nous donnons un exemple d'histogramme sur 8 orientations uniquement. De plus, le voisinage pris en compte est de taille 8×8.

(3) *Sélection des orientations* – La plus grande valeur de cet histogramme, l'orientation principale, ainsi que toutes celles au dessus de 80% de ce maximum, les orientations secondaires sont considérées comme les orientations du point. Ainsi, nous pouvons avoir plusieurs orientations

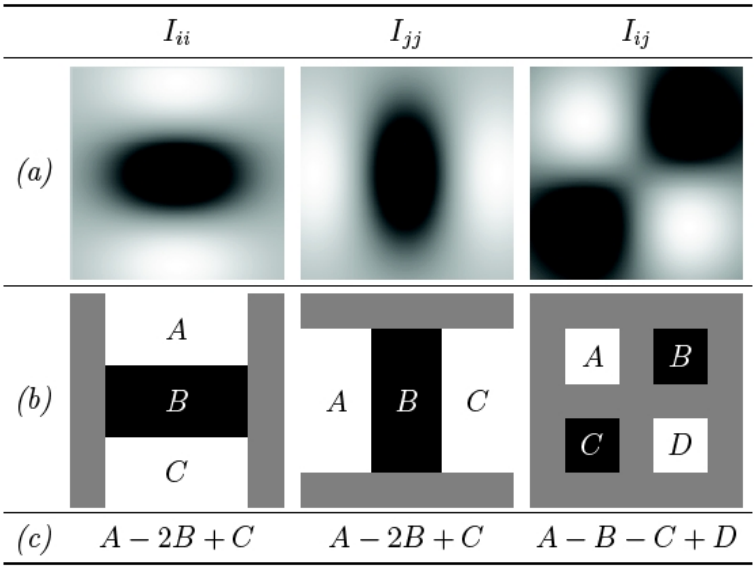


FIGURE A.7. – Illustration du *Box filtering* – La ligne (a) montre les images des masques de convolution pour calculer I_{ii}, I_{jj} et I_{ij} en un point donné. Plus le niveau de gris est élevé, plus la valeur du masque est élevée (ces valeurs pouvant être négatives, le gris représente 0). Les valeurs de ces masques sont approchées et tronquées pour obtenir les masques de la ligne (b). Ainsi, le calcul de la convolution peut être remplacé par l'opération indiquée en (c). A est la somme des valeurs des pixels de l'image inclus dans le rectangle A lorsque le masque est centré sur p. Cette somme peut être calculée rapidement à l'aide de l'image intégrale.

pour le même point. En pratique, les auteurs signalent que seuls 15% aux maximum des points, se voient attribués plusieurs orientations.

(4) *Interpolation des orientations* – Pour mieux localiser le maximum, une interpolation est réalisée au voisinage de chaque maximum, en prenant en compte les 3 valeurs les plus proches de ce maximum, cf. figure A.8.(b).

Calcul du descripteur – Le descripteur proposé s'inspire du modèle visuel primaire humain qui est plus sensible aux changements d'orientation du gradient qu'aux déplacements (pour permettre la reconnaissance d'un objet quelque soit le point de vue). Ainsi, le descripteur utilisé est construit de la manière suivante :

(1) *Utilisation de l'orientation et de l'histogramme déjà estimés* – Les mêmes données, cf. figure A.8.(a), sont utilisés, à trois différences près :

(a) L'histogramme ne possède plus 36 mais seulement 8 orientations.

(b) Une rotation, relative à l'orientation du point d'intérêt, est appliquée aux coordonnées ainsi qu'à l'orientation de chaque point (ce qui est estimé comme dans la figure A.8.(b)). Ceci, dans le but de rendre la description invariante aux rotations.

(c) La participation de chaque point dans le calcul de l'histogramme local est pondérée par sa distance au point d'intérêt caractérisé (et non plus de simple poids gaussiens), c'est-à-dire par $1-d$ dans chaque dimension où d est la distance du point au point central.

(2) *Première normalisation du vecteur-descripteur* – Les valeurs sont recadrées entre 0 et 1 afin que le descripteur soit invariant aux changements affines d'illumination.

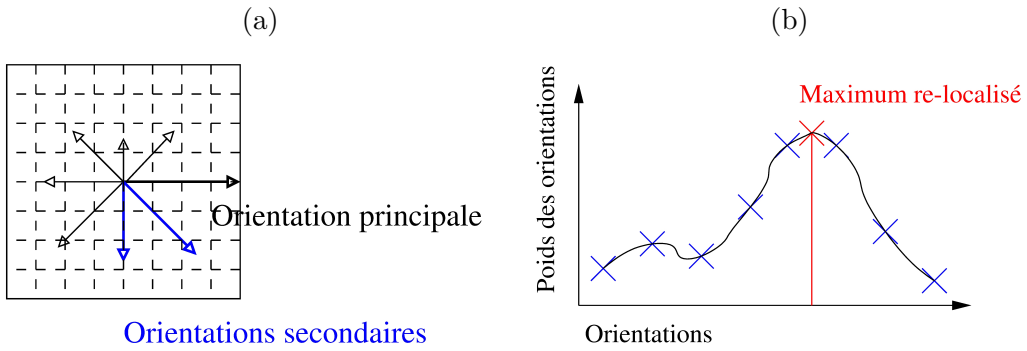


FIGURE A.8. – Histogramme des orientations pour le descripteur SIFT – Un exemple sur 8 orientations seulement où il y a une orientation principale ainsi que 2 orientations secondaires (80% de l'orientation principale). Dans cet exemple, nous utilisons un voisinage 8×8, mais, dans la pratique le descripteur SIFT s'appuie sur un voisinage 4×4.

(3) *Seuillage des forts gradients* – Cette étape est réalisée afin de prendre en compte des changements d'illumination non-linéaires. Les auteurs ont expérimentalement choisi un seuil très restrictif à 0.15. Ceci implique la normalisation suivante.

(4) *Deuxième normalisation du vecteur-descripteur* – De la même manière que la première normalisation.

Les auteurs proposent un descripteur de taille 128 en utilisant un voisinage 4×4 et 8 orientations.

Appariement – Pour la mise en correspondance, une simple distance peut être utilisée. Dans la littérature, nous pouvons citer l'utilisation d'un produit scalaire (entre les deux descripteurs de 128), de la mesure ZNCC ou d'une distance L_2.

HoG, *Histogram of Oriented Gradient* [Dalal 05a]

Cette approche comporte de grandes similitudes avec le descripteur SIFT et c'est pourquoi nous commencerons par ce descripteur. En effet, le principe est de caractériser un point par la distribution locale des orientations des gradients dans son voisinage. Donc, de la même manière que SIFT, il est nécessaire de calculer des histogrammes locaux des orientations des gradients. Plus précisément les étapes de construction de ce détecteur sont :

1. Division du voisinage du point en sous-imagettes régulières, appelées cellules ;
2. Calcul des histogrammes des orientations des gradients pour chaque cellule ;
3. Normalisation des histogrammes pour les rendre robustes aux changements d'illumination ;
4. Combinaison des histogrammes pour former le descripteur.

L'étape (3) est optionnelle. L'avantage de cette approche est de faire intervenir des outils simples de la vision. En effet :

- Les calculs se font sur les images en couleur, avec le système RGB.
- Les dérivées sont calculées par simple différence (avec les filtres $[-1, 0, 1]$ et $[-1, 0, 1]^T$ et il n'est pas nécessaire de filtrer les images au préalable).
- Les histogrammes sont calculés sur 9 bins uniquement et en pondérant par l'intensité du gradient (comme SIFT).
- La normalisation est réalisée sur des ensembles de cellule, que les auteurs appellent des blocs.
- Il est conseillé de prendre des cellules de taille 6×6 (c'est plus petit que pour SIFT) et des blocs de 3×3.

BRIEF, *Binary Robust Independent Elementary Features* [Calonder 10]

Le principe de ce descripteur est de construire une chaîne de bits, relativement courte à l'opposé des descripteurs SIFT ou HoG, permettant de résumer la région autour du point pour ensuite utiliser une distance de Hamming pour effectuer les comparaisons. Un descripteur court pouvant être simplement comparé à un autre, doit, d'après les auteurs, permettre d'être plus rapide et plus efficace que les descripteurs traditionnellement utilisé (SIFT, HoG) et classiquement comparé en utilisant la norme L_2 ou la mesure ZNCC.

Le principe utilisé pour construire la chaîne de bits repose sur la simple comparaison des intensités le long de lignes. Plus précisément, les auteurs proposent le codage suivant, avec (x, y) les coordonnées du pixel étudié :

$$\text{BRIEF}(x, y) = \sum_{1 \leq i \leq N_b} 2^{i-1} \tau(x, y, x_i, y_i), \tag{A.25}$$

avec N_b le nombre de bits formant le descripteur et (x_i, y_i) une position dans le voisinage du point (x, y). De plus la fonction τ est définie par :

$$\tau(x, y, x_i, y_i) = \begin{cases} 1 & \text{si } I(x, y) < I(x_i, y_i) \\ 0 & \text{sinon.} \end{cases} \tag{A.26}$$

Les deux éléments à fixer sont : N_b le nombre de bits (les auteurs ont testé 128, 256 et 512) et surtout le sous-ensemble de points à prendre en compte dans le voisinage. Pour ce dernier aspect, les auteurs ont testés 5 stratégies différents : un tirage suivant une loi uniforme, suivant une loi gaussienne en (x, y), une loi gaussienne séparément sur chaque canal (en x puis en y), un tirage aléatoire suivant une grille du cercle polaire pour introduire une forme de quantification, une répartition uniforme sur une grille du cercle polaire. Avant de calculer ce descripteur et afin d'être robuste aux bruits dans l'image, les auteurs proposent de filtrer l'image au préalable (avec un filtre gaussien par exemple).

Il est important de noter qu'actuellement, une approche utilisant BRIEF a été développée. Il s'agit du détecteur et descripteur ORB, *Oriented FAST and Rotated BRIEF* [Rublee 11]. Cette approche combine le détecteur FAST que nous avons présenté au § A.1.4 avec le descripteur BRIEF. Les auteurs font remarqué que ce détecteur et ce descripteur ne sont pas robustes au changement d'orientation des primitives suivies, comme peuvent l'être SIFT ou SURF. Ils ajoutent donc une façon de prendre en compte l'orientation aussi bien pour le détecteur FAST que pour le descripteur BRIEF. Pour cela, voici les étapes suivies pour l'amélioration des points détectés par FAST :

1. Calcul de N points d'intérêt en utilisant FAST avec un rayon de 9 ;
2. Ordonnancement de ces points en utilisant la mesure de Harris ;
3. Calcul de l'orientation de chaque point en utilisant les moments associés à chaque région circulaire de chaque point sélectionné.

Afin d'être robuste au changement d'échelle, tous ces calculs sont réalisés pour toutes les images d'une pyramide d'images. Enfin, le descripteur BRIEF peut être amélioré en prenant en compte l'orientation associée à chaque point d'intérêt. De plus, les auteurs ont ajouté une seconde phase d'apprentissage pour réduire la présence de points d'intérêt avec une grande corrélation entre eux et ainsi augmenter la variance entre points d'intérêt détectés.

LBP, *Local Binary Pattern* [Ojala 02]

Les motifs locaux binaires, ou *Local Binary Pattern*, LBP, ont été introduits dans le contexte de la classification de texture. Le principe consiste à prendre en considération la distribution des pixels dans un voisinage circulaire, plus précisément, les niveaux de gris des pixels se situant sur un cercle, dont il faut définir le rayon et le pas d'échantillonnage. Les niveaux de gris de cette distribution sont centrés par rapport au niveau de gris du pixel central, en remplaçant toutes les valeurs négatives par 0

et toutes valeurs positives par 1. Ce centrage doit permettre une première invariance aux changements de niveaux de gris. Enfin, ces valeurs sont accumulées en utilisant un facteur binomial associé à chaque niveau, afin de former un code binaire, et ainsi, si nous considérons $\mathbf{p}$ le point pour lequel nous calculons le descripteur, nous définissons :

$$\mathrm{LBP}(\mathbf{p}, P, R) = \sum_{p=0^{P-1}} s(I(\mathbf{p}^{'}) - I(\mathbf{p}))2^p, \tag{A.27}$$

avec P le nombre de pixels considérés pour le motif local, placés sur un cercle de rayon R. Afin d'être invariant aux rotations, les auteurs effectuent $(P-1)$ décalage du code LBP de l'équation (A.27) afin de sélectionner celui de plus faible valeur. De plus, les auteurs identifient le nombre de transitions dans le code (c'est-à-dire le nombre de fois où la valeur de deux bits consécutifs est différente) et considèrent qu'au delà de deux transitions, tous les motifs sont égaux à $P = 1$. Au final, le descripteur associé à une région est l'histogramme de tous les codes obtenus dans son voisinage.

Détecteur s'appuyant sur la covariance [Tuzel 06]

Tous les descripteurs présentés jusque là possèdent de nombreux éléments et sont donc de taille importante, ce qui rend leur utilisation assez coûteuse. Une autre façon de rendre leur estimation et leur utilisation moins coûteuse est de, d'une part de proposer des traitements rapides ou accélérés, et, d'autre part, de proposer un descripteur contenant moins d'élément. C'est le cas du détecteur proposé dans [Tuzel 06] où les auteurs proposent d'utiliser seulement les matrices de covariance des régions d'intérêt étudiés en prenant en compte neuf éléments : position, couleur, dérivées premières et secondes suivants les lignes et les colonnes. De plus, des techniques d'accélération des calculs sont utilisées, notamment en s'appuyant sur les images intégrales, cf. section A.1.6.

A.3. Rappels de géométrie différentielle

A.3.1. Notations

Nous utilisons les notations suivantes :

- $\mathbf{F}_x$: la dérivée partielle d'ordre 1 $\frac{\partial \mathbf{F}}{\partial x}$ d'une fonction vectorielle $\mathbf{F}$ relative à la variable x.
- De manière similaire, $\mathbf{F}_{xy}$: la dérivée partielle d'ordre 2 $\frac{\partial^2 \mathbf{F}}{\partial x \partial y}$ de $\mathbf{F}$ relative aux variables x et y.

De plus, nous supposons une caméra perspective calibrée où les coordonnées des points de l'image sont exprimées dans l'image normalisée, c'est-à-dire que $\mathbf{x} = (x, y)$ est obtenu à partir de $[x, y, 1] = \mathtt{K}^{-1}[u, v, 1]$, où (u, v) sont les coordonnées du pixel et $\mathtt{K}$ est la matrice de calibrage standard [Hartley 04] (matrice triangulaire supérieure).

A.3.2. Définitions: Carte différentielle – Plan tangent – Carte de Gauss

Nous notons $\mathcal{F} \subset \mathbb{R}^3$ une *surface régulière*[2] dont la paramétrisation est donnée par une carte différentielle. $\mathbf{F} : U \subset \mathbb{R}^2 \to \mathcal{F}$ avec

$$\mathbf{F}(x, y) = [X(x, y), Y(x, y), Z(x, y)]^\top \tag{A.28}$$

À chaque $\mathbf{x} = [x, y]^\top \in U$, une carte est associée. Elle est notée : $\mathrm{d}\mathbf{F}_\mathbf{x} : \mathbb{R}^2 \to \mathbb{R}^3$, et nous l'appelons la différentielle de $\mathbf{F}$ en $\mathbf{x}$. Elle est définie dans [do Carmo 76, p. 128]. Nous notons $\mathbf{v} \in \mathbb{R}^2$ un vecteur et $\boldsymbol{\alpha} : (-\epsilon, \epsilon) \to U$ une courbe différentiable qui satisfait $\boldsymbol{\alpha}(0) = \mathbf{x}$ et $\boldsymbol{\alpha}'(0) = \mathbf{v}$. Ainsi, la courbe $\mathbf{F} \circ \boldsymbol{\alpha}$ de $\mathbb{R}^3$ est également différentiable. Avec ces éléments, nous définissons

$$\mathrm{d}\mathbf{F}_\mathbf{x}(\mathbf{v}) = (\mathbf{F} \circ \boldsymbol{\alpha})'(0)$$

2. La définition d'une surface régulière dans R^n est donnée dans [Gray 06, p. 281-286].

qui nous fournit une première approximation linéaire, c'est-à-dire du premier ordre, de $\mathbf{F}(\mathbf{x}+\mathbf{v})$ avec $\mathbf{v}$ suffisamment petit. Ceci est illustré dans la figure A.9(a). Le sous-espace vectoriel $\mathrm{d}\mathbf{F}_{\mathbf{x}}(U) \subset \mathbb{R}^3$ a une dimension 2 est il correspond au plan contenant tous les vecteurs tangents de $\mathbf{F}$ avec $\mathbf{P} = \mathbf{F}(\mathbf{x})$. Il s'agit du *plan tangent* de $\mathcal{F}$ en $\mathbf{P}$ et il est noté $T_{\mathbf{P}}(\mathcal{F})$.

Dans [do Carmo 76, p129], il est prouvé que cette définition ne dépend pas du choix pour $\boldsymbol{\alpha}$. De plus, le fait que $(\mathbf{F} \circ \boldsymbol{\alpha})'(0) = \mathbf{F}'(\boldsymbol{\alpha}(0))\boldsymbol{\alpha}'(0)$ implique que $\mathrm{d}\mathbf{F}_{\mathbf{x}}(\mathbf{v})$ est linéaire en $\mathbf{v}$. En particulier, pour les bases canoniques de $\mathbb{R}^2$ et $\mathbb{R}^3$, nous avons

$$\mathrm{d}\mathbf{F}_{\mathbf{x}}(\mathbf{v}) = \mathtt{J}_{\mathbf{F}}(\mathbf{x})\mathbf{v}$$

Ceci implique que la *matrice jacobienne* de $\mathbf{F}$ en $\mathbf{x}$, de taille 3×2 est définie par :

$$\mathtt{J}_{\mathbf{F}}(\mathbf{x}) = \begin{bmatrix} \mathbf{F}_x(\mathbf{x}) & \mathbf{F}_y(\mathbf{x}) \end{bmatrix} \tag{A.29}$$

avec $\mathbf{F}_x$ et $\mathbf{F}_y$ des colonnes. Ceci implique également que le sous-espace vectoriel $\mathrm{d}\mathbf{F}_{\mathbf{x}}(U)$ est de dimension 2.

