Analyse automatique d'images, de l'expert à l'apprenti

A Benoit

UNIVERSITÉ SAVOIE MONT BLANC

Manuscrit HDR

Pour obtenir le diplôme d'

HABILITATION à DIRIGER des RECHERCHES

de l'UNIVERSITE SAVOIE MONT BLANC

Spécialité : **STIC Traitement de l'Information**

Arrêtés ministériels : Arrêtés ministériels des 23 novembre 1988 et 13 février 1992

Présenté par

Alexandre BENOIT

Analyse automatique d'images, de l'expert à l'apprenti

HDR soutenue publiquement le 11 décembre 2019 ,
devant le jury composé de :

M. Mihai DATCU
Professeur, German Aerospace Center (DLR), Président
Mme. Jenny BENOIS-PINEAU
Professeur, Université de Bordeaux, Rapporteur
M. Frédéric JURIE
Professeur, Université Caen-Normandie, Rapporteur
M. Christophe GARCIA
Professeur, Institut National des Sciences Appliquées de Lyon, Rapporteur
M. Georges OPPENHEIM
Professeur, Université Paris Est, Membre
M. Patrick LAMBERT
Professeur, Université Savoie Mont Blanc, Membre

Table des matières

1 Introduction **4**
- 1.1 Vers l'apprentissage automatique . . . 5
- 1.2 Contributions . . . 6
- 1.3 Éléments de contexte . . . 8

2 Prétraitement d'images **11**
- 2.1 Prétraitement avec contrainte de visualisation . . . 12
  - 2.1.1 Compression de dynamique . . . 13
  - 2.1.2 Atténuation des effets de brume (« Dehazing »). . . . 17
- 2.2 Prétraitement pour l'analyse d'images . . . 20
  - 2.2.1 Filtrage spatio-temporel . . . 21
  - 2.2.2 Prétraitement de données hyperspectrales . . . 24
- 2.3 Synthèse et perspectives de recherche sur le prétraitement des données . . . 27

3 Extraction et fusion de caractéristiques **30**
- 3.1 Extraction de caractéristiques dans les images et séquences vidéo . . . 32
  - 3.1.1 Descripteurs Sacs de mots spatio-temporels . . . 33
  - 3.1.2 Vers des descripteurs obtenus par apprentissage . . . 41
- 3.2 Fusion de descripteurs . . . 43
  - 3.2.1 Fusion de descripteurs pour l'indexation sémantique . . . 44
  - 3.2.2 Fusion de descripteurs pour la modélisation de similarités entre vidéos 47
- 3.3 Synthèse et perspectives sur la description et fusion de descripteurs . . . 50

4 Apprentissage automatique et réseaux de neurones profonds **53**
- 4.1 Apprentissage . . . 55
  - 4.1.1 Apprentissage supervisé depuis peu d'exemples . . . 55
  - 4.1.2 Apprentissage semi-supervisé actif . . . 60
- 4.2 Architectures de réseaux de neurones profonds compactes . . . 65
  - 4.2.1 Réseaux de neurones compacts . . . 67
  - 4.2.2 Convolutions indexées pour l'imagerie non conventionnelle . . . 70
- 4.3 Synthèse et perspectives sur l'apprentissage et les réseaux de neurones . . . 72

5 Synthèse et Perspectives **76**

6 Liste des publications de l'auteur (post-thèse) **90**

Remerciements

Passer son Habilitation à Diriger les Recherches est un moment particulier dans la vie d'un chercheur. Il permet de s'accorder (enfin) un temps privilégié pour une prise de recul, une réflexion sur les activités passées, leur lien avec le présent et permet une projection vers les futures orientations de recherche. Je l'ai apprécié alors qu'il se positionne dans une tranche de vie déjà riche en activités de recherches, d'enseignements, de tâches administratives et bien sûr, heureusement, familiales. Le jeu entre cette émulation ambiante et l'exercice de synthèse est donc bien différent de celui que l'on vit lors de la thèse de doctorat. De la cohérence du travail d'une thèse à la cohérence de travaux de recherche associés à des co-encadrements de thèses et projets sur une dizaine d'années, la prise de recul se doit d'être plus importante. À ce titre, je ne peux que remercier tous ceux qui ont contribué dans cette aventure.

Je commence par remercier tout particulièrement Patrick Lambert pour toutes ces années de travail commun au sein du LISTIC et pour ses conseils et relectures de ce document. Nos échanges scientifiques et amicaux ont dès le début été des plus agréables et le soutien a toujours été sans failles.

Mes remerciements vont également à tous les membres du laboratoire pour tous ces échanges dans une ambiance toujours sympathique. Nous formons une équipe certes réduite, mais la complémentarité et la diversité de nos thèmes de recherche permettent une émulation permanente qui nous fait tous progresser.

J'adresse mes sincères remerciements aux membres de jury de cette Habilitation à Diriger les Recherches : Jenny Benois-Pineau, Mihai Datcu, Frédéric Jurie, Christophe Garcia, Georges Oppenheim et Patrick Lambert. Merci pour ce temps passé à rapporter et échanger sur les travaux présentés. La discussion lors de la soutenance a été riche et intéressante et a permis d'approfondir encore ma réflexion sur les orientations de recherche à venir.

Merci à tous les doctorants que j'ai pu co-encadrer, chronologiquement Tiberius Sabin Strat, Nicolas Voiron, Rizlène Raoui, Amina Ben Hamida, Mikaël Jacquemont et Emna Amri. Merci pour ces échanges scientifiques et amicaux qui nous ont permis de progresser et de faire des propositions scientifiques intéressantes.

Des remerciements particuliers à des confrères et amis comme Georges Oppenheim et Gatthas Badih pour nos échanges, leurs qualités humaines et scientifiques et leurs nombreux conseils.

Et bien sûr, mes remerciements vont à ma famille et en particulier ma femme et mes filles qui m'ont soutenu tout au long de cette aventure, et qui m'ont supporté dans tous les sens du terme dans la phase de rédaction de ce manuscrit.

Chapitre 1

Introduction

Capter, percevoir, analyser, classer, estimer, prédire, détecter, reconnaître, raisonner, agir, voici seulement quelques verbes pour décrire ce que nous, êtres vivants, réalisons chaque jour si facilement, parfois automatiquement. Notre organisme nous propose un ensemble de capacités de perception, d'analyse et d'interaction dont nous aimerions doter les machines.

Nous avons ainsi su développer ces dernières décennies des outils numériques qui nous alimentent maintenant par un flux quasi constant de données. Principalement visuelles et audio dans le monde multimédia, ce sont des « données » de toute nature en général, de la valeur signalée par un capteur de température à celle indiquée par le compteur de consommation de nos véhicules.

Pour l'ensemble de ces informations captées, maintenant massives, nous exprimons le besoin d'automatiser ces tâches déjà énoncées, si faciles, si automatiques, mais désormais devenues trop chronophages pour l'homme. Nous développons ainsi des outils de substitution que certains nommeront « IA, Intelligence Artificielle », qui se heurtent aux mêmes problèmes que le vivant. Il s'agit d'utiliser des capteurs, de prendre en compte leurs imprécisions et défauts, pour décrire ensuite les informations perçues, les organiser, les associer à d'autres pour répondre à nos besoins. Ces tâches sont difficiles. Elles sont optimisées par le vivant à l'échelle d'une vie ou de générations par les processus d'adaptation, d'apprentissage et d'évolution. Un des principaux défis est de franchir le fossé sémantique entre l'information dite de « bas niveau sémantique » que sont les données captées (couleurs, intensité lumineuse, etc.), à une information de « haut niveau sémantique » que sont nos interprétations (les objets, les actions, etc.). Nous voulons ainsi résoudre ces problèmes dans le monde numérique, en quelques années, plus vite, plus précisément... en nous inspirant parfois du vivant.

Dans ce contexte, les travaux présentés dans ce document concernent l'analyse de données numériques, principalement les images et séquences vidéo. Nous nous intéresserons à des problèmes de reconnaissance de concepts visuels, des objets aux actions, et traverserons au fil des cas d'études rencontrés, les différentes étapes de traitement, de la sortie du capteur à la prédiction.

1.1 Vers l'apprentissage automatique

Des méthodes construites par l'expert à l'apprentissage automatique de bout en bout. Nous vivons un réel changement de paradigme.

Dans le domaine de l'imagerie numérique et de la compréhension de leur contenu, une date clef, 2012, marque un événement relativement rare dans la vie d'un chercheur, celui d'un véritable virage méthodologique et technologique.
Jusqu'à cette transition et encore aujourd'hui, dans certains domaines, les différentes communautés de recherche ont développé des chaines de traitement que nous qualifierons dans ce document de méthodes « expertes », choisies, définies et paramétrées par des experts du domaine. Ces méthodes sont capables de réaliser l'analyse de données brutes, par exemple les pixels d'une image, en une cascade de quelques processus f_i bien maitrisés pour générer une prédiction y_s, par exemple, la reconnaissance d'objets. Nous pouvons illustrer et prendre comme fil conducteur dans ce document une structuration classique pour les tâches de détection, catégorisation et bien d'autres :

- $y_1 = f_1(x, \theta_1)$, un processus de prétraitement des données brutes x,
- $y_2 = f_2(y_1, \theta_2)$, l'extraction de caractéristiques permettant de répondre au problème,
- $y_s = f_3(y_2, \theta_3)$, la génération de la prédiction attendue, par exemple une catégorisation ou une régression.

Chaque processus f_i est un bloc fonctionnel bien choisi, réglé avec un nombre de paramètres θ_i petit, de l'ordre de la dizaine, et optimisé de façon séquentielle, du processus proche des données au processus de prédiction final. Une telle chaine est optimisée sur une quantité de données que nous qualifierons de « raisonnable », de l'ordre de la centaine au millier d'exemples. Les deux processus f_1 et f_2 peuvent être considérés comme des changements de variables dont l'objectif est d'introduire une invariance à certaines transformations T des données d'entrée. Cette invariance consiste à obtenir des réponses $f_i(x') \approx f_i(x)$ pour $x' = T(x)$. Le prétraitement f_1 consistera à trouver les invariants permettant d'être robuste à des groupes de transformation des données liées au capteur (sensibilité, contraste, etc.) et dans une certaine mesure au contexte d'acquisition (ambiance lumineuse, nébulosité de la scène, etc.). La recherche d'invariants liés à des groupes de transformation plus complexes liés aux objets contenus dans la scène visuelle, leur déformation, occultation, etc. sera considérée par des processus plus évolués définis dans f_2. Le processus de décision f_3 gagnera en simplicité et pertinence si les précédents ont parfaitement réduit la variabilité des données et pourra idéalement se réduire à un simple processus de décision linéaire.

En apparence à l'opposé, la tendance post 2012 est à l'apprentissage automatique dit profond ou « Deep Learning » qui a pour objectif de réaliser les mêmes tâches, mais en ne s'appuyant que sur une seule « fonction » non linéaire $y_s = f_d(x, \theta_d)$ dotée de très nombreux paramètres θ_d, de l'ordre du million. Cette fonction est une composition séquentielle et parallèle, d'opérateurs simples. La plupart sont des modèles numériques de neurones dont les paramètres sont ajustables. C'est la longueur de la séquence de ces opérateurs, en général grande, qui a amené à la dénomination de « réseaux de neurones profonds ». Si l'on reprend l'analogie avec la recherche d'invariants, f_d est une composition de changements de variables qui réduit progressivement la variabilité des données, à de multiples échelles et de façon hiérarchisée [Mal16]. L'ensemble de ses paramètres est optimisé automatiquement en confrontant le réseau à des données d'apprentissage en considérant que l'on peut extraire une forme de « régularité » dans cette masse de données. Cela permet alors à f_d de se rapprocher d'une solution optimale, une approximation du modèle réel qui lie les données et à la tâche, lui permettant de prédire correctement dans la grande majorité des cas d'usage. On parle alors de capacité à généraliser. Ce type de méthode amène donc à changer complètement de paradigme en cherchant à optimiser les millions de paramètres d'un seul processus à partir de millions d'exemples.

Ces réseaux profonds apparaissent alors comme des structures centralisées, compactes, monolithiques. On les qualifie souvent de « boites noires », n'étant pas encore capable d'en comprendre le fonctionnement exact. Décomposer le processus f_d en un ensemble de sous-processus structurés de façon similaire au découpage évoqué ci-dessus (f_1 prétraitement, f_2 description et f_3 prédiction) présente alors certains intérêts :

— Un premier est lié à l'infrastructure associée à ces processus. Classiquement centralisée et potentiellement hébergée dans des infrastructures de calcul massif (cloud et data center), cette structure monobloc peut être décomposée de façon à rapprocher certains processus près de la source de données. On parle alors de « Edge Computing ». Ceci ouvre à la possibilité de distribuer la charge de calcul et d'alléger la charge du processus centralisé. Cela rend également possible la prise de décisions locales et la réduction des coûts de transport par émission de données de plus haut niveau sémantique potentiellement plus compressibles que les données brutes. Un exemple est la voiture autonome qui doit, localement, prendre des décisions pour gérer son mouvement dans son environnement tout en échangeant avec des systèmes distants permettant d'optimiser la gestion des flottes de véhicules et de gérer le trafic dans son ensemble.

— Un second intérêt est d'ordre sociétal. Il est lié au besoin d'explication et de justification des décisions prises par les systèmes automatiques. Décomposer une chaine de traitement complexe en sous-éléments dont les entrées et sorties sont identifiées est une des stratégies permettant de répondre à ce besoin de compréhension.

— Un autre intérêt, sociétal également, est lié à la confidentialité des données. Être capable de transmettre, sur les canaux de communication, seulement l'information strictement nécessaire à la tâche permet de limiter la diffusion des données contextuelles pouvant être porteuses d'information privée. Également, l'encodage des données transmises permet une forme de sécurisation.

Dans ce contexte, mes travaux sur la période 2008-2019 se sont principalement focalisés sur l'analyse et la compréhension des images et séquences vidéo. Ils couvrent des problématiques abordées par des approches expertes et par apprentissage profond. Ces dernières pourront paraitre parfois désuètes face à leurs homologues actuelles. En revanche, cette revue de travaux permet, dans une certaine mesure, d'identifier l'intérêt de chaque méthode et leurs potentielles complémentarités pour ouvrir à de nouvelles pistes de recherche et pour répondre aux besoins énoncés ci-dessus.

1.2 Contributions

Mes travaux ont été réalisés en collaboration avec six doctorants. Trois ont soutenu (Sabin Tiberius Strat, Nicolas Voiron, Rizlène Outach Raoui) et trois sont en cours d'avancement (Amina Ben Hamida, Emna Amri et Mikaël Jacquemont) auxquels se sont ajoutés un certain nombre de stages de master. Des collaborations avec les entreprises AboutGoods Company et Total ont permis de renforcer le lien entre la recherche et les entreprises. Également, des collaborations académiques ont permis d'étendre le champ d'expertise avec notamment l'université de Sfax en Tunisie, l'université Polytechnica de Bucarest en Roumanie, l'université norvégienne de sciences et technologie (NTNU) et les collègues du consortium de laboratoires français IRIM au sein du GDR ISIS.

Les propositions concernent les différentes étapes d'une chaine de traitement, le prétraitement des données, leur description et l'apprentissage pour la prédiction.

Mes travaux seront présentés en trois parties, chacune décrivant une étape de l'analyse

de données, de la source à la prédiction comme l'illustre la figure 1.1 :

— La première est liée aux processus de prétraitement des données brutes. Elle consiste à extraire des données captées, une information pertinente renforcée nécessaire à la suite des analyses (renforcement des textures, atténuation des perturbations, etc.).

— La seconde consiste en l'extraction et la fusion de caractéristiques permettant de décrire les données. On attend à ce niveau de traitement l'obtention d'informations pertinentes (présence de textures, de parties d'objets, etc.) pour la résolution du problème et présentant une forme d'invariance à la variabilité des données (position, contexte, etc.).

— La troisième et dernière étape consiste en la synthèse d'une prédiction pour répondre à un objectif par exemple une classification (reconnaissance d'objets). Dans mes travaux, la prédiction est considérée dans un processus d'apprentissage permettant d'optimiser l'association des caractéristiques extraites à l'objectif applicatif.

Nous verrons qu'une des perspectives est l'intégration de méthodes d'apprentissage visant à optimiser des critères spécifiques à chaque étape. Cette proposition nous amènera à des considérations plus globales qui traitent de l'apprentissage multi objectif, l'apprentissage asynchrone et la justification des prédictions des systèmes automatiques.

L'ensemble des travaux présentés est lié à une problématique d'analyse de contenus visuels. Un certain nombre d'entre eux est influencé par les neurosciences et plus précisément le modèle biologique de la perception visuelle. Cette interaction a été initiée dans mes travaux de thèse et s'est poursuivie ensuite. Plus précisément, un modèle numérique de la rétine humaine développé au laboratoire GIPSA a inspiré une partie des travaux présentés. Le virage vers l'apprentissage profond est lui aussi influencé par les neurosciences et le modèle biologique. Ces thématiques sont fortement connectées et tendent à l'avenir à se rapprocher encore pour ouvrir à des perspectives intéressantes. Nous prendrons donc le temps de décrire de façon synthétique le modèle de rétine sur lequel certains travaux se sont appuyés.

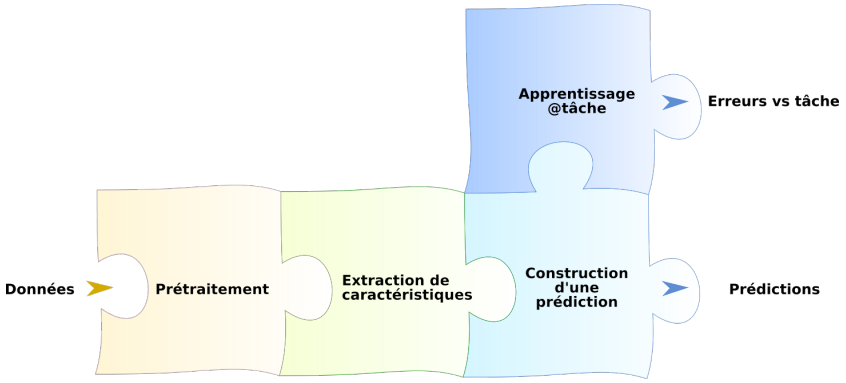


FIGURE 1.1 – Organisation typique d'une chaine de traitement selon laquelle sont développés les travaux présentés dans ce document.

1.3 Éléments de contexte

Les travaux présentés ont été développés sur ces dix dernières années en intégrant un certain nombre d'éléments de contexte que nous résumerons ici. Elle concerne l'exploitation de l'expérience acquise sur le modèle de rétine biologique, le travail collaboratif au sein du consortium de laboratoire IRIM, la transition vers l'apprentissage profond et la problématique de la ressource de calcul.

S'inspirer du vivant, un modèle numérique de rétine.

S'inspirer du vivant pour filtrer et renforcer l'information captée.

Mes travaux de thèse se sont appuyés sur un modèle numérique de la rétine humaine développé au laboratoire GIPSA [Ben07]. Finalisé dans mes premières années au laboratoire LISTIC, le modèle a été diffusé sous la forme d'un article de journal [Ben+10] et du module logiciel dédié « bioinspired » au sein de la bibliothèque de traitement d'image libre OpenCV [1]. Une partie de mes travaux sur le prétraitement de données et la synthèse de descripteurs de contenus décrits aux chapitres 2 et 3 s'appuie sur ces productions et en élargit le champ d'applications. Nous synthétiserons ici les principales propriétés de ce modèle afin d'alléger la suite du document.

Ce modèle numérique reproduit une partie connue des processus de traitement de l'information visuelle captée par les photorécepteurs et transitant entre les différentes couches neuronales au sein de la rétine. En particulier, au niveau de la couche plexiforme externe, les photorécepteurs interagissent avec les cellules dites horizontales et bipolaires et réalisent un filtrage spatio-temporel à variables non séparables dont la fonction de transfert est représentée sur la figure 1.2. Elle présente différentes propriétés essentielles à la suite des travaux qui peuvent se résumer ainsi :

— un blanchiment des fréquences spatiales à fréquence temporelle nulle. Cette propriété compense la tendance en $1/f$ du spectre spatial des images naturelles. L'effet est alors une forte atténuation de la composante continue et un renforcement des détails de l'image fixe, ou des détails statiques dans le plan de la caméra de séquences vidéo.

— une coupure des hautes fréquences spatio-temporelles typiques du bruit de capteur pour une image comme pour une séquence vidéo.

1. http://opencv.org

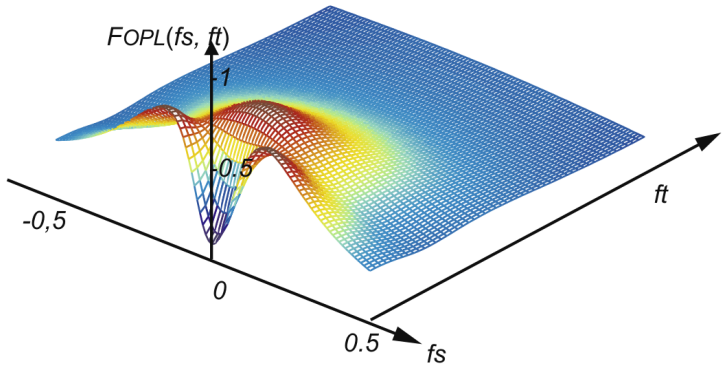


FIGURE 1.2 – Fonction de transfert de la couche plexiforme externe (F_{OPL}) du modèle de rétine. f_s dénote les fréquences spatiales et f_t les fréquences temporelles.

— une tendance filtre passe-bas spatiale pour les fréquences temporelles hautes. Outre l'effet classique de réduction de bruit, d'une façon générale, cette propriété permet de couper les hautes fréquences spatiales de l'information transitoire. Les objets en mouvement sont donc perçus de façon floue.

En sortie de ce filtre, deux canaux de sorties sont générés :

— le canal parvocellulaire. Cette sortie correspond, sur le modèle biologique, à la vision centrale. L'information de détail statique dans le plan de la caméra est renforcée. Le mouvement et le bruit spatio-temporel sont fortement atténués.

— la canal magnocellulaire. Correspondant plutôt à la vision périphérique, celui-ci est dédié à la perception de mouvement. Seuls les signaux de basses fréquences spatiales et à fréquence temporelle non nulle sont transmis et renforcés. Notons que l'énergie sur ce canal est liée de façon linéaire à la vitesse.

Nous illustrerons ce modèle au travers de ses extensions présentées dans les chapitres suivants de l'amélioration d'image à l'extraction de descripteurs spatio-temporels.

Participation collaborative à un concours international d'indexation vidéo

Un travail collaboratif au sein du consortium de laboratoires IRIM pour progresser et répondre à des problématiques d'analyse à grande échelle.

Certaines de mes contributions sur le prétraitement des données, l'extraction et la fusion de caractéristiques ont été valorisées par la participation au concours international annuel TRECVid. Ce sont les travaux principalement présentés dans les sections 2.2.1, 3.1 et 3.2.1. Nos contributions ont été intégrées à la participation commune du consortium de laboratoires français IRIM sur la période 2012 à 2017. Cette équipe, décentralisée et multi compétence, a ainsi pu rassembler jusqu'à une dizaine de laboratoires durant quelques années. Dans ce cadre collaboratif, chaque laboratoire a pu proposer l'intégration d'éléments de son domaine d'expertise afin de constituer une chaine complète de traitement de très bon niveau scientifique et technique. Sur des problèmes d'analyse de vidéo à grande échelle, l'équipe IRIM a pu être régulièrement classée parmi les 4 meilleures équipes tout en rendant possible la comparaison d'une grande variété de méthodes en différents points de sa chaine de traitement.

D'un point de vue général, la participation en équipe à ce type de concours internationaux est intéressante sur plusieurs points :

— l'enrichissement mutuel par la collaboration. Les équipes peuvent progresser et s'adapter plus rapidement sur un problème difficile et proposer des solutions innovantes.

— la diffusion des savoirs et compétences au travers de communications communes et des journées thématiques. À ce titre, nous avons pu organiser un certain nombre de journées au sein du GDR ISIS.

— la possibilité de confronter ses approches sur des cas pratiques réels et appliqués à grande échelle. Cette comparaison peut se faire à deux niveaux. D'un point de vue interne, au sein de l'équipe IRIM, les méthodes d'un certain niveau de la chaine de traitement peuvent être comparées au sein d'un système unifié. D'un point de vue externe, les propositions de l'équipe IRIM sont comparées sur le plan des performances globales avec celles de grandes équipes de recherche et industrielles internationales.

Transition vers l'apprentissage profond

Mes travaux s'inscrivent dans une période de changement de paradigme scientifique et un enrichissement des domaines applicatifs.

