Apprentissage statistique de classes sémantiques pour l'interprétation d'images aériennes

Hicham Randrianarivo

le cnam

ONERA
THE FRENCH AEROSPACE LAB

THÈSE DE DOCTORAT DU CONSERVATOIRE DES ARTS ET MÉTIERS

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Hicham Randrianarivo

Pour obtenir le grade de

DOCTEUR

Spécialité

Informatique

Apprentissage statistique de classes sémantiques pour l'interprétation d'images aériennes

---

Soutenue le 15 décembre 2016 devant un jury composé de :

Rapporteur	**Pr.Valérie Gouet-Brunet**	(IGN)
Rapporteur	**Pr. Mihai Datcu**	(DLR (Allemagne))
Examinateur	**Pr. Nicole Vincent**	(Université Paris-Descartes)
Examinateur	**Dr. David Picard**	(ENSEA)
Directeur de thèse	**Pr. Michel Crucianu**	(CNAM)
Encadrant	**Dr. Bertrand Le Saux**	(ONERA)
Encadrant	**Dr. Marin Ferecatu**	(CNAM)

Table des matières

1 Introduction

Sommaire

1.1 Problématique

L'observation de la Terre par des satellites ou des moyens aériens permet d'obtenir des renseignements essentiels sur l'occupation du sol, sur les océans ou sur l'atmosphère. Grâce à elle nous pouvons suivre et prédire l'évolution de la météorologie, gérer l'occupation des sols, détecter une pollution en mer ou dans l'atmosphère ou encore suivre les mouvements de population pour prévenir les épidémies. Les applications intéressent des acteurs de tous les domaines : militaires, civiles et commerciaux. La tendance des moyens d'observation est à une augmentation de la résolution des images (aujourd'hui couramment 0,5 m/pixel) et de la fréquence d'acquisition. Grâce à *Google Earth* tout le monde peut aujourd'hui observer n'importe quel endroit de la Terre avec une très grande précision.

Nous allons d'abord présenter certaines applications d'hier et d'aujourd'hui section 1.1.1 avant de détailler les enjeux de la télédétection section 1.1.2.

1.1.1 Quelques applications

Un peu d'histoire : L'observation de la terre depuis le ciel a des origines anciennes. Lors de la bataille de Fleurus en 1794 un ballon d'observation fût pour la première fois utilisé pour observer les positions ennemies. Les objectifs étaient tactiques mais les développements commerciaux suivirent. En 1858 Gaspard-Félix Tournachon (dit Nadar) monta dans un ballon dirigeable pour

photographier des quartiers de la ville de Paris (cf. figure 1.1). Le pionnier de la photographie fût donc aussi le pionnier de l'imagerie aérienne. Ces clichés sont entrés dans l'histoire de l'art.

FIGURE 1.1 – Nadar élevant la Photographie à la hauteur de l'Art., lithographie d'Honoré Daumier parue dans Le Boulevard, le 25 mai 1863.

Observation des océans : L'altimétrie radar fournit des données topographiques de précision sur la surface de la mer et ses variations ; celles-ci permettent, entre autres applications, de connaître la circulation océanique (courants marins). La mission satellitaire franco-américaine TOPEX/Poseidon est la plus connue : les altimètres de TOPEX/Poseidon mesurent le niveau de la mer, la hauteur des vagues et les vents sur les océans avec une répétitivité de dix jours. La position des satellites par rapport au centre de la Terre est connue avec une précision de 3 cm, ce qui permet des mesures de la topographie océanique extrêmement précises. La figure 1.2 illustre comment le satellite TOPEX/Poseidon à permit de mesurer la variation du phénomène El Niño en 1997.

Observations des milieux urbains : Au tournant de l'an 2000, le transfert des technologies de la télédétection militaire vers les applications civiles donne naissance à des satellites d'observation de la Terre à très haute résolution. Le satellite IKONOS, exploité par la société privée Space Imaging Corp., en est l'exemple le plus remarquable : il permet l'acquisition d'images à la résolution de 1 m en mode panchromatique (1 seule bande spectrale) et de 4 m en mode multispectral. La fusion des deux types de données fournit des images couleur dont les applications sont comparables à celles des photographies aériennes. La figure 1.3 nous montre un exemple d'une image capturée par le satellite IKONOS en Australie.

1.1.2 Enjeux actuel de l'observation de la Terre et de la télédétection

Aujourd'hui de nombreux enjeux sont liés à l'imagerie aérienne et satellitaire que ce soit dans la société, d'un point de vue scientifique et technologique ou encore stratégique.

D'un point de vue sociétal les enjeux sont tout d'abord militaires. Lors d'un conflit la capacité de pouvoir observer les mouvements des effectifs du camp opposé donne un avantage stratégique certain. Toujours dans le cadre de conflit armés, la télédétection permet de pouvoir observer sur le théâtre d'opérations les différents dommages occasionnés comme la destruction de bâtiments ou de cibles. Elle permet aussi de pouvoir vérifier la vérité de faits invérifiables autrement comme dans le cas où des groupes terroristes ont détruit des monuments historiques (exemple :à Nabû en Iraq [1]). Dans ce cas, L'United Nations Educational, Scientific and Cultural Organization

1. http://www.un.org/apps/newsFr/storyF.asp?NewsID=37450

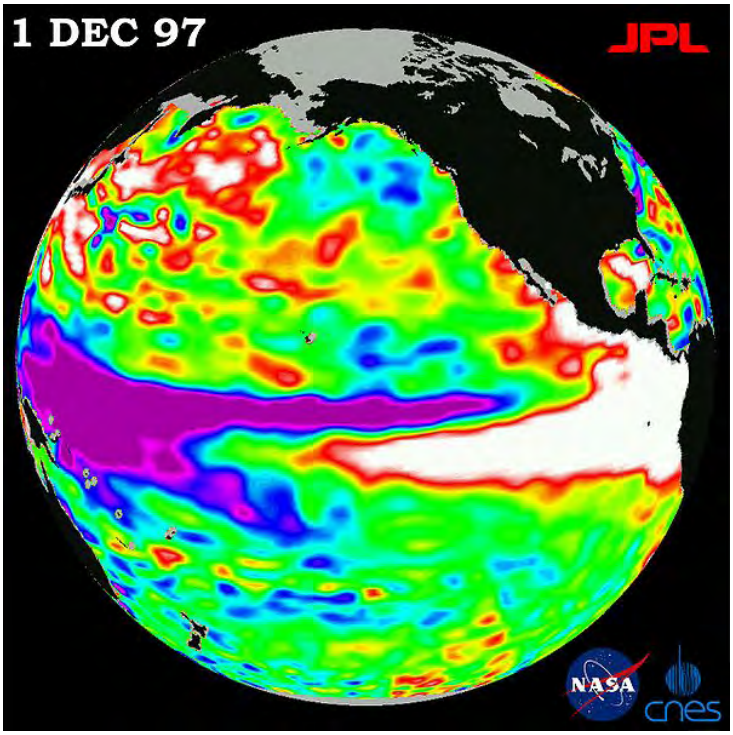


FIGURE 1.2 – Carte dressée avec les données fournies par Topex/Poseidon met en évidence les différences de hauteur de l'Océan Pacifique générées par le courant El Niño en 1997. Les zones en blanc reflètent un rehaussement par rapport au niveau moyen compris entre 14 à 32 cm tandis que les parties en violet traduisent des dépressions d'au moins 18 cm.

(UNESCO) s'est aidé d'images satellitaires et aériennes pour aider à confirmer la destruction des monuments. Mais les apports de la télédétection ne s'arrêtent pas là, elle permet aussi de mettre en oeuvre la surveillance de zones sensibles. Certaines zones comme les côtes peuvent être la cible d'actes de piraterie et il est crucial de pouvoir les surveiller pour prévenir ou réagir à une attaque.

En plus des enjeux militaires il existe aussi des enjeux humanitaires forts. La télédétection permet la mise en place en place de missions de sauvetage par les Organisations Non-Gouvernementales qu'il serait impossible d'organiser autrement d'un point de vue logistique et sécuritaire. Elle permet par exemple de ne pas installer un camp de réfugiés trop près des zones de conflit, de pouvoir suivre et modéliser les flux de réfugiés et prévoir où installer des camps. Grâce à la télédétection la recherche d'accès aux sources d'eau potable ou le suivi de la propagation d'une épidémie devient plus aisée.

D'un point de vue scientifique les enjeux de l'imagerie aérienne et satellitaire sont tout aussi importants. Une des sciences s'appuyant en grande partie sur l'imagerie satellitaire est la climatologie. Grâce aux données collectées par les satellites, nous sommes capables de prédire la météo et ainsi bénéficier d'informations cruciales pour les secteurs de l'économie météo-dépendants comme l'agriculture ou certains commerces. Les images satellitaires nous permettent aussi de pouvoir surveiller différents phénomènes naturels qui surviennent sur la planète comme la fonte des glaces, la dérive des icebergs ou l'évolution de la biomasse. La surveillance de ces phénomènes est d'une importance vitale pour comprendre et évaluer les conséquences de l'activité humaine sur la planète.

Grâce à la télédétection nous pouvons également suivre les différents événements sismiques sur la

FIGURE 1.3 – *Gold coast* en Australie observée par le satellite IKONOS.

planète et ainsi prévenir les risques liés aux tremblements de terres ou aux éruptions volcaniques. Nous pouvons suivre les évolutions des sols et sous-sols et ainsi définir les zones habitables ou les zones sujettes aux affaissements. La télédétection permet aussi l'observation des mers et des océans. Grâce à elle nous pouvons suivre l'évolution de phénomènes océaniques et leurs impacts sur le climat comme *El Niño* et *La Niña*. Avec elle nous pouvons cartographier les zones touchées par la pollution marine comme le *7e continent* ou continent de plastique dans l'océan pacifique [2]. Dans le domaine de l'urbanisme l'usage de l'imagerie aérienne et satellitaire est une ressource essentielle pour la gestion des espaces. Elle permet de cartographier l'occupation des terrains pour des espaces urbains et naturels mais a aussi des applications pour la gestion et l'analyse des flux de véhicules.

C'est pour cela que de nombreuses agences gouvernementales et des industries mettent en place de grands programmes pour la mise en orbite de satellites. Le satellite comme moyen d'acquisition de données est une technologie en pleine expansion à tel point que les images sont maintenant mises à disposition pour stimuler le développement de nouvelles les applications. Par exemple le programme d'observation de la Terre Copernicus mis en place par l'European Space Agency (ESA) donne libre accès à plusieurs giga-octets de données journalières. Ces nouveaux de moyens de distribution des images permettent d'alimenter l'offre en images satellitaires et ainsi créer de nouvelles demandes et applications que ce soit pour l'industrie ou pour la recherche.

Aujourd'hui les satellites peuvent prendre n'importe quel endroit de la planète à une fréquence élevée (1 image par jour) ce qui signifie qu'il faut pouvoir traiter et interpréter ces images rapidement. De plus la quantité d'information contenue dans ces images est très importante. Par exemple l'un des nombreux satellites d'observation de la Terre Sentinel-2 produit des images à une résolution de 20 m/pixel pour une largeur de bande d'acquisition de 290 km

2. http://www.lemonde.fr/planete/article/2012/05/09/le-7e-continent-de-plastique-ces-tourbillons-de-dechets-dans-les-oceans_1696072_3244.html

Pour répondre aux nouveaux challenges créés par l'imagerie aérienne et satellitaire, le besoin de pouvoir traiter ces données automatiquement est né. Il faut d'une part les traiter dans les délais les plus brefs possibles mais aussi avoir à disposition des méthodes pouvant tirer avantage de la masse de données disponibles. Les traitements sur lesquelles nous allons nous concentrer sont les traitements permettant de faire la classification des différentes régions de l'image. Il s'agit de pouvoir donner une étiquette à un ou plusieurs pixels de l'image uniquement à partir des informations extraites de l'image.

Les deux principales façons d'obtenir des images de télédétection sont l'imagerie satellitaire et l'imagerie aérienne. L'imagerie satellitaire se distingue de l'imagerie aérienne par l'étendue des zones vues par le satellite par rapport à un avion. L'imagerie aérienne a pour avantage une très grande résolution spatiale qui permet d'observer des objets avec beaucoup plus de détails que ne le ferait un satellite.

1.2 Imagerie satellitaire

L'imagerie satellitaire est le principal outil permettant l'observation de notre planète à large échelle. Ainsi un grand nombre d'entre eux gravitent autour de la Terre mais tous n'ont pas la même finalité ce qui influent fortement sur les images produites. Dans *Les Bases de l'Imagerie Satellitaire*, Nicolas classifie les satellites d'observation en deux grands ensembles, les satellites à imageur optique et les satellites à imageur radar.

1.2.1 Satellites optiques

(Nicolas, 2012) propose de raffiner la classification des satellites à imageur optique en 3 grandes familles en fonction de leurs champs de visée et de la résolution spatiale du capteur.

Certain satellites optique embarquent des capteurs pouvant capturer des longueurs d'onde au-delà du spectre visible (comme par exemple les infrarouges) on parle alors d'imagerie multi spectrale ou hyper-spectrale en fonction du nombre de bandes capturées.

Les satellites à imageur optique sont classés de la façon suivante :

- Imageur à champ de limbe à limbe Les applications de ces imageurs se situent principalement dans l'acquisition de donnée météorologique ou dans l'observation de zones océanographique. Généralement ces capteurs possèdent une résolution à l'échelle kilométrique. Des exemples de satellites à champs de limbe à limbe sont le Geostationary Operational Environmental Satellite (GEOS) ou encore Meteosat.
- Imageur à champ limité

  Ces imageurs se différencient des imageurs à champ de limbe à limbe par les zones plus limitées qu'ils peuvent imager mais surtout par leurs résolutions bien meilleur de l'échelle du décamètre. Ces satellites permettent néanmoins une couverture homogène et régulière de la surface du globe et sont utilisés pour la surveillance de la biosphère grâce à des capteurs hyper spectraux. Des exemples de satellites embarquant des capteurs à champ limité sont les programmes LANDSAT ou encore SPOT
- Imageur à champ limité haute résolution

  Plus récemment une nouvelle génération de capteurs à vue le jour, il s'agit des capteurs haute résolution dont la résolution spatiale peut descendre en dessous du mètre par pixel. Ces satellites sont dédié à l'observation de zones très précise de la planète. Les programmes

Ikonos et Worldview sont des exemples de satellite embarquant des capteurs très haute résolution.

1.2.2 Satellites RADAR

L'imagerie radar est considérée comme étant complémentaire de l'image optique, de par le principe d'acquisition des images les informations obtenues sont d'une tout autre nature. Grâce à son antenne un imageur RADAR va capturer la réponse des ondes électromagnétiques émise par l'antenne du satellite pour former une image. Les imageurs RADAR s'affranchissent ainsi de certaines contraintes des imageurs optique tel que la nécessité de l'éclairement solaire. Le radar s'affranchit aussi des désagréments causés par les nuages car les ondes électromagnétiques les traversent. Cependant le principal défaut des imageurs RADAR est la résolution des images obtenues qui dépend directement de la taille de l'antenne utilisée. Plus la taille de l'antenne sera grande et meilleur sera la résolution de l'image. Mais la dimension de l'antenne est limitée par l'espace disponible dans un satellite. Un radar traditionnel ne donne que des résolutions de plusieurs centaines de mètres alors qu'en utilisant le principe de la synthèse d'ouverture Synthetic Aperture Radar (SAR) nous pouvons aller jusqu'à des résolutions décimétrique.

1.3 Imagerie aérienne

Grâce à l'évolution des capteurs en imagerie aérienne, il est possible aujourd'hui d'avoir des images avec une résolution inférieure à 5 cm/pixel. À ce niveau de détails il est possible de distinguer à l'oeil nu les différents objets composant une image. Grâce à ce gain d'informations il devient possible de transférer des méthodes d'analyse d'images utilisées en vision par ordinateur vers des images aériennes.

Les méthodes usuelles d'analyse d'images vues du sol sont généralement développées pour des images où les objets d'intérêts occupent une large portion de l'image comme dans figure 1.4b. Alors qu'en imagerie aérienne, un grand nombre d'objets peuvent être représentés dans l'image comme l'illustre la figure 1.4a. Par conséquence les objets d'intérêt n'occupent généralement qu'une petite portion de l'image ce qui rend leur détection très difficile. Cependant l'imagerie satellitaire possède plusieurs autres spécificités qui lui sont propres par rapport aux images comme celle de la figure 1.4b. Premièrement le point de vue des images satellitaires change une grande partie des approches pour l'analyse d'images. Dans ce type d'images Le changement de point de vue n'existe pas comme dans d'autre type d'image, les objets sont toujours selon le même point de vue ce qui peut faciliter l'apprentissage de modèles basé sur l'apparence des objets. Cependant cela introduit d'autres problèmes comme les problèmes d'échelles et d'orientations qui sont beaucoup plus présents en imagerie satellitaire. En effet la taille d'un objet va dépendre de la résolution du capteur et le transfert de modèles entre capteurs de résolutions différentes n'est pas toujours possible. De plus les objets dans les images satellitaires peuvent être orienté à 360° ce qui va nous obliger à développer des méthodes qui doivent être robustes à la rotation des objets dans l'image.

(a)

(b)

FIGURE 1.4 – L'image Fig. 1.4a représente une extraite de notre base d'images annotée. L'image est à une résolution de 10 cm/pixel et contient plusieurs milliers d'objets que l'on souhaitera identifier individuellement. L'image Fig. 1.4b représente un groupe de 4 personnes, dans cette images les objets à identifier sont les personnes, le frisbee et les voitures en arrière-plan.

1.4 Contexte de la thèse

Nous avons vu que les images issues de capteurs aériens et satellitaires sont disponibles de plus en plus facilement et en très grande quantité. Ces images contiennent énormément d'informations qui pourront être exploitées pour différentes applications tel que la planification de missions de sauvetage, l'aménagement du territoire ou l'étude de l'environnement. Cependant le traitement de ces images nécessite énormément de ressources humaines. De plus même pour un opérateur humain expérimenté, interpréter et annoter les images pour une utilisation future peut se révéler fastidieux. C'est pour cela que se développent de plus en plus des systèmes de recherche et d'analyse d'images par le contenu qui permettent une automatisation de la tâche d'indexation du contenu des images. Ces systèmes se révèlent particulièrement utiles dans les cas où la mise en place des annotations est difficile : quand par exemple les images sont de grande taille ou sont en très grand nombre. Du point de vue de la recherche scientifique, les tâches d'indexation et de classification du contenu d'une image sont un sujet de recherche très actif dans les communautés vision par ordinateur et télédétection.

Classification statistique en télédétection : De nombreuses méthodes ont été mise au point pour classifier le contenu de ces images qu'elles soient issues de capteurs optiques, multi spectraux ou radar. Elles se basent sur des statistiques de l'image. Nous en présentons quelques-unes. (DATCU et al., 2003) proposent un système de recherche dans des grandes bases de données appelé Knowledge-driven Information Mining (KIM). KIM est un système qui supporte les interactions homme machine pour intégrer de manière adaptative des informations dans les résultats des requêtes de l'utilisateur. (GOMEZ-CHOVA, TUIA, MOSER, & CAMPS-VALLS, 2015) présentent une analyse détaillée de la classification d'images multimodales pour la télédétection. Ils analysent principalement les différentes méthodes de fusion ou de combinaison de données provenant de différentes sources pour augmenter la précision des méthodes de classification du contenu. Les auteurs décrivent un ensemble de méthodes utilisées pour la classification d'images multimodales en télédétection et mettent en avant les récentes avancées pour ce problème grâce à des méthodes issues du traitement du signal et de l'apprentissage statistique. (PENATTI, SILVA, VALLE, GOUET-BRUNET, & TORRES, 2014) proposent une méthode de classification des images basées sur le modèle Bag Of Words (BOW). La méthode baptisée Word Spatial Arrangement (WSA) permet de modéliser la répartition des mots visuels

dans le modèle BOW.

Apprentissage interactif : Un axe de recherche en pleine expansion pour la recherche et classification d'images par le contenu est l'approche par boucle de pertinence. Cette approche permet à un utilisateur expert de corriger les prédictions du système de manière interactive pendant l'apprentissage du modèle de classification. (Crawford, Tuia, & Yang, 2013) présentent une analyse des méthodes d'apprentissage interactive pour la classification des données acquises par télédétection. Les auteurs analysent plusieurs méthodes pour la sélection des exemples d'apprentissage de manière interactive. Ils évoquent aussi les différentes approches spécifiques à la télédétection pour gérer les problèmes rencontrés lorsque les données proviennent de capteurs différents (multimodales), de différentes localisations et/ou de plusieurs temporalités (multi temporelles) (Blanchart, Ferecatu, & Datcu, 2011) proposent une méthode de détection d'objets dans des images satellitaires basée sur de l'apprentissage interactif. La méthode d'apprentissage du modèle développé par les auteurs propose une approche où les exemples d'apprentissage sont des patches extraits à différentes échelles, de la plus grossière à la plus fine, qui serviront d'exemples d'apprentissage en fonction du retour de l'utilisateur. Les auteurs proposent aussi un algorithme d'apprentissage permettant de propager automatiquement les exemples d'apprentissage entre les différentes échelles. (Le Saux, 2014) propose une méthode de détection d'objets dans des images aériennes qui peut corriger le manque d'exemples d'apprentissage ou les exemples mal annotés grâce à l'apprentissage interactif. L'intervention de l'utilisateur expert dans la procédure d'apprentissage permet de choisir les exemples d'apprentissage les plus informatifs pour le modèle.

Approche orientée objet en télédétection : Avec l'augmentation de la résolution des capteurs les approches consistant à détecter directement les objets dans l'image et non plus seulement classifier les pixels sont devenus très populaires. Avec les capteurs à très haute résolution les objets de petite taille dans l'image comme les voitures fournissent assez d'information pour être détectées individuellement. Cette abondance d'informations permet l'utilisation de méthodes de détection d'objets utilisées par la communauté vision par ordinateur (Grabner, Nguyen, Gruber, & Bischof, 2008). Un type particulier d'objet dans une image aérienne est le chemin entre deux points, (Ratliff, Bagnell, & Zinkevich, 2007) propose une modélisation graphique de ce type d'objet et entraîne un classifieur du chemin optimal permettant d'éviter les obstacles entre deux points.

L'intérêt de cette approche par rapport à la tâche de classification des pixels de l'image est de pouvoir retrouver individuellement chacune des instances de l'objet d'intérêt. Cela permet d'autre type d'analyse du contenu d'une image comme compter automatiquement les véhicules sur une route ou localiser spécifiquement un bâtiment.

Détection et reconnaissance d'objets en vision par ordinateur : Avec la disponibilité des images aériennes très haute résolution il est possible d'utiliser des méthodes populaires de vision par ordinateur pour détecter des objets ou classifier les pixels d'une image (segmentation sémantique). Les méthodes tels que (Dalal & Triggs, 2005 ; Felzenszwalb, Girshick, McAllester, & Ramanan, 2010) requièrent que les objets dans l'image contiennent une certaine quantité de données pour détecter les objets de façon efficace. Bien qu'il soit possible d'utiliser ces méthodes avec des résultats encourageants, elles ne sont pas adaptées aux particularités de l'imagerie aérienne d'où la nécessité d'adapter ces méthodes pour l'imagerie aérienne. De plus les performances des méthodes de reconnaissance visuelle des objets dans une image aérienne peuvent être amélioré par la modélisation d'informations de haut niveau

tel que le contexte d'une image (Biederman, 1972) Cependant les informations de contexte extraits d'une image aérienne sont très différentes de celles pouvant être extraits d'une image en vue de face. Par exemple la position relative à un autre objet à la verticale ou la taille relative entre objets sont des informations de contexte utiles pour quantifier le contexte dans le cas des images comme celle de la figure 1.4b (Cinbis & Sclaroff, 2012) mais inutiles dans le cas de l'imagerie aérienne. Ces méthodes ayant fait leur preuves pour les tâches de classification et de détection d'objets dans des images naturels peuvent se montrer intéressantes si elles sont adaptées à l'imagerie aérienne. Une image aérienne contient des centaines d'objets ce qui suggère que les scènes aériennes contiennent un contexte entre les objets très riches que l'on peut exploiter pour améliorer la connaissance de la scène (Porway, Wang, & Zhu, 2010).

Positionnement : Cette thèse a pour objectif de développer des méthodes pour la classification du contenu d'une image qui soient adaptées aux spécificités de l'imagerie aérienne et satellitaire. Nous allons tirer parti de la résolution toujours croissante de ces images pour transférer des idées et des principes qui ont leur origine dans la vision par ordinateur. Pour ce faire nous avons choisi de nous concentrer sur deux types d'approches : la détection d'objets et la classification contextuelle.

L'approche pour la détection d'objets que nous proposons vise à modéliser finement les détails des objets tout en gardant l'information de l'organisation spatiale de ses éléments comme (Dalal & Triggs, 2005 ; Penatti et al., 2014). Nous proposons un détecteur basé sur un mélange de ces modèles (développant une idée de (Felzenszwalb et al., 2010)) et nous proposons une manière efficace de l'entraîner pour être particulièrement performant (Randrianarivo, Le Saux, & Ferecatu, 2015 ; Randrianarivo, Le Saux, Crucianu, & Ferecatu, 2016). Les détecteurs obtenus sont robustes aux changements d'échelles en fonction de la résolution de l'image ou aux changements d'orientations des objets dans l'image.

Dans un deuxième temps nous proposons une approche pour intégrer l'information contextuelle des images aériennes décrites par (Porway et al., 2010) dans la classification de scènes. Notre approche utilise un modèle graphique local pour caractériser et apprendre le contexte. L'apprentissage est effectué par des méthodes à noyaux structurelles ce qui étend à un cadre voisin (les objets au lieu des chemins) leurs utilisation dans (Ratliff et al., 2007).

1.5 Contributions à l'analyse d'images aériennes

Ce manuscrit de thèse se décompose en deux grandes parties. La partie I où est étudiée la problématique de la détection d'objets dans des images aériennes avec un mélange de modèles discriminatifs. La partie II où est approfondie le problème de la segmentation sémantique d'une image et de l'utilisation du contexte pour cette tâche.

La partie I est découpée de la façon suivante :

Le chapitre 2 constitue une introduction sur les méthodes usuellement utilisées pour l'analyse d'images aérienne et satellitaire. Dans ce chapitre nous présentons d'abord différentes méthodes utilisées pour la classification du contenu des images. Ensuite nous présentons les différentes méthodes pour l'évaluation d'un modèle statistique en classification et en détection. Le chapitre se conclut avec une présentation des différentes bases d'images que nous avons utilisées dans cette thèse.

Le chapitre 3 décrit nos contributions pour la modélisation de sous-catégories visuelles dans une catégorie sémantique. Nous présentons un modèle pour l'apprentissage de sous-catégories

visuelles au sein d'une catégorie sémantique en se basant sur une combinaison des métadonnées et des informations visuelles d'un exemple d'apprentissage (Randrianarivo et al., 2015). L'idée de modéliser les sous-catégories visuelles dans une catégorie sémantique a déjà été utilisées pour différentes tâches de détection d'objets dans des images. Notre approche se différencie des approches existantes en combinant deux types d'informations dans une méthode de partitionnement hiérarchique. Nous proposons aussi une procédure pour estimer le nombre de sous-catégories visuelles en se basant sur le Bayesian Information Criteron (BIC). Ce chapitre commence par une analyse de l'apparence des objets dans une base d'image et définit deux notions importantes : les catégories sémantiques et les catégories visuelles. Ensuite nous présentons différentes méthodes pour la recherche de sous-catégories visuelles au sein d'une catégorie sémantique et pour terminer nous évaluons nos résultats de recherche de sous-catégories visuelles.

Le chapitre 4 décrit notre méthode pour détecter des objets dans des images aériennes en utilisant des mélanges de modèles discriminatifs. Nous présentons comment est appris un modèle d'apparence pour une catégorie d'objets et une étude sur les performances des modèles des différentes sous-catégories pour la détection des objets dans l'image. Nous commençons le chapitre par une description de la signature utilisée pour quantifier l'apparence d'un objet dans une image. Puis nous montrons comment à partir d'un mélange de modèles d'apparence il est possible d'obtenir un détecteur d'objets performant (Lagrange et al., 2015 ; Randrianarivo et al., 2015 ; Randrianarivo, Le Saux, Crucianu, & Ferecatu, 2016). Les travaux de (Lagrange et al., 2015) ont été récompensés par une 2^{e} place à l'IEEE GRSS/IADF-TC 2015 Zeebrugge où notre méthode de détection de véhicules présentée était celle qui donnait les meilleurs résultats. Dans ce chapitre nous aborderons le sujet de la détection d'objets dans un contexte multimodales (Randrianarivo, Le Saux, & Ferecatu, 2014). Puis nous décrirons nos résultats expérimentaux sur différentes bases d'image aériennes.

La partie II se divise de la façon suivante : Le chapitre 5 porte sur l'état de l'art en modélisation d'informations contextuelles en reconnaissance d'images. Dans ce chapitre nous récapitulons les différentes méthodes permettant de modéliser les interactions entre les objets dans une image.

Le chapitre 6 quand à lui présente nos contributions sur la modélisation, l'évaluation et l'apprentissage d'informations contextuelles entre objets dans des images aériennes. Dans ce chapitre nous montrons dans un premier temps comment quantifier les différentes interactions entre les superpixels d'une image. Contrairement aux approches de la littérature en imagerie aérienne nous avons adopté pour une approche par sous-graphes locaux aux superpixels plutôt qu'une approche qui modélise le contexte global de la scène (Randrianarivo, Le Saux, Audebert, Crucianu, & Ferecatu, 2016). Nous commençons par décrire la tâche de segmentation sémantique et la méthode utilisée. Ensuite nous décrivons la procédure pour l'extraction du graphe de contexte de la scène et des sous-graphes locaux aux superpixels. Nous détaillons la procédure d'apprentissage utilisant des sous-graphes comme exemples d'apprentissage et finalement nous présentons une méthode pour la prédiction de la catégorie sémantique d'un superpixel basée sur les prédictions de chacun de ses voisins pour robustifier les prédictions. Finalement le chapitre 7 présente nos conclusions et décrit de possibles futures voies de recherche.

Première partie

Détection d'objets

2

Analyse d'image aérienne et satellitaire

Sommaire

De nombreuses équipes de recherche se sont penchées sur le problème de l'analyse du contenu des images aériennes et satellitaires. De nombreuses méthodes existent qu'elles concernent l'imagerie panchromatique, multi/hyper spectrale ou que les images soient à basse, haute ou très haute résolution. En effet en fonction de la nature du capteur utilisé et de la résolution des images à analyser différentes méthodes ont montré leur efficacité. C'est pour cela que dans un premier temps nous passons en revue un ensemble de méthodes de référence pour l'analyse d'image en section 2.1. Une autre problématique est l'analyse des résultats obtenus par ces méthodes. De nombreuses méthodes existent en fonction de la tâche à évaluer : en section 2.2 nous présentons les différentes métriques que nous utiliserons pour l'évaluation de nos méthodes. Enfin dans la section 2.3 nous présentons les différentes bases d'images que nous avons utilisées pour évaluer nos méthodes, trois bases d'images ont été utilisées pour la tâche de détection d'objets et une quatrième pour la tâche de segmentation sémantique.

2.1 Classification en télédétection

L'analyse automatique d'images issues de la télédétection est une problématique très importante que ce soit au niveau industriel ou scientifique. Elle consiste à extraire des images des informations permettant à une personne de pouvoir interpréter la carte facilement. Par exemple en donnant une catégorie aux pixels d'une image ou encore la reconnaissance automatique de chemins prenant en compte les obstacles entre deux points dans l'image sont des exemples d'analyses qui facilitent l'interprétation de l'image par l'opérateur. La mise en oeuvre de des méthodes d'analyse d'images aériennes suit différentes approches en fonction des capteurs et de la résolution des images. Les méthodes basées sur la classification automatique de différentes régions de l'image sont les plus populaires. (Qiao Weng, 2009) classe ces méthodes selon 6 grandes approches elles-mêmes divisées en différents sous-ensembles : par pixel (17 sous-ensembles), sous pixellique (7 sous ensembles), par région (6 sous ensembles), basée contexte (13 sous ensembles), basée sur la connaissance du domaine (6 sous ensembles), approche ensembliste de classifieurs (14 sous ensembles). Une autre manière de classer le contenu d'une image aérienne est d'utiliser une approche par objet. (Blaschke et al., 2014) développent le concept d'objet en télédétection et montrent les avantages que cette approche a sur les méthodes basées pixels ou régions. Dans cette thèse nous nous intéressons en particulier aux approches objets et régions pour l'analyse d'images. Ces deux types d'approches se basent sur une procédure en deux étapes. L'étape d'apprentissage qui peut se résumer par les étapes suivantes :

1. Constitution d'une base d'apprentissage
2. Extraction des caractéristiques des exemples
3. Apprentissage du modèle

L'étape d'évaluation :

1. Extraction des caractéristiques des images
2. Application du modèle

Pour ces deux étapes nous avons besoin de définir comment extraire des caractéristiques des images et définir le modèle qui va être utilisé pour la reconnaissance. Dans les problèmes de classification d'images, l'une des premières étapes consiste à extraire une représentation de l'image à l'aide de descripteur bas niveau tel que le Histogram of Oriented Gradients (HOG) se basant sur la forme ou le Locally Binary Pattern/Motifs Locaux Binaires (LBP) se basant sur la texture. Une représentation intermédiaire comme le BOW peut être calculées pour augmenter l'expressivité des descripteurs. L'étape suivante est l'apprentissage du modèle à partir des descripteurs de l'image. Le choix de la méthode d'apprentissage est grandement dépendante de la nature de la description des données mais ces dernières années les approches par Support Vector Machines (SVM) (Mountrakis, Im, & Ogole, 2011) et par *random forest* (Belgiu & Drăguţ, 2016) sont les plus populaires. Un panorama des méthodes utilisées pour les images panchromatique est donnée par (Lu & Weng, 2007) Une analyse des méthodes utilisées pour les images hyper spectrales est donnée par (Camps-Valls, Tuia, Bruzzone, & Benediktsson, 2014 ; Gomez-Chova, Tuia, Moser, & Camps-Valls, 2015)

2.1.1 Par pixel

Depuis les débuts de l'analyse d'images en télédétection dans les années 70, l'approche consistant à analyser individuellement chaque pixel de l'image dans le but de lui donner un label est la principale approche pour l'analyse d'images. Cette approche se justifie principalement en imagerie satellitaire car il n'est pas rare que la résolution d'un pixel de ces images soit du même ordre de grandeur que les objets d'intérêt. Les approches qui se concentrent sur les pixels tirent aussi

avantage de la capacité de certains capteurs à fournir un grand nombre de bandes spectrales pour un pixel. Ces capteurs permettent de compenser le manque d'information spatiale par des informations spectrales qui peuvent mettre en avant des informations autres que les informations de couleur.

Parmi ces approches pour classifier les pixels certaines se basent sur les informations locales autour d'un pixel pour en extraire des informations de plus haut niveau comme la texture ou le contexte.

Les méthodes remarquables en imagerie panchromatique sont :

(Haralick, Shanmugan, & Dinstein, 1973 ; Marceau, Howarth, Dubois, & Gratton, 1990) utilisent des Grey Level cooccurence Matrices/Matrices de Co-occurence des Niveaux de Griss (GLCMs) pour quantifier la texture des différentes zones d'une image. Ensuite les auteurs utilisent alors soit un test statistique, soit un classifieur pour donner un score aux différentes régions de l'image.

(Hay, Niemann, & McLean, 1996) utilisent une approche basée sur des variogrammes comme une mesure de texture alternative aux GLCMs. Le variogramme permet en plus d'ajouter une information spatiale à l'information de texture.

(Gerhardinger, Ehrlich, & Pesaresi, 2005) utilisent un filtre sur la forme d'une voiture pour reconnaître les véhicules dans des images haute résolution (60 cm/pixel). La réponse du filtre sur l'image permet ainsi de localiser les véhicules sur l'image. Le seuil de décision est estimé à partir d'échantillons de véhicules.

(Montoya-Zegarra, Wegner, Ladický, & Schindler, 2015) proposent une méthode de segmentation sémantique des pixels d'une image aérienne utilisant des informations de contexte pour décrire la configuration entre les pixels d'une image en fonction de leurs catégorie. La méthode se concentre particulièrement sur les relations entre les catégories bâtiments et routes pour modéliser le contexte.

Dans cette thèse, le travail décrit chapitre 6 sur la segmentation sémantique d'images aérienne est comparable aux travaux de (Montoya-Zegarra et al., 2015). Comme eux nous avons décidé d'incorporer une modélisation des interactions intraclasses pour améliorer les performances du classifieur. Cependant contrairement à eux nous avons choisi un modèle qui prend en compte les relations entre les superpixels extraits de l'image et nous avons choisi de quantifier le contexte sous la forme d'un vecteur descripteur pour l'intégrer dans un modèle graphique qui décrit un superpixel et son voisinage.

Les méthodes remarquables en imagerie hyper spectrale sont :

Une mesure commune pour reconnaître la végétation en imagerie hyper-spectrale est d'utiliser entre autre le Normalize Difference Vegetation Index (NDVI). Cet indice ainsi que de nombreux autre sont étudiés dans (Haboudane, 2004).

(Melgani & Bruzzone, 2004) ont proposé une méthode qui permet de sélectionner automatiquement les bandes les plus informatives vis-à-vis de la variance en utilisant la Principal Component Analysis/Analyse en Composante Principales (PCA). Les pixels sont ensuite classés en utilisant les bandes sélectionnées comme entrée pour des SVMs.

L'approche par pixel est l'approche la plus étudiée pour la classification du contenu d'images en télédétection. Ceci est principalement dû à la taille des objets d'intérêt de l'image qui est du même ordre de grandeur que la résolution spatiale des pixels. Cependant avec l'augmentation de la résolution des images disponibles de nouvelles méthodes émergent pour analyser ces nouvelles données. Dans les images très haute résolution la variabilité des pixels d'une même classe d'objets augmente énormément. Les méthodes par pixel ont alors beaucoup de difficultés à modéliser et reconnaître une classe d'objet au niveau du pixel. (Hay et al., 1996) nomment ce problème

Le Problème de Haute Résolution. L'étape suivante consiste à trouver comment regrouper des ensembles de pixels pour obtenir des régions de l'image correspondant à des classes d'objets d'intérêt.

2.1.2 Par région

Bien que la classification individuelle des pixels d'une image donne de bonnes performances en imagerie basse résolution, cette approche tend à être plus problématique en haute et très haute résolution. Avec l'augmentation de la résolution et des détails des objets présents dans une image la variance des pixels appartenant à une même classe tend à augmenter. Il devient alors beaucoup plus difficile de trouver des caractéristiques communes aux pixels pour les classifier. Une manière d'augmenter le nombre d'informations discriminative est de considérer des groupes de pixels homogènes pour la classification. Les approches populaires de la littérature se divisent en deux approches principales : les méthodes supervisées et les méthodes non supervisées. Les méthodes de classification non supervisée divisent des données hétérogènes en groupes homogènes. Les méthodes de classification supervisées s'appuient sur un ensemble d'exemples annotés pour réaliser la classification.

Méthodes non supervisées : Les méthodes d'apprentissage de modèles de modèle non supervisées ont été jusqu'à très récemment très étudiées pour la classification de données et plus particulièrement de grandes masses de données. Ces méthodes de classification minimisent un critère sur les données qui sera utilisé pour diviser les données en différents sous-ensembles. Le grand avantage de ces méthodes est que l'apprentissage d'un modèle ne dépend pas de la constitution d'une base d'apprentissage qui peut être difficile à mettre en place. Cependant c'est à l'utilisateur final de donner une signification sémantique à chacun des sous-ensemble retrouvé par la méthode. Parmi les méthodes de classification non supervisées des régions d'images aérienne et satellitaire nous pouvons citer :

(Lorette, Descombes, & Zerubia, 2000) proposent d'extraire les zones urbaines d'images satellitaires en modélisant les informations de texture. Ils proposent de caractériser la texture d'une image en utilisant un ensemble de modèles markoviens dont les paramètres sont estimés en fonction de la direction estimée du pixel. Les auteurs utilisent un algorithme des *fuzzy C-means* modifié pour prendre en compte l'entropie comme critère de partitionnement.

(Cao, Yang, & Mao, 2005) propose une méthode en deux étapes pour la détection de structures humaines dans des images aériennes. Les auteurs utilisent d'abord l'erreur fractale pour caractériser les constructions humaines. L'erreur fractale est un indice efficace pour différencier les constructions (grande erreur fractale) de la végétation (petite erreur fractale) (Selvage, Chenoweth, & Cooper, 1994). En conjonction avec l'erreur fractale les auteurs utilisent la Discret Cosinus Transform/Transformée Discrète en Cosinus (DCT) pour caractériser les bords des zones texturées de l'image. La détection de contours par DCT donnant de meilleurs résultats que les détecteurs de contours classique (Randen & Husoy, 1999). Finalement la segmentation est faite en utilisant un algorithme basée sur la fonctionnelle de Mumford-Shah et dont les contraintes utilisent à la fois l'erreur fractale et les contours extraits par la DCT.

(Molinier, Laaksonen, Member, & Häme, 2007) proposent de détecter différentes classes d'objets dans des images satellitaires haute résolution (QuickBird 60 cm/pixel) en utilisant un système interactif. Ils proposent d'extraire différentes caractéristiques des images comme la texture ou un histogramme des orientations des contours pour caractériser

les constructions. Des cartes de Kohonen (Kohonen, Schroeder, & Huang, 2001) sont alors apprises et utilisées pour donner un score sur la présence ou non des constructions dans l'image.