Soit $\mathbf{P} = \mathbf{F}(\mathbf{x})$ un point de $\mathcal{F}$. Nous posons $\mathbf{N} : \mathcal{F} \subset \mathbb{R}^3 \to \Sigma \subset \mathbb{R}^3$ la carte différentielle qui attribue à chaque $\mathbf{P}$, le vecteur de coordonnées $\mathbf{N}(\mathbf{P})$, sur la sphère unité Σ. Ce vecteur correspond à la normale unitaire de $\mathcal{F}$ en $\mathbf{P}$ et se calcule de la manière suivante :

$$\mathbf{N}(\mathbf{P}) = \frac{\mathbf{F}_x(\mathbf{x}) \times \mathbf{F}_y(\mathbf{x})}{\|\mathbf{F}_x(\mathbf{x}) \times \mathbf{F}_y(\mathbf{x})\|} \text{ avec } \mathbf{x} = \mathbf{F}^{-1}(\mathbf{P}) \tag{A.30}$$

Cette carte est également appelée la carte de Gauss, *Gauss map*, de $\mathcal{F}$.

La carte de Gaus est un recalage entre les deux surfaces $\mathcal{F}$ et Σ et la définition de la différentielle est étendue à cette carte. La différentielle de la carte de Gauss $\mathcal{F}$ au point $\mathbf{P}$ est la carte donnée par $\mathrm{d}\mathbf{N}_{\mathbf{P}} : T_{\mathbf{P}}(\mathcal{F}) \subset \mathbb{R}^3 \to \mathbb{R}^3$. Ainsi, si nous notons $\mathbf{V} \in T_{\mathbf{P}}(\mathcal{F})$ un vecteur dans la plan tangent de $\mathcal{F}$ en $\mathbf{P}$ et $\boldsymbol{\beta} : (-\epsilon, \epsilon) \to \mathcal{F}$ une courbe différentiable sur la surface $\mathcal{F}$ qui vérifie $\boldsymbol{\beta}(0) = \mathbf{P}$ et $\boldsymbol{\beta}'(0) = \mathbf{V}$. Ainsi $\mathbf{N} \circ \boldsymbol{\beta}$ en $\mathbb{R}^3$ est également différentiable. Ensuite, nous définissons :

$$\mathrm{d}\mathbf{N}_{\mathbf{P}}(\mathbf{V}) = (\mathbf{N} \circ \boldsymbol{\beta})'(0) \tag{A.31}$$

Cette expression permet de représenter le travail de $\mathbf{N}$. Encore une fois, il a été prouvé [do Carmo 76, p129], que la définition que nous avons choisie pour $\mathrm{d}\mathbf{N}_{\mathbf{P}}$ ne dépend pas du choix de la courbe $\boldsymbol{\beta}$.

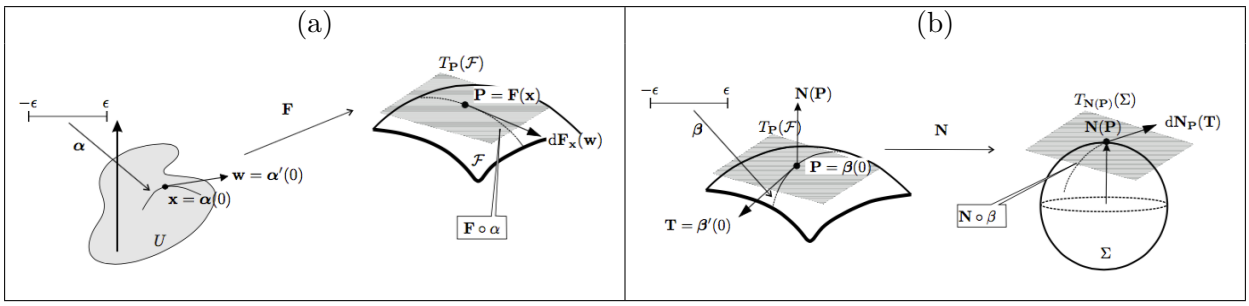


FIGURE A.9. – Illustration d'une carte gaussienne associée à une surface – En (a) La surface régulière $\mathcal{F}$ et sa différentielle au point $\mathbf{x}$ dans la direction $\mathbf{w}$ est paramétrée par la carte $\mathbf{F}$. En (b), $\mathbf{N}$ est la carte gaussienne, c'est-à-dire, la correspondance entre la surface $\mathcal{F}$ et la sphère unité Σ. Nous illustrons également la différentielle $\mathbf{P}$ dans la direction $\mathbf{T}$.

De la même façon que la différentielle assoicée à une carte, tel que nous venons de l'expliquer, le fait d'avoir $(\mathbf{N} \circ \boldsymbol{\beta})'(0) = \mathbf{N}'(\boldsymbol{\beta}(0))\boldsymbol{\beta}'(0)$ garantit que $\mathrm{d}\mathbf{N}_{\mathbf{P}}(\mathbf{V})$ est linéaire en $\mathbf{V} = \boldsymbol{\beta}'(0)$. Le sous-espace

vectoriel $\mathrm{d}\mathbf{N}_\mathbf{P}(T_\mathbf{P}(\mathcal{F})) \subset \mathbb{R}^3$ possède deux dimensions : il s'agit du plan $T_{\mathbf{N}(\mathbf{P})}(\Sigma)$ qui contient tous les vecteurs tangents à la sphère unité $\mathbf{N}(\mathbf{P})$, appelé le plan tangent de Σ en $\mathbf{N}(\mathbf{P})$. Ainsi $\mathrm{d}\mathbf{N}_\mathbf{P}$ est compris dans $\mathrm{d}\mathbf{N}_\mathbf{P}(T_\mathbf{P}(\mathcal{F})) = T_{\mathbf{N}(\mathbf{P})}(\Sigma)$. Plus précisément, nous pouvons facilement montrer que $T_{\mathbf{N}(\mathbf{P})}(\Sigma)$ et $T_\mathbf{P}(\mathcal{F})$ sont des plans parallèles et ainsi, la différentielle de $\mathbf{N}$ est définie par $\mathrm{d}\mathbf{N}_\mathbf{P} : T_\mathbf{P}(\mathcal{F}) \to T_\mathbf{P}(\mathcal{F})$.

A.3.3. Courbures : formes fondamentales

Soit $\mathbf{T} \in T_\mathbf{P}(\mathcal{F})$ un *vecteur unitaire* qui correspond à une direction danns le plan tangent de $\mathcal{F}$ en $\mathbf{P}$. Soit $\mathscr{C}$ la courbe obtenue en prenant une coupe de $\mathcal{F}$ avec une section normalisée de $\mathcal{F}$ en $\mathbf{P}$, le long de la direction $\mathbf{T}$ [3], c'est-à-dire, le plan passant par $\mathbf{P}$ et parallèle simultanément à $\mathbf{N}(\mathbf{P})$ et $\mathbf{T}$. La *courbure normale* de $\mathcal{F}$ dans la direction unitaire $\mathbf{T}$ est la courbure de $\mathscr{C}$ en $\mathbf{P}$, donnée par [do Carmo 76, p. 144]:

$$\kappa_N(\mathbf{T}) = -\mathbf{T} \cdot \mathrm{d}\mathbf{N}_\mathbf{P}(\mathbf{T}) \tag{A.32}$$

Cette expression montre comment $\mathbf{N}$, les courbes $\mathcal{F}$, se comportent dans le voisinage de $\mathbf{P}$. Une remarque importante est que le *rayon de courbure* de $\mathscr{C}$ en $\mathbf{P}$ est égal à : $1/|\kappa_N|$.

Les *courbures principales* κ_1, κ_2 de $\mathcal{F}$ en $\mathbf{P}$ peuvent se définir comme les extrema de la fonction (A.32) dans la direction $\mathbf{T} \in T_\mathbf{P}(\mathcal{F})$ avec la contrainte $\|\mathbf{T}\| = 1$. Les directions correspondantes sont appelées les *directions principales* de $\mathcal{F}$ en $\mathbf{P}$. Dans [Koenderink 92], il est exprimé le fait que les coefficients κ_1 et κ_2 sont des paramètres déterminants pour décrire complètement la forme de la surface au second ordre, modulo un mouvement rigide. Cela signifie que les courbures principales sont invariantes à la paramétrisation de la surface.

À présent, nous considérons un nouveau système de coordonnées 3D avec $\mathbf{P} = \mathbf{F}(\mathbf{x})$ pour origine. Tout vecteur $\mathbf{T}$ sur le plan tangent $T_\mathbf{P}(\mathcal{F})$ peut s'écrire :

$$\mathbf{T} = u\mathbf{F}_x(\mathbf{x}) + v\mathbf{F}_y(\mathbf{x}) = \mathtt{J}_\mathbf{F}(\mathbf{x}) \begin{bmatrix} u \\ v \end{bmatrix} \tag{A.33}$$

avec $\mathtt{J}_\mathbf{F}(\mathbf{x})$ la matrice Jacobienne de $\mathbf{F}$ définie par (A.29) et (u, v) les *coordonnées locales* de $\mathbf{T}$. À partir de maintenant, uv en indice, comme dans $\mathbf{T}_{uv}$, indique les coordonnées locales d'un vecteur.

Première forme fondamentale d'une surface

Soit (u, v), la norme d'un vecteur quelconque $\mathbf{T}_{uv}$ sur $T_\mathbf{P}(\mathcal{F})$ est donnée par :

$$\|\mathbf{T}_{uv}\| = \sqrt{[u, v]\mathtt{I}_\mathbf{P}[u, v]^\top} \tag{A.34}$$

avec

$$\mathtt{I}_\mathbf{P} = \begin{bmatrix} \mathbf{F}_x(\mathbf{x}) \cdot \mathbf{F}_x(\mathbf{x}) & \mathbf{F}_x(\mathbf{x}) \cdot \mathbf{F}_y(\mathbf{x}) \\ \mathbf{F}_x(\mathbf{x}) \cdot \mathbf{F}_y(\mathbf{x}) & \mathbf{F}_y(\mathbf{x}) \cdot \mathbf{F}_y(\mathbf{x}) \end{bmatrix}$$

La forme quadratique sur $T_\mathbf{P}(\mathcal{F})$

$$\mathfrak{I}_\mathbf{P}(u, v) \doteq [u, v]\mathtt{I}_\mathbf{p}[u, v]^\top \tag{A.35}$$

est appelée la *première forme fondamentale* de $\mathcal{F}$ [do Carmo 76, p94].

Seconde forme fondamentale d'une surface

Soit $\mathbf{T}_{uv}$ une direction sur le plan tangent $T_\mathbf{P}(\mathcal{F})$, en coordonnées locales 2D. Dans [do Carmo 76, p156], la différentielle de la carte de Gauss de $\mathcal{F}$ en $\mathbf{P} = \mathbf{F}(\mathbf{x})$ dans la direction $\mathbf{T}_{uv}$ s'écrit, en

3. Nous pouvons vérifié que tout plan passant par $\mathbf{P}$ est parallèle à $\mathbf{T}$ d'après le théorème de Meusnier [Gallier 01, p482].

coordonnées 3D standard :

$$\mathrm{d}\mathbf{N}_\mathbf{P}(\mathbf{T}_{uv}) = [\mathbf{N}_x(\mathbf{x}) \mid \mathbf{N}_y(\mathbf{x})] \begin{bmatrix} u \\ v \end{bmatrix}$$

De plus, si nous notons $\mathrm{d}\mathbf{N}_\mathbf{P}(u,v)$, la différentielle de $\mathbf{N}$ en $\mathbf{P} = \mathbf{F}(\mathbf{x})$ dans la direction $\mathbf{T}_{uv}$, exprimée en coordonnées locales 2D, c'est-à-dire telle que $\mathrm{d}\mathbf{N}_\mathbf{P}(\mathbf{T}_{uv}) = \mathtt{J}_\mathbf{F}(\mathbf{x})]\mathrm{d}\mathbf{N}_\mathbf{P}(u,v)$, alors, nous obtenons :

$$\mathrm{d}\mathbf{N}_\mathbf{P}(u,v) = \mathtt{I}_\mathbf{P}^{-1}\mathtt{II}_\mathbf{P} \begin{bmatrix} u \\ v \end{bmatrix} \tag{A.36}$$

où

$$\mathtt{II}_\mathbf{P} = \begin{bmatrix} \mathbf{N}_x(\mathbf{x}) \cdot \mathbf{F}_x(\mathbf{x}) & \mathbf{N}_x(\mathbf{x}) \cdot \mathbf{F}_y(\mathbf{x}) \\ \mathbf{N}_x(\mathbf{x}) \cdot \mathbf{F}_y(\mathbf{x}) & \mathbf{N}_y(\mathbf{x}) \cdot \mathbf{F}_y(\mathbf{x}) \end{bmatrix} \tag{A.37}$$

La preuve est donnée dans [do Carmo 76, p156].

La forme quadratique

$$\mathfrak{II}_\mathbf{P}(u,v) = [u,v]\mathtt{II}_\mathbf{P}[u,v]^\top \tag{A.38}$$

est appelée la *seconde forme fondamentale de* $\mathcal{F}$ [do Carmo 76, p143]. Ceci implique directement que (A.32) peut s'exprimer en coordonnées locales et ainsi donner :

$$\kappa_N(u,v) = [u,v]\mathtt{II}_\mathbf{P}[u,v]^\top \tag{A.39}$$

Pour chaque vecteur non unitaire $\mathbf{V}(u,v)$ suivant $T_\mathbf{P}(\mathcal{S})$, afin d'obtenir une direction unitaire, il suffit de diviser (A.39) par le carrée de l'expression (A.34). Ainsi, la *courbure normale* dans la direction unitaire $\frac{\mathbf{V}}{\|\mathbf{V}\|}$ est donnée par

$$\kappa_N(u,v) = -\frac{\mathbf{V}_{uv} \cdot \mathrm{d}\mathbf{N}_\mathbf{P}(\mathbf{V}_{uv})}{\|\mathbf{V}_{uv}\|^2} = \frac{\mathfrak{II}_\mathbf{P}(u,v)}{\mathfrak{I}_\mathbf{P}(u,v)} \tag{A.40}$$

Solutions de courbures principales pour des formes fermées

Les courbures principales κ_1 et κ_2 de $\mathcal{S}$ en $\mathbf{P}$ peuvent être définies comme les extrema de la fonction (A.40) avec les coordonées (u,v) comme variables.

Si nous notons (A.40) comme le quotient généralise de Rayleigh, il a été montré, cf. [Parlett 98, p18], que κ_N possède un extremum en $(\hat{u},\hat{v})$ seulement si $\kappa_N(\hat{u},\hat{v})$ est une racine de $\det(\mathtt{II}_\mathbf{P} - \kappa_N\mathtt{I}_\mathbf{P})$ ou, de manière équivalente, seulement si $\kappa_N(\hat{u},\hat{v})$ est une valeur propre de la matrice $2\times 2 : \mathtt{I}_\mathbf{P}^{-1}\mathtt{II}_\mathbf{P}$. Cette matrice n'est pas symétrique mais a toujours des valeurs propres réelles [Gallier 01, p500]. Ainsi, les courbures principales sont les deux valeurs propres κ_α $(\alpha = 1,2)$ de la matrice $\mathtt{I}_\mathbf{P}^{-1}\mathtt{II}_\mathbf{P}$. Les directions principales 3D sont $\mathbf{T}_\alpha = \mathtt{J}_\mathbf{F}(\mathbf{x})\mathbf{e}_\alpha$ où $\mathbf{e}_\alpha$ sont les vecteurs propres correspondants.

Nous pouvons ainsi formuler la proposition suivante qui est fondamentale dans notre travail :

Proposition 3 *La courbure principale* κ_α $(\alpha = 1,2)$ *en* $\mathbf{P}$ *associée à la direction unitaire 3D principale* $\mathbf{T}_\alpha$ *est égale à l'amplitude absolue de la différentielle de la carte de Gauss en ce point, soit :*

$$|\kappa_\alpha| = \|\mathrm{d}\mathbf{N}_\mathbf{P}(\mathbf{T}_\alpha)\| \tag{A.41}$$

Proof. Comme la norme euclidienne est invariante aux changements de coordonnées euclidiens, sans perte de généralié, nous pouvons choisir une nouvelle paramétrisation telle que : $\tilde{\mathbf{S}}(\tilde{x},\tilde{y}) = [\tilde{x},\tilde{y},\tilde{Z}(\tilde{x},\tilde{y})]^\top$, pour une nouvelle fonction de profondeur $\tilde{Z}$, c'est à dire un repère orthonormé 3D dont l'origine est $\mathbf{P}$ et le plan $\tilde{x}\tilde{y}$ correspond au plan tangent $T_\mathbf{P}(\mathcal{F})$. De manière générale, nous ajoutons le symbole $\tilde{}$ pour distinguer les nouvelles représentations des anciennets, à l'exception des

courbures principales qui sont indépendants des paramétrages. Nous rappelons que : $\tilde{\mathbf{T}}_\alpha = \mathtt{J}_{\mathbf{F}}(\mathbf{x})\tilde{\mathbf{e}}_\alpha$, où $\tilde{\mathbf{e}}_\alpha$ est le vecteur propre associé et nous notons que la matrice de la nouvelle forme fondamentale $\tilde{\mathtt{I}}_{\mathbf{P}}$ est la matrice identité. En conséquence, en partant du fait que κ_α est une valeur propre de la matrice 2×2 : $\widetilde{\mathtt{I}}^{-1}_{\tilde{\mathbf{P}}}\widetilde{\mathtt{II}}_{\tilde{\mathbf{P}}}$, nous avons :

$$\begin{aligned}
(\widetilde{\mathtt{I}}^{-1}_{\tilde{\mathbf{P}}}\widetilde{\mathtt{II}}_{\tilde{\mathbf{P}}})\tilde{\mathbf{e}}_\alpha &= \kappa_\alpha\tilde{\mathbf{e}}_\alpha \\
\Leftrightarrow \mathrm{d}\tilde{\mathbf{N}}_{\tilde{\mathbf{P}}}(\tilde{\mathbf{e}}_\alpha) &= \kappa_\alpha\tilde{\mathbf{e}}_\alpha \\
\Leftrightarrow [\tilde{\mathbf{S}}_{\tilde{\mathbf{x}}} \mid \tilde{\mathbf{S}}_{\tilde{\mathbf{y}}}]\mathrm{d}\tilde{\mathbf{N}}_{\tilde{\mathbf{P}}}(\tilde{\mathbf{e}}_\alpha) &= \kappa_\alpha[\tilde{\mathbf{S}}_{\tilde{x}} \mid \tilde{\mathbf{S}}_{\tilde{y}}]\tilde{\mathbf{e}} \\
\Leftrightarrow \mathrm{d}\tilde{\mathbf{N}}_{\mathbf{P}}(\tilde{\mathbf{T}}_\alpha) &= \kappa_\alpha\tilde{\mathbf{T}}_\alpha \\
\Rightarrow \|\mathrm{d}\tilde{\mathbf{N}}_{\mathbf{P}}(\tilde{\mathbf{T}}_\alpha)\|^2 &= \kappa_\alpha^2
\end{aligned}$$

■

A.3.4. Analyses de l'indicatrice de Dupin

Dans ce que nous avons proposé, nous détectons les points tels que cette différence soit élevée :

$$CS(x, y) = \kappa_1(x, y) - \kappa_2(x, y) \tag{A.42}$$

Nous appelons (A.42) la saillance curviligne, *curvilinear saliency*, CS. Curviligne signifie que l'élément se trouve sur une ligne courbe.