Mes travaux se sont progressivement orientés vers les réseaux de neurones profonds et l'apprentissage. Cette transition correspond au mouvement général observé dans le domaine de l'analyse d'image et est concomitante avec une volonté de se rapprocher d'autres activités du laboratoire. Dans ce contexte, le domaine d'application a évolué en effectuant une transition progressive de l'analyse d'image multimédia à des données plus spécifiques comme l'imagerie satellitaire et l'astrophysique qui permettent de renforcer les compétences du laboratoire et ses collaborations. L'intérêt scientifique est également de s'intéresser à l'adaptation de méthodologies issues du domaine multimédia à des données et tâches d'autres domaines. Il est intéressant dans ce cadre d'étudier les caractéristiques communes à ces différents domaines et celles qui les rendent spécifiques. Les approches par transfert de représentations actuellement valorisées par l'apprentissage profond montrent les surprenantes capacités de discrimination des caractéristiques apprises sur un domaine et appliquées sur d'autres. Il manque cependant encore une formalisation de ces propriétés et il apparait comme nécessaire d'avancer sur ce point.
La thèse d'Amina Ben Hamida et plus récemment celles d'Emna Amri et de Mikael Jacquemont permettent de s'intéresser à ces questions et ouvrent maintenant à de nouvelles collaborations, du monde industriel avec Total à la physique avec le Laboratoire d'Annecy de Physique des Particules (LAPP).

Gestion des moyens et plateforme de calculs

La recherche de moyens de calculs a permis de faire évoluer le centre de calcul de l'université.

L'ensemble des travaux présentés est dépendant d'une puissance de calcul suffisante et des moyens humains associés. Un exemple typique est la participation à des compétitions internationales comme TRECVid pour laquelle une petite équipe locale de recherche doit déployer et évaluer ses contributions scientifiques sur un problème à grande échelle. La recherche de moyens pour mettre en production les propositions scientifiques sur des systèmes de calculs suffisants pour respecter les temps imposés par le concours exige alors un certain niveau de technicité et d'organisation des travaux.

C'est dans ce cadre que le laboratoire a pu renforcer son implication dans la ferme de calcul MUST de l'université. Cette plateforme, intégrée à la grille de calcul exploitant les données scientifiques issues du Large Hadron Collider (LHC), est initialement dédiée à la physique. MUST est à ce jour une grille de calcul mutualisée qui compte quelques milliers de noeuds CPU. Au travers de projets et avec le soutien de l'équipe gérant la plateforme, nous avons tout d'abord pu adapter son usage dans le but de l'exploiter dans le cadre des compétitions annuelles TRECVid. Avec l'arrivée de l'apprentissage profond, j'ai ensuite été à l'initiative de projets qui ont amené la plateforme à évoluer vers une hybridation des noeuds de calcul. La plateforme est désormais capable de coordonner des tâches appliquées sur les processeurs classiques (CPU) ainsi que d'autres tâches réalisées sur des processeurs graphiques (GPU). Ces processeurs graphiques capables de réaliser des calculs massivement parallèles, très adaptés à l'apprentissage profond sont maintenant intégrés aux services proposés par l'infrastructure. Ce type de grille hybride est relativement innovant dans l'infrastructure du LHC et ouvre à des travaux plus spécialisés sur le calcul distribué qui pourront être lancés avec d'autres forces du laboratoire.

Chapitre 2

Prétraitement d'images

Mes contributions couvrent l'usage de modèles de réseaux de neurones « experts, issus des neurosciences » et s'ouvrent maintenant à des modèles « optimisés par apprentissage ». Les travaux s'intéressent au prétraitement pour l'optimisation des signaux avec et sans contrainte de visualisation.

Capter, percevoir, les premières étapes d'une analyse commencent par une phase d'acquisition d'information. Cependant, la donnée captée se présente en général sous une forme brute et est sujette à différentes perturbations systématiques ou transitoires pouvant impacter la tâche à réaliser. Ces perturbations peuvent être liées à la variabilité introduite par les conditions d'acquisition (un paysage ensoleillé, dans la brume, au soleil couchant). Les données peuvent également être sensibles aux caractéristiques des capteurs (sensibilité, bruits). Une approche classique est alors de définir un processus de prétraitement. Il a pour objectif de réduire la variabilité des données causée par ces perturbations, voire de renforcer l'information pertinente nécessaire à la suite du processus. Cette tâche présente cependant un aspect générique qui n'est pas dédié à une tâche ou à un scénario spécifiques. Nos yeux et leur rétine, par exemple, renforcent l'information pertinente nécessaire à réaliser une grande partie de nos activités quotidiennes dans une grande variété de situations et sont en cela génériques sur cet ensemble d'activités.
Sur ces dernières décennies, de nombreux travaux de recherche se sont intéressés à la tâche de prétraitement. Leur définition dépend de l'état de nos connaissances sur les données à capter, les capteurs mis en jeux et les éventuelles contraintes de contrôle que l'on doit imposer. Nous pouvons distinguer deux approches :

- les approches « expertes » : si l'on dispose de suffisamment d'aprioris sur la nature des données, leur variabilité, les types de perturbations auxquelles se confronter, etc. Il est alors possible de définir un prétraitement adapté. Ce seront des opérations de filtrage ou des transformations bien choisies que je nommerai approches « d'expertes » dans la suite de ce chapitre. L'avantage est alors que la méthode de prétraitement choisie sera suffisamment contrainte pour n'être contrôlée que par quelques paramètres dont l'impact sur les traitements sera généralement maitrisé. Un écueil potentiel est cependant le biais introduit par le choix des prétraitements sur l'efficacité globale du système. S'ils ne permettent d'optimiser que la partie connue de la relation entre les données et l'objectif, alors il est possible que des facteurs non connus importants impactent les performances.
- les méthodes s'appuyant sur un apprentissage pour découvrir un prétraitement adapté. Dans un contexte plus général, la relation entre une tâche à réaliser et les données sur lesquelles s'appuyer n'est pas parfaitement connue. Il en va de même pour le prétraitement. Quelle information faut-il filtrer ou renforcer ? Quelques aprioris peuvent exister, mais ne permettent pas de définir avec précision les prétraitements à appliquer sur les données. On peut alors définir un prétraitement capable de

s'adapter par un processus d'optimisation. Les réseaux de neurones profonds entrent dans ce cadre. Ce type de prétraitement possède alors un plus grand nombre de degrés de liberté qui pourront être contraints ou optimisés en introduisant des critères spécifiques au prétraitement, par exemple l'amélioration de la qualité des images.

Nous verrons dans ce chapitre plusieurs contextes applicatifs de prétraitement de l'information visuelle auxquels j'ai pu m'intéresser. Nous distinguerons, dans la section 2.1, des méthodes de prétraitement introduisant une contrainte visant à optimiser la qualité perçue des images en sortie de ce processus. Dans la section 2.2, nous nous focaliserons sur des prétraitements visant à alimenter de façon pertinente des chaines d'analyse d'image, par filtrage et renforcement de l'information brute.

Au travers de ces travaux, nous verrons les avantages et écueils potentiels des approches de prétraitement « expertes » ou « entrainables ». Ces travaux m'ont convaincu que la tâche de prétraitement des données reste un problème spécifique qu'il faut considérer en complément des objectifs principaux. Le traiter seul sans prise en compte suffisante des tâches en aval limitera potentiellement les performances de ces objectifs finaux. Également, comme on le retrouve systématiquement dans les approches par apprentissage profond, intégrer directement le prétraitement au sein de la chaine de traitement sans introduire de critères spécifiques réduit fortement la maîtrise et la compréhension globales des processus réalisés.

Une approche intéressante, somme toute déjà proposée dans les modèles biologiques, consiste à voir le prétraitement comme un processus dont la structure peut être prédéfinie en amont, mais qui possède une grande flexibilité d'adaptation ajustable par apprentissage. Comme décrit dans [MWK16], cet apprentissage peut alors être modulé par des tâches de haut niveau (se déplacer, reconnaitre, etc.) comme par des tâches de plus bas niveau (compenser les défauts du capteur, leur dérive dans le temps, extraire et adapter dans le temps des représentations).

2.1 Prétraitement avec contrainte de visualisation

Nous considérons ici les méthodes de traitement d'images visant à améliorer la qualité perçue.

L'amélioration d'image est un domaine applicatif classique du prétraitement qui vise à améliorer l'aspect visuel des acquisitions pour l'utilisateur. C'est une tâche que l'on qualifie souvent de bas niveau sémantique, car elle est plus liée au traitement de l'information au niveau des pixels et est, dans une moindre mesure, dépendante du contenu de haut niveau sémantique qui est capturé (objets, scène, etc.). Le prétraitement d'image peut ainsi introduire des critères spécifiques ayant pour but de maximiser la qualité visuelle, la perception de rendu naturel de la scène et d'autres critères subjectifs liés à la perception d'observateurs (humains). L'introduction de contraintes de visualisation permet donc de générer des produits exploitables aussi bien par des chaines d'analyse que par des opérateurs humains. Un point intéressant réside dans le fait que l'opérateur peut alors mieux comprendre les informations traitées par les processus en aval. Ceci ouvre ainsi à l'explication et à la justification de leurs réponses par une connaissance sur les propriétés de leurs données d'entrée. De nombreux types de prétraitement sont proposés dans la littérature et peuvent être réalisés conjointement ou en cascade. On trouve ceux en sortie de capteur visant à convertir ses signaux bruts en une information adaptée à un support. Dans ce cadre, la correction de dynamique, le filtrage du bruit de captation, le dématricage (« demosaicing ») sont classiques et souvent dépendants d'une connaissance fine du capteur.

D'autres filtrages plus spécifiques sont parfois requis pour renforcer une information jugée pertinente. Ces traitements peuvent être intégrés à la phase d'acquisition ou appliqués

en aval. Je me suis intéressé à deux cas :

— une méthode experte, décrite dans la section 2.1.1. Elle concerne la compression de dynamique pour l'adaptation d'images à grande plage dynamique (High Dynamic Range, HDR) vers des images à dynamique réduite classique (Standard Dynamic Range, SDR), affichées sur nos écrans (TV, smartphone, etc.).

— une méthode par apprentissage, décrite dans la section 2.1.2. qui vise à réduire les effets de brume. La brume est un phénomène physique qui impacte la transmission de la lumière et réduit la visibilité des textures et détails des objets. La réduction de ses effets est particulièrement intéressante pour des applications pouvant impacter la sécurité des biens et des personnes comme la conduite autonome.

2.1.1 Compression de dynamique

Nous nous intéressons au problème de l'adaptation de la dynamique d'une image au support de diffusion tout en maintenant la perception d'image naturelle.

Sur la période 2007-2008, au sein du projet Futur'Image issu d'un partenariat entre le laboratoire IRCCyN, Orange et Technicolor, je me suis intéressé à l'adaptation d'images à grande plage dynamique (HDR) pour un rendu sur des écrans à plage dynamique différente, plus réduite, comme nos écrans d'ordinateur (SDR).

La gamme de luminance naturelle que nous sommes en mesure de percevoir peut évoluer d'une ambiance fortement ensoleillée à celle d'un clair de lune soit un rapport de 1/1e9 ou 90dB. Notre système sensoriel a su développer des stratégies d'adaptation à cette grande plage de luminance et notre besoin est d'avoir à disposition des outils de captation et de restitution dotés de capacités similaires. Les outils d'imagerie grand public, capteur et écran, utilisent classiquement un codage sur 8bits par canal couleur soit 256 niveaux différents pour représenter l'information. Ce type d'image est dit à faible plage dynamique (SDR), car il ne peut capter ou restituer l'ensemble de l'information qui pourrait être perçue directement de nos yeux. Des appareils et algorithmes plus spécialisés sont maintenant capables de créer des images à grande plage dynamique. En revanche, les écrans de restitution présentent encore aujourd'hui une dynamique plus réduite. Un traitement intermédiaire d'adaptation de dynamique est alors nécessaire. On parle d'opérateur de mappage de ton local ou « tone mapping ». Dans le cas de la compression de dynamique, l'objectif est d'obtenir, sans effets de sur ou sous exposition, l'affichage de l'ensemble des détails d'une image à grande dynamique sur un écran à faible dynamique.

Les méthodes de la littérature s'appuient sur les neurosciences à des degrés de précision variés, de la compression des forts gradients [FLW02] à des modèles plus aboutis intégrant les modèles de sensibilités aux contrastes du système visuel [MS08]. Cette tâche de compression de dynamique est principalement liée à l'information de luminance et s'expose alors aux risques de distorsion suivants observables sur la figure 2.1 :

— sur l'information couleur, des phénomènes de saturation introduisent une perception d'image non naturelle. On trouve, selon les algorithmes, des saturations sur des motifs locaux, sous la forme de halos colorés ou de façon ponctuelle par amplification du bruit lié à la couleur.

— sur l'information de luminance. Compresser la dynamique de façon linéaire a pour effet d'atténuer la perception des détails de l'image, autrement dit l'énergie des fréquences spatiales portant l'information de texture. Il est alors intéressant de renforcer ou maintenir cette énergie tout en compressant la dynamique globale. Les méthodes visant à répondre à ces deux objectifs font alors face au risque d'introduction de halos très visibles autour des objets texturés qui réduisent la perception d'image naturelle.

La rétine comme un outil naturel de compression de dynamique

Comment compresser la dynamique des images et vidéo en s'inspirant de la vision biologique ? Contribution de postdoctorat.

Dans le cadre du projet Futur'Image, je me suis intéressé à un ensemble de propriétés intéressantes de la voie parvocellulaire du modèle de rétine étudié en thèse. Ce modèle a été complété pour répondre à cette problématique. L'objectif était de proposer un opérateur capable de préserver la constance des couleurs, de réduire les effets de halo et également de pouvoir traiter les contenus vidéo. Le modèle proposé est illustré sur la figure 2.2. Ce modèle s'appuie sur le modèle de rétine présentée au paragraphe 1.3. Le modèle ne traitant initialement qu'une information monochromatique est appliqué sur des images couleur en l'appliquant après un étage de multiplexage des plans chromatiques. Cela permet de corriger la dynamique de luminance tout en conservant l'information colorimétrique. Cette dernière est recouvrée en fin de processus par l'application d'une opération de démultiplexage. Tout comme le modèle biologique, sa modélisation gagne à s'appuyer sur l'échantillonnage couleur réalisé par le capteur. Le multiplexage simule donc cet échantillonnage sur des images classiques. Les propriétés du système sont alors les suivantes :

— du point de vue de la compression de la luminance, les processus d'adaptation locaux de la dynamique au niveau des photorécepteurs et des cellules ganglionnaires de ce canal permettent d'adapter *localement* la dynamique tout en préservant les contrastes dans les zones claires et sombres. Le filtrage spatio-temporel intermédiaire du modèle de rétine réalise le blanchiment spectral. Un premier avantage est l'atténuation globale de la luminance moyenne de l'image qui applique alors une compression globale de la dynamique. Le second gain est relatif au renforcement des fréquences spatiales des détails et l'élimination du bruit haute fréquences. Le dernier point est important, car le bruit couleur est alors fortement limité. Également, les effets de halo dans les zones contrastées peuvent être modulés jusqu'à leur annulation. L'image compressée conserve ainsi un aspect naturel.

— l'information couleur s'inspire également du modèle biologique. Comme illustré sur la figure 2.2, l'échantillonnage couleur réalisé par les photorécepteurs tout comme les capteurs numériques est modélisé par une opération de multiplexage des canaux couleurs (un échantillonnage de Bayer par exemple) avant l'injection des données dans le modèle de rétine. L'information est finalement démultiplexée en sortie de rétine. Cette stratégie montre à quel point ce prétraitement peut se rapprocher de la sortie du capteur numérique.

— le prétraitement des vidéos est naturellement intégré au modèle par le filtrage spatio-temporel du modèle. Par sa nature de filtre passe-bas temporel, le modèle lisse les instabilités temporelles de la luminance comme montré sur la figure 2.3.

.

Résultats et limites

Un modèle « expert » de rétine spatio-temporelle efficace et générique. Des systèmes d'acquisition et métriques de performances en maturation.

Ces travaux ont été diffusés dans les articles [Ben+09 ; Ben+10]. L'intérêt de la méthode est le recours très limité à un ajustement des paramètres du modèle de rétine pour une compression satisfaisante sur un large panel de scènes visuelles. Une limite observée est classique des « systèmes génériques ». Il s'agit du besoin d'optimisation des hyper paramètres pour atteindre un rendu optimal pour une configuration donnée, pour un capteur et un affichage spécifique ou pour des conditions d'observation spécifiques.

Du point de vue de l'évaluation des performances du modèle, les métriques fiables d'évaluation en termes de qualité perçue par l'observateur humain ne sont apparues, pour ces méthodes, que quelques années plus tard. Une collaboration entre le laboratoire IRCCyN et Rafal Mantiuk à l'université de Bangor a notamment permis de proposer la méthode HDR-

FIGURE 2.1 – À gauche, une image HDR compressée par une correction gamma (2.6) ; au centre une compression par le modèle [MS08] ; à droite, compression obtenue avec l'approche proposée [Ben+09]. Par une analyse fine de cette figure (zoom et bonne qualité d'affichage ou d'impression), nous pouvons observer des effets de halos, de saturation et bruits couleur plus prononcés sur l'image centrale.

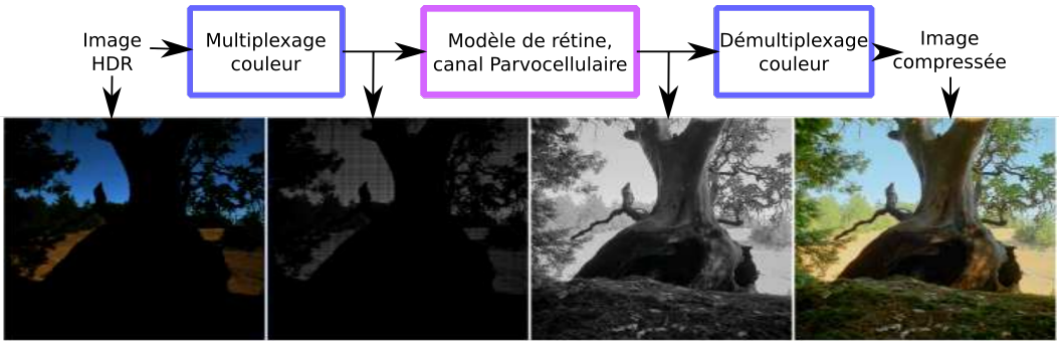


FIGURE 2.2 – Illustration du processus de compression de dynamique sur une image HDR couleur. De gauche à droite, l'image HDR à compresser, affichée après une simple compression linéaire, l'image HDR après multiplexage de ses canaux couleur en une image HDR en niveaux de gris, l'image multiplexée compressée, l'image compressée après démultiplexage de l'information colorimétrique.

VPD2.2 [Nar+14]. Également, une étude comparative récente propose une catégorisation et évaluation des approches de l'état de l'art sur le thème de la compression de dynamique de vidéo et montre les évolutions proposées et l'intérêt d'opérateurs spécialisés [EMU17]. La compression d'images HDR est un sujet de recherche qui reste très ouvert, car il dépend également des problématiques d'acquisition d'images HDR et de leur propre évaluation de qualité. Ces thèmes amont sont toujours très actifs et les propositions de protocoles sont encore récentes [LTG17].

D'un point de vue applicatif, nous nous sommes restreints aux prétraitements réalisés au niveau de la rétine et n'avons pas considéré les processus d'analyse visuelle plus centraux en aval de la rétine. Ceux-ci pouvant appliquer à la fois des renforcements d'information, mais également des coupures, le choix de leur intégration dans le modèle reste dépendant de l'application finale à considérer.

Enfin, le modèle proposé est limité par la connaissance du modèle biologique et par nos

FIGURE 2.3 – Illustration de la compression de dynamique sur une séquence temporelle sur une scène HDR de conduite automobile. L'occultation du soleil en fin de séquence n'introduit pas de forte variation d'ambiance lumineuse sur le reste de l'image.

approximations réalisées lors de la définition du modèle numérique. Nous avons par exemple considéré l'application du canal parvocellulaire sur l'ensemble de l'image. Or, l'observateur humain restreint son usage au niveau de sa vision centrale et use de stratégies d'explorations par saccades oculaires libres ou guidées par une tâche afin de couvrir la zone d'observation. Les filtrages spatiaux temporels conduisent alors à des sensibilités différentes. Par exemple, suivre du regard un objet en mouvement permet de réduire le flou de bougé introduit par le modèle et permet d'extraire les détails au détriment du contexte environnant, ce que n'intègre pas notre proposition. Également, la densité des cellules sur la rétine n'est pas homogène et conduit à des filtrages également non homogènes alors que notre approche considère un traitement homogène.

Ces limites sont donc à prendre en compte selon le cadre applicatif. Aussi, certaines limites peuvent être contournées par optimisation, moyennant l'introduction d'une métrique liée à l'objectif applicatif. Deux niveaux d'optimisation sont possibles :

— du point de vue macroscopique, par optimisation des hyper paramètres du modèle. Des paramètres globaux de gains des filtres, leurs constantes spatiales et temporelles sont à ce titre ajustables. Ce cas de figure est intéressant pour trouver une configuration optimale pour un cadre applicatif donné, par exemple optimiser le modèle pour des matériels spécifiques.

— du point de vue microscopique, par une optimisation fine de chaque paramètre du modèle, jusqu'aux coefficients des filtres. Si le modèle constitue le premier étage d'une chaine d'analyse d'image optimisée par apprentissage, il peut être intéressant d'intégrer des critères d'optimisation spécifiques à ce prétraitement pour satisfaire la contrainte de visualisation. Cela est possible du fait de la dérivabilité du modèle. Au sein d'une chaine d'analyse complète, ces critères d'optimisation viendront alors compléter les critères liés à l'objectif applicatif final. Ceci nous amène alors à un problème d'optimisation multicritères, un champ de recherche encore très ouvert sur lequel nous reviendrons dans le chapitre 4.

Explicabilité et justification des décisions obtenues à l'aide de la méthode proposée

La méthode proposée est « experte ». Elle s'appuie sur un modèle de rétine identifié qui répond au problème de compression de dynamique. Elle est en cela explicable. Par ailleurs, la fonction de transfert de chaque composant du modèle est connue et paramétrée

manuellement. Les réponses produites par le modèle proposé sont donc justifiables.

2.1.2 Atténuation des effets de brume (« Dehazing »).

Atténuer la brume, une perturbation non homogène et dépendante du contenu des scènes visuelles observées.

Je me suis récemment intéressé à un aspect plus spécifique du prétraitement : la réduction des effets de brume dans les images. Ces travaux sont actuellement initiés dans une collaboration avec Jean-Baptiste Thomas à l'université de Dijon et l'université norvégienne de sciences et technologie (NTNU). Ils sont soutenus par des stages d'ingénieurs et de master. Ils font suite aux travaux de Jessica El Khoury durant sa thèse de doctorat [El 16].

Les effets de brumes sont généralement liés à des poussières, fumées ou gouttelettes d'eau. Ces perturbations atmosphériques diffusent la lumière émise par les sources et les rayons lumineux renvoyés par les éléments de la scène visuelle avant la captation. Il en résulte un effet de voile qui atténue le contraste des images. La particularité de cette perturbation est son aspect non uniforme. L'effet de brume dépend d'une part de la distance entre les objets et le capteur. D'autre part, sa densité n'est pas constante dans l'espace. Contrairement aux approches d'amélioration d'image classiques qui supposent en général une indépendance entre le signal et la perturbation, le processus de correction des effets de brume doit prendre en compte une réelle dépendance. Les domaines d'applications de ce problème sont relativement nombreux et couvrent l'aide à la conduite de véhicule, la surveillance de milieux en conditions difficiles jusqu'au domaine multimédia afin de présenter des contenus de la meilleure qualité possible. L'atténuation des effets de brume est donc plutôt un prétraitement dont le but est d'augmenter la visibilité des détails importants dans l'image qui permettra de maximiser l'efficacité des processus en aval.

Vers un modèle physique des effets de brume.

Le phénomène physique est approché par le modèle de Koschmieder qui présente quelques limites à dépasser.

L'atténuation des effets de brume est abordée jusque là par des approches d'amélioration d'image classiques comme l'égalisation adaptative d'histogrammes, [XLJ09] mais également par inversion du phénomène physique sous-jacent [HST11] proposé par Koschmieder en 1924 [Kos24]. Ce modèle est défini selon l'équation 2.1 qui détermine l'image I captée par un observateur en fonction de l'image de référence sans brume J associée à des informations liées au contexte de la scène visuelle.

$$\begin{gathered} I(x,y) = J(x,y) \times t(x,y) + I_\infty \times (1 - t(x,y)) \\ avec \\ t(x,y) = \mathrm{e}^{-\beta d(x,y)} \end{gathered} \tag{2.1}$$

En tout point (x, y), le rayonnement global de la scène I_∞ et la transmittance $t(x, y)$ prenant en compte la distance des objets à la caméra $d(x, y)$ et le coefficient de diffusion de la brume β permettent de déterminer l'information captée par l'observateur.