Méthodes supervisées : Plus récemment les méthodes d'apprentissage de modèles supervisés ont connus une grande popularité. Le principe de ces méthodes est de classifier les données d'apprentissage en fonction de catégories prédéfinies. L'utilisation de méthodes supervisées présente l'avantage de donner un sens sémantique à la tâche de classification. En effet les données sont classées en catégories bien définit tel que "végétation", "voiture" ou "bâtiment" ce qui simplifie énormément la tâche d'interprétation des images pour l'utilisateur en fin de chaîne. La réalisation de la classification de régions de l'image en fonction de catégories sémantiques est aussi appelée segmentation sémantique de l'image. L'apprentissage à partir des données brutes de l'image ne permet généralement d'obtenir une classification de l'image satisfaisante. Il faut d'abord passer par une phase de description des régions de l'image et c'est sur cette description des régions de l'image que seront appris les modèles. Ainsi nous avons identifié comme méthodes :

(Unsalan & Boyer, 2004) ont pour but de différencier des zones urbaines de zones de végétation en imagerie satellitaire haute résolution. Ils utilisent une combinaison de caractéristiques extraites d'images panchromatiques et hyper spectrales. Les caractéristiques extraites des images sont différentes mesures statistiques à la fois sur les lignes composant différents environnements urbains pour les images panchromatiques et sur l'indice NDVI pour les images hyper-spectrales. La classification se faisant avec des classifieurs tel que des classifieurs Bayésiens ou des Fenêtres de Parzen.

(Inglada, 2007) propose de détecter des constructions dans des images satellite haute résolution (2,5 m/pixel) en couplant les caractéristiques géométriques des objets avec une SVM. L'une des originalités de ce travail est que l'auteur a développé un système générique pour la détection d'objets indépendamment du type d'objet recherché. L'auteur propose deux types de descripteur pour caractériser les objets dans une fenêtre d'analyse. D'abord des descripteurs bas-niveau capturant les propriétés géométriques des objets (un descripteur basé sur les moments géométrique (Flusser, 2000) et un descripteur basé sur la transformée de Fourier-Mellin (Derrode & Ghorbel, 2001)). Ensuite un ensemble de descripteurs hauts niveau composé d'histogrammes sur des mesures comme les distances des barycentres des régions *fermées* ou l'entropie des orientations des lignes dans la fenêtre d'analyse.

2.1.3 Par objets

Une autre façon de classifier le contenu d'une image est d'adopter une approche orientée objet. Au lieu de classifier toutes les régions/pixels de l'image, nous partons du principe qu'une image est composée de différentes classes d'objets d'intérêt. En plus des objets, une classe est souvent ajoutée que l'on nommera *fond*. Pour la tâche de détection d'objets, il s'agit de distinguer une classe d'intérêt (par exemple la classe voiture) de la classe fond qui représente tout ce que n'est pas un objet. Une analyse détaillée de la classification d'objets dans des images aérienne et satellitaire a été effectuées par (Blaschke et al., 2014). Cependant contrairement aux auteurs de (Blaschke et al., 2014) nous ne considérerons comme objets que des ensembles connexes de pixels qu'un humain peut assimiler à un concept sémantique simple tel que "voiture" ou "bâtiment" :

(Grabner, Nguyen, Gruber, & Bischof, 2008) proposent une méthode pour la détection de véhicules dans des images aériennes. Les auteurs utilisent 3 types de caractéristiques

pour apprendre un modèle de voitures : des caractéristiques de Haar (Viola & Jones, 2004), des HOGs (Dalal & Triggs, 2005) et une version simplifiée des LBPs (Ojala, Pietikainen, & Maenpaa, 2002). L'apprentissage du modèle se fait avec une version interactive du *boosting* (Schapire, 2002), c'est à dire qu'un humain sert d'oracle pour le classifieur afin corriger les exemples mal classés.

(Sirmaçek & Ünsalan, 2009) proposent une méthode de détection de bâtiments dans des images satellitaires basée sur l'extraction de points d'intérêt et le groupement de ceux-ci en utilisant des méthodes graphiques. Les auteurs choisissent deux patchs représentant deux bâtiments avec des illuminations différentes comme modèles de bâtiments. Ils utilisent la méthode proposé par (Lowe, 2004) pour trouver les points d'intérêts de l'image et les décrire en utilisant le Scale-Invariant Feature Transform (SIFT). Les zones urbaines de l'image de test sont extraites de l'image en utilisant une méthode de *sub-graph matching*. Les bâtiments sont redéfinis individuellement en utilisant une méthode de *graphcut* sur les nuages de points obtenus.

(Sun, Sun, Wang, Li, & Li, 2012) proposent de détecter des avions dans des images aériennes hautes résolutions en utilisant un modèle de sac de mots visuels modifié pour prendre en compte des contraintes spatiales. Les auteurs extraits des points d'intérêt des exemples d'apprentissage en utilisant la méthode proposée par (Lowe, 2004) qui seront utilisés comme mots visuels pour l'apprentissage d'un dictionnaire. Un modèle de l'objet d'intérêt est ensuite appris en encodant les exemples d'apprentissage avec le dictionnaire appris. Le modèle d'objet est entraîné à partir des exemples encodé en utilisant une SVM linéaire. La détection se faisant avec une simple glissante dans l'image qui est testé contre le modèle entraîné. Afin d'introduire de la robustesse aux rotations et ajouter des informations sur la structure des objets, les auteurs utilisent un système de coordonnée polaire et une fenêtre glissante circulaire.

Ces approches contrairement à la méthode de détection d'objets présentée dans le chapitre 4 ne prennent pas en compte la variabilité interclasse d'une classe d'objets à détecter. La modélisation de la variance interclasse, comme étudiée dans le chapitre 3, permet l'apprentissage de modèles robustes aux variations d'orientations et d'apparence. De plus la méthode présentée dans le chapitre 4 se base sur un modèle visuellement facilement interprétable par un humain ce que permet d'analyser le modèle appris et d'utiliser une méthode de *template matching* rapide pour détecter les objets.

2.1.4 Par apprentissage profond

Grâce à l'augmentation en nombre des données disponibles une famille de méthodes pour la classification se distingue : les méthodes utilisant l'apprentissage profond. Les méthodes de classification basée sur l'apprentissage profond ont besoin en règle général de plusieurs millions d'exemples d'apprentissage pour apprendre une représentation fiable des entrées. Comme elles apprennent directement une représentation des données, il n'est pas utile de chercher à extraire des descripteurs des images et on peut utiliser les pixels bruts de l'image comme entrée. Les descripteurs appris par le réseau sont souvent beaucoup plus discriminants que les descripteurs de type HOG ou SIFT mais nécessite une gigantesque quantité de données pas toujours accessible.

(Mnih & Hinton, 2010) proposent une méthode de détection de routes dans des images aériennes haute résolution. Ils classifient individuellement les pixels de l'image en utilisant un réseau de neurones profond et appliquent un post-traitement pour assurer la cohésion spatial des objets à détecter. Les auteurs pointent l'abondance de cartes où sont référencées les routes ce qui leur a permis de mettre en oeuvre leur méthode.

(Campos-Taberner et al., 2016) décrit les résultats du IEEE GRSS/IADF-TC 2015 Zeebrugge (DFC2015) où les deux équipes gagnantes du concours 2D ont présentés des méthodes basées en grande partie sur du *deep-learning*. Les auteurs présentent une comparaison des différentes méthodes de l'état de l'art pour la classification d'images aériennes. Les principaux résultats montrent que les approches par *deep-learning* surpassent les autres méthodes pour la classification de régions (segmentation sémantique).

(L. Zhang, Zhang, & Kumar, 2016) passent en revue les différentes approches utilisant le *deep-learning* en télédétection. Ils commencent par une présentation des différents types algorithmes existant pour le *deep learning* puis enchaînent sur leurs applications pour la télédétection. Ils mettent en avant un cadre générale pour utiliser le *deep-learning* en prenant en compte la nature (optique ou hyperspectrale) des images à classifier. Les auteurs ont aussi analyser l'utilisation des méthodes de *deep-learning* en tant qu'extracteur de descripteurs. Ils comparent les résultats de classification utilisant les descripteurs appris à plusieurs descripteurs de la littérature utilisés en télédétection pour montrer la supériorité des descripteurs appris.

Les méthodes présentées dans cette section ont prouvé leur efficacité par le passé mais présentent aussi un certain nombre de limites. Dernièrement les méthodes inspirées par la vision par ordinateur ont rencontré un certain succès en télédétection. Ce succès est principalement dû à l'arrivée de nouvelle données à très haute résolution permettant de nouvelles approches tel que la détection d'objets dans les images.

Dans cette thèse nous avons choisis dans un premier temps d'investiguer les méthodes de détection d'objets pour l'imagerie aérienne en identifiant ce qui fait leur force en vision et comment les utiliser efficacement en imagerie aérienne. Dans un second temps nous avons cherché un moyen d'améliorer les méthodes de segmentation sémantique en explorant la piste du contexte entre les objets dans les images aériennes qui a été très peu étudiée.

2.2 Évaluation des modèles

2.2.1 Définitions

Une fois un modèle l'objet appris une question importante se pose est : Comment évaluer les performances d'un modèle pour une tâche donnée ? Pour cela nous définissons la fonction de score $f(x \mid w) = \mathrm{p}(Y = 1 \mid X = x; w)$ dont les valeurs sont comprises entre $[0\,;1]$ avec w le vecteur des paramètres du modèle. Soit une quantité T définit par $f\colon X \to \mathbb{R}$ et correspondant aux sorties du modèle. Soit un ensemble de N exemples annotés $(X, Y) = \{(x_n, y_n) \mid n = 1 \dots N\}$, l'application du modèle sur cet ensemble produit un vecteur de scores $T \in \mathbb{R}^n$. L'évaluation du modèle a pour but d'analyser la probabilité jointe des ensembles (Y, T) pour déterminer les performances de classification du modèle. Afin d'analyser la probabilité jointe (Y, T), nous définissons 4 quantités qui modèlent les probabilités d'apparition des différents événements :

- La probabilité d'un événement True Positive/Vrai Positif (TP) pour un seuil de décision t. Les TPs sont les exemples positifs classé comme positifs par le modèle

$$\mathrm{p}(T > t, Y = 1) \tag{2.1}$$

- La probabilité d'un événement False Positive/Faux Positif (FP). Les FPs sont les exemples négatifs classés comme positifs par le modèle

$$\mathrm{p}(T > t, Y = 0) \tag{2.2}$$

- La probabilité d'un événement True Negative/Vrai Négatif (TN). Les TNs sont les exemples négatifs classés comme négatifs par le modèle

$$\mathrm{p}(T \leq t, Y = 0) \tag{2.3}$$

- La probabilité d'un événement TN. Les False Negative/Faux Négatifs (FNs) sont les exemples positifs classés comme négatifs par le modèle

$$\mathrm{p}(T \leq t, Y = 1) \tag{2.4}$$

En pratique ces quantités sont estimées empiriquement à partir des annotations en comptant les occurrences des événements. On fait l'hypothèse que les exemples X sont Indépendants et Identiquements Distribués (i.i.d) et on utilise le théorème de Glivencko-Cantelli pour approcher les probabilités des différents événements. Le théorème de Glivencko-Cantelli énonce qu'une loi de probabilité peut être révélée par la connaissance d'un grand nombre d'échantillon de ladite loi de probabilité. La probabilité d'apparition d'un événement TP, FP, TN ou FN est donnés par :

$$\mathrm{p}(T > t, Y = 1) \approx \frac{\#TP(t)}{N} \tag{2.5}$$

$$\mathrm{p}(T > t, Y = 0) \approx \frac{\#FP(t)}{N} \tag{2.6}$$

$$\mathrm{p}(T \leq t, Y = 1) \approx \frac{\#TN(t)}{N} \tag{2.7}$$

$$\mathrm{p}(T \leq t, Y = 1) \approx \frac{\#FN(t)}{N} \tag{2.8}$$

où $\#TP(t)$, $\#FP(t)$, $\#TN(t)$, $\#FN(t)$ désignent le cardinal de l'ensemble des éléments de TP, FP, TN, FN pour un seuil t.

Une fois ces différents événements définis, il devient possible d'utiliser un large ensemble de métriques pour évaluer les performances du modèle. Dans cette thèse, nous nous intéresserons aux métriques mettant en oeuvre les notions de précision et rappel. Dans le contexte de la détection d'objets dans des images, les mesures de précision et de rappel sont particulièrement bien adaptées pour l'évaluation des performances d'un modèle où le nombre de TNs (tous les ensembles connexes de pixels du fond de l'image) est immense par rapport aux autres événements (TPs, FPs et FNs) (Davis & Goadrich, 2006).

2.2.2 Précision et Rappel

La connaissance de la probabilité des différents événements permet d'introduire deux nouvelles métriques que sont la précision et le rappel. La précision est une métrique quantifiant le taux de vrais positifs parmi tous les échantillons testés. Elle évalue la capacité du détecteur à trouver le plus grand nombre de cibles possibles. Le rappel est une métrique quantifiant le taux de vrais positifs parmi les échantillons prédits comme étant positifs. Il évalue la capacité du détecteur à renvoyer le moins de fausses alarmes possibles.

Ces métriques sont formellement définies de la façon suivante :

- Expression de la précision $prec(t)$

$$\begin{aligned} prec(t) &= \mathrm{p}(Y = 1, T \geq t) \\ &= \frac{\mathrm{p}(Y = 1, T > t)}{\mathrm{p}(T > t)} \\ &= \frac{\mathrm{p}(Y = 1, T > t)}{\mathrm{p}(T > t, Y = 1) + \mathrm{p}(T > t, Y = 0)} \\ &\approx \frac{TP(t)}{TP(t) + FP(t)} \end{aligned} \tag{2.9}$$

- Expression du rappel $rec(t)$

$$\begin{aligned} rec(t) &= \mathrm{p}(T > t \mid Y = 1) \\ &= \frac{\mathrm{p}(T > t, Y = 1)}{\mathrm{p}(Y = 1)} \\ &= \frac{\mathrm{p}(T > t, Y = 1)}{\mathrm{p}(T > t, Y = 1) + \mathrm{p}(T < t, Y = 1)} \\ &= \frac{TP(t)}{TP(t) + FN(t)} \end{aligned} \tag{2.10}$$

Ces métriques sont déjà des mesures permettant d'évaluer le comportement d'un modèle sur une base de test. Elles permettent de donner une idée sur le comportement du modèle vis-à-vis des exemples positifs de la base de test. Cependant en combinant ces deux métriques nous pouvons obtenir une interprétation plus générale des performances du modèle. La précision et le rappel pour un modèle donné ont des comportements opposés par rapport au seuil t utilisé. Pour un seuil t très petit un grand nombre d'échantillons testés seront classés comme étant des échantillons positifs, cela implique que parmi ces échantillons un grand nombre seront les FPs. Par contre un grand nombre de cibles seront détectées, on a donc une précision basse et un rappel élevé. Pour un seuil t très grand au contraire le détecteur classifiera peu d'échantillons comme positifs mais avec une très grande confiance ce qui signifie peu de FPs . Par contre le détecteur pourra louper certaines cibles avec un score de confiance plus faible, on a donc une précision élevée et un rappel faible.

2.2.3 Précision Moyenne

Parmi les métriques possibles à partir de $prec(t)$ et $rec(t)$ l'aire sous la courbe précision rappel est un outil particulièrement efficace pour évaluer un modèle. De plus la courbe précision-rappel introduit une notion de classement entre les exemples de test par rapport à la sortie du modèle. La courbe précision-rappel une courbe paramétrée est ainsi définit par un ensemble de seuils t

$$\mathrm{PR} = \{(rec(t), prec(t)) \mid t \in \mathbb{R}\} \tag{2.11}$$

Au fur et à mesure que le seuil t augmente, le rappel tend vers 0 et a contrario la précision doit tendre vers 1. Si la précision tend vers 1, cela signifie que l'exemple de test avec le plus haut score est un TP ce qui est le comportement souhaitable pour le modèle. Par contre si la précision ne tend pas vers 1 alors que le rappel tend vers 0, cela signifie que l'exemple avec le score le plus

haut n'est pas un TP. Cela implique que la valeur de la précision autour de $rec(t) \sim 0$ est sujette à une grande variance et que la construction de la courbe en sera perturbée.

Une fois la courbe précision-rappel construite, il est utile de pouvoir la réduire à une seule valeur pour une analyse plus aisée. Une façon de faire est de calculer l'aire sous la courbe pour obtenir un score de performance du modèle. L'aire sous la courbe précision-rappel peut-être interprétée comme une moyenne pondérée de la précision pour un seuil donné ou encore la fraction d'exemples positifs que de modèle pourra reconnaître pour un seuil donné. Dans la littérature ce score est appelé Average Precision/Précision Moyenne (AP) ou Area Under the Curve (AUC) et est défini de la façon suivante :

$$\text{ap} = \int_{\mathbb{R}} prec(t)\, \mathrm{d}P(T \leq t) \tag{2.12}$$

Le calcul de l'intégrale équation (2.12) n'est cependant pas trivial. Pour une valeur donnée du rappel $rec(t)$, il peut exister plusieurs valeurs de la précision $prec(t)$. Le calcul de l'aire sous la courbe précision-rappel n'est pas trivial car la définition équation (2.11) ne garanti pas que pour une valeur de $rec(t)$ il n'existe qu'une seule valeur de $prec(t)$. Il existe donc plusieurs méthodes pour calculer cette aire, (Boyd, Eng, & Page, 2013) ont en fait une étude détaillée. Dans nos propres expériences nous calculons l'aire sous la courbe précision-rappel en gardant les points avec la précision la plus élevée pour chaque valeur de t.

2.2.4 Taux de classification

Le taux de classification (*accuracy*) est le score qui mesure le taux de bonnes classifications d'un modèle sur l'ensemble des prédictions. Ce taux est défini par :

$$\text{acc} = \frac{\#TP + \#TN}{\#TP + \#TN + \#FP + \#FN} \tag{2.13}$$

Ce score de classification se prête particulièrement bien à l'évaluation des modèles multiclasses où l'on veut mettre en avant la capacité du modèle à prédire la bonne catégorie parmi N autres catégories.

2.2.5 f_1-score

Le f_1-score est une mesure de l'exactitude des prédictions binaires réalisées par un modèle. Il s'agit cas particulier du f_β-score où $\beta = 1$. Il est donné par :

$$f_beta(t) = (1 + \beta^2)\frac{prec(t) \times rec(t)}{(\beta^2 prec(t)) + rec(t)} \tag{2.14}$$

Ce score peut être interprété comme une moyenne pondérée entre la précision et le rappel. Il permet de donner un aperçu de la capacité du modèle à retrouver précisément les éléments à classifier au-dessus d'un certain score t

2.2.6 Intersection over Union (IoU)

Lors de la détection d'objets dans des images il faut pouvoir établir pour un seuil t quel sont les vrais détections des fausses alarmes. Pour cela nous utilisons la définition de *bonne détection* tel que définie par le Pascal Visual Object Challenge (Pascal VOC) (Everingham, Gool, Williams, Winn, & Zisserman, 2010). Les hypothèses de détection dans une image sont assignées aux objets de la vérité terrain et sont considérées comme de bonnes ou mauvaises détections en fonction du recouvrement entre la détection et la vérité terrain. Pour être considérée comme une bonne détection, la boite englobante d'une détection A doit recouvrir la boite englobante de la vérité terrain B avec un score supérieur à 0,5 selon la formule suivante :

$$\text{IoU}(A, B) = \frac{\text{aire}(A \cup B)}{\text{aire}(A \cap B)} \tag{2.15}$$

2.2.7 Bayesian Information Criteron

Dans le chapitre 3 nous présentons une méthode de modélisation des sous-catégories visuelles au sein d'une catégorie sémantique. Cependant la méthode proposée ne permet pas de déterminer a priori le nombre de sous-catégories visuelles. Pour trouver le nombre de sous-catégories visuelles permettant de maximiser les performances du détecteur nous proposons d'utiliser le BIC. Le BIC permet d'estimer le nombre de composantes d'un mélange en fonction du nombre de paramètres et du nombre d'échantillons à partitionner. L'avantage du BIC sur d'autres critères comme le Akaike Information Criteron (AIC) est qu'il prend en compte le nombre d'échantillons dans la partition ce qui permet d'obtenir des partitions avec un nombre équilibré d'éléments par partie. Le BIC est donné par :

$$\text{BIC} = 2 \times \ln \hat{L} + k \times \ln(n) \tag{2.16}$$

Où $\hat{L}$ est la fonction de vraisemblance du modèle, k le nombre de paramètres du modèle à estimer et n le nombre d'observations. Le nombre optimale de composantes pour la Gaussian Mixture Model/Mélange de Modèles Gaussiens (GMM) est le nombre k qui minimise le BIC, les autres paramètres découlant de ce paramètre.

2.3 Bases d'images

Pour l'évaluation de nos méthodes de détection d'objets et de segmentation sémantique nous utilisons des bases d'images aériennes ou satellitaires publiques. Ces images permettent de tester nos méthodes sur des scènes complexes et difficiles. Elles contiennent un grand nombre d'objets d'intérêt que l'on doit détecter/classifier. Les pixels n'appartenant pas à un objet d'intérêt définissent un fond, complexe et très texturé, qui va gêner la détection. Elles différent également par le type de contenu : les villes imagées sont très différentes les unes des autres d'un point de vue architectural comme le montre la figure 2.1.

Afin de pouvoir entraîner et évaluer nos méthodes nous avons besoin qu'une vérité-terrain existe pour ces images. La vérité terrain est la réalisation par un interprète du détourage manuel des objets dans l'image pour la détection d'objets ou d'une carte de classification dense pour la segmentation. Ces vérités-terrain ont un double intérêt pour nous. D'une part elles constituent

FIGURE 2.1 – Aperçu de la diversité des bases de données utilisées pour les expériences. La première ligne montre des zooms sur les images aériennes des différentes villes imagées (de gauche à droite) : Christchurch (Nouvelle-Zélande), Zeebrugge (Belgique), Thetford Mines (Canada) et Vaihingen (Allemagne). Les deux premières bases sont des images Rouge-Vert-Bleu (RGB), une image multimodale et pour la dernière des images InfraRouge-Rouge-Vert (IR-R-G). La deuxième ligne montre des vues au sol des rues de chaque ville où l'on trouve différents types de bâtiments suivant la localisation.

l'objectif à atteindre par nos méthodes. D'autre part elles sont utilisables en partie pour entraîner des méthodes d'apprentissage statistique.

Nous utilisons deux types de bases :

1. Celles mises à disposition par la communauté scientifique comme les bases IEEE GRSS/IADF-TC 2014 Thetford Mines (DFC2014) et Vaihingen (Vaihingen) qui possèdent déjà une vérité terrain.
2. Des bases d'images disponibles publiquement comme ONERA Christchurch (Christchurch) et DFC2015 pour lesquelles nous avons été amenés à créer la vérité-terrain associée.

Utiliser des bases d'images accessibles au plus grand nombre permet en outre de pouvoir nous comparer avec le reste de la communauté.

Ces bases d'images sont variées entre elles tant par la résolution que par les capteurs qui les ont générées. Dans tous les cas il s'agit d'images aériennes à très haute résolution. C'est à dire que la résolution des images est d'au moins 20 cm/pixel. L'avantage est que les objets contiennent suffisamment de détails pour utiliser des méthodes issues de la vision par ordinateur. Les différentes bases permettront de valider la robustesse des méthodes à différentes échelles. Enfin, dans le cas de la base DFC2014 les images optiques sont complétées par des images multispectrales de la même zone afin de permettre la mise en place de méthodes de fusion de données multimodales.

Dans la suite nous présentons en détail chacune des bases.

2.3.1 ONERA Christchurch

La base d'image Christchurch est composée d'images aériennes orthonormées d'une résolution de 15 cm/pixel. Elle est issue du jeu de données « NZAM : New Zealand Aerial Mapping Limited, Christchurch after earthquake on 22 February 2011 », p.d. capturées après le tremblement de terre qui a eu lieu le 22 février 2011. 4 images d'une taille de plus ou moins 5000×4000 pixels/image ont été annotées par le laboratoire Département Traitement de l'Information et Modélisation (DTIM) de l'Office National d'Étude et de Recherche Aérospatiale (ONERA) (RANDRIANARIVO, LE SAUX, & FERECATU, 2013). Les objets annotés dans l'image se répartissent de la manière

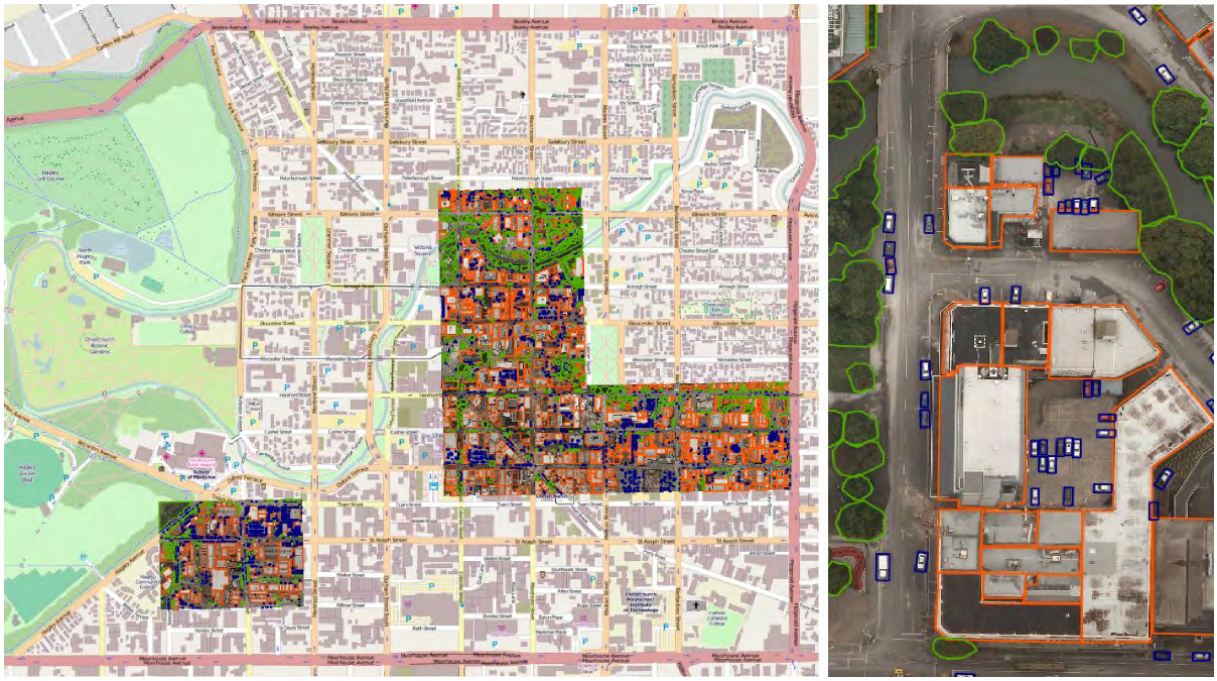


FIGURE 2.2 – Base d'images sur la ville de Christchurch et vérité-terrain créée par nos soins.

suivante : bâtiments (797 instances), voitures (2357 instances) et végétation (938 instances). Une vue des images surimposées sur une carte et des exemples d'objets de la carte sont montrés figure 2.2

La principale particularité de cette base d'images est que Christchurch est une ville anglo-saxonne et donc qu'elle possède un plan en damier (hippodamien). À cause de cette particularité un grand nombre de véhicules dans l'image se trouvent soit à l'horizontale soit à la verticale ce qui peut entraîner un biais pour la mise au point d'une méthode de détection. Cette base d'images a été utilisée pour évaluer notre méthode de détection d'objet en portant un aspect particulier aux voitures.

2.3.2 IEEE GRSS/IADF-TC 2014 Thetford Mines

Dans l'Institute of Electrical and Electronics Engineers (IEEE) IEEE Geoscience and Remote Sensing Society (GRSS), l'Image Analysis and Data Fusion (IADF) organise une compétition le Data Fusion Contest (DFC) [1].

Le DFC2014 avait pour thème la fusion d'images infrarouge à extrêmement haute résolution et des images optiques. La figure 2.3 montre une vues d'ensemble de la base d'image que nous utiliserons pour nos expérimentation.

1. http://www.grss-ieee.org/community/technical-committees/data-fusion

FIGURE 2.3 – Aperçu de la base mise à disposition pour le DFC2014

La compétition du IEEE GRSS/IADF-TC 2014 Thetford Mines [2] est organisée par le comité technique IADF de l'IEEE Geoscience and Remote Sensing Society. L'IADF met à disposition une base composée d'une image optiques haute résolution et d'une image infrarouge.

Les images couvrent une zone urbaine près de Thetford Mines au Québec. Nous avons utilisé cette base pour faire de la détection d'objets et pour cela nous avons nous-mêmes réalisé la vérité terrain. La vérité terrain dans ce cas est l'ensemble des boîtes englobantes qui délimitent certains objets d'intérêt dans l'image (voitures et végétation).

L'image optique possède une résolution de 20 cm/pixel pour une taille de 4386×3769 pixels. L'image infrarouge a été acquise par un imageur à 84 canaux qui couvre les longueurs d'ondes allant de 7,8 à 11,5 µm à une résolution de 1 m/pixel pour une taille de 874×751 pixels.

Pour l'évaluation de la méthode, une partie des images optiques et infrarouges sont données comme base d'entraînement et une partie totalement différente sert pour l'évaluation.

2.3.3 IEEE GRSS/IADF-TC 2015 Zeebrugge

Pour l'année 2015, le DFC avait pour thème la fusion d'images Light Detection And Rangings (LiDARs) à extrêmement haute résolution et des images optiques. La figure 2.4 montre un extrait de la base d'image que nous utiliserons pour nos expérimentations.

2. http://www.grss-ieee.org/community/technical-committees/data-fusion/2014-ieee-grss-data-fusion-contest/

(a) (b)

FIGURE 2.4 – IEEE GRSS/IADF-TC 2015 Zeebrugge. figure 2.4a est l'une des tuiles optiques à une résolution de 5 cm/pixel tandis que la figure 2.4b correspond au Digital Surface Model (DSM) à une résolution de 10 cm/pixel. La base est principalement composée de ports et de zones résidentielles.

Les images ont été acquises le 13 Mars 2011 à partir d'une plateforme aérienne à une altitude de 300 m au-dessus d'une zone résidentielle et d'un port de Zeebrugge en Belgique (51°33′ N, 3°20′ E). Les données ainsi collectées ont été géo-référencées dans le système WGS84. Les orthophotos ont été prises au nadir à une résolution spatiale d'environ 5 cm/pixel. Les balayage, angle et fréquence du laser sont de 125 Hz, 20° et 40 Hz respectivement. La densité de points du capteur LiDAR est approximativement de 65 points/m^2 pour un écart entre les points d'environ 10 cm. Le nuage de points 3D et le DSM sont tous les deux fournis.

La base est composées de 7 tuiles orthorectifées (cf. figure 2.4) composées de :

- Une image optique couleur orthorectifiée de taille 10 000 pixels × 10 000 pixels (format GeoTIFF, RGB à 5 cm/pixel de résolution)
- Une portion du DSM de taille maximale 5000 pixels × 5000 pixels (format GeoTIFF, valeurs à virgule flottante, 10 cm/pixel de résolution)
- Un fichier texte listant les points 3D au format XYZI (latitude, longitude, élévation et intensité LiDAR).

Lors du DFC2015, l'équipe de l'ONERA/DTIM a créé une vérité-terrain avec les classes présentes dans le tableau 2.1, sous la forme de tuiles 10 000 pixels × 10 000 pixels (cf. figure 2.5)

TABLE 2.1 – Vérités-terrain pour la tâche de segmentation sémantique du DFC2015.

Class	*Color*	*RGB code*	nb. pixels	% pixels	objects
Asphalte		(255,255,255)	$235.5 * 10^6$	33.6 %	
Bâtiments		(0,0,255)	$57.6 * 10^6$	8.2 %	
Vegetation basse		(0,255,255)	$75.5 * 10^6$	10.8 %	
Arbre		(0,255,0)	$13.9 * 10^6$	2.0 %	
Voiture		(255,255,0)	$3.2 * 10^6$	0.5 %	955
Fouillis		(255,0,0)	$54.8 * 10^6$	7.8 %	
Bâteau		(255,0,255)	$4.7 * 10^6$	0.7 %	275
Eau		(0,0,128)	$201.0 * 10^6$	28.7 %	

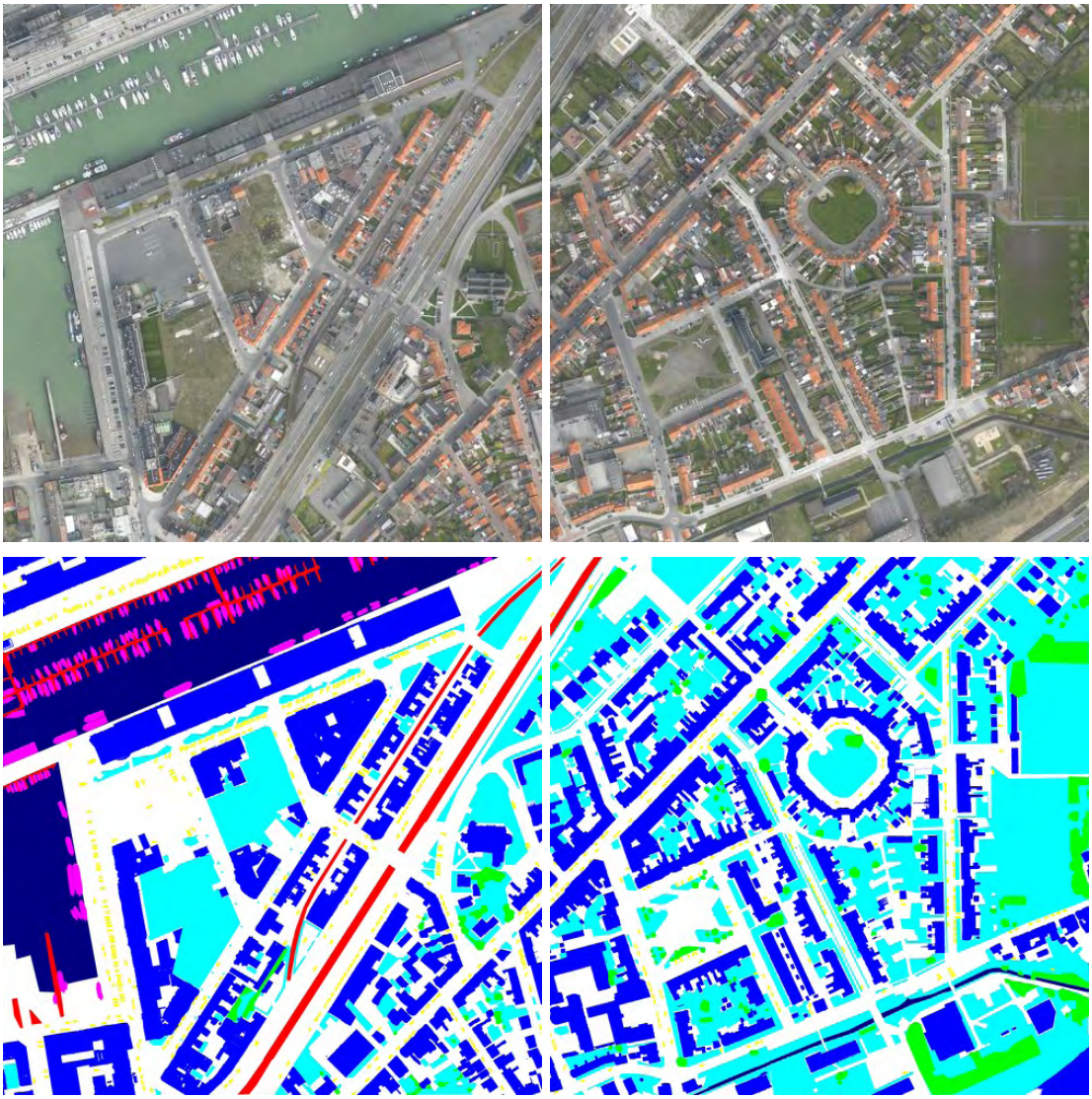


FIGURE 2.5 – Étiquetage sémantique pour le DFC2015 : tuiles orthorectifiées et cartes de labels fournies avec (LAGRANGE et al., 2015).

Pour les expérimentations, nous avons choisi 4 tuiles qui ont constitué l'ensemble d'apprentissage,

1 qui a servi d'ensemble de validation et enfin les deux dernières ont servi d'ensemble de test.

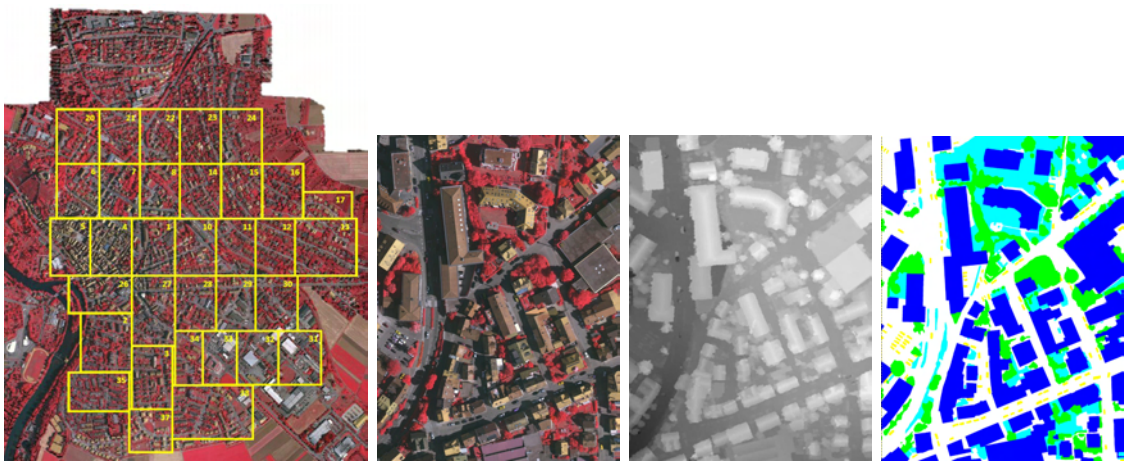
Contrairement à la base Christchurch décrite en section 2.3.1 , la ville de Zeebrugge n'a pas de plan en damier. Pour les véhicules dans l'image cela signifie que le biais sur l'orientation des véhicules est grandement diminué. Les données du modèle d'élévation et LiDAR ne sont pas utilisées par notre méthode pour la détection.

2.3.4 ISPRS Working Group III/4 : Vaihingen segmentation sémantique

Au sein de l'International Society for Photogrammetry and Remote Sensing (ISPRS) le Working Group III/4 (WGIII/4) est chargé de l'analyse de scènes 3D. La base d'image Vaihingen (Rottensteiner et al., 2012) a été mise à disposition par ce groupe dans le cadre d'un concours de segmentation sémantique [3].

Table 2.2 – Vérité-terrain pour la base d'image Vaihingen.

Class	*Color*	*RGB code*	nb. pixels	pixels	objets
Asphalte		(255,255,255)	$21.8 * 10^6$	27.9 %	
Bâtiment		(0,0,255)	$20.4 * 10^6$	26.1 %	732
Végétation basse		(0,255,255)	$16.3 * 10^6$	20.8 %	
Arbre		(0,255,0)	$18.1 * 10^6$	23.3 %	
Voiture		(255,255,0)	$0.9 * 10^6$	1.2 %	1119
Fouillis		(255,0,0)	$0.5 * 10^6$	0.6 %	



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

Figure 2.6 – Base ISPRS Vaihingen pour la segmentation sémantique. figure 2.6a est une vue d'ensemble des 33 tuiles. Les figures 2.6b à 2.6d sont respectivement l'orthophoto à une résolution de 9 cm/pixel, le DSM et la vérité-terrain du concours (asphalte, bâtiments, végétation basse, arbres, voitures et fouillis) pour la tuile #1.

La base d'images est composée de 33 images orthorectifiées capturées au-dessus de la ville de Vaihingen en Allemagne (cf. figure 2.6)

3. http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html

Les images de cette base ont la particuliarités d'être en IR-R-G au lieu de RGB. L'ajout de l'infrarouge permettant une meilleure caractérisation des catégories liées à la végétation.