Nous considérons à présent les indicatrices de Dupin mais plus précisément les ellipses et les hyperboles réelles. Nous ne prenons pas en compte le cas des paraboles car, dans les expériences réelles, la condition $\kappa_2 = 0$ ne pourra jamais être vérifiée exactement.

De nombreuses mesures peuvent décrire ces formes de coniques [Koenderink 92]. Nous avons choisi d'introduire la quantité CS, cf. équation (A.42), qui permet de traiter de manière unique les ellipses et les hyperboles en utilisant les interprétations suivantes.

Soit le demi grand axe et le demi petit axe d'une indicatrice de Dupin notés $r_{\mathrm{maj}} = \max(\rho_1^2, \rho_2^2)$ et $r_{\mathrm{min}} = \min(\rho_1^2, \rho_2^2)$ respectivement, où ρ_1 et ρ_2 sont les rayons de coubure des courbes obtenues pour la section normale de $\mathcal{D}$ en $\mathbf{P}$ le long des directions principales.

Proposition 4 *Le terme CS dans (4.14) est le carré du rapport entre l'excentricité e de la conique de Dupin et son petit axe, soit :*

$$\kappa_1 - \kappa_2 = \left(\frac{e}{r_{min}}\right)^2 \tag{A.43}$$

où

$$e = \sqrt{1 \pm \frac{r_{min}^2}{r_{maj}^2}} \quad \textit{avec} \begin{cases} - : & \textit{si la conique est une ellipse.} \\ + : & \textit{si la conique est une hyperbole.} \end{cases}$$

L'excentricité peut être interprétée comme la fraction de distance le long du grand axe où le foyer se situe. Le terme CS est normalisé en divisant le carré de l'excentricité par le carrée du demi petit axe. Il est important de noter que (A.43) fonctionne aussi bien pour les ellipses que pour les hyperboles de Dupin.

A.3.5. K, courbure gaussienne de $\mathcal{D}$: preuve

Nous avons : $\mathtt{I}^{-1}_{\mathbf{P}}\mathtt{II}_{\mathbf{P}} = \frac{1}{\alpha}\mathsf{M}$ avec

$$\mathsf{M} \triangleq \begin{bmatrix} (Z_y^2+1)Z_{xx} - Z_xZ_yZ_{xy} & (Z_y^2+1)Z_{xy} - Z_xZ_yZ_{yy} \\ (Z_x^2+1)Z_{xy} - Z_xZ_yZ_{xx} & (Z_x^2+1)Z_{yy} - Z_xZ_yZ_{xy} \end{bmatrix} \tag{A.44}$$

Soit λ_1 et λ_2 les valeurs propres ordonnées ($\lambda_1 \geq \lambda_2$), obtenues par $-\mathsf{M}$ comme définies dans (A.44), pour obtenir $\kappa_i = \|\nabla Z\|\,\lambda_i$, $i = 1, 2$. Comme les deux valeurs propres de $-\mathsf{M}$ s'écrivent :

$$\lambda_{\pm} = \frac{1}{2}\left(-\operatorname{trace}(\mathsf{M}) \pm \sqrt{(\operatorname{trace}\mathsf{M})^2 - 4\det\mathsf{M}}\right) \tag{A.45}$$

nous avons $\lambda_1 = \lambda_+$ et $\lambda_2 = \lambda_-$. De plus, comme

$$\kappa_1 - \kappa_2 = \|\nabla Z\|\,(\lambda_1 - \lambda_2) \tag{A.46}$$

nous obtenons bien la définition de la salliance curviligne.

A.3.6. Preuve de la proposition 2

Les valeurs propres ordonnées de $\mathtt{I}^{\mathcal{I}}_{\mathbf{P}} = \mathtt{J}^{\top}_{\mathbf{I}}\mathtt{J}_{\mathbf{I}}$ peuvent se déduire des valeurs propres de $\nabla I \nabla I^{\top}$, c'est-à-dire, $\|\nabla I\|^2$ et 0, et donc, $\mu_1 = \|\nabla I\|^2 + 1$ et $\mu_2 = 1$. Ceci permet de conclure la preuve.

A.4. Résultats de détection de primitives avec MCS et MFC

Méthodes	MFC	MCS	PCBR [Deng 07]	Harris [Harris 88]	SIFT [Lowe 04]	SURF [Bay 08]
Car	21	29	40	77	85	71
Shoe	34	52	67	106	111	108
Plane	26	23	19	43	46	47
T-Rex	20	17	25	100	143	46
Elephant	21	41	55	91	114	74
Fhydrant	15	23	35	86	74	67
Jeep	29	31	42	67	74	89
Mug	35	56	65	133	134	145
Teddy	19	24	31	69	77	101
Pistol	18	16	26	96	44	73

TABLE A.1. – Distance de Hausdorff moyenne des détecteurs de points d'intérêt pour la base de test *web collection* – Plus cette valeur est faible, et plus le détecteur est précis. Il s'agit de la moyenne calculée entre toutes les images réelles avec différentes textures et éclairage et toutes les images de profondeur associée. Ce calcul est réalisé pour tous les points de vue envisagés.

Méthodes	MFC	MCS	PCBR [Deng 07]	Harris [Harris 88]	SIFT [Lowe 04]	SURF [Bay 08]
Plane	47	48	59	63	68	73
Bicycle	71	75	79	101	93	100
Boat	62	68	75	77	87	76
Bus	106	110	117	123	131	127
Car	80	85	98	100	113	108
Diningtable	84	85	96	118	118	111
Motorbike	62	64	78	96	94	86
Sofa	70	75	77	98	93	92
Train	101	108	121	123	133	127
Tvmonitor	96	102	104	104	111	105
Chair	92	105	115	108	107	112
Bottle	78	84	87	90	92	97

TABLE A.2. – Distance de Hausdorff moyenne des détecteurs de points d'intérêt pour la base de test *PASCAL3D+* – Plus cette valeur est faible, et plus le détecteur est précis. Il s'agit de la moyenne calculée entre toutes les images réelles avec différentes textures et éclairage et toutes les images de profondeur associée. Ce calcul est réalisé pour tous les points de vue envisagés.

Annexe B.

Éléments nécessaires à la mise en correspondance

B.1. Mesures de corrélation

B.1.1. Corrélation croisée

Ce paragraphe concerne les mesures s'appuyant sur un produit scalaire entre les deux vecteurs contenant les niveaux de gris des voisins de chaque point apparié. Plus précisément, la mesure de corrélation croisée normalisée, notée NCC (*Normalized Cross-Correlation*), est donnée par :

$$\mathrm{NCC}(\mathbf{f}_g, \mathbf{f}_d) = \frac{\mathbf{f}_g \cdot \mathbf{f}_d}{\|\mathbf{f}_g\|\|\mathbf{f}_d\|}. \tag{B.1}$$

Les valeurs de $\mathrm{NCC}(\mathbf{f}_g, \mathbf{f}_d)$ appartiennent à l'intervalle [0;1].

Il existe une version centrée de cette mesure, donnée par :

$$\mathrm{ZNCC}(\mathbf{f}_1, \mathbf{f}_2) = \frac{(\mathbf{f}_1 - \overline{\mathbf{f}_1}) \cdot (\mathbf{f}_2 - \overline{\mathbf{f}_2})}{\|\mathbf{f}_1 - \overline{\mathbf{f}_1}\|\|\mathbf{f}_2 - \overline{\mathbf{f}_2}\|} \tag{B.2}$$

où $\mathbf{f}_w, w = 1, 2$ correspond aux vecteurs des niveaux de gris dans les voisinages respectifs des points appariés. Il s'agit de la corrélation centrée normalisée, *Zero-mean Normalised Cross Correlation*.

Pour ces deux mesures, la normalisation est nécessaire car, sans normalisation, plus les niveaux de gris sont élevés, plus les scores sont élevés. De plus, cela permet une invariance de type gain. Cette mesure correspond au coefficient de corrélation linéaire classique en statistiques. Grâce au centrage, elle a l'avantage de présenter une invariance de type biais.

B.1.2. Mesures classiques

Ces mesures s'appuient sur des outils de statistiques classiques et regroupent des mesures utilisant la distribution des différences des niveaux de gris des deux fenêtres, comme les distances ou norme L_P. Ces mesures de corrélation, notées D_P (*Distances*), sont données par :

$$\mathrm{D}_P(\mathbf{f}_g, \mathbf{f}_d) = \|\mathbf{f}_g - \mathbf{f}_d\|_P{}^P \text{ avec } P \in \mathbb{N}^*. \tag{B.3}$$

Ce sont des mesures de dissimilarité et leurs valeurs appartiennent à l'intervalle $[0; \mathrm{I}_{\max}^P N_f]$.

Les distances localement centrées, mesures proposées dans [Aschwanden 92], sont les mesures qui obtiennent les résultats les plus performants dans nos travaux précédents. Le principe des mesures localement centrées et notées LD_P (*Locally scaled Distances*), est de ramener la moyenne des niveaux de gris de la fenêtre de corrélation dans l'image droite à celle de la fenêtre de corrélation dans l'image gauche. Pour cela, chaque niveau de gris de l'image droite est multiplié par le rapport des moyennes

des niveaux de gris des fenêtres de corrélation gauche et droite. Ces mesures sont données par :

$$\mathrm{LD}_P(\mathbf{f}_g, \mathbf{f}_d) = \mathrm{D}_P\left(\mathbf{f}_g, \frac{\overline{\mathbf{f}_g}}{\overline{\mathbf{f}_d}}\mathbf{f}_d\right) = \left\|\mathbf{f}_g - \frac{\overline{\mathbf{f}_g}}{\overline{\mathbf{f}_d}}\mathbf{f}_d\right\|_P^P. \tag{B.4}$$

Ce sont des mesures de dissimilarité et leurs valeurs appartiennent à l'intervalle $[0; {\mathrm{I}_{\max}}^P N_f]$.

La mesure LSAD, *Locally scaled Sum of Absolute Differences*, celle qui obtient les meilleures résultats, est donnée par :

$$\mathrm{LSAD}(\mathbf{f}_g, \mathbf{f}_d) = \mathrm{LD}_1(\mathbf{f}_g, \mathbf{f}_d) = \mathrm{D}_1\left(\mathbf{f}_g, \frac{\overline{\mathbf{f}_g}}{\overline{\mathbf{f}_d}}\mathbf{f}_d\right) = \left\|\mathbf{f}_g - \frac{\overline{\mathbf{f}_g}}{\overline{\mathbf{f}_d}}\mathbf{f}_d\right\|_1. \tag{B.5}$$

B.1.3. Corrélation des champs de gradients

Les mesures présentées jusqu'à maintenant utilisaient directement les niveaux de gris. Mais, il existe également des mesures s'appuyant sur les images dérivées. La mesure la plus performante, appelée corrélation des champs de gradients, notée GC (*Gradient field Correlation*) et proposée par [Crouzil 96], utilise la similarité de vecteurs gradients et est donnée par :

$$\mathrm{GC}(\mathbf{f}_g, \mathbf{f}_d) = \frac{\sum_{p=-N_v}^{N_v}\sum_{q=-N_h}^{N_h} \|\nabla I_g(i+p, j+q) - \nabla I_d(i+p, v+q)\|}{\sum_{p=-N_v}^{N_v}\sum_{q=-N_h}^{N_h} (\|\nabla I_g(i+p, j+q)\| + \|\nabla I_d(i+p, v+q)\|)}. \tag{B.6}$$

Il s'agit d'une mesure de dissimilarité dont les valeurs appartiennent à l'intervalle $[0; \infty[$. Cette mesure est invariante au biais. Cette mesure GC est résistante aux bruits.

B.1.4. CENSUS, mesure de recensement

Sur le même principe que dans le paragraphe précédent, certaines mesures reposent sur le rang ou l'ordre des niveaux de gris au lieu de s'appuyer directement sur les niveaux de gris. Dans cette catégorie, la plus célèbre est **census**. Elle s'appuie sur la transformation de recensement qui permet d'obtenir une chaîne de bits représentative des pixels contenus dans la fenêtre de corrélation. Cette chaîne rend compte des pixels dont l'intensité est inférieure à celle du pixel central. La transformation, notée R_τ, est définie par :

$$\mathrm{R}_\tau(\mathbf{p}) = \bigotimes_{k\in[0;N_f-1]} \xi(I^{N_f/2}, I^k), \tag{B.7}$$

où $\bigotimes$ correspond ici à la concaténation. Le terme $\xi(I^{N_f/2}, I^k)$ vaut 1 si $I^{N_f/2} < I^k$ ou 0 dans le cas contraire.

Au final, la mesure de corrélation associée, notée CENSUS, utilise la distance de Hamming :

$$\mathrm{CENSUS}(\mathbf{f}_1, \mathbf{f}_2) = \sum_k^{N_f-1} \mathrm{D}_{Ham}(\mathrm{R}_\tau(\mathbf{p}_1^k), \mathrm{R}_\tau(\mathbf{p}_2^k)). \tag{B.8}$$

où la distance de Hamming est donnée par :

$$\mathrm{D}_{\mathrm{Ham}}(\mathbf{a}_1, \mathbf{a}_2) = \sum_{k=0}^{N_f-1} \mathrm{sgn}\,|a_1^k - a_2^k|. \tag{B.9}$$

La distance de Hamming donne une valeur comprise dans l'intervalle $[0; N_f]$. Ainsi, la mesure CENSUS est une mesure de dissimilarité et ses valeurs appartiennent à l'intervalle $[0; N_f^2]$.

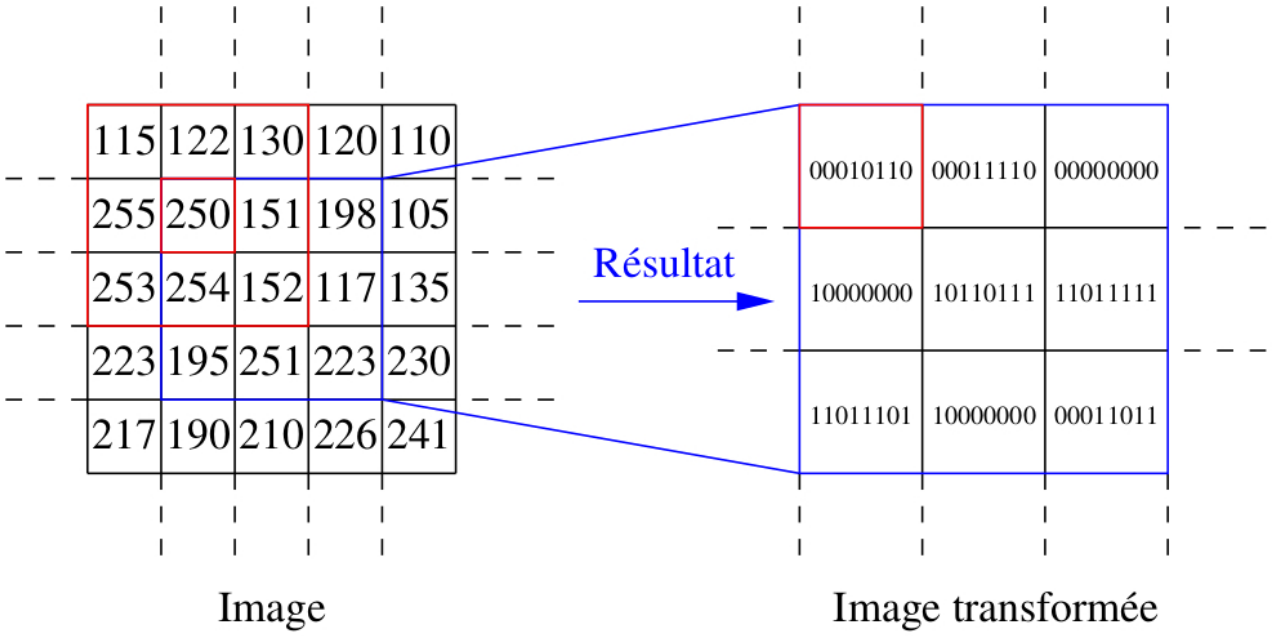


FIGURE B.1. – Mesure CENSUS – Chaque couleur est remplacée par un code binaire, calculé dans son voisinage et qui représente son voisinage. Plus précisément, 1, respectivement 0, indique que le pixel du voisinage possède un niveau de gris plus élevé, respectivement moins élevé, que celui du pixel. Le code binaire ainsi obtenu est comparé à celui du pixel correspondant candidat en utilisant la distance de Hamming distance. Ici nous supposons que le voisinage pris en compte est de taille 3×3.

B.1.5. *Smooth Median Powered Deviation*

Dans nos travaux, nous nous sommes particulièrement intéressés au problème des occultations qui se traduit par le fait que dans le voisinage d'un pixel situé à la frontière d'un changement de profondeur, certains pixels appartiennent au premier niveau de profondeur et d'autres au deuxième niveau de profondeur, ce qui peut perturber la mise en correspondance et provoquer des résultats erronés. Pour prendre en compte ce problème, les mesures robustes s'appuient sur le principe suivant : les pixels qui ont une profondeur différente du pixel étudié sont considérés comme des données aberrantes. Ainsi, ces mesures utilisent des outils de statistiques robustes qui sont beaucoup moins sensibles aux données aberrantes que les outils des statistiques classiques.