Ce modèle physique a cependant quelques limites. Il suppose par exemple une brume homogène. Également, il a été montré dans [ETM16] que le modèle est inadapté aux forts niveaux de brumes. Enfin, l'atténuation des effets de brume est un problème mal posé dans le cas de prise de vue libre, car la distance à la caméra et les propriétés de la brume ne sont en général pas connues. Dans de telles conditions les méthodes d'inversion directe du modèle peuvent difficilement donner un niveau de performance suffisant.
Les recherches actuelles visent donc à affiner le modèle physique afin d'étendre le champ d'application actuel [LS16] et d'associer la tâche de correction de brume à des tâches de niveau sémantique supérieure comme la détection et reconnaissance d'objets. La difficulté

principale est cependant liée à la disponibilité de données permettant d'identifier un modèle fiable et de le valider. La littérature s'appuie sur des bases de données présentant différentes scènes visuelles avec et sans brume, à des degrés de calibration variables. Actuellement, seules quelques bases de données sont diffusées. La plupart reposent sur une simulation de la brume comme les bases FRIDA (Foggy Road Image DAtabase) [Tar+12] et FRIDA2 [Tar+10]. Ce type de scènes synthétiques est cependant biaisé, car il s'appuie sur le modèle de Koschmieder pour simuler la brume. Seules quelques bases de données comme I-Haze [Anc+18a], O-HAZE [Anc+18b] et CHIC [ETM16] proposent des scènes réelles, mais sont restreintes à un jeu limité de scènes et contextes. La base CHIC est une des plus maitrisées, car elle propose une gradation contrôlée des niveaux de brume et des chartes de calibration Macbeth placées à différentes distances de la caméra comme le montre la figure 2.4. Elle offre alors la possibilité de définir de nouveaux modèles des effets de brume et permet également de comparer finement différentes méthodes de correction. Cette base présente tout de même quelques limites comme la quantité de scènes visuelles disponibles réduites et une relative homogénéité de la brume.

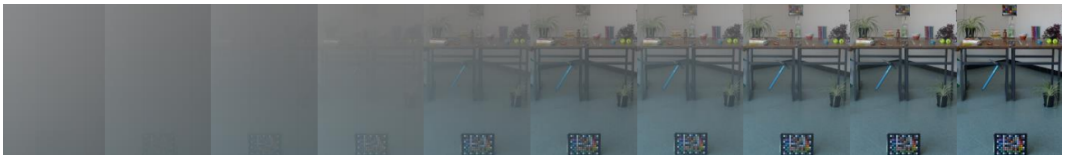


FIGURE 2.4 – Un exemple de scène visuelle de la base CHIC présenté avec ses 10 niveaux calibrés de brume.

Un comparatif de méthodes

Comment comparer les méthodes et évaluer les limites du modèle physique ? Contributions issues d'une collaboration avec NTNU et l'université de Dijon.

Dans ce travail collaboratif engagé avec NTNU, nous nous sommes intéressés à la comparaison de différentes approches de la littérature capables de corriger les effets introduits pas la brume. En particulier, nous avons comparé des approches classiques à des méthodes basées sur les réseaux de neurones profonds optimisés par apprentissage. Ce type d'étude a pour but de vérifier dans quelle mesure les méthodes utilisées actuellement basées sur le modèle de brume de Koschmieder ont atteint leurs limites et si des méthodes par apprentissage peuvent les dépasser. Ces travaux préliminaires se sont appuyés sur la base CHIC et permettent des mesures comparatives entre :

- les approches classiques comme DCP [HST11] qui inverse le modèle de Koschmieder et CLAHE [XLJ09] basée sur des égalisations d'histogrammes adaptatives.
- des réseaux de neurones intégrant le modèle physique. DehazeNet par exemple [Cai+16] estime la transmittance et l'intègre ensuite dans le modèle physique. D'autres réseaux comme AODNet [Li+17] estiment directement l'image sans brume en reformulant l'équation du modèle physique. L'efficacité de ces approches dépend de la structure des réseaux et leur entrainabilité. Mais elle dépend aussi fortement de la qualité du modèle physique utilisé.
- des réseaux de neurones n'intégrant pas de modèle physique. Ceux-ci sont entrainés dans le seul but de « traduire » une image avec brume en la même image sans brume. Ne plus se conformer à un modèle physique peut être intéressant si sa fiabilité est réduite, mais cela implique que le réseau puisse identifier sans information supplémentaire le modèle physique sous-jacent. L'efficacité de cette approche dépend alors

des données, de leur diversité, de la structure du modèle et de son optimisation afin d'assurer une généralisation dans toutes les situations attendues. Une proposition récente est le réseau CycleDehaze [EGE18] qui s'appuie sur les réseaux génératifs entrainés de façon non supervisée. Dans cette approche, deux réseaux sont définis, l'un $F : X \rightarrow Y$ capable d'atténuer la brume et l'autre $G : Y \rightarrow X$ capable de l'ajouter. Une contrainte cyclique introduite dans la phase d'entrainement force la reconstruction fidèle du contenu original X dégradé par la brume après application des deux fonctions de traduction $\tilde{X} = G(F(X))$ et inversement $\tilde{Y} = F(G(Y))$.

La principale métrique utilisée dans ces études est le PSNR. Cette mesure est la plus utilisée dans la littérature bien qu'elle ne soit pas directement corrélée à la qualité perçue par un observateur humain. La nature des erreurs mesurée par le PSNR est variée (bruit, motifs locaux différents). Ce qui rend discutable cette métrique est le fait que des images de qualité perçue différente peuvent donner la même valeur de PSNR. On peut cependant considérer qu'une valeur supérieure à 20dB indique une qualité acceptable. Nous nous sommes appuyés sur des mesures plus spécifiques liées à la fidélité couleur, et des critères liés aux textures et contours des images dans [CBT18].

Résultats et limites

Les premiers travaux publiés [CBT18 ; CTB18] ont montré que les approches par apprentissage profond donnent des niveaux de performance au moins équivalents aux approches classiques. Dans le cas du réseau DehazeNet par exemple, la préservation des couleurs et la qualité de l'estimation de la transmittance peuvent dépasser les approches classiques, mais la qualité des détails est dégradée. Ces résultats sont intéressants si l'on prend en compte le fait que le réseau a été entrainé de façon biaisée par l'usage de bases de données pour lesquelles la brume a été simulée par le modèle de Koschmieder. On peut donc supposer obtenir de meilleurs résultats par un apprentissage sur des scènes prises dans des conditions réelles. Néanmoins, les limites du modèle physique impactent les performances des réseaux qui s'y conforment. Tout comme les approches classiques basées sur le modèle physique, la correction des forts niveaux de brume n'est pas satisfaisante.
Nous avons mené des expérimentations plus récentes sur les réseaux n'intégrant pas explicitement le modèle physique. Les résultats préliminaires obtenus avec le réseau CycleDehaze montrent que de telles approches sont capables de dépasser les limitations du modèle physique en particulier pour les forts niveaux de brume comme montrés sur la figure 2.5.

Les méthodes basées sur les réseaux de neurones peuvent rivaliser avec les approches classiques. Leur usage ouvre à l'apprentissage implicite du modèle de brume afin de contourner les limites du modèle physique identifié.

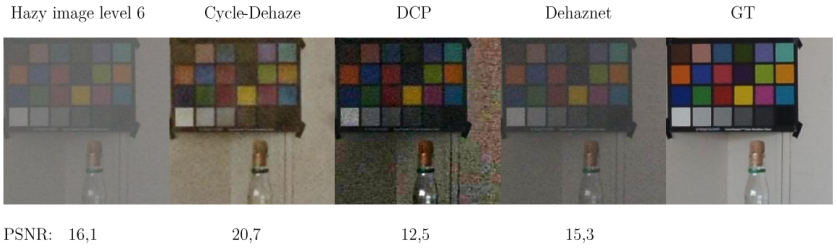


FIGURE 2.5 – Comparaison de méthodes d'atténuation de brume sur une portion d'image de la base CHIC avec la référence sans brume (GT, image de droite). Les mesures de PSNR donnent une indication de qualité (les valeurs hautes suggèrent une meilleure qualité).

Les méthodes par apprentissage montrent cependant une grande sensibilité aux données

d'entrainement. Celles-ci doivent être les plus variées possible de façon à focaliser l'apprentissage sur la tâche d'atténuation de brume plutôt que sur l'apprentissage du contenu des images.

Une piste intéressante pour contrer le manque de données calibrées pourra consister à associer la tâche d'atténuation des effets de brume à des tâches connexes comme la modélisation de la physique de l'interaction entre la lumière et la matière. Elle peut également être associée à des tâches de plus haut niveau comme la détection et la classification d'objets, la reconnaissance de scènes, etc. Ces autres tâches peuvent permettre de faire converger les processus de prétraitement vers une solution relativement générale. Celle-ci pourra alors être affinée sur la gestion des problèmes de brume par la prise en compte de ses critères d'optimisations spécifiques. Nous abordons là à nouveau le thème de l'apprentissage multitâches sur lequel nous reviendrons dans le chapitre 4.

Explicabilité et justification des décisions obtenues à l'aide des méthodes proposées

L'atténuation des effets de brume est un problème important pour des cadres applicatifs comme la conduite autonome de véhicules. Il est donc critique de connaître l'impact d'un tel prétraitement sur des applications comme la détection de piétons, d'obstacles, etc. Les méthodes « expertes » étudiées sont basées sur le modèle physique. Leurs propriétés et limites étant connues, la réponse de ces méthodes est donc justifiable. En revanche, les modèles basés sur des apprentissages sont des structures neuronales profondes et complexes. Leur fonction de transfert est fortement non linéaire et ne peut actuellement être analysée qu'expérimentalement sur un jeu de test contrôlé comme la base CHIC. En revanche, une analyse fine exige une grande diversité de scènes visuelles ce qui pose encore problème. Il est donc actuellement difficile de justifier finement les réponses de tels systèmes. Les travaux initiés avec NTNU pourront s'orienter vers cette problématique.

2.2 Prétraitement pour l'analyse d'images

Nous nous intéressons ici au prétraitement et au sous-échantillonnage de données 2D+X.

Le prétraitement dans une chaine d'analyse de données est de nature variée. L'objectif initial est, comme discuté dans la section précédente, d'améliorer la qualité des données par filtrage de signaux. Néanmoins, d'un point de vue plus général, le prétraitement peut également inclure des processus de ré-échantillonnage, de détection et gestion des données manquantes ou aberrantes. Ceci implique alors des transformations parfois fortes des signaux et de leurs dimensions.

Dans une chaine complète d'analyse de données, le prétraitement est en général suivi d'une étape dite d'extraction de caractéristiques qui sera décrite au chapitre suivant. Cette étape consiste à extraire des motifs (on parle de « patterns » ou « features »), trouver des points de rupture, des tendances, etc. La frontière entre le prétraitement et l'extraction de caractéristiques est donc relativement floue. Cependant, nous considèrerons que le prétraitement se limite à des transformations génériques des données est n'est pas spécifique à une application. Il se restreint au renforcement de l'information jugée pertinente et à la réduction de la variabilité des données liée au capteur et au contexte d'acquisition (luminosité, nébulosité, etc.).

Dans ce cadre, je me suis intéressé à des approches spécifiques de prétraitement pour des tâches d'analyse d'image de haut niveau sémantique : pour la tâche de reconnaissance de concepts sémantiques, dans les séquences d'images, présentée dans la section 2.2.1 et pour la reconnaissance de types de sols depuis des images aériennes hyperspectrales que nous

illustrerons dans la section 2.2.2. Pour ces deux cadres applicatifs, l'extraction d'une information pertinente dans les données brutes est une première problématique. L'information locale dans ce type d'image est fortement liée ou corrélée à son voisinage spatial, temporel et spectral. Avant le processus d'extraction de caractéristiques locales, le prétraitement peut alors consister à améliorer l'information locale par filtrage et à intégrer l'information pertinente liée au voisinage. Cela permet alors un sous-échantillonnage simplifiant les données brutes tout en réduisant la perte d'information.

2.2.1 Filtrage spatio-temporel

Nous nous intéressons au prétraitement pour normaliser les données vidéo et fiabiliser leur indexation.

Dans le cadre de la thèse de Sabin Tiberius Strat, nous nous sommes intéressés à des méthodes de description des vidéos pour répondre à un problème d'indexation sémantique. Une telle tâche est relativement complexe, car une vidéo est un média présentant une très grande richesse de données. Elle est une agrégation de différents plans constituant un arc narratif et chaque plan représente une variété de concepts sémantiques (scène, objets, actions, etc.). Leur indexation est donc difficile et une approche classique vise alors à sous-échantillonner fortement les vidéos jusqu'à ne considérer que l'analyse de quelques images par plan. La description de ces plans s'expose alors à un fort risque de perte d'une information pertinente, typiquement l'information temporelle. Une approche permettant de réduire cette perte est de définir un prétraitement capable d'intégrer l'information temporelle liée au voisinage du plan.

Cette problématique a été considérée dans le cadre du concours TRECVid. Il s'appuie sur de grands corpus de vidéo multimédia d'origines diverses (séries télévisées, vidéo amateur, etc.) et l'un des objectifs est alors d'évaluer l'applicabilité de méthodes d'analyse automatique sur des corpus réels à grande échelle. Plus spécifiquement, la tâche SIN (« Semantic INdexing ») du concours consiste en l'indexation de plans vidéo. Ces plans sont de résolution, vitesse d'échantillonnage, compression et condition d'acquisition très variées comme l'illustre la figure 2.6.

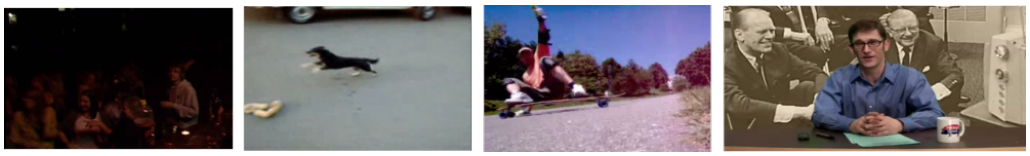


FIGURE 2.6 – Exemples d'images clefs issues du jeu de données du concours TRECVid. La qualité d'image est variable (encodage, résolution). La prise de vue n'est pas contrainte et la caméra peut être fixe ou mobile.

Les méthodes utilisées sur la période 2012-2015 de nos participations étaient basées sur les sacs de mots visuels [Siv+05]. Elles s'appuient sur une description de patchs locaux des images réalisée par des opérateurs comme SIFT [Low04]. Leur application pour l'analyse vidéo est possible lorsqu'il s'agit de décrire des images clefs, mais l'information temporelle est alors occultée.

Vers des sacs de mots spatio-temporels

Comment réduire la variabilité des données liée à la captation et intégrer une information temporelle ? Contributions de thèse de Sabin Tiberius Strat.

Nous avons proposé un prétraitement basé sur le modèle de rétine spatio-temporelle résumé dans la section 1.3. Il vise à normaliser les données et introduire une information temporelle dans des chaines de traitement basées sur des descripteurs locaux de la famille SIFT. Comme montré sur la figure 2.7, pour un extrait vidéo appliqué en entrée du système, le prétraitement rétinien est suivi de la synthèse des sacs de mots qui sont finalement exploités par un classifieur chargé de reconnaitre les concepts recherchés.

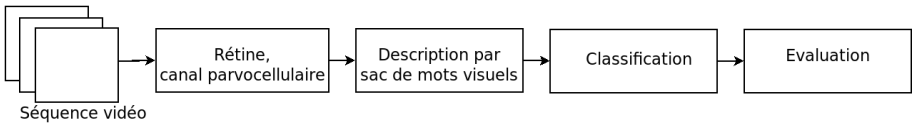


FIGURE 2.7 – Prétraitement d'extraits vidéo par un modèle de **rétine spatio-temporelle** pour une tâche d'indexation sémantique. On représente ici l'application directe sans détailler la phase d'entrainement de la chaine, plus de détails dans [Str+12b ; Str+14].

Dans ce contexte, le filtrage réalisé par le canal parvocellulaire de la rétine est appliqué juste en amont du descripteur standard SIFT ou ses variantes. Les objectifs sont les suivants :

— réduire la variabilité des données liées à l'acquisition. On profite là d'un paramétrage générique du modèle de rétine qui transforme le jeu de données complet de façon homogène et normalise chaque extrait vidéo dans une configuration similaire de luminance moyenne et niveau de bruit spatio-temporel. Le spectre de l'information spatiale statique dans le plan caméra est également blanchi et les détails sont ainsi renforcés. Le processus de description des contenus aval pourra alors extraire une information standardisée.

— intégrer, dans l'image clef décrivant l'extrait vidéo, une information liée au contexte spatio-temporel. Pour cela, le modèle de rétine est appliqué sur la séquence d'images en amont de l'image clef cible. L'information statique dans le plan de la caméra est renforcée et celle liée au mouvement est lissée. Le descripteur appliqué en sortie du prétraitement décrira alors plus finement le contenu focalisé par la caméra.

L'efficacité du préfiltrage proposé ne peut être fiablement validée que d'un point de vue global. On évalue donc la performance de la chaine de traitement complète sur la tâche d'indexation. La métrique de performance est alors la métrique définie comme référence par les organisateurs du concours. Il s'agit d'une précision moyenne inférée adaptée aux grands jeux de données [YKA08]. Le travail au sein du consortium IRIM a grandement facilité cette évaluation en proposant une chaine de traitement de très bon niveau scientifique et un processus de validation travaillant sur un très grand corpus de données, de l'ordre de quelques centaines de milliers d'extraits vidéo dans lesquels 346 concepts doivent être détectés.

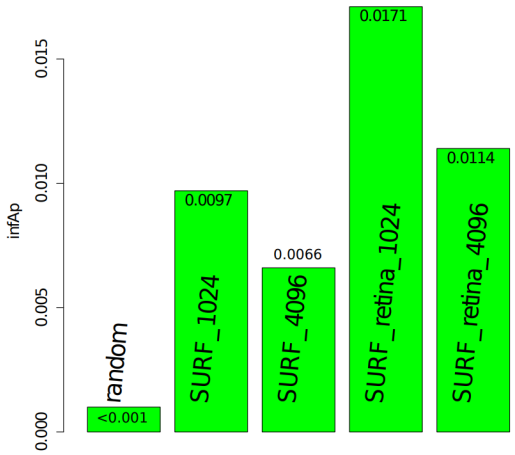


FIGURE 2.8 – Performance sur la tâche d'indexation pour un descripteur basé sur une approche « sacs de mots » utilisant le descripteur SURF. L'usage du modèle de rétine en amont permet un gain significatif de performance pour différentes configurations du descripteur (les valeurs 1024 et 4096 représentent la taille du dictionnaire de mots visuels). Plus de détails dans [Str+14].

Résultats et limites

Un prétraitement spatio-temporel « expert » basé sur un modèle de rétine permet un gain significatif des performances de la chaine d'indexation.

Les travaux publiés dans [Str+12b ; Str+14 ; SBL13] détaillent et comparent l'approche classique de description par sac de mots visuels à notre proposition utilisant un prétraitement. La participation au concours TRECVid a également permis d'évaluer et diffuser la méthode au travers des publications liées au concours [Bal+12 ; Bal+13 ; Bal+14 ; Le +15].

Les descripteurs SURF et SIFT sont utilisés comme outils de description locale au début de la chaine sac de mots et l'on observe, dans les deux cas, un gain significatif lorsque le prétraitement proposé est appliqué. Bien que ces descripteurs aient été conçus pour être robustes dans une certaine mesure à des variabilités bas niveau comme la luminance, la rétine renforce encore cette robustesse et valorise l'information pertinente en intégrant également l'information temporelle. La figure 2.8 montre les niveaux de performances pour des sacs de mots basés sur SURF, différentes configurations des sacs de mots et l'usage de la rétine. Le prétraitement rétinien permet un gain significatif. Nous pourrons tout de même noter le niveau de performance très faible sur la tâche lorsqu'un seul descripteur est utilisé. Ces mesures sont relativement standard sur ce problème d'indexation, car une seule méthode de description n'est pas suffisante pour franchir le fossé sémantique qui sépare les données brutes et les concepts à détecter. C'est pourquoi le consortium IRIM proposait plus d'une cinquantaine de méthodes de description comme nous le verrons dans le chapitre suivant.

La méthode « experte » proposée présente quelques limites principalement liées à l'adéquation de son paramétrage aux propriétés des données à traiter. La résolution spatiale des contenus vidéo a peu d'impact, car l'amélioration d'image appliquée par la rétine est relativement indépendante des contenus et de leur échelle pour la base de données considérée. En revanche, le prétraitement est sensible à la régularité de l'échantillonnage temporel. Les vidéos pouvant être réencodées à des fréquences temporelles variables, l'altération du contenu

temporel peut alors entrainer au niveau de la rétine la création d'artéfacts susceptibles de fausser la description du contenu en aval. C'est une limite classique des prétraitements, lorsqu'appliqué sur des données non calibrées.

À ce niveau de la chaine d'indexation, nous retiendrons que le prétraitement « expert » proposé et basé sur le modèle de rétine permet un gain de performance. La normalisation des informations sur les images clefs prétraitées est une première raison, car les données transmises à l'étape aval de description ont alors une variabilité liée à l'acquisition fortement réduite. Également, l'intégration de la composante temporelle a son importance, mais n'est pas pleinement exploitée par le processus qui suit. Le chapitre 3 présentera des propositions spécifiques pour la description de contenu spatio-temporel.

Explicabilité et justification des décisions obtenues à l'aide des méthodes proposées

Tout comme la méthode de compression de dynamique basée sur le modèle de rétine, la connaissance de la fonction de transfert du prétraitement proposé permet de justifier les réponses de ce prétraitement.

2.2.2 Prétraitement de données hyperspectrales

Les images hyperspectrales représentent une information locale très riche qui s'oppose à une connaissance en général faible sur ces données.

Nous nous sommes intéressés à l'imagerie hyperspectrale dans le cadre de la thèse d'Amina Ben Hamida. Ce type d'imagerie permet la captation d'images par un échantillonnage régulier et dense de l'information spectrale. Des capteurs proposant une description de chaque pixel par plus d'une centaine de bandes spectrales sont ainsi proposés pour décrire le spectre visible et au-delà à une résolution spectrale extrêmement fine. Contrairement aux capteurs multi spectraux classiques sensibles à un nombre de bandes spectrales plus larges, moins denses et échantillonnées de façon moins régulière, ce type de captation rend possible une caractérisation fine des matériaux. On trouve un fort intérêt en imagerie satellitaire et aéroportée (cf. figure 2.9) pour différents usages, de la reconnaissance de sols jusqu'à la détection et quantification de composants des sols et particules aériennes. L'imagerie hyperspectrale a d'autres champs d'application comme l'analyse de carottes prélevées dans les sols pour quantifier localement les composants des sols et leur évolution au cours du temps [Jac+19a].

L'intérêt pour ce type de capteur est grand compte tenu des enjeux actuels comme le réchauffement climatique, la surveillance des sols, la détection de changement, etc. En revanche, la richesse d'information proposée par ce type d'imagerie est difficile à appréhender.

Un prétraitement classique mis en oeuvre dans la littérature consiste en un fort sous échantillonnage. Plusieurs méthodes sont généralement admises, et reposent sur une forme de sélection de bandes spectrales d'intérêt. Cette sélection peut consister en une recherche des bandes corrélées pour lesquelles l'expert ou une méthode de maximisation d'information sélectionne une bande représentante [Hab18]. D'autres méthodes non supervisées de réduction comme l'analyse en composantes principales permettent de ne retenir que les dimensions portant l'information dominante. Il existe dans toutes ces méthodes un risque de perte d'information important pouvant impacter l'efficacité des processus de traitement appliqué en aval.

Par ailleurs, l'information liée au voisinage spatial est en général peu prise en compte. Différentes raisons souvent dépendantes du contexte applicatif sont en général données :

— le fait que certains capteurs soient de résolution spatiale faible réduit l'intérêt de considérer le voisinage. Par exemple, le capteur Hyperion présente une résolution

FIGURE 2.9 – Exemple d'images hyper spectrales captées par l'imageur AVIRIS. a) l'université de Pavia en Italie, b) le centre-ville de Pavia, c) le centre spatial Kennedy. Cette représentation colorée est basée sur la sélection de seulement 3 bandes spectrales dans le spectre visible.

spatiale de 30m et l'on peut estimer que le contenu d'un seul pixel est déjà suffisamment hétérogène pour écarter l'usage de l'information apportée par le voisinage.
— si au contraire la résolution spatiale est très fine comme sur le capteur aéroporté AVIRIS, la corrélation spatiale entre pixels voisins peut être considérée comme suffisante pour ne considérer que chaque pixel indépendamment.
— d'une façon générale, le traitement conjoint de l'information spatiale et de l'information spectrale est difficile, car cela nous plonge dans les problèmes d'analyse de données en grandes dimensions. Les outils mathématiques sont alors moins nombreux et les coûts de calculs limitent rapidement leur applicabilité.