3

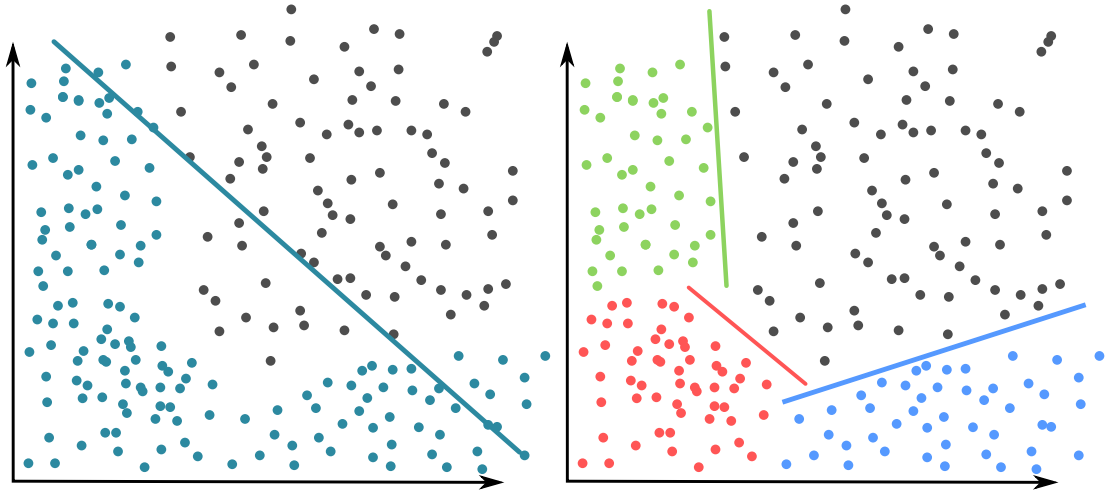
Recherche de sous-catégories visuelles

Sommaire

Dans ce chapitre nous nous intéressons à la recherche de sous-catégories visuelles dans une catégorie sémantique d'objets de manière non supervisée. Lors de la réalisation d'une base d'image un compromis est réalisé entre deux impératifs : maximiser le nombre de catégories que la base va contenir et minimiser la complexité de la réalisation de la vérité terrain qui augmente avec le nombre de catégories et le nombre d'objets par catégorie. C'est principalement pour ces raisons que les catégories d'objets dans les base d'images sont définies par des concepts sémantiques larges tel que "voiture" ou "végétation". Au sein d'une catégorie sémantique, par exemple les voitures, il est ainsi possible d'affiner la classification en fonction des modèles de véhicules puis en fonction de la couleur du véhicule etc.

L'un des principaux problèmes pour la détection d'objets dans des images est l'apprentissage d'un modèle d'une catégorie qui soit suffisamment représentatif de l'ensemble des instances de la classe (et qui puisse se généraliser à de nouvelles instances). Une manière de simplifier ce problème est de chercher de manière non supervisée des sous-catégories d'objets qui permettent de regrouper ensemble les instances d'un objet visuellement similaires. Une catégorie est alors définie par plusieurs modèles qui représentent chacun un sous-ensemble des exemples d'apprentissage (cf. figure 3.1)

Chaque sous-catégorie est représentée par un sous-modèle qui la modélise très précisément par rapport à un modèle appris sur l'ensemble des exemples. La difficulté ici est de trouver un critère permettant de quantifier la similarité entre deux exemples pour les placer dans une même sous-catégorie. Si les instances des sous-catégories ne sont pas visuellement proches l'approche décrite précédemment n'est plus valable. Cependant les méthodes d'apprentissage statistique



(a) (b)

FIGURE 3.1 – On observe sur la figure 3.1a la représentation d'une catégorie d'objets. Trouver une séparatrice linéaire qui divise parfaitement les exemples positifs en bleu et les exemples négatifs en noir est impossible dans ce cas de figure car la frontière de décision est non-linéaire. La figure 3.1b représente la même catégorie d'objets mais les sous-catégories ont été mises en évidence. Apprendre une séparatrice linéaire pour chaque sous-ensemble vis-à-vis du reste de la base est un problème beaucoup plus simple.

ont généralement besoin d'un certain nombre de données pour apprendre un modèle robuste. Quand nous divisons les exemples entre les sous-catégories nous diminuons le nombre d'exemples d'apprentissage par modèle et donc diminuons la capacité de généralisation du modèle. Plusieurs approches pour résoudre ce problème ont été étudiées pour des problématiques de détection d'objets. (DIVVALA, HOIEM, HAYS, EFROS, & HEBERT, 2009) montre que ce qui rend le détecteur d'objets de (FELZENSZWALB, GIRSHICK, MCALLESTER, & RAMANAN, 2010) si performant réside principalement dans le mélange de modèles qu'il utilise. (MALISIEWICZ, GUPTA, & EFROS, 2011) a poussé le concept de sous-catégories en apprenant un modèle par instance d'objet et montré que cette approche permettait d'améliorer un détecteur d'objets basé sur du HOG. (GU, ARBELÁEZ, LIN, YU, & MALIK, 2012) a proposé une approche très similaire à la nôtre en modélisant des sous-catégories pour la détection d'objets, cependant pour chaque sous-catégorie ils choisissent un exemple source et recherchent les autres exemples dont la pose correspond à cette exemple. Notre méthode elle s'appuie sur des critères permettant d'estimer automatiquement la pose des objets dans l'image sans avoir besoins d'exemples de références.

Nous allons d'abord présenter notre analyse des catégories d'objets dans une base d'images en section 3.1, puis nous présenterons notre méthode pour apprendre automatiquement comment retrouver les sous-catégories visuelles d'un classe sémantique en section 3.2 et finalement nous présenterons les résultats de notre méthode en section 4.4.

3.1 Analyse de catégories sémantiques

3.1.1 Études des objets dans une catégorie

Dans une base d'image, pour des besoins de classification les objets sont regroupés selon différentes catégories sémantique facilement interprétable par l'homme. Ces catégories sont désignés par des noms génériques tel que "voiture", "chien" ou encore "bâtiment". Ainsi dans une catégorie sémantique il existe une importante variation d'apparence entre les différentes instance d'objets qui la compose.

(a) (b) (c)

FIGURE 3.2 – Dans les bases d'images, les chien des différentes races sont labellisé sous la même catégorie sémantique "chien" alors que visuellement chaque race possède ses propres particularités physique qui la rend très différente d'une autre race.

Par exemple la figure 3.2 montre cette variation d'apparence pour la catégorie sémantique "chien". Les chiens issues de différentes races peuvent être très différents physiquement. C'est cette différence d'apparence qui explique en partie la difficulté pour détecter des objets dans des images. Pour la détection d'objets un problème encore plus difficile à résoudre est de modéliser la pose d'un objet dans l'image. Le problème de l'estimation de la pose d'un objet dans une image est aussi l'une des raisons pour laquelle la détection d'objet est un problème difficile. Nous observons aussi que les chiens ont différentes poses dans chacune de ces images. La modélisation de la variation de pose des objets est l'une des grandes difficultés pour modéliser les objets dans une image. C'est l'une des raisons qui rend les approches de type correspondance de modèles rigides peu performantes car ces méthodes ne modélisent généralement qu'un seul aspect de l'objet. Dans ce travail, nous considérons tous les éléments qui font varier l'apparence des instances d'une catégorie sémantique comme étant les caractéristiques des sous-catégories visuelles que nous voulons capturer. Dans le cas de l'imagerie aérienne, c'est à dire de l'imagerie vue du haut, la variation d'apparence d'un objet est généralement moins importante que dans des images du type de la figure 3.2. Cependant la variation de la pose des objets, principalement l'orientation dans ce cas précis, quant à elle est très importante car les objets d'une même catégorie peuvent être dans toutes les orientations possibles. Un exemple est les voitures de la figure 3.3.

Dans cette image nous pouvons observer un large échantillon des différentes poses que peuvent avoir un objet d'intérêt dans une image aérienne. Nous observons aussi une légère variation de taille entre les différents modèles de voitures. D'après cette analyse, les instances d'une catégories sémantique ont une apparence qui varie fortement en fonction de plusieurs critères comme l'orientation ou la taille. Maintenant que sont identifiés les principales causes qui perturbent la reconnaissance d'un objet, nous allons chercher une méthode permettant de caractériser ces variations pour obtenir des sous-catégories homogènes.

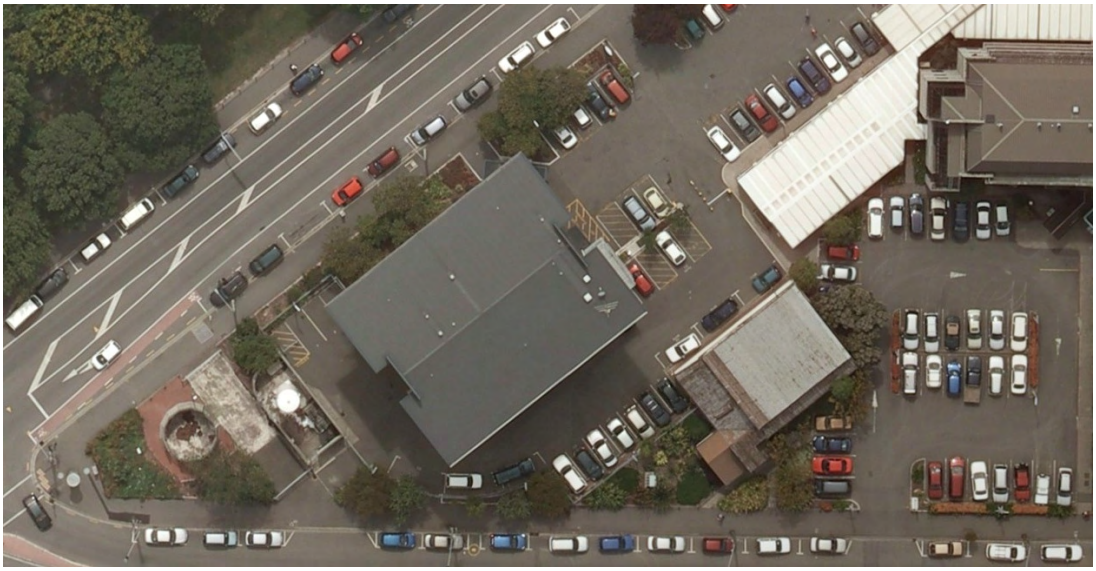


FIGURE 3.3 – Extrait d'une des images d'apprentissage de Christchurch. Cette image donne une idée des principales variation d'apparence que peuvent avoir les objets de la catégorie voiture.

3.1.2 Caractérisation de sous-catégories d'objets

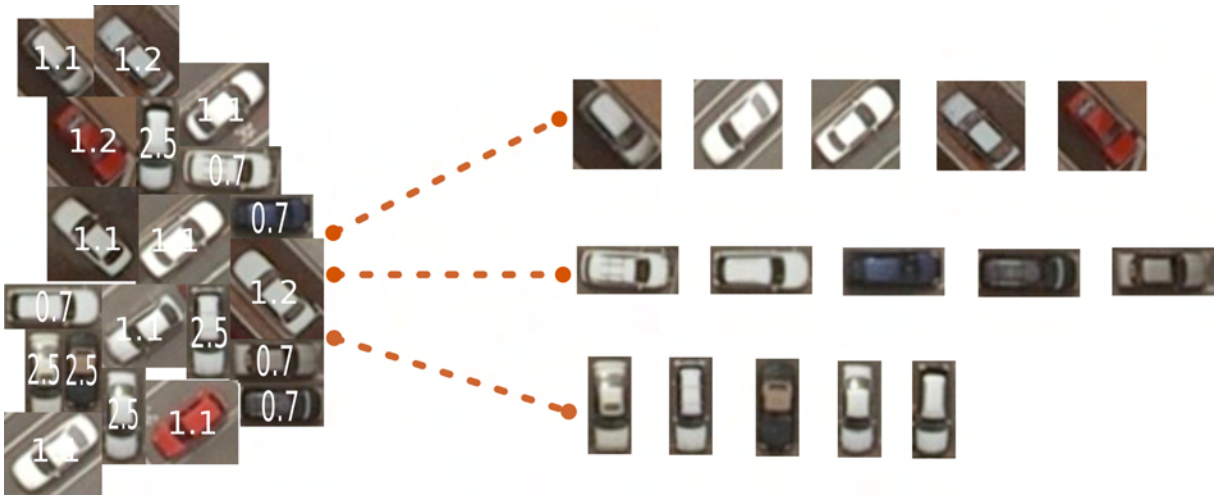


FIGURE 3.4 – À gauche nous avons les exemples positifs de la catégorie sémantique que nous voulons analyser, à droite nous avons le partitionnement de cette catégorie en différentes sous-catégories visuelles.

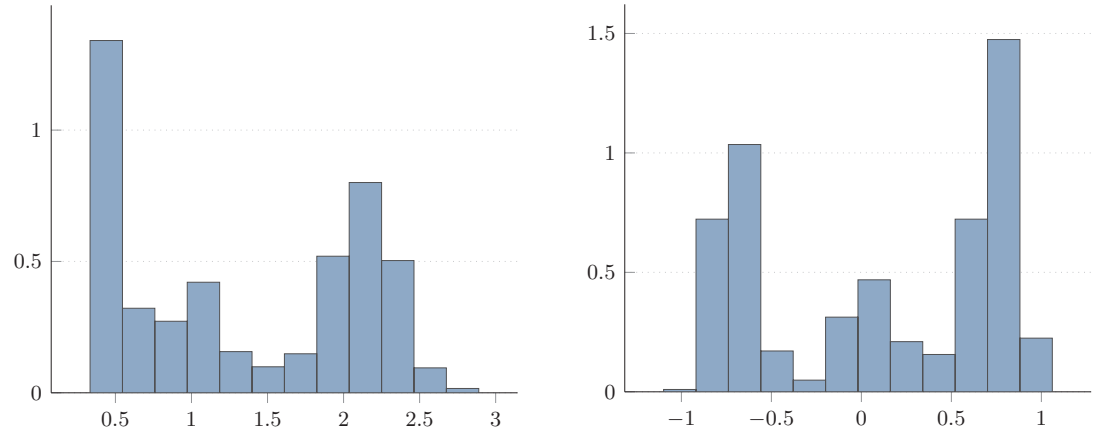
La recherche de sous-catégorie visuelles passe par une étape de caractérisation des objets de la catégorie sémantique. Cette caractérisation consiste en l'extraction de caractéristiques qui vont permettre de regrouper ensemble les instances ayant une pose similaire comme dans la figure 3.4. L'estimation de la pose d'un objet dans une image est une thématique de recherche très étudiée en vision par ordinateur. Cependant pour notre application nous n'avons pas besoin d'estimer cette pose avec une grande précision mais plutôt de trouver une estimation de la pose qui puisse permettre de retrouver une orientation approximative de l'objet. Dans un premier temps nous avons utilisé les informations disponible dans les annotations pour résoudre ce problème. En effet juste avec le rapport de format nous pouvons estimer efficacement l'orientation d'un objet. Ce rapport est défini de la façon suivante.

$$r(x) = \frac{\text{largeur}(x)}{\text{hauteur}(x)} \tag{3.1}$$

Selon la valeur de r, on peut estimer une orientation générale de l'objet comme par exemple si l'objet est orienté horizontalement, verticalement ou en diagonal. La figure 3.5a montre un histogramme de la distribution de ce rapport dans la base d'images Christchurch. Cette figure nous montre que tel quel ce rapport ne peut être utilisé pour la recherche de sous-catégories. Si nous utilisation le rapport de cette manière les algorithmes de partitionnement ne pourront pas trouver des sous-catégories homogènes. Par exemple la sous-catégorie des exemples verticaux à un rapport compris dans $[0, 1]$ et la catégorie des exemples diagonaux un rapport aux alentours de 1. Une grand partie des exemples ont un rapport compris entre 0 et 1 alors que la gamme les valeurs possible pour le rapport est définie sur R^+ donc en appliquant un algorithme de partitionnement uniquement sur le rapport de format ces instances seront très probablement classées dans la même sous-catégorie malgré le fait qu'elles ne soient pas visuellement homogènes. Pour pallier à ce problème nous redéfinissons la caractéristique d'orientation r par :

$$r(x) = \log\left(\frac{\text{largeur}(x)}{\text{hauteur}(x)}\right) \tag{3.2}$$

La fonction logarithme permet de traiter de rendre symétrique par rapport à 0 les valeurs du rapport de format entre 0 et 1 en les projetant entre $[-\infty, 0]$ et les valeurs entre $[1, \infty]$ en les projetant entre $[0, \infty]$. La figure 3.5b montre la nouvelle distribution de la caractéristique d'orientation sur la base Christchurch. On remarque que maintenant des modes dans l'histogramme sont clairement visibles. Ces modes correspondent aux orientations horizontales, verticales et diagonales des objets dans l'images.



(a) (b)

FIGURE 3.5 – Histogrammes normalisés des caractéristiques d'orientations des objets voitures dans la base d'image. La figure 3.5a est l'histogramme utilisant le rapport de format tandis que la figure 3.5b est l'histogramme construit à partir du logarithme du rapport de format.

Bien que cette caractéristique permette de facilement retrouver une orientation principale d'un objet elle souffre d'un inconvénient majeur, elle n'utilise aucune information visuelles liée à l'objet. Comme le contenu de l'image n'est pas utilisé par cette caractéristique, la classification dans les différentes sous-catégories dépend largement de la qualité des annotations. De plus ne pas utiliser d'informations visuelle entraîne une certaine confusion lors de la recherche des

sous-catégories. La figure 3.6 montre un exemple de problème que l'on rencontre en créant des sous-catégories uniquement à partir des annotations. Les instances de cette sous-catégorie ont bien tous un rapport de format très proche cependant elles différent en orientation ce qui produit une sous-catégorie encore non homogène visuellement.

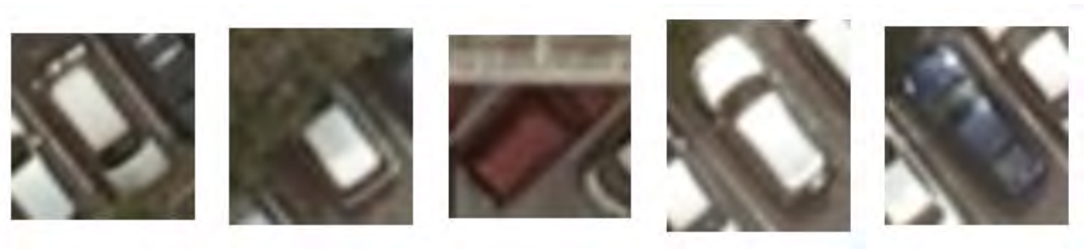


FIGURE 3.6 – Exemple d'instances de la catégories voitures dont l'orientation estimée est diagonale. On remarque que les objets de cette sous-catégorie bien que tous diagonaux ne sont pas orienté dans la même direction

Une des manière de pallier ce problème serait donc d'affiner les sous-catégories trouvées grâce à la caractéristique d'orientation avec un second partitionnement qui serait basé sur les informations visuelles de l'image. Un partitionnement appliqué directement sur les valeurs des pixels d'une image serait trop bruité pour obtenir des sous-catégories satisfaisantes visuellement. À la place d'appliquer le partitionnement directement dans l'espace des pixels nous préférons utiliser un descripteur du contenu d'une image de l'état de l'art. Le descripteur choisit pour cette tâche est le HOG (DALAL & TRIGGS, 2005). Le HOG est un descripteur d'apparence de l'objet, il permet de quantifier la distribution de l'orientation des contours de l'objet. Dans notre cas nous voulons regrouper ensemble les instances d'une catégorie ayant une orientation similaire, donc les instances ayant un descripteur HOG similaire. La figure 3.7 montre une visualisation du descripteur pour des instances de voitures. Visuellement on observe que les descripteurs varient en fonction de l'orientation de la voiture dans l'image. Notre intuition est qu'un algorithme de partitionnement bien choisi doit permettre de retrouver ces orientations que nous observons.

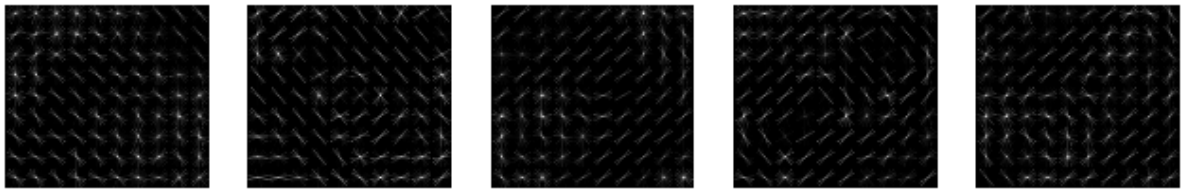


FIGURE 3.7 – Visualisation du HOG d'instances de voitures dont l'orientation estimée est diagonal.

Bien que le HOG offre une description de l'apparence utilisable directement par un algorithme de partitionnement la recherche de sous-catégories directement à partir du HOG ne donne pas de très bons résultats. La variance des instances dans la base est en effet trop élevée et le HOG pas assez discriminative pour obtenir des sous-catégories homogènes. Notre méthode de recherche de sous-catégories va combiner hiérarchiquement le partitionnement à partir des annotations et le partitionnement à partir des informations visuelles pour obtenir des sous-catégories visuellement homogènes.

3.2 Apprentissage de sous-catégories visuelles

3.2.1 Algorithmes de partitionnement

3.2.1.1 Gaussian Mixture Model/Mélange de Modèles Gaussiens

Nous présentons dans cette section une méthode de partitionnement basée sur un mélange de modèles. Notre objectif étant de partitionner les descripteurs d'apparence des objets d'une catégorie sémantique. Le partitionnement par mélange de modèle modélise un ensemble d'échantillons $\{x_i \mid i = 1, \ldots, N\}$ comme étant les observations d'une variable aléatoire X définit sur un espace $\mathbb{R}$ suivant une loi de mélange fini p. Cette loi de mélange est défini comme une combinaison linéaire de lois de probabilités $\{p_k \mid k = 1, \ldots, K\}$. Le mélange de modèles s'exprime en fonction des coefficients $\{\pi_k \mid k = 1 \ldots, K\}$ de la combinaison linéaire :

$$p(x) = \sum_{k=1}^{K} \pi_k p_k(x) \tag{3.3}$$

avec

$$\sum_{k=1}^{K} \pi_k = 1 \tag{3.4}$$

$$\pi_k > 0 \ \forall k \in 1, \ldots, K \tag{3.5}$$

Les p_k étant les composantes du mélange et les coefficients π_k étant la proportion de la composante p_k dans le mélange.

Nous devons maintenant estimer l'ensemble des paramètres $\theta = \{(\pi_k, \alpha_k) \mid k = 1, \ldots, K\}$ du mélange. Dans la suite, nous considérerons que les composantes du mélange p_k suivent des lois normales $p_k \sim \mathcal{N}(\mu_k, \sigma_k)$ dont les paramètres servent à définir les paramètres α_k du mélange : $\alpha_k = (\mu_k, \sigma_k^2)$.

Nous cherchons à déterminer les valeurs du paramètre θ à partir d'un ensemble d'observations $(x_1, \ldots, x_N)$ d'une variable aléatoire X qui suit une loi $p(X, \theta)$. Une méthode d'estimation de paramètres consiste à trouver la valeur de θ qui maximise la log-vraisemblance de $p(X, \theta)$

$$l(\theta; x) = \sum_{i=1}^{N} \sum_{k=1}^{K} \ln(\pi_k p(x_i, \alpha_k)) \tag{3.6}$$

En pratique l'optimisation des paramètres θ est difficile car le plus souvent il n'existe pas le formule analytique pour résoudre le problème d'optimisation et les algorithmes d'optimisation de type Newton peuvent être compliqués à mettre en oeuvre sur ce problème. Une méthode plus efficace existe, il s'agit de l'algorithme Expectation-Maximization (EM). L'algorithme EM est un algorithme conçu pour résoudre spécifiquement les problèmes de maximisation de vraisemblance dans le cas de données manquantes (Dempster, Laird, & Rubin, 1977). La mise en oeuvre de l'algorithme EM se met en place simplement si on peut exprimer la log-vraisemblance complété des observations.

Une information cachée dans le modèle est que chaque observation appartient à l'une des K catégories du mélange. À chaque observation x_i on peut associer une variable latente $z_{i,k}$ qui vaut 1 si x_i appartient à la catégorie K et 0 sinon. Cependant comme les variables $z_{i,k}$ ne sont pas directement observable on dit qu'elles sont latentes. À partir de cette information nous pouvons

réécrire équation (3.6) en fonction des variables latentes. Cette fonction est la log-vraisemblance complétée :

$$l(\theta; x, z) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{i,k} \ln(\pi_k p(x_i, \alpha_k)) \tag{3.7}$$

L'algorithme EM se décompose en deux étape pour maximiser équation (3.7) :

- Une étape d'espérance E qui consiste à calculer

$$q_{i,k}^{t+1} = \mathbb{E}[z_{i,k} \mid x_i; \theta_n] \tag{3.8}$$

Cette étape peut aussi être interprétée comme la probabilité conditionnelle qu'une observation x_i ait été générée par la composante k du mélange.

$$\begin{aligned} q_{i,k} &= p(Z_{i,k} = 1 \mid X = x_i; \theta) \\ &= \pi_k \frac{p(x_i; \alpha_k)}{p(x_i; \theta)} \end{aligned} \tag{3.9}$$

- Une étape de maximisation $\mathbb{M}$ qui consiste à calculer :

$$\theta^{t+1} = \arg\max_{\theta} l\Big(x, q^{n+1}; \theta\Big) \tag{3.10}$$

Pour le modèle gaussien, le calcul des paramètres (π_k, μ_k, σ_k) est donné à l'itération t par :

$$\pi_k^{t+1} = \frac{\sum_i^n q_{i,k}^{t+1}}{n} \tag{3.11}$$

$$\mu_k^{t+1} = \frac{1}{n_k^{t+1}} \sum_i^n q_k^{t+1} \tag{3.12}$$

$$\sigma_k^{t+1} = \frac{1}{n_k^{t+1}} = \sum_i^n q_{i,k}^{t+1} (x_i - \mu_k^{t+1})(x_i - \mu_k^{t+1}) \tag{3.13}$$

La recherche des paramètres θ qui maximisent la log-vraisemblance s'arrête soit lorsque la stationnarité du critère est atteinte soit après un nombre d'itération prédéfinit. La stationnarité découle de la croissance de la fonction $l(\theta; x, z)$.

Avec le mélange de modèle l'appartenance à une sous-catégorie d'une observation x_i se calcul de façon similaire à la phase d'espérance.

$$k' = \arg_{k \in K} \max \pi_k p(x_i; \alpha_k) \tag{3.14}$$

3.2.1.2 Partitionnement avec la méthode de Ward

La méthode de Ward est une méthode de partitionnement hiérarchique se basant sur la minimisation de l'inertie des partions. Les méthodes de partitionnement hiérarchique sont une famille de méthodes où le partitionnement des individus est obtenu par agrégation successif des partitions.

Ces méthodes peuvent être représentées par un arbre de classification (ou dendrogramme) pour visualiser les différentes partitions. En règle générale l'arbre de classification est construit en partant des feuilles jusqu'à la racine : on les appelle des méthodes montantes. Cependant de manière plus marginale il existe des méthodes dites divisives ou descendantes qui partent de la racine. Les partitions dans un dendrogramme sont obtenues par agrégation des noeuds les plus proches par rapport à une métrique. L'algorithme 1 détaille la méthode de classification.

Algorithme 3.1 : Algorithme général pour le partitionnement hiérarchique

Data : $X = \{x_i \mid i = 1, \ldots, N\}$ la population à partitionner, K le nombre de partitions à trouver
Result : Arbre de classification P
begin
  Initialisation des partitions $P = \{p_i \mid 1, \ldots, N\} \leftarrow \{x_1, \ldots, x_n\}$
  while $|P| > K$ **do**
    $(p_i, p_j) = \arg\min_{A,B \in P} \Delta(A, B)$
    $p_i \leftarrow$ `fusion`(p_i, p_j)
    `supprimer`(p_j)
  end
end

On peut noter qu'en fonction des données, le choix de la métrique va fortement influencer la constitution des différentes partitions. La méthode partitionne une population selon un critère qui va chercher un minimum local de l'inertie intraclasse (et selon le théorème de Huygens cela revient à chercher un maximum local de l'inertie inter classes) à chaque itération. Pour cela on définit une mesure de dissimilarité $\Delta(p_i, p_j)$ entre deux partitions. Cette mesure de dissimilarité quantifie la perte d'inertie lors du regroupement de deux partitions. Elle est définie de la façon suivante pour une paire de partitions (p_i, p_j) :

$$\begin{aligned}\Delta(A, B) &= \sum_{i \in A \cup B} \|x_i - \vec{m}_{A \cup B}\|^2 - \sum_{i \in A} \|\vec{x}_i - \vec{m}_A\|^2 - \sum_{i \in B} \|\vec{x}_i - \vec{m}_B\|^2 \\ &= \frac{n_A n_B}{n_A + n_B} \|\vec{m}_a - \vec{m}_B\|^2\end{aligned} \tag{3.15}$$

Avec $\vec{m}_j$ et n_j étant respectivement le centre et le nombre d'éléments dans la partition j. À la première itération de l'algorithme ce coût est nul car chaque partition est constituée d'un seul élément. On note aussi que la méthode de Ward cherche à augmenter le coût $\Delta(p_i, p_j)$ du plus petit incrément possible pour cela lorsque deux partitions ont une perte d'inertie égale la méthode choisira de grouper les partitions avec le plus petit nombre d'éléments.

3.2.2 Partitionnement pour la recherche de catégories visuelles

Notre méthode de partitionnement des instances d'une classe s'applique en deux temps. Dans un premier temps nous partitionnons les instances de la catégorie en utilisant un GMM sur le rapport de format des instances. Cela nous donne un premier sous-ensemble de sous-catégories où les éléments d'une sous-catégories possèdent une orientation similaire. On remarque qu'à un facteur d'échelle près les éléments de ces sous-catégories ont des dimensions très similaires ce qui nous aidera plus tard pour l'apprentissage du modèle de la sous-catégorie. À partir de ce premier

sous-ensemble de sous-catégories nous allons pouvoir appliquer un deuxième partitionnement sur l'apparence visuelle des éléments. Grâce à la première étape de partitionnement les éléments des premières sous-catégories ont déjà des poses visuellement proches. Ce second partitionnement va permettre d'obtenir des sous-catégories plus fines où toutes les voitures sont orientées vers la gauche où vers le haut par exemple. Pour se faire nous redimensionnons d'abord les éléments de la sous-catégories pour qu'ils aient les mêmes dimensions. Comme les éléments des sous-catégories ont déjà des dimensions très proches on minimise ainsi les risques de modifier l'apparence des objets lors du redimensionnement. Ensuite nous calculons le descripteur HOG de chaque élément et nous appliquons la méthode de Ward pour le partitionnement des descripteurs. La méthode de Ward va agréger ensemble les exemples dont les descripteurs HOG sont proches vis-à-vis de la distance euclidienne. Les exemples proches en utilisant la distance euclidienne sont généralement les descripteurs dont la distribution des gradients est proche et donc des exemples ayant une apparence visuelle proche. La figure 3.8 illustre la répartition dans les différentes sous-catégories des instances de la catégorie voiture pendant la procédure de partitionnement.

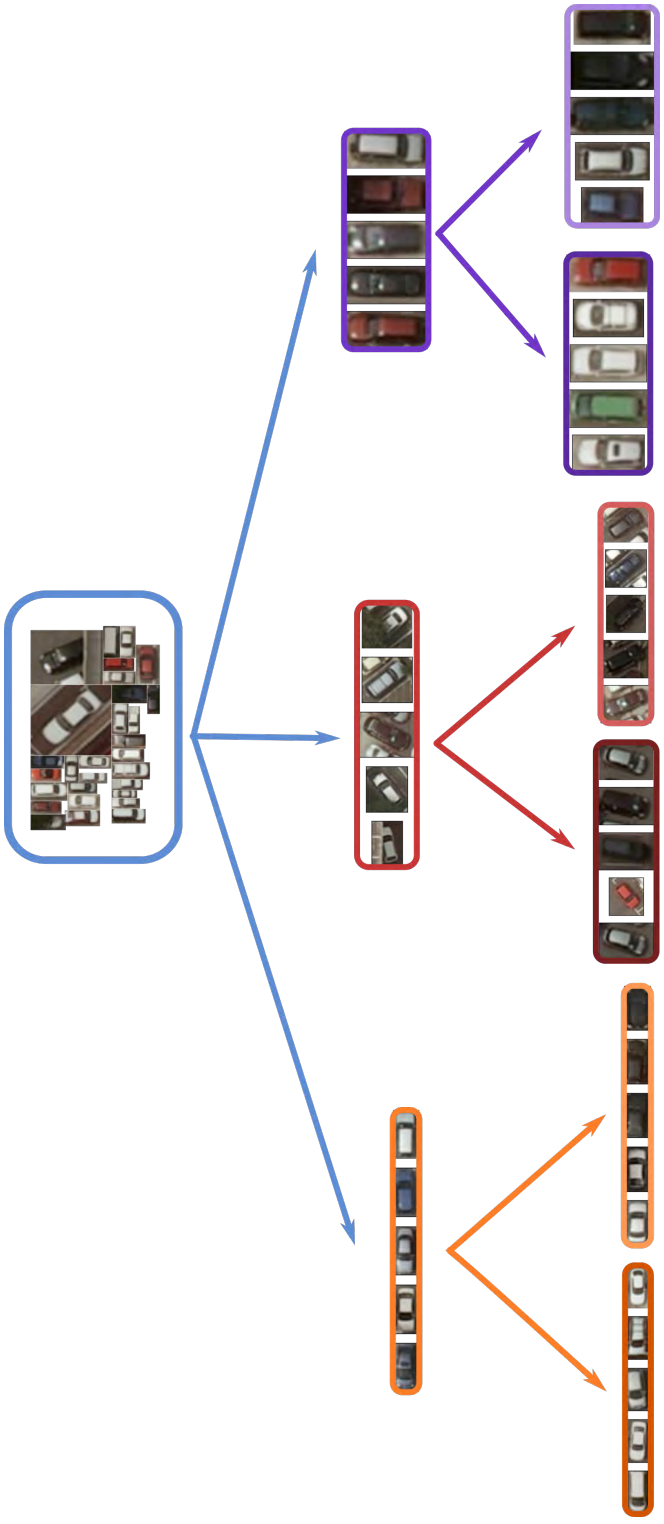


FIGURE 3.8 – Vue d'ensemble de notre méthode de recherche de sous-catégories visuelles. Le premier étage de la méthode est un partitionnement sur le rapport de format de l'instance, le second étage est un partitionnement réalisé sur l'apparence de l'instance

3.3 Résultats expérimentaux

Cette section présente les différents résultats que nous avons obtenus avec notre méthode. Afin de tester la généralisation de notre méthode nous la testons sur différentes bases d'images aérienne. Nous présentons d'abord les résultats qui nous ont permis de choisir le nombre de modèles pour le GMM. Puis pour les deux méthodes de partitionnement nous allons montrer les différentes sous-catégories obtenues.

3.3.1 Choix du nombre de modèles dans le mélange

Le problème principal du partitionnement des exemples d'apprentissage est que l'on doit prendre en compte deux finalités qui peuvent être contradictoires. D'un côté nous devons trouver une partition des exemples où les éléments de chaque partition sont visuellement très proche, un cas extrême étant le cas où chaque partition n'est constituée que d'un seul élément. D'un autre côté il faut que chaque partition contienne suffisamment d'exemples pour pouvoir apprendre un modèle de la sous-catégorie statistiquement robuste. La robustesse du modèle dépendant fortement du nombre d'exemples d'apprentissage dans la partition.

Pour estimer le nombre de paramètres dans le mélange permettant au mieux de représenter une sous-catégorie en fonction du descripteur d'orientation nous utilisons les méthodes suivantes :

- Un histogramme pour visualiser la répartition du logarithme du rapport de format. Cet histogramme permet de visualiser les différentes modalités présentes dans les données où chacune des modalités correspondant à une sous-catégorie
- Un graphique représentant la valeur du BIC pour différents paramètres de la GMM. Comme le partitionnement par GMM peut produire des partitions légèrement différentes entre deux essais, nous calculons le BIC sur 5 itérations et représentons la valeur moyenne ainsi que l'écart type sur le graphique. L'étoile représente le nombre de composantes qui minimise le BIC et donc le nombre de sous-catégories optimale. Pour illustrer notre méthode nous allons étudier la catégorie voitures présente dans les bases d'images.

3.3.1.1 Répartition des rapports de format

Dans un premier temps nous voulons pouvoir visualiser la répartition des orientations dans la base d'images. Connaître cette répartition nous permettra d'estimer visuellement en fonction des modes de l'histogramme le nombre de composantes dans la GMM nécessaire pour modéliser la catégories voitures. La figure 3.9 montre la répartition des orientations pour la classe voitures dans la base d'images de Christchurch. Afin de faciliter la recherche des modes de l'histogramme, nous affichons sur l'histogramme l'estimation par un noyau gaussien Parzen, 1962 de la distribution des orientations. Pour la catégorie voiture nous observons sur figure 3.9 qu'il existe 3 modes principaux que nous interprétons comme étant 3 sous-catégories cachées dans la catégorie voitures.

La figure 3.10 montre le même résultat mais sur les données extraites de la base d'image du DFC2015. Cette fois ci l'histogramme semble trop bruité pour déterminer de manière sûre le nombre de modes. Bien que la visualisation de l'histogramme des descripteurs d'orientations permette d'avoir un a priori sur les données à partitionner, nous allons utiliser un critère d'évaluation qui nous permettra de récupérer automatiquement le nombre de sous-catégories cachées en fonction du descripteur d'orientation.

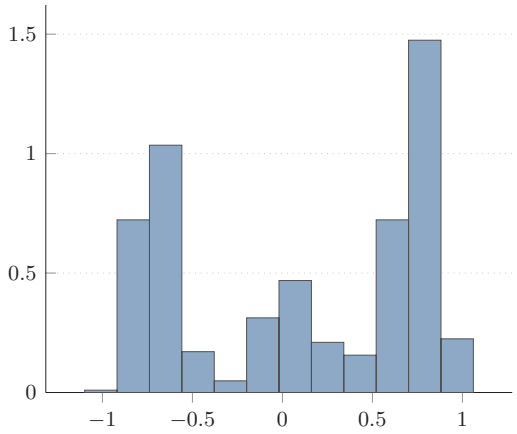


FIGURE 3.9 – Répartition des orientations pour la catégorie voitures sur la base d'images Christchurch. Nous observons sur cette figures 3 modes principaux

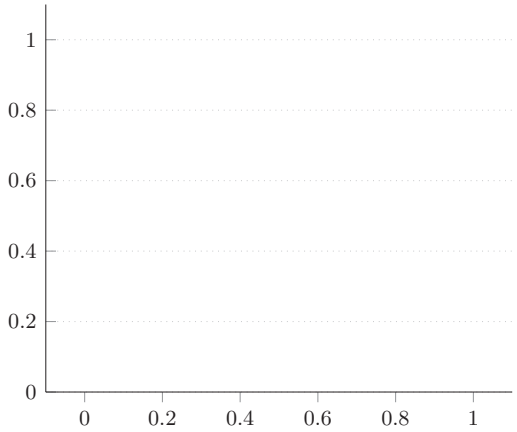


FIGURE 3.10 – Répartition des orientations pour la catégorie voitures sur la base d'images du DFC2015. Nous ne pouvons observer directement le nombre de modes.

3.3.1.2 Choix du nombre de modèles

La caractérisation de la catégorie sémantique se fait en la subdivisant en un ensemble de sous-catégories visuelles. Un paramètre critique pour de notre méthode est le nombre de sous-catégories visuelles que nous allons utiliser pour décrire la catégorie sémantique. Pour la première étape qui consiste à trouver des sous-catégories en fonction d'un descripteur d'orientations nous utilisons le BIC pour estimer le nombre de composantes du mélange. Ce critère pénalise les mélanges avec un grand nombre de composantes tout en favorisant les mélanges où la vraisemblance aux données est forte. Le choix de favoriser les mélanges avec peu de composantes se justifie par le fait que plus un mélange contient de composantes, moins ces composantes contiennent d'instances d'un objet ce qui rend moins robuste le modèle final pour la détection.

Les figure 3.11 et figure 3.14 montrent les résultats obtenus pour la catégorie d'objets voitures. Dans ces expérimentations, nous avons appris des sous-catégories en utilisant notre descripteur d'orientations ainsi qu'un modèle GMM. Nous avons fait varier le nombre de composantes dans le mélange et calculer le BIC. Afin d'assurer la répétabilité de la méthode, nous avons répété ce processus 5 fois. Les figures 3.11 et 3.14 montrent la moyenne et la variance du BIC pour chacune des composantes. La figure 3.11 montre le résultat du calcul du BIC sur la base d'images Christchurch. Sur cette figure, on observe que la valeur du BIC est optimale pour 3 (cf. figure 3.12)

et 5 à 15 composantes. On note que pour 1, 2 et 4 les valeurs du BIC laissent à penser que les sous-catégories seront très bruitées comme sur la figure 3.13

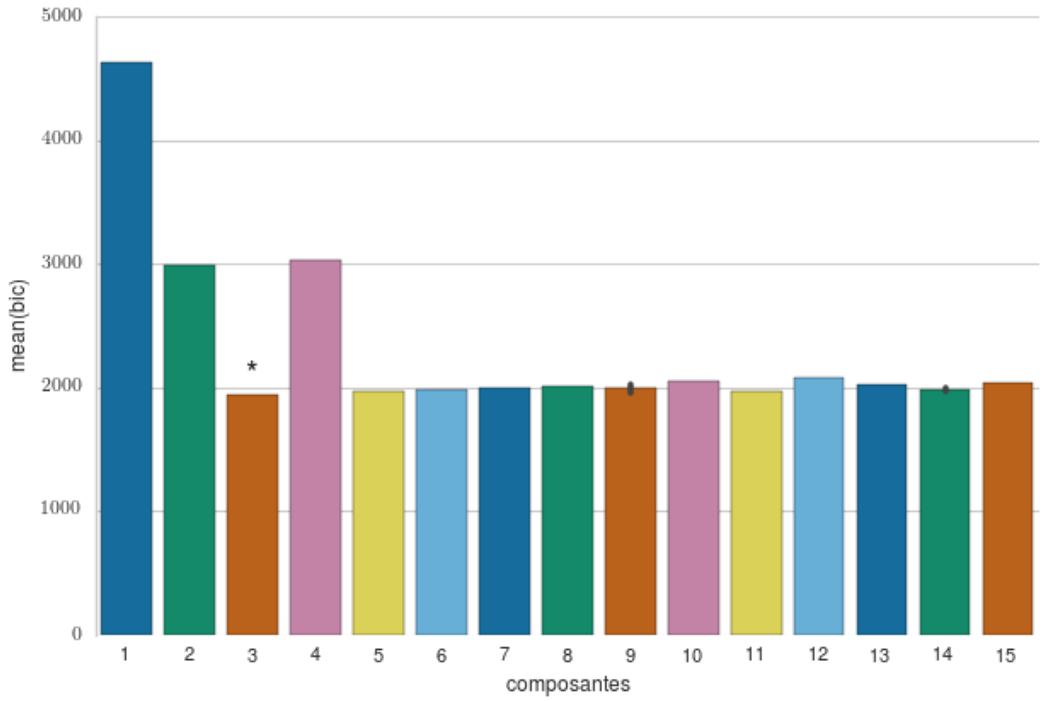


FIGURE 3.11 – BIC moyen sur la base d'image Christchurch pour la catégorie voitures. Les mélanges varient de 1 à 15 composantes par mélange. Le tiret noir au-dessus d'une barre représente la variance du score et l'étoile le mélange avec le nombre optimal de composantes vis-à-vis du critère.