En particulier, la mesure s'appuyant sur l'estimateur SMAD (*Smooth Median Absolute Deviation*), qui correspond à une estimation robuste de la variance, proposé dans [Rousseeuw 92] est celle qui nous permet d'obtenir les résultats les plus performants. Elle est définie par :

$$\mathrm{SMAD}(\mathbf{f}_g, \mathbf{f}_d) = \sum_{k=0}^{h-1} \left(|\mathbf{f}_g - \mathbf{f}_d - \mathrm{med}(\mathbf{f}_g - \mathbf{f}_d)|^2 \right)_{k:N_f-1} . \tag{B.10}$$

Les valeurs de cette mesure de dissimilarité appartiennent à l'intervalle $[0; {\mathrm{I}_{\max}}^2 h]$ [1].

1. Nous avons choisi $h = 50\%$ dans toutes les évaluations, suite aux résultats exposés dans [Chambon 05a, Chap. 3]

B.1.6. Information mutuelle

L'information mutuelle ou *Mutual Information*, MI, est également utilisée dans la littérature, en particulier pour les approches globales, puisqu'elle est robuste aux distorsions radiométriques. Il est important de noter que dans cas, dans ces méthodes globales, il n'y a pas de zones d'agrégation mais il s'agit d'approche point par point. Dans la littérature, il est mis en avant que le fait de se passer de zones d'agrégation doit permettre de ne pas générer d'artefacts au niveau des zones de discontinuités. La difficulté majeure de la méthode MI est la nécessité d'avoir une carte de disparités initiale. Mais pour résoudre cette difficulté, il est classique de prendre en compte, dans un premier temps, les résultats obtenus par une approche classique puis d'effectuer un processus itératif de mise à jour, à l'aide de l'information mutuelle. Afin de comparer nos travaux à ce type d'approche, nous avons choisi d'utiliser l'approche de [Hirschmüller 08].

B.2. Représentation de la couleur

De nombreux travaux portent sur le processus de formation des couleurs et le système de perception des couleurs de la vision humaine, cf. [Levine 85, chap. 3, 4 et 7], [Carron 95, chap. 1], [Brun 96, chap. 1], [Hardeberg 99, chap. 1] et [Chambah 01, chap. 2], mais, ici, seuls les systèmes de représentation de la couleur seront abordés. Les termes utilisés sont :

- *La luminance* – Il s'agit de la notion d'émettre plus ou moins de lumière.
- *La teinte* – Elle correspond à une couleur, une longueur d'onde dominante.
- *La saturation* – Elle indique le niveau de coloration d'une teinte, la pureté de la couleur.

La teinte et la saturation définissent la chrominance d'une couleur. Un système de représentation de la couleur est un système de coordonnées qui permet de représenter la couleur. Ce système, d'après la CIE (Commission Internationale de l'Éclairage), doit posséder trois composantes par analogie au système visuel humain qui comporte trois cônes différents, réceptifs à trois longueurs d'ondes différentes. Une couleur est représentée par les valeurs de ces trois composantes.

Cinq catégories de systèmes peuvent être distinguées [Vandenbroucke 00, chap. 2] :

- les systèmes de primaires ;
- les systèmes luminance-chrominance ;
- les systèmes perceptuels ;
- les systèmes d'axes indépendants ;
- les systèmes hybrides.

Les quatre premières catégories sont classées dans les systèmes classiques, par opposition aux systèmes hybrides. Nous allons présenter les systèmes les plus couramment utilisés pour chaque catégorie.

B.2.1. Systèmes de primaires

Ces systèmes s'appuient sur l'utilisation de trois couleurs fondamentales (les primaires). Ils se différencient par le choix des primaires et du blanc de référence (codage du blanc). Le système le plus courant est RGB (*Red Green Blue*) ou RVB (Rouge Vert Bleu), introduit par la CIE. Les composantes couleur de ce système sont définies par $R \in [R_{min}; R_{max}]$, $G \in [G_{min}; G_{max}]$ et $B \in [B_{min}; B_{max}]$. Pour modéliser certaines couleurs du domaine visible, la composante R doit parfois être négative, cf. [Brun 96, p. 15], or, généralement, les intervalles de variation sont $R \in [0; 255]$, $G \in [0; 255]$ et $B \in [0; 255]$. Ce qui signifie que certaines couleurs ne peuvent pas être modélisées. Il y a autant de systèmes RGB que de blancs de référence, cf. [Vandenbroucke 00, p. 195] et [Trémeau 04, p. 82].

Un autre système a été proposé par la CIE, cf. [Cie 15.2 86], il s'agit de XYZ. Ce système a été proposé pour prendre en compte le problème des valeurs négatives de la composante R du système RGB. Chaque composante de ce système est une combinaison linéaire des composantes RGB. Pour chaque système RGB, une matrice de passage a donc été définie pour passer des coordonnées R, G

et B aux coordonnées X, Y et Z, cf. [Vandenbroucke 00, p. 195–197]. Ce système est rarement utilisé directement car il sert plutôt de système de transition pour passer d'un système RGB vers un autre système. Ici, nous indiquons la transformation la plus couramment employée, cf. [Vandenbroucke 00, p. 196] et [Trémeau 04, p. 90] :

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.607 & 0.174 & 0.200 \\ 0.299 & 0.587 & 0.114 \\ 0.000 & 0.066 & 1.116 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \tag{B.11}$$

Lorsque les intervalles de variation du système RGB sont $R \in [0;255]$, $G \in [0;255]$ et $B \in [0;255]$, les intervalles de variation des composantes XYZ sont $X \in [0;250.16]$, $Y \in [0;255]$ et $Z \in [0;301.41]$. Les trois composantes de ces systèmes sont liées à la luminance. Or, deux couleurs peuvent avoir le même caractère chromatique avec des luminances différentes. Pour ne tenir compte que de la chrominance, les coordonnées peuvent être normalisées de la façon suivante, cf. [Vandenbroucke 00, p. 52–56] :

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B} \quad \text{et} \quad b = \frac{B}{R+G+B}. \tag{B.12}$$

$$x = \frac{X}{X+Y+Z}, \quad y = \frac{Y}{X+Y+Z} \quad \text{et} \quad z = \frac{Z}{X+Y+Z}. \tag{B.13}$$

Cette normalisation permet d'obtenir des coordonnées comprises entre 0 et 1 et comme $x+y+z=1$, la couleur peut être représentée dans un plan. Ces nouvelles coordonnées sont indépendantes des variations scalaires de la luminance de la couleur étudiée. En effet, si chaque composante R, G et B est multipliée par le même scalaire, alors les valeurs de r, g et b ne changent pas. L'inconvénient de cette normalisation est qu'elle n'est pas définie pour $R=G=B=0$. Dans ce cas particulier, une solution est de prendre $r=g=b=\frac{1}{3}$, cf. [Garbay 79].

Enfin, le dernier système de primaires, CMY (*Cyan Magenta Yellow*), utilisé dans [Gevers 98, Trémeau 03], s'appuie sur une synthèse soustractive et est utilisé pour la télévision et les imprimantes. Une quatrième composante peut être ajoutée (composante « noire »). La transformation suivante peut être utilisée, cf. [Trémeau 04, p. 89] :

$$\begin{pmatrix} C \\ M \\ Y \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \tag{B.14}$$

Les intervalles de variation sont $C \in [0;510]$, $M \in [0;510]$ et $Y \in [0;510]$.

Un des inconvénients d'un système de synthèse additive est qu'il ne permet pas de modéliser certaines couleurs (certaines nuances de vert s'obtiennent en ajoutant des lumières bleues et vertes et en retranchant du rouge, cf. [Brun 96, p. 15] or les couleurs sont représentées avec des valeurs positives). Dans de nombreuses applications en imagerie, comme l'affichage, l'impression, la restauration d'images, etc., il est souhaitable que les propriétés d'un système de représentation de la couleur se rapprochent le plus possible de celles de la vision humaine. Si nous notons D, la distance qui sépare deux couleurs dans un système donné et si nous appelons différence, la différence entre deux couleurs perçue par l'œil humain, un système qui se rapproche d'un système visuel humain doit posséder la propriété (P) :

> $\forall\ \mathbf{c}_1,\ \mathbf{c}_2,\ \mathbf{c}_3$, trois couleurs telles que la différence de couleur perçue entre $\mathbf{c}_1$ et $\mathbf{c}_3$ est la même que celle perçue entre $\mathbf{c}_2$ et $\mathbf{c}_3$ alors $\mathrm{D}(\mathbf{c}_1,\mathbf{c}_3) = \mathrm{D}(\mathbf{c}_2,\mathbf{c}_3)$.

Tous ces systèmes de primaires ne possèdent pas de distance D qui vérifie la propriété (P) et constituent donc une approche limitée pour ces applications. C'est pourquoi d'autres systèmes possédant cette propriété (P) ont été proposés, comme les systèmes luminance-chrominance.

B.2.2. Systèmes luminance-chrominance

Parmi les systèmes luminance-chrominance, quatre types sont différenciés :
- les systèmes perceptuellement uniformes (une distance D vérifiant (P) peut leur être associée) ;
- les systèmes de télévision que nous ne détaillerons pas ici puisqu'ils sont particulièrement dédiés à la télévision et sont largement abordés dans [Vandenbroucke 00, p. 64–67] ;
- les systèmes antagonistes ;
- les autres systèmes.

Tous ces systèmes possèdent une composante de luminance et deux composantes de chrominance.

Systèmes perceptuellement uniformes

Parmi les systèmes perceptuellement uniformes, deux systèmes sont proposés par la CIE : CIELUV noté aussi Luv et CIELAB noté aussi Lab, cf. [Trémeau 04, p. 95–102]. Ces deux systèmes se déduisent de XYZ par les transformations suivantes :

$$L = \begin{cases} 116\left(\frac{Y}{Y_b}\right)^{\frac{1}{3}} - 16 & \text{si } \frac{Y}{Y_b} > 0.01 \\ 903.3\,\frac{Y}{Y_b} & \text{sinon,} \end{cases} \tag{B.15}$$

$$u = 13L(u' - u'_b) \text{ avec } u' = \frac{4X}{X + 15Y + 3Z}, \tag{B.16}$$

$$v = 13L(v' - v'_b) \text{ avec } v' = \frac{9Y}{X + 15Y + 3Z}, \tag{B.17}$$

$$a = 500\left(f\left(\frac{X}{X_b}\right) - f\left(\frac{Y}{Y_b}\right)\right), \tag{B.18}$$

$$b = 200\left(f\left(\frac{Y}{Y_b}\right) - f\left(\frac{Z}{Z_b}\right)\right), \tag{B.19}$$

$$\text{avec } f(x) = \begin{cases} x^{\frac{1}{3}} & \text{si } x > 0.008856 \\ 7.787x + \frac{16}{116} & \text{sinon.} \end{cases}$$

Les termes X_b, Y_b, Z_b, u'_b et v'_b sont à associer au blanc de référence. Généralement, le blanc de référence est codé $(255\ 255\ 255)^T$, en RGB. Ainsi en utilisant l'équation (B.11), nous obtenons : $X_b = 250.155$, $Y_w = 255$ et $Z_w = 301.41$. Les intervalles de variation sont donc : $L \in [0; 100]$, $u \in [-131.95; 220.8]$, $v \in [-139.05; 121.47]$, $a \in [-137.72; 96.14]$ et $b \in [-99.23; 115.65]$.

Systèmes antagonistes

Les systèmes antagonistes de [Faugeras 79] ont pour but de représenter le mieux possible la perception visuelle humaine. En s'appuyant sur des études physiologiques montrant que l'activité ou la réponse des cônes récepteurs de la rétine humaine, notés LMS (*Long Medium Short*) est proportionnelle non pas à l'intensité du stimulus mais à son logarithme, Faugeras utilise le logarithme des trois signaux L, M et S. Un système semblable s'appuyant sur le système RGB a aussi été proposé par [Garbay 81] :

$$A = \frac{1}{3}(\log(R) + \log(G) + \log(B)), \tag{B.20}$$

$$C_1 = \frac{\sqrt{3}}{2}(\log(R) - \log(G)) \text{ et} \tag{B.21}$$

$$C_2 = \log(B) - \frac{\log(R) + \log(G)}{2}. \tag{B.22}$$

Les intervalles de variation sont donc $A \in [0; \log(255)]$, $C_1 \in [-\frac{\sqrt{3}}{2}\log(255); \frac{\sqrt{3}}{2}\log(255)]$ et $C_2 \in [-\log(255); \log(255)]$. Même si la modélisation de la non linéarité de la réponse des cônes de la rétine humaine n'est pas conservée, généralement, le système de [Swain 91] est préféré car il n'utilise pas le logarithme :

$$A = \frac{R+G+B}{3}, \tag{B.23}$$

$$C_1 = \frac{\sqrt{3}}{2}(R-G) \text{ et} \tag{B.24}$$

$$C_2 = B - \frac{R+G}{2}. \tag{B.25}$$

Les intervalles de variations sont donc : $A \in [0; 255]$, $C_1 \in \left[-\frac{\sqrt{3}}{2}255; \frac{\sqrt{3}}{2}255\right]$ et $C_2 \in [-255, 255]$.

Autres systèmes

Enfin, d'autres systèmes semblables ont été proposés par [Simonetto 05], ou [Carron 95, p. 17]. Nous présentons le système le plus couramment utilisé, celui de Carron :

$$Y = \frac{R+G+B}{3}, \tag{B.26}$$

$$Ch_1 = R - \frac{G+B}{2} \text{ et} \tag{B.27}$$

$$Ch_2 = \frac{\sqrt{3}}{2}(B-G). \tag{B.28}$$

Ainsi, les intervalles de variation sont $Y \in [0; 255]$, $Ch_1 \in [-255; 255]$ et $Ch_2 \in \left[-\frac{\sqrt{3}}{2}255; \frac{\sqrt{3}}{2}255\right]$.

B.2.3. Systèmes perceptuels

Les systèmes perceptuels distinguent trois grandeurs : la luminance, la teinte et la saturation. Ils sont notés HSI (*Hue Saturation Intensity*). De nombreuses propositions ayant été faites pour le calcul de ces trois composantes, cf. [Hanbury 03, Vandenbroucke 00, p. 74–82], nous ne présentons que les équations les plus utilisées :

$$H = \begin{cases} \pi & \text{si } R = G = B \\ \arccos \dfrac{\frac{1}{2}((R-G)+(R-B))}{\sqrt{(R-G)^2+(R-B)(G-B)}} & \text{si } B \leq G \\ 2\pi - \arccos \dfrac{\frac{1}{2}((R-G)+(R-B))}{\sqrt{(R-G)^2+(R-B)(G-B)}} & \text{sinon,} \end{cases} \tag{B.29}$$

$$S = \begin{cases} 0 & \text{si } R = G = B = 0 \\ 1 - \dfrac{3\min(R,G,B)}{R+G+B} & \text{sinon} \end{cases} \text{ et} \tag{B.30}$$

$$I = \frac{R+G+B}{3}. \tag{B.31}$$

Avec ces équations, les intervalles de variation sont $H \in [0; 2\pi]$, $S \in [0; 1]$ et $I \in [0; 255]$.

B.2.4. Systèmes d'axes indépendants

Dans les systèmes de primaires, les trois composantes sont fortement corrélées car elles possèdent toutes une information commune qui est la luminance. Des systèmes d'axes indépendants ont été

proposés. Ils s'appuient sur l'utilisation de l'analyse en composantes principales qui permet d'obtenir des composantes décorrélées. Les deux systèmes les plus utilisés sont :

- Le système d' [Ohta 80] dont les composantes sont définies par :

$$I_1 = \frac{R+G+B}{3}, \tag{B.32}$$

$$I_2 = \frac{R-B}{2} \text{ et} \tag{B.33}$$

$$I_3 = \frac{2G-R-B}{4}, \tag{B.34}$$

  avec $I_1 \in [0;255]$, $I_2 \in [-\frac{255}{2};\frac{255}{2}]$ et $I_3 \in [-\frac{255}{2};\frac{255}{2}]$. Ce système est fondé sur une approximation de la transformation de Karhunen-Loeve, cf. [Levy 00], où I_1 est la composante la plus discriminante.
- Le système $H_1H_2H_3$ de [Braquelaire 97] dont les composantes sont définies par :

$$H_1 = R+G, \tag{B.35}$$

$$H_2 = R-G \text{ et} \tag{B.36}$$

$$H_3 = B - \frac{R+G}{2}, \tag{B.37}$$

  où $H_1 \in [0;510]$, $H_2 \in [-255;255]$ et $H_3 \in [-255;255]$. Dans [Levine 85, chap. 7], les auteurs montrent que les trois directions privilégiées par la vision humaine sont les directions : rouge-vert, bleu-jaune et blanc-noir. Ce système met en évidence ces trois directions privilégiées.

B.2.5. Systèmes hybrides

L'approche hybride de Vandenbroucke [Vandenbroucke 00, chap. 4] consiste à choisir, pour un ensemble d'images données, parmi toutes les composantes couleur de tous les systèmes de représentation de la couleur, les trois composantes les plus pertinentes pour une application donnée, en utilisant une technique d'apprentissage supervisé. Ces trois composantes forment le système hybride.

B.2.6. Synthèse des systèmes de représentation de la couleur

Parmi les systèmes qui viennent d'être brièvement présentés, nous reprenons les dix suivants (cf. tableau B.1 et B.2) : RGB, XYZ, CMY, Luv, Lab,AC_1C_2, YCh_1Ch_2, HSI, $I_1I_2I_3$ et $H_1H_2H_3$. Ce sont les plus représentatifs et les plus utilisés en traitement d'images. La remarque qui peut être faite est la suivante : les modèles luminance-chrominance Lab et Luv, et le modèle perceptuel HSI se rapprochent le plus de la vision humaine mais il reste à déterminer si cela a une influence sur les résultats obtenus avec des outils de détection de contours ou de segmentation tels que vus dans les chapitres précédents. Cela dépend du type d'images et de l'importance de la couleur dans les images étudiées ainsi que le traitement souhaité. Il faut également noter qu'utiliser la couleur, à la place des niveaux de gris, peut avoir un impact sur le temps d'exécution puisqu'il faut traiter trois fois plus de données et que, de plus, le passage vers un de ces systèmes de représentation est coûteux en temps de calcul.

B.3. Résultats liés à la mise en correspondance globale utilisant la couleur

La figure B.2 permet de présenter toutes les images testées ainsi que les cartes de distorsions radiométriques que nous avons estimées.