À l'opposé de ces méthodes classiques de réduction de données, d'autres approches consistent au contraire en la décomposition de l'information en une distribution de données filtrées. Pour de telles approches, la difficulté est alors liée au choix de la base d'ondelettes permettant de décrire l'information pertinente pour la tâche cible. Des travaux comme [Kav+14] montrent des résultats intéressants, mais le choix de la base d'ondelettes est dépendant des données et des objectifs et exige une expertise.

Des approches par apprentissage comme les réseaux de neurones conservent cette idée, mais permettent une optimisation de la base de filtres de décomposition en fonction des données et de la tâche à réaliser. On peut ainsi espérer obtenir une décomposition plus pertinente. En revanche, la difficulté est cette fois la disponibilité de données en quantité suffisante pour permettre un apprentissage optimal de la base de décomposition.

Vers un prétraitement adapté, optimisé par apprentissage

Comment sous échantillonner les données hyperspectrales ? Contributions de la thèse d'Hamina Ben Hamida.

Nous nous sommes intéressés à l'usage de l'imagerie hyperspectrale aéroportée pour la classification des sols au niveau du pixel. Il s'agit en cela d'une tâche de segmentation sémantique. L'approche considérée vise à appliquer le principe de l'apprentissage profond en traitant directement la donnée brute. La tâche d'apprentissage sera détaillée au chapitre 4. Nous nous focalisons ici sur les prétraitements réalisés par les premières couches du réseau. La structure de ces premières couches a été spécifiée dans le but de réaliser un

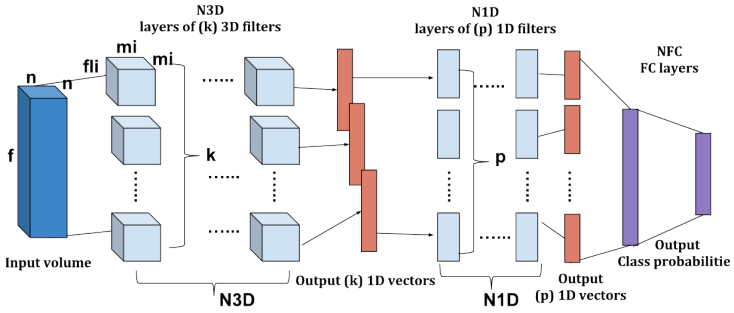


FIGURE 2.10 – Principe général de l'architecture neuronale proposée dans [Ben+18b]. Une cascade de couches de convolutions 3D décompose l'information spatio spectrale appliquée en entrée jusqu'à générer des descriptions 1D spatio spectrales. Une cascade de couches de convolutions 1D suivie de couches de neurones densément connectés finalise la description des données avant la classification finale.

sous-échantillonnage spatio spectral minimisant la perte d'information.
L'apprentissage profond dans ce type de contexte est actuellement difficile du fait de la faible quantité de données à disposition. Seules quelques images annotées sont mises à disposition, mais sont de dimensions spatiales faibles. Les travaux de la littérature basés sur les réseaux de neurones s'appuient sur une réduction des données réalisée par un prétraitement expert amont choisi parmi ceux discutés au paragraphe précédent. Par exemple [Hu+15] se limite à l'analyse spectrale de chaque pixel sans considérer le voisinage spatial et [Mak+15] applique une réduction de dimension avant l'usage d'un réseau de neurones en s'appuyant sur une analyse en composantes principales.

Notre approche a consisté en l'usage de couches de neurones de convolution 3D dans les premières couches du réseau. Ces opérateurs traitent ainsi l'information spatio spectrale localement. Ils utilisent des noyaux de convolutions d'étendue spatiale et spectrale réduite appliqués avant un sous échantillonnage. Leur optimisation réalisée par un processus d'apprentissage permet un prétraitement adapté avant l'application des couches neuronales aval. Ces traitements appliqués dans les premières couches d'un réseau profond peuvent être considérés comme des prétraitements à part entière s'ils sont génériques à une variété d'applications. Nous nous sommes dans cette première étude focalisés sur la segmentation sémantique, mais leurs réponses peuvent également être considérées pour d'autres tâches comme l'identification et la quantification des différents composants des sols (souvent dénommés « end members ») [Kum+12] sur les mêmes jeux de données.

Résultats et limites

Des convolutions spatio spectrales apprises pour sous-échantillonner de façon adaptée.

Les travaux conduits dans le cadre de la thèse d'Amina Ben Hamida publiés dans [Ben+18b ; Ben+16] ont permis de valider l'approche.

La figure 2.10 montre l'architecture générale des réseaux que nous avons proposée. Un ensemble de couches de convolution spatio spectrales (3D) associées à des étapes de sous échantillonnage réalise un prétraitement. Ils délivrent alors une information spatio spectrale sous la forme de vecteurs 1D sur lesquels s'appuient les couches suivantes (convolutions 1D et couches densément connectées) pour réaliser la tâche de classification. L'architecture globale et son optimisation seront discutées plus en détail au chapitre 4 dans la section 4.1.1.

Le choix d'opérateurs spécifiques, les convolutions 3D, en début de réseau permet de réaliser un prétraitement adapté aux données d'entrée échantillonnées de façon dense dans leurs dimensions spatiales et spectrales. Le prétraitement permet alors, par les filtrages 3D, d'agréger l'information locale avant sous échantillonnage réduisant ainsi la perte d'une information potentiellement pertinente. La tâche d'analyse des étages suivants est simplifiée et peut être réalisée par un réseau réduit en termes de nombre de neurones et de paramètres associés.

En revanche, l'aspect générique des prétraitements proposés n'est pas montré dans ces premiers travaux. Ce sera fait par la considération de tâches annexes comme la quantification des matériaux pour chaque pixel [Kum+12] sur le même type de données. Cela sera rendu possible par la définition de processus dédiés à cette tâche, une structure neuronale dédiée par exemple, qui sera alimentée par la sortie de ces prétraitements. Nous abordons là le thème de l'apprentissage multitâches discuté au chapitre 4.

Explicabilité et justification des décisions obtenues à l'aide de la méthode proposée

Le prétraitement proposé basé sur des convolutions 3D est optimisé par apprentissage. La réponse de ces filtrages est donc dépendante des données d'entrainement et du ou des critères d'optimisation. Les modèles de prétraitement proposés sont constitués de peu de couches d'opérateurs. Il est donc possible d'identifier la fonction de transfert des filtrages de prétraitement réalisés. En revanche, des cascades de tels opérateurs accompagnés de non-linéarités comme on les trouve classiquement dans les réseaux de neurones profonds poseront rapidement des problèmes d'identification des filtres équivalents ainsi que de leur justification.

2.3 Synthèse et perspectives de recherche sur le prétraitement des données

Isoler le prétraitement permet de spécifier ses propres critères à optimiser. Associer des approches expertes et des approches entrainables permet de dépasser nos connaissances partielles sur les modèles physiques à appréhender.

Le prétraitement est dit « bas niveau » car il est très fortement lié à la physique du capteur et aux conditions d'acquisition. Il vise à réduire la variabilité liée à ces facteurs. Nous avons vu dans ce chapitre un ensemble de méthodes au travers de deux problématiques, le prétraitement pour le filtrage et le renforcement de l'information ainsi que le prétraitement introduisant des contraintes de visualisation. Nous avons abordé ces approches par des méthodes expertes et d'autres basées sur un apprentissage.

Il est en général possible d'introduire des critères permettant de qualifier la qualité des prétraitements, en particulier en introduisant une contrainte de visualisation ou des connaissances sur un modèle physique sous-jacent. Ceci permet, dans une chaine de traitement complète, de mieux comprendre voire de justifier les réponses du système global par connaissance des informations transmises en sortie de prétraitement. Cette approche est classique sur les chaines d'analyse traditionnelles et les méthodes expertes. Leur limite potentielle est cependant liée à leur capacité à respecter fidèlement la contrainte ou le modèle physique sous-jacent.

Au contraire, les approches basées sur les réseaux de neurones profonds tendent à occulter la définition explicite d'un prétraitement. Leur optimisation consiste généralement à ajuster simultanément l'ensemble du réseau pour minimiser un critère lié à l'objectif global du système. Sur des tâches de haut niveau sémantique comme la reconnaissance d'images,

le prétraitement est donc réalisé de façon implicite, mais n'est pas clairement localisé dans la structure de traitement. Différents risques apparaissent alors :

— une difficulté à isoler le processus de prétraitement ce qui empêche sa décentralisation et son rapprochement vers la source de données.
— une forte limitation de notre possibilité d'explication des traitements réalisés et de justification des réponses du système global.
— une non-généricité du prétraitement du fait de l'usage d'un critère d'optimisation principalement lié à l'objectif global du système.

Nous avons vu dans ce chapitre qu'il est possible de considérer des approches par apprentissage dédiées au prétraitement. Ceci permet de revenir à une structuration classique de chaine de traitement et réduit alors les risques que nous venons de citer. Des perspectives intéressantes se dessinent alors.

Pistes de recherche

En apprentissage profond, introduire explicitement un critère d'optimisation du prétraitement en complément du critère d'optimisation global à la tâche.

Une première piste concerne l'association des approches expertes avec celles obtenues par apprentissage. Sur les problèmes de prétraitement en particulier, les approches expertes sont souvent une approximation d'un modèle physique. Il est alors intéressant de les associer à des modèles plus flexibles capables de surmonter l'imprécision apportée par cette approximation. Cela peut se faire sous différentes approches parmi lesquelles :

— une utilisation conjointe. Pour satisfaire un critère sur la tâche de prétraitement (correction de bruit, renforcement de signal, atténuation de la brume ou autre), une solution peut consister en la fusion des réponses d'un modèle expert avec celles d'un modèle obtenu par apprentissage. Dans ce cadre, une analyse des contributions de chaque modalité pourra aider à identifier les processus appris complétant le modèle expert. La connaissance sur le modèle physique sous-jacent pourra alors progresser.
— une approche professeur/élève. Si la fiabilité du modèle expert est bonne, l'optimisation d'un modèle entrainable imitant ses réponses peut être intéressante pour réduire les temps de calcul et affiner la précision sur les cas limite. L'entrainement sur une grande variété de données peut également amener à affiner la précision par un effet de régularisation. On peut alors considérer ce type d'approche sous l'angle de la réduction de modèle.
— une expertise partielle pour structurer le modèle entrainable. Contraindre la structure d'un modèle entrainable par la prise en compte d'un certain nombre d'aprioris permet de faciliter la convergence de ses paramètres vers une solution plus performante tout en réduisant le coût énergétique de l'entrainement. C'est une approche que l'on trouve dans les systèmes biologiques et notamment le système nerveux. Au-delà de la plasticité de ces systèmes, leur structuration préalablement planifiée permet une adaptation accélérée à l'environnement.

Ces différentes pistes associent donc une expertise à des modèles entrainables. Ceci est une approche permettant d'aller vers une meilleure compréhension des modèles obtenus par apprentissage et une justification de leurs réponses.

Une seconde piste de recherche est directement liée à l'apprentissage et s'intéresse à l'aspect multitâches. Comme le défend [MWK16], les systèmes nerveux biologiques sont naturellement multitâches. D'un capteur aux différentes aires corticales avec lesquelles il peut échanger, l'ensemble des réseaux de neurones les reliant sont capables de réaliser de multiples tâches, locales et globales permettant de répondre aux différentes sollicitations de l'individu pour évoluer dans son environnement. Le processus de prétraitement peut être alors vu comme un processus local pour lequel des critères d'optimisation spécifiques

peuvent être définis. Ce peut être, par exemple, la réduction des effets de brume. Les sorties de ce prétraitement alimentent des processus de plus haut niveau, par exemple la détection d'objets. Une dynamique de l'apprentissage entre ces différentes tâches peut alors être définie en fonction des données à traiter et de la connaissance qui leur est attribuée pour optimiser l'ensemble du système. Cette proposition permet de faire évoluer la chaine de traitement classique illustrée en introduction sur la figure 1.1 vers l'architecture proposée sur la figure 2.11. L'avantage de cette proposition est la possibilité de décentraliser complètement le processus de prétraitement et son optimisation spécifique et ouvre ainsi à l'apprentissage distribué.

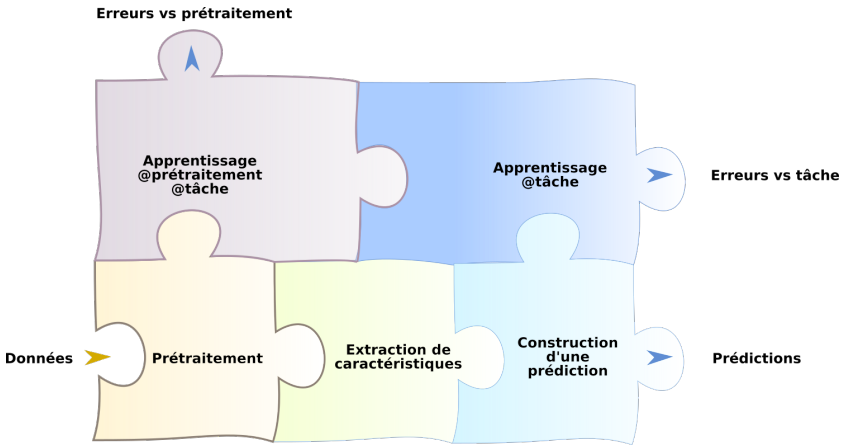


FIGURE 2.11 – Intégration de processus d'optimisations pour le prétraitement dans un système d'analyse complet. Le prétraitement est optimisé sur ses propres critères (suivre un modèle physique ou autre) et est également influencé par la tâche principale.

Chapitre 3

Extraction et fusion de caractéristiques

Mes contributions concernent la proposition de méthodes de description spatio-temporelles des vidéos et des méthodes de fusion tardives.

Extraire des caractéristiques pertinentes, rechercher des motifs, identifier des tendances... une fois l'information captée et prétraitée viennent alors les tâches d'analyse de l'information. Ces objectifs sont en général préliminaires à l'étape finale de prédiction. Elles sont typiques des chaines de traitement classiques et le sont également, de façon implicite, avec les approches actuelles basées sur les réseaux de neurones profonds. L'objectif principal est de générer une forme d'abstraction de haut niveau sémantique des données brutes sur laquelle la prédiction, devant en général répondre à un problème de niveau sémantique encore supérieur, pourra s'appuyer. Cette représentation intermédiaire vise ainsi à franchir le fossé sémantique qui sépare les données brutes de l'objectif final. Elle présente d'une façon plus générale de nombreux intérêts :

— « trouver des invariants, linéariser ». L'objectif le plus important est de projeter les données brutes de grandes dimensions dans un espace pour lequel le problème posé peut être résolu par des processus simplifiés, linéaires. Cet espace devrait alors réduire la variabilité des données tout en maintenant l'information essentielle permettant de séparer les différentes situations.

— « justifier ». Il devient nécessaire de garantir une explication des décisions prises par le système prédictif global. La synthèse de caractéristiques intermédiaires explicables doit aider à répondre à cet enjeu sociétal devenu critique, maintenant que l'on affecte des tâches de plus en plus sensibles aux systèmes automatiques.

— « encoder ». Représenter une information brute sous la forme d'un code dont le sens et l'usage sont connus seulement du système aval permet de garantir une forme de sécurité des données. Par ailleurs, si la description est restreinte à l'information pertinente uniquement liée à la tâche, l'information de contexte porteuse d'information potentiellement privée ne sera pas transmise. On aborde là le thème de la préservation de la confidentialité des données, un autre enjeu sociétal majeur.

— « compresser ». La réduction des coûts de transport est un objectif important, que ce soit au sein d'une infrastructure de calcul centralisée comme en mode décentralisé (l'informatique connectée, l'internet des objets, etc.). On trouve alors un intérêt réel à ne transmettre sur les canaux de communication qu'une information de volumétrie la plus réduite possible.

L'extraction de caractéristiques est donc un processus $f : x \rightarrow z$ qui consiste en une forme de changement de variable des données brutes x pour la synthèse d'une description z

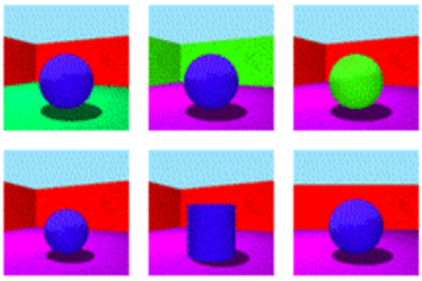


FIGURE 3.1 – Exemples d'images dans un « cas simple » pour lequel une scène de synthèse est contrôlée par un jeu de variables réduit : forme, taille, couleur de l'objet central, couleur des éléments du fond, etc. L'objectif est d'identifier ces paramètres de façon automatique. Illustration reprise de [Loc+18].

répondant à tout ou partie aux intérêts énoncés précédemment. La difficulté du problème est grande, car les données brutes sont en général de grande dimension et à grande variabilité. L'identification de caractéristiques ayant un sens est alors un problème difficile. D'une façon générale, si l'on souhaite identifier les paramètres indépendants contrôlant les données brutes, comment les quantifier sans connaissance a priori ? Des travaux récents sur des scènes synthétiques simples et bien contrôlées, illustrées sur la figure 3.1, montrent la difficulté du problème : dans ce cadre présentant, à première vue, des données à faible variabilité, aucune solution actuelle ne permet d'identifier automatiquement les variables de contrôle sans l'introduction d'apriori sur les données ou sur le modèle [Loc+18].

Cette tâche d'extraction de caractéristiques pertinentes a pourtant une histoire longue. Si l'on se restreint à l'analyse des images, elle couvre l'identification de caractéristiques bas niveau, les contours et couleurs par exemple, à une description de plus haut niveau comme les objets. On trouve d'ailleurs de fortes connexions entre l'analyse des images numériques et l'étude de notre perception en neurosciences :

— En neurosciences, dès 1953, les travaux de Barlow [Bar53] suggèrent l'existence de neurones sensibles à des contrastes locaux au niveau de la rétine de la grenouille. Hubel et Wiesel [HW62] ont ensuite montré sur le chat l'existence de cellules neuronales spécialisées, au niveau du cortex visuel, capables de détecter des contours orientés. De Valois, en 1985, mis en évidence dans [WV85] l'organisation fine du cortex visuel primaire pour décomposer efficacement l'information visuelle par orientations et bandes de fréquences. Des travaux récents semblent montrer que, lorsque nous observons un objet ayant subi une transformation, une déformation, un mouvement, ou autre, des opérateurs cognitifs génériques sont capables de modéliser la transformation de la représentation de l'objet [WIC18].

— En analyse d'image numérique, le cheminement est proche. L'extraction de caractéristiques a commencé par l'extraction d'attributs simples (la couleur, les contours, leur orientation, etc.). Le détecteur de Canny-Deriche en 1986 [Can86] est une proposition marquante. Des représentations plus complexes de l'information visuelle ont ensuite été proposées. Un exemple intéressant est la méthode des sacs de mots visuels au début des années 2000 [Siv+05]. La tendance actuelle est aux réseaux de neurones profonds optimisés par apprentissage. Sur cette dernière avancée, on valorise en particulier la grande capacité de transférabilité des caractéristiques [PY10] entre différents domaines et tâches.

Les travaux en imagerie numérique montrent qu'il est relativement difficile de trouver

une unique fonction de description f satisfaisante pour traiter des cas difficiles comme l'indexation d'images et de vidéos. Une alternative consiste alors à répondre de façon plus pertinente au problème posé en réalisant une étape de fusion. On considère dans ce cadre, pour un échantillon x, une variété de méthodes d'extraction de caractéristiques $\{f_1(x), f_2(x), ...f_j(x)\}$ et au moins une application de prédiction qui associe une caractéristique à un score, par exemple de classification, soit $S : f \rightarrow o$. On distingue alors la fusion précoce U_p qui réalise une agrégation d'information dans l'espace des caractéristiques avant l'application de l'étape de prédiction. Le score final est alors $o_p = S(U_p(f_1(x), f_2(x), ...f_j(x)))$. Une autre approche est la fusion dite tardive qui vise à agréger l'information dans l'espace sémantique, les scores de prédiction, obtenus depuis chaque caractéristique. Nous obtenons alors le score $o_t = U_t(S(f_1(x)), S(f_2(x)), ...S(f_j(x))$. Un ensemble de solutions hybrides complète la panoplie d'outils pour associer les caractéristiques pertinentes à différents niveaux du processus global.

Mes travaux se sont intéressés aux deux problèmes : la génération de caractéristiques décrite dans la section 3.1, et leur fusion présentée en section 3.2 afin de répondre à des problèmes d'indexation et de classification dans les images et séquences vidéo. Ces travaux se déclinent sous la forme de propositions « expertes » et d'autres, obtenues par apprentissage. Après leur résumé, nous pourrons, comme nous l'avons fait au chapitre précédent, discuter de l'intérêt de ces approches et en dégager des pistes de recherche.

3.1 Extraction de caractéristiques dans les images et séquences vidéo

Extraire des caractéristiques décrivant les concepts sémantiques invariantes à leurs transformations.

Extraire des caractéristiques pertinentes peut être ramené au problème de trouver une transformation $f(x)$ capable de projeter des données brutes dans un espace de plus faible dimension permettant de séparer de façon simple les catégories, tendances, etc. L'étage final de décision ou de prédiction pourra alors se réduire idéalement à un simple processus linéaire, par exemple identifier une droite sur un problème de régression, appliquer un SVM linéaire sur un problème de classification, etc. Le verrou scientifique principal est lié au fait qu'un même concept peut être représenté par une grande variété d'exemples, chacun d'aspect très différent (objet déplacé, occulté, déformé, placé dans des ambiances différentes d'éclairage, de scène, etc.). Il faudrait alors être capable d'intégrer, dans f, les groupes de transformations, translations, rotations et déformations complexes qui entrent en jeu. Mais leur identification reste difficile.

Les approches traditionnelles « expertes » d'analyse d'image génèrent des descriptions basées sur seulement quelques niveaux d'abstraction ou d'échelles. Parmi les approches historiques, on retrouve les méthodes basées sur des décompositions multi échelles en ondelettes [DV02] et les sacs de mots [Siv+05]. Ces derniers ont été étendus à l'analyse multi échelles, par exemple par analyse pyramidale [LSP06]. Cependant, quelle que soit l'approche, le choix des méthodes de description est difficile et ne s'appuyer que sur une seule limite l'efficacité à des sous-ensembles des données. C'est ainsi que s'est développé un ensemble de travaux sur la recherche de méthodes de descriptions. Leur combinaison par différentes méthodes de fusion permet ensuite de générer une description globale aux propriétés d'invariance plus pertinentes.

Les approches basées sur les réseaux de neurones profonds optimisés par apprentissage présentent une alternative intéressante. L'unique fonction $f_d(x, \theta_d)$ du réseau, composée d'un grand nombre de couches d'opérateurs et de milliers de paramètres associés, θ_d peut être vue comme l'agrégation d'un grand nombre de descripteurs. Les dernières couches du

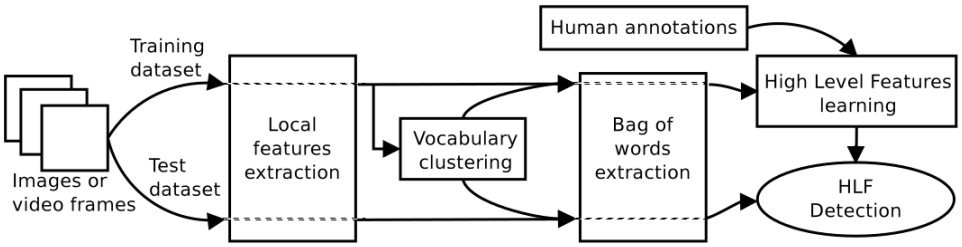


FIGURE 3.2 – Chaine de traitement basée sur la description par sacs de mots visuels. Reprise de [Str13].

réseau combinent alors ces descriptions et délivrent potentiellement des invariants pertinents.

Nous décrivons, dans la section 3.1.1, différentes méthodes d'extraction de caractéristiques « expertes » proposées dans le cadre de la thèse de Sabin Tiberius Strat. Également, des méthodes par apprentissages, considérées au travers des thèses de Rizlène Raoui et d'Amina Ben Amida et de travaux avec Nicolas Voiron, sont décrites dans la section 3.1.2.