La ville de Christchurch est une ville anglo-saxonne où la plupart des routes sont perpendiculaires entre elles. Cela signifie que les principales orientations des véhicules seront horizontales, verticales et en diagonale. La visualisation des sous-catégories trouvée avec un mélange de 3 composantes est présenté en figure 3.12, on observe bien que les orientations principales des véhicules sont retrouvées. Cependant comme ici nous n'avons pas utilisé d'information visuelle, on observe que la sous-catégorie correspondant aux véhicules diagonaux est très bruitée car ces véhicules peuvent être soit orientés vers la gauche, soit vers la droite.

Les figures 3.14 et 3.15 montrent les même résultats que précédemment mais sur la base du DFC2015. Ici le nombre optimal de sous-catégorie est 2. Cela s'explique car dans la ville de Zeebrugge contrairement à Christchurch les rues ne sont pas agencées selon un plan hippodamien, donc en moyenne les voitures sont plutôt orientées vers à verticale et plutôt orientées à l'horizontal comme montré figure 3.15.

3.3.2 Résultats de partitionnement

Nous allons maintenant observer les résultats obtenus en incorporant des informations visuelles dans la procédure de partitionnement. À partir des sous-catégories obtenues à partir des orientations nous allons appliquer un partitionnement sur chacune des nouvelles sous-catégories en utilisant la méthode de Ward et le descripteur HOG.

FIGURE 3.12 – Partitionnement des exemples de la base Christchurch en 3 sous-catégories pour la classe voiture. Pour chacune des sous-catégories nous avons d'abord affiché l'image moyenne de la sous-catégorie pour visualiser globalement la variation d'apparence puis 5 exemples tirés aléatoirement dans la sous-catégorie.

Comme précédemment la première ligne de l'image montre un exemple moyen de la sous-catégorie et la seconde ligne de l'image un échantillon aléatoire de la partition.

Nous allons évaluer visuellement les sous-catégories obtenues et juger à partir de l'image moyenne et d'échantillons la qualité des sous-catégories obtenues .

Sur la base d'images Christchurch le nombre optimal de composantes dans le mélange est de 3 en utilisant la méthode décrite en section 3.2.1.1. Nous avons testé différents nombres de sous-clusters et la valeur qui a donné visuellement les meilleurs résultats est de diviser encore une fois les sous-catégories en deux nouvelles sous-catégories pour un mélange de 3 sous-catégories. La figure 3.16 montre que ce partitionnement permet de retrouver l'orientation des voitures (haut, bas, droite, gauche).

Sur la base d'images du DFC2015 le nombre optimal de composantes dans le mélange est de 2 là encore une sous division en 2 nouvelles sous-catégorie donne visuellement les résultats les plus plaisants. La figure 3.17 montre les sous-catégories obtenues et encore une fois on remarque que les orientations des véhicules sont bien retrouvées dans les sous-catégories.

FIGURE 3.13 – Partitionnement des exemples de la base Christchurch en 4 sous-catégories pour la classe voiture.

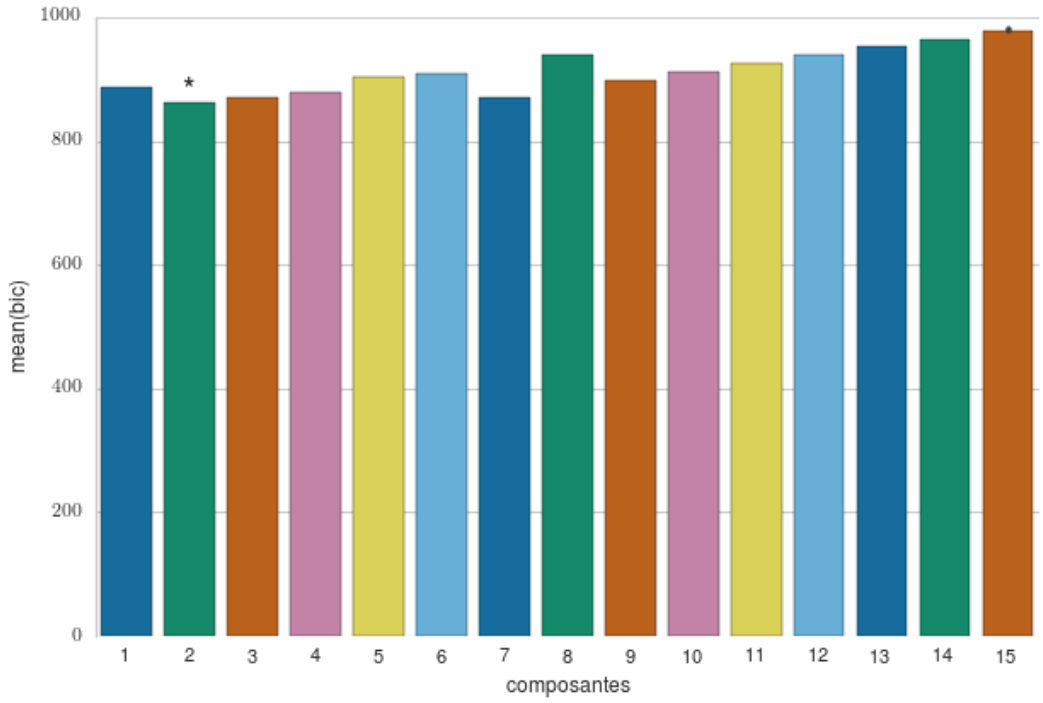


FIGURE 3.14 – BIC moyen sur la base d'image du DFC2015 pour la catégorie voitures. Les mélanges varient de 1 à 15 composantes par mélange. Le tiret noir au-dessus d'une barre représente la variance du score et l'étoile le mélange avec le nombre optimal de composantes vis-à-vis du critère.

FIGURE 3.15 – Partitionnement des exemples de la base DFC2015 en 2 sous-catégories pour la classe voiture.

FIGURE 3.16 – Partitionnement des exemples de la base Christchurch en 2×3 sous-catégories pour la classe voiture. Nous avons amélioré le partitionnement des sous-catégories trouvées en figure 3.12 pour finalement obtenir des sous-catégories plus homogènes. On observe cette fois ci que chacune des sous-catégories est composée de véhicules ayant tous la même orientation

FIGURE 3.17 – Partitionnement des exemples de la base DFC2015 en 2×2 sous-catégories pour la classe voiture.

3.4 Conclusions

Dans ce chapitre nous avons présenté le problème que pose la variation d'apparence des exemples dans une catégorie sémantique pour la reconnaissance d'objets. Les classes sémantique d'objets regroupent généralement des objets dont l'apparence peut énormément varier et, comme c'est le cas des images de télédétection, avoir de multiples orientations. Cependant les méthodes de description d'images usuelles tel que le HOG ne sont pas invariantes à la rotation des objets dans l'image mais sont très efficaces dans les cas où les variations d'apparences des objets. Nous avons donc présenté une analyse sur la variation de l'apparence des objets dans deux bases d'images aériennes (Christchurch et DFC2015). Nous avons déterminé que dans les images aériennes les principales variations d'apparence des objets provenait de la rotation des objets et qu'une méthode permettant de pallier ce problème est de modéliser des sous-catégories où les variations d'apparence sont réduites. Nous avons ainsi proposé un modèle de partitionnement permettant de retrouver dans une catégorie d'objets sémantiques différentes sous-catégories visuelles où l'apparence des objets est très similaire. Cette modélisation des sous-catégories permet d'utiliser des signatures comme le HOG pour apprendre un modèle robuste à la rotation des objets dans l'image. L'apprentissage de modèles sur des sous-catégories visuellement homogènes pour la détection d'objets étant le sujet du chapitre 4.

4 Détection d'objets

Sommaire

La détection d'objets dans des images est en vision par ordinateur la tâche consistant à la fois à reconnaître et localiser un objet dans une image. Idéalement on pourrait souhaiter attribuer une classe à chaque pixel de l'image. Pour des raisons pratiques une solution plus simple avait été adoptée dans le concours de détection d'objets Pascal VOC (Everingham, Gool, Williams, Winn, & Zisserman, 2010) : placer autour de chaque objet une boîte englobante.

Une approche courante de détection d'objets consiste à utiliser une fenêtre glissante qui parcourt l'image et à appliquer un classifieur basé sur un modèle d'apparence de l'objet en chaque position. La fenêtre glissante permet la localisation spatiale de l'objet dans l'image. Le modèle d'apparence de l'objet quant à lui permet d'émettre une hypothèse sur la présence ou non de l'objet dans la fenêtre.

Dans un premier temps (section 4.1) nous présentons le Discriminatively Trained Part Based Models (DPM) (Felzenszwalb, Girshick, McAllester, & Ramanan, 2010) qui constitue la base de notre approche. Il utilise un modèle d'apparence qui modélise les sous-catégories visuelles d'une classe d'objets d'intérêt. Nous proposons ensuite notre méthode

proprement dite : le Discriminatively Trained Mixture of Models (DtMM) (section 4.2). Il s'appuie sur une modélisation fine de l'apparence des sous-catégories. Nous présenterons des résultats expérimentaux section 4.4 sur des données aériennes optiques et multimodales.

4.1 Modèle à Parties Déformable (Discriminatively Trained Part Based Models)

Le DPM est un modèle d'apparence pour la détection d'objets proposé par Felzenszwalb et al., 2010. Ce modèle s'appuie sur trois composantes majeures pour décrire une catégorie d'objets :

1. Une signature robuste aux changements de couleurs et aux petites déformations le HOG
2. Un modèle de déformation d'une catégorie d'objets pour capturer les grandes déformations.
3. Une modélisation des sous-catégories basée sur le rapport d'aspect.

4.1.1 Descripteur Histogram of Oriented Gradients amélioré

La tâche de détection d'objets dans des images est grandement tributaire de la capacité du modèle à pouvoir discerner un objet d'intérêt de ce que l'on considère comme étant le fond de l'image. A cause des variations d'apparence des objets dans une image il est difficile d'apprendre un modèle directement à partir des pixels de l'image. C'est pour cela que l'on passe par une phase de description des exemples d'apprentissage par un vecteur signature qui va permettre de quantifier l'apparence visuelle de l'objet tout en étant plus robuste aux variations d'apparence de l'objet dans les exemples d'apprentissage. Une signature ayant fait ses preuves pour la détection d'objets dans des images est le HOG (Dalal & Triggs, 2005) . Le HOG permet la description d'une image sous la forme d'un vecteur quantifiant la forme des objets dans l'image. Le vecteur HOG est calculé en suivant les étapes suivantes :

1. Extraction d'un descripteur des gradients pour chaque pixel.
2. Agrégation des descripteurs par cellules de taille fixe
3. Normalisation et troncature des descripteurs des cellules par bloc.
4. (Optionnel) Réduction de dimension

Les 3 premières étapes correspondent à la signature HOG originale de (Dalal & Triggs, 2005), la dernière correspond à la version améliorée de (Felzenszwalb et al., 2010).

4.1.1.1 Description pixellique

Le calcul du descripteur de l'image passe par une extraction des informations de contour au niveau du pixel. Soit $\rho(x, y)$ l'intensité et $\theta(x, y)$ l'orientation du gradient pour un pixel localisé à une position (x, y) dans une image I. Le calcul de ρ et θ se fait en utilisant les filtres suivants $G_x = [-1\ 0\ 1]$ et $G_y = [-1\ 0\ 1]^T$ (Canny, 1986). L'intensité du gradient en (x, y) est donnée par :

$$\rho(x, y) = \|G_x\|^2 + \|G_y\|^2 \tag{4.1}$$

L'orientation du gradient quant à elle est discrétisée en p valeurs prenant en compte ou non la sensibilité au contraste :

- Codage de l'orientation avec la sensibilité au contraste

$$B_s(x,y) = \text{round}\left(p\frac{\theta(x,y)}{2\pi}\right) \mod p \tag{4.2}$$

- Codage de l'orientation sans la sensibilité au contraste

$$B_i(x,y) = \text{round}\left(p\frac{\theta(x,y)}{\pi}\right) \mod p \tag{4.3}$$

Pour une image couleur les valeurs de ρ et *theta* sont calculées pour chaque canal et n'est gardé que la valeur du gradient la plus forte. La valeur de l'intensité du gradient donne une mesure de la force du contour capturé. Chaque pixel est représenté par un vecteur $F(x,y)$ épars de taille p. Avec $b \in \{0,\dots,\}$ les différentes classes d'un histogramme des orientations du gradient. Le vecteur descripteur d'un pixel à une position (x,y) s'écrit :

$$F(x,y)_b = \begin{cases} \rho(x,y) & \text{si} \quad b = B(x,y) \\ 0 & \text{sinon} \end{cases} \tag{4.4}$$

4.1.1.2 Agrégation spatiale

Une fois les vecteurs $F(x,y)$ d'une image de taille $w \times h$ calculés ils sont agrégés ensemble pour apporter de la robustesse aux petites déformations et réduire la taille du vecteur descripteur de l'image. Une grille dense de cellules de taille fixe $k \times k$ est définie sur toute l'image. Une cellule C est indexée dans la grille par ses coordonnées $\{(i,j) \mid 0 \geq i \geq \lfloor (w-1)/k \rfloor, 0 \geq j \geq \lfloor (w-1)/k \rfloor\}$

Parmi les différentes approches qui existent pour agréger les descripteurs d'une cellule, (Dalal & Triggs, 2005) ont choisi d'utiliser la méthode dite du *soft-binning*. Avec cette méthode les pixels d'une cellule contribuent au descripteurs des quatre cellules voisines en utilisant une interpolation bilinéaire.

4.1.1.3 Normalisation et troncature

La normalisation des descripteurs d'une cellule permet d'ajouter une certaine invariance à l'intensité des gradients dans celle-ci. Pour leur signature (Dalal & Triggs, 2005) utilisent 4 facteurs de normalisation $N_{\delta,\gamma}$ avec $(\delta,\gamma) \in \{-1,1\}^2$.

$$N_{\delta,\gamma}(i,j) = (\|C(i,j)\|^2 + \|C(i+\delta,j)\|^2 + \|C(i,j+\gamma)\|^2 + \|C(i+\delta,j+\gamma)\|^2)^{1/2} \tag{4.5}$$

Chacun des termes de l'équation (4.5) mesure l'énergie du gradient dans un bloc de quatre cellules auquel appartient (i,j). Le vecteur normalisé obtenu est ensuite tronqués à α par une fonction T_α. Le vecteur HOG final H est obtenu par la concaténation des descripteurs normalisés et tronqué pour chacune des cellules d'un bloc.

$$H(i,j) = \begin{cases} T_\alpha\left(C(i,j)/N_{1,1}(i,j)\right) \\ T_\alpha\left(C(i,j)/N_{1,-1}(i,j)\right) \\ T_\alpha\left(C(i,j)/N_{-1,1}(i,j)\right) \\ T_\alpha\left(C(i,j)/N_{-1,-1}(i,j)\right) \end{cases} \tag{4.6}$$

4.1.1.4 Réduction de dimensions

Les paramètres usuellement utilisés pour calculer le descripteur HOG sont les suivants : $p = 9$ orientations non sensibles au contraste, cellules de taille $k = 4$ et paramètre de troncature $\alpha = 0.2$. Ces paramètres donnent un vecteur descripteur pour un bloc de taille 36 qui est celui le plus communément utilisé. (Girshick, 2012) a collecté un grand nombre de HOGs descripteurs sur un grand nombre d'image puis les a analysé en utilisant une PCA. La figure 4.1 montre les vecteurs propres les plus important décrivant l'espace des HOGs.

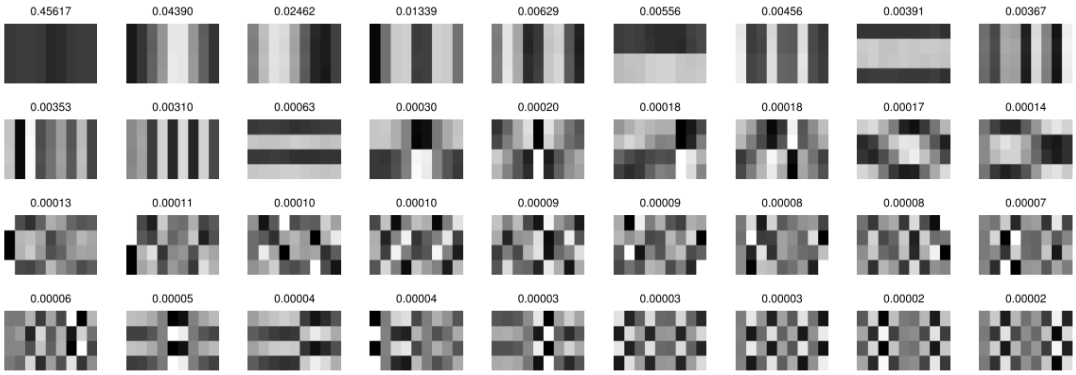


Figure 4.1 – Vecteurs propres par ordre d'importance de descripteurs HOG (Girshick, 2012). Pour la visualisation les vecteurs sont affichés sous la forme d'une matrice 4×9 où chaque ligne correspond au descripteur normalisé d'une cellule. On remarque que les vecteurs propres associés aux plus grande valeurs propres sont constants soit le long des colonnes, soit le long des lignes.

En observant cette figure on observe que l'essentiel des informations capturées par la signature peut-être capturée par le sous espace généré par les onze plus grandes valeurs propres extraites. (Girshick, 2012) a aussi montré que l'utilisation de ce descripteur de taille réduite donne les mêmes performances de détection sur la base Pascal 2007 que le descripteur HOG original. Cependant un descripteur de taille plus petite permet un gain de temps pour la détection et l'apprentissage du modèle. Mais ces avantages sont perdus à cause du coût de projection du descripteur sur la nouvelle base. (Girshick, 2012) a donc proposé une méthode de réduction de dimension moins coûteuse que la PCA basé sur une observation empirique de la distribution des HOGs de la base Pascal 2007. La figure 4.1 montre que les représentations matricielles des vecteurs propres sont approximativement constantes selon les lignes ou les colonnes. Il exploite cette structure particulière pour approximer une projection du descripteur HOG sur la base formée par les onze vecteurs propres les plus informatifs. Les plus grand vecteurs propres sont définis par un ensemble de vecteurs épars avec une seule colonne ou ligne à 1 dans leurs représentation matricielle $V = \{u_1, \ldots, u_9\} \bigcup \{v_1, \ldots, v_4\}$

$$u_k(i,j) = \begin{cases} 1 & \text{si} \quad j = k \\ 0 & \text{sinon} \end{cases} \tag{4.7}$$

$$v_k(i,j) = \begin{cases} 1 & \text{si} \quad j = k \\ 0 & \text{sinon} \end{cases} \tag{4.8}$$

(Girshick, 2012) définit un descripteur à 13 dimensions en faisant le produit scalaire entre le HOG à 36 dimensions et chacun des u_k et v_k. Il a montré qu'utiliser cette approximation donnait les même résultats en détection que le HOG original ou l'utilisation de la PCA mais pour un coût de calcul beaucoup moins important que la projection de la signature sur les valeurs propres les plus importantes. Le vecteur résultant de cette approximation peut être interprété comme un vecteur dont 9 composantes encodent les informations les orientations et 4 composantes encodent l'énergie du gradient dans les différentes zones autour de la cellule.

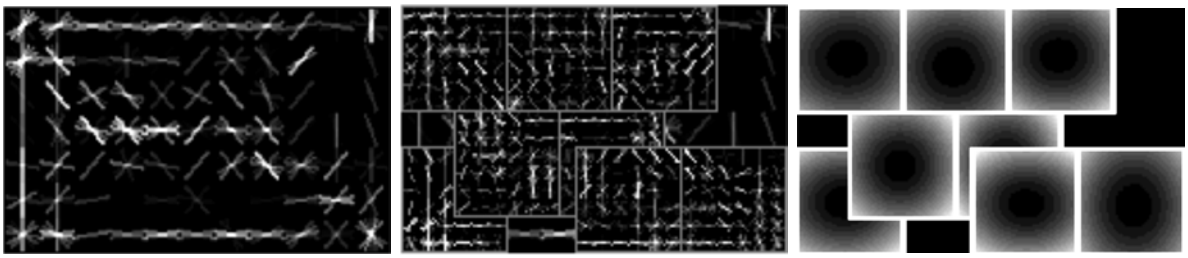
En pratique certaines catégorie d'objets sont plus sensibles aux variations de contraste que d'autres lors de la détection. Le descripteur final contient donc à la fois la description non sensibles au contraste d'une cellule C calculée sur 9 orientations et celle sensible au contraste D calculée sur 18 orientations. Le descripteur d'un bloc normalisé et tronqué est calculé comme pour équation (4.6) est calculé à la fois pour C et D. Ceci nous donne un descripteur de taille $4 \times (9 + 18) = 108$ auquel est appliqué la méthode de réduction de dimension décrite plus haut. Le descripteur final étant un vecteur de taille $4 + (9 + 18) = 31$ dont 27 composantes encodent les orientations (9 non sensibles au contexte, 18 sensibles au contexte) et 4 encodent l'énergie des gradients dans les cellules qui composent le bloc.

4.1.2 Modèle

4.1.2.1 Modèle déformable

Le DPM présenté par (Felzenszwalb et al., 2010) représente une catégorie d'objet sous une forme à la fois globale et locale. L'objet est représenté globalement car le modèle est composé d'un modèle *racine* qui décrit l'objet entièrement et est noté F_0. L'objet est aussi représenté localement car le modèle est composé de N parties $\{P_i \mid i = 1, \ldots, N\}$ qui encodent à la fois la structure sous-jacente de l'objet et décrivent localement son apparence. Chaque partie P_i est constituée de 4 éléments (F_i, x_i, y_i, d_i) avec :

- F_i le filtre décrivant visuellement l'élément i du modèle
- (x_i, y_i) la position de la partie relativement au centroïde de la racine
- $\vec{d_i} = (dx_i, dy_i)$ le coût de déformation des parties



(a) (b) (c)

FIGURE 4.2 – Illustration du DPM pour la catégorie bâtiments. La figure 4.2a est une représentation de la racine du modèle : on peut reconnaitre la forme générale d'un toit avec le faîte qui ressort longitudinalement. La figure 4.2b représente les différentes parties du modèle. La figure 4.2c illustre le coût du placement des parties lors de la détection. Pour chaque partie le coût est proportionnel à l'éloignement du centre de la partie, donc un carré uniformément noir correspond à une grande latitude de déplacement tandis qu'un distribution piquée implique un placement relatif précis.

Tel que nous l'avons décrit le DPM natif est bi-échelle. Cependant lorsqu'il est appliqué aux images il devient multi-échelle. La position des filtres du modèle est alors décrite par un triplet $p_i = (x_i, y_i, l_i)$ qui sont les coordonnées du filtre F_i. Les coordonnées du filtre sont données dans une pyramide de descripteurs $L = \{l_0, \ldots, l_K\}$

4.1.2.2 Score du modèle

Nous décrivons ici comment sont calculés les scores de détection d'un modèle $M = (N_0, P_1, \ldots, P_N)$. Le calcul du score est utile pour l'optimisation du modèle lors de l'apprentissage et pour évaluer les détections. Chacun des descripteurs HOG du modèle peut être vu comme un filtre F_i qui va être appliqué à une image. Le score d'un filtre F_i avec une fenêtre G pour une position (x, y) dans l'espace des descripteurs est donné par la formule de la corrélation croisée

$$F_i \star G[x, y] = \sum_{u,v} F_i[u, v] \times G[x + u, y + v] \tag{4.9}$$

À partir de cette formule on exprime le score d'un filtre F_i du modèle à une position p dans une image I par :

$$\mathrm{S}_{i,l}(x, y) = s_{i,l}(x, y) - c_i(x, y) \tag{4.10}$$

composé de deux éléments :

- Le score d'apparence du filtre pour une position p donnée est :

  $$s_{i,l}(x, y) = F_i \star I_{\Phi,l}[x, y] \tag{4.11}$$

  avec I_Φ, l la représentation de l'image dans l'espace des descripteurs à une échelle l
- Le coût de déplacement par rapport à (x_i, y_i) la position apprise de la i-ieme partie

$$c_i(x,y) = \langle d_i, \phi_d(x,y)\rangle + b \tag{4.12}$$
$$\phi_d(x,y) = [\psi(x,y); \psi(x^2,y^2)] \tag{4.13}$$
$$\psi(x,y) = (x,y) - (2(x_i,y_i) + v_i) \tag{4.14}$$

avec d_i est un vecteur à 4-dimensions quantifiant le coût de déplacement par rapport à la position (x,y). Usuellement $d = (0,0,1,1)$ ce qui correspond à un coût de déplacement quadratique. Évidement si $i = 0$ (cas du filtre *racine*) le coût de déplacement est nul.

À l'apprentissage comme en détection on cherche à optimiser ce score : c'est à dire maximiser la corrélation entre les filtres et les images et minimiser le coût de placement des parties. Le score de détection de chaque partie est donc donné par la fonction suivante :

$$D_{i,l}(x,y) = \max_{dx,dy} S_{i,l}(x+dx, y+dy) \tag{4.15}$$

Le score final f de détection du modèle (racine et parties) est donné par :

$$D_l(x,y) = S_{0,l}(x,y) + \sum_{i=1}^{N} D_{i,2l}(x,y) \tag{4.16}$$

4.1.2.3 Mélange de modèles

Felzenszwalb a remarqué que le modèle est mis en défaut lorsque les objets ont une trop grande variance d'apparence. Il a donc proposé une heuristique : pour chaque objet il construit des sous-catégories visuellement homogènes en utilisant simplement leurs rapport hauteur sur largeur des boîtes englobante des objets (ratio d'aspect) (Felzenszwalb et al., 2010). La méthode de partitionnement des exemples d'apprentissage est détaillé dans algorithme 2. Les exemples d'apprentissage positifs sont divisés en plusieurs partitions. Chaque partition sert à apprendre un DPM distinct. Il définit donc un mélange de modèles $M = (M_1, \ldots, M_T)$ composé de T modèles comme décrit en section 4.1.2.

Algorithme 4.1 : Algorithme de partitionnement des exemples d'une catégorie présenté par (Felzenszwalb et al., 2010)

Data : $X = \{x_1, \ldots, x_N\}$ les objets d'une catégorie dans la base d'image, K nombres de partitions désiré.
Result : $\{P_1, \ldots, P_K\}$ les exemples divisé en K partitions.
for $i = 1, \ldots, N$ **do**
  $r_i \leftarrow$ `aspect_ratio`(x_i)
end
$X = \{x_i \mid r_i > r_{i-1}, i = 1, \ldots, N\}$ `// Ordonne selon l'aspect-ratio`
for $k = 1 \ldots, K$ **do**
  $P_k \leftarrow \{x_i \mid i = k \times N, \ldots, k \times N + {}^N/_K\}$
end

4.1.3 Apprentissage

Pour entrainer le modèle décrit en section 4.1.2, une formulation de type SVM est donnée en section 4.1.3.1. Comme ce problème est semi-convexe (section 4.1.3.2), une approche pour l'optimisation alternée est décrite en section 4.1.3.3

4.1.3.1 Description du problème

Le score d'un modèle peut se réécrire en fonction d'un exemple à tester x_n de la façon suivante :

$$f_w(x_n) = \max_{z \in Z(x)} \langle w, \Psi(x_n, z) \rangle \tag{4.17}$$

où w est le vecteur de poids du modèle, z est le vecteur positions des parties du modèle et $z(x_n)$ définit l'ensemble des positions possibles des différentes parties pour un exemple d'apprentissage x_n et Ψ une fonction qui extrait d'un exemple d'apprentissage x_n une représentation à parties suivant les positions dans z.

Le problème d'optimisation qui en découle s'écrit de façon analogue à celle du SVM :

$$L_D(w) = \frac{1}{2}\|w\|^2 + C \sum_{n=1}^{N} \max(0, 1 - y_n f_w(x_n)) \tag{4.18}$$

où $D = \{(x_n, y_n) \mid n = 1, \ldots, N\}$. Si pour un exemple d'apprentissage l'espace des variables cachées peut être réduit à $|Z(x_n)| = 1$ alors f_w est une fonction linéaire en w et la minimisation de L_D se passe comme pour un SVM linéaire.

4.1.3.2 Semi-convexité

La minimisation de l'équation (4.18) est un problème non-convexe. La fonction de perte du problème d'apprentissage est définie par :

$$l_{f_w}(x_n, y_n) = \begin{cases} \max(0, 1 - f_w(x_n)) & \text{si} \quad y_n = +1 \quad \text{(4.19a)} \\ \max(0, 1 + f_w(x_n)) & \text{si} \quad y_n = -1 \quad \text{(4.19b)} \end{cases}$$

Pour les exemples positifs, équation (4.19) est le max entre une fonction constante (et donc convexe) 0 et une fonction concave $(1 - f_w(x_n))$. Pour les exemples négatif, équation (4.19) est le max entre deux fonction convexe (0 et $1 + f_w(x_n)$ car $f_w(x_n)$ est linéaire en w). Comme le max entre deux fonctions convexe est convexe, le problème est convexe en w pour les exemples négatifs. Comme le problème est convexe pour une partie des exemples d'apprentissage et non-convexe pour une autre partie, (Felzenszwalb et al., 2010) ont appelé cette propriété la semi-convexité. Cependant si pour les exemples positifs on peut restreindre l'ensemble des exemples négatifs $Z(x_n)$ à une seule valeur ($|Z(x)| = 1$) alors l'équation (4.17) est linéaire. En combinant cette propriété avec la définition de la semi-convexité on peut considérer la fonction l_{f_w} comme convexe.

4.1.3.3 Optimisation

La minimisation de équation (4.18) se fait en deux étapes, d'abord il faut restreindre pour chaque exemple positif x_n l'ensemble des variables cachées $Z(x_n)$ à une seule valeur possible. Ce nouvel ensemble où chaque exemple positif est limité à une seule configuration cachée est appelé Z_p. Z_p permet de définir un nouvel ensemble d'apprentissage $D(Z_p)$ dérivé de D et définit par $D(Z_p) = \{(x_n, y_n = 1) \mid Z_p(x_n) = \{z_n\}\}$ Ce nouvel ensemble permet de définir un nouveau problème de minimisation défini par :

$$L_D(w) = \min_{Z_p} L_{D(Z_p)}(w) \tag{4.20}$$

Cela nous permet de minimiser équation (4.18) en passant par équation (4.20). En pratique ce problème est résolut en optimisant une approche où les deux problèmes de minimisation sont coordonnée :

1. Résolution des variables cachées : Optimisation de $L_{D(Z_p)}$ sur Z_p en recherchant la configuration z_n qui maximise le score du classifieur.

$$z_n = \arg\max_{z \in Z(x_n)} \langle w, \Psi(x_n, z) \rangle \tag{4.21}$$

2. Optimisation du classifieur : Optimisation des poids w en résolvant le problème convexe équation (4.20).

4.2 Discriminatively Trained Mixture of Models

Dans cette section nous allons présenter la méthode que nous avons développée pour faire de la détection d'objets dans des images aériennes : Mélange de modèles entrainés discriminativement ou Discriminatively Trained Mixture of Models. Nous allons d'abord présenter la procédure que nous avons mise en place pour l'apprentissage d'un modèle en section 4.2.1. Ensuite nous présentons comment à partir d'un modèle nous réalisons la détection en section 4.2.2

4.2.1 Apprentissage d'un mélange de modèles

Notre approche pour la détection d'objets s'appuie sur un mélange de modèles de classification décrit dans le chapitre 3. Chacun de ces modèles est basé sur la comparaison entre une zone rectangulaire de l'image avec un modèle d'apparence de l'objet à détecter. Il est donc important d'apprendre un modèle de la catégorie d'objets d'intérêt qui soit robuste aux variations d'apparence de l'objet. Nous allons décrire notre procédure pour l'apprentissage d'un modèle robuste ayant pour finalité de donner un score de détection d'un objet aux différentes positions possibles dans l'image. Un schéma synoptique de notre approche DtMM se trouve en figure 4.3

4.2.1.1 Support Vector Machines linéaires

Nous rappelons ici brièvement le principe et la formule des machines à vecteurs de support. Surtout nous motivons l'optimisation dans l'espace primal pour apprendre des modèles d'apparence.

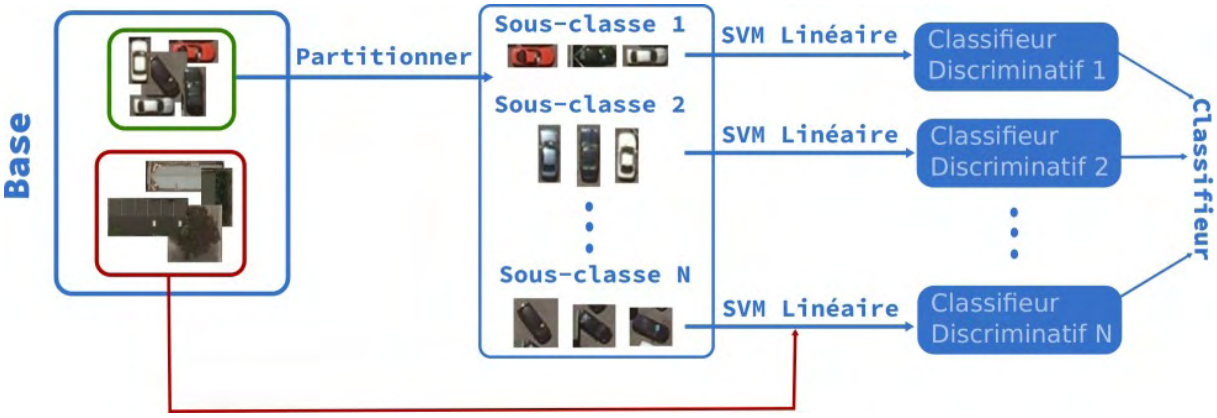


FIGURE 4.3

Il existe un large éventail de méthode permettant l'apprentissage d'un modèle à partir d'un ensemble exemples. Une méthode d'apprentissage statistique très populaire pour ce genre de problème est les SVMs. Les SVMs reposent sur l'apprentissage d'un hyperplan qui sépare les exemples en fonction de leurs catégorie. Pour un ensemble d'apprentissage labellisé $\{(x_n, y_n) \mid n = 1, \ldots, N\}$, $x_n \in \mathbb{R}^d$ et $y_n \in \{-1, 1\}$, avec L la fonction de perte du SVM (*hinge loss*) et p à valeur 1 ou 2, le problème primal d'optimisation du SVM s'écrit :

$$w^* = \min_w \frac{1}{2}\|w\|_2^2 + C\sum_n^N \xi_n^p \tag{4.22}$$

$$\text{t.q.} \quad y_n(\langle w, x_n\rangle + b) \geq 1 - \xi_n \tag{4.23}$$

$$\xi_n = L(y_n, \langle w, x_n\rangle + b) \tag{4.24}$$

$$\xi_n \geq 0 \tag{4.25}$$

La fonction de classification s'écrit

$$f(x) = \langle w, x\rangle + b \tag{4.26}$$

L'équation (4.22) est en règle générale résolue en utilisant la formulation duale du problème. La formulation duale du problème est préférée pour les raisons suivantes :

1. Elle permet d'intégrer les différentes contraintes directement dans la fonction objectif et ainsi de simplifier l'optimisation.
2. Elle s'exprime en fonction de produit scalaires entre les exemples d'apprentissage, ce qui rend naturel l'introduction de fonctions noyaux pour apprendre un hyperplan non linéaire.

La formulation duale du SVM s'écrit de la façon suivante :

$$\max_{\alpha_i} \sum_i \alpha_i - \frac{1}{2}\sum_{j,k} \alpha_j\alpha_k y_j y_k \langle x_j, x_k\rangle \tag{4.27}$$

$$\text{t.q.} \quad 0 \leq \alpha_i \leq C \tag{4.28}$$

$$\sum \alpha_i y_i = 0 \tag{4.29}$$

La fonction de classification du problème équation (4.27) s'écrit alors :

$$f(x) = \sum_{n}^{N} \alpha_n y_n \langle x_n, x \rangle + b \tag{4.30}$$

En pratique la complexité de l'équation (4.30) dépend du nombre de vecteurs de support. Les vecteurs de support sont les exemples d'apprentissage x_n dont le coefficient α_n est différent de 0. Normalement le nombre K de vecteurs de support est très inférieur au nombre d'exemple d'apprentissage N.

La résolution du problème du SVM peut se faire dans l'espace primal ou l'espace dual. Fondamentalement quel que soit l'espace choisi pour optimiser le problème les solutions obtenues sont très proches (Chapelle, 2006).

La formulation primale présente plusieurs avantages dans notre cas :

- L'apprentissage du modèle visuel d'apparence se fait avec un grand nombre d'exemples d'apprentissage. Trouver une solution exacte au problème dual de la SVM peut devenir insoluble en raison de problèmes numériques. On passe généralement par des approximations pour apprendre le modèle (Bordes, Bottou, & Gallinari, 2009 ; Shalev-Shwartz, Singer, Srebro, & Cotter, 2010 ; Bottou, 2010). Dans ce cas l'optimisation dans le primal réduit l'influence des erreurs d'approximation sur la solution optimale.
- L'optimisation dans l'espace primal peut se faire de manière incrémentale c'est à dire que le modèle visuel peut être mis à jour un exemple à la fois. Cette propriété permet de mettre en place la procédure de mise à jour du modèle avec des exemples négatifs difficiles détaillée dans la section 4.2.1.3.
- L'extraction d'un modèle visuel d'apparence est triviale à partir de la normale à l'hyperplan. Cela permet de visualiser les filtres appris sous une forme compréhensible par un opérateur humain (cf. figure 4.2a). Cette dimension est primordiale dans notre application qui nécessite l'acceptation de l'outil par les interprètes.
- Optimiser dans l'espace primal optimise directement la représentation du modèle visuel d'apparence.

4.2.1.2 Recherche efficace d'hyperparamètres

Nous avons mis en oeuvre une procédure rapide (ou du moins en temps contraint) d'optimisation du modèle par Tree-structured Parzen Estimator (TPE) que nous explicitons dans la suite.

En apprentissage statistique et en vision par ordinateur, une grand partie des performances du système mis en place dépend du bon choix des hyperparamètres de la méthode. La sélection du modèle des hyperparamètres se fait traditionnellement par une recherche exhaustive selon une grille de paramètres. Trouver les ensembles de paramètres dans la grille qui maximisent le score de classification d'un modèle dépend de la connaissance du modèle par l'utilisateur mais aussi de ses connaissances de la base étudiée. De plus les ensembles de paramètres trouvés pour une base ne peuvent généralement pas être transféré à une autre base et doivent être réestimer pour chaque bases. La recherche exhaustive bien que relativement efficace présente le désavantage d'être extrêmement couteuse en calcul. La complexité d'une recherche exhaustive étant en $O(()\,(M-1)N^2)$ avec M le nombre d'hyperparamètres de la méthode et N les valeurs possibles de ces hyperparamètres. Ces dernières années une approche alternative est apparue pour apprendre efficacement les hyperparamètres d'un système en se basant principalement sur l'optimisation bayésienne (Snoek, Larochelle, & Adams, 2012 ; James Bergstra,

Bardenet, Bengio, & Kegl, 2011 ; J Bergstra, Yamins, & Cox, 2013). L'approche par optimisation bayésienne estime la probabilité $\mathrm{p}(s \mid c, \mathcal{H})$ qu'une configuration c retourne un score s à partir d'un historique des couples (scores, configurations), les preuves de cet algorithme sont décrites en (Hutter, Hoos, & Leyton-Brown, 2011). L'Expected Improvement (EI) au-dessus d'un seuil μ est une heuristique permettant de proposer de nouvelle configurations (D. R. Jones, 2001). L'algorithme génère des configurations c qui optimisent l'EI pour un seuil μ fixé :

$$\begin{aligned} EI(c) &= \int_{y<\mu} y\, \mathrm{p}(y \mid c, \mathcal{H})\, \mathrm{d}y \\ &= E_{y<\mu}[y \mid c, \mathcal{H}] \end{aligned} \tag{4.31}$$

La méthode TPE proposée par (James Bergstra et al., 2011) permet de générer une configuration d'hyperparamètres selon un modèle probabiliste. Ils formalisent l'optimisation des hyperparamètres d'un modèle comme la minimisation d'une fonction de perte sur un espace où les configurations sont représentées par des arbres structurés. Les feuilles des arbres représentant les différents hyperparamètres à optimiser comme le facteur de coût C de la SVM où le nombre de couches dans un réseau de neurones. Pour cela un espace de recherche des configurations des hyperparamètres doit être défini par l'utilisateur. Les espaces de recherche des hyperparamètres sont définis par une densité de probabilité défini par l'utilisateur (J Bergstra et al., 2013). L'approche par TPE consiste d'abord à définir un modèle $P(c|s, \mathcal{H})$ la probabilité qu'une configuration d'hyperparamètres c atteigne un score s avec un historique $\mathcal{H}$. Sous certaines conditions cette approche revient à optimiser l'équation (4.31). La probabilité P est définie en fonction de deux densités :

$$\mathrm{p}(c \mid s, \mathcal{H}) \begin{cases} \mathrm{l}(x) & \text{si} \quad s < \mu \\ \mathrm{g}(x) & \text{si} \quad s \geq \mu \end{cases} \tag{4.32}$$

Avec $\mathrm{l}(x)$ la densité formée en utilisant les configurations $\{c_n\}$ dont la valeur de la fonction de perte $f(x_n)$ est inférieure à μ et $\mathrm{g}(x)$ la densité formée en utilisant les configurations $\{c_n\}$ dont la valeur de la fonction de perte est supérieure à μ. L'algorithme TPE choisit automatiquement la valeur de μ en fonction des quantiles des précédentes observations stockés dans $\mathcal{H}$. (James Bergstra et al., 2011) ont montré que pour maximiser l'EI il fallait que les configurations dans l'historique $c_n \in \mathcal{H}$ aient une grande probabilité $\mathrm{l}(x)$ et une petite probabilité $\mathrm{g}(x)$. Une itération de l'algorithme consiste en un grand nombre de tirages en fonction de $\mathrm{l}(x)$ et une évaluation du tirage en calculant $\mathrm{g}(x)/\mathrm{l}(x)$. Le résultat de l'itération étant la configuration c_n^* avec le plus grand EI.