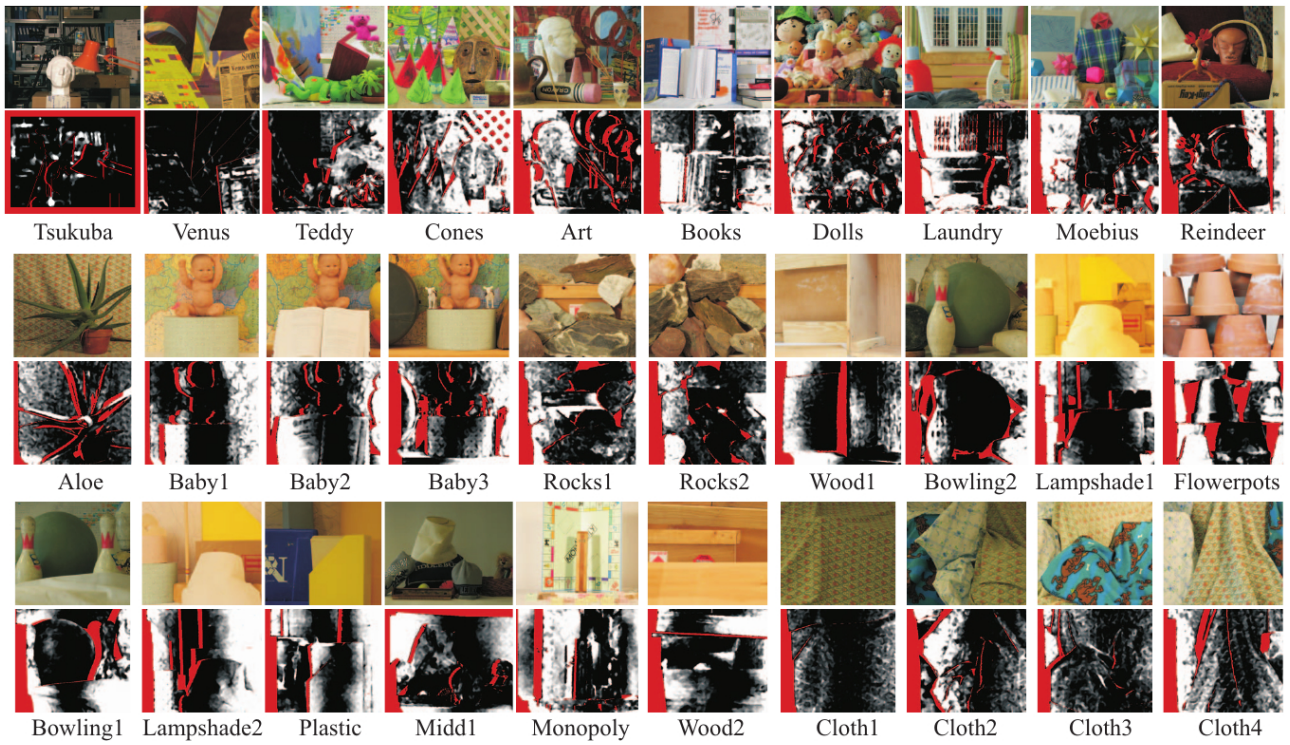
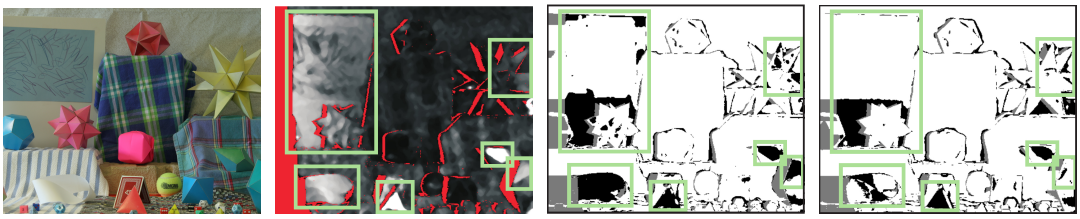


FIGURE B.2. – Tests pour les coûts globaux d'appariement – Voici les 30 images testées avec les cartes de distorsions radiométriques associées. Plus l'intensité est élevée et plus la distorsion radiométrique est élevée. Les pixels rouges correspondent aux pixels occultés.

Pour illustrer les différences de résultats obtenus avec différents systèmes de représentation de la couleur, nous reprenons l'exemple de l'image Moebius dans la figure B.3. Comme espéré, ce résultat nous permet de visualiser le gain de performance obtenu dans les zones de distorsions radiométriques, les rectangles en vert dans l'image, en particulier avec le système de représentation de la couleur qui permet les résultats les plus performants, le système *Luv*.



(a) (b) (c) (d)

FIGURE B.3. – Cartes des erreurs de disparités obtenues pour l'image Moebius – En (a), nous présentons l'image originale alors qu'en (b), il s'agit de la carte des zones de fortes distorsions radiométriques (les pixels clairs). En (c) le coût utilise les niveaux de gris alors qu'en (d) il s'agit du système *Luv*. Dans tous les cas, nous avons utilisé la norme L_1. Les erreurs de disparités correspondent aux pixels en noir (dans les zones non occultées) et en gris (dans les zones occultées). Nous remarquons que ces erreurs se produisent approximativement dans les mêmes zones de l'image, quelque soit le coût utilisé. Comme attendu, ces zones semblent correspondre aux zones de fortes distorsions radiométriques montrées en (d), dans les rectangles verts. Enfin, les résultats obtenus avec les niveaux de gris sont visuellement de moins bonne qualité que ceux obtenus avec le système *Luv*.

Référence	Type	Définition des composantes	Intervalles
XYZ [Cie 15.2 86]	Système de primaires	$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.607 & 0.174 & 0.200 \\ 0.299 & 0.587 & 0.114 \\ 0.000 & 0.066 & 1.116 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	$X \in [0; 250.16]$ $Y \in [0; 255]$ $Z \in [0; 301.41]$
CMY [Trémeau 04, p. 89]		$\begin{pmatrix} C \\ M \\ Y \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	$C \in [0; 510]$ $M \in [0; 510]$ $Y \in [0; 510]$
Luv [Trémeau 04, p. 95–102]	Système perceptuellement uniforme	$L = \begin{cases} 116 \left(\frac{Y}{Y_b}\right)^{\frac{1}{3}} - 16 & \text{si } \frac{Y}{Y_b} > 0.01 \\ 903.3 \frac{Y}{Y_b} & \text{sinon} \end{cases}$ $u = 13L(u' - u'_b)$ avec $u' = \frac{4X}{X+15Y+3Z}$ $v = 13L(v' - v'_b)$ avec $v' = \frac{9Y}{X+15Y+3Z}$	$L \in [0; 100]$ $u \in [-131.95; 220.8]$ $v \in [-139.05; 121.47]$
Lab [Trémeau 04, p. 95–102]		$a = 500 \left(f\left(\frac{X}{X_b}\right) - f\left(\frac{Y}{Y_b}\right)\right)$ et $b = 200 \left(f\left(\frac{Y}{Y_b}\right) - f\left(\frac{Z}{Z_b}\right)\right)$ avec $f(x) = \begin{cases} x^{\frac{1}{3}} & \text{si } x > 0.008856 \\ 7.787x + \frac{16}{116} & \text{sinon} \end{cases}$	$L \in [0; 100]$ $a \in [-137.72; 96.14]$ $b \in [-99.23; 115.65]$
AC_1C_2 [Swain 91]	Système antagoniste	$\begin{pmatrix} A \\ C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} & 0 \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	$A \in [0; 255]$ $C_1 \in \left[-\frac{\sqrt{3}}{2}255; \frac{\sqrt{3}}{2}255\right]$ $C_2 \in [-255; 255]$
YCh_1Ch_2 [Carron 95, p. 17]	Système luminance-chrominance	$\begin{pmatrix} Y \\ Ch_1 \\ Ch_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	$Y \in [0; 255]$ $Chr_1 \in [-255; 255]$ $Chr_2 \in \left[-\frac{255\sqrt{3}}{2}; \frac{255\sqrt{3}}{2}\right]$

Table B.1. – Systèmes les plus représentatifs parmi ceux abordés dans ce manuscrit.

Référence	Type	Définition des composantes	Intervalles
HSI [Hanbury 03]	Système perceptuel	$H_1 = \left(\frac{\frac{1}{2}((R-G)+(R-B))}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right)$ $H = \begin{cases} \pi & \text{si } R=G=B \\ \arccos H_1 & \text{si } B \leq G \\ 2\pi - \arccos H_1 & \text{sinon} \end{cases}$ $S = \begin{cases} 0 & \text{si } R=G=B=0 \\ 1 - \frac{\min(R,G,B)}{R+G+B} & \text{sinon} \end{cases}$ $I = \frac{R+G+B}{3}$	$I \in [0;255]$ $S \in [0;1]$ $H \in [0;2\pi]$
$I_1I_2I_3$ [Ohta 80]	Système d'axes indépendants	$\begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & -\frac{1}{2} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	$I_1 \in [0;255]$ $I_2 \in [-127.5;127.5]$ $I_3 \in [-127.5;127.5]$
$H_1H_2H_3$ [Braquelaire 97]		$\begin{pmatrix} H_1 \\ H_2 \\ H_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ -\frac{1}{2} & 0 & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	$H_1 \in [0;510]$ $H_2 \in [-255;255]$ $H_3 \in [-255;255]$

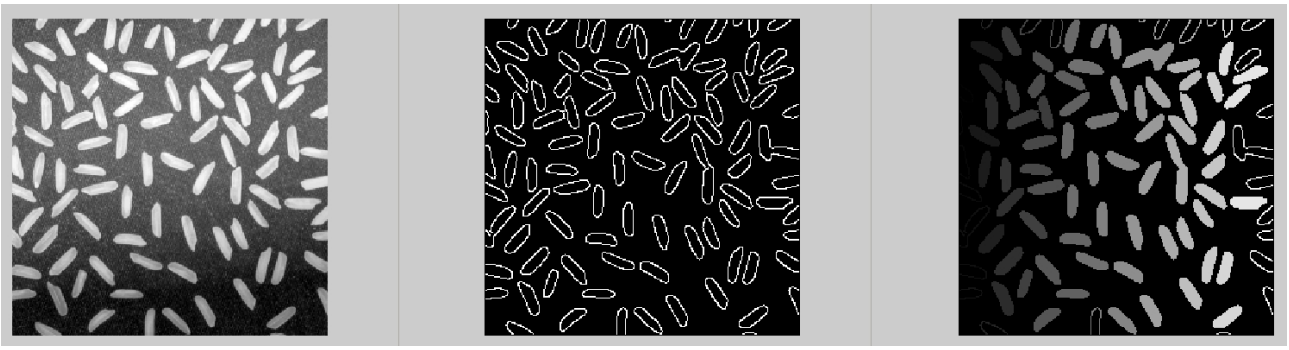
Table B.2. – Systèmes les plus représentatifs parmi ceux abordés dans ce manuscrit (suite).

Annexe C.

Méthodes de segmentation existantes

C.1. Approches par contours

Des approches classiques aux contours actifs – Les plus anciennes s'appuient sur des approches bas niveau de détection de contours, cf. figure C.1. Les techniques les plus avancées dans ce domaine concernent les contours actifs [Kass 88]. Ils sont très souvent employées en imagerie médicale où les contours sont très faiblement contrastées.



(a)



(b)

FIGURE C.1. – Approche classique par détection de contours puis fermeture (colonne du milieu). En (a), cette approche propose des résultats exploitables alors qu'en (b), nous illustrons la limite de cette approche. Il s'agit de résultats obtenus en séance de travaux pratiques dans le cadre de l'Unité d'Enseignement ≪ Image, Modélisation et Rendu ≫.

Principe – Un **contour actif ou *snake*** possède les caractéristiques suivantes :

- Il est caractérisé par des **énergies qui représentent ses propriétés physiques** ainsi que son **adéquation à la forme**.

- Son **énergie élastique** permet de préciser la configuration souhaitée, ce qui permet de prendre en compte une **information *a priori* sur la forme**.
- En lui attribuant une inertie, il acquiert un **comportement dynamique**, ce qui lui permet d'utiliser des **connaissances *a priori* sur le mouvement** de la forme dans une séquence d'images ou plus simplement dans la séquence de déformation.

Étapes des approches par contours actifs – Une approche par contour actif suit les étapes suivantes, en notant $\mathcal{C}$ la courbe active (fermée ou non) :

(1) Initialisation du contour ;

(2) Déformation par itérations successives jusqu'à convergence vers le contour réel.

La déformation s'effectue en étudiant l'évolution d'une énergie E qui représente la qualité de la courbe vis-à-vis du contour recherché. Les itérations successives sont généralement réalisées jusqu'à ce qu'il y ait peu de changements de l'énergie de la courbe.

La première étape de cet algorithme influence beaucoup la qualité du résultat. De plus, plus le contour est éloigné de celui recherché, plus l'algorithme sera coûteux.

Énergie à minimiser – Plus formellement, Le contour est représenté dans le plan par une forme paramétrique (fermée ou non), noté $v(s) = (x(s), y(s)), s \in [0, 1]$. Cette courbe est donc déformée de manière itérative afin de minimiser une fonction d'énergie E qui a pour forme générale :

$$E(\mathcal{C}) = \int_0^L \left(\underbrace{w_1 \|\mathcal{C}'(v)\|^2 + w_2 \|\mathcal{C}''(v)\|^2}_{\text{Forces internes}} + \underbrace{\mathcal{P}(\mathcal{C}(v))}_{\text{Force externe ou image}} \right) dv.$$

où

- $[0, L]$ est le domaine de définition de la courbe C ;
- L est la longueur de la courbe.

Les deux premiers termes correspondent aux **forces internes** avec w_1, w_2 les poids de chaque élément (dérivées première et seconde). Plus précisément w_1 représente le coefficient d'**élasticité** (résistance à l'allongement) de la courbe (cela pénalise la longueur de la courbe) et w_2 correspond au coefficient de **flexibilité** qui pénalise la courbure du contour. Ces deux termes agissent donc sur l'aspect régulier de la courbe. Là aussi, les choix pour w_1 et w_2 sont délicats et dépendent de la forme recherchée : plus ou moins long, plus ou moins rigide.

Le troisième terme correspond à la **force externe ou image** (*a priori* de forme) qui définit un potentiel d'attraction ou d'énergie d'attache aux données.

Choix possibles – Différents choix sont donc à réaliser pour mettre en œuvre une segmentation par contours actifs :

- **Choix pour $\mathcal{P}$**
  (1) *Basé niveaux de gris :* $\mathcal{P}(\mathcal{C}(v)) = I(v)$;
  (2) *Basé gradients:* $\mathcal{P}(\mathcal{C}(v) = \phi(\|\nabla I(v)\|)$ avec ϕ fonction décroissante.
- **Type de contours**
  - *Contours géométriques :* évolution planaire de la courbe
  - *Contours géodésiques :* cas particulier des contours actifs avec :

$$E(\mathcal{C}) = \int_\omega \left(\|\mathcal{C}'(v)\|^2 + \mathcal{P}(\mathcal{C}(v)) \right) dv$$

Forme discrète de l'énergie – Ceci est une définition **en continu**. Pour l'appliquer, il est nécessaire d'utiliser une **version discrète**. Pour cela, il faut utiliser des points de contrôle sur la courbe. Cela signifie qu'il faut choisir **un pas d'échantillonnage** et une forme qui peut être polygonale ou de type *spline*, par exemple. Dans ce cas, voici le type d'algorithme qui peut être mis en œuvre, cf. algorithme 5.

Algorithme basique de calcul de contours actifs – Pour faire évoluer la courbe, différentes méthodes d'optimisation sont possibles : descente de gradient, recuit simulé.

Algorithme 5 : Segmentation par contours actifs – La condition d'arrêt peut être le fait que l'énergie tend vers zéro. Pour faire évoluer le point d'énergie maximale, il est déplacé dans son voisinage en utilisant, par exemple, un algorithme de descente de gradient.

```
1 Fonction segmentation_par_contours_actifs (I, C)
  Entrées : C : le contour initial avec ces points de contrôle p_i
  Sorties : Image segmentée
2 Calculer l'énergie pour chaque point p_i de la courbe C
3 Ordonner la liste des points par énergies croissantes
4 tant que la condition d'arrêt n'est pas atteinte faire
5     Évoluer le point d'énergie maximale p_max d
6     Attribuer la nouvelle énergie à p_max
7     Si la distance est trop importante entre p_max et ses voisins, ajouter des points intermédiaires
```

C.2. Approches par régions

Nous distinguons trois catégories : par histogramme, par croissance de régions, par division/fusion.

C.2.1. Approches par histogramme

Ces approches sont les plus simples, les plus adaptées dans le cas à plusieurs dimensions (hyperspectrales, couleurs) mais malheureusement, elles sont souvent sensibles au bruit. Le principe est d'isoler des pics de l'histogramme et d'effectuer un seuillage des zones correspondantes. Dans le domaine de la segmentation, l'approche la plus utilisée pour déterminer les seuils de segmentation est l'approche par mélange de gaussiennes ou *Gaussian Mixture Model*, GMM, en utilisant l'algorithme d'Espérance-Maximisation, EM. Pour déterminer le seuil, il va falloir estimer les paramètres suivants : moyennes, variances et amplitudes de chaque distribution gaussienne.

L'histogramme d'une image bimodale peut être vue comme l'estimation de la fonction densité de probablilité de l'intensité : $p(x)$. Cette densité est la superposition de deux densités unimodales correspondant aux deux régions à segmenter : $p(x) = P_1 p_1(x) + P_2 p_2(z)$, avec p_1 et p_2, deux densités de probabilité gaussiennes, centrées respectivement en μ_1 et μ_2 (avec $\mu_1 < \mu_2$), et d'écart type respectif σ_1 et σ_2. Bien évidemment, nous avons $P_1 + P_2 = 1$. Un seuil S peut être choisi de manière à séparer les deux modes. L'erreur de classification faite sera alors, avec E_1, la probalilité de se tromper en classant 2 en 1, et *vice versa* : $E(S) = P_2 E_1(S) + P_1 E_2(S)$.

La minimisation de cette erreur est atteinte lorsque : $P_1 p_1(T) = P_2 p2_(T)$. Le résultat de cette équation passe par la résolution d'un polynôme du second degré.