3.1.1 Descripteurs Sacs de mots spatio-temporels

Extraire des caractéristiques de flux vidéo, décrire les concepts et leur mouvement.

Dans le cadre de la thèse de Sabin Tiberius Strat, nous nous sommes intéressés à la description de contenus vidéo par sacs de mots visuels. Le cadre applicatif et d'évaluation était le concours international TRECVid [1] sur la tâche d'indexation sémantique « Semantic INdexing, SIN ». Cette tâche vise à détecter des concepts de haut niveau sémantique, des objets aux actions. Ces travaux réalisés sur la période 2012-2015 se situent dans une période encore très prolifique aux méthodes de classification basées sur les sacs de mots visuels introduites par Zisserman [Siv+05]. Cette approche est illustrée sur la figure 3.2 et peut être résumée ainsi :

1. L'extraction de descripteurs « sacs de mots visuels » décrivant chaque contenu. Ils sont obtenus par une première phase de description de zones d'intérêts localisées dans le contenu à analyser. Pour cela, des opérateurs comme ceux de la famille des SIFT [Low04] décrivent localement les orientations des contours. Un dictionnaire de mots visuels est ensuite défini depuis un ensemble de descriptions locales obtenues depuis une base de données d'apprentissage. Identifiés par un clustering, les centroïdes alors nommés « mots visuels » résument les descriptions similaires. Ils représentent une abstraction de plus haut niveau sémantique des motifs extraits de l'image. Par un comptage de leurs occurrences dans une image, on obtient alors un histogramme ou « sac » de mots visuels donnant une description générale du contenu.
2. une classification supervisée : les histogrammes de mots visuels extraits sur les exemples annotés d'une base d'entrainement sont utilisés pour optimiser des classifieurs supervisés. Les classifieurs sont ensuite utilisés pour obtenir des scores de classification des concepts recherchés (High Level Features, HLF sur la figure 3.2) sur des échantillons non connus.

Comme présenté au chapitre précédent dans la section 2.2.1, la détection de concepts sémantiques liés aux objets et aux actions dans les vidéos peut nécessiter l'introduction de

1. http://trecvid.nist.gov

l'information temporelle au sein des sacs de mots. La littérature rapporte des propositions intéressantes comme la détection et description des points d'intérêts spatio-temporels avec les STIP [LL03], l'ajout de l'information locale de mouvement en complément de la description spatiale avec MOSIFT [CH09] ou encore la description de trajectoires de points d'intérêts [BDP11].

Nous avons ainsi proposé un éventail d'extensions complémentaires visant à décrire les objets statiques tout comme les actions. Nous résumons dans les paragraphes suivants deux contributions. Une première est basée sur un usage avancé de la rétine présentée au paragraphe 1.3. Elle étend les descripteurs de la famille des SIFT en intégrant des comportements spatio-temporels. Une seconde proposition est basée sur l'analyse de trajectoires dédiées à la représentation du mouvement.

Usage avancé du modèle de rétine

Comment extraire une description générique pertinente focalisée sur les zones saillantes des vidéos ? Contributions de thèse de Sabin Tiberius Strat.

La rétine biologique permet, de façon générique, de décrire à la fois le contenu statique et l'information de mouvement. Cela est lié à la décorrélation entre ces deux informations opérées par ce capteur et ses prétraitements. Comme nous l'avons vu, deux canaux d'information sont alors disponibles, le canal parvocellulaire portant une information statique, détaillée et colorée et le canal magnocellulaire transmettant l'information temporelle transitoire. Nous avons montré le gain apporté par l'usage du canal parvocellulaire en tant que simple prétraitement au chapitre précédent. La prise en compte de l'information fournie par le canal magnocellulaire ouvre à la proposition de nouvelles descriptions spatio-temporelles génériques.

Nous avons ainsi proposé, dans [Str+14], un ensemble de stratégies intégrant le canal magnocellulaire comme outil de sélection de zones d'intérêt. La sélection réalisée ne se limite pas aux seuls éléments en mouvement par rapport au plan de la caméra comme le font les approches de la littérature. Elle vise les objectifs suivants :

— localiser les zones riches en information (les zones texturées), qu'elles soient statiques ou en mouvement.
— localiser les zones de mouvement.

Comme le montre la figure 3.3, le canal magnocellulaire est parfaitement adapté à cette problématique. Dans le régime transitoire du modèle de rétine, dans les premiers instants d'exposition à un contenu, ce canal renvoie une énergie forte sur les zones texturées. C'est ensuite dans le régime stable de la rétine que le canal magnocellulaire révèle plus particulièrement l'information liée au mouvement. Nous avons ainsi proposé une méthode de seuillage localement adaptée sur ce canal. Elle repose sur une différence de gaussiennes spatiales qui met en valeur les zones de l'image rapportant des événements temporels importants par rapport à leur voisinage. Cette méthode extrait des zones de « saillance spatio-temporelle » dans l'image en s'appuyant sur des informations de bas niveau. Elle se veut rapide et moins spécialisée que des modèles de saillance visuelle au sens classique du terme comme le modèle de Itti et Koch [IKN98]. Ces modèles représentent, sous la forme de carte de chaleur, les zones d'intérêt sur lesquelles un observateur humain focalise son attention. Ils sont obtenus par une modélisation experte des processus réalisés au niveau de la rétine jusqu'au cortex visuel, voire par apprentissage [CBA19]. Notre approche est plus bas niveau, mais a l'avantage d'intégrer l'information temporelle pour focaliser la description sac de mots sur les zones de saillance spatio-temporelle.

Nous définissons alors différentes stratégies de fusion précoce des informations à disposition en sortie de rétine : l'information de détails statiques, l'information transitoire et les zones d'intérêts. Le résultat est la proposition de différents descripteurs spatio-temporels.

FIGURE 3.3 – Extraits d'une séquence vidéo représentant un fond statique et un journaliste sur lequel est appliqué le modèle de rétine décrit au paragraphe 1.3. (a) et (b) montrent des images de l'extrait durant le régime transitoire puis stable du modèle. Les sorties parvocellulaire (c,d) et magnocellulaire (e,f) extraites à ces mêmes instants montrent la réponse du modèle dans ses deux régimes. La sortie magnocellulaire permet la sélection de zones d'intérêt (g,h). Les zones détectées correspondent aux zones texturées dans la phase transitoire du modèle de rétine et se focalisent ensuite sur le mouvement dans la phase stable. Sur cette scène à fond statique, i) montre l'importance accordée aux différentes zones de l'image par cumul des masques sur la période de la séquence traitée. Repris de [SBL14c].

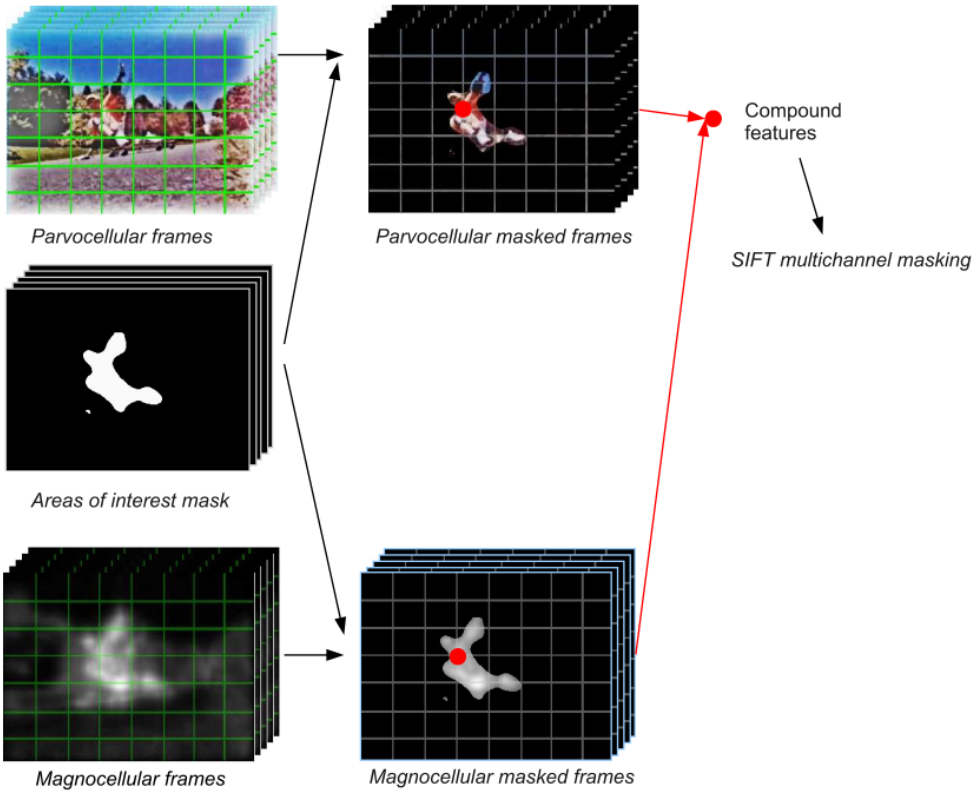


FIGURE 3.4 – Descripteur spatio-temporel « SIFT_multichannel_masking » basé sur le modèle de rétine. Le canal magnocellulaire génère un masque de sélection de zones d'intérêts. Le canal parvocellulaire et le canal magnocellulaire sont décrits localement, à de multiples échelles par le descripteur SIFT. Leurs descriptions en une même position spatiale sont agrégées avant l'étape de synthèse des mots visuels. Repris de [Str13].

Le plus abouti est le descripteur « SIFT_multichannel_masking » capable de décrire, sur les zones d'intérêt, l'information spatiale et l'information transitoire. Chaque canal de sortie du modèle de rétine est décrit à de multiples échelles par le descripteur SIFT sur les zones d'intérêt comme illustré sur la figure 3.4. Leur agrégation locale permet ensuite la synthèse de mots visuels. La description globale de la séquence vidéo s'appuie sur une collecte de ces descripteurs sur une fenêtre temporelle de l'ordre de la seconde. Elle agrège alors une description du contexte donnée principalement dans les premiers instants d'exposition et une description du mouvement, principalement en fin de séquence. Nous synthétiserons les résultats et limites de ces approches avec celles des descripteurs trajectoires présentés ci-après.

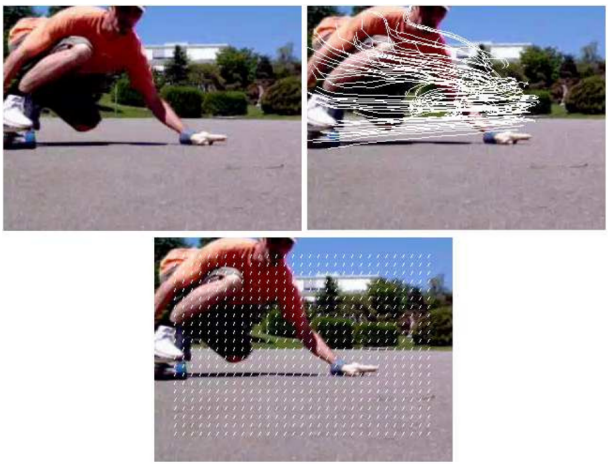


FIGURE 3.5 – Extraction de trajectoires sur une scène de sport en extérieur. Le mouvement de la caméra est estimé depuis une grille dense de points d'intérêts (en bas). L'extraction de trajectoires (à droite) est réalisée sur des points d'intérêts du type GFFT ([JT94]) en compensant le mouvement de caméra. Reprise de [Str13].

Descripteurs Sac-de-Mots de trajectoires

Comment extraire une description spécifique au mouvement ? Contributions de thèse de Sabin Tiberius Strat.

Au-delà d'une description globale de l'information spatio-temporelle, une information plus spécifique liée au mouvement est souvent nécessaire pour décrire finement les actions et types de mouvement. Pour cela, les trajectoires sont alors des indices intéressants. Une trajectoire décrit le comportement spatio-temporel précis d'un motif spatial (objet, texture) sur une durée potentiellement grande. Un ensemble de trajectoires dans une séquence vidéo peut alors décrire les actions et mouvements qui la composent.
Sur ce principe, nous avons proposé des descripteurs sacs de mots capables de représenter les trajectoires de points d'intérêt. Chaque trajectoire d'un extrait vidéo est considérée comme une caractéristique locale. L'ensemble des trajectoires recueillies dans la séquence est alors agrégé dans un sac de mots. La chaine de traitement proposée est la suivante :

1. un ensemble de points d'intérêt Ω_p, constitués de motifs simples (coins) sont détectés par le détecteur de Shi-Tomasi [JT94]. Ils sont complétés par un ensemble de points Ω_r échantillonnés régulièrement sur l'image.
2. l'algorithme de flot optique de Lucas et Kanade [LK81] suit l'ensemble des points détectés dans le temps. Le mouvement de caméra est déterminé depuis Ω_r à l'aide de la méthode RANSAC [FB81].
3. Après compensation du mouvement de la caméra, les trajectoires sont calculées pour chaque point d'intérêt de l'ensemble Ω_p. Seules les trajectoires de longueurs suffisantes et liées à un mouvement d'objet sont alors conservées.
4. chaque trajectoire est décrite sous la forme d'histogrammes discutés ci-après.
5. la séquence vidéo est décrite globalement par un « sac de mots-trajectoires ».

La figure 3.5 montre les trajectoires obtenues sur une scène du corpus TRECVid qui présente un mouvement de la personne au premier plan et un autre de la caméra.
L'originalité de ce travail a été principalement liée aux différents types de description des trajectoires sélectionnées. Nous avons proposé la construction d'histogrammes de directions

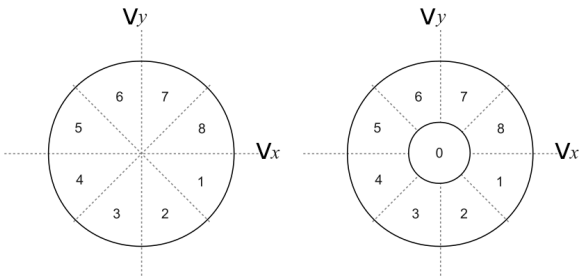


FIGURE 3.6 – Décomposition en orientations et vitesses appliquée tout au long des trajectoires sur les deux dimensions spatiales. À gauche, la vitesse est intégrée par pondération de l'orientation en chaque point échantillonné de la trajectoire. À droite, les vitesses faibles, voire nulles, sont identifiées explicitement. Repris de [Str13].

du mouvement et des accélérations des points d'intérêt. Nous nous sommes appuyés sur des décompositions en plages d'orientations spécifiques comme illustrées sur la figure 3.6. On obtient alors un ensemble de mesures décrivant les notions d'amplitude et d'accélération tout au long de la trajectoire auxquelles est ajoutée l'information de durée. Un ensemble de combinaisons de ces informations est alors défini pour réaliser des descriptions « sacs de mots trajectoires ».

Résultats et limites

Des descripteurs complémentaires ont été proposés. Un premier, générique basé sur la rétine spatio-temporelle, un second basé sur les trajectoires.

Ces contributions ont été validées dans le cadre des participations au concours TREC-Vid. La masse et l'hétérogénéité des corpus multimédias associés sont intéressantes pour étudier la performance des descripteurs proposés. Dans ce cadre, une grande variété de concepts sémantiques, jusqu'à 346, doivent être reconnus. Ils correspondent à des scènes, objets, personnes et à des actions parfois associées dont les occurrences sont rares et de proportions très hétérogènes. Nous nous plaçons donc dans un cadre d'usage réaliste avec un corpus de données typique d'une situation réelle.

Intégrés au sein du consortium de laboratoires français IRIM, nous avons pu bénéficier d'une infrastructure globale commune de haut niveau scientifique permettant alors de comparer les descripteurs de façon unifiée. La figure 3.7 montre le niveau de performance moyen des différents types de descripteurs proposés par le consortium. La métrique de performance est une précision moyenne inférée [YKA08] qui qualifie la performance moyenne de classification de chaque méthode sur l'ensemble du corpus. En nous référant à la publication IRIM de la participation 2015 du concours [Le +15], nous pouvons observer que les descripteurs très spécialisés comme ceux spécifiques aux visages et ceux spécifiques à l'analyse du mouvement sont pénalisés en moyenne (précision moyenne de l'ordre de 7% pour les trajectoires). Leur description n'est pertinente que pour certains concepts recherchés, mais n'est pas générique sur l'ensemble. Les descripteurs comme SIFT et VLAT le sont plus et donnent des niveaux de performances individuels entre 12% et 15% de précision moyenne. Enfin, des descripteurs basés sur des réseaux de neurones profonds pré-entrainés montrent des niveaux de performance nettement supérieurs, autour de 20%. Notons que l'ensemble des niveaux individuels de performance restent bas sur la tâche considérée. La fusion de ces propositions décrite dans la section suivante présente alors un intérêt certain.

Les propositions que nous avons faites sont, d'une façon générale, des descripteurs de dimension faible. La longueur des vecteurs produits pour chaque échantillon vidéo est d'une

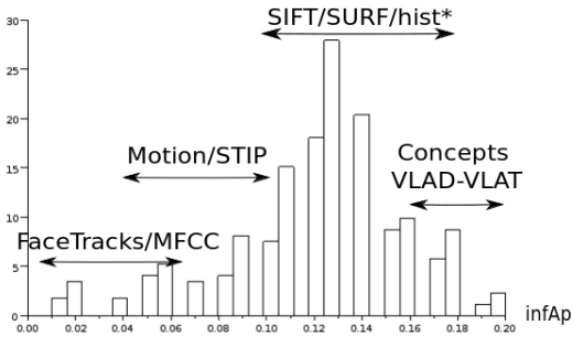


FIGURE 3.7 – Histogramme de la performance moyenne inférée pour la variété des descripteurs proposés par le consortium IRIM sur la tâche Semantic INdexing du concours IRIM. La majorité des descripteurs sont des sacs de mots classiques de la famille SIFT et ont une gamme de performances similaire. Des descripteurs de plus haut niveau (VLAD et VLAT), moins nombreux, ont une précision moyenne supérieure. Des descripteurs plus spécialisés focalisés sur l'analyse du mouvement, du son ou des visages ont une performance moyenne plus faible sur l'ensemble des concepts considérés.

taille de l'ordre de 10^3. Les approches basées sur la rétine sont relativement génériques et celles basées sur les trajectoires sont spécialisées :

— les descripteurs basés sur la rétine que nous avons proposés et diffusés dans [Str+14] présentent des niveaux de performances dans la gamme haute des descripteurs SIFT classiques (13% de précision moyenne). La taille de ces descriptions présente l'avantage d'être compacte. La figure 3.8 compare le nombre de concepts à détecter pour lesquelles différentes approches se révèlent plus pertinentes. Dans ce cas d'étude, trois descripteurs sont comparés, une méthode SIFT standard et deux approches spatio-temporelles, celle par prétraitement rétine présentée en section 2.2.1 et celle plus élaborée, présentée dans cette partie. Nous observons que les méthodes spatio-temporelles sont plus adéquates dans 80% des cas considérés. En revanche, on peut supposer une complémentarité entre les différents descripteurs permettant de dépasser les performances du meilleur descripteur.

— les descripteurs de trajectoires sont très spécialisés sur les concepts associés au mouvement. Ceux-ci représentent, dans la tâche TRECVid SIN, environ un tiers des concepts recherchés. Nous avons pu montrer la pertinence de ces descripteurs spécialisés pour des concepts spécifiques, par exemple « manger », « athlète », etc. par rapport à d'autres descriptions plus génériques [SBL14a ; SBL14b].

Néanmoins, le niveau de performance de ces descripteurs experts, spécialisés comme génériques reste trop faible pour répondre aux objectifs auxquels nous nous sommes intéressés. Le niveau d'abstraction des descripteurs proposés reste trop limité, des méthodes plus pertinentes sont donc nécessaires.

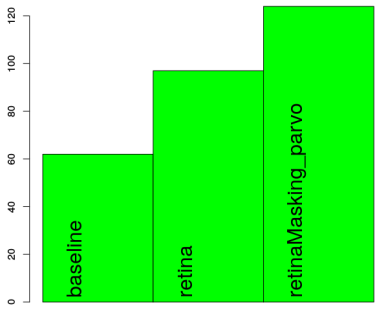


FIGURE 3.8 – Histogramme représentant le nombre de concepts pour lesquels chacun des trois descripteurs comparés donne la meilleure détection. « baseline » est basé sur SIFT seul, « retina » est le prétraitement rétine présenté au paragraphe 2.2.1 et « retinaMasking_parvo » est le descripteur décrit dans cette partie. Chacun est entrainé et évalué dans les mêmes conditions et la taille de leur description est identique, 1024. Le descripteur répondant le plus souvent est le plus générique sur la tâche TRECVid SIN.

Explicabilité et justification des décisions obtenues à l'aide des méthodes proposées

La compréhension fine des descripteurs « experts », dont ceux proposés dans ces travaux, reste limitée. Pour les approches étudiées basées sur les sacs de mots visuels, nous avons une certaine maitrise des descriptions locales. Les propriétés des opérateurs locaux comme SIFT sont connues et peuvent servir à expliquer et justifier les caractéristiques des informations extraites sur les données. La synthèse des sacs de mots visuels est plus problématique. La définition des mots visuels s'appuie sur un clustering non supervisé d'une collection de signatures locales sélectionnées dans une base d'entrainement. Les sacs de mots obtenus sur des données de test dépendent donc des propriétés des données utilisées pour le clustering, son algorithme et de la méthode d'affectation des mots visuels pour la synthèse des sacs de mots. Cet ensemble de facteurs est en général optimisé sur des critères de performance de classification sur une base de données de référence. L'étude fine du sens des mots visuels n'est en général pas considérée. Il existe tout de même des propositions permettant d'aller vers une justification des sacs de mots qui s'appuient sur des contraintes au niveau de la synthèse des mots visuels et leurs combinaisons en sacs de mots comme [NOF12]. Cependant, son application sur des corpus de données très hétérogènes est plus difficile à mettre en oeuvre. Dans ce cadre, on peut dire que ce type de méthode de description est d'un niveau de transparence faible.

3.1.2 Vers des descripteurs obtenus par apprentissage

Le transfert de méthodes de descriptions apprises sur de nouveaux problèmes est-il pertinent ? Contributions des thèses d'Amina Ben Hamida, Rizlène Raoui et de travaux avec Nicolas Voiron.

Nous nous sommes ensuite intéressés à la comparaison entre les descripteurs « experts » que nous venons de voir et les approches reposant sur des réseaux de neurones profonds optimisées par apprentissage. Dans le cadre de l'extraction de descripteurs de contenus visuels, nous nous sommes plus particulièrement intéressés au transfert d'apprentissage formalisé d'une façon générale dans [PY10]. Le transfert de réseaux de neurones consiste à exploiter des réseaux préalablement optimisés sur une tâche et une masse de données relativement proches des nouveaux objectifs, dans notre cas, la classification d'images. Ces réseaux, comme le désormais classique AlexNet [KSH12], sont constitués de nombreuses couches neuronales qui réalisent une cascade de projections non linéaires des données brutes. Les dernières couches de ces réseaux donnent une représentation de haut niveau sémantique. Lorsque le pré-entrainement a été réalisé sur des données présentant une variabilité importante mettant en valeur des transformations typiques du problème (translations, rotations, variations de luminosité, etc.), les descriptions obtenues sont potentiellement suffisamment génériques pour décrire une grande variété de nouvelles images en faisant abstraction de ces transformations tout en conservant une information discriminante sur les objets de la scène visuelle.

Nous avons pu comparer ces approches dans différents travaux :

— sur des problèmes d'analyse de vidéo multimédia dans le cadre du concours TRECVid sur des tâches d'indexation sémantique (tâche SIN) [Le +15] et de détection d'instances à partir de peu d'exemples (tâche INS) [Man+16 ; Man+17]. Les résultats montrent qu'un pré-entrainement sur une tâche de classification d'images statiques à grande variabilité de fond et d'objets comme celles du concours ILSVRC 2012 est pertinent pour produire des caractéristiques discriminantes applicables sur des contenus spatio-temporels. Ces méthodes se révèlent souvent plus performantes que les méthodes expertes développées jusque là.

— sur le thème de l'observation de la terre dans le cadre de la thèse d'Amina Ben Hamida. Pour répondre à un problème de classification d'images aéroportées, le transfert de représentations apprises sur des données multimédias s'avère plus pertinent que des approches par sacs de mots optimisées sur les données cibles[Ben+15]. Les images sont de même nature, mais les contenus et points de vue sont différents. Les représentations extraites par transfert restent les plus discriminantes avec un écart de précision significatif.