4.2.1.3 Recherche d'exemples d'apprentissage difficiles

La qualité du modèle visuel d'apparence dépend à la fois de la méthode utilisée pour apprendre le modèle et des exemples utilisés lors de l'apprentissage. Généralement les exemples d'apprentissage positifs sont fournis via les annotations de la base d'image. Pour les exemples négatifs plusieurs stratégies peuvent être mises en oeuvre :

1. Si la base d'image contient les annotations d'autres catégories, ces exemples peuvent êtres utilisés comme exemples négatifs.

2. On peut extraire des exemples négatifs aléatoires à partir des images où n'apparaissent pas les objets d'intérêt.
3. S'il n'existe pas d'image sans objets d'intérêt on peut extraire des exemples négatifs à partir de positions aléatoires dans les images de la base d'entraînement à condition que les boites englobantes de ces exemples ne recouvrent pas un pourcentage r d'un exemple positif (généralement on prend $r = 70\,\%$ de recouvrement)

L'objectif de notre méthode est de pouvoir discriminer du fond de l'image les objets d'intérêt. Il semble plus intéressant de choisir comme exemples d'apprentissage négatifs des patchs représentant le fond de l'image afin d'apprendre au modèle à discriminer les objets d'intérêt du fond. Cependant dans une image aérienne où toutes les catégories d'objets sont présentes dans toutes les images la méthode 2 ne peut pas s'appliquer. Nous utilisons donc la méthode 3 pour constituer les exemples d'apprentissage négatifs. L'apprentissage du modèle se fait dans un premier temps avec les exemples positifs annotés et un ensemble d'exemples $\mathcal{H}_0 = \{x_n \mid n = 1, \ldots, N^-\}$ d'exemples d'apprentissage négatifs. La procédure de sélection des exemples négatifs est détaillée dans l'algorithme 3

Algorithme 4.2 : Recherche d'exemples négatifs difficiles

Data : h, w la taille d'une fenêtre, Exemples d'apprentissage positifs $\mathcal{X}^P = \{(x_n, y_n) \mid y_n = +1, n = 1, \ldots, N^+\}$, I Liste des images de la base d'apprentissage
Result : $\mathcal{X}^N$ l'ensemble des exemples négatifs
$\mathcal{X}^N \leftarrow \emptyset$
while $|\mathcal{X}^N| \leq N_{negs}$ **do**
  $Im \leftarrow$ `image_aleatoire(`I`)`
  $bbox =$ `boite_englobante(`x`,`y`,`h`,`w`)`
  if $bbox \in Im$ **then**
    if $\frac{bbox \bigcup x_p}{bbox \bigcap x_p} < 0.3 \ \forall x_p \in \mathcal{X}^P$ **then**
      $\mathcal{X}^N \leftarrow \mathcal{X}^N + Im(bbox)$
    end
  end
end

La qualités des exemples d'apprentissage négatifs influe directement sur la capacité du modèle à discriminer les objets d'intérêt du fond de l'image. C'est pour cela qu'une approche commune en détection d'objets est d'augmenter la qualité des exemples négatifs en utilisant le modèle appris précédemment. Lors de la phase de détection à chaque position de la fenêtre glissante est associé un score de détection. Une manière de rechercher les exemples négatifs difficiles est de classer comme tel toutes les fenêtres ne contenant pas d'objet d'intérêt mais dont le score de détection est supérieur à un seuil t. Notre méthode pour la recherche d'exemples négatifs est une répétition des étapes suivantes :

- Détection sur les ensembles de *train* et de *val*.
- Évaluation des détections de l'ensemble *val*.
- Recherche des *faux-positifs* dans la base de *train*.
- Mise à jour du modèle.

Les *faux-positifs* sont les détections ne contenant pas un objet d'intérêt mais dont le score de détection s est supérieur au plus petit score s_{tpmin} des détections qui contiennent un objet

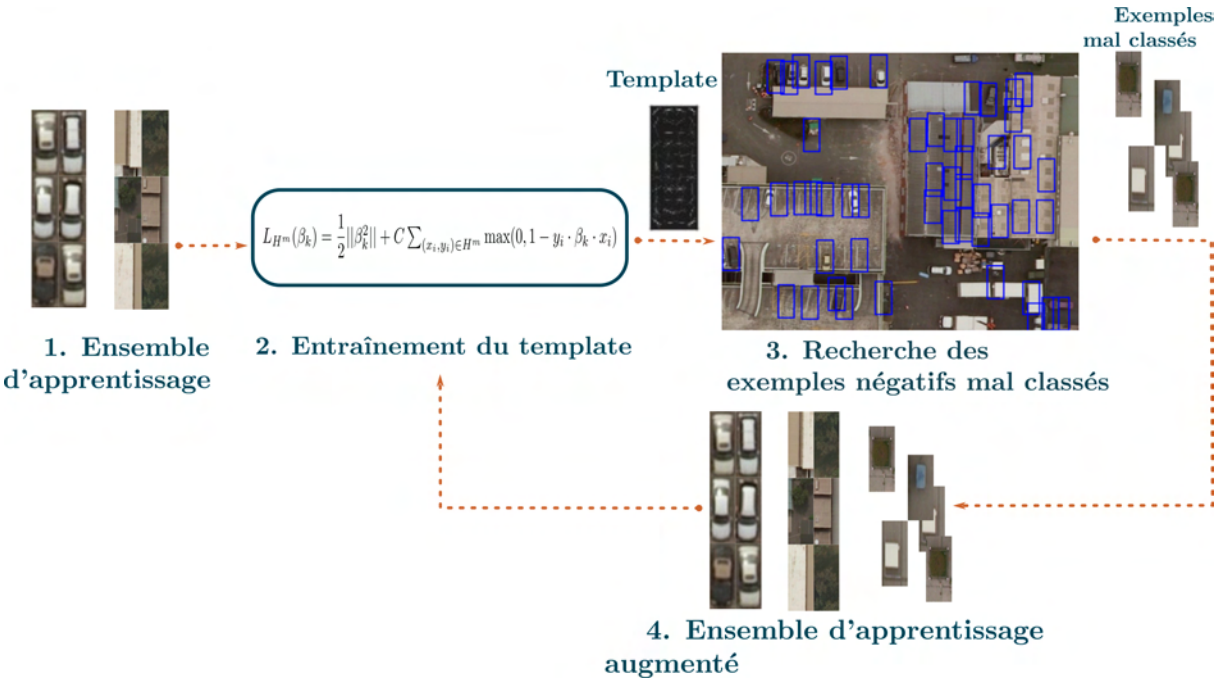


FIGURE 4.4 – Entraînement de l'approche DtMM. À partir d'un ensemble d'exemples positifs et négatifs de l'objet d'intérêt un premier modèle est appris. Ensuite une procédure itérative permet d'affiner le modèle en ajoutant les zones des images mal classées à l'ensemble d'entraînement.

d'intérêt. Les détections de l'ensemble de *train* classées comme des *faux-positifs* sont utilisées comme nouvel ensemble d'exemples négatifs $\mathcal{H}_1$ (cf. figure 4.4). Le modèle est mis à jour avec les anciens exemples positifs auxquels sont ajoutés les nouveaux exemples négatifs provenant de $\mathcal{H}_1$.

Nous modélisons le problème d'optimisation où les exemples négatifs ne forment pas un ensemble d'entraînement statique par :

$$w^* = \arg\min_w \frac{1}{2}\|w\|^2 + \frac{C}{|\mathcal{X}^+|}\sum_{x\in\mathcal{X}^+} l(x, 1; w) + \frac{C}{|Z(\mathcal{I}; w)|}\sum_{x\in Z(\mathcal{I};w)} l(x, -1; w) \qquad (4.33)$$

w le vecteur de poids que l'on cherche à optimiser, $\mathcal{X}^+$ l'ensemble des exemples positifs du problème d'apprentissage, Z la fonction qui permet de trouver les exemples négatifs difficiles dans un ensemble d'images $\mathcal{I}$ sachant w. Cette fonctionnelle en l'état n'est pas résoluble par les méthodes d'optimisation convexe usuelles. Pour la résoudre nous avons mis en place une procédure d'optimisation en deux temps où :

1. $Z(\mathcal{I}; w)$ est calculé. Cette fonction recherche dans les images les FPs (c'est à dire les zones des images où le classifieur pense qu'il y a un objet mais qui sont en fait du fond). Ensuite elle renvoie un nouvelle ensemble d'exemples négatifs.
2. Le vecteur de poids est optimisé en utilisant l'algorithme Stochastic Gradient Descent/Descente de Gradient Stochastique (SGD) de (BOTTOU, 2010).

Ces étapes sont répétées tant que le score de détection du modèle w augmente sur l'ensemble de validation (cf. Algorithme 4).

Cet algorithme est particulièrement adapté à notre méthode car un modèle déjà appris peut être mis à jour avec de nouveaux exemples. La mise à jour peut se faire sans avoir besoin de garder en mémoire les exemples d'apprentissage négatifs des passes précédentes.

Algorithme 4.3 : Mise à jour du modèle

Data : w_0 les poids du modèle, $\mathcal{X}^P$ les exemples d'apprentissage positifs, ensembles d'images *train* et *val*
Result : w_t modèle
$ap_t \leftarrow 0$
$ap_{t-1} \leftarrow 0$
while $ap_t < ap_{t-1}$ **do**
  $detections_{train}$, $detections_{val} \leftarrow$ `détecter`(w, *train*, *val*)
  $tp, fp \leftarrow$ `evaluer`($detections_{train}$, *train*)
  $ap_{t-1} \leftarrow ap_t$
  $ap_t \leftarrow$ `precision_moyenne`($detections_{val}$)
  $\mathcal{H}_t \leftarrow fp \; \forall fp > \texttt{min}(tp)$
  $w_t \leftarrow$ `mettre_à_jour`(w_t, $\mathcal{X}^P$, $\mathcal{H}_t$)
end

4.2.1.4 Validation croisée

L'évaluation de modèles statistiques appris sur une population d'éléments est un problème fréquemment rencontré en apprentissage statistique. A moins de posséder une base d'apprentissage infinie, les performances des méthodes d'apprentissage sont biaisées par les bases sur lesquelles elles ont été apprises. (Arlot & Celisse, 2010) ont réalisé une analyse des méthodes d'évaluation de modèles statistiques se basant sur la Cross-Validation/Validation Croisée (CV). La CV est une division des données d'apprentissage en plusieurs sous-ensembles. Certains sous-ensembles servant alors à entrainer le modèle, les autres à le tester. Le calcul de l'erreur de prédiction se fait alors sur l'ensemble de test. Ces étapes d'entraînement et de test sont généralement répétées plusieurs fois et les valeurs des erreurs sur les ensembles de test sont agrégées. Nous avons utilisé une méthode de CV appelée *K-fold cross validation* pour choisir l'ensemble des hyperparamètres qui maximisent le score d'AP lors de l'entraînement des détecteurs. En plus de la sélection de modèle la CV est la méthode utilisé pour évaluer la performances des modèles (diviser sa base d'image en train/val/test est une autre procédure de CV). L'avantage du *K-fold* dans notre cas est que cette méthode est très simple à mettre en place. Elle consiste à diviser la base d'apprentissage en K sous-ensemble aléatoirement puis à entraîner le modèles sur $K-1$ ensembles et tester le modèle sur le dernier. La procédure est répétée jusqu'à ce que tous les ensemble aient servit d'ensemble de test. Grâce à cette méthode, on peut artificiellement calculer l'erreur du modèle sur l'ensemble de la base d'apprentissage et ainsi réduire le biais de l'erreur calculée. Selon (Arlot & Celisse, 2010) la valeur de K la plus suffisamment efficace (celle qui demande le moins de calculs tout en calculant une erreur fiable) est $K = 5$.

4.2.2 Détection avec un mélange de modèles

Nous formalisons le problème de détecter des objets dans des images sous la forme d'un problème de correspondance avec un modèle appris. Notre méthode de détection d'objets se base sur un ensemble de modèles qui produisent un ensemble de cartes de détections des objets dans l'image. Les différentes cartes de détection sont finalement fusionnées en fonction des scores de détection pour produire la carte finale de détection

4.2.2.1 Mise en correspondance du modèle

Afin que notre méthode puisse retourner des hypothèses de détection nous devons mettre en place une méthode permettant de retourner un score de détection pour chaque localisation de la fenêtre glissante dans l'image. (Dalal & Triggs, 2005) utilisent une simple fenêtre glissante sur l'image. Le contenu de la fenêtre est ensuite transformé en utilisant le descripteur HOG puis un score de détection est calculé à partir du modèle d'apparence. (Felzenszwalb et al., 2010) utilisent une approche appelée Distance Transform (DT) proposée par (Felzenszwalb & Huttenlocher, 2012) entre le modèle et l'image entière dans l'espace des HOGs. La méthode renvoie une carte de distance entre le modèle et une grille de positions du modèle dans l'espace des HOGs. La méthode que nous avons utilisée pour la mise en correspondance du modèle avec les différentes régions de l'image est basée sur la Normalized Cross Correlation/Corrélation Croisée (NCC) de (Lewis, 1995). La version habituellement utilisée de la NCC s'écrit de la façon suivante :

$$\gamma(u,v) = \frac{\sum_{x,y}[f(x,y) - \bar{f}_{x,y}][t(x-u,y-v) - \bar{t}]}{\sqrt{\sum_{x,y}[f(x,y) - \bar{f}_{x,y}]^2 \sum_{x,y}[t(x-u,y-v) - \bar{t}]^2}} \tag{4.34}$$

Où $\bar{\cdot}$ est la valeur moyenne d'une fonction. Différents classifieurs appris sur des domaines différents et avec des tailles différentes nécessitent généralement l'apprentissage d'une calibration afin de pouvoir être comparé entre eux. Avec la NCC les scores de détection sont déjà normalisés entre 0 à 1 ce qui permet une comparaison direct des scores de détection des différents modèles. La méthode de NCC de (Lewis, 1995) permet de réaliser la NCC dans l'espace de Fourier et ainsi pouvoir utiliser la Fast Fourier-transform (FFT) pour réaliser des convolutions rapides.

4.2.2.2 Fusion des hypothèses

La méthode de détection d'objets que nous avons développée agrège les hypothèses de détection de plusieurs modèles. Chaque position du modèle dans l'image dont le score de détection est au-dessus d'un certain seuil est considéré comme une hypothèse de détection. Parce que nous utilisons une méthode de mise en correspondance basée sur la corrélation entre un modèle et une zone de l'image, plusieurs zones contiguës renvoient de haut scores. Comme le montre la figure 4.5 de multiples hypothèses de détection provenant des différents modèles se retrouvent sur les objets d'intérêt.

Lors de l'évaluation du modèle une seule de ces détections sera considérée comme un TP, les autres seront considérées comme de faux positifs. L'objectif est de proposer le minimum d'hypothèses permettant de retrouver l'ensemble des objets d'intérêt. Nous avons mis en place une méthode de fusion des hypothèses provenant de multiples modèles. Plusieurs méthodes existantes étaient utilisables. Certaines méthodes proposent de prendre l'hypothèse de détection correspondant à la détection moyenne ou médiane dans une région, d'autres proposent d'utiliser un *K-means* ou une régression logistique pour prédire la boîte englobante à partir des hypothèses renvoyées par les modèles. Dans notre cas la détection d'objets se base sur la mise en correspondance avec un filtre modèle. Les objets d'intérêt doivent donc avoir une boîte englobante dont la taille et la forme correspondent au modèle à un facteur d'échelle σ près. Redimensionner les hypothèses de détection ne fait donc pas sens dans notre cas car si le modèle est suffisamment expressif il existe une boîte englobante qui recouvre quasi parfaitement l'objet d'intérêt. L'hypothèse de détection avec le plus haut score devrait être celle qui recouvre le mieux l'objet d'intérêt. De plus en imagerie aérienne les objets d'une même catégorie ont un recouvrement quasi nul (il n'y

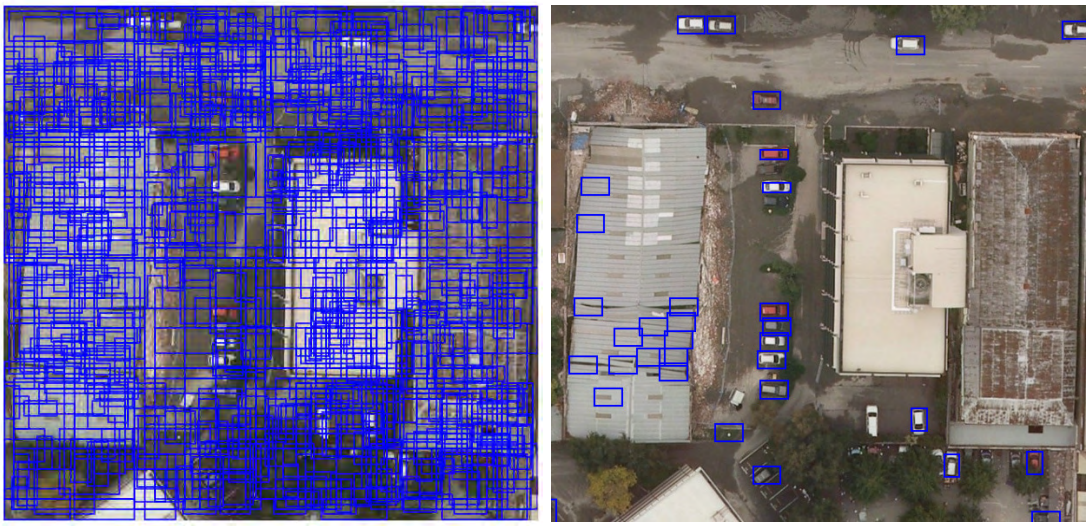


FIGURE 4.5 – À gauche hypothèses de détections provenant de plusieurs modèles avant la fusion. A droite hypothèses filtrés par la Non Maximum Suppression (NMS) : plusieurs orientations de véhicules sont conservées (y compris les faux positifs sur le toit).

a pas de voitures ou de bâtiments empilés). Nous utilisons donc un algorithme permettant de sélectionner dans un voisinage donnée la détection avec le plus haut score et de supprimer toutes celles qui la recouvrent au-delà d'un certain seuil. Pour cela nous utilisons la NMS proposée par (NEUBECK & VAN GOOL, 2006) qui est détaillée en Algorithme 5. Un exemple de résultats est proposé en figure 4.5

Algorithme 4.4 : Non Maximum Suppression

Data : $D = \{(d_n, s_n), i = 1, \ldots, N\}$ les positions et les scores des détections à fusionner
Result : $D' = \{(d_m, s_m), i = 1, \ldots, M\}$
$D \leftarrow$ `sort_by_score(`D`)`
$D' \leftarrow \emptyset$
for $n \in 1, \ldots, |D|$ **do**
  $D' = D' + (d_n, s_n)$
  for $k \in n + 1 \ldots N$ **do**
    $o \leftarrow$ `overlap(`d_n, d_m`)`
    if $o > 0.5$ **then**
      $D \leftarrow D - (d_m, s_m)$
    end
  end
end

4.3 Détection d'objets multimodale

En imagerie aérienne et satellitaire il est courant d'avoir pour une zone donnée des images provenant de différents capteurs. Une fusion efficace de ces données permet de combiner de manière avantageuse l'information fournie par les capteurs Ici nous proposons une méthode de détection d'objets utilisant les informations du domaine optique et du domaine infrarouge. Pour

le domaine optique, nous utilisons le détecteur HOG qui était la base du DtMM présenté en section 4.2. Pour le domaine infrarouge, nous utilisons un détecteur basé sur une fenêtre glissante et une description du contenu par vecteurs de Fisher (Perronnin, Sánchez, & Mensink, 2010) que nous allons décrire par la suite (section 4.3.1).

Le problème principal est de normaliser les sorties de chacun des détecteurs ce qui est plus compliqué que le modèle mono-capteur décrit auparavant en section 4.2.2.1. Pour combiner les prédictions des deux détecteurs nous étudions deux types de calibration et retenons celle de Platt (Platt, 1999) qui nous donne les meilleurs résultats. Enfin pour la fusion des boites englobantes nous utilisons à nouveau la NMS (Neubeck & Van Gool, 2006).

4.3.1 Vecteurs de Fisher pour la détection

Le descripteur d'objets permettant d'obtenir les performances de l'état de l'art en détection d'objets est le HOG. Ce descripteur cependant à besoins des informations d'un grand nombre de pixels afin d'extraire une signature discriminante. L'imagerie infrarouge ne possède qu'une résolution d'1 m/pixel. La faible résolution spatial de ces image est cependant compensée par une grande résolution spectrale. Les images infrarouges sont acquises par un capteur possédant 84 canaux (dimension $D = 84$). Le capteur couvre les longueurs d'ondes allant de 7,8 à 11,5 µm. Nous considérons que chaque spectre (équivalent du pixel pour les images) est un des descripteurs de l'objet. Le nombre de descripteurs dépend du nombre de pixels capturés par l'imageur pour l'objet. Le problème est d'extraire d'un ensemble de pixels une description discriminante de l'objet à étudier. Notre méthode de détection d'objets va s'appuyer sur le vecteur de Fischer. Le vecteur de Fischer permet d'extraire une signature d'un ensemble de descripteurs.

4.3.1.1 Noyau de Fisher

Soit un ensemble d'échantillons $\{x_n \mid n = 1, \ldots, N\}$ de N observations $x_n \in \mathcal{X}$. Soit f_θ la densité de probabilité qui modélise le processus génératif des éléments de $\mathcal{X}$ avec $\theta = \{\theta_1, \ldots, \theta\} \in \mathbb{R}^M$ En statistique la fonction de score est la fonction définie par le gradient du logarithme de la fonction de densité (la log-vraissemblance $L_\theta(X)$)

$$G_\theta^X = \nabla\theta \log f_\theta(X) \tag{4.35}$$

Le gradient de la log-vraissemblance décrit comment chacun des paramètres de f_θ doit être modifié pour s'adapter aux données X. $G_\theta^X \in R^M$ dépend entièrement du nombre de paramètres M dans θ et non de la taille des échantillons. La matrice d'information de Fisher $F_\theta \in \mathbb{R}^{M\times M}$de la densité de probabilité $f\theta$ est donnée par la formule suivante

$$F_\theta = \mathrm{E}_{\mathrm{X}\sim\mathrm{u}_\theta}\left[G_\theta^X G_\theta^{X'}\right] \tag{4.36}$$

De cette matrice (Jaakkola & Haussler, 1999) ont donné une expression du noyau de Fisher :

$$K_{FK}(X, Y) = G_\theta^{X'} F_\theta^{-1} G_\theta^Y \tag{4.37}$$

La matrice F_θ étant semi définie positive, on peut utiliser la décomposition de Cholesky de son inverse $F_\theta = L_\theta^{'} L_\theta$ pour exprimer K_{FK} soit la forme d'un produit scalaire.

$$\begin{aligned} K_{FK}(X,Y) &= (L_\theta G_\theta^X)'(L_\theta G_\theta^Y) \\ &= \mathcal{G}_\theta^{X\prime} \mathcal{G}_\theta^Y \end{aligned} \tag{4.38}$$

Le vecteur $\mathcal{G}_\theta^X$ est appelé vecteur de Fischer normalisé de X

4.3.1.2 Vecteur de Fischer

L'apprentissage d'un modèle linéaire utilisant le vecteur de Fischer comme descripteur est équivalent à l'apprentissage d'un modèle non linéaire utilisant un noyau de Fisher pour comparer deux ensembles de pixels. Il nous faut pouvoir appliquer cette transformation à nos images (directement sur les pixels de l'image). Pour cela $\mathcal{G}_\theta^X$ est écrit sous forme discrète

$$\mathcal{G}_\theta^X = \sum_n L_\theta \nabla \log u_\theta(x_n) \tag{4.39}$$

Pour chaque descripteur (ou pixel) x_n il faut calculer $\phi_{FK} = \sum_m L_\theta \nabla \log u_\theta(x_n)$ Le modèle u_θ utilisé pour décrire la distribution des pixels dans l'image est un GMM de paramètres $\theta = (w_m, \mu_m, \Sigma_m \mid m = 1, \ldots, M)$ où w_m est le poids de la m^{e} composante du mélange, μ_m la moyenne et Σ_m la matrice de covariance. Le modèle de mélange s'écrit :

$$u_\theta(x) = \sum_m^M w_m u_m(x) \tag{4.40}$$

avec

$$u_m(x) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_m)'\Sigma_m^{-1}(x-\mu_m)\right) \tag{4.41}$$

$$\sum_m^M w_m = 1 \tag{4.42}$$

$$w_m \geq 0 \tag{4.43}$$

Les paramètres de la GMM sont entraînés directement à partir des valeurs des pixels des images d'apprentissage. L'algorithme EM est utilisé pour l'estimation des paramètres du mélange. Afin de simplifier l'estimation des matrices de covariance Σ_m nous avons considéré que les différentes bandes dans l'image sont indépendantes et donc que les matrices Σ_m sont diagonales ($\sigma_m^2 = \text{diag}(\Sigma_m)$). Le gradient d'un pixel x_n s'écrit de la façon suivante :

$$\nabla_{w_m} L_{u_\theta}(x_n) = \gamma_n(m) - w_m \tag{4.44}$$

$$\nabla_{w_m} L_{\mu_\theta}(x_n) = \gamma_n(m) \left(\frac{x_n - \mu_m}{\sigma_m^2} \right) \tag{4.45}$$

$$\nabla_{w_m} L_{\sigma_k \theta}(x_n) = \gamma_n(m) \left(\frac{(x_n - \mu_m)^2}{\sigma_m^3} - \frac{1}{\sigma_m} \right) \tag{4.46}$$

avec

$$\gamma_n(m) = \frac{w_m u_{\theta_m}(x_n)}{\sum_j^M w_j u_{\theta_j}(x_n)} \tag{4.47}$$

Nous avons une formulation pour le gradient. Il faut maintenant une formulation pour L_θ définit par $L'_\theta L_\theta = F_\theta^{-1}$ (Sanchez et al., 2013) ont montré que la matrice F_θ est diagonale, on peut donc exprimer L_θ comme étant une matrice diagonale dont les éléments sont l'inverse de la racine carrée de F_θ. On obtient pour chacun des paramètres du mélange l'expression suivante :

$$\mathcal{G}_{w_m}^X = \frac{1}{\sqrt{w_m}} \sum_n^N (\gamma_n(m) - w_m) \tag{4.48}$$

$$\mathcal{G}_{\mu_m}^X = \frac{1}{\sqrt{w_m}} \sum_n^n \gamma_n(m) \left(\frac{x_n - \mu_m}{\sigma_m} \right) \tag{4.49}$$

$$\mathcal{G}_{\sigma_m}^X = \frac{1}{\sqrt{w_m}} \sum_n^N \gamma_n(m) \frac{1}{\sqrt{2}} \left(\frac{(x_n - \mu_m)^2}{\sigma_m^2} - 1 \right) \tag{4.50}$$

$\mathcal{G}_{w_m}^X$ est un scalaire alors que $\mathcal{G}_{\mu_m}^X$ et $\mathcal{G}_{\sigma_m}^X$ sont des vecteurs de dimension D. Le vecteur de Fischer pour l'ensemble de descripteurs $\{x_n \mid n = 1, \ldots, N\}$ est la concaténation des gradients pour chacune des composantes du mélange. La dimension du vecteur de Fischer est de $(2D + 1) \times M$. Pour éviter que le nombre de descripteurs utilisés influe trop fortement sur le vecteur de Fischer, le vecteur est normalisé par le nombre de descripteurs. Le vecteur de Fischer s'exprime donc de la façon suivante :

$$\mathcal{G}_\theta^X = \frac{1}{N} \left[\mathcal{G}_w^X ; \mathcal{G}_\mu^X ; \mathcal{G}_\sigma^X \right] \tag{4.51}$$

4.3.2 Calibration des classifieurs

La fusion de multiples classifieurs s'effectue en combinant les scores de classification des modèles. Les scores renvoyés par ces modèles ne sont pas directement comparables entre eux. La comparaison entre des modèles a priori incompatibles peut se faire en passant par une étape de calibration des classifieurs. La calibration d'un classifieurs consiste à produire une sortie probabiliste à partir du score de classification des échantillons.

4.3.2.1 Calibration de Platt

(Platt, 1999) propose d'estimer les probabilités de sortie d'un classifieur en utilisant une sigmoïde. L'estimation de la probabilité de sortie d'un classifieur $f(x)$ à travers une sigmoïde se fait de la façon suivante :

$$\mathrm{p}(y = 1 \mid f) = \frac{1}{1 + \exp(Af + b)} \tag{4.52}$$

Les paramètres A, B sont les paramètres d'une régression logistique. La régression logistique est apprise grâce au problème d'optimisation suivant :

$$A^*, B^* = \arg\min_{A,B} -\sum_n y_n \log(p_n) + (1 - y_n) \log(1 - p_n) \tag{4.53}$$

avec

$$p_n = \frac{1}{1 + \exp(Af(x_n) + B)} \tag{4.54}$$

Entraîner la régression logistique sur le même ensemble d'apprentissage que celui du classifieur peut biaiser le modèle. Une modèle de calibration appris à partir des données d'apprentissage va avoir pour sortie des valeurs soit très proches de 0 soit très proches de 1. Pour avoir une bonne estimation des probabilités il faut que la calibration se fasse sur un nouveau jeu de données.

4.3.2.2 Calibration isotonique

La calibration de (Platt, 1999) estime de bonnes probabilités mais seulement dans le cas particulier où les scores de sortie du modèle suivent une sigmoïde comme dans le cas des SVMs. (Zadrozny & Elkan, 2002) propose la méthode de calibration isotonique. Cette méthode d'estimation de probabilités se veut plus générique dans le sens où les scores de sortie d'un classifieur ne doivent pas suivre une fonction particulière. Cependant il faut que les scores de sortie soient croissant monotone (isotonique) pour estimer de bonnes probabilités. Le modèle de régression isotonique apprend une fonction en escalier dont les valeurs varient entre 0 à 1. Pour un ensemble de scores et leurs label $\{(s_n, y_n) \mid n = 1, \ldots, N\}$ la regression isotonic apprend les scores $\{z_n \mid n = 1, \ldots, N\}$ via le problème d'optimisation :

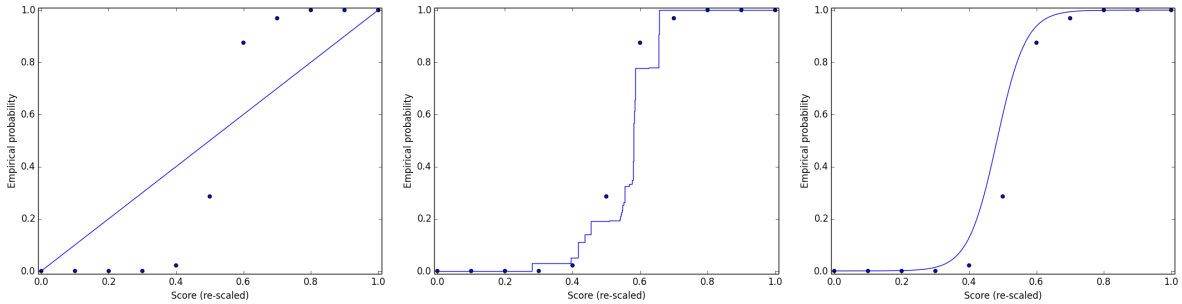
$$\min_{z_1,\ldots,z_n} \sum_{n=1}^{N} (y_n - z_n)^2 \tag{4.55}$$

$$z_n \leq z_{n+1} \quad \forall n \in \{1, \ldots, N-1\} \tag{4.56}$$

La recherche des valeurs des marches z_n est résolue en utilisant l'algorithme Pair Adjacent Violators (PAV) proposé par (Ayer, Brunk, Ewing, Reid, & Silverman, 1955). La régression isotonique à besoins d'un grand nombre d'échantillons pour apprendre les probabilités du modèle. Si le nombre d'exemple d'apprentissage est trop petit, la taille des marches de la fonction sera trop grande et la calibration donnera des probabilités incorrectes.

La figure 4.6 montre comment la calibration influe sur les scores de sortie les classifieurs. Cette figure montre la distributions des probabilités de sortie d'une SVM. On a affiché 3 méthodes de calibrations différentes. La figure 4.6a est une normalisations des scores entre 0 à 1

La figure 4.6b représente la fonction en escalier apprise par la régression isotonique. Sur cette figure on peut remarquer qu'aux alentours de 0,5 la calibration a appris peu de probabilités. Cela s'explique car les scores calibrés sont les sorties d'une SVM et la zone autour de 0,5 correspond à un score de 0 en sortie. Le principe de la SVM étant d'avoir un minimum de point x qui viole la contrainte $|\langle w, x\rangle + b| < 1$ (dans le cas où $\|w\|_2^2 = 1$) il est normal d'avoir du mal à estimer des



(a) Normalisation (b) Calibration isotonique (c) Calibration de Platt

FIGURE 4.6 – Illustration des différentes méthodes de calibration des classifieurs. Pour chacune des figures les points sont les scores bruts d'une SVM. La courbe quant à elle représente la nouvelle répartition des scores après calibration.

probabilités aux alentours de la frontière de décision. Entraîner la calibration sur un ensemble *val* distinct de l'ensemble *train* permet de réduire ce défaut. Bien que cette méthode de calibration soit généralisable à un grand nombre de méthodes de classification, elle estime mal les probabilités aux endroits où le classifieur sort peu de scores. La figure 4.6c représente la calibration apprise par la régression logistique. L'allure de la fonction de sortie est connue (sigmoïde) et seul les paramètres de cette fonction sont à estimer. Par rapport à la régression isotonique cette méthode présente l'avantage d'estimer de façon continue les probabilités de sortie du modèle à calibrer alors que la calibration isotonique estime un nombre probabilités de sortie correspondant au nombre de points utilisé pour l'apprentissage. Dans le cas où les données d'apprentissage pour la calibration sont abondantes, la calibration isotonique estime de meilleures probabilités, dans le cas contraire la méthode de Platt est plus appropriée.

4.3.2.3 Mise en place de la calibration

L'apprentissage de la calibration des modèles se fait après un premier entraînement des modèles. La base est divisée en trois sous-ensembles distinct : *train*, *val*, *test*. Les modèles sont appris une première fois sur l'ensemble de *train*. Ils sont ensuite calibrés sur l'ensemble de *val*. Finalement les modèles sont test sur l'ensemble de *test*.

Domaine optique : À partir l'ensemble de *train* des images optique de la base on apprend un DtMM. On applique le DtMM sur l'ensemble de *val* et on récupère les hypothèses de détection. On catégorise les hypothèses de détection en deux groupes les TPs et les FPs. La calibration est apprise en considérant les TPs comme les exemples positifs et les FPs comme les exemples négatifs de la régression.

Domaine infrarouge : Les vecteurs de Fischer sont appris et extrait à partir de la base de *train*. Les ensembles de pixels sont définis à partir de la carte de segmentation servant de vérité terrain. Chaque ensemble de pixels représente une instance de la catégorie d'objet étudiée. Une SVM linéaire est apprise sur les ensembles de pixels. Apprendre une SVM linéaire à partir de vecteurs de Fisher revient à apprendre une SVM non-linéaire en utilisant un noyau de Fisher sur un groupe de pixels. La calibration est apprise à partir des scores de classification de la SVM sur l'ensemble de *val*.

4.3.3 Fusion des classifieurs

Différents classifieurs ont été appris sur différents type d'images. Les images étant géoréréfencées, nous projetons les détections infrarouges dans le référentiel de l'image optique, en gardant les scores de détection originaux.

Pour pouvoir fusionner les décisions des différents classifieurs il faut que les scores soient comparables. La fusion des classifieurs se fait à partir des scores calibrés (probabilités) des classifieurs. Si plusieurs hypothèses de détection se chevauchent, l'hypothèse de détection avec la probabilité la plus haute est considérée comme une bonne détection. Les autres détections sont considérées comme des faux positifs et sont donc éliminées. La fusion des hypothèses provenant des différents classifieurs est illustrée par la figure 4.7. La figure montre comment à partir des hypothèses de détection dont les scores ont été calibrés nous effectuons la fusion des détections des deux domaines.

FIGURE 4.7 – Illustration de la fusion entre les hypothèses de détection du domaine optique et du domaine infrarouge. Les hypothèse de détection du domaine infrarouge sont projetées dans le domaine optique puis la fusion grâce à la NMS est effectuée.

4.4 Résultats expérimentaux

Nous présentons dans cette partie plusieurs expériences permettant d'analyser notre méthode de détection d'objets dans des images aériennes. Nous évaluons notre méthode sur plusieurs jeux de données afin de montrer la robustesse de la méthode. Les jeux de données contiennent des images à différentes résolutions spatiales et de plusieurs villes dans le monde (cf. figure 4.8) Dans un premier temps nous rappelons les caractéristiques des données que nous allons utiliser en section 6.4.1. Ensuite nous présentons les résultats de notre méthode DtMM pour la détection d'objets et l'évaluation des modèles sur des image optiques en section 4.4.2. Finalement nous présentons les résultats de nos travaux pour la détection d'objets multimodale en section 4.3.

4.4.1 Données et contenu des expériences

Trois jeux de données de grande taille à diverses résolutions, acquis sur 3 continents différents, ont été utilisés pour valider les méthodes de détection d'objets que nous venons de présenter. Tous ont des vérités-terrain associées.

Le DtMM (voir section 4.2) a été testé sur deux jeux de données aériennes à très haute résolution.

La base Christchurch est décrite en section 2.3.1 : elle a été acquise sur la ville néo-zélandaise de Christchurch en 2011. Sa résolution est de 15 cm/pixel. Elle est composée de 4 images. Parmi ces 4 images, 2 sont utilisées comme base d'apprentissage, 1 est utilisée comme image de validation et la dernière comme image de test.

La base mise à disposition lors du DFC2015 a été acquise sur la ville de Zeebrugge en Belgique (cf. section 2.3.3). Sa résolution est de 5 cm/pixel Elle est constituée de 6 images : 2 servent de base d'entraînement, 2 d'ensemble de validation et les 2 dernières d'ensemble de test. Pour l'évaluation de notre méthode nous nous sommes restreints à la détection de véhicules dans l'image.

Enfin pour la détection multimodale nous avons utilisé la base rendu publique pour le DFC2014. Cette base est constituée d'une unique zone dont les 2/3 des pixels sont labellisés et le dernier tiers à servi pour l'évaluation du challenge. Elle a été acquise au-dessus de la ville de Thetford Mines au Québec. La résolution de l'image optique haute résolution est de 20 cm/pixel tandis que l'image multispectrale à une résolution de 1 m/pixel.

(a) Christchurch (b) DFC2015 (c) DFC2014

FIGURE 4.8 – Aperçu des trois bases utilisées pour valider la méthode DtMM : l'apparence visuelle du contenu varie d'une image à l'autre.

4.4.2 Résultats de l'approche DtMM

Dans cette section nous allons présenter les résultats obtenus par notre méthode en détection d'objets. Nous présentons d'abord les résultats bruts de détection : courbes précision-rappel

pour l'évaluation quantitative et résultats de détection sur l'image pour l'évaluation qualitative (section 4.4.2.1). Puis nous mettons en évidence diverses composantes du fonctionnement de la méthode : évolution du nombre de faux positifs lors de la recherche d'exemples difficiles, analyse de la réponse des filtres du mélange, etc.

4.4.2.1 Résultats de détection

La figure 4.9 montre la vérité-terrain de l'image de test que nous avons utilisé. Cette image compte plusieurs centaines de voitures, bâtiments et voitures. On observe que les objets apparaissent à différentes tailles et à différentes orientations dans l'image. La figure 4.10 montre des résultats de détection du DtMM pour plusieurs catégories d'objets. Un DtMM a été appris par catégorie et les résultats de détection de toutes les catégories ont été affichés sur l'image.