Le résultat classique pour une même variance σ est : $S = \frac{x_1+x_2}{2} + \frac{\sigma^2}{x_1+x_2} \ln(\frac{P_2}{P_1})$.

Fusion/Division

Principe – Le principe est de diviser ou agréger les régions suivant un prédicat, de manière itérative jusqu'à une certaine convergence.

Algorithme Il s'écrit très simplement :

- Soit R_i, issue d'une première segmentation par division,
- Soit $P(R_i)$ un prédicat logique donnant `vrai` si la région satisfait un certain critère (homogénéité),

(1) Diviser en régions chaque R_i telle que $P(R_i)$ est `faux` ;

(2) Fusionner toutes régions R_j et R_k telles que $P(R_j \cup R_k)$ est `vrai` ;

(3) Arrêter dès que la fusion/division est impossible.

Le prédicat utilisé peut être par exemple le contraste de la région qui s'appuie souvent sur une analyse de l'écart-type ou de l'entropie de la région (représente la quantité d'information contenue dans la région).

Détails Les points à déterminer pour cette approche sont :

- La manière d'enchaîner les divisions/fusions (parfois toutes les divisions puis, toutes les fusions, ou fusions/divisions simultanées) ;
- La manière de diviser les régions (souvent en 4) ;
- Le prédicat de fusion/division. Voici des exemples de prédicat :

(1) Le **contraste** : $P(R_i) = |\max(R_i) - \min(R_i)| < S$;

(2) Les **différences limitées** : $\forall (x, y)$ pixels voisins $|I(x) - I(y)| < S$;

(3) L'**écart-type**, la **moyenne**, le **nombre de pixels**, le **contour/périmètre** ...

C.2.2. Croissance de régions

Ce type d'approche est souvent utilisée en imagerie médicale où le praticien a besoin d'un outil d'aide et non un outil totalement automatique. En effet, il s'agit souvent d'approches semi-automatiques puisque des points germes sont fournis pour permettre la croissance.

Principe Le principe est le suivant :

(1) Nous possédons des amorces : « cœur des régions ».

(2) Pour chaque amorce **a** de coordonnées (x_A, y_A) :

Propager l'amorce en agglomérant tous les pixels **p** tels que : $|I(x_A, y_A) - I(x, y)| < S$.

La propagation peut être séquentielle ou parallèle (pour ne pas privilégier une région). Le critère de propagation présenté est le plus simple mais il peut être adapté et prendre en considération d'autres critères : gradient, texture.

Détails Il y a différents points à à déterminer (en utilisant l'histogramme ou des *a priori*) :

- le choix des amorces (automatique ou non) ;
- le choix du seuil ;
- le choix de la propagation.

De plus, l'ordre dans lequel sont ajoutés les pixels dans une région a une influence sur le résultat. Toutefois, cette implémentation est relativement simple et les temps d'exécution sont rapides.

C.2.3. Ligne de partage des eaux

Cette approche a été souvent utilisée dans le domaine des cartes d'élévation numérique ou *Digital Elevation Models*. Une illustration de son comportement est donné dans la figure C.2.

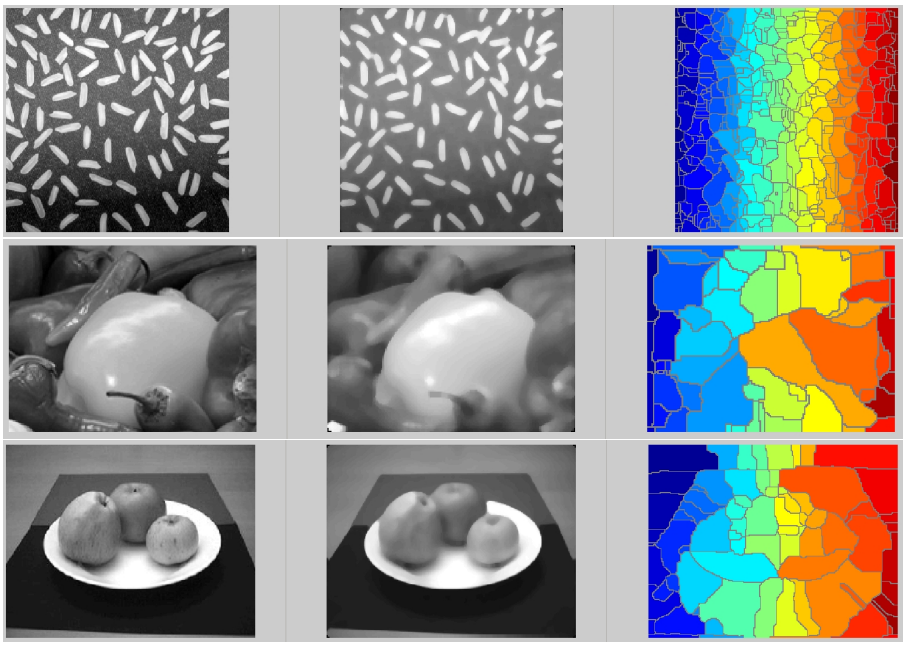


FIGURE C.2. – Approche par ligne de partage des eaux : résultats obtenus sur différents types d'images. Il s'agit de résultats obtenus en séance de travaux pratiques dans le cadre de l'Unité d'Enseignement « Image, Modélisation et Rendu ».

Principe L'algorithme de la ligne de partage des eaux (LPE), proposé par [Vincent 91], utilise la **description des images en termes géographiques**. En effet, une image peut être vue comme un relief, pour peu que le niveau de gris soit assimilé à une altitude.

Une ligne de partage des eaux est définie comme la **crête formant la limite entre deux bassins versants**. Pour l'obtenir, il faut imaginer l'immersion d'un relief dans l'eau, en précisant que l'eau ne peut pénétrer dans les vallées que par ses minima. La ligne de partage des eaux est représentée par les **points où deux lacs disjoints se rejoignent au cours de l'immersion**. Une des difficultés à la mise en œuvre de cette analogie intuitive est qu'elle laisse beaucoup de liberté quant à sa formalisation. De plus, l'efficacité des lignes de partage des eaux en tant qu'outil de segmentation **dépend grandement des marqueurs de départ**, c'est-à-dire des minima. Sans traitements préalables, le plus souvent, le résultat obtenu est une sur-segmentation de l'image. Une segmentation conforme au but recherché nécessite donc un filtrage adéquat des minima, qui constituent les marqueurs de départ.

Mise en œuvre Il est nécessaire :

- de définir une **grille** : pas de difficulté mais nécessité de définir le type de voisinage : 4/8 connexe, hexagones ;
- d'établir la notion de **chemin** : pas de difficulté mais cela dépend du voisinage choisi ;
- de choisir **minimum/plateau** : délicat car il s'agit surtout d'être capable d'identifier les critères adaptés au problème à traiter ;
- de définir comment construire les bassins, soit, **les lignes**. Il y a deux possibilités : par les chemins (difficiles à implémenter) ou par immersions.

C.3. Approches inspirées des approches de classification

C.3.1. k-moyennes

Principe L'algorithme des k-moyennes (ou *k-means*) est très connu en classification et très souvent utilisé pour la segmentation. Nous supposons **le nombre de classes connu** N. Nous cherchons donc la meilleure répartition des pixels en N classes.

Algorithme Il s'agit d'un **algorithme itératif** dont les principales étapes sont :

(1) Découper l'image en N régions suivant un critère c (à définir).

(2) Estimer les centres C_k, $k \in [1, N]$, de ces régions.

(3) Tant que les centres sont modifiés, faire :

  (a) Pour chaque pixel de l'image, trouver le centre C_{k^*} le plus proche, au sens du critère c, et donner au pixel considéré l'étiquette k^*.

  (b) Mettre à jour les centres C_k des N régions ainsi constituées.

Il n'existe pas de preuve de convergence de cet algorithme vers l'optimum global. D'ailleurs, le résultat dépend fortement de l'initialisation.

Les régions obtenues par cette technique ne sont pas toujours connexes, cela dépend de ce qui est pris en compte dans le calcul des centres (la couleur seule, la position seule, ou une combinaison des deux). Pour conclure, il existe des **variantes de cet algorithme**, où un aspect spatial peut être ajouté en ne prenant pas aléatoirement k couleurs, mais k **centres dans l'image**, cf. les différents résultats donnés dans la figure C.3. Plus précisément, une version plus subtile de cet algorithme est l'algorithme de *mean-shift* qui ne nécessite pas de connaître ou du moins de choisir k (méthode par minimisation d'une fonctionnelles, abordée au paragraphe suivant).

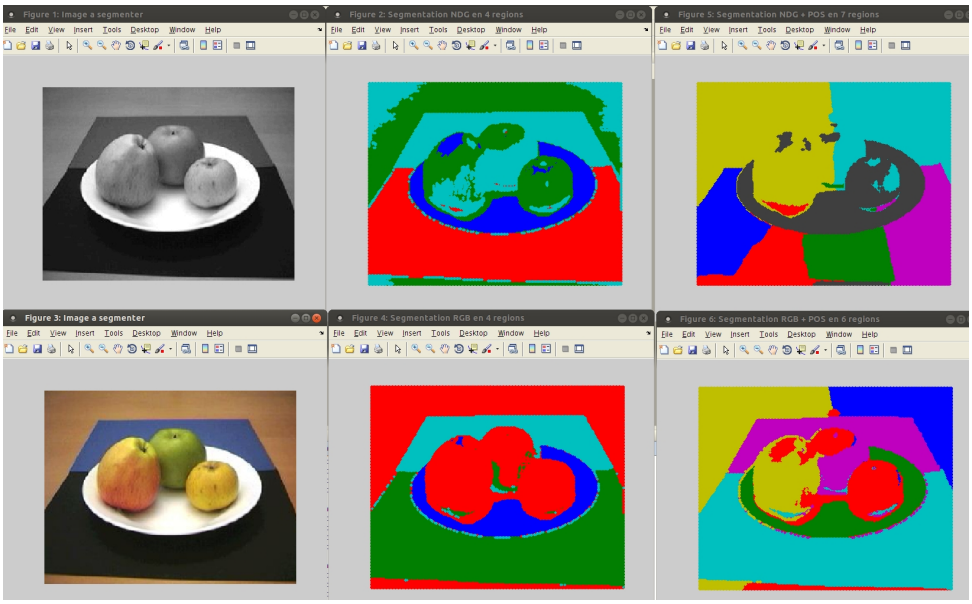


FIGURE C.3. – Approches par k-moyennes avec différents critères – Il s'agit de résultats obtenus en séance de travaux pratiques dans le cadre de l'Unité d'Enseignement « Image, Modélisation et Rendu ».

C.3.2. *Mean-Shift* [Comaniciu 02]

Historique Le *mean-shift* a été proposé en mathématiques pour **estimer les modes principaux d'une fonction**. Il a été repris par Comaniciu et Meer dans [Comaniciu 99, Comaniciu 02], qui ont montré l'intérêt d'appliquer cet outil à la segmentation d'images.

Principe Cette technique de segmentation est une des plus fréquemment utilisées dans la littérature. Elle présente le gros avantage d'être non supervisée, c'est-à-dire qu'il n'est pas utile de préciser à l'avance le nombre de régions recherchées. En pratique, nous disposons d'un nuage de points de $\mathbb{R}^d$ dont nous souhaitons estimer les modes, sachant que le *mode* d'un nuage de points correspond à un **maximum local de sa fonction de densité**. Une manière de trouver les modes consiste à « **suivre la direction du gradient** » de la fonction de densité.

Approche non-paramétrique Pour estimer les modes, dans le cas de la méthode *mean-shift*, une approche **non paramétrique** est utilisée, c'est-à-dire que nous ne faisons pas d'hypothèse sur le modèle des données. La solution pour ne faire aucune hypothèse consiste à utiliser une analyse locale par fenêtre, appelée *fenêtre de Parzen*. Soit un échantillon de n observations $\mathbf{x}_i$, $i = 1 \dots n$. Une estimation de la fonction de densité $\hat{f}(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^d$, est donnée par :

$$\hat{f}(\mathbf{x}) = \frac{1}{n\,h^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \tag{C.1}$$

où :

- Le *noyau* K doit posséder les propriétés suivantes : il doit être positif, borné, d'intégrale égale à 1 et à décroissance rapide.
- Le paramètre $h > 0$ permet de régler la taille du voisinage pris en compte. Ce paramètre a une influence sur le nombre de modes détectés. Plus h est petit, plus le nombre de modes détectés augmente. Cela permet de « démystifier » la segmentation non supervisée : au lieu de fixer *a priori* le nombre de régions, nous fixons un paramètre dont la valeur influe directement sur ce nombre.

Nous pouvons choisir le **noyau d'Epanechnikov** (également appelé *noyau parabolique*), qui est un noyau à support borné défini par :

$$K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) = \begin{cases} \dfrac{d+2}{2\,c_d}\left(1 - \dfrac{\|\mathbf{x} - \mathbf{x}_i\|^2}{h^2}\right) & \text{si } \|\mathbf{x} - \mathbf{x}_i\| < h \\ 0 & \text{sinon} \end{cases} \tag{C.2}$$

où c_d désigne le volume de la sphère unité en dimension d et $\|.\|$ désigne la distance euclidienne. L'avantage d'utiliser ce noyau est que l'estimation du gradient de la densité s'écrit alors comme suit :

$$\widehat{\nabla f}(\mathbf{x}) = \frac{1}{n\,h^d}\,\frac{d+2}{c_d\,h^2} \sum_{\mathbf{x}_i \in S_h(\mathbf{x})} (\mathbf{x}_i - \mathbf{x}) = \frac{1}{n\,h^d}\,\frac{d+2}{c_d\,h^2}\,n_{\mathbf{x}} \left[-\mathbf{x} + \frac{1}{n_{\mathbf{x}}} \sum_{\mathbf{x}_i \in S_h(\mathbf{x})} \mathbf{x}_i\right] \tag{C.3}$$

Dans cette expression, $S_h(\mathbf{x})$ désigne l'ensemble des $\mathbf{x}_i$ se trouvant dans la sphère (en dimension d) de rayon h, centrée en $\mathbf{x}$, et $n_{\mathbf{x}} = \#\{S_h(\mathbf{x})\}$.

Pour appliquer cette approche au cas de la segmentation d'images couleur, à chaque itération, il faut remplacer chaque $\mathbf{x}$ par $\mathbf{x} + \delta_{\mathbf{x}}\,\widehat{\nabla f}(\mathbf{x})$, où $\delta_{\mathbf{x}}\,\widehat{\nabla f}(\mathbf{x})$ constitue un « petit déplacement » dans la direction du gradient, pour se rapprocher du mode le plus proche. En choisissant $\delta_{\mathbf{x}} = \left[\frac{1}{n\,h^d}\,\frac{d+2}{c_d\,h^2}\,n_{\mathbf{x}}\right]^{-1}$,

cela revient d'après (C.3) à remplacer $\mathbf{x}$ par la moyenne $M_h(\mathbf{x})$:

$$M_h(\mathbf{x}) = \frac{1}{n_{\mathbf{x}}} \sum_{\mathbf{x}_i \in S_h(\mathbf{x})} \mathbf{x}_i \tag{C.4}$$

Voilà pourquoi cette méthode s'appelle *mean-shift*, qui signifie à peu près « remplacement par la moyenne ».

Étapes – L'algorithme *mean-shift* se décrit de manière détaillée comme dans l'algorithme 6.

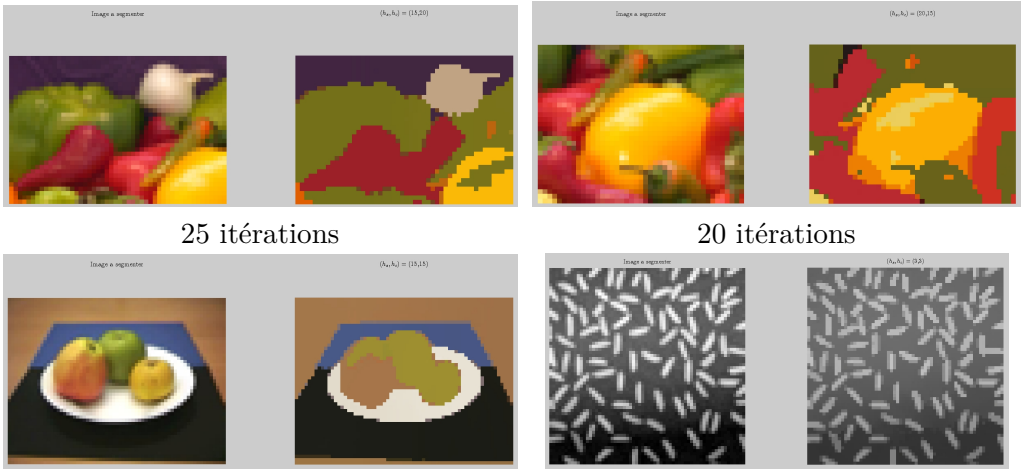
Algorithme 6 : Algorithme *mean-shift*.

```
Fonction estimation_modes (I, k)
  Entrées : k : nombre de classes
  Sorties : image segmentée
k ← 1
tant que ∃x, ‖x − M_h(x)‖ > ϵ et k ≤ k_max faire
    pour chaque x faire
        Calcul de la moyenne M_h(x) : cf. (C.4)
        x ← M_h(x)
    k ← k + 1
```

En pratique, pour déterminer $S_h(\mathbf{x})$, la couleur et la position sont traitées différemment : à l'intérieur d'une fenêtre de taille $(2\,h_s + 1) \times (2\,h_s + 1)$ centrée en $\mathbf{x}$, il faut considérer les pixels $\mathbf{x}_i$ ayant une couleur proche de celle de $\mathbf{x}$, c'est-à-dire tels que $\|I(\mathbf{x}_i) - I(\mathbf{x})\| \leq h_c$. Seule la couleur est modifiée au fil des itérations, ce qui signifie que $\mathbf{x}$, $\mathbf{x}_i$ et $M_h(\mathbf{x})$ sont des vecteurs de $\mathbb{R}^3$. L'algorithme dépend donc de quatre paramètres :

- h_s et h_c : seuil spatial et seuil colorimétrique qui définissent l'ensemble $S_h(\mathbf{x})$.
- ϵ et $k_{\max}$: paramètres qui permettent de contrôler l'arrêt de l'algorithme.

Une illustration du comportement de l'algorithme est donnée dans la figure C.4.