— sur un problème industriel d'analyse de tickets de caisse dans le cadre de la thèse CIFRE de Rizlène Raoui et une collaboration avec la société AboutGoods Compagny. Les tâches de détection de tickets dans des images acquises de façon non contrôlée, la détection et la classification de leur logo ont pu être obtenues par transfert de réseaux pré-entrainés [Rao+16 ; Rao+17 ; Mas+19]. Cette approche s'avère tout à fait adaptée pour répondre à des besoins industriels contraints par des taux d'erreur et des délais de mise en oeuvre réduits.

Résultats et limites

Les descriptions obtenues par l'apprentissage profond s'avèrent pertinentes dans des cas d'usage variés et dépassent les approches expertes.

Les résultats de ces travaux montrent la capacité des réseaux neuronaux à produire des descriptions de qualité et génériques pour un large éventail de problèmes. Des propriétés d'invariance (rotation, échelle, luminosité) plus robustes que les approches expertes sont obtenues. La capacité de généralisation est impressionnante comparée aux propositions antérieures et suscitent un intérêt croissant dans le monde industriel.
Les réseaux de neurones ont trouvé, dans la phase de pré-entrainement, des invariants leur permettant de répondre au problème sur lequel ils ont été initialement optimisés. Ils sont capables de générer une description similaire pour des transformations complexes des objets à reconnaitre. Plus formellement, pour un objet x générant la description $f(x)$, une transformation $x' = T(x)$ donne une description relativement proche de l'original soit $f(x') \approx f(x)$. Lors du transfert, ces invariants conservent tout de même l'information pertinente nécessaire pour répondre à de nouveaux problèmes. Ces descriptions « latentes » obtenues sont donc d'un réel intérêt. Elles montrent la capacité de systèmes ayant acquis des connaissances et des compétences sur des données et une tâche à les appliquer sur de nouvelles tâches ou domaines. En revanche, certaines similarités sur les données et les tâches entre le problème initial et le nouveau problème doivent être partagées. Cependant, la difficulté est de savoir les identifier et les transférer de façon optimale. Nous reviendrons sur ces questionnements au chapitre 4. D'une façon générale, les communautés de recherche déjà très active sur ce domaine tentent de formaliser les invariants et capacités de généralisation apportés par les réseaux de neurones profonds [Mal16].

Nous retiendrons pour cette partie l'intérêt des méthodes de descriptions par transfert de réseaux de neurones. En revanche, il ne s'agit là que d'une autre approche pour décrire les contenus qui vient compléter les méthodes classiques comme les sacs de mots que nous avons vus précédemment. À ces méthodes s'ajoutent par exemple les approches par apprentissage non supervisée comme les autoencodeurs [KW13] qui n'ont pas encore été considérés dans ces travaux. D'une façon générale, quelles que soient les méthodes utilisées, classiques ou par apprentissage, décrire un contenu à l'aide d'une seule description reste limitant lorsque le problème est difficile. Les niveaux de performance, obtenus sur la tâche TRECVid SIN, le confirment [Le +15]. La fusion d'une variété de descripteurs est alors une piste d'amélioration des performances à laquelle nous nous intéresserons dans la section suivante.

Explicabilité et justification des décisions obtenues à l'aide des méthodes proposées

Les travaux sur l'explication et la justification des réseaux de neurones profonds n'en sont qu'à leurs débuts et peu de solutions convaincantes sont proposées jusque là. Des études de sensibilité [ZF14] et des méthodes d'inversion de la fonction réseau de neurones [Sel+16] apportent des premières pistes d'intérêt. Elles permettent d'identifier, dans l'espace d'origine, l'information qui a le plus contribué à la réponse en sortie du réseau. Des approches plus générales comme [Bb18] interprètent les réseaux de neurones comme des variétés de splines qui partitionnent les données d'origine. Les réseaux adoptent alors pour chaque partition des comportements linéaires spécifiques. Dans le cas de l'extraction de description, ces approches peuvent donner des informations pertinentes pour expliquer les signatures obtenues. D'une façon plus générale, comme discuté au chapitre 4, il peut également être intéressant d'introduire des contraintes durant la phase d'entrainement des réseaux pour forcer les descriptions à se conformer à des propriétés attendues. Des propositions, par

exemple basées sur les autoencodeurs variationnels [KW13], apportent des premières pistes. Cependant, la désintrication automatique des différents paramètres contrôlant les données reste encore difficile [Loc+18].

3.2 Fusion de descripteurs

Agréger des caractéristiques pour décrire de façon plus pertinente.

Décrire, par une seule méthode, une information complexe comme les images et vidéo est, comme nous l'avons vu, potentiellement peu efficace. La difficulté est principalement liée au fossé sémantique qui sépare les données brutes de l'objectif. Les méthodes de description sont alors souvent trop spécifiques à une partie du problème réduisant ainsi leur pertinence pour ses autres composantes. Également, il peut être nécessaire de prendre en compte plusieurs sources d'informations hétérogènes afin de répondre de façon pertinente.

L'agrégation d'une variété de descriptions est alors une solution permettant de traiter ces problématiques. La littérature propose différentes méthodes de fusion que l'on peut structurer en trois approches comme proposé dans [Liu+14] :

— la fusion précoce. Dans cette configuration, des informations de bas niveau sémantique sont agrégées relativement tôt dans le processus. Le résultat de cette fusion, après une étape de description intermédiaire, est transmis à l'outil de décision final, par exemple un classifieur. Notre proposition *SIFT_multichannel_masking* en est un exemple : des descriptions locales SIFT extraites sur chaque canal de sortie de la rétine sont agrégées par concaténation avant une description par sac de mots puis une classification. Ce type de fusion est relativement simple et ne nécessite qu'une seule étape de représentation des caractéristiques agrégées. Cependant, dans le cas général, les données de bas niveau à agréger sont souvent de nature différente et de grande dimension. La synthèse d'une représentation pertinente peut alors poser problème.

— la fusion tardive. On définit des « experts », conformément à notre chapitre de livre [Str+12a], chacun capable de réaliser indépendamment la tâche. Ces experts produisent des scores, par exemple de classification, que l'étage de fusion agrège en post-traitement pour optimiser la performance globale de la tâche visée. Les méthodes peuvent être des sommes pondérées ou non, des opérations de type min, max, vote majoritaire, boosting, etc. L'objectif est de réduire le biais et la variance des erreurs de prédiction et de permettre une meilleure généralisation. Ce type de fusion présente alors un coût de calcul potentiellement réduit et permet une parallélisation des calculs. Des travaux comme [Kit+98] montrent que l'opérateur somme permet de limiter l'influence des erreurs des classifieurs individuels. Également, Ng et Kantor [NK00] montrent que des « experts » générant des prédictions dissimilaires, mais ayant des niveaux de performance moyens équivalents sont de bons candidats pour une fusion tardive.

— des méthodes de fusion hybrides. Ces méthodes réalisent une fusion à un ou plusieurs niveaux intermédiaires de la chaine de traitement. Les approches à noyaux multiples [Pin+12] en sont un exemple tout comme les réseaux de neurones profonds qui réalisent à de multiples niveaux des agrégations d'informations locales (convolutions, spatial pooling, etc.) et des agrégations de descriptions différentes par combinaison non linéaire de leurs entrées. Les limites potentielles de ces méthodes sont alors liées au coût de calcul et au nombre de paramètres à optimiser, souvent très important.

Chaque approche a donc ses avantages et inconvénients. Ainsi, le cadre applicatif, ses données et objectifs permettent de déterminer les approches les plus intéressantes. Sur les

problèmes complexes devant mettre en jeu une grande variété de méthodes de description comme nous avons pu le voir dans le concours TRECVid, ces méthodes de fusion s'unissent naturellement. Certains descripteurs réalisent des fusions précoces. Ils sont combinés par fusion tardive à d'autres descripteurs, certains pouvant également réaliser en interne des fusions hybrides comme les réseaux de neurones. Notons que les travaux d'Opitz sur les réseaux de neurones montraient déjà l'intérêt d'une fusion tardive d'une variété de réseaux de neurones [OS96]. Il montra alors qu'une meilleure généralisation est atteignable en réalisant une fusion tardive par vote majoritaire ou moyenne de réseaux donnant des réponses corrélées négativement aux autres. Ceci est ensuite confirmé par les travaux de Kantor [NK00] et montre l'intérêt de la fusion tardive y compris pour des méthodes hybrides.

Quelle que soit la méthode de fusion, elle doit, pour être efficace, prendre en compte les liens entre les méthodes de description. Elle peut s'appuyer sur des critères de similarité ou dissemblances :

— sur la nature des descripteurs. Ceux-ci peuvent s'appuyer sur des processus de description similaires, mais utiliser des paramétrages différents.
— sur leurs réponses. Les descripteurs peuvent donner des réponses corrélées, complémentaires, contradictoires.

À ces critères s'ajoutent les incertitudes liées à ces sources d'information, leurs types (mesures catégorielles ou continues), leurs intervalles, etc. On aborde là le thème de recherche spécifique de la fusion d'information. Il couvre les méthodes probabilistes et non probabilistes comme la théorie de l'évidence (ou de Dempster-Shafer) [Sha76] pour prendre en compte des interactions entre les sources. Dans un cadre de l'analyse de données massives ou « big data », le coût de calcul de ces méthodes peut cependant limiter leur usage.

Dans notre contexte, nous avons privilégié des méthodes d'agrégation et combinaison plus simples permettant le passage à l'échelle. Les thèses de Sabin Tiberius Strat et de Nicolas Voiron, nous ont permis de nous intéresser à la fusion tardive automatique de descripteurs. Nous l'avons appliquée dans deux cadres, l'indexation de concepts sémantiques dans les séquences vidéo (objets, actions, etc.), décrite dans la section 3.2.1, et la modélisation de mesures de similarité entre vidéos, présentée dans la section 3.2.2.

3.2.1 Fusion de descripteurs pour l'indexation sémantique

Comment agréger un grand nombre de caractéristiques très hétérogènes ? Contributions de thèse de Sabin Tiberius Strat.

La fusion de descripteurs complémentaires a été tout d'abord étudiée sur un problème à grande échelle dans le cadre du consortium IRIM pour la participation au challenge TRECVid. Les équipes de recherche du consortium étaient capables de fournir jusqu'à une centaine de descripteurs de nature très hétérogène : des histogrammes couleur, des sacs de mots visuels et des descriptions par réseaux de neurones. Plusieurs classifieurs étaient optimisés pour chacun d'eux, pour la reconnaissance des concepts sémantiques cible. Nous disposons ainsi d'une collection « d'experts », certains étant spécialisés comme les descripteurs de mouvement et d'autres étant plus génériques comme les sacs de mots basés sur SIFT ou les réseaux de neurones. Une fusion tardive est alors nécessaire afin d'augmenter le niveau global de performance de reconnaissance de l'ensemble des concepts cibles.

Dans le cadre de la thèse de Sabin Tiberius Strat, nous avons proposé une méthode de fusion automatique multi niveaux. Comme illustré sur la figure 3.9, on définit une première étape de fusion par différentes méthodes classiques (Adaboost, somme pondérée, sélection du meilleur expert par concept sémantique cible) et une méthode originale par fusion agglomérative. Un dernier étage de fusion combine ces fusions intermédiaires pour donner la prédiction finale.

Notre fusion par clustering agglomératif [Str+12a] s'appuie sur les travaux de Ng et

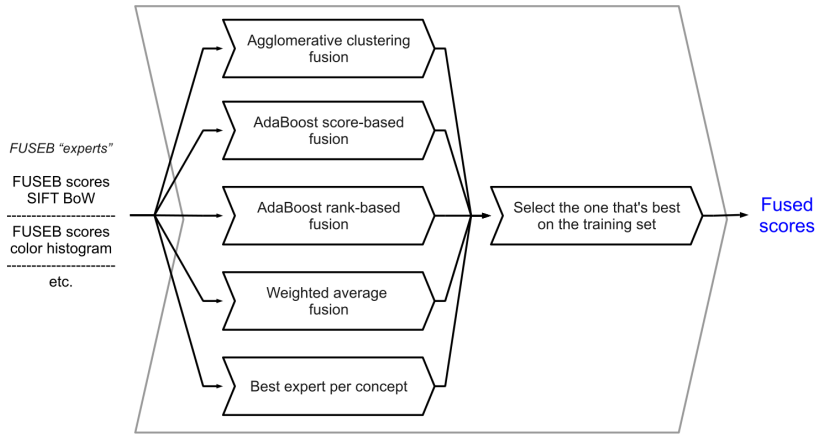


FIGURE 3.9 – Stratégie globale de fusion développée dans le cadre de la thèse de Sabin Tibérius Strat [Str13].

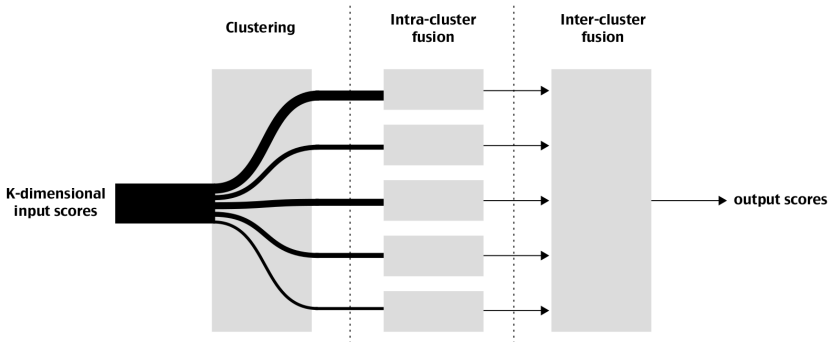


FIGURE 3.10 – Stratégie de fusion agglomérative développée dans le cadre de la thèse de Sabin Tibérius Strat [Str13].

Kantor [NK00]. On cherche dans cette approche à construire et à agréger une collection d'experts forts et dissimilaires. Comme illustré sur la figure 3.10, nous définissons alors une fusion tardive en trois étapes :

— détection et regroupement des « experts » similaires. Cette étape se fait par analyse des corrélations entre les réponses des experts. Cette étape inclut également une sélection des experts donnant un niveau de performance minimum suffisant pour un concept donné.
— fusion intragroupe par moyenne. Les membres de chaque famille d'experts similaires sont agrégés. On obtient alors une collection de descripteurs de plus haut niveau ayant des niveaux de performance similaires et décorrélés.
— fusion finale inter-groupe par somme pondérée afin d'obtenir les scores de classification finaux.

Résultats et limites

La fusion tardive de score de classifieurs est pertinente. Prendre en compte leur variété permet de dépasser le meilleur classifieur.

Dans le cadre du consortium IRIM et la participation au concours TRECVid, nous avons comparé différentes méthodes de fusion basées sur la même diversité « d'experts ». Les méthodes de fusion automatiques développées dans nos travaux ont pu être comparées à une méthode hiérarchique manuelle « experte » [Bal+12] qui se base sur une connaissance fine des propriétés de chaque descripteur. Les résultats de ces travaux sont publiés dans le chapitre de livre [Str+12a].

Lorsque l'on fait la fusion d'approches de type sac de mots générique et spécialisé, par exemple les descripteurs SIFT basés sur le modèle de rétine que nous avons développée et les descripteurs de trajectoires, nous obtenons un gain de l'ordre de 30% par rapport à un l'usage d'un seul bon expert.

Comparés à une méthode de fusion hiérarchique manuelle experte, les niveaux de performance se rapprochent, mais restent au bénéfice de la méthode manuelle. La méthode automatique la plus efficace est celle basée sur les méthodes de boosting. Ces observations sont intéressantes, car elles montrent que des méthodes automatiques plus efficaces sont encore à développer dans ce contexte pour atteindre voire dépasser les méthodes manuelles.

Si l'on s'intéresse maintenant à la fusion de descripteurs avancés de types différents et de niveaux de performance également très différents, par exemple, les « sacs de mots » experts et les réseaux de neurones profonds optimisés par apprentissage, nos dernières participations à la tâche d'indexation sémantique montrent des éléments intéressants. Dans [Le +15], nous observons que la fusion d'experts basés uniquement sur les réseaux de neurones n'est pas aussi performante que la fusion de l'ensemble des descripteurs à disposition, y compris les « méthodes expertes » ayant des niveaux de performance moyens très inférieurs. La fusion apporte donc systématiquement un gain, dans le cadre expérimental considéré. La prise en compte de l'ensemble des experts donne ainsi une augmentation relative de performance de 15% par rapport à celle obtenue par le meilleur expert, basé sur les réseaux de neurones. Un compromis entre performance et coût de calcul est donc à prendre en compte.

Retenons que les approches de fusion tardive proposées exploitent la complémentarité entre les différents experts et permettent d'obtenir des gains de performance intéressants. L'intérêt principal est une meilleure généralisation du système d'indexation sur des concepts et contenus très hétérogènes. Les observations faites dans la littérature [NK00 ; Kit+98 ; OS96] se confirment donc pour des variétés de descripteurs et corpus de données différents, la fusion tardive d'experts aux sensibilités différentes est efficace.

Explicabilité et justification des décisions obtenues à l'aide des méthodes proposées

Les méthodes de fusion proposées sont finalement des approches « expertes », car les critères d'agrégation sont clairement définis. Ils sont explicables, car il est possible de révéler les sélections, agrégations et pondérations des experts réalisées par ces approches. La justification est également assurée, car les critères d'agrégation et de sélection sont maitrisés. En revanche, leurs résultats s'appuient sur une agrégation d'informations provenant de systèmes amont potentiellement moins transparents comme discuté précédemment.

3.2.2 Fusion de descripteurs pour la modélisation de similarités entre vidéos

Comment agréger des mesures objectives de dissemblance pour modéliser la perception des distances ? Contributions de la thèse de Nicolas Voiron.

Modéliser la perception des systèmes biologiques comme celles de l'humain est un vieux problème. On peut distinguer différents niveaux de modélisation. On trouve tout d'abord les modélisations « bas niveau » relatives à notre sensibilité aux informations captées simples (texture, couleur, mouvement, etc.). Dans ce cadre, nous avons vu la modélisation de la rétine pour répondre à une problématique de prétraitement au chapitre précédent. On trouve également des modélisations de plus haut niveau cognitif (organisation de l'information, comportements, etc.). Dans cet autre cadre, nous nous sommes intéressés, par exemple dans ce chapitre, à la modélisation de l'organisation de l'information. La catégorisation d'échantillons, la reconnaissance de sa classe prédéfinie par l'humain est une solution que nous venons de voir. Elle débouche sur des applications de type système d'indexation ou de classification automatique. Une autre approche à laquelle nous nous sommes intéressés consiste à modéliser la perception des distances ou dissemblances entre les échantillons. La notion de catégories nommées et prédéfinies n'apparait alors plus directement et l'on s'oriente vers une organisation des contenus selon une mesure d'éloignement (distance ou dissemblance) entre chaque élément. Une finalité est la possibilité d'identifier des groupes d'échantillons ressemblants, des « clusters ». Ce type de classification automatique n'introduit pas de connaissance sur les catégories à identifier. Elle est plus clairement identifiée sous le terme « clustering » en langue anglaise.

C'est dans ce cadre que les travaux de thèse de Nicolas Voiron se sont intéressés à la modélisation de la dissemblance perçue par les opérateurs humains par fusion de mesures de dissemblances calculées, objectives. La finalité est de maximiser la concordance, au sens du tau de Kendal, entre les mesures de référence et leur modélisation. Partants de variables explicatives, nous avons proposé un ensemble de mesures de dissemblances normalisées. La modélisation de la mesure d'éloignement de référence est alors réalisée par une technique de sélection et de fusion s'appuyant sur une corrélation de rang des mesures d'éloignement. Des corrélations comme le tau de Kendal permettent d'identifier les mesures objectives concordant le plus avec la référence. En revanche, selon leur type, différentes problématiques apparaissent. Les mesures discrètes peuvent être très pertinentes, mais ne permettent pas toujours de différencier certains échantillons, car un certain nombre d'ex aequo peuvent apparaitre. Les mesures continues n'ont en général pas ce problème, mais leur pertinence peut être renforcée par des mesures discrètes. Enfin, le coût de calcul peut devenir trop important si une approche force brute est employée pour trier l'ensemble des mesures. Nous avons alors proposé une méthode permettant une sélection itérative des meilleures mesures permettant d'augmenter le taux de concordance global entre les mesures fusionnées et la référence à modéliser.

Le point fort de ces travaux est la possibilité de manipuler des variables explicatives de nature et d'ordre de grandeur différents. Nous avons pour cela proposé l'utilisation du coefficient gamma de Goodman et Kruskal décrit par Podani [Pod97] afin de quantifier et de comparer la qualité des ordonnancements entre chaque mesure de dissemblance calculée et celles de référence donnée par l'humain. Nous avons utilisé cette mesure de corrélation dans une approche basée sur un tri lexicographique des mesures de dissemblance. L'algorithme proposé est le suivant et décrit dans l'algorithme 1 :
Partant d'un ensemble Ω de paires d'échantillons, on considère pour chacune les mesures de dissemblances d_h données par une expertise humaine ainsi que les mesures de dissemblances calculées d_i d'un ensemble initial D. On définit enfin la liste ordonnée des dissemblances D_f, initialement vide, qui sera utilisée par le modèle. Itérativement, on calcule sur Ω_r, une

copie de l'ensemble initial Ω, les gamma de Goodman et Kruskal entre la référence humaine d_h et chaque dissemblance d_i de l'ensemble D. On identifie la dissemblance calculée d_s ayant le plus grand gamma avec la référence c.-à-d. celle ayant le plus grand accord. Elle est alors ajoutée à D_f et retirée de D. Elle est également utilisée pour classer toutes les paires d'échantillons possibles qui sont alors éliminés de Ω_r. Cette stratégie est appliquée itérativement jusqu'à ce que toutes les dissemblances soient utilisées ou jusqu'à l'obtention d'un classement total (sans ex aequo). On calcule alors la mesure de dissemblance finale fusionnée d_f, par normalisation, en divisant tous les rangs de classement par le nombre total de paires.

Input: Ω, ensemble des paires d'échantillons du corpus de données.
Input: d_h, mesures de dissemblance de référence sur Ω à modéliser.
Input: $D = \{d_1, d_2, ...d_n\}$, ensemble des mesures de dissemblance calculées sur Ω.
Result: d_f, dissemblance fusionnée

$\Omega_r = \Omega$
$D_f = \emptyset$, liste des dissemblances sélectionnées.
while Ω_r *non vide ET* D *non vide* **do**
  $L_\gamma = gamma(D, d_h)$; ▷ mesure des gamma entre d_h et chaque élément de D
  $d_s = argmax(L_\gamma)$; ▷ détection de la mesure d_s donnant le plus grand gamma
  $D.remove(d_s)$;
  $D_f.append(d_s)$; ▷ d_s est déplacée dans la liste des dissemblances sélectionnées
  $\omega = sorted(\Omega_r, d_s)$; ▷ détection des paires classées par d_s
  $\Omega_r.remove(\omega)$; ▷ les paires classées par d_s sont enlevées de l'ensemble Ω.
end
$d_f = rank(\Omega, D_f)/card(\Omega)$

Algorithm 1: Modélisation d'une dissemblance de référence par fusion itérative de mesures de dissemblance calculées. Travaux de thèse de Nicolas Voiron [Voi15].

Résultats et limites

Nous avons proposé une approche itérative par corrélation de rang limitant les coûts de calcul.

Nous nous sommes intéressés à la modélisation de la perception des distances par des opérateurs humains, entre des vidéos. Mesurer des distances entre films complets est a priori plus complexe que mesurer des distances entre les plans de ces mêmes films comme nous l'avons fait au paragraphe 3.1. Nous avions constaté la nécessité d'agréger une variété de descripteurs pour catégoriser les plans. Cela est également nécessaire pour la catégorisation d'une séquence complète composée de multiples plans. Les descripteurs sont potentiellement de nature différente et peuvent intégrer des informations plus globales par exemple l'année de production, le genre, le type de public visé, etc. Le problème étant difficile, nous nous sommes placés dans un contexte spécifique et avons considéré la mesure de dissemblance entre les films du festival international du film d'animation présenté chaque année à Annecy et organisé par l'organisme CITIA (Image & industries créatives) [2]. Les films sont de nature très hétérogène, mais présentent une certaine régularité : ce sont des films d'animation uniquement. Ils sont élaborés par différentes techniques, de l'image de synthèse au dessin au crayon et sont de durée relativement courte. Ces propriétés ainsi que d'autres variables explicatives (le genre, le type de public et le pays de production, etc.) sont connues

2. CITIA : https://www.citia.org/

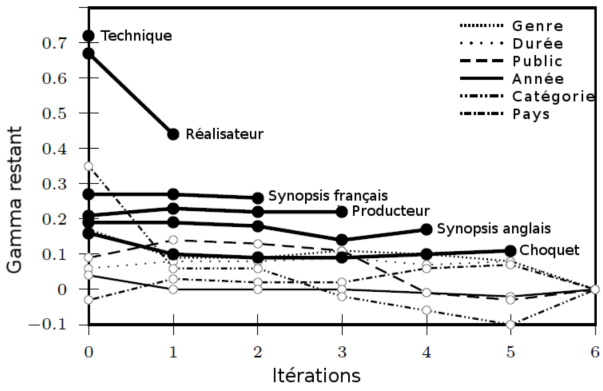


FIGURE 3.11 – Évolution du gamma de Goodman et Kruskal entre la référence humaine et chaque mesure de dissemblance sur l'ensemble des paires d'échantillons restant à trier au fil des itérations. L'interruption des courbes montre à quelle itération une certaine dissemblance calculée a été intégrée à la collection qui sera utilisée pour la fusion. Reprise de [VBL12].

pour chaque film ce qui nous donne un référentiel suffisant pour mener une étude sur leur structuration.