La catégorie végétation est la classe que le détecteur a le plus de mal à modéliser. On retrouve des détections d'arbre un peu partout dans l'image sur quasiment toutes les zones texturées. Mais on remarque que les zones de végétation sont les endroits où le détecteur place le plus de boites ce qui signifie que le modèle ne les place pas aléatoirement. Pour la classe bâtiment, on remarque que la méthode place globalement les boîtes englobantes sur les bâtiments de l'image. Cependant la méthode manque de précision et place souvent plusieurs boites sur les bâtiments là où on n'en voudrait qu'une. Ces deux résultats peuvent s'expliquer par le fait que le HOG encode principalement des informations sur la forme des objets et que pour reconnaître un bâtiments où un arbre dans une image aérienne les informations de texture et/ou de couleurs sont très importantes mais ne sont pas suffisamment prises en compte par la signature et la méthode.

Sur la figure 4.11 on visualise les détections de véhicules sur l'image de test de Christchurch. On peut noter la qualité des boites englobantes placées sur les voitures le long des rues et sur les parkings. La précision est telle que le comptage des véhicules est même possible. Il y a très peu de faux positifs que ce soit sur les rues ou la végétation. Cependant l'algorithme se trompe parfois et confond des vasistas sur les toits ou des places de parking vides avec notre cible. En effet ces objets ont une forme rectangulaire de la même dimension que nos modèles.

La figure 4.12 montre un zoom de la figure 4.11 sur la zone de stationnement située dans le coin en bas à droite de l'image. Cet agrandissement permet d'attester de la qualité des détections de véhicules. Pour une grande majorité des véhicules de l'image, la boite englobante et quasi-parfaitement placée autour du véhicules. Quelques véhicules n'ont pas été détectés car ils sont en partis caché par un bâtiment ou alors ils ne ressortent pas assez par rapport à la route et la méthode n'arrive pas à les détecter.

La figure 4.13 montre un zoom des résultats de détection sur une des images de test du DFC2015. Pour ce travail nous avons voulu tester les capacités de généralisation du modèle ayant produit les résultats de la figure 4.11. Pour rappel le modèle est un modèle à 5 composantes appris sur la base Christchurch et les exemples d'apprentissage ont été extraits d'images a une résolution de 15 cm/pixel. Le modèle est ensuite testé sur la base DFC2015 où les images ont une résolution de 5 cm/pixel. Les résultats de détection montrent que notre méthode de détection d'objets est suffisamment robuste pour permettre de transférer un détecteur appris à dans une certaine ville du monde pour détecter des véhicules dans une ville à l'autre bout du monde imager avec un capteur différent.

Nous montrons sur la figure 4.14 les résultats de différentes expériences de détection d'objets en faisant varier le nombre de modèles dans le mélange. Il s'agit des courbes *précision-rappel* tel que défini en section 2.2.2 et les scores de AP tel que défini en section 2.2.3. L'approche de référence est un détecteur à base de HOG et de SVM. Cette méthode simple (Michel

et al., 2011) obtient une précision moyenne de 0,36. Tous les mélanges de modèles DtMMs (de 2 à 7 modèles) obtiennent de meilleurs précisions moyennes, de 0,44 à 0,77 pour le DtMM à 5 modèles. Visuellement la courbe précision-rappel de ce mélange (en violet) a incontestablement une meilleure aire sous la courbe que les autres. Cependant, même avec d'autres nombres de modèles dans le mélange, les résultats sont nettement supérieurs à l'approche de référence. Cela laisse supposer que même si le nombre de modèle optimal n'est pas trouvé, l'approche DtMM est très robuste.

Intuitivement plus le mélange comporte de modèles plus il devrait être performant. Cependant quand le nombre de modèles augmente, en moyenne la quantité de données d'apprentissage pour chaque modèle diminue ce qui a pour incidence de diminuer la capacité de généralisation du modèle. C'est pour cela que le meilleur compromis dans ce cas-là semble être un mélange de modèles à 5 composantes. Un autre cas remarquable est le cas du mélange de modèles à 4 composantes. Étonnamment ce modèle donne des performances beaucoup plus faibles que les autres mélanges. Nous expliquons ce phénomène par la raison suivante : dans ce cas l'algorithme de partitionnement produit des sous-catégories qui mélangent différentes orientations d'objets et donc ne permettent pas d'apprendre un modèle robuste. On peut estimer visuellement la qualité des sous-catégories sur les figures 3.12 et 3.13.

4.4.2.2 Analyse de la méthode

Durant la phase d'apprentissage du modèle plusieurs indicateurs peuvent être visualisés pour donner une idée sur les futures performances du modèle. La procédure d'apprentissage se base sur une procédure itérative où le modèle cherche les nouveaux exemples négatifs qui vont lui permettre d'améliorer ses performances de détection à chaque itération. La figure 4.15 montre l'évolution du nombre d'exemples négatifs difficiles sur l'ensemble de validation en fonction du nombre de passes sur la base d'images. La figure 4.15 montre l'évaluation d'un mélange de modèles à 5 composantes. Pour 4 des 5 modèles du mélange on observe que le nombre de *hard-negatifs* diminue à chaque itération sauf pour le modèle n° 2. Les modèles performants se spécialisent de plus en plus vers le bon objet à reconnaître (véhicule à la bonne orientation). Au contraire le modèle n° 2 diverge et assemble des exemples visuellement très différents les uns des autres. Ceci dit cela n'est absolument pas dommageable pour la méthode globale. En effet la fusion des sorties des modèles éliminera les réponses du modèle divergent.

La figure 4.16 illustre comment chacun des modèles du mélange permet de récupérer une partie des véhicules à détecter. Sur les cartes de chaleurs figures 4.16b à 4.16f les zones en bleu sont les endroits où les détecteurs renvoient un score faible de présence de voiture et les zones en rouges les endroits où la probabilité d'avoir une voiture est très forte. Cette figure montre que chacun des templates est spécialisé dans la détection d'un sous-ensemble des voitures de l'image. En faisant le lien avec la forme des modèles montrés au chapitre 3, on vérifie que chaque modèle trouve des détections qui ont la même orientation que les exemples de sa sous-catégorie d'apprentissage. La figure 4.16a montre le résultat de la fusion des détections provenant des 5 templates sur une partie agrandie de l'image complète.On peut noter que les parkings comportent bien des véhicules garés selon différents angles.

4.4.3 Détection d'objets multimodales

Dans cette section nous présentons les résultats obtenus pour le DFC2014. Il s'agit ici de combiner les informations optiques avec les informations infrarouges pour analyser le contenu d'une image. Les figures 4.17 et 4.19 montres les courbes *précision-rappel* pour les détecteurs de voitures et

de végétation. Les figures 4.18 et 4.20 montrent des exemples de détections renvoyé par notre méthode. On remarque que pour la détection de voitures l'infrarouge apporte très peu pour la détection. Cela est dû au fait que l'imageur infrarouge à une résolution de 1 m/pixel et donc les voitures sont très peu résolues et la quantité d'information disponible ne permet pas de caractériser les voitures. De plus à moins que les moteurs des voitures ne soient chauds les voitures se confondent avec l'asphalte en imagerie infrarouge. Par contre, la fusion de l'infrarouge et de l'optique donne des résultats probants pour la détection d'arbres dans l'image.

FIGURE 4.9 – Vérité terrain de l'image de test de la base Christchurch. Les voitures sont entourées d'une boite englobante bleue, les bâtiments d'une boite englobante jaune et les arbres d'une boite englobante verte.

FIGURE 4.10 – Détections du DtMM sur l'image de test de la base Christchurch. Les détections de voiture sont en bleu, celles de bâtiments sont jaune et celles de végétation en vert. On observe que d'un point de vue global le détecteur arrive à reconnaitre les objets des différentes catégorie même si il est peu précis dans le cas des bâtiments et des arbres.

FIGURE 4.11 – Hypothèses de détections renvoyées par le DtMM pour la détection de véhicules sur l'image de test de la base Christchurch.

FIGURE 4.12 – Zoom sur l'image de test de la base Christchurch.

FIGURE 4.13 – Zoom sur une des images de test de la base Zeebrugge.

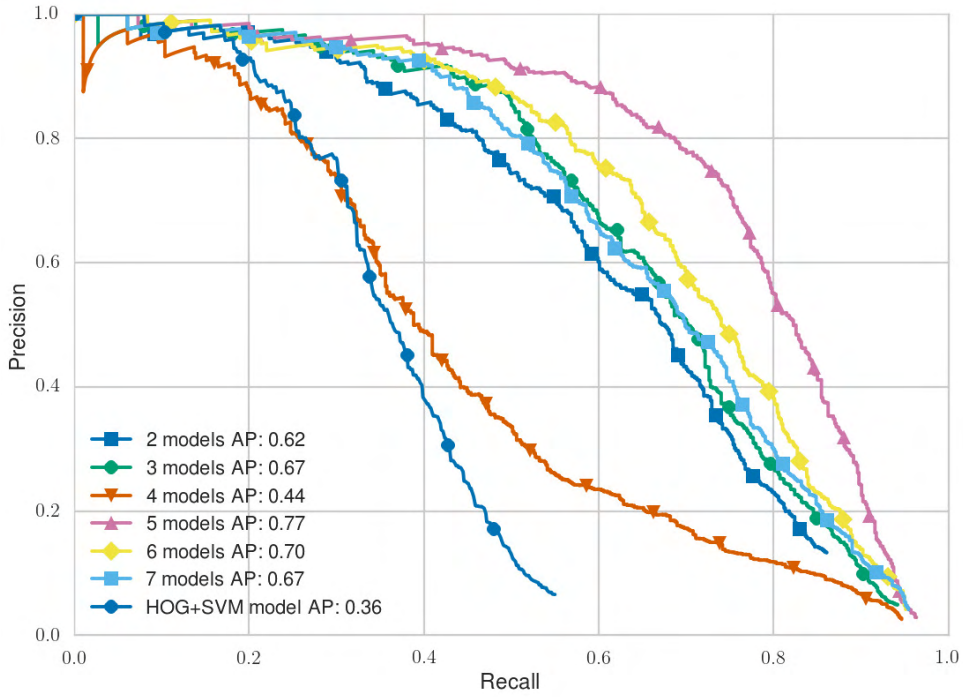


FIGURE 4.14 – Courbes précision-rappel pour la détection de véhicules sur la base DFC2015 pour différents mélanges de modèles. La courbe "HOG+SVM" correspond à la méthode de (DALAL & TRIGGS, 2005)

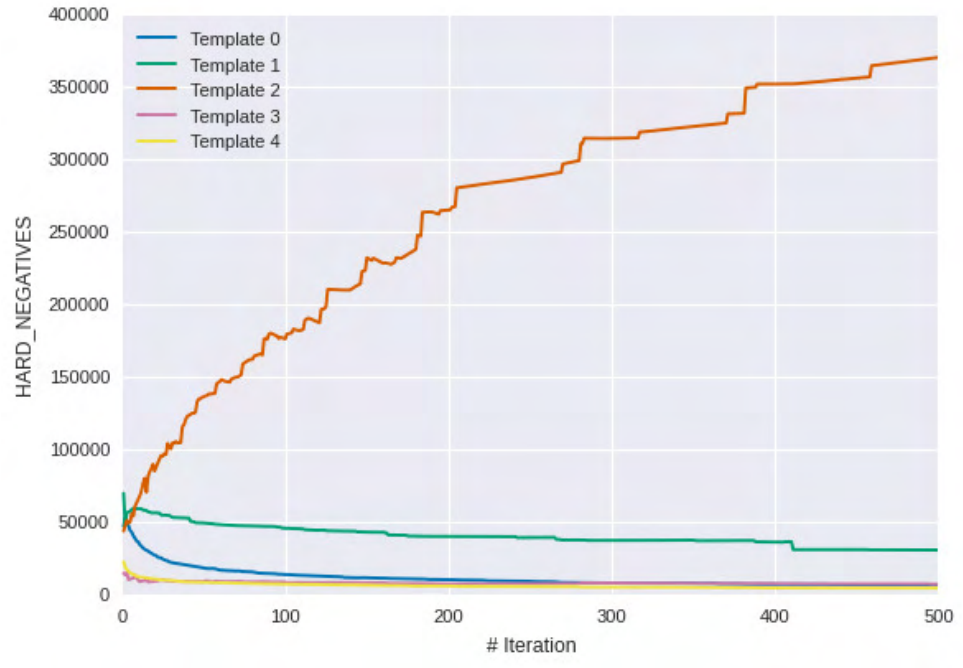
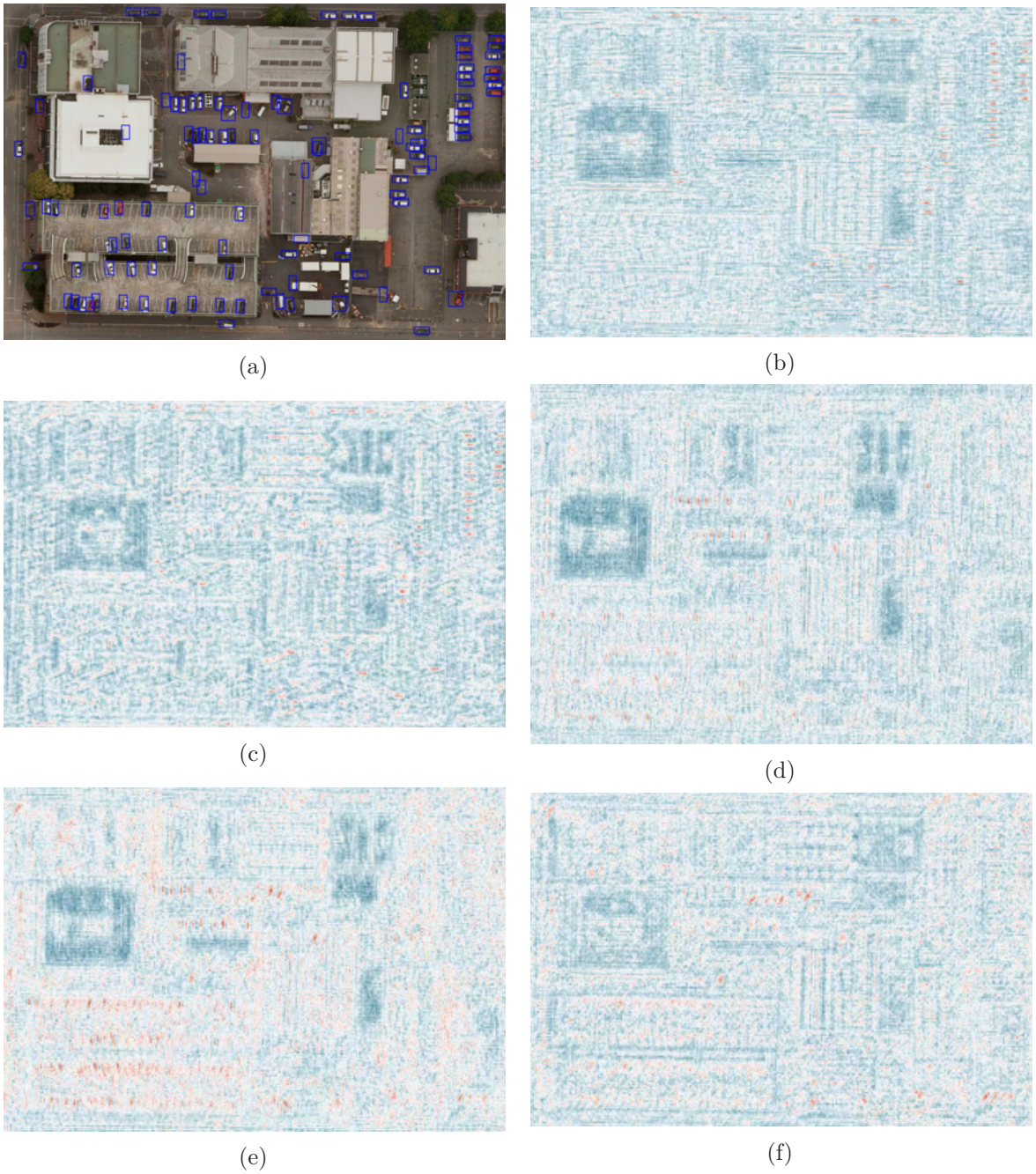


FIGURE 4.15 – Évolution nombre de *hard-negatifs* par template en fonction du nombre d'itérations lors de la phase d'apprentissage



(a) (b) (c) (d) (e) (f)

FIGURE 4.16 – Sur cette figure, nous montrons un résultat de détection accompagné des cartes de chaleur permettant de produire ce résultat. Chaque carte correspond à la réponse d'un modèle sur la zone étudiée. On observe sur ces images un score de corrélation entre un modèle et une zone de la taille du modèle dans l'image. Plus un pixel est bleu et plus la zone et le modèle sont décorrélé, à l'inverse plus un pixel est rouge et plus la zone de l'image et le modèle sont corrélé. Une forte corrélation entre le modèle et une zone de l'image signifie qu'un objet d'intérêt se trouve dans cette zone avec une forte certitude.

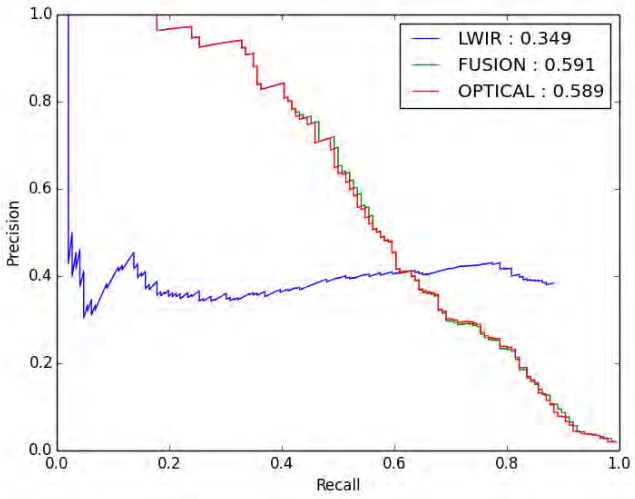


FIGURE 4.17 – Courbes précision-rappel pour la détection véhicules. La courbe rouge représente les performances de détection dans le domaine visible . La courbe bleue les performances dans le domaine infrarouge. La courbe verte montre la fusion des deux détecteurs basée sur la NMS

FIGURE 4.18 – Détections de véhicules utilisant notre méthode de fusion de modèles. Les détections en bleu sont les TPs , les détections en vert sont les FPs.

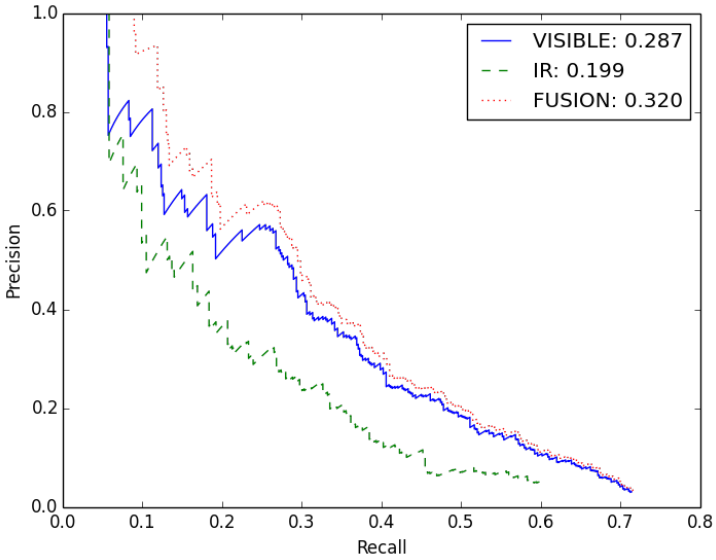


FIGURE 4.19 – Courbes précision-rappel pour la détection d'arbres dans l'image. La courbe bleue représente les performances de détection dans le domaine visible . La courbe verte les performances dans le domaine infrarouge. La courbe rouge montre la fusion des deux détecteurs basée sur la NMS

FIGURE 4.20 – Détections d'arbres utilisant notre méthode de fusion de modèles. Les détections en bleu sont les TPs , les détections en vert sont les FPs.

4.5 Conclusions

Dans ce chapitre nous avons présenté différentes méthodes de détection d'objets dans des images aériennes. Nous avons montré que la modélisation des sous-catégories permet de produire des mélanges de détecteurs spécialisé sur des sous-ensemble de la base d'image mais dont la fusion a permet une meilleure généralisation des performances de détection. Nous avons présenté nos expériences sur deux bases d'images optiques aériennes très haute résolution et une base d'image combinant de l'optique et de l'infrarouge. Une manière de réduire le coût computationnel de l'utilisation du mélange de modèle serait de rendre la signature des images invariantes à la rotation. De plus notre méthode permet la détection d'objets ayant des formes bien définies comme les voitures mais montre des performances limitées pour les objets dont les formes varient fortement comme les bâtiments.

Deuxième partie

Contexte

5

Contexte dans les images

Sommaire

Dans ce chapitre après une brève introduction en section 5.1, nous commencerons par décrire les différents niveaux de contextes qui existent dans une image ainsi que les méthodes permettant de le modéliser et l'utiliser pour la classification d'images. La section 5.2 décrit méthodes pour modéliser le contexte global d'une image. La section 5.3 décrit les différentes méthodes pour modéliser le contexte local dans une image.

5.1 Introduction

Ce chapitre a pour sujet les travaux existant en modélisation du contexte en vision par ordinateur Le contexte est une composante essentielle de la vision humaine, il est donc naturel d'essayer de transposer ce principe en vision artificiel. En vision naturel le contexte apporte un gain d'information pour aider à la reconnaissance d'un objet ou d'une scène. L'objectif de la modélisation du contexte en vision par ordinateur est d'imiter la vision naturel et de proposer des modèles de contexte permettant d'améliorer les capacités des systèmes de reconnaissance d'objets dans des images. Les méthodes de reconnaissance d'objets basées uniquement sur les propriétés visuelles d'un objets, comme la couleur, la forme, la répartition des pixels etc., sont sujettes à des variations de performances.

Ces variations sont principalement causées par :

1. L'occultation de l'objet d'intérêt par un autre objet ou par du fond
2. Le bruit d'acquisition de l'image
3. La variation de poses que peut avoir un objet
4. La variation d'illumination entre les images

Une façon de rendre les système de détection d'objets plus robuste aux variations précédentes est de combiner à la description visuelle une informations sémantique de plus haut niveau que l'on nommera le contexte. Le contexte peut être divisé en deux niveau d'abstraction : *global* et *local*. Le contexte global décrit les interactions entre les composantes dans l'image entière. Le contexte

local décrit les interactions entre les composantes de l'image localement avec pour référence un objet ou une région d'intérêt.

Deux articles passant en revue les apports du contexte pour la reconnaissance d'objets sont (Galleguillos & Belongie, 2010) et (Divvala, Hoiem, Hays, Efros, & Hebert, 2009)

(Galleguillos & Belongie, 2010) décrivent les différents éléments de contexte qui existent dans les images et proposent une classification de ces éléments en fonction de leurs usage pour la reconnaissance.

- Le contexte sémantique qui est la probabilité de cooccurence des différentes catégories d'objets à la fois entre eux et dans une scène. Les auteurs appuient leurs déclarations en citant l'étude de Palmer qui porte sur l'influence des images antérieures dans la reconnaissance d'objets. Les conclusions des auteurs sont que l'identification d'un objet par un humain est plus précise si les images que le sujet a vu antérieurement sont en rapport avec l'objet à reconnaître, par exemple reconnaître une miche de pain après avoir vu une cuisine. Alors qu'au contraire si le sujet a vu antérieurement des images qui n'ont rien à voir avec l'objet à reconnaître, comme une fourchette et un caisson de basse, le sujet sera moins précis pour la reconnaissance.
- Le contexte spatial en rapport avec la position des objets. Ce contexte très utile pour les scènes de la vie de tous les jours est définit par les positions possibles/improbables pour un objet dans une scène. Bar a examiné les conséquences des relations spatiales sur les humains pour la reconnaissance d'objets et en a déduit :
  1. Si la position de l'objet dans l'image a une interprétation unique alors la reconnaissance va grandement en bénéficier
  2. Une modélisation correcte des relations spatiales entre les objets permettent de réduire les erreurs pour la reconnaissance d'une instance d'un objet.

Ces observations montrent que les contexte sémantique et spatiale sont très liés.

- Le contexte d'échelle en rapport avec la taille des objets. Un apriori sur la taille des objets permet d'éliminer les détections en fonction de leurs taille et d'éviter d'utiliser des approches multi-échelles.

Ensuite les auteurs classifient les modèles de contexte en fonction de leurs interactions avec l'image. Si l'interaction entre le modèle et l'image se fait sur la globalité de l'image on va parler de contexte globale. Si l'interaction entre le modèle et l'image se fait au niveau de l'objet on va parler de contexte local. Le contexte global consiste à exploiter la configuration de la scène dans l'image entière comme source d'informations contextuelles. Différentes méthodes existent pour quantifier la structure globale d'une scène comme la moyenne des descripteurs de la scène ou les statistiques sur les valeurs des pixels.Un descripteur de contexte global très utiliser est le *gist* proposé par (Oliva & Torralba, 2001). Par contre le contexte globale a des difficultés à reconnaître la scène dans des images où les objets sont trop nombreux. Avec le nombre d'objets qui augmente plusieurs scènes différentes peuvent partager les mêmes configurations et sembler similaires.

Le contexte local quant à lui exploite les informations disponibles dans le voisinage proche de l'objet à détecter. Ces informations peuvent êtres les pixels, les régions de l'image ou encore les objets environnants. L'un des avantages du contexte global sur le contexte local est que la représentation globale de la scène génère moins de données que les multiples représentations locales et donc constituent un gain de mémoire et de coût de calcul.

Le principe du contexte local est d'extraire les informations autour d'une région d'intérêt dans une sous partie de l'image. Ce type de contexte permet de capturer finement les interactions entre les différentes catégorie d'objets dans les images. Par rapport au contexte global où la scène est représentée entièrement, le contexte locale décompose une scène en plusieurs éléments et un contexte particulier peut-être extrait pour différentes régions d'intérêt. Cependant une première phase d'extraction des zones d'intérêt est nécessaire pour identifier les régions d'intérêt dans l'image. Par exemple en détection d'objet il s'agira d'extraire les boîtes englobantes ou en segmentation sémantique d'extraire les superpixels.

(Divvala et al., 2009) proposent une étude empirique de l'apport d'informations contextuelles pour la détection d'objets dans des images. Les auteurs évaluent l'apport des représentations contextuelles sur la base Pascal VOC 2008 proposée par (Everingham, Gool, Williams, Winn, & Zisserman, 2010). Dans un premier temps les auteurs définissent les sources de contexte que l'on peut trouver dans une image (local, global, 3D, sémantique, photogrammétrique, d'illumination, météorologique, géographique, temporel et culturel)

Ces sources de contexte sont utilisées pour améliorer les aspects suivants de la détection d'objets :

- La présence d'un objet
- L'apparence d'un objet
- La localisation d'un objet
- La taille de l'objet
- Le support spatial de l'objet

Les auteurs utilisent le détecteur d'objets proposé par (Felzenszwalb, Girshick, McAllester, & Ramanan, 2010) pour établir le seuil de performance d'un détecteur sans informations de contexte. Les erreurs les plus fréquentes rencontrées par ce détecteur étant de mal localiser un objet dans l'image ou de ne pas arriver à détecter les objets de petite taille comme l'ont montré (Hoiem, Chodpathumwan, & Dai, 2012). Pour leurs évaluations les auteurs combinent différentes sources de contexte pour analyser quelles sont les sources contextuelles qui ajoutent de l'information et dans quels cas. Le modèle de contexte des auteurs permet à la fois de pondérer le score de détection d'une hypothèse mais aussi de redimensionner les boîtes englobante de l'hypothèse de détection. Selon les expériences conduites par les auteurs, le contexte sémantique, le contexte global et le contexte global sont responsable en moyenne de 70 % des améliorations de performances par rapport au détecteur d'objets seul. La méthode utilisée pour modifier la taille des boîtes englobantes permet d'augmenter l'AP de 2.1pour les objets naturels (vaches, moutons, chevaux etc.) et de 0.8point pour les objets faits par l'homme (vélos, bateaux, bouteilles etc.)

5.2 Contexte global

Des études en psychologies tendent à montrer que le processus de reconnaissance est basé sur une représentation hiérarchique du contexte. Le contexte s'organise de manière hiérarchique à partir d'une structure globale de l'image jusqu'au détails les plus fins. Le contexte global exploite la structure entière de l'image comme source d'informations additionnelle pour la représentation de la scène. La structure du contexte dans l'image doit permettre d'extraire un résumé de la configuration des objets. De récentes recherches sur la modélisation de la structure du contexte dans une image ont été utilisées pour comme information préalable pour les tâches de localisation et reconnaissance d'objets dans des images (Murphy, Torralba, & Freeman, 2003 ;

Oliva & Torralba, 2001 ; a. Torralba, Murphy, Freeman, & Rubin, 2003 ; Antonio Torralba, 2003 ; Porway, Wang, & Zhu, 2010 ; A. Torralba, Oliva, & Freeman, 2010).

(Oliva & Torralba, 2001) proposent un modèle de description d'une scène qui encode la configuration global de la scène sous la forme d'un vecteur dont le nombre de dimensions est réduite. Ce descripteur permet de caractériser une scène sans avoir à passer par les classiques étapes de segmentation ou de reconnaissance des régions de l'image. Pour se passer de ces étapes, les auteurs définissent d'abord L'enveloppe spatiale est représentée par les relations entre les bordures des surfaces de la scène ainsi que les textures et les motifs représentés par les éléments de la scène (murs, fenêtre, voitures etc.). Les auteurs proposent de capturer l'enveloppe spatiale de la scène à grâce aux notions suivantes : naturalité, ouverture, rudesse, expansion et robustesse de la scène. À partir de ces notions les auteurs génèrent un vecteur à 960 dimensions qui encode la structure de la scène.

(Choi, Torralba, & Willsky, 2012) proposent de modéliser le contexte global d'une image en utilisant une structure de type arbre. Les auteurs énoncent l'hypothèse suivante : les dépendances entre objets peuvent être modélisées de façon hiérarchique. Une manière de représenter cette hiérarchie de manière parcimonieuse est d'utiliser une structure de type arbre. Les noeuds de l'arbre représentent les différentes catégories d'objets. Les relations de parenté entre les noeuds représentent les relations de contexte. Le modèle proposé par (Choi et al., 2012) décrit les relations de cooccurence et spatiales entre les différentes catégories d'objets dans une image. Il modélise aussi bien les corrélations positives que les corrélations négatives. Le modèle de contexte est d'un modèle donnant à priori la probabilité qu'un objet d'une catégorie définie apparaisse dans une image sachant les objets qui l'entourent. La probabilité qu'un objet apparaissent dans une image tout en sachant les autres objets qui l'entourent est calculée en utilisant deux modèles de probabilités jointes. Le premier modèle de probabilités jointes est donné par la modélisation des cooccurences des objets dans l'image. Le second modèle de probabilités jointes est quant à lui donné par la modélisation de la hauteur relative de l'objet dans l'image. Chacun des noeuds modélisent un modèle de probabilités jointes en agrégeant les deux modèles précédent. Le score de contexte d'un objet étant donné par la multiplication des probabilités jointes entre la racine de l'arbre et le noeud de la catégorie de l'objet d'intérêt. Pour compléter le modèle de contexte un modèle d'apparence est ajouté utilisant le *gist* de (Oliva & Torralba, 2001). Le *gist* de la scène permet d'ajouter implicitement une information sur la catégorie de la scène et de pouvoir différencier par exemple les scènes intérieurs des scènes extérieurs.

(Porway et al., 2010) présentent une méthode de modélisation du contexte pour l'interprétation d'images aériennes. Le modèle proposé par les auteurs organise les objets en groupes dans une hiérarchie en fonction de l'apparence des objets mais aussi de leurs configuration dans l'image. Ce modèle détecte les objets dans l'image et les organise simultanément dans une représentation hiérarchique du contexte de la scène. Cette représentation regroupe automatiquement ensemble les objets par groupe et encode comment les objets sont reliés entre eux.

Les principales contributions des auteurs sont :

- Modélisation d'une scène en utilisant un modèle hiérarchique sur 3 couches implémenté par un graphe
- Algorithme d'apprentissage pour ajouter/apprendre les contraintes entre les noeuds du graphe.

La représentation hiérarchique de la scène se fait en utilisant un graphe dont les noeuds correspondent aux éléments extraits de la scène. Les auteurs se limitent à une représentation sur 3 niveaux :

1. Le premier niveau est représenté par la racine du graphe. Ce noeud représente tous les objets et les groupes d'objets de la scène.
2. Les noeuds du niveau suivant représentent des groupes d'objets. Ces groupes sont composés d'objets d'une même catégorie comme les voitures d'un parking ou les bâtiments d'un quartier.
3. Les noeuds du dernier niveau représentent les objets d'un groupe

Dans cette représentation hiérarchique, une scène peut être composée par un ou plusieurs groupe de chaque catégorie. Les groupes étant eux même composés d'un ou plusieurs objets de chaque catégorie. Cette représentation seule ne fait que comptabiliser les objets et les groupes d'objets dans l'image. Les relations de contexte entre les objets sont ensuite ajoutées comme des liens entre les noeuds du graphe. Le modèle de contexte enrichit le modèle hiérarchique en ajoutant des contraintes entre les objets. Ce modèle contrôle que l'apparence des objets ainsi que leurs configurations spatiales suivent certaines contraintes. Les contraintes sur les objets, aussi appelées relations statistiques, sont apprises à partir d'informations extraites d'un ensemble d'images d'apprentissage. Ces relations statistiques peuvent être la distribution de la taille relative entre deux voitures ou de leurs distances.

Les modèles de contexte global d'une image en vision par ordinateur servent généralement à améliorer l'interprétation de la scène entière. L'utilisation de ces méthodes requièrent que les objets de la scène soient prédit avec une grande confiance car leurs position est critique pour extraire la structure de la scène.

5.3 Contexte local

Contrairement au contexte global qui englobe tous les éléments d'une image, le contexte local est définit localement une image. Cela implique qu'une même image peut comporter une multitude de contexte locaux en fonction des régions d'intérêt considérées. Une image comporte des régions d'intérêt sur plusieurs niveaux. Du niveau sémantique le plus bas qui est le pixel en passant par des niveaux intermédiaires (groupes de pixels, segments d'image etc.) jusqu'au niveau le plus élevé qui est un objet. Chacun de ces niveaux fournit des informations permettant d'extraire le contexte d'une image. Dans cette thèse, nous nous intéressons aux relations pouvant exister localement entre les différentes régions d'intérêt dans une image. Une relation de contexte local entre plusieurs régions d'intérêt dans l'image est la modélisation de l'interaction entre deux régions d'intérêt ou plus. Dans la littérature plusieurs méthodes existent pour modéliser les relations locales entre régions d'intérêt dans une image (Galleguillos & Belongie, 2010 ; Divvala et al., 2009). Ces relations peuvent être modélisées sous la forme d'un ensemble non ordonné de vecteurs (Cinbis & Sclaroff, 2012), d'un arbre (Malisiewicz & Efros, 2009) ou encore d'un modèle graphique (Volpi & Ferrari, 2015).

(Cinbis & Sclaroff, 2012) propose un modèle de contexte pour la détection d'objets dans des images. Les auteurs identifient deux sources principales de contexte, le contexte de la scène et le contexte des objets Le contexte de la scène est capturé par le *gist* de (Oliva & Torralba, 2001). Le contexte des objets est quant à lui modélisé à partir des proposition renvoyés par un détecteur d'objets. Les auteurs utilisent les relations entre paires d'objets comme source de contexte. Les relations entre les paires d'objets sont encodées sous la forme d'un ensemble de caractéristiques qui décrivent la relation dans la paire. Les auteurs proposent les caractéristiques suivantes pour capturer le contexte entre une paire d'objet :

1. Le score de détection de la classe. Un vecteur de taille le nombre de classes présent dans la base et dont toute les composantes sont à zéro excepté les composantes correspondant aux classes de la paire d'objets. Ces composantes sont quant à elles initialisées avec le score de détection des objets.
2. Les relations spatiales entre les objets (y relatif, y relatif en valeur absolue, hauteur et largeur relative, distance).
3. Le ratio de chevauchement entre les objets.
4. Scores relatifs

Pour l'apprentissage du modèle de contexte à partir des pairs d'objets les auteurs proposent une méthode d'apprentissage basée sur la méthode du boosting appelée SetBoost.

(Malisiewicz & Efros, 2009) présentent leurs travaux sur l'apprentissage d'un Memory Index (*memex*) visuelle pour modéliser les associations entre des instances d'objets dans le but d'améliorer les systèmes de reconnaissance d'objets. Les auteurs rappellent la définition du contexte tel que donnée par Aristote et qui est communément admise dans les sciences formelles. Selon Aristote les catégories sont définies comme des ensembles discrets d'entités qui partagent un certain nombre de propriétés. Cette définition rend différentes catégories mutuellement exclusives et induit que tous les membres d'une catégorie sont égaux entre eux. Bien que communément admise cette définition présente un certain nombre de défauts. (Wittgenstein, 1953) a souligné que ce n'était certainement pas comme cela que notre cerveau marche au quotidien. (Rosch, 2013) a montré que nous ne divisions pas le monde qui nous entoure en catégories bien définies qui partagent des propriétés communes mais plutôt que nous nous basions sur un ensemble de similarités pour percevoir notre environnement. (Bar, 2009) émet l'hypothèse que dans le cerveau la reconnaissance d'un objet ne se fait pas en reconnaissant explicitement la catégorie de l'objet mais plutôt qu'elle passe par un ensemble d'associations avec des objets déjà vu similaire. Bar met ainsi en évidence l'importance les analogies et des associations dans le processus de reconnaissance d'un objet. Partant de ces constats, Malisiewicz et Efros propose une méthode de modélisation des associations entre les instances des objets dans une base d'images appelée *memex* visuelle en hommage à (Bush, 1945). Le *memex* visuel est modélisé sous la forme d'un graphe dont les nœuds sont les instances des objets dans la base et les arêtes représentent les relations entre les instances des objets. Les arêtes du graphe représentent deux types de relations, les similarités visuelles et les associations contextuelles entre les noeuds. (Malisiewicz & Efros, 2009) contrairement à (Bush, 1945) propose d'apprendre automatiquement les relations contextuelles et de similarité dans le *memex*.

(Volpi & Ferrari, 2015) proposent une méthode de segmentation sémantique d'images satellitaire utilisant des informations contextuelles. . Leur méthode se base sur la modélisation des interactions locales entre les classes en utilisant des Conditional Random Fieldss (CRFs). Les auteurs proposent de modéliser les interactions entre les classes se basant sur le contexte géométrique/spatial entre les pixels d'une image. Ils considèrent autour d'un groupe de pixels d'intérêts une ou plusieurs zones concentriques de rayons prédéfinies. Les auteurs modélisent les relations entre les classes en utilisant deux caractéristiques se basant sur un modèle de Potts.

Dans le cadre des modèles graphique probabiliste le modèle de Potts favorise les noeuds voisins d'un graphe à partager le même label. Le descripteur de contexte de (Volpi & Ferrari, 2015) est décomposé en deux termes :

1. Un terme mesurant le contraste entre deux noeuds.
2. Un terme estimant un score de compatibilité entre les catégories

Dans ce chapitre nous avons présenté différentes méthodes pour la modélisation du contexte dans des images. La modélisation du contexte suit deux grandes approches, soit on modélise le contexte globale d'une scène et dans ce cas le contexte sert à aider la classification de la scène entière. Soit on modélise le contexte au niveau local et dans ce cas le contexte peut être utilisé d'une part pour décrire la scène globalement en agrégeant les différents contextes locaux ou d'autre part pour décrire le contexte localement à un objet ou une région de l'image. Dans cette thèse nous avons choisi de représenter le contexte au niveau local entre les superpixels extraits d'une image. Nous avons modélisé le contexte sous la forme d'un vecteur de contexte entre des pairs d'objets et décrivons le contexte local à un superpixel sous la forme d'un graphe des relations entre des ensembles de paires. Les méthodes présentées ici utilisent directement les sortis des classifieurs pour prédire la catégorie d'une image, dans notre cas nous utilisons la redondances des prédictions pour augmenter la robustesse aux erreurs de notre méthode.

6

Segmentation contextuelle

Sommaire

6.1 Introduction

Dans ce chapitre nous présentons notre approche pour intégrer l'information contextuelle dans la classification et la détection d'objets en imagerie aérienne. Nous sommes partis d'une méthode de classification sémantique de l'état de l'art basée sur des réseaux de neurones profonds qui atteint de très bonnes performances. Nous la décrivons précisément en section 6.1.2. En section 6.2 nous montrons comment nous pouvons caractériser le contexte dans le formalisme de cette méthode. Nous proposons une approche par apprentissage du modèle graphique pour construire un nouveau classifieur plus performant en section 6.3. Afin de donner une vue d'ensemble de l'approche nous la décrivons brièvement dans la suite.

6.1.1 Aperçu de la méthode

L'idée principale de l'utilisation du contexte est que certaines classes ou objets sont fréquemment co-localisés. Notre approche va chercher à modéliser statistiquement ce principe.