25 itérations — 20 itérations

23 itérations — 13 itérations

FIGURE C.4. – Application de l'algorithme de *Mean-Shift* sur différentes images.

C.3.3. *Support Vector Machine*

En apprentissage supervisé, les séparateurs à vaste marge ou machines à vecteur de support, *Support Vector Machine*, ont été introduits pour permettre la séparation entre différentes classes, c'est-à-dire la classification. Cette approche s'appuie sur deux concepts : la notion de marge maximale et la notion de fonction noyau, qui ne sont pas des notions récentes. Toutefois, leur utilisation conjointe comme fondement a été introduite seulement en 1992, par [Boser 92]. Depuis, cette technique a été largement employée et nous recommandons cet ouvrage pour en obtenir les détails les plus fins [Duda 01, chapitre 5, page 259].

Brièvement, l'intérêt de cette approche repose sur deux aspects pour réaliser la séparation des classes :

(1) La prise en compte d'une marge maximale entre la frontière de séparation, l'hyperplan, et les échantillons les plus proches, les vecteurs supports, contenus dans chaque classe ;

(2) La capacité à traiter les cas où la séparation n'est pas linéaire en transformant l'espace de représentation des données d'entrées en un espace de dimension plus grande. Cela veut dire que nous supposons qu'il existe une dimension pour laquelle la séparation devient linéaire.

Annexe D.

Présentation résumée de l'ensemble de la carrière

Pour une vision synthétique de l'ensemble de la carrière, une frise, figure D.1, est disponible en page 291.

D.1. Période présentée

Dans ce manuscript, j'ai cherché à synthétiser les travaux de recherche de **janvier 2006 à décembre 2019**, soit la période après la soutenance de thèse, en décembre 2005. Pendant cette période, j'ai été :

- **Depuis septembre 2011 : Enseignante-Chercheuse** à l'ENSEEIHT-INPT, École Nationale Supérieure en Electronique, Electrotechnique, Informatique et Télécom - Institut National Polytechnique de Toulouse, où j'ai été affectée à l'IRIT, équipe VORTEX, *Visual Object from Reality To EXpression*. Depuis janvier 2018, l'équipe VORTEX a été ré-organisée afin de permettre la création de deux nouvelles équipes, dont REVA, Réel Expression et Vie Artificielle, à laquelle j'appartiens.
- **Décembre 2007 à août 2011 : Chargée de Recherche** à l'IFSTTAR, Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux, sur le site de Nantes.
- **Septembre 2006 à novembre 2007 : Post-doctorante** à Télécom Paris, sous la direction d'Isabelle Bloch, dans le laboratoire LTCI, Laboratoire de Traitement et Communication de l'Information.
- **Septembre 2005 à août 2006 : ATER**, Attachée Temporaire d'Enseignement et de Recherche (en demi-poste), à l'Université Paul Sabatier, où j'ai travaillé avec Alain Crouzil, à l'IRIT, Institut de Recherche en Informatique de Toulouse.

D.2. Intéractions dans les différentes équipes de recherche

Mon parcours professionnel m'a permis de travailler avec de nombreuses personnes dont voici une liste que j'espère exhaustive :

- Depuis la prise de poste à l'ENSEEIHT (2011), je travaille essentiellement avec les collègues suivants : Vincent Charvillat, professeur, Pierre Gurdjos, ingénieur de recherche et Géraldine Morin, professeure.
- Pendant le poste de chargée de recherche à l'IFSTTAR (2007-2011), j'ai travaillé avec Christian Gourraud, Jean-Marc Moliard, Philippe Nicolle, techniciens supérieurs, et Vincent Baltzart, chargé de recherche. J'ai aussi travaillé avec Aurélien Cord, chargé de recherche sur le site de Paris. J'ai également pu travailler comme collaboratrice externe privilègiée avec Jérôme

Idier, directeur de recherhce CNRS, à l'IRCCYN, Institut de Recherche en Communications et Cybernétique de Nantes.
- Pendant le post-doctorat (2006-2007) à Télécom Paris, j'ai travaillé avec Elsa Angelini, MCF, Isabelle Bloch, professeure et Antonio Moreno, doctorant.

Tout le travail développé dans ce manuscrit découle, d'une part, d'un travail personnel et, d'autre part du travail en collaboration avec toutes ces personnes. J'espère ainsi transmettre au mieux le travail réalisé en respectant les efforts également fournis par mes collègues et co-auteurs.

D.3. Encadrements de recherche

J'espère que ce manuscrit respecte le travail fourni par tous les doctorants, post-doctorants et stagiaires de master 2 encadrés et co-encadrés.

D.3.1. Thèses soutenues

Tout ce travail a été développé au cours des directions et/ou encadrements de thèses de :

(1) Damien Mariyanayagam, à l'IRIT (équipe VORTEX/REVA), [Mariyanayagam 20]
- *Période* : novembre 2015 à mars 2020 (53 mois)
- *Financement* : Région Occitanie
- *Co-direction* : 40%, avec P. Gurdjos (40%) et V. Charvillat (20%)
- *Titre* : Utilisation d'un marqueur et de son environnement pour le suivi pour la réalité augmentée
- *Publications associées* : [Mariyanayagam 18, Mariyanayagam 19]
- *Actuellement* : ATER à l'ENSEEIHT

(2) Bastien Durix, à l'IRIT (équipe VORTEX), [Durix 17]
- *Période* : septembre 2014 à décembre 2017 (39 mois)
- *Financement* : MESR
- *Co-direction* : 40%, Géraldine Morin (60%)
- *Titre* : Squelettes pour la reconstruction 3D : de l'estimation de la projection du squelette dans une image 2D à la triangulation du squelette en 3D
- *Publications associées* : [Durix 15a, Durix 16a]
- *Après thèse* : Ingénieur de recherche chez `Fitting box`

(3) Marie-Anne Bauda, à l'IRIT (équipe VORTEX), [Bauda 16]
- *Période* : avril 2012 à juin 2016 (50 mois)
- *Financement* : ANRT (thèse cifre)
- *Co-direction* : 60%, Vincent Charvillat (20%) et Pierre Gurdjos (20%)
- *Titre* : Compréhension de scènes urbaines par combinaison d'information 2D/3D
- *Publications associées* : [Bauda 15b, Bauda 15a]
- *Actuellement* : Chef de projet R&D chez Akka Technologie

(4) Rabih Amhaz, à l'IFSTTAR, mais sur le site de l'IRIT-ENSEEIHT, [Amhaz 15]
- *Période* : septembre 2011 à mars 2015 (42 mois)
- *Financement* : Demi financement IFSTTAR [1] + demi financement région Pays de la Loire
- *Co-encadrement* : 40%, direction Jérôme Idier (40%) et co-encadrement Vincent Baltazart (20%)
- *Titre* : Détection de fissures par traitement d'images pour la surveillance des chaussées et des ouvrages d'art
- *Publications associées* : [Amhaz 14, Amhaz 16]
- *Actuellement* : Enseignant chercheur à l'ECAM Strasbourg-Europe

1. Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux, site de Nantes.

(5) Guillaume Galès, à l'IRIT (équipe TCI), [Gales 11]
- *Période* : septembre 2007 à juillet 2011 (46 mois dont 4 mois arrêt maladie)
- *Financement* : MESR
- *Co-encadrement* : 30%, direction Patrice Dalle et co-encadrement avec Alain Crouzil
- *Titre* : Mise en correspondance de pixels pour la stéréovision binoculaire par propagation d'appariements de points d'intérêt et sondage de régions
- *Publications associées* : [Gales 10b, Gales 10a]
- *Actuellement* : Ingénieur de recherche chez Foundry, Irelande

D.3.2. Thèse en cours

Actuellement, je co-encadre une thèse :
- Thierry Malon, à l'IRIT (équipe VORTEX/REVA)
  - *Période* : Depuis octobre 2017
  - *Financement* : Projet européen H2020
  - *Encadrement* : 40%, avec A. Crouzil (30%) et V. Charvillat (30%)
  - *Titre* : Reconstruction de mouvements articulés à partir de reconstruction 4D de scènes de crime
  - *Publication associée* : [Malon 18, Guyot 19, Malon 19b, Malon 20]

D.3.3. Encadrements de post-doctorants

Les travaux présentés sont aussi le résultat de l'encadrement de deux post-doctorats :

(1) Chafik Bakkay, à l'IRIT, équipe VORTEX
- *Période* : septembre 2016 à août 2017 (12 mois)
- *Financement* : Prestation pour un projet FUI, avec la société SiConsult
- *Encadrement* : 100%
- *Sujet* : Détection de papillons dans des images de pièges
- *Publications associées* : [Bakkay 18a, Bakkay 18b]
- *Actuellement* : Chercheur à l'IRT Saint-Exupéry

(2) Hatem Rashwan, à l'IRIT, équipe VORTEX
- *Période* : novembre 2014 à novembre 2016 (24 mois)
- *Financement* : Région Occitanie
- *Encadrement* : 60%, avec Pierre Gurdjos (20%), Géraldine Morin (10%) et Vincent Charvillat (10%)
- *Sujet* : Mise en correspondance 2D/3D
- *Publications associées* : [Rashwan 16, Rashwan 17, Rashwan 19a]
- *Actuellement* : Enseignant chercheur à l'Université de Tarragone, Espagne

D.3.4. Encadrements de stagiaires de master 2

Il est important de citer également tous les étudiants de Master 2 encadrés pendant cette période :
- Emilie Jalras
  - *Période* : du 30/03/2015 au 25/09/2015
  - *Financement* : équipe VORTEX
  - *Encadrement* : 100%
  - *Titre* : Mise en correspondance par corrélation avec fenêtres adaptatives basées superpixels.
- Lukasz Jasinski
  - *Période* : du 01/09/2014 au 01/06/2015
  - *Financement* : bourse erasmus
  - *Encadrement* : 100%

  - *Titre* : Parallélisation d'un algorithme de détection de fissures dans des images de chaussées
- Maya Alsheh Ali
  - *Période* : du 15/03/2011 au 15/09/2011
  - *Financement* : Bourse du gouvernement syrien
  - *Encadrement* : 50%, co-encadrement avec Jérôme Idier
  - *Titre* : Caractérisation de la texture des chaussées par un modèle auto-régressif non causal 2D (ARNC 2D)
- Hanan Salam
  - *Période* : du 01/02/2010 au 15/09/2010
  - *Financement* : IFSTTAR
  - *Encadrement* : 50%, co-encadrement avec Jérôme Idier
  - *Titre* : Détection automatique de fissures dans des images de chaussée par modélisation markovienne
- Noémie Vanetti
  - *Période* : du 01/02/2008 au 15/09/2008
  - *Financement* : IFSTTAR
  - *Encadrement* : 50%, co-encadrement avec Philippe Nicolle
  - *Titre* : Géolocalisation par photogrammétrie des désordres d'ouvrages d'art sur des images acquises par un drone d'hélicoptère
- Roberta Brocardo
  - *Période* : du 01/02/2007 au 15/09/2007
  - *Financement* : Télécom Paris
  - *Encadrement* : 50%, co-encadrement avec Isabelle Bloch
  - *Titre* : Introduction of a Dynamic Lung Model ina CT-PET Landmark-based Registration
- Guillaume Galès
  - *Période* : du 01/09/2006 au 01/06/2007
  - *Financement* : Aucun
  - *Encadrement* : 50%, co-encadrement avec Alain Crouzil
  - *Titre* : Détection et mise en correspondance de points d'intérêt pour la propagation de germes
- Benoît Bocquillon
  - *Période* : du 01/09/2003 au 01/06/2004
  - *Financement* : Aucun
  - *Encadrement* : 50%, co-encadrement avec Alain Crouzil
  - *Titre* : Obtention de la vérité terrain pour la mise en correspondance stéréoscopique

D.4. Collaborations

D.4.1. Collaborations internationales

- **Depuis janvier 2017 : Collaboration avec l'université de la Catalogne, Espagne (plusieurs séjours courts)** pour travailler avec Hatem Rashwan dans la poursuite des travaux de post-doctorat, sur l'amélioration d'une approche de mise en correspondance 2D/3D, cf. [Rashwan 19a]. Cette collaboration a permis le depôt d'un projet européen (rejeté) et un projet ANR (en cours).
- **Janvier 2016 : Invitation (4 semaines) à *National University of Singapour (NUS)*** de Wei Tsang Ooi pour visiter le *Multimedia Systems Lab*. L'objectif de cette mission a été de renforcer la collaboration existante entre les équipes REVA et le groupe de recherche à Singapour en abordant notamment les aspects vision par ordinateur étudiés à NUS. Cela a permis la prise de contact avec Terence Sim sur les aspects de reconnaissance de visages, et Michael Brown sur les aspects de cartes de focus (abordées/utilisées dans le cadre de la publication [Rashwan 19a]).

De plus, j'ai pu également aborder avec des chercheurs du laboratoire IPAL [2], des travaux sur les maisons intelligentes. Ainsi, j'ai co-encadré 2 stages de master 1 effectués en juin/juillet 2016 sur la mise en œuvre d'une approche pour faciliter l'utilisation de tests pour évaluer la santé mentale de patients.

- **Avril 2011 : Invitation (1 semaine) de l'Université de Maynooth (Irlande)** pour réaliser un séminaire sur mes travaux de recherche et pour travailler avec John McDonald et Guillaume Galès, sur la détection et l'appariement de points d'intérêt pour la mise en correspondance par propagation [Gales 12].
- **Août 2010 : Invitation (3 semaines) de l'équipe *Media processing*, Université de Vienne, Autriche**, pour travailler avec Michael Bleyer sur la mise en correspondance globale d'images couleur [Bleyer 10a].
- **2006 – 2007 : Collaboration avec l'Université de Floride (États-Unis)**, avec Anand Santhanam et Jannick Rolland, nous avons collaboré sur les travaux de recherche sur le recalage multimodal de volumes du poumon en prenant en compte un modèle de respiration [Chambon 07a, Chambon 08, Moreno-Ingelmo 08, Chambon 11e].

D.4.2. Collaborations nationales

- **2014 – 2017 :** Céline Roudet et Lionel Garnier, maîtres de conférences à l'Université de Bourgogne, Dijon, laboratoire Le2i [3]. Nous avons co-encadré la thèse de Bastien Durix sur la reconstruction 3D basé squelette et ces travaux ont donné lieu à des publications communes [Durix 15b, Durix 15a, Durix 16b, Durix 16a].
- **2011 – 2016 :** Vincent Baltazart, chargé de recherche IFSTTAR [4], Nantes, et Jérôme Idier, directeur de recherche CNRS, IRCCyN [5]. Nous avons co-encadré la thèse de Rabih Amhaz sur la détection de fissures sur chaussées et ouvrage d'art par une approche s'appuyant sur la recherche de chemins minimaux dans des images. Ces travaux ont été publiés dans [Amhaz 13b, Amhaz 13a, Amhaz 14, Amhaz 16].
- **2013 :** Stéphanie Prévost, maître de conférences, université de Reims. Nous avons mis en commun les travaux réalisés en mise en correspondance faisant intervenir plusieurs méthodes et plus particulièrement la propagation d'appariements fiables [Prévost 13b, Prévost 13a].
- **2007-2011 :** Aurélien Cord, chargé de recherche IFSTTAR[4], Paris. Nous avons collaboré sur la mise en œuvre d'une approche de détection de défauts de surface appliquées aux chaussées par apprentissage en utilisant une technique de type *AdaBoost* [Cord 11].
- **2010 :** Nicolas Coudray et Argyro Karathanou, post-doctorants à l'université de Haute-Alsace. Nous avons travaillé sur une approche par ligne de partage des eaux afin de détecter les fissures dans des images de chaussées [Coudray 10].
- **2007 :** Adrien Bartoli, professeur à l'université d'Auvergne. Nous avons collaboré sur une approche de reconstruction s'appuyant sur un modèle de type plaque mince ou *Thin-Plate Spline* (TPS). Les contributions portent sur la prise en compte de la contrainte épipolaire et la modélisation des déformations dues à la perspective [Bartoli 07, Bartoli 10].
- **2006 – 2007 :** Antonio Moreno, ingénieur de recherche INSERM, et Isabelle Bloch, professeur à Télécom Paris. Il s'agit des mêmes travaux que ceux partagés avec A. Santhanam et J. Rolland [Chambon 07a, Chambon 08, Moreno-Ingelmo 08, Chambon 11e].

2. Image & Pervasive Access Lab - Singapour, UMI CNRS 2955 - CNRS, National University of Singapore, A*STAR Institute for Infocomm Research, Université Joseph Fourier, Université Pierre et Marie Curie, Institut Mines-Telecom
3. Laboratoire Electronique, Informatique et Image.
4. Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux, site de Nantes.
5. Institut de Recherche en Communications et Cybernétique de Nantes.

D.4.3. Collaborations au sein du laboratoire IRIT

- **Depuis mai 2017 :** Julien Pinquier et Christine Senac, maîtres de conférences (équipe SAMOVA [6]), Florence Sedes, professeur, et André Peninou, maître de conférences, (équipe SIG [7]). Nous collaborons sur le traitement de données audios et vidéos issues de multiples caméras, via un projet européen appelé VICTORIA, *VIdeo analysis for investigations of Crimininal and TerrORIst Activities* [Malon 18, Guyot 19].
- **Depuis septembre 2006 :** Alain Crouzil, maître de conférences (équipe MINDS [8]). Nous avons co-encadré la thèse de Guillaume Galès sur la mise en correspondance dense robuste aux occultations. Nos travaux ont été publiés dans [Gales 10b, Gales 10a, Gales 10c, Gales 12]. Nous co-encadrons la thèse de Thierry Malon sur le suivi d'objets articulés dans des vidéos prises de différents points de vue, avec recouvrement, dans le cadre du projet VICTORIA. Ces travaux ont premis les publications suivantes [Malon 18, Malon 19a, Malon 19b, Malon 20].