Dans les travaux publiés dans [VBL12], nous avons pu définir 12 mesures de dissemblance depuis les variables explicatives à disposition sur la base de données CITIA considérée. Certaines sont discrètes, car s'appuient sur des données catégorielles (le type de public, le genre, etc.), d'autres sont continues, car se basent sur des mesures plus fines par exemple l'écart temporel entre les dates de deux productions de films, ou une distance entre descriptions de leur contenu par sacs de mots, etc.

La figure 3.11 montre l'évolution du gamma au fil des itérations de la méthode. Les valeurs des gamma restant pour les descripteurs non encore sélectionnés sont calculées sur les paires jusque là non séparables par les descripteurs déjà retenus. On observe que certains descripteurs initialement corrélés avec la référence sur le jeu complet de paires d'échantillons, mais également fortement corrélés à certains descripteurs sélectionnés aux étapes précédentes cèdent leur rang à des descripteurs plus discriminants sur les paires restantes. La méthode de fusion proposée se rapproche ainsi des propositions de Kantor [NK00] et montre l'intérêt d'associer des descripteurs aux comportements dissimilaires.
La méthode proposée présente certains avantages :

- les variables explicatives peuvent être de nature et d'échelles différentes.
- elle permet de ne sélectionner que les mesures de dissemblance utiles.
- elle permet une réduction significative du nombre de tris à réaliser par rapport à une méthode force brute. Cela est permis par l'approche itérative qui, au fil des itérations, réduit le nombre de dissemblances à comparer et le nombre de paires d'échantillons à trier.

Les limites sont potentiellement liées au coût de calcul. Bien que la méthode proposée réduise significativement les coûts par rapport à une approche force brute, l'application de la méthode sur de grands corpus de données et une grande variété de descripteurs peut être limitée. Par exemple dans des situations semblables à celles du concours TRECVid, le nombre d'échantillons se compte en millions. Appliquer ces algorithmes de tri en prenant en compte des dizaines de descripteurs exige alors des ressources de calcul importantes.

Explicabilité et justification des décisions obtenues à l'aide des méthodes proposées

La fusion réalisée est « experte », explicable et justifiable. L'utilisateur a accès à la liste des descripteurs pris en compte pour la prédiction finale et à leur pondération respective. La sélection des descripteurs est justifiée et expliquée par la comparaison des concordances entre ceux-ci et la référence à modéliser (gamma de Goodman et Kruskal).

3.3 Synthèse et perspectives sur la description et fusion de descripteurs

Les travaux vont vers l'extraction de caractéristiques apprises. La fusion de classifieurs est toujours nécessaire.

La synthèse de descriptions intermédiaires entre les données brutes et l'étape finale de prédiction est une étape critique dans les chaines de traitement de données. Elle permet le franchissement du fossé sémantique entre l'information brute et la réponse attendue, souvent de haut niveau sémantique.

Nous nous sommes intéressés à la synthèse de descripteurs par une variété d'approches :

— des méthodes intégrant l'information spatio-temporelle pour l'indexation des vidéos. Partant d'une structure standard de description basée sur les sacs de mots, nous avons proposé des stratégies de fusion précoce de l'information temporelle avec l'information spatiale par l'usage d'un modèle numérique de rétine spatio-temporelle. Nous avons également proposé des descripteurs spécifiquement liés au mouvement pour la description de la trajectoire des objets dans la scène.

— des méthodes de description de haut niveau sémantique par l'usage de réseaux de neurones pré-entrainés. Nous avons pu mesurer la pertinence des descriptions fournies par des réseaux transférés sur de nouveaux problèmes.

Ces approches permettent d'obtenir une représentation compacte, un code, des données initiales. Elles présentent, chacune, certaines propriétés d'invariance. Cependant, dans des situations difficiles comme l'indexation vidéo, une seule méthode, même parmi les plus avancées, comme les réseaux de neurones, ne généralise pas suffisamment pour donner une réponse pertinente dans une grande variété de situations.

Il est alors nécessaire de réaliser une fusion de descripteurs complémentaires. L'objectif est alors d'associer les propriétés d'invariance de chaque descripteur afin d'augmenter la capacité de généralisation à une variété plus large de situations. Nous avons proposé un ensemble de méthodes de fusion tardives agrégeant les prédictions d'experts reposant sur chaque méthode de description. Ces approches présentent finalement une structure commune :

— une étape de synthèse d'experts de haut niveau, pertinents, mais dont les réponses sont dissimilaires. Cela est réalisé par des méthodes d'agrégation des descripteurs corrélés ou par sélection d'un unique représentant.

— une fusion des experts de haut niveau pour générer la prédiction finale.

Les méthodes de fusion proposées permettent alors une meilleure généralisation des prédictions. Nous avons alors observé l'intérêt des descriptions basées sur des méthodes par apprentissage. Ces méthodes ont des capacités de généralisation individuelles initialement importantes qui dépassent en général les « méthodes expertes » classiques, même fusionnées. En revanche, la fusion de ces deux types de descriptions permet souvent d'augmenter encore le niveau de performance.

Du point de vue de la capacité à expliquer et justifier les réponses des systèmes proposés, les méthodes de fusion proposées sont « expertes » et présentent un niveau de transparence

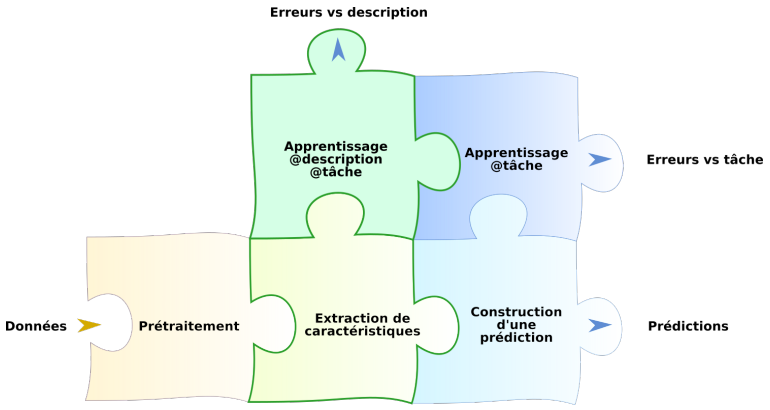


FIGURE 3.12 – Perspectives de travail sur la description des données : décrire les données par des méthodes par apprentissage multitâche. L'optimisation globale du système est complétée par une optimisation sur la synthèse des descriptions intermédiaires s'appuyant sur ses propres critères.

intéressant. À l'inverse, les méthodes de description par sacs de mots et par apprentissage restent des boites noires. Leurs réponses sont très dépendantes des propriétés des jeux de données ayant servi à leur construction et leur adéquation avec les données d'utilisation. Leur fonction de transformation des données en une description est non linéaire et reste difficile à analyser, quelle que soit l'approche étudiée. Cette difficulté devient un thème de recherche à part entière d'intérêt scientifique et sociétal.

Pistes de recherche

L'objectif est d'introduire des critères d'optimisation spécifiques à l'extraction de caractéristiques en complément du critère d'optimisation de la tâche.

Nous avons vu que les descriptions obtenues par des systèmes optimisés par apprentissage donnent des niveaux de performance très intéressants. Leur capacité de généralisation réduit l'intérêt des méthodes expertes souvent trop spécialisées et moins à même de s'adapter à une grande variété de situations. Les réseaux de neurones en particulier intègrent la notion de fusion hybride. Un premier avantage est tout d'abord la capacité à agréger des données hétérogènes au sein d'une même architecture. Un exemple fort est par exemple l'approche de [Kai+17] qui propose la synthèse d'une description commune de différentes modalités pour réaliser différentes tâches. On peut également interpréter les réseaux de neurones comme une grande variété de sous-systèmes qui sont agrégés à différents niveaux de l'architecture [Gom+19a]. Les réseaux de neurones présentent donc un réel intérêt, mais une de leurs limites est la justification et l'explication de leurs prédictions.

Pour la synthèse de description de données, une piste de recherche qui me semble intéressante est un travail sur des méthodes de description pertinente pouvant être utilisées pour justifier les prédictions des systèmes qui les utilisent. Une piste générale est envisagée : elle consiste en l'introduction de contraintes spécifiques pour l'optimisation des méthodes de descriptions. Ceci constitue alors une nouvelle tâche à associer à l'optimisation globale du système comme illustré sur la figure 3.12. Deux axes complémentaires peuvent être envisagés :

— une approche supervisée : pour une variété de problèmes, il est intéressant d'obtenir des représentations intermédiaires ayant un sens. Par exemple, [Xu+18] génère des prédictions intermédiaires interprétables recombinées ensuite pour répondre à un

problème de plus haut niveau. Ce type d'approche est déjà engagé, par exemple dans les travaux de thèse de Mikael Jacquemont sur un problème d'astrophysique. La prédiction de certaines grandeurs des phénomènes observés, comme l'énergie de rayonnements gamma, dépend de grandeurs intermédiaires liées à l'observation, par exemple l'orientation par rapport à la caméra et au point d'impact du rayonnement au sol. Il est alors intéressant de modéliser explicitement ces mesures intermédiaires pour faciliter la régression des paramètres d'intérêt.

— une approche non supervisée : dans une vaste majorité de situations, il est difficile de connaître les caractéristiques qui contrôlent les données (la forme, la taille des objets, leur nature, les propriétés du contexte, etc.). Il est alors intéressant de les découvrir automatiquement. Les approches du type autoencodeur variationnel [KW13] apportent des solutions intéressantes. Elles permettent de découvrir des encodages de propriétés des données et de les modifier pour générer de nouvelles données personnalisées [LSW15 ; Hou+16]. En revanche, le problème général de l'identification de chaque propriété des données sans supervision ou connaissances a priori reste un problème difficile comme discuté dans [Loc+18]

Ces deux approches sont complémentaires. Comme nous l'avons évoqué au chapitre 2, les modèles physiques ou les contraintes issues de l'expertise sont parfois incomplets ou issus de connaissances partielles sur le problème. L'approche supervisée pour l'application de contraintes présente alors les mêmes limites et il peut être intéressant de compléter les descriptions supervisées par d'autres, obtenues de façon non supervisée. On ajoute ainsi de la flexibilité au système global. L'étude des propriétés de l'information transitant par le canal non supervisé et sa complémentarité avec le canal supervisé pourrait alors amener à une meilleure compréhension du modèle physique sous-jacent.

Les cadres applicatifs sont nombreux, du multimédia à la physique et l'imagerie satellitaire. Chacun d'eux amène également à s'intéresser à des problématiques connexes comme l'informatique décentralisée ou « Edge computing ». Les données captées sont massives et potentiellement porteuses d'informations sensibles, voire confidentielles, et il peut être intéressant de rapprocher la synthèse de leur description vers la source avant de les faire cheminer sur les réseaux vers le système de prédiction centralisé. Ce concept présente des intérêts importants qui incluent la réduction des coûts de transport de l'information vers le système central par le cheminement de descriptions compactes. Il contribue également à la préservation de la confidentialité par la transmission encodée de l'information seulement liée à la tâche.

Chapitre 4

Apprentissage automatique et réseaux de neurones profonds

Mes travaux se sont focalisés sur l'apprentissage pour les réseaux de neurones.

Apprendre, inculquer, raisonner, optimiser ... trouver la relation entre l'information source et l'objectif repose souvent sur l'identification de processus voire de règles dans une phase d'apprentissage, depuis des données d'exemple. Ces processus sont associés à des systèmes capables d'intégrer cette relation pour ensuite l'appliquer sur de nouvelles données à un certain degré de précision. On parle alors de l'apprentissage automatique, un thème scientifique majeur ouvert depuis quelques décennies, qui couvre un grand nombre d'approches et de problèmes. Sa naissance est fortement associée aux travaux fondateurs d'Alan Turing sur la machine universelle en 1936 [Tur37] puis sur la notion d'apprentissage automatique proposée en 1950 [Mac50]. Dans la même période, les travaux en neuroscience de Mac Cullogh de 1943 [MP43] initient la compréhension et la modélisation des réseaux de neurones. Ces travaux ouvrirent alors la voie aux travaux sur l'intelligence artificielle mise sur le devant de la scène, par notamment Mac Carthy et Minsky, au milieu des années 1950.

L'apprentissage automatique est un apprentissage statistique, mis en oeuvre de différentes façons selon la disponibilité de données d'entrainement et de connaissances qui leur sont ou non associées. On parle alors d'apprentissage supervisé, non supervisé et d'apprentissage par renforcement sur lesquels nous reviendrons. L'apprentissage, d'une façon générale, vise à répondre à un problème souvent mal posé qui est celui d'identifier à partir d'un ensemble d'observations Ω, la loi f à laquelle les données obéissent sur des problèmes de classification, de régression et plus généralement de prédiction. L'ensemble des observations ne permettant généralement pas d'identifier f sur l'ensemble des cas possibles, on recherche alors un estimateur paramétrique $\hat{f}_\theta$ le plus proche possible de f. L'apprentissage consiste alors à chercher le paramétrage θ de l'estimateur minimisant un critère de coût ou risque L calculé sur l'ensemble des observations Ω entre les prédictions données par l'estimateur et la réponse mesurée de f comme montré sur l'équation 4.1.

$$\begin{aligned} L(\theta) &= \frac{1}{|\Omega|} \sum_{x \in \Omega} l\left(f_\theta(x), f(x)\right) \\ \hat{f}_\theta &= \underset{f_\theta}{\operatorname{argmin}}\, L(\theta) \end{aligned} \tag{4.1}$$

Cet estimateur $\hat{f}_\theta$ est une fonction paramétrique qui peut être basée sur des arbres, des

méthodes à noyaux, des réseaux de neurones, etc. La méthode considérée doit avoir une flexibilité suffisante pour s'adapter à tous les cas d'exemples de Ω et devrait idéalement avoir les capacités suffisantes pour permettre d'interpoler entre chacun d'eux.

Le processus d'optimisation ou d'entrainement consiste donc à minimiser le critère L. Une variété de méthodes existe, comme les méthodes de Newton, la descente de gradient... Nous nous focaliserons sur cette dernière, actuellement la plus utilisée pour résoudre des problèmes à grand nombre de paramètres, notamment dans le cas des réseaux de neurones, tout en présentant un coût de calcul plus réduit [GBC16]. La descente de gradient est itérative et consiste à confronter l'estimateur dans son état courant à des lots d'échantillons et de le corriger dans la direction permettant de réduire ses erreurs. L'estimateur peut ainsi être confronté au jeu complet des données d'entrainement Ω ou à des sous-ensembles, souvent nommés « batchs », de taille $b << card(\Omega)$. La méthode de descente de gradient suit alors la formulation de base donnée par l'équation 4.2. À chaque itération t, les estimations sont mises à jour pour l'itération suivante en les ajustant en fonction du gradient de l'erreur du modèle sur les échantillons du batch pondéré par un coefficient λ fixant, en quelque sorte, la vitesse de l'apprentissage. Un aspect stochastique peut être amené par le tirage des échantillons constituant le batch.

$$\theta_{t+1} = \theta_t - \lambda \left(\frac{1}{b} \sum_{i=1}^{b} \nabla_{\theta_t} l \left(\hat{f}_{\theta_t}(x), f(x) \right) \right) \tag{4.2}$$

La difficulté associée est la construction de l'estimateur. Quelle structure, quel nombre de paramètres doit-il avoir ? Nous nous sommes intéressés aux réseaux de neurones profonds. Ces approches présentent des capacités d'adaptation à une grande variété de problèmes. Ce sont des systèmes en général sur-paramétrés qui, entrainés avec une masse de données conséquente et variée, ont la capacité de généraliser leurs prédictions à une grande variété d'exemples de tests. Leurs taux d'erreurs peuvent être bien plus faibles que les méthodes phares présentées jusque là, par exemple, les méthodes « expertes » de la famille des sacs de mots présentées au chapitre 3 sur les problèmes de reconnaissance d'image. En revanche, ceci nous amène à un problème d'optimisation non convexe pour lequel de nombreuses solutions existent. Ben Arous [Cho+15] montre que le nombre de minina locaux augmente de façon exponentielle avec le nombre de paramètres. Dans la phase d'apprentissage, la quête d'un bon minimum est donc délicate. Cependant, le minimum global sur le jeu d'entrainement ne doit pas être le critère principal, car ce minimum amène généralement au sur-apprentissage.

Au travers des travaux de thèse d'Amina Ben Hamida, Rizlène Raoui, Emna Amri, Nicolas Voiron et Mikaël Jacquemont, nous nous sommes intéressés à la structuration des réseaux de neurones et à leur méthode d'optimisation. Les contributions portent tout d'abord sur des propositions de méthodes d'entrainement à coût réduit présentées dans la section 4.1. Un second volet de travaux présenté dans la section 4.2 s'intéresse aux architectures de réseaux légères en termes de nombre de paramètres permettant de réduire également l'empreinte du coût de calcul.

4.1 Apprentissage

Une variété d'approches selon l'état des connaissances sur les données.

D'une manière générale, l'apprentissage peut se décomposer en plusieurs approches qui répondent à différentes problématiques :

- l'apprentissage supervisé pour lequel une base de données d'entrainement annotée présente une collection de couples $(x, f(x))$. La relation entre ces couples de référence est alors estimée par $\hat{f}_\theta$. La principale difficulté de cette approche est la disponibilité d'une base de données qui couvre l'ensemble des situations auxquelles sera confronté l'estimateur sur des données de travail. La littérature et les travaux présentés au chapitre 3, paragraphe 3.1.2 rapporte l'intérêt des réseaux de neurones par leur capacité à apprendre, de façon souvent supervisée des caractéristiques transférables et pertinentes d'un domaine à un autre.
- l'apprentissage non supervisé pour lequel un système est optimisé sur des données pour lesquelles aucune annotation n'est disponible. Les objectifs dans ce cas de figure sont différents. Ils peuvent consister en l'extraction de caractéristiques générales propres aux données, en la modélisation de leur distribution, leur partitionnement, etc. Pour les réseaux de neurones, les approches de type autoencodeur ont constitué l'approche principale qui vise à trouver des couples de transformations $f : x \to z$ $g : z \to \hat{x}$ avec $\hat{x} \sim x$ permettant de trouver des représentations pertinentes z des données. L'apprentissage de représentation de niveau d'abstraction croissant par une cascade d'autoencodeurs [Ben+07] a été une méthode de référence. Des approches génératives plus récentes telles que les autoencodeurs variationnels [KW13] et les méthodes adverses [Goo+14] permettent d'apprendre directement des représentations à haut niveau d'abstraction.
- l'apprentissage par renforcement pour lequel l'apprenant interagit avec son environnement pour atteindre son objectif. L'optimisation s'appuie sur des pénalités et bonifications associées à la pertinence des séquences d'actions réalisées. On ne parle plus directement de base d'apprentissage, mais d'expériences par essai/erreur dans l'environnement de travail [SB+98].

Ces méthodes d'apprentissage peuvent s'associer pour répondre à de nouvelles problématiques. Par exemple, l'apprentissage semi-supervisé, l'apprentissage actif, l'apprentissage multitâche, etc. L'ensemble de ces approches se rassemble autour d'une idée commune que nous avons évoquée dans les chapitres précédents, trouver des invariants dans les données, tout en conservant l'information permettant de répondre au problème posé.

Sur ce thème de l'apprentissage, nos travaux ont été déclinés selon deux orientations complémentaires. La première, présentée dans la section 4.1.1, porte sur l'apprentissage supervisé dans des situations pour lesquelles peu de données d'entrainement sont à disposition. La seconde, présentée dans la section 4.1.2, concerne l'association de l'apprentissage semi-supervisé actif à des méthodes de propagation automatique de la connaissance.

4.1.1 Apprentissage supervisé depuis peu d'exemples

Un ensemble de méthodes qui s'appuie souvent sur des prérequis.

L'apprentissage supervisé est un cas favorable appliqué sur une base de données composée d'exemples pour lesquels la réponse attendue est parfaitement connue. C'est par cette approche que les réseaux de neurones sont revenus sur le devant de la scène en 2012, avec les travaux de Krizhevsky [KSH12] et la structure « profonde » en huit couches AlexNet. De grands jeux de données comme la base ImageNet [1], composés de millions d'images représentant quelques milliers de catégories d'objets dans des situations très variées ont alors

1. http://www.image-net.org/

permis d'entrainer des réseaux composés de millions de paramètres répartis. Sur les architectures neuronales récentes, les neurones se répartissent sur des dizaines voir des centaines de couches de neurones.

Cependant, dans un cas d'étude réel, cette quantité de données annotée peut s'avérer faible. Il est alors nécessaire de définir des stratégies d'apprentissage spécifiques. La littérature montre différentes approches dont les travaux présentés dans [Sch+19] proposent la structuration suivante :

- l'apprentissage de métriques. On entraine des réseaux à estimer une distance entre des échantillons. Cette distance, potentiellement de haut niveau sémantique, peut être donnée par des opérateurs humains, des experts. Les réseaux profonds siamois [CHL+05] sont une référence classique qui permet dans une phase préalable d'entrainement sur un grand jeu de données de définir des distances entre échantillons sur des critères de similarité/dissemblance. Ces réseaux peuvent alors être utilisés dans un nouveau domaine en présentant de nouveaux exemples et en mesurant leurs distances [KZS15]. Le paragraphe suivant 4.1.2 propose à ce titre une méthode permettant de réduire les coûts d'obtention de la base d'entrainement.
- le « Meta learning ». Cette approche vise à optimiser un système en le confrontant à différentes tâches et différents jeux de données, potentiellement de petites tailles. Cette approche permet alors, de façon itérative, de transférer la connaissance d'un problème à un autre et de l'adapter, voire de l'enrichir. Dans ce cadre, le transfert d'apprentissage (« Transfer learning »), présenté au chapitre 3 au paragraphe 3.1.2, est une forme de Meta learning simplifiée ne s'appuyant que sur une seule étape de pré-entrainement pour laquelle une tâche complexe est apprise sur un jeu de données à grande variabilité. Ce réseau est ensuite transféré partiellement sur un nouveau problème pour lequel des données plus spécifiques, mais moins nombreuses sont à disposition. La première optimisation permet de synthétiser des descripteurs suffisamment pertinents et génériques pour être capables de décrire de façon pertinente un nouveau domaine de données et s'adapter à de nouvelles tâches par une nouvelle phase d'entrainement minimale.
- l'augmentation de données. Cette approche vise à apprendre à générer de nouveaux exemples pertinents pour enrichir la base de données d'entrainement. Le travail de Chen [Che+19] est un bon exemple qui présente une méthode de génération d'exemples d'entrainement obtenus par l'apprentissage de déformations pertinentes des échantillons d'entrainement originaux. Les approches génératives [Goo+14 ; KW13] sont également possibles, mais requièrent en revanche une collection suffisamment riche, variée et équilibrée d'exemples.
- les méthodes sémantiques. Ces dernières permettent un entrainement depuis peu d'exemples en s'appuyant sur des métadonnées : attributs, nom de catégorie, description textuelle... [Sch+19] montre par exemple l'intérêt d'entrainer des réseaux de neurones à combiner des descriptions visuelles et textuelles pour répondre à plusieurs tâches et générer ainsi des représentations génériques permettant un apprentissage final depuis un nombre très limité d'exemples.