Deux types d'information sont utilisés pour la classification :

1. Les informations visuelles qui modélisent l'apparence d'une région
2. Les informations contextuelles qui modélisent les relations entre une région et ses voisines

Le contexte entre les différentes régions d'une image est modélisé sous la forme d'un modèle graphique qui va décrire les interactions entre elles. Les noeuds du modèle graphique sont les vecteurs décrivant l'apparence d'une région. Les arêtes du modèles encodent les relations contextuelles qui existent entre deux régions. Nous entraînons un modèle statistique sur les graphes extraits de données d'entraînement afin de construire un classifieur. Pour cela nous utilisons le modèle Structured Support Vector Machines (SSVM) qui est une extension du modèle SVM aux espaces structurés comme les espaces de graphes. Le SSVM permet ainsi de prédire les catégories de tous les nœuds d'un graphique en utilisant à la fois les informations visuelles et contextuelles. La Figure 6.1 illustre notre méthode de représentation du contexte par modèle graphique. L'image est au préalable segmentée en petites régions d'intérêt. Localement pour chaque région on considère les régions voisines, c'est à dire celles situées dans un cercle de rayon fixé. On définit un graphe local dont les différents noeuds sont les régions précitées. L'ensemble des graphes locaux constituent les données traitées (apprentissage et prédiction) par la SSVM.

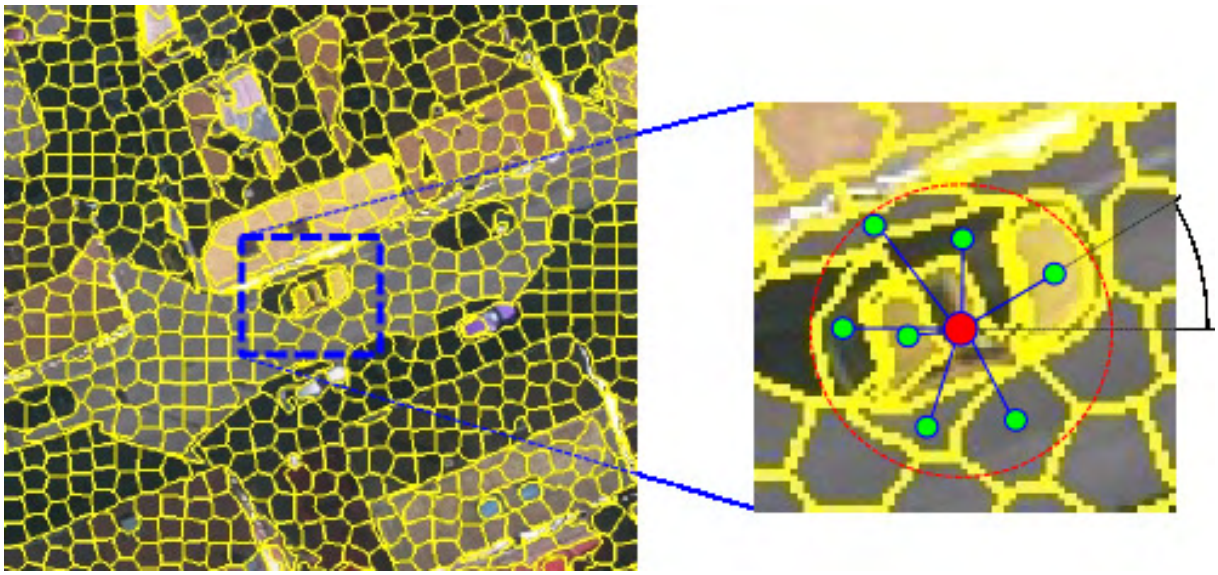


FIGURE 6.1 – Modèle de représentation des interactions locales entre les superpixels (cf. section 6.1.2) extraits d'une image. Pour chaque superpixel d'intérêt (noeud rouge) nous construisons un graphe local de relations avec les noeuds voisins (en vert). Ces relations sont caractérisées par divers attributs tel que distance, orientation, etc. (cf. section 6.2)

6.1.2 Segmentation sémantique d'images

Nous utilisons comme méthode de référence l'algorithme de segmentation sémantique proposé dans (LAGRANGE et al., 2015 ; AUDEBERT, LE SAUX, & LEFÈVRE, 2016). Il combine segmentation en superpixel, description par réseaux de neurones profond et SVM. Plus précisément les différentes étapes sont :

1. L'image est divisée en petite régions d'intérêt (les superpixels) en utilisant le Simple Linear Iterative Clustering (SLIC) (ACHANTA, SHAJI, & SMITH, 2012).
2. Un patch rectangulaire de taille $N \times N$ est extrait autour de chaque superpixel.

3. Le patch est redimensionné à une taille de 228×228 puis on lui applique un réseau de neurones convolutif sur le modèle d'AlexNet (Krizhevsky, Sulskever, & Hinton, 2012). Celui-ci était entraîné pour une tâche de classification d'images génériques, on enlève donc la dernière couche de classification pour ne garder que les couches convolutives qui permettent de fournir un vecteur descripteur de l'apparence de la région.

Les descripteurs extraits grâce au réseau AlexNet sont ensuite utilisés pour apprendre une SVM multiclasses qui réalisera les prédictions sur de nouveaux superpixels. La figure 6.2 illustre les différentes couches du réseau AlexNet.

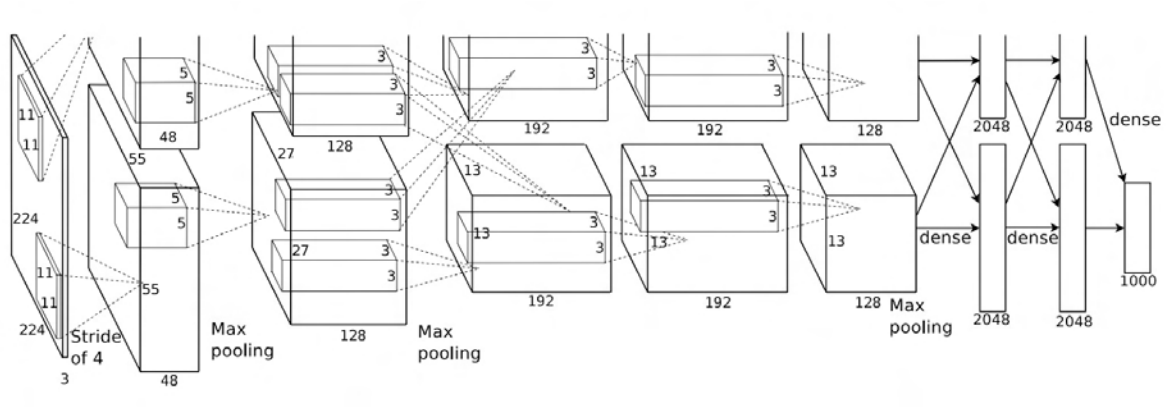


Figure 6.2 – Illustration des différentes couches d'un réseau AlexNet. Figure extraite de (Krizhevsky, Sulskever, & Hinton, 2012). Pour une approche descriptive la dernière couche de classification de taille 1000 est enlevée.

Cette approche fait l'hypothèse que tous les pixels d'un superpixel ont la même classe. En pratique c'est ce qui constitue son principal inconvénient. En effet, un superpixel peut recouvrir une zone correspondant à deux objets différents (par exemple dans une zone d'ombre) ce qui va poser des problèmes tant à l'entraînement qu'à la prédiction car on attribue à tous les superpixels la classe prépondérante. Inversement une même région sémantique peut être découpée en plusieurs superpixels qui sont classés indépendamment et peut-être de manière différente. Néanmoins cette méthode obtient des résultats comparables aux approches les plus performantes sur des bases de données de référence (« 2015 IEEE GRSS Data Fusion Contest », p.d. ; Rottensteiner et al., 2012) et surclasse des approches par descripteurs déterministes (c'est à dire non appris comme les HOGs) et SVMs.

Nous allons maintenant voir comment nous pouvons introduire de la dépendance entre superpixels voisins pour contrecarrer les erreurs de sur-segmentation.

6.2 Représentation du contexte local

6.2.1 Caractérisation du contexte dans des images

Dans le but de pouvoir modéliser l'information contextuelle entre deux objets nous devons d'abord identifier dans l'image quelles sont les relations de contexte entre les objets et comment nous pouvons les quantifier.

Suite à l'analyse de la littérature menée en chapitre 5 nous avons retenu les mesures ci-dessous. En général elles se basent sur une quantification de la similarité d'apparence ou des relations spatiales.

Pour chaque superpixel SP^n on note $x^n \in \mathbb{R}^d$ son vecteur d'apparence (dans le cas présent d est la dimension de sortie du réseaux convolutif décrit en section 6.1.2). De plus le centroïde de SP^n en coordonnées image est noté (u^n, v^n).

1. *Distance* : La distance euclidienne entre les centroïdes des deux noeuds (couramment utilisé dans les modèles graphiques markoviens comme par exemple dans (Schistad Solberg, Taxt, & Jain, 1996) et plus récemment dans (Volpi & Ferrari, 2015)). La relation de contexte dépend en effet de la distance entre deux objets, plus les objets seront éloignés moins ils auront d'influence l'un sur l'autre.

$$d(\mathrm{SP}^n, \mathrm{SP}^m) = \sqrt{|u^n - u^m|^2 + \|v^n - v^m\|^2} \tag{6.1}$$

2. *Orientation* : Angle réalisé par les centroïdes des deux noeuds avec le point de coordonnées $(0, 0)$ de l'image (Volpi & Ferrari, 2015).

$$\theta(\mathrm{SP}^n, \mathrm{SP}^m) = \mathrm{atan2}(|u^n - u^m|, |v^n - v^m|) \tag{6.2}$$

3. *Distance normalisée* : Nous introduisons la distance normalisée par la longueur de la diagonale de la boite englobante de l'union des deux noeuds. Cela permet de minimiser l'importance de la taille des objets lors de la comparaison de distances. En effet deux objets de grande taille auront une grande distance absolue entre leurs centroïdes alors que la distance normalisée permet de mieux rendre compte de leurs relation et de les comparer avec des objets de petite taille.

$$\bar{d}(\mathrm{SP}^n, \mathrm{SP}^m) = \frac{d(c^n, c^m)}{\mathrm{diag}(\mathrm{BB}(\mathrm{SP}^n \bigcup \mathrm{SP}^m))} \tag{6.3}$$

4. *Similarité visuelle* : Cette caractéristique permet de quantifier la ressemblance entre les apparences de deux superpixels. Nous utilisons la forme exponentielle introduite dans (Krähenbühl & Koltun, 2012) sous le nom *kernel of smoothness*.

$$s(\mathrm{SP}^n, \mathrm{SP}^m) = \exp\left(-\gamma \sum_{i=0}^{d} (x_i^n - x_i^m)^2\right) \tag{6.4}$$

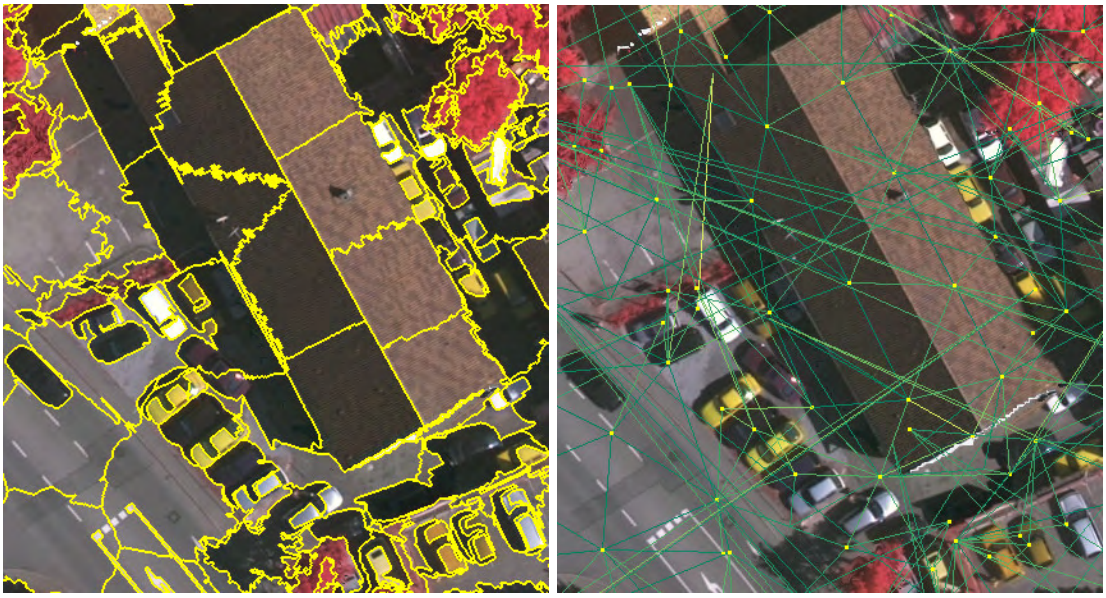
Ces quatre valeurs permettent de construire un vecteur descripteur de la relation pour chaque paire de superpixels. Ce vecteur contient beaucoup moins d'informations que les vecteurs d'apparence des noeuds cependant chaque noeud a plusieurs relations avec ses voisins ce qui constitue une caractérisation du contexte somme toute conséquente. Elle a vocation à être complémentaire de l'information d'apparence des noeuds.

6.2.2 Construction du graphe d'une image

Pour chacune des images de la base d'apprentissage on va créer un graphe global de contexte (cf. figure 6.3). Il représente les relations entre les différents objets ou superpixels dans l'image. Chaque objet dans l'image correspond à un noeud dans le graphe. Chacun des noeuds est représenté par un vecteur signature de l'objet qui quantifie son apparence et qui est extrait grâce au réseau de neurones. Une relation entre deux objets correspond à une arête. Afin de limiter la complexité du graphe, on ne considère que les objets au-delà d'un certain rayon r d'un noeud ne vont pas contribuer au contexte de l'objet. Dans le graphe cela se traduit par une absence

d'arête entre un noeud et les noeuds situés à une distance supérieure à r. Chacune des arêtes du graphe représente le contexte local entre deux objets. Ce contexte est capturé par le descripteur décrit en section 6.2.1. Les graphes locaux que nous avons décrits sur la figure 6.1 sont donc des sous-graphes de ce graphe global. Ils représentent le contexte local d'un objet avec sur les noeuds la représentation visuelle d'un objet et sur les arêtes la représentation contextuelle.

6.3 Apprentissage sur des graphes de contexte



(a) (b)

FIGURE 6.3 – Visualisation du contexte local entre les superpixels d'une image. La figure 6.3a représente une segmentation en superpixels de l'image et la figure 6.3b représente les liens de contexte entre les différents superpixels. La couleur des liens entre les superpixels représente l'importance du contexte entre les superpixels.

Nous présentons maintenant l'apprentissage d'un modèle à sortie structurée le SSVM tout d'abord de manière théorique en section 6.3.1 puis son application dans le cadre des graphes de contexte tel que défini à la section précédente en section 6.3.2.

6.3.1 Modèles à sorties structurées

Dans cette section nous allons décrire la procédure d'apprentissage permettant d'apprendre les paramètres d'un modèle structuré. Les méthodes structurés permettent l'apprentissage de modèles pour des problème où les variables d'entrées présentent une structure comme les chaines de Markov ou les graphes.

Le but de la prédiction structurée est de prédire un objet structuré $y \in \mathcal{Y}$ comme les noeuds d'un graphe où une séquence de mots pour un exemple $x \in \mathcal{X}$ en entrée. L'approche standard pour la prédiction structurée consiste à définir une fonction de projection jointe $\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^D$ qui encode un exemple et un classifieur linéaire w définit par la fonction de score suivante :

$$\begin{aligned} h_w(x) &= \arg\max_{y \in \mathcal{Y}} g(x, y, w) \\ &= \arg\max_{y \in \mathcal{Y}} \langle w, \phi(x, y) \rangle \end{aligned} \tag{6.5}$$

Pour un ensemble d'apprentissage $\mathcal{D} = \{(x^n, y^n) \mid n = 1, \ldots, N\}$ les paramètres du classifieur linéaire sont estimés en résolvant le problème suivant :

$$\min_{w \in \mathbb{R}^D} \frac{\lambda}{2}\|w\| + \frac{C}{N}\sum_{n=1}^{N} \Delta(y^n, h_w(x^n)) \tag{6.6}$$

Cependant dans le cas de la prédiction structurée, résoudre l'équation (6.6) est un problème infaisable à cause de $\Delta(y^n, h_w(x^n))$. Généralement la fonction de coût Δ est constante par morceaux ce qui rend les méthodes d'optimisation par descente de gradients inutiles. Cependant utiliser une borne supérieure convexe de la fonction de coût permet de résoudre l'équation (6.6) (T. Zhang, 2004) qui est réécrite :

$$\min_{w \in \mathbb{R}^D} \frac{\lambda}{2}\|w\| + \frac{C}{N}\sum_{n=1}^{N} l(x^n, y^n, w) \tag{6.7}$$

Avec la fonction l qui est une borne supérieure convexe de la fonction Δ et qui est définie par :

$$l(x^n, y^n, w) = \max_{y \in \mathcal{Y}} \Delta(y^n, y) - g(x^n, y^n, w) + g(x^n, y, w) \tag{6.8}$$

L'Équation (6.8) généralise la *Hinge loss* dans le cas de sorties structurées (Taskar, Guestrin, & Koller, 2003). L'utilisation de l'équation (6.8) permet de voir le problème d'optimisation structurée comme étant une extension du problème du SVM aux sorties structurées. La fonction l juge si la prédiction y^n du modèle w sur l'entrée structurée x^n est *bonne* ou *suffisamment similaire* des valeurs du vecteur de labels à prédire. Un avantage du SSVM sur le SVM est que la fonction de coût Δ peut être définie arbitrairement tant que les propriétés d'une fonction de coût sont respectées. (Nowozin, 2010) présente ces fonctions de coûts très utilisées en vision par ordinateur :

Zero-One loss : $\Delta(y, y') = [\![y \neq y']\!]$. Cette fonction vaut 0 si y et y' sont différent sinon 1. Il s'agit de la fonction la plus usuellement utilisées pour les problèmes de classification multiclasses. Elle est peu utilisée pour faire de la prédiction structurées où on ne recherche pas forcément à prédire exactement les *vrais* valeurs de la sortie se révèle être très difficile.

Hamming loss : $\Delta(y, y') = \frac{1}{M}\sum_{n=1}^{M} [\![y_i \neq y'_i]\!]$. Cette fonction est principalement utilisée dans les modèles qui sont utilisés pour faire de la segmentation sémantique. Ici chacun des y_i représentent la catégorie d'un pixel ou d'un superpixel et la valeur de Δ est le nombre moyen d'erreurs de classification que fait le modèle.

Area Overlap : $\Delta(y, y') = \frac{\text{area}(y \cap y')}{\text{area}(y \cup y')}$. Il s'agit de la fonction standard utilisée pour la localisation d'un objet dans une image pour la compétition Pascal (Everingham, Gool, Williams, Winn, & Zisserman, 2010). Ici y et y' sont les coordonnées des boîtes englobantes de la prédiction et de la vérité terrain.

6.3.2 Application aux graphes de contexte

À partir de chacune des images, un graphe est extrait mettant en relation les superpixels de l'image avec leurs voisins. Chacun des superpixels est décrit par un vecteur descripteur extrait comme décrit en section 6.1.2. Les superpixels sont modélisés par les noeuds du graphe et les relations entre les superpixels par les arrêtes du graphe. C'est sur les arêtes du graphe que sont stockée les descripteurs de contexte décrit en section 6.2.1. À partir de chaque noeud du graphe de l'image, on extrait les sous-graphes locaux des noeuds situés a une distance inférieure à r pixels du noeud d'intérêt. Un sous-graphes de M noeuds est noté par $x^n = \{(x_i^n, y_i^n) \mid i = 0, \ldots, M\}$. Avec x_0^n le descripteur du superpixel que l'on souhaite classifier et $\{x_i^n \mid i = 1, \ldots, M\}$ sont les descripteurs des superpixels voisins. Avec les $y^n = \{y_i^n \mid i = 0, \ldots, M\}$ sont les sorties structurées qui correspondent aux catégories des x_i^n noeuds du graphe. L'ensemble des couples $X = \{(x^n, y^n) | n = 1, \ldots, N\}$ vont servir d'exemples d'apprentissage pour le SSVM.

L'évaluation de la fonction de score équation (6.5) est un problème d'optimisation sur des graphes qu'il faut résoudre. Plusieurs bibliothèque pour résoudre les problèmes d'inférences sur des modèles graphiques existent et ont été analysées par (Kappes et al., p.d.). Les méthodes que nous avons utilisé sont :Quadratic Pseudo-Boolean Optimization (QPBO) (Kolmogorov & Rother, 2007) et Alternating Directions Dual Decomposition (AD^3) (Martins et al., 2012).Pour nos expériences, nous avons choisi d'utiliser la méthode QPBO pour résoudre le problème d'inférence sur les graphes. Les solutions proposées par AD^3 sont beaucoup plus précises mais cette méthode est beaucoup plus lente que QPBO et nous devons apprendre le modèle sur un grand nombre d'exemples d'apprentissage.

Dans le cadre des modèles graphique, le problème d'inférence équation (6.5) se réécrit en fonction des noeuds et des arêtes du modèle sous la forme suivante :

$$h_w(x) = \arg\max_{y \in \mathcal{Y}} \sum_{i=0}^{N} \langle w, \phi_{node}(x_i, y_i) \rangle + \sum_{i,j=0}^{N} \langle w, \phi_{edge}(x_i, x_j, y_i, y_j) \rangle \tag{6.9}$$

Avec ϕ_{node} et ϕ_{edge} les fonctions de projection jointe pour les noeuds et les arêtes d'un graphe x

Parmi les paramètres du modèle nous avons choisis comme fonction de coût la *Hamming Loss* car c'est la plus utilisée pour les tâches de segmentation sémantique La *Hamming Loss* calcule la moyenne des erreurs de prédiction de la catégorie de chacun des noeuds du sous-graphe.

Pour l'apprentissage nous avons utilisé deux algorithme d'optimisation pour apprendre les modèles structurés :Block-Coordinate Frank-Wolfe (BCFW) (Lacoste-julien, Jaggi, Schmidt, & Pletscher, 2013)et Subgradient Descent (SD) (Ratliff, Bagnell, & Zinkevich, 2007). Sachant le nombre d'exemples, les deux méthodes ne sont pas toujours applicables. Dans les cas où la taille des données d'apprentissage X est très grande, la méthode SD est préférée car elle est peu gourmande en mémoire mais va converger lentement en $O\left(\frac{1}{\epsilon}\right)$ où ϵ est l'erreur du modèle sur l'ensemble d'entraînement. Dans les cas où le taille du problème était modérée, la méthode BCFW était préférée car elle permettait d'obtenir un modèle beaucoup plus rapidement en $O\left(\sqrt{\epsilon}\right)$ mais consommait plus de mémoire que SD. Dans les expériences nous utilisons l'implémentation de (Müller & Behnke, 2014) distribuée sous la forme de la bibliothèque *PyStruct*.

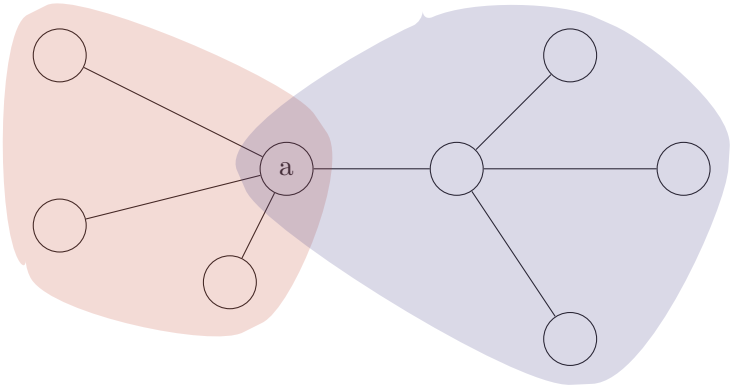


FIGURE 6.4 – Prédiction sur deux sous-graphes de contexte. Comme le noeud (a) est commun aux deux graphes (il est un voisin commun aux deux superpixels) le modèle va prédire pour ce noeud une catégorie alors qu'il appartient au graphe bleu puis une catégorie alors qu'il appartient au graphe rouge. Les prédiction sont accumulées et la catégorie final du noeud (a) sera la catégorie ayant reçu le plus de votes.

6.3.3 Prédiction d'une catégorie

Lors de la prédiction pour tous graphe local le modèle renvoie la prédiction des labels de l'ensemble des noeuds. C'est à dire du superpixel dont on veut prédire la catégorie mais aussi tous ses voisins (cf. figure 6.4).

Nous exploitons cette redondance en accumulant les prédictions pour chaque superpixel. La catégorie assignée au final étant la catégorie ayant reçu le plus de vote de la part des voisins. Ainsi la catégorie d'un superpixel est déterminée par la prédiction que le modèle fait sur lui le superpixel d'intérêt mais aussi par les différentes prédictions que les voisins lui assignent en fonction du contexte.

6.4 Résultats expérimentaux

Nous présentons dans cette partie plusieurs expériences permettant d'analyser et de montrer l'apport du contexte pour la classification. Nous évaluons les différentes caractéristiques de contexte ainsi que leurs combinaisons et nous comparons une approche contextuelle à une approche simple sur un jeux de données de référence.

6.4.1 Données et contenu des expériences

La méthode de segmentation sémantique a été testée sur le jeu de données fournit par l'ISPRS dans le cadre du banc d'essai de cartographie sémantique nommé (GERKE, 2015). Il contient

TABLE 6.1 – Vérité terrain pour la tâche classification sémantique for la base d'images Vaihingen décrite en section 2.3.4.

Asphalte	Bâtiments	Végétation basse	Arbres	Voitures	Fouillis

des images aériennes sur 3 canaux IR-R-G et une vérité terrain définie pour chaque pixel selon 6 classes : routes, bâtiments, végétation basse, arbres voitures et une classe fouillis pour le reste. Cette base est décrite complètement en section 2.3.4. Nous utilisons une partie des images de cette base pour l'apprentissage et une autre pour l'évaluation. Plus précisément nous découpons la base de la manière suivante : Les tuiles 1, 5, 7, 11, 17, 23, 26, 28, 34 et 37 servent pour l'ensemble d'apprentissage (les tuiles en question peuvent être vues sur la figure 2.6a). Les tuiles 13, 21 et 30 servent pour la validation du modèle et enfin les tuiles 3, 15 et 32 servent pour évaluer les performances du modèle en prédiction. Pour l'évaluation des performances, nous utilisons deux mesures décrites en section 2.2.5 et section 2.2.4

1. Une mesure de performances de classification par classe : le f_1-score (cf.section 2.2.5)
2. Une mesure de performance globale sur toutes les classes : le taux de classification (cf. section 2.2.4)

6.4.2 Évaluation des caractéristiques de contexte local

Nous avons défini différentes caractéristiques de contexte utilisables pour l'imagerie aérienne en section 6.2. Nous avons évalué leurs qualités intrinsèques sur une tâche de classification de relations. Étant donné qu'il y a 6 classes de base, une paire d'objets $p_{m,n}$ a le couple d'étiquettes ce qui définit 21 nouvelles catégories compte tenu des permutations. La tâche de classification de relations consiste donc à prédire la fonction $h(x^n, x^m) \rightarrow \{1, \ldots, 6\}^2$

Pour chaque paire d'objets nous avons calculé les différentes mesures (distance, similarité visuelle etc.). Nous construisons donc un ensemble d'apprentissage avec ces différentes mesures et les catégories de couple de classes. Une SVM multiclasses simple est utilisée pour apprendre et tester les classifieurs basés sur les différentes mesures et leurs combinaisons possibles. Les résultats sont compilés dans le tableau 6.2. Il en ressort que d'une part seules les caractéristiques de contexte ne permettent pas de résoudre la tâche de classification de relations, ce qui était prévisible. D'autre part cela nous permet néanmoins d'identifier les mesures informatives et leur combinaison optimale : similarité visuelle, distance et orientation. On remarque que ne serait-ce que pour caractériser les relations entre régions, un mélange d'informations visuelles et spatiales conduit aux meilleurs résultats.

6.4.3 Évaluation de l'apport du contexte

À partir des caractéristiques choisies à partir des résultats du tableau 6.2 (mesure de similarités, distance et orientations) nous avons extrait un ensemble de sous-graphes labellisés $(X = \{x^n\}_{n=1}^N, Y = \{y^n\}_{n=1}^N)$ des images d'apprentissage. À partir de cet ensemble nous avons appris un modèle structuré que nous avons utilisé pour la tâche de classification des superpixels d'une image. Dans le but de quantifier l'apport d'ajout d'informations contextuelles, nous comparons notre méthode avec deux autre méthodes où le contexte n'est pas utilisé. La première méthode a laquelle nous nous comparons est une méthode de classification basée sur la description des superpixels en utilisant le réseaux convolutif AlexNet pour des régions de taille 32×32 centrées le superpixel d'intérêt (*Baseline32*). La seconde méthode est analogue à la première mais en se basant sur des patches deux fois plus étendus (64×64). Avec des patches deux fois plus étendus, le réseau encode des informations sur les pixels voisins du superpixel à encoder ce qui constitue une forme de contexte. La méthode *Contexte Structurel* est la méthode de classification contextuelle que nous présentons section 6.3 Pour les deux méthodes, un SVM linéaire est utilisé pour réaliser la prédiction de la catégorie du superpixel.

TABLE 6.2 – Comparaison selon le taux de classification des différentes caractéristiques de contexte local et leurs combinaisons. Le meilleur taux de classification est obtenu pour la combinaison d'indices de similarité visuelle d'une part et de distance et d'orientation d'autre part. *Dist* est la distance définit par le point 1 de la section 6.2.1, *DistNorm* correspond au point 3, *SimVis* au point 4 et *Orientation* au point 2.

Caractéristique de contexte	Taux de classification	Écart type
Dist	18.17	0
DistNorm	10.51	1.11
SimVis	6.88	0.8
Orientation	19.01	1.01
Dist/DistNorm	10.51	1.1
Dist/SimVis	6.91	0.8
Dist/Orientation	**19.03**	**1.03**
DistNorm/SimVis	4.88	0.54
DistNorm/Orientation	7.87	0.85
SimVis/Orientation	8.08	1.37
Dist/DistNorm/SimVis	4.91	0.54
Dist/DistNorm/Orientation	7.99	1.01
Dist/SimVis/Orientation	8	1.43
DistNorm/SimVis/Orientation	5.11	0.79
Dist/DistNorm/SimVis/Orientation	5.22	0.87

TABLE 6.3 – Comparaison selon le f_1-score et le taux de classification des différentes méthodes de segmentation sémantique sur l'ensemble de validation.

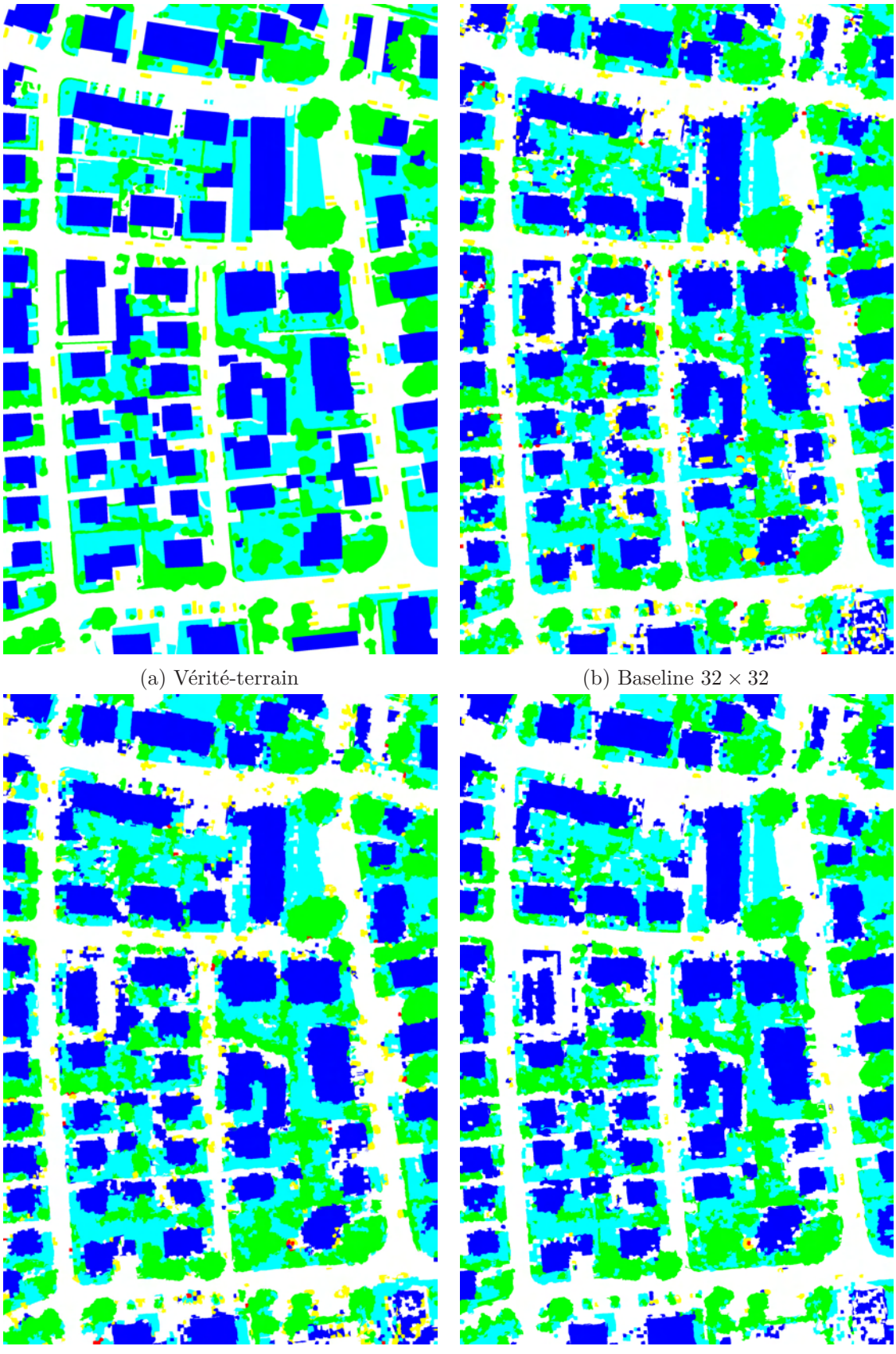
	f_1-score					Acc.
Model	Imperv.	Build.	Veget.	Tree	Cars	
Baseline 32	81.26	81.58	**62.71**	77.88	40.10	76.33
Baseline 64	81.13	82.36	62.46	76.13	**41.03**	75.98
Contexte Structurel	**82.00**	**82.40**	58.18	**78.38**	32.46	**78.36**

Le tableau 6.3 montre des différents scores de classification pour les méthodes testées. On observe que pour 3 catégories sur 5 notre modèle donne de meilleurs scores que les autres méthodes. Pour les catégorie végétation basse et voitures les scores sont par contre en dessous des autres méthodes ce qui peut s'expliquer par un contexte plus varié et donc moins informatif pour ces deux catégories. Avec le taux de classification sur toutes les classes, notre méthode donne de meilleurs scores que les deux autres méthodes.

Nous validons les résultats quantitatifs de tableau 6.3 par une visualisation des résultats de segmentation des différentes méthodes sur les 3 images de validation. Pour chacun des test nous présentons d'abord la vérité terrain utilisée pour valider les résultats puis nous montrons les différentes méthodes que nous comparons. Sur les figures 6.5 à 6.7 nous montrons les résultats de classification sémantique sur les 3 tuiles de validation. La première chose que l'on peut remarquer sur ces images est que le modèle de contexte produit une carte de classification beaucoup moins bruitée que les deux autres méthodes. Notre méthode permet aussi de réduire de beaucoup certaines classifications aberrantes comme des prédictions de voitures au milieu des bâtiments.

Cette propriété est particulièrement visible sur la figure 6.8. Cette figure est un agrandissement de la zone située en bas à droite de la tuile n° 3.

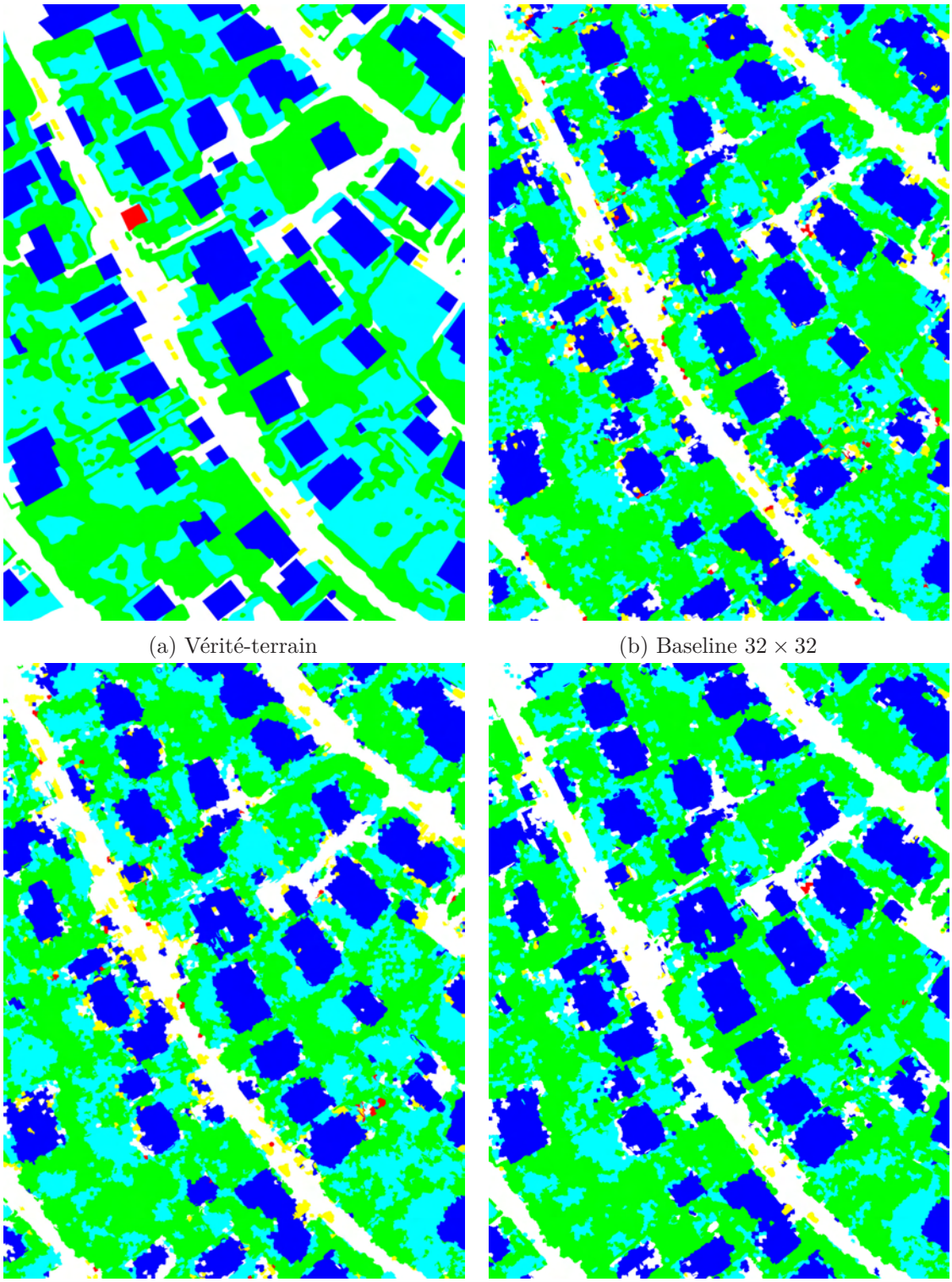
Nous présentons ensuite une comparaison qualitative de notre méthode avec plusieurs autres méthodes de segmentation sémantique en figure 6.9.