D.5. Responsabilités scientifiques

D.5.1. Projets européens

- **2017 – 2020 : Participation au projet européen VICTORIA, *VIdeo analysis for investigations of Crimininal and TerrORIst Activities*, H2020**
  Ce projet porte sur l'analyse des contenus de audios et vidéos collectées par les agences de sécurité nationale et régionale, suite à un acte criminel ou à une attaque terroriste. Ce projet est le cadre d'une collaboration avec les enseignants-chercheurs des équipes MINDS, SAMOVA et SIG de l'IRIT, cf. les collaborations internes décrites précédemment. Ce projet permet aussi le financement de la thèse de Thierry Malon.
- **2016 – 2018 : Participation au projet européen LADIO, *Live Action Data Input/Output***
  Ce projet a permis en partie de financer les travaux réalisés avec Pierre Gurdjos sur la reconstruction 3D, notamment via la mise en correspondance de données 2D/3D, cf. [Rashwan 16, Rashwan 17, Rashwan 19a].
- **2011–2014 : Participation au projet européen « TRIMM – *Tomorrow's Road Infrastructure Monitoring and Management***
  Le porteur du projet était Dean Wright (Irelande). Au début, j'ai rédigé et pris la responsabilité d'un lot « Monitoring surface condition ». Ce projet a été soumis lorsque j'étais encore chargée de recherche à l'IFSTTAR, ainsi, une fois arrivée à l'ENSEEIHT dans mes nouvelles fonctions, je suis devenue consultante pour ce projet.

D.5.2. Implication internationale

- **2011 – 2019 : Relectures pour des revues internationales (plus de 65 papiers relus) :** Pattern Recognition (PR), IEEE Digital Signal Processing (DSP), International Journal of Remote Sensing (IJRS), Electronic Letters on Computer Vision and Image Analysis (ELCVIA), IET (Institution of Enginnering and Technology), parties *Image Processing* et *Computer Vision*, IEEE Digital Signal Processing, Biomedical Engineering (BME), International Journal of Remote Sensing (IJRS), Electronic Letters on Computer Vision and Image Analysis (ELCVIA), IET (Institution of Enginnering and Technology) Image Processing, Computer-Aided Civil and Infrastructure Engineering (CACIE), Transactions on Aerospace and Electronic Systems (T-AES).

6. Structuration, Analyse et Modélisation de documents Vidéo et Audio.
7. Systèmes d'Informations Généralisés.
8. coMputational imagINg anD viSion.

- **2014 – 2017 : Relectures pour des conférences internationales :** International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2020, European Signal Processing Conference (EUSIPCO) 2017, IEEE International Conference on Image Processing (ICIP) 2014, Scandinavian Conference on Image Analysis (SCIA), 2011, IEEE Computer Vision and Pattern Recognition (CVPR), 2010.
- **Octobre 2014 : Invitation pour participer à une session spéciale :** « *Detection of Road Surface Degradations* » pour *IEEE International Conference on Image Processing*, ICIP, Paris, 2014, organisée par Paulo Correia et Henrique Oliveira de l'Institut des Télécommunications, Lisbonne, Portugal.
- **2013 : Rapporteure externe** pour le stage de master 2 Recherche dirigé par Eric McClean, NUI Maynooth, Irlande.
- **2010 – 2011 : Présidente de session** aux conférences VISAPP.

D.5.3. Projets nationaux

- **2016 – 2017 : Responsable de la sous-traitance d'une partie d'un projet FUI VITI (du 01/09/2016 au 31/08/2017)**
  Le projet a pour but de diminuer l'utilisation des phytosanitaires pour la culture de la vigne. Pour cela, le projet souhaite introduire un dispositif d'acquisition d'images de pièges des parasites et les traiter automatiquement afin d'analyser l'évolution du parasite. Nous avons donc mis en œuvre un algorithme répondant à cette problématique. Cette sous-traitance a permis de financer le post-doctorat de Chafik Bakkay, cf. publication [Bakkay 18a, Bakkay 18b].
- **2015 – 2017 : Participation au projet région/FEDER [9] « INVISO » – Application de réalité augmentée pour l'usine du futur**
  Ce projet a permi le financement, en partie, de la thèse de Damien Mariyanayagam et est relatif aux publications [Mariyanayagam 17, Mariyanayagam 18, Mariyanayagam 19].
- **2014 – 2015 : Responsable scientifique d'un projet régional « MobVille – Application mobile pour la signalisation des alertes urbaines »**
  Ce projet était porté par la société `Imajing`. Nous avons développé une approche permettant l'appariement 2D/3D. Ce projet a financé le post-doctorat d'Hatem Rashwan, cf. publications [Rashwan 16, Rashwan 17, Rashwan 19a].
- **2012 – 2015 : Responsable d'une opération de recherche IFSTTAR 11R125**
  Avec Vincent Baltazart en tant que chargée de recherche IFSTTAR, nous avons pris en charge une opération sur la « Détection automatique de défauts par traitement du signal et de l'image – Application à la caractérisation de fissures surfaciques (débouchantes) et internes (non-débouchantes) dans les matériaux et structures du génie civil ». Après mon départ en septembre 2011, je suis restée **consultante** pour l'IFSTTAR (**convention de collaboration IRIT – IFSTTAR**). Cette opération a permi de financer pour moitié la thèse de Rabih Amhaz.

D.5.4. Implication nationale

- **2021 : Organisation d'une conférence nationale :** ORASIS.
- **2019 : Participation au comité d'organisation des GTMG :** Journées du Groupe de Travail en Modélisation Géométrique, https://gtmg2019.sciencesconf.org/.
- **2017 : Expertise pour l'ANRT [10]** pour une demande de thèse CIFRE.
- **2015 – 2020 : Relectures pour des conférences nationales** : ORASIS, 2015, 2017, 2019, GRETSI 2017.
- **2013–2019 : Comités de programme pour une conférence nationale :** ORASIS

9. Fonds Européen de Développement Régional.
10. Association Nationale de la Recherche et de la Technologie

- **2013 : Relectures pour une revue nationale** : Traitement du Signal, TS.
- **2010 : Expertise pour l'appel à projet ANR** TecSan, Technologies pour la Santé.
- **2010 – 2020 : Participation à des jurys de thèse (sans compter les thèses encadrées) en tant qu'examinatrice**
  - Thèse de Lamees Nasser Khalafallah Mahmoud sur « Une méthode de débruitage basée sur la méthode des dictionnaires pour une segmentation robuste de noyaux bruités et denses dans des images biologique 3D de microscopie », Université de la Sorbonne, juillet 2019.
  - Thèse de Salah Eddine Kabbour sur « Human Ear 3D Reconstruction Using Cell Phone Camera », Université de Rennes, juin 2019.
  - Thèse de Yizhen Lao « 3D Vision Geometry for Rolling Shutter Cameras », Université de Clermont Ferrand, mai 2019.
  - Thèse de Mathias Gallardo sur « Contributions to monocular deformable 3D reconstruction: curvilinear objects and multiple visual cues », Université de Clermont Ferrand, septembre 2018.
  - Thèse de Tien Sy Nguyen sur « Extraction de structures fines sur des images texturées : application à la détection à la détection automatique de fissures sur des images de surface de chaussées », Université d'Orléans, 2010.
- **2010 – 2020 : Comités de sélection pour des postes MCF**
  - Université de Bourgogne, Dijon, 2020
  - Université de Toulouse (ENSEEIHT), 2018
  - Université de Bourgogne, Le Creusot (annulé pour cause de suppression du poste), 2017
  - Université de Paris Sud, 2016
  - Université de Paris Sud, 2015
  - Université de Toulouse (ENSEEIHT), 2015
  - Université de Rouen (INSA), 2014
  - Université de Toulouse (ENSEEIHT), 2014
  - Université de Toulouse (ENSEEIHT), 2013
  - Université d'Auvergne (IUT), 2010
- **2018 – 2020 : Comissions interdisciplinaires (promotion locale):** pour l'INPT.

D.5.5. Vulgarisation scientifique

- **2021 :** Invitation à participer au café des Sciences au Quai des Savoirs (Toulouse) par l'association Femmes & Sciences. Cette participation consiste à présenter sa carrière et ses objectifs scientifiques auprès d'autres chercheuses dans divers domaines avant de constituer des binômes qui animeront un café ensemble. La difficulté de cette tâche est de faire une présentation sans support électronique pour vulgariser les recherches, puis, de répondre aux questions des personnes venues assister au café.
- **2016–2019 : Participation aux journées informatique et ingénierie des classes préparatoires aux grandes écoles** du lycée Déodat de Séverac à Toulouse. La participation consiste à assister aux démonstrations et présentations des projets d'ingénierie (en informatique/mathématiques appliquées) des étudiants des CPGE afin de leur faire un retour, puis, une présentation des travaux de recherche est réalisée afin de les sensibiliser au monde de la recherche et à l'utilisation des mathématiques en recherche en informatique.

D.6. Enseignement

Au cours de ma carrière, avant mon arrivée à l'ENSEEIHT, j'ai pu enseigner à des publics très variés : à l'université (Toulouse, Paris), dans des IUT (Saint Nazaire) ou dans des écoles (Polytech

Nantes et école centrale de Nantes), dans des filières variées : informatique, physique, génie civil. La plupart du temps, le public visé concernait des étudiants en licence. Le tableau D.1 résume la totalité de ces enseignements.

Matière	*Formation*	*Type*	*Quotat*
Algorithmique, programmation C	Licence 3	TD, TP	140
Base de données	Licence 3	Cours, TD, TP	42
Caml, OCaml	Licence 1 et 2	TP	189
Vision par ordinateur et traitement d'images	Master 1 et 2	Cours, TP	50
Prolog	Licence 3	TP	13
Bureautique	Licence 1	TP	16
		TOTAL	**450 éq. TD**

TABLE D.1. – Enseignements réalisés de 2002 à 2011 – J'ai eu l'envie et la chance de pouvoir enseigner dès mon année de DEA, en 2002. À chaque poste occupé, j'ai réalisé des enseignements : durant la thèse, en ATER, en post-doctorat et pendant mes années de chargé de recherche à l'IFSTTAR. Pendant ces années, j'ai suivi des enseignements déjà mis en place mais j'ai également proposé des sujets de travaux pratiques, de bureau d'études, des supports de cours et des travaux dirigés.

À partir du moment où j'ai enseigné à l'ENSEEIHT, j'ai réalisé des enseignements essentiellement pour la licence 3 et les masters 1 et 2. Toutefois, mon envie d'avoir des publics variés m'a amenée à enseigner à l'Université Jean Jaurès et également à la prépa des INP. Le tableau D.2 me permet de présenter les quotas moyens réalisés chaque année depuis mon arrivée, sachant que j'ai eu une demi charge la première année, deux tiers de charge la seconde année, et un temps complet jusqu'en 2016, date à laquelle, j'ai commencé un temps partiel (80%).

Matière	*Formation*	*Type*	*Quota/an*
Vision par ordinateur, traitement d'images	Info, Électronique 3	Cours/TP	45
HTML5/CSS/JavaScript	Info	Cours/TP	22
Architecture des ordinateurs/VHDL	Info 1 et 2	TD/TP	70
Ada	Info 1	TD/TP	68
Caml	Info 1	TD/TP	20
Probabilité et Statistiques	Info 1	TP	12
Python	Prépa INP 1	TP	30

TABLE D.2. – Enseignements réalisés depuis 2011 – Il s'agit des moyennes annuelles par enseignement réalisé, sachant que les matières fluctuent d'une année à l'autre ainsi que la charge : 96h en 2011, 144h en 2012, 192h avant 2016 et 154h depuis 2016.

D.7. Investissement pédagogique et administratif

- **2017 à 2020 : Adaptation d'une partie des enseignements en proposant des APP (Apprentissage par Projet et Problèmes)**. Pour cela, j'ai réalisé des demandes de financements (1000 euros de matériels interactifs + 15h équivalent TD) auprès de l'ENSEEIHT chaque année (demandes acceptées).
- **Mars 2016 : Participation à des MOOC :**
  - Enseigner avec Moodle,

  - Enseigner et Former dans le Supérieur.
- **Depuis 2015 : Responsable des relations internationales, pour le département Sciences du Numérique :** Cette responsabilité entraîne une décharge de 24h équivalent TD. De septembre 2015 à décembre 2016, cela a été une co-responsabilité avec l'ancienne responsable afin de réaliser le transfert des connaissances et d'assurer le service pendant mon congé maternité. Actuellement, cette responsabilité entraîne la gestion d'une soixantaine de demandes de départs à l'étranger (pour une moyenne de 35 départs effectifs) et d'une trentaine de demandes d'arrivées d'étudiants étrangers (au plus une dizaine seront acceptés).
- **Depuis 2015 : correspondante pour la plateforme d'apprentissage en ligne `moodle`** pour le département Sciences du Numérique. Cette responsabilité implique la création des nouveaux liens de cours, semestre, ... pour les enseignants du département mais également la formation pour promouvoir l'utilisation de cette plate-forme.
- **Janvier 2014 à décembre 2017 : Membre du conseil de département (Informatique et Mathématiques Appliquées).**
- **Depuis janvier 2014 : suivi des étudiants en Apprentissage :** suivi de 2 à 3 étudiants en formation par apprentissage, cela revient à 12h de décharge lorsque l'étudiant est en première année, puis 8h en deuxième et troisième année.
- **Depuis 2012 : Responsable de deux unités d'enseignement « Vision et Synthèse d'Images » et « Traitement de la vidéo », en troisième année ENSEEIHT.**
- Enfin, il est à noter que, suite à la demande de la direction, dans le cadre d'une **réforme de l'ENSEEIHT**, le département Informatique et Mathématiques Appliqués, auquel j'appartiens, a démarré une fusion avec le département Réseaux et Télécommunications. Le nouveau département a été mis en place pour la rentrée 2017-2018, seulement pour la première année. Cela signifie qu'il y a eu une forte participation (réunions, groupes de travail) pour réaliser cette fusion et, depuis la rentrée 2017–2018, une refonte/une adaptation des enseignements. Cette fusion s'est poursuivi sur les 2 années suivantes et cela a impliqué de nombreuses heures de travail pour fusionner également les relations internationales, dont j'ai pris la responsabilité en septembre 2015.

La figure D.1 permet de présenter une synthèse de la carrière. Au dessus de la frise principale, il s'agit des deux post-doctorants encadrés, et, au dessous de la frise principale, ce sont les six doctorants. En pervenche, j'indique tous les étudiants financés via un projet : MOBVILLE est un projet région, VITI, un projet FUI, VICTORIA un projet H2020, INVISO, un projet région et TRIMM un projet FUI. Pour MOBVILLE, j'étais représentante du projet, pour l'IRIT, alors que pour tous les autres, j'étais participante, à l'exception de VITI où il s'agissait d'une prestation. Ainsi, pour chaque étudiant, ceux indiqués en pervenche correspondent à ceux financés par un projet, celle indiquée en bleu pastel correspond à une bourse cifre, et enfin, ceux indiqués en vert pastel, sont ceux financés par une bourse MSER.

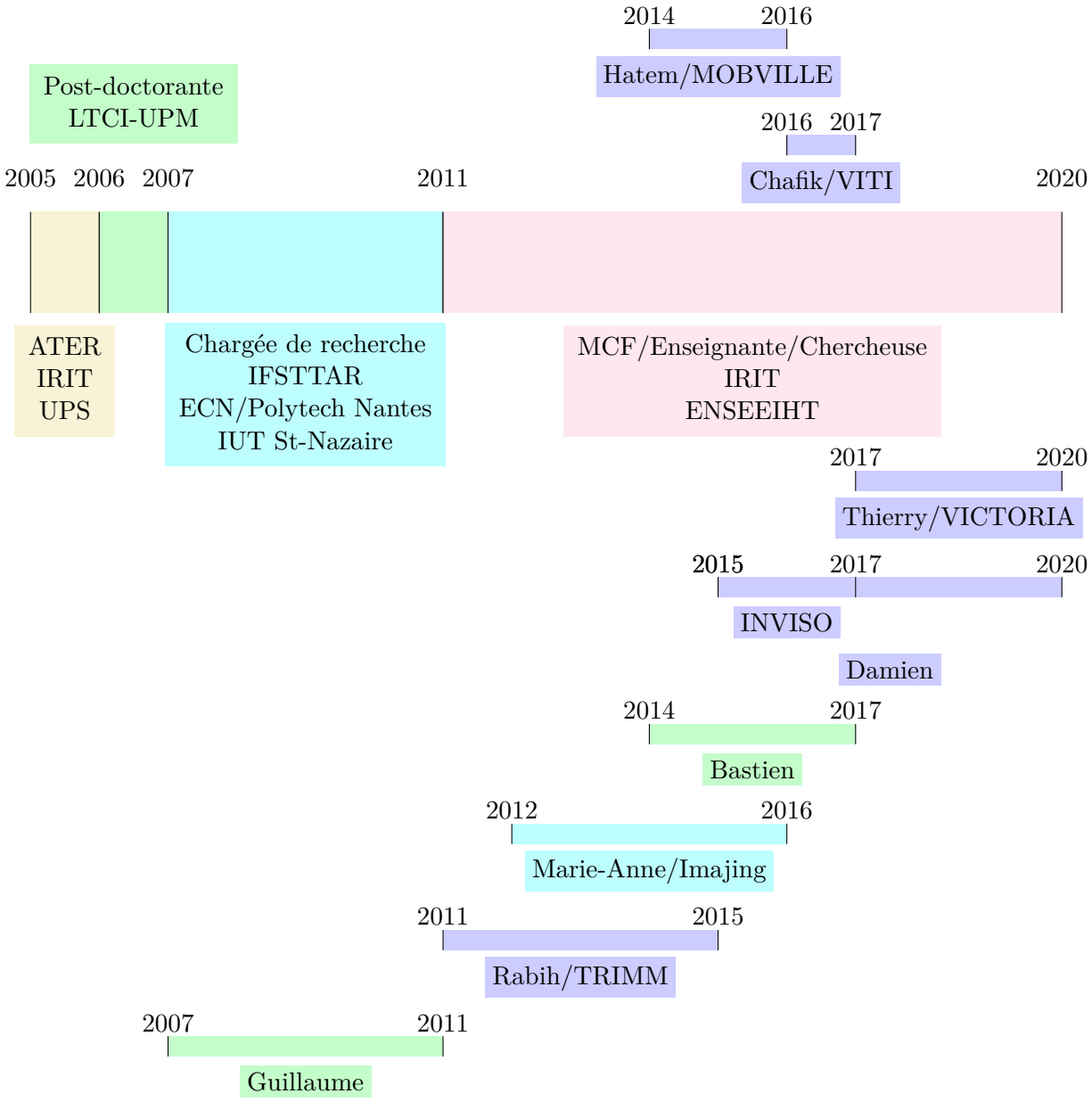


FIGURE D.1. – Carrière résumée en une frise.