Pour attaquer un nouveau problème, la plupart de ces méthodes peuvent partir d'un réseau préalablement entrainé sur d'autres données issues d'un domaine proche. Cependant, certaines situations s'appuient sur des données spécifiques qui ne le permettent pas (un nouveau format de données, un nouveau capteur, etc.). Il est alors nécessaire d'entrainer directement des réseaux sur des jeux de données spécifiques. Cette situation est classique en apprentissage profond lorsque la base de données est conséquente. Elle est en revanche plus originale lorsque très peu de données sont disponibles. Des stratégies d'apprentissage

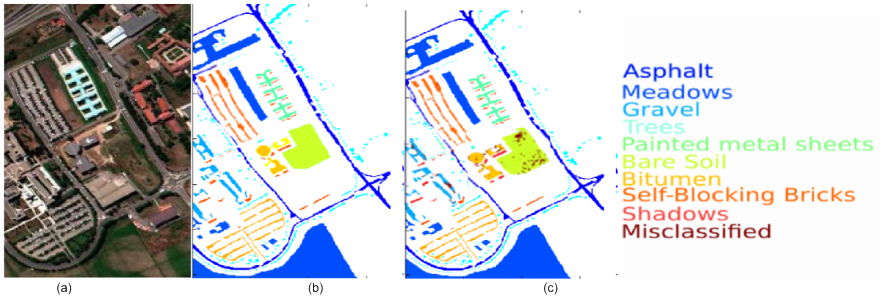


FIGURE 4.1 – Classification au niveau pixel de l'image hyperspectrale « Université de Pavia » (image RGB à gauche (a), vérité terrain au milieu (b) et prédictions des classes proposées dans [Ben+18b] à droite (c).

sur très peu d'exemples sont alors nécessaires. C'est dans ce cadre que nous nous sommes intéressés à la classification des sols depuis des images hyperspectrales.

Apprentissage supervisé depuis peu d'exemple pour la segmentation sémantique

Comment apprendre directement depuis des données rares et spécifiques ? Contributions de thèse d'Amina Ben Hamida.

Dans le cadre de la thèse d'Amnina Ben Hamida, nous nous sommes intéressés à une tâche de segmentation sémantique d'images pour laquelle il est nécessaire de réaliser la classification de chaque pixel. Le domaine d'étude est l'imagerie hyperspectrale des sols en prise de vue aéroportée. Dans ce cadre, les images annotées et diffusées dans le domaine public sont encore rares. On parle de « jeux de données » pour de simples images de dimensions faibles comme les images du capteur AVIRIS [2]. Par exemple, l'image de l'université de Pavia montrée sur la figure 4.1 est de résolution 610*340 pixels pour une résolution spatiale de 1.3m par pixel, chacun décrit par 103 bandes spectrales. La vérité terrain consiste en une catégorisation de chaque pixel en 9 classes et reste limitée à environ 20% de l'ensemble des données.

Ce type de contexte contraste fortement avec celui, plus favorable, présenté dans la littérature dans le domaine multimédia pour lequel de grands jeux de données annotés sont disponibles, par exemple CAMVid [BFC09] et Cityscapes [3]. Les méthodes d'apprentissage supervisées utilisées dans ce cadre favorable consistent en la création de lots d'échantillons (« batchs ») de b imagettes de taille $h \times w$ pour lesquels le modèle prédit la probabilité de chaque classe c pour chaque pixel x_p de chaque imagette. Un critère d'optimisation l typique suit alors l'équation 4.3 qui représente une mesure moyenne d'entropie croisée entre les probabilités de classes de référence $f(x_{p,c})$ et prédite $\hat{f}_\theta(x_{p,c})$ sur l'ensemble des pixels du batch. Un terme de régularisation sur les paramètres optimisés $L_r(\theta)$, par exemple une norme L^2, est généralement ajouté et son influence est pondérée par un coefficient λ_r. Dans cette configuration, le critère est moyenné sur $N = h \times w \times b$ pixels et correspond à la méthode A sur la figure 4.2.

$$l\left(\hat{f}_\theta, f\right) = \frac{1}{N}\sum_{p=1}^{N}\sum_{c=1}^{C} f(x_{p,c}) log\left(\hat{f}_\theta(x_{p,c})\right) + \lambda_r L_r(\theta) \tag{4.3}$$

2. aviris.jpl.nasa.gov
3. https://www.cityscapes-dataset.com/

Cette stratégie est intéressante dans le cas général, car la taille de batch b n'a plus besoin d'être grande pour calculer une erreur moyenne sur une grande variété de situations si les imagettes garantissent la présence d'un nombre de classes varié dans des situations variées. Cela peut être le cas sur certains jeux de données comme CAMVid [4] [BFC09]. En revanche, sur des images par exemple de haute résolution, cette diversité locale peut se réduire et ne garantit alors plus un apprentissage fiable sans augmentation de la variété des situations. Une solution consiste alors à sous-échantillonner les images. Une autre approche consiste à augmenter b en assurant une bonne sélection des imagettes. Néanmoins, lorsque les annotations sont rares et partielles comme montré sur la figure 4.1b, il peut être difficile d'obtenir une variété suffisante d'imagettes dont tous les pixels soient annotés. Le nombre d'imagettes est alors réduit et la distribution des catégories peut être fortement déséquilibrée. Les méthodes de rééquilibrage classiques comme [LSD14] qui applique une pondération des erreurs en fonction de la proportion de chaque catégorie ont alors une efficacité limitée.

Dans cette configuration, notre proposition [Ben+18b] a consisté à réaliser une stratégie d'entrainement spécifique que l'on peut résumer ainsi :

- on considère des imagettes dont la taille $n \times n$ correspond à celle du champ récepteur des apprenants, dans notre cas, des réseaux de neurones profonds. On entend par champ récepteur, l'ensemble des pixels (voisins) qui impactent la catégorisation du pixel central de l'imagette.
- la base d'entrainement consiste en un ensemble d'imagettes de taille $n \times n$ dont les centres sont choisis depuis l'ensemble annoté Ω. Ces pixels centraux sont tirés aléatoirement tout en maintenant une distribution uniforme de la représentation des classes.
- pour chaque imagette du batch d'entrainement, le critère d'erreur n'est évalué que sur les pixels de l'ensemble Ω, c.-à-d. les pixels centraux. Les autres pixels de l'imagette ne servent alors qu'à synthétiser la réponse du pixel central. Dans l'équation 4.3, l'erreur est alors moyennée sur $N = b$ ce qui correspond à la taille du batch.

Cette stratégie correspond à la méthode B illustrée sur la figure 4.2. Elle permet de garantir un équilibre des classes dans la phase d'entrainement. Par ailleurs, cette approche permet d'introduire des données non annotées. Comme la prédiction n'est évaluée que sur le pixel central, les pixels de Ω à la frontière des zones annotées peuvent être considérés. Leur classification est alors influencée par les pixels voisins potentiellement non annotés. Cette approche permet ainsi de montrer au modèle des relations de voisinage spatiales plus riches. Ceci est particulièrement intéressant dans le cas de données partiellement annotées pour lesquelles les photo-interprètes établissent la vérité terrain en ne sélectionnant que des régions d'intérêt ne contenant qu'une unique classe connue. Ceci est illustré sur la figure 4.2 pour laquelle l'imagette représentée par le cadre rose est centrée sur un pixel, représentant la catégorie « métal peint », entouré d'une zone non annotée elle-même voisine d'une autre classe connue. Le modèle peut alors apprendre cette relation spatiale.

Résultats et limites

Une sélection précise des données d'entrainement permet d'apprendre efficacement.

Nos travaux publiés dans [Ben+18b ; Ben+16] ont montré la pertinence de l'approche sur un problème de catégorisation des sols depuis une captation d'images hyperspectrales aéroportées. Cette méthode est associée à une architecture de réseau de neurones spécifique, composée d'un nombre très réduit de paramètres, décrite au paragraphe 4.2.1. Les expérimentations conduites avec des architectures à différents degrés de complexité ont montré la

4. http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/

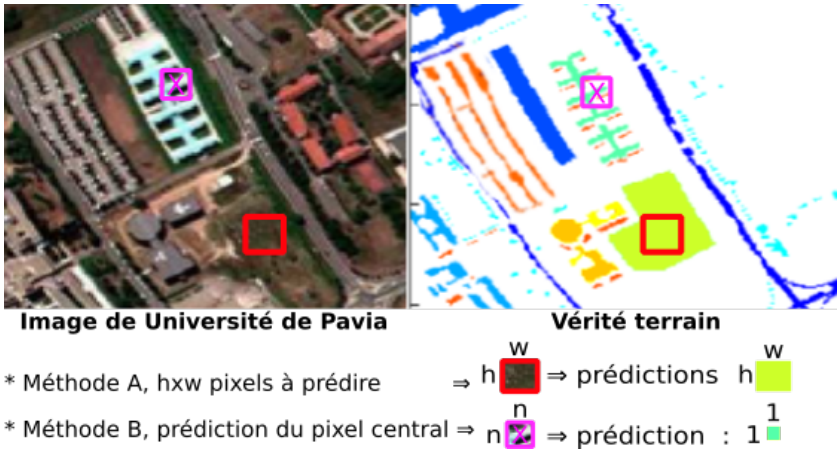


FIGURE 4.2 – Extraction d'échantillons d'apprentissage pour la segmentation sémantique. La méthode A est l'approche classique, une imagette est extraite dans les zones annotées (zones de vérité terrain à droite en couleur). La vérité terrain est disponible sur l'ensemble du patch et le réseau doit la prédire intégralement. La méthode B, proposée dans [Ben+18b] extrait des patchs de taille correspondant à la taille $n \times n$ du champ récepteur du réseau de neurones utilisé et ne prédit que la classe du pixel central. L'imagette doit donc être centrée sur un pixel annoté, mais il peut couvrir les zones voisines non expertisées.

rapidité de convergence de l'apprentissage associée à de bonnes capacités de généralisation.

La figure 4.3 montre l'évolution de l'exactitude (« Accuracy ») de la catégorisation en fonction du taux d'échantillons utilisés pour l'entrainement, pour la meilleure architecture retenue. On observe, pour trois jeux de données, une très rapide convergence vers le taux de performance maximum du réseau. Ces résultats montrent qu'il est possible d'atteindre une bonne qualité de catégorisation des sols depuis des images hyperspectrales avec un coût d'annotation réduit.

Nous nous sommes également intéressés à la transférabilité des modèles entrainés en les appliquant à d'autres jeux de données. Les réseaux sont intégralement entrainés depuis 5% des pixels annotés sur le jeu de donnée initial. Dans la phase de transfert, seule la couche de sortie du réseau est remplacée et entrainée sur 5% du jeu de données final. Les paramètres de la partie transférée restent figés ce qui reste un cas non optimal de transfert [Yos+14]. On observe dans ces conditions un écart de performance inférieur à 2% par rapport à une optimisation directe sur le jeu de données final. La transférabilité et la généralisation obtenues grâce à cette méthode d'apprentissage sont donc intéressantes. La régularité des données a pu être identifiée par le réseau de neurones associé en ne faisant appel qu'à un jeu de données d'entrainement limité.

Enfin, des travaux plus poussés sont encore à mener pour identifier et comprendre l'impact du choix des échantillons d'entrainement sur les performances du système. Une méthode de sélection optimale est à identifier.

Ceci nous amène à nous intéresser à la problématique de l'annotation de données pour l'apprentissage. Nous allons, dans ce qui suit, décrire une autre contribution permettant d'augmenter efficacement, et de façon interactive, la connaissance sur les jeux de données.

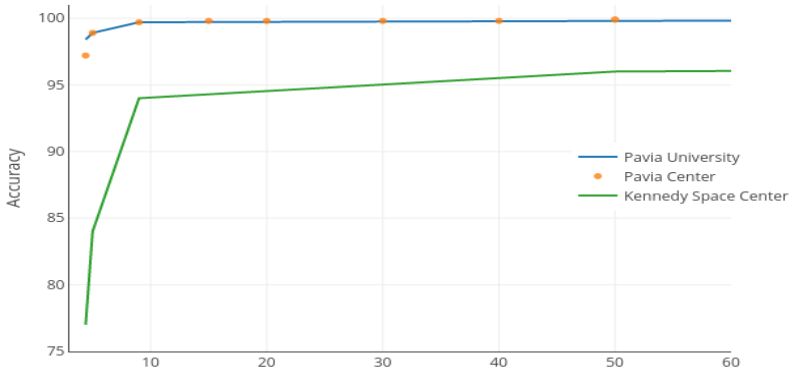


FIGURE 4.3 – Exactitude de la catégorisation des sols en fonction du taux d'échantillons utilisés pour l'entrainement pour un réseau de neurones à 8 couches présenté au paragraphe 4.2.1. Les performances sont présentées pour 3 jeux de données hyperspectrales (Université de Pavia, Centre-ville de Pavia et le centre spatial Kenedy). Extrait de [Ben+18b].

4.1.2 Apprentissage semi-supervisé actif

Apprendre depuis une connaissance partielle et l'étendre.

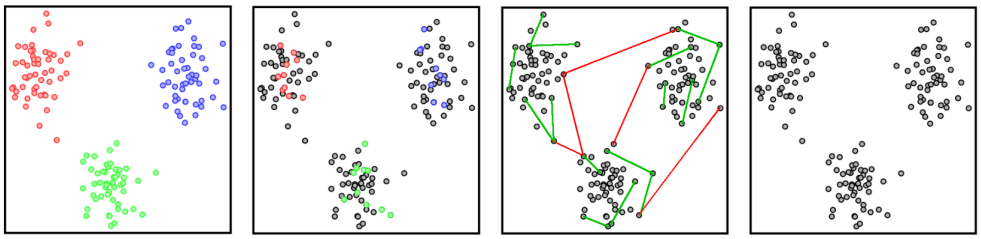
L'apprentissage supervisé s'appuie sur l'annotation complète d'un jeu de données. Son obtention est coûteuse, car elle fait en général appel à une expertise (humaine) nécessitant un travail long et fastidieux. Une alternative consiste à n'utiliser qu'une connaissance partielle sur les données. Deux types d'annotations sont alors possibles selon [Jai10] :

— l'annotation partielle par étiquettes pour laquelle chaque échantillon est annoté indépendamment. Le cas d'étude de la section précédente illustré sur la figure 4.1 en est un exemple.

— l'annotation partielle par liens ou contraintes entre échantillons. Par exemple, des liens de similarités ou dissemblances entre paires, triplets voire quadruplets [LTC13].

La figure 4.4 illustre ces différentes configurations.

Dans de telles configurations, l'apprentissage semi-supervisé est une méthode qui intègre les connaissances partielles et les associe à l'agrégation des échantillons non connus. Initialement proposée par Blum [BM98] en 1998, cette approche a été sujette à une grande variété de nouvelles propositions comme le montre une revue de l'état de l'art récente [CBP19].
Une extension naturelle de ces approches est l'apprentissage actif qui permet d'étendre dynamiquement la connaissance sur les données en proposant d'expertiser de nouveaux échantillons pertinents. Ce concept repose sur une détection d'échantillons spécifiques, par exemple, ceux considérés comme ambigus et ceux permettant de maximiser la qualité du modèle [Mel+15 ; Su+15]. On peut également citer des approches visant à réduire l'incertitude des prédictions [XJC16]. L'ensemble de ces méthodes d'apprentissage actif vise ainsi à maximiser la qualité de l'annotation, pour maximiser l'efficacité de l'entrainement des modèles tout en réduisant le coût d'annotation par les experts humains. On trouve également d'autres usages comme celui de rééquilibrer la distribution des annotations dans les jeux de données (catégories, similarités, etc.), ce facteur étant critique pour l'apprentissage



(a) Totalement étiquetée. (b) Partiellement étiquetée. (c) Partiellement contrainte. (d) Aucune connaissance.

FIGURE 4.4 – Illustration des différents types de connaissances sur un jeu de données selon [Jai10], reprise de [Voi15]. L'annotation par étiquette est complète sur l'imagette (a), partielle sur l'imagette (b) et chacune montre une couleur par catégorie. L'annotation par liens est montrée sur l'imagette (c) avec en vert les liens de similarité « Must Link (ML) » et en rouge les liens de dissemblance « Cannot Link (CL) ».

supervisé [Ert+07].

Il apparait ainsi que l'apprentissage semi-supervisé actif est un moyen d'optimiser un système depuis des données dont l'état des connaissances croît progressivement par l'introduction d'une expertise. Cette expertise restant coûteuse, il devient intéressant de maximiser ses effets.

Propagation automatique de la connaissance pour l'apprentissage semi-supervisé actif

Comment optimiser l'usage de la connaissance ? Contributions de la thèse de Nicolas Voiron.

Dans le cadre de la thèse de Nicolas Voiron [Voi15], nous avons proposé une approche qui vise à augmenter automatiquement la connaissance sur les données. Cette connaissance se présente sous la forme de contraintes ou liens de similarité (« Must Link, ML ») et de dissemblance (« Cannot Link, CL ») entre paires d'échantillons d'une base de données. Cet ensemble peut donc être vu comme un graphe dont les sommets représentent les échantillons et les liens, la connaissance. Pour un jeu de données de n échantillons, le nombre de liens du graphe complet est $\frac{n(n-1)}{2}$. Ceci confirme l'aspect potentiellement fastidieux de l'annotation. La méthode proposée vise alors à déduire automatiquement de nouveaux liens lors de l'ajout de nouvelles connaissances. Cette propagation automatique présente deux intérêts principaux :

- elle permet de réduire la sollicitation de l'expert et les coûts associés. Cela permet également de supprimer ses éventuelles contradictions. On peut alors considérer l'expert comme un « oracle ».
- dans la recherche de coupes dans le graphe des données, la propagation peut accélérer la convergence vers le partitionnement final par l'injection d'un nombre plus important de contraintes. Ceci est illustré sur la figure 4.5 pour laquelle un graphe, sa matrice de similarité et son bi-partitionnement sont montrés avec différentes quantités de contraintes, avant et après propagation. La méthode de partitionnement utilisée dans l'exemple est une méthode simple respectant peu les contraintes [XJC16]. Le fait de déduire de nouvelles contraintes permet de mieux guider la méthode vers une coupe appropriée.

L'état de l'art montre des règles classiques de déduction de contraintes connexes depuis celles déjà connues. La figure 4.6 illustre deux règles typiques : $ML + ML \Rightarrow ML$ ainsi

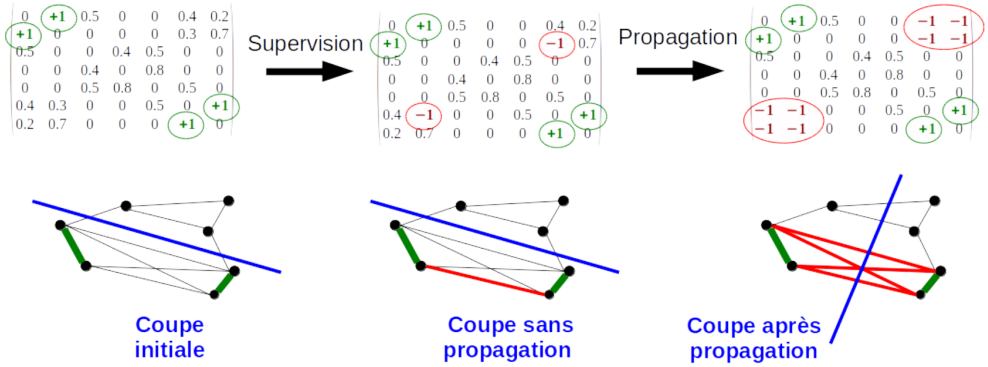


FIGURE 4.5 – Intérêt de la propagation pour le partitionnement de données. En haut, une matrice de similarité, en bas le graphe associé et sa coupe en bleu. En vert et rouge sont représentées les contraintes connues, respectivement similarité et dissemblances. Illustration reprise de [Voi15].

FIGURE 4.6 – Règles de propagation classiques : à gauche la règle 1 : $ML + ML \Rightarrow ML$ et à droite la règle 2 : $ML + CL \Rightarrow CL$. Reprise de [Voi15].

que $ML + CL \Rightarrow CL$. Cependant, le cas $CL + CL \Rightarrow ?$ reste toujours indéterminé. Une première contribution de nos travaux a été de montrer que pour tout n-partitionnement, si l'on trouve un n-simplexe composé de $\left(\frac{n(n-1)}{2} - 1\right)$ liens CL, alors le dernier est forcément de type ML comme illustré sur la figure 4.7. Une troisième règle de propagation est alors applicable.

La seconde contribution a consisté en la généralisation de la propagation automatique des contraintes. Pour un état donné du graphe, un bouclage spécifique des 3 règles de propagation assure la déduction automatique de l'ensemble des contraintes possibles. Cette stratégie a alors été intégrée dans un système d'apprentissage semi-supervisé actif illustré sur la figure 4.8. Une méthode de clustering choisie réalise un partitionnement des données d'entrée. L'expert, désormais oracle, ajoute des contraintes sur les paires qui lui sont proposées. Enfin, la propagation finalise l'enrichissement des connaissances du graphe avant un nouveau cycle.

Résultats et limites

Propager la connaissance et garantir une convergence plus rapide des méthodes de partitionnement.

Nos travaux ont été publiés dans [Voi+15a ; Voi+15b ; Voi+16]. Différentes méthodes de clustering ont pu être comparées : des approches basées sur le clustering spectral et une approche par réseaux de neurones profonds. Ces travaux ont montré que la méthode de propagation automatique proposée accélère systématiquement la convergence vers le partitionnement final de données. L'effet, sur les méthodes de clustering utilisées, dépend de leurs propriétés intrinsèques. Nous avons alors montré que l'usage de la propagation automatique de contraintes remet en cause le choix des méthodes. Des approches peu

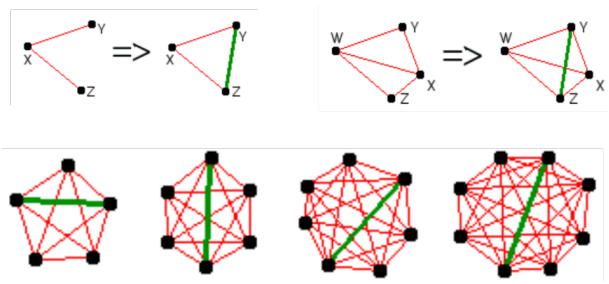


FIGURE 4.7 – Proposition d'une troisième règle de propagation automatique dans le cas d'un n-partitionnement : dans le n-simplexe $\left(\frac{n(n-1)}{2} - 1\right) \times CL \Rightarrow ML$. En haut à gauche, propagation dans le cas d'un 3-simplexe. En haut à droite, propagation dans le cas du 4-simplexe. En bas, résultat de propagation à des ordres supérieurs. Reprise de [Voi15].

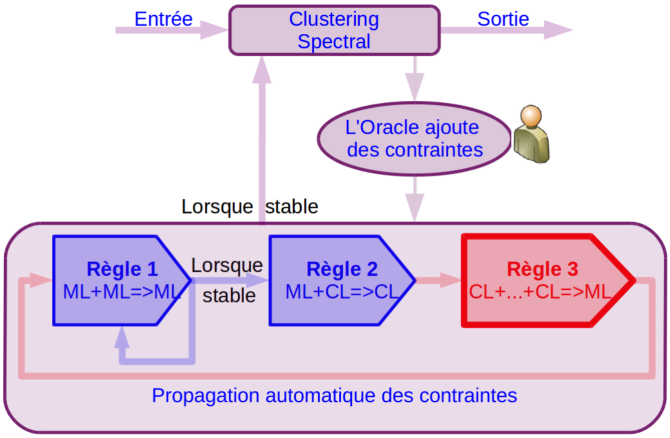


FIGURE 4.8 – Méthode de propagation automatique de contraintes pour le clustering de données. Reprise de [Voi15].

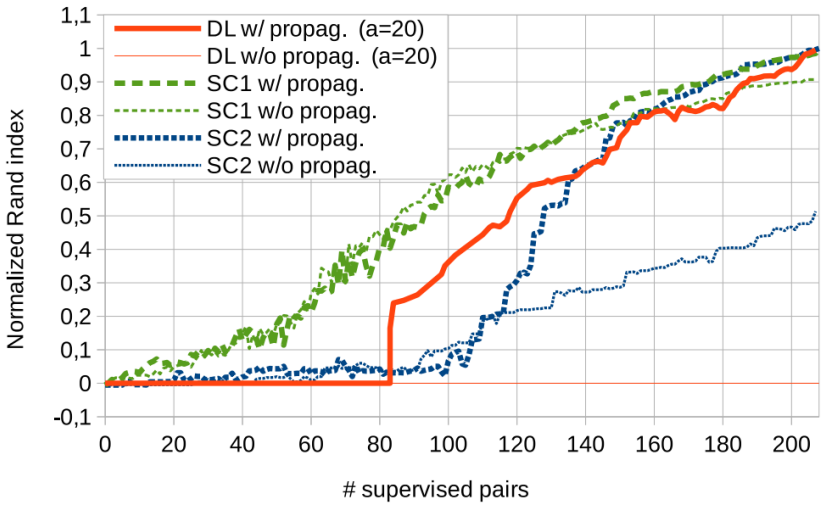


FIGURE 4.9 – Sur le jeu de données « Sonar » de l'UCI [DG17], mesure de la qualité de partitionnement en fonction du nombre de paires supervisées. Différentes méthodes de clustering sont comparées avec ou sans propagation (marqueurs fins). « DL » est une méthode par réseau de neurones profonds de type siamois apprenant les similarités et dissemblances entre paires d'échantillons. « SC1 » est la