(a) Vérité-terrain (b) Baseline 32×32

(c) Baseline 64×64 (d) Contexte Structurel

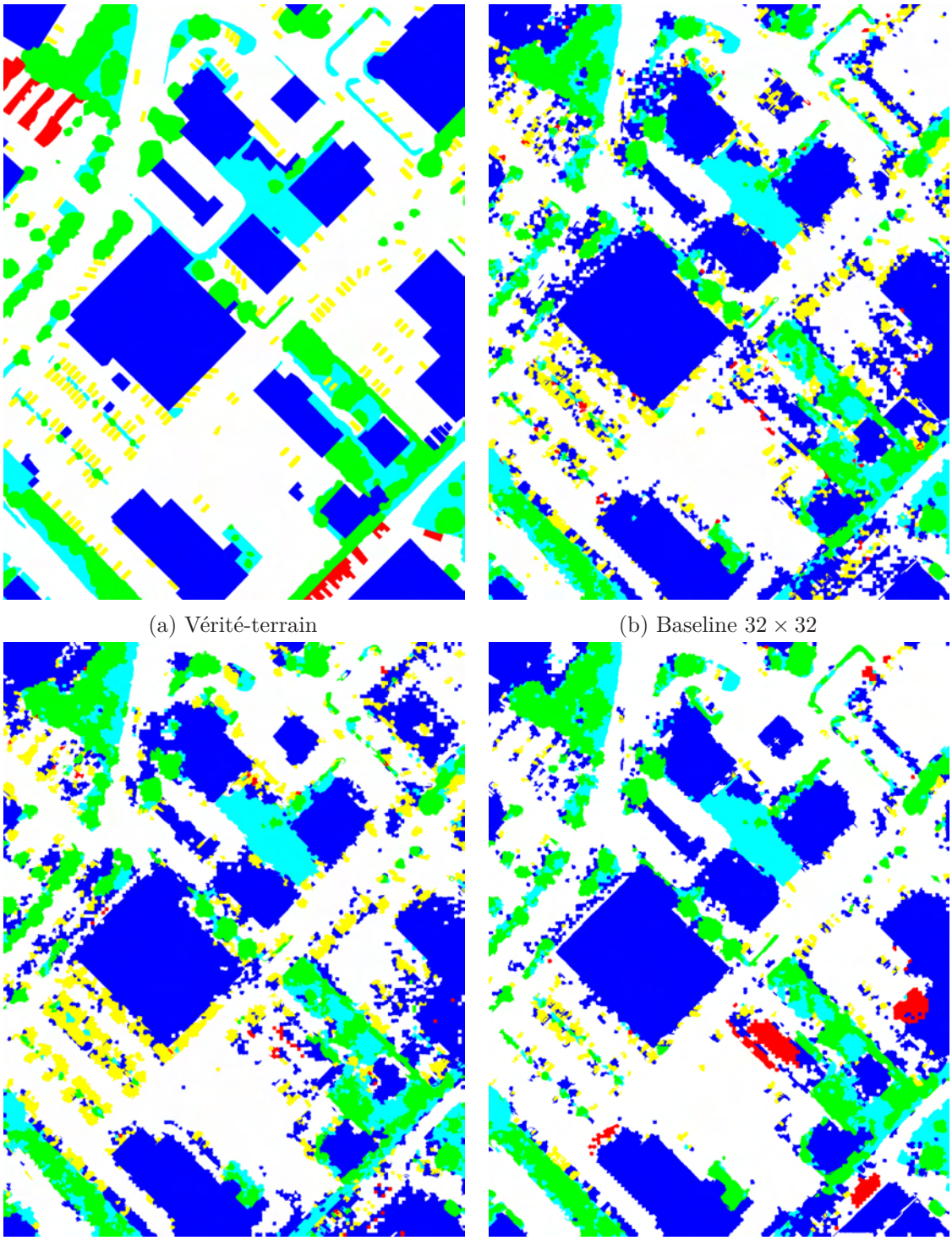
FIGURE 6.5 – Cartes sémantiques prédites pour la tuile 3



(a) Vérité-terrain (b) Baseline 32×32

(c) Baseline 64×64 (d) Contexte Structurel

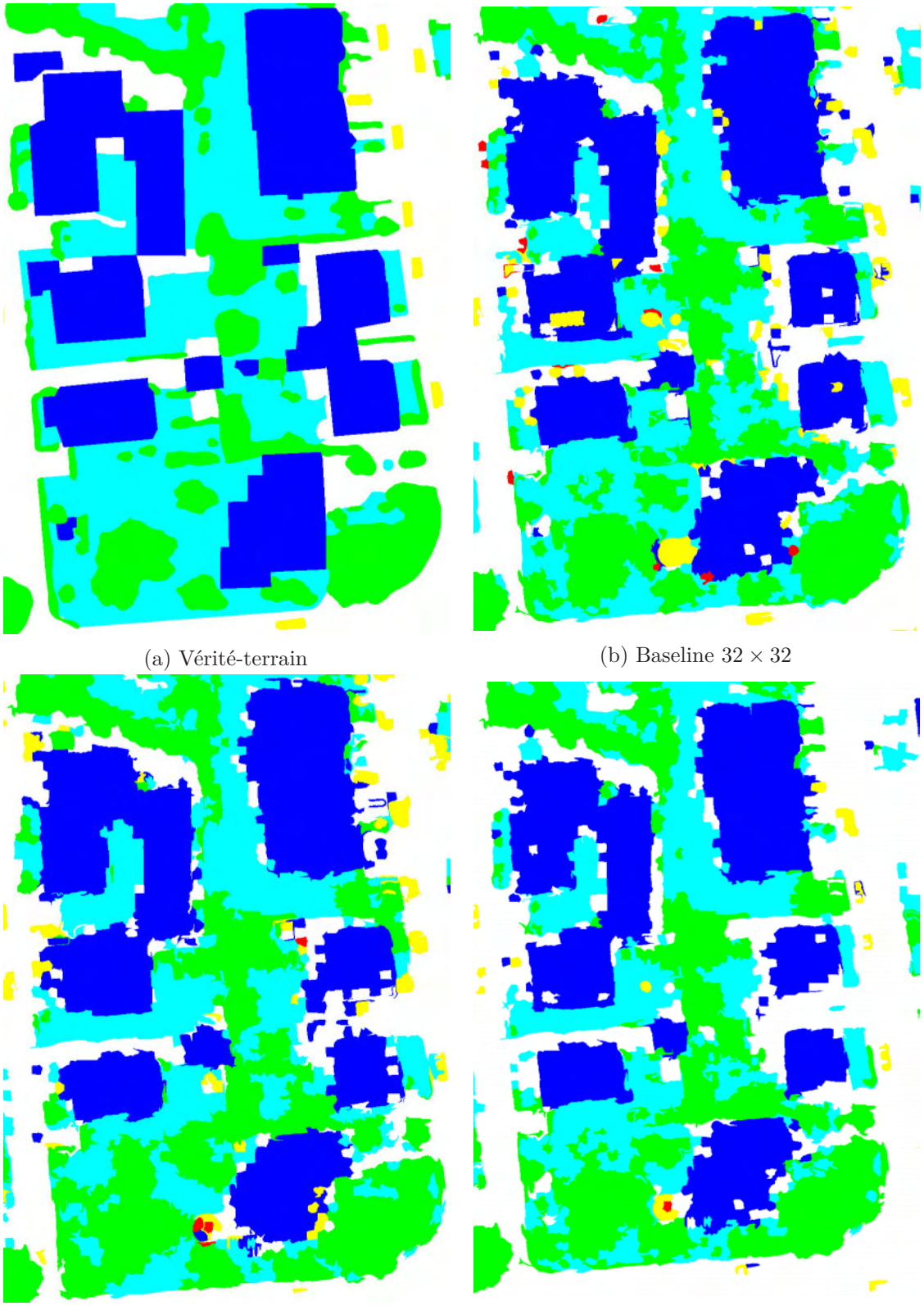
FIGURE 6.6 – Cartes sémantiques prédites pour la tuile 15



(a) Vérité-terrain (b) Baseline 32×32

(c) Baseline 64×64 (d) Contexte Structurel

FIGURE 6.7 – Cartes sémantiques prédites pour la tuile 32



(a) Vérité-terrain

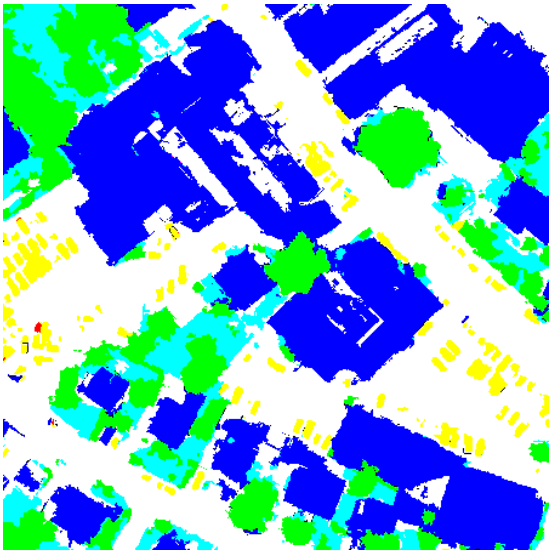
(b) Baseline 32×32

(c) Baseline 64×64

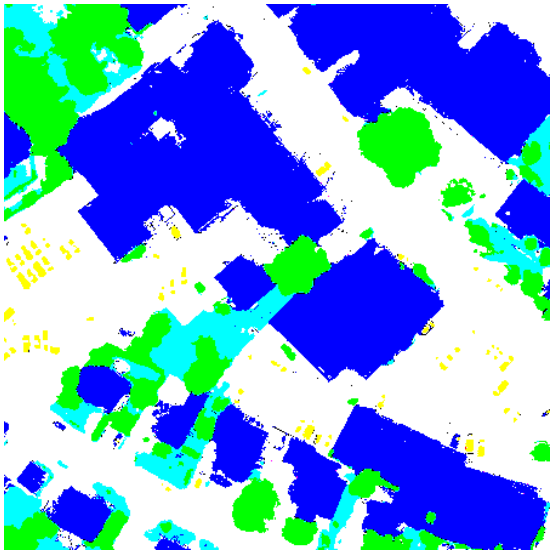
(d) Contexte Structurel

FIGURE 6.8 – Zoom sur la tuile 3

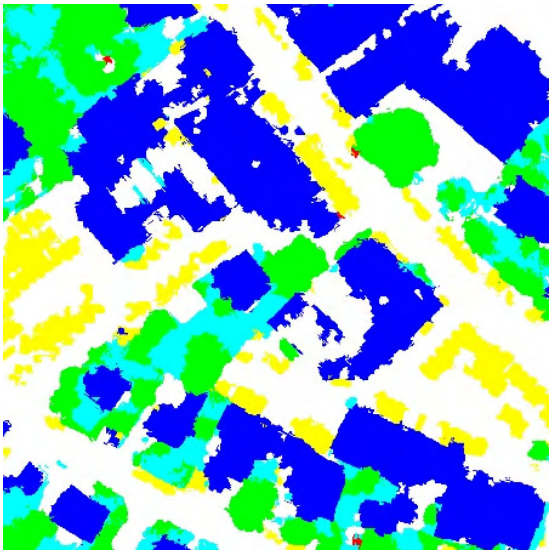
(a) Données IR-R-G



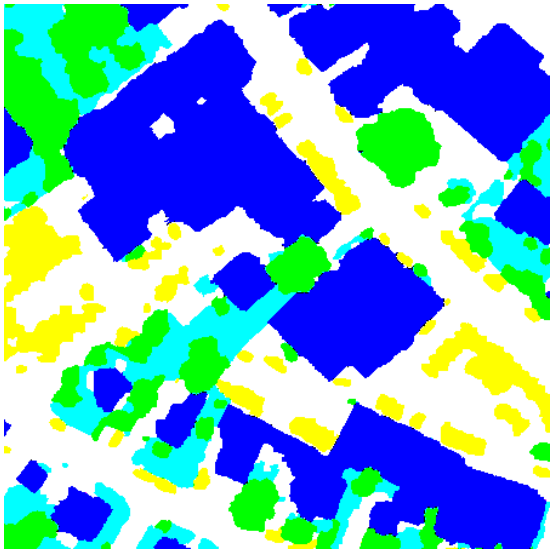
(b) SVL (Gerke, 2015)



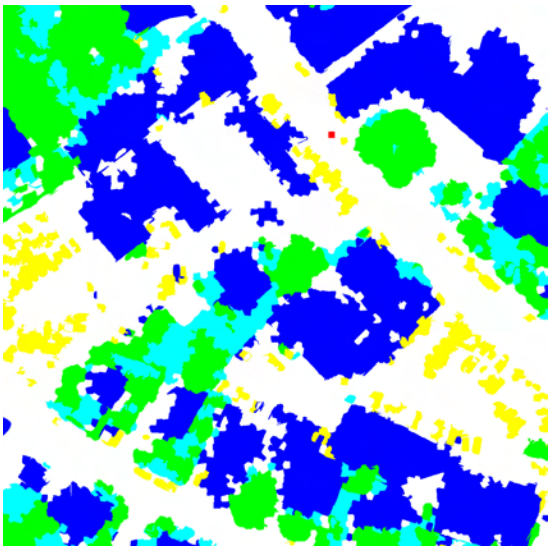
(c) RF + CRF (Quang, Thuy, Sang, & Binh, 2015)



(d) Multi-échelles (Audebert, Le Saux, & Lefèvre, 2016)



(e) CNN + RF + CRF (Paisitkriangkrai, Sherrah, Janney, & Van-Den Hengel, 2015)



(f) Notre méthode

Figure 6.9 – Segmentations provenant de différentes méthode sur un extrait de l'ensemble de test Vaihingen.

6.5 Conclusions

Dans ce chapitre nous avons présenté nos travaux sur la modélisation du contexte dans des images aériennes. Notre méthode s'articule selon deux grands axes : d'abord une quantification du contexte via un vecteur descripteur puis l'apprentissage d'un modèle graphique de contexte. Nous avons montré que l'utilisation du contexte pour la classification de régions d'une image aérienne permet d'améliorer les performances du classifieur. Le principal avantage de l'introduction du contexte dans la classification est de rendre les résultats plus facilement interprétables par un humain plutôt que le gain en performances pur. De plus les performances du modèle vont beaucoup dépendre de la qualité des descripteurs visuels utilisés. Si les descripteurs visuels sont peu performants alors le contexte va permettre de compenser leurs performances. Cependant si les descripteurs visuels sont déjà très performants, alors le gain en performance lié au contexte sera moins significatif.

7

Conclusions et perspectives

Depuis quelques années, les images aériennes et satellitaires à très haute résolution constituent une source de données précieuse pour de nombreuses applications. Cependant par manque de temps et de ressources il n'est pas toujours possible d'exploiter totalement les informations contenues dans ces images. Ceci est l'une des raisons pour laquelle l'analyse d'images aériennes et satellitaires est un vaste domaine de recherche avec de nouvelles avancées chaque jour. Récemment nous avons vu augmenter la disponibilité des bases d'images aériennes très haute résolution. Dans ces images, les objets sont suffisamment résolus pour nous permettre d'utiliser des méthodes d'analyse d'image populaires dans la communauté vision par ordinateur comme le descripteur HOG pour la détection d'objets et le *deep learning* pour la classification de pixels.

Dans cette thèse, nous nous sommes concentrés sur les approches utilisées par la communauté vision par ordinateur pour l'analyse d'images et comment les adapter aux particularités de l'imagerie aérienne et satellitaire. Nous avons choisi d'aborder le problème de la classification du contenu d'une image aérienne selon deux grandes approches : la première approche consistait à décrire le contenu d'une image en analysant et en détectant automatiquement les objets présents dans celle-ci, la seconde approche consistait à décrire le contenu d'une image grâce aux prédiction d'un classifieur sur les différentes régions de l'image.

Dans nos travaux, nous avons proposé une méthode de détection d'objets dans des images basée sur un mélange de modèles et une méthode de segmentation sémantique se basant sur un classifieur appris en utilisant à la fois des informations visuelle et des informations contextuelles. Nous avons mis en évidence que lors de la constitution d'une base d'apprentissage pour une tâche de reconnaissance visuelle, les catégories sémantiques tel que "voiture" ou "bâtiments" ne sont pas suffisantes pour caractériser les exemples d'apprentissage. Il existe dans ces catégories sémantique très variées des catégories visuelles que les modèles de reconnaissance ne sont pas toujours capables de gérer. Typiquement les modèles basés sur le descripteur HOG qui est très efficace pour la détection d'objets dans des images sont très sensibles aux changements de pose ou d'orientation des objets dans la base d'images. La modélisation de sous-catégories visuelles où les variations de pose, d'orientation et d'apparence sont réduites permettent d'utiliser ces descripteurs très spécialisés dans le cadre de l'imagerie aérienne. L'apprentissage d'un mélange de modèles se basant sur un ensemble de sous-catégories donne un détecteur d'objets robuste aux changements d'orientation des objets dans l'image. Nous avons appliqué cette idée à la détection de véhicules dans des images aériennes. Nous avons choisi de nous concentrer sur les véhicules car ce sont les objets dans dont la forme est bien définit contrairement aux bâtiments ou aux arbres. Cependant notre méthode permet aussi de détecter ces objets dans des images mais donne des résultats beaucoup moins performants. Au cours de nos expérimentations nous avons remarqué

que le principal problème pour reconnaître les objets dans une image venait de la représentation de la catégorie de l'objet. Les descripteurs d'images traditionnels que nous avons à disposition tel que le HOG permettent de faire de la détection d'objets dans des images aériennes mais ne décrivent pas suffisamment fortement les objets pour détecter tous les objets d'une image. Pour augmenter les performances de détection des méthodes basées sur ces descripteurs, nous devons mettre en place des stratégies comme la modélisation de sous-catégories visuelles. Cependant le HOG à l'avantage d'être peu coûteux à calculer sur une image et la formulation de ce descripteur sous la forme d'un filtre permet de l'utiliser efficacement en conjonction avec des méthodes de mise en correspondance de modèles efficaces sur des images de grande dimension comme la NCC. Les méthodes permettant une représentation forte des objets comme les méthodes basées une description des objets utilisant le *deep learning* ont montré de très bonnes performances pour la classification d'images. Il existe des méthodes permettant de faire de la détection d'objets dans des images en utilisant le *deep learning* avec des performances très supérieures aux méthodes basées les descripteurs classiques. Cependant pour la localisation des objets dans l'image ces méthodes se basent sur des techniques de proposition de boîtes englobantes qui sont inutilisables en imagerie aérienne. La localisation des objets dans l'image doit donc se faire via une fenêtre glissante mais le nombre de fenêtres à tester pour avoir un détecteur d'objets efficace est beaucoup trop grand pour pouvoir utiliser le *deep learning* de cette façon. Une future orientation serait de pouvoir adapter les méthodes de détection d'objets basées sur du *deep learning* en imagerie aérienne. Une façon de faire serait d'adapter les méthodes de proposition de boîtes englobantes aux images aériennes afin de proposer un ensemble restreint de localisation où peut se trouver un objet.

La suite de nos travaux a concerné la segmentation sémantique d'image aérienne utilisant à la fois des informations visuelles et des informations de contexte. Pour ces travaux nous avons décidé de baser notre méthode sur une segmentation de l'image en superpixels et une description du contenu des superpixels utilisant le *deep learning* et plus particulièrement le réseaux de neurones convolutif AlexNet. La description du contexte est faite en quantifiant les informations de contexte entre les superpixels de l'image. Nous avons postulé que le contexte entre les superpixels d'une image peut être modélisé et nous avons proposé une méthode pour quantifier et apprendre ce contexte.

Dans nos expériences, le contexte est modélisé par un ensemble de caractéristiques spatiales et visuelles que nous combinons sous la forme d'un vecteur quantifiant le contexte entre une paire de superpixels. Nous avons évalué les différentes combinaisons de caractéristiques avec l'idée que certaines combinaisons permettent de mieux reconnaître la relation entre deux superpixels que d'autres. Dans le cas des images aériennes nous avons montré que la distance et l'orientation entre deux superpixels sont les caractéristiques les plus efficaces pour reconnaitre le contexte. Nous avons aussi montré que même la distance ou l'orientation entre les superpixels donnent de bonnes performances de reconnaissance des catégories des paires de superpixels. L'idée de modéliser le contexte en utilisant des paires de superpixels nous a permis d'utiliser le formalisme des modèles graphiques pour modéliser le contexte d'un superpixel. Cela a permis une représentation flexible du contexte qui dépend entièrement des relations d'un superpixel avec ses voisins. De plus le modèle graphique nous a permis de combiner les informations visuelles des superpixels avec les informations de contexte entre les superpixels. Afin d'apprendre un prédicteur à partir des graphes de contexte nous avons utilisé une méthode d'apprentissage appelée SSVM. La SSVM permet de prédire les catégories de l'ensemble des noeuds d'un graphe. Nous avons utilisé cette propriété pour proposer une méthode de prédiction de la catégorie d'un superpixel qui prend en compte les prédictions des graphes des voisins. Notre méthode de segmentation sémantique utilise un réseaux AlexNet pour décrire de contenu d'un superpixel. L'utilisation d'un réseaux comme extracteur de descripteur est une pratique courante car les descripteurs issus d'un réseaux de neurones profonds sont bien souvent de meilleur qualité que les descripteurs traditionnel.

Cependant cette utilisation d'un réseau de neurones ne permet pas d'en exploiter pleinement les capacités. Une des forces des réseaux de neurones profond est d'apprendre toute la chaîne de traitement d'une image depuis l'extraction des caractéristiques jusqu'à sa classification. Une voie d'amélioration de notre méthode serait d'intégrer la SSVM au réseau pour que l'extraction des caractéristiques soit directement guidé par la tâche de classification. L'introduction de modèles à sorties structurées dans un réseau de neurones profond est un des défis que cherche à relever la communauté apprentissage statistique dernièrement. De nouvelles architectures de réseaux de neurones profond ont vu le jour pour les tâches de segmentation sémantique. Cependant ces réseaux utilisent encore très peu les relations spatiales entre les exemples lors de la phase d'apprentissage. Une perspective intéressante serait d'intégrer les modèles à sorties structurées dans ces nouvelles architectures pour permettre au réseau d'apprendre une représentation du contexte entre les exemples de la base d'apprentissage.

Table des figures

Publications

Campos-Taberner, M., Romero-Soriano, A., Gatta, C., Camps-Valls, G., Lagrange, A., Le Saux, B., ... Tuia, D. (2016). Processing of Extremely High-Resolution LiDAR and RGB Data : Outcome of the 2015 IEEE GRSS Data Fusion Contest–Part A : 2-D Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1–13.

Lagrange, A., Le Saux, B., Beaupere, A., Boulch, A., Chan-Hon-Tong, A., Herbin, S., ... Ferecatu, M. (2015). Benchmarking classification of earth-observation data : from learning explicit features to convolutional networks. In *Ieee international geoscience and remote sensing symposium (invited talk in the special session on data fusion)* (T. 2015-Novem, p. 4173–4176).

Randrianarivo, H., Le Saux, B., Audebert, N., Crucianu, M., & Ferecatu, M. (2016). Structural classifiers for contextual semantic labeling of aerial images. In *Conference on big data from space.*

Randrianarivo, H., Le Saux, B., Crucianu, M., & Ferecatu, M. (2016). Discriminatively-trained model mixture for object detection in aerial images. In *10th esa-eusc-jrc conference on image information mining.*

Randrianarivo, H., Le Saux, B., & Ferecatu, M. (2013). Urban structure detection with deformable part-based models. In *Ieee international geoscience and remote sensing symposium* (p. 200–203).

Randrianarivo, H., Le Saux, B., & Ferecatu, M. (2014). Multimodal classification with deformable part-based models for urban cartography. In *Ieee international geoscience and remote sensing symposium (invited talk in the special session on data fusion)* (p. 203–206).

Randrianarivo, H., Le Saux, B., & Ferecatu, M. (2015). Détection de véhicules en imagerie aérienne par mélange de modèles discriminatifs. In *Gretsi* (p. 1–4).

Bibliographie

Achanta, R., Shaji, A., & Smith, K. (2012). SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(11), 2274–2281.

Arlot, S. & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Audebert, N., Le Saux, B., & Lefèvre, S. (2016). HOW USEFUL IS REGION-BASED CLASSIFICATION OF REMOTE SENSING IMAGES IN A DEEP LEARNING FRAMEWORK ? In *Ieee international geoscience and remote sensing symposium*.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An Empirical Distribution Function for Sampling with Incomplete Information. *The Annals of Mathematical Statistics*, *26*(4), 641–647.

Bar, M. (2004, août). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–29.

Bar, M. (2009). The proactive brain : memory for predictions. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *364*(1521), 1235–43.

Belgiu, M. & Drăguţ, L. (2016). Random forest in remote sensing : A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31.

Bergstra, J. [J], Yamins, D., & Cox, D. (2013). Making a science of model search : Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*.

Bergstra, J. [James], Bardenet, R., Bengio, Y., & Kegl, B. (2011). Algorithms for Hyper-Parameter Optimization. In *Advances in neural information processing systems* (p. 2546–2554).

Biederman, I. (1972). Perceiving real-world scenes. *Science*.

Blanchart, P., Ferecatu, M., & Datcu, M. (2011). CASCADED ACTIVE LEARNING FOR OBJECT RETRIEVAL USING MULTISCALE COARSE TO FINE ANALYSIS. In *Ieee international conference on image processing*.

Blaschke, T., Hay, G. J., Kelly, M., Lang, S., Hofmann, P., Addink, E., ... Tiede, D. (2014). Geographic Object-Based Image Analysis - Towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing*, *87*, 180–191.

Bordes, A., Bottou, L., & Gallinari, P. (2009). SGD-QN : Careful Quasi-Newton Stochastic Gradient Descent. *The Journal of Machine Learning Research*, *10*, 1737–1754.

Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. In *International conference on computational statistics*.

Boyd, K., Eng, K. H., & Page, C. D. (2013). Area under the precision-recall curve : Point estimates and confidence intervals.

Bush, V. (1945). Think As We May. *The atlantic monthly*.

Camps-Valls, G., Tuia, D., Bruzzone, L., & Benediktsson, J. A. (2014). Advances in hyperspectral image classification : Earth monitoring with statistical learning methods. *IEEE Signal Processing Magazine*, *31*(1), 45–54.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *8*(6), 679–698.

Cao, G., Yang, X., & Mao, Z. (2005). A two-stage level set evolution scheme for man-made objects detection in aerial images. In *Ieee conference on computer vision and pattern recognition* (T. 1, p. 474–479).

Chapelle, O. (2006). Training a Support Vector Machine in the Primal. *Neural Computation*, *1*(1), 1–22.

Choi, M. J., Torralba, A., & Willsky, A. S. (2012, février). A tree-based context model for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(2), 240–52.

Cinbis, R. & Sclaroff, S. (2012). Contextual object detection using set-based classification.

Crawford, M. M., Tuia, D., & Yang, H. L. (2013). Active learning : Any value for classification of remotely sensed data ? *Proceedings of the IEEE*, *101*(3), 593–608.

Dalal, N. & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *Ieee conference on computer vision and pattern recognition*.

Datcu, M., Daschiel, H., Pelizzari, A., Quartulli, M., Galoppo, A., Colapicchioni, A., ... D'Elia, S. (2003, décembre). Information mining in remote sensing image archives : system concepts. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(12), 2923–2936.

Davis, J. & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *International conference on machine learning* (p. 233–240). New York, New York, USA : ACM Press.

Dempster, A. A., Laird, N. N., & Rubin, D. D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*, *39*(1), 1–38.

Derrode, S. & Ghorbel, F. (2001). Robust and Efficient Fourier–Mellin Transform Approximations for Gray-Level Image Reconstruction and Complete Invariant Description. *Computer Vision and Image Understanding*, *83*(1), 57–78.

Divvala, S., Hoiem, D., Hays, J., Efros, A., & Hebert, M. (2009, juin). An empirical study of context in object detection. In *Ieee conference on computer vision and pattern recognition* (p. 1271–1278). Ieee.

Everingham, M., Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010, septembre). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010, septembre). Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1627–1645.

Felzenszwalb, P. F. & Huttenlocher, D. P. (2012). Distance Transforms of Sampled Functions. *Theory of Computing*, *8*(1), 415–428.

Flusser, J. (2000). On the independence of rotation moment invariants. *Pattern Recognition*, *33*(9), 1405–1410.

Galleguillos, C. & Belongie, S. (2010, juin). Context based object categorization : A critical survey. *Computer Vision and Image Understanding*, *114*(6), 712–722.

Gerhardinger, a., Ehrlich, D., & Pesaresi, M. (2005). Vehicles detection from very high resolution satellite imagery. *International Archives of Photogrammetry and Remote Sensing*, *36*(3), 83–88.

Gerke, M. (2015). *Use of the Stair Vision Library within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen)* (rapp. tech. N° 2000).

Girshick, R. (2012). *From Rigid Templates To Grammars : Object Detection With* (thèse de doct.).

Gomez-Chova, L., Tuia, D., Moser, G., & Camps-Valls, G. (2015). Multimodal Classification of Remote Sensing Images : A Review and Future Directions. *Proceedings of the IEEE*, *103*(9), 1560–1584.

Grabner, H., Nguyen, T. T., Gruber, B., & Bischof, H. (2008). On-line boosting-based car detection from aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *63*(3), 382–396.

Gu, C., Arbeláez, P., Lin, Y., Yu, K., & Malik, J. (2012). Multi-component models for object detection.

Haboudane, D. (2004, avril). Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies : Modeling and validation in the context of precision agriculture. *Remote Sensing of Environment*, *90*(3), 337–352.

Haralick, R., Shanmugan, K., & Dinstein, I. (1973). Textural features for image classification.

Hay, G., Niemann, K., & McLean, G. (1996). An object-specific image-texture analysis of H-resolution forest imagery. *Remote Sensing of Environment*, *55*(2), 108–122.

Hoiem, D., Chodpathumwan, Y., & Dai, Q. (2012). Diagnosing error in object detectors.

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6683 LNCS*, 507–523.

Inglada, J. (2007, août). Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. *ISPRS Journal of Photogrammetry and Remote Sensing*, *62*(3), 236–248.

Jaakkola, T. & Haussler, D. (1999). Exploiting generative models in discriminative classifiers. In *Advances in neural information processing systems*.

Jones, D. R. (2001). A Taxonomy of Global Optimization Methods Based on Response Surfaces. *Journal of Global Optimization*, *21*, 345–383.

Kappes, J. H., Andres, B., Hamprecht, F. A., Schnörr, C., Nowozin, S., Batra, D., ... Rother, C. (p.d.). A Comparative Study of Modern Inference Techniques for Discrete Energy Minimization Problems. In *Ieee conference on computer vision and pattern recognition*.

Kohonen, T., Schroeder, M. R., & Huang, T. S. (2001, janvier). Self-Organizing Maps. *Proceedings of the IEEE*.

Kolmogorov, V. & Rother, C. (2007). Minimizing nonsubmodular functions with graph cuts - A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Krähenbühl, P. & Koltun, V. (2012). Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems* (p. 1–9).

Krizhevsky, A., Sulskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in neural information processing systems* (p. 1–9).

Lacoste-julien, S., Jaggi, M., Schmidt, M., & Pletscher, P. (2013). Block-Coordinate Frank-Wolfe Optimization for Structural SVMs. In *International conference on machine learning* (T. 28).

Le Saux, B. (2014). Interactive design of object classifiers in remote sensing. In *International conference on pattern recognition* (p. 2572–2577).

Lewis, J. P. (1995). Fast Normalized Cross-Correlation. *Vision Interface*, *1995*(1), 1–7.

Lorette, A., Descombes, X., & Zerubia, J. (2000). Texture Analysis through a Markovian Modelling and Fuzzy Classification : Application to Urban Area Extraction from Satellite Images. *International Journal of Computer Vision*, *36*(3), 221–236.

Lowe, D. G. (2004, novembre). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Lu, D. & Weng, Q. [Q.]. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, *28*(5), 823–870.

Malisiewicz, T. & Efros, A. (2009). Beyond categories : The visual memex model for reasoning about object relationships. In *Advances in neural information processing systems* (p. 1–9).

Malisiewicz, T., Gupta, A., & Efros, A. a. (2011, novembre). Ensemble of exemplar-SVMs for object detection and beyond. In *Ieee international conference on computer vision* (p. 89–96). Ieee.

Marceau, D. D. J., Howarth, P. J. P., Dubois, J. J.-m. M., & Gratton, D. D. J. (1990). Evaluation Of The Grey-level Co-occurrence Matrix Method For Land-cover Classification Using Spot Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *28*(4), 513–519.

Martins, A. F. T., Pt, A., Figueiredo, M. A. T., Aguiar, P. M. Q., Smith, N. A., & Xing, E. P. (2012). AD 3 : Alternating Directions Dual Decomposition for MAP Inference in Graphical Models *. *The Journal of Machine Learning Research*.

Melgani, F. & Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, *42*(8), 1778–1790.

Michel, J., Grizonnet, M., Malik, J., Bricier, A., Lahlou, O., & France, T. C. (2011). Local feature based supervised object detection : Sampling, Learning and Detection. *IEEE Geoscience and Remote Sensing Letters*, 2381–2384.

Mnih, V. & Hinton, G. E. (2010). Learning to detect roads in high-resolution aerial images.

Molinier, M., Laaksonen, J., Member, S., & Häme, T. (2007). Detecting Man-Made Structures and Changes in Satellite Imagery With a Content-Based Information Retrieval System Built on Self-Organizing Maps. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(4), 861–874.

Montoya-Zegarra, J. A., Wegner, J. D., Ladický, L., & Schindler, K. (2015). Semantic Segmentation of Aerial Images in Urban Areas With Class-Specific Higher-Order Cliques. *ISPRS Journal of Photogrammetry and Remote Sensing*, *II-3/W4*(March), 127–133.

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing : A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247–259.

Müller, A. C. & Behnke, S. (2014). PyStruct -Learning Structured Prediction in Python. *The Journal of Machine Learning Research*, *15*, 2055–2060.

Murphy, K., Torralba, A., & Freeman, W. (2003). Using the forest to see the trees : a graphical model relating features, objects and scenes. In *Advances in neural information processing systems* (T. 53, *3*, p. 107–114).

Neubeck, A. & Van Gool, L. (2006). Efficient Non-Maximum Suppression. In *International conference on pattern recognition* (T. 3, *1*, p. 850–855).

Nicolas, J. M. (2012). *Les Bases de l'Imagerie Satellitaire*.

Nowozin, S. (2010). Structured Learning and Prediction in Computer Vision. *Foundations and Trends in Computer Graphics and Vision*, *6*(3-4), 185–365.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002, juillet). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987.

Oliva, A. & Torralba, A. [Antonio]. (2001). Modeling the shape of the scene : A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

2015 IEEE GRSS Data Fusion Contest. (p.d.). http ://www.grss-ieee.org/community/technical-committees/data-fusion/data-fusion-contest/.

NZAM : New Zealand Aerial Mapping Limited, Christchurch after earthquake on 22 February 2011. (p.d.). http ://nzam.com.

Paisitkriangkrai, S., Sherrah, J., Janney, P., & Van-Den Hengel, A. (2015). Effective semantic pixel labelling with convolutional networks and Conditional Random Fields. In *Ieee conference on computer vision and pattern recognition* (T. 2015-Octob, p. 36–43).

Palmer, T. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*(5), 519–526.

Parzen, E. (1962). On estimation of a probability density function and mode.

Penatti, O. A. B., Silva, F. B., Valle, E., Gouet-Brunet, V., & Torres, R. D. S. (2014). Visual word spatial arrangement for image retrieval and classification. *Pattern Recognition*.

Perronnin, F., Sánchez, J., & Mensink, T. (2010). Improving the fisher kernel for large-scale image classification.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*.

Porway, J., Wang, Q., & Zhu, S. C. (2010, novembre). A Hierarchical and Contextual Model for Aerial Image Parsing. *International Journal of Computer Vision*, *88*(2), 254–283.

Quang, N. T., Thuy, N. T., Sang, D. V., & Binh, H. T. T. (2015). An Efficient Framework for Pixel-wise Building Segmentation from Aerial Images. In *Proceedings of the sixth international symposium on information and communication technology* (p. 282–287). SoICT 2015. New York, NY, USA : ACM.

Randen, T. & Husoy, J. (1999, avril). Filtering for texture classification : a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(4), 291–310.

Ratliff, N. D., Bagnell, J. A., & Zinkevich, M. a. (2007). (Online) Subgradient Methods for Structured Prediction. *Artificial Intelligence and Statistics*.

Rosch, E. (2013). Principles of Categorization. *Readings in Cognitive Science : A Perspective from Psychology and Artificial Intelligence*, 312–322.

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., & Breitkopf, U. (2012). the Isprs Benchmark on Urban Object Classification and 3D Building Reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, *I-3*(September), 293–298.

Sanchez, J., Perronnin, F., Mensink, T., Verbeek, J., Classification, I., Jorge, S., ... Jakob, M. (2013). Image Classification with the Fisher Vector : Theory and Practice. *International Journal of Computer Vision*.

Schapire, R. E. (2002). The Boosting Approach to Machine Learning An Overview. *MSRI Workshop on Nonlinear Estimation and Classification*.

Schistad Solberg, A. H., Taxt, T., & Jain, A. K. (1996). A Markov Random Field Model for Classification of Multisource Satellite Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *34*(1).

Selvage, J., Chenoweth, D., & Cooper, B. (1994, mars). Fractal error for detecting man-made features in aerial images. *Electronics Letters*, *30*(7), 554–555.

Shalev-Shwartz, S., Singer, Y., Srebro, N., & Cotter, A. (2010, octobre). Pegasos : primal estimated sub-gradient solver for SVM. *Mathematical Programming*, *127*(1), 3–30.

Sirmaçek, B. & Ünsalan, C. (2009). Urban-Area and Building Detection Using SIFT Keypoints and Graph Theory. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(4), 1156–1167.

Snoek, J., Larochelle, H., & Adams, R. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in neural information processing systems* (p. 1–9).

Sun, H., Sun, X., Wang, H., Li, Y., & Li, X. (2012). Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. *IEEE Geoscience and Remote Sensing Letters*, *9*(1), 109–113.

Taskar, B., Guestrin, C., & Koller, D. (2003). Max-margin Markov networks. In *Advances in neural information processing systems* (p. 25–32).

Torralba, a., Murphy, K., Freeman, W., & Rubin, M. (2003). Context-based vision system for place and object recognition. In *Ieee international conference on computer vision* (T. 1, p. 273–280).

Torralba, A. [A.], Oliva, A., & Freeman, W. T. (2010). Object recognition by scene alignment. *Journal of Vision*, *3*, 196–196.

Torralba, A. [Antonio]. (2003). Contextual priming for object detection. *International Journal of Computer Vision*, *53*(2), 169–191.

Unsalan, C. & Boyer, K. (2004, décembre). Classifying land development in high-resolution Satellite imagery using hybrid structural-multispectral features. *IEEE Transactions on Geoscience and Remote Sensing*, *42*(12), 2840–2850.

Viola, P. & Jones, M. J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision*, *57*(2), 137–154.

Volpi, M. & Ferrari, V. (2015). Structured prediction for urban scene semantic segmentation with geographic context. In *Joint urban remote sensing event*.

Weng, Q. [Qiao]. (2009). *Remote Sensing and GIS Integration : Theories, Methods, and Applications : Theory, Methods, and Applications*. McGraw-Hill, New-York.

Wittgenstein, L. (1953). *Philosophical investigations*. Blackwell Oxford.

Zadrozny, B. & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *International conference on knowledge discovery and data mining* (p. 694). New York, USA : ACM Press.

Zhang, L., Zhang, L., & Kumar, V. (2016). Deep learning for Remote Sensing Data. *IEEE Geoscience and Remote Sensing Magazine*, (june), 18.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, *32*, 56–85.

Résumé. — Ce travail concerne l'interprétation du contenu des images aériennes optiques panchromatiques très haute résolution. Deux méthodes pour la classification du contenu de ces images ont été développées. Une méthode basée sur la détection des instances des différentes catégories d'objets et une autre méthode basée sur la segmentation sémantique de l'image utilisant un modèle des relations de contexte entre les superpixels extraits de l'image. La méthode de détection des objets dans une image très haute résolution est basée sur l'apprentissage d'un mélange de modèles d'apparence d'une catégorie d'objets à détecter puis de la fusion des hypothèses renvoyées par le mélange. Nous proposons une méthode de partitionnement des exemples d'apprentissage de la base en sous-catégories visuelles basée sur une procédure en deux étapes qui utilise les métadonnées des exemples et leurs apparences. Cette phase de partitionnement permet d'apprendre des modèles d'apparence où chaque modèle est spécialisé dans la reconnaissance d'une sous-catégorie visuelle de la base et dont la fusion permet de généraliser les détections à l'ensemble de la classe sémantique. Les performances du détecteur ainsi obtenues sont évaluées sur plusieurs bases d'images aériennes très haute résolution à des différentes résolutions et en plusieurs endroits du monde. La méthode de segmentation sémantique contextuelle développée utilise une combinaison des descriptions visuelles des superpixels extraits d'une image et des informations de contexte extraits entre les superpixel. La représentation du contexte entre les superpixels est obtenu en utilisant une représentation par modèle graphique entre les superpixels voisins. Les noeuds du graphes correspondant à la représentation visuelle d'un superpixel et les arêtes la représentation contextuelle entre deux voisins. Enfin nous présentons une méthode de prédiction de la catégorie d'un superpixel en fonction des décisions données par les voisins pour rendre les prédictions plus robustes. La méthode a été testé sur une base d'image aérienne très haute résolution.

Mots-clés : Imagerie aérienne ; Détection d'objets ; Segmentation sémantique ; Sous-catégorie visuelle ; Mélange de modèles ; Modèle de contexte ; Modèle graphique

Abstract. — This work is about interpretation of the content of very high resolution aerial optical panchromatic images. Two methods are proposed for the classification of this kind of images. The first method aims at detecting the instances of a class of objects and the other method aims at segmenting superpixels extracted from the images using a contextual model of the relations between the superpixels. The object detection method in very high resolution images uses a mixture of appearance models of a class of objects then fuses the hypothesis returned from each model. We develop a method that clusters training samples into visual subcategories based on a two stages procedure using metadata and visual information. The clustering part allows to learn models that are specialised in recognizing a visual subcategory and whose fusion allow to generalize the detection to the whole semantic category. The performances of the method are evaluate on several dataset of very high resolution images at several resolutions and several places. The method proposed for contextual semantic segmentation use a combination of visual description of a superpixel extract from the image and contextual information gathered between superpixels. The contextual representation is based on a graph where the nodes are the superpixels and the edges are the relations between two neighbors. Finally we predict the category of a superpixel using the predictions made by of the neighbors using the contextual model in order to make the prediction more reliable. We test our method on a dataset of very high resolution images.

Keywords: Aerial Images; Object detection; Semantic segmentation; Visual subcategories; Mixture of models; Contextual model; Structural SVM

Office National d'Études et de Recherches Aérospatiales
Chemin de la Hunière et des Joncherettes – BP 80100
91123 PALAISEAU Cedex
Tél. : +33 1 80 38 60 60 - Fax : +33 1 80 38 65 10
http://www.onera.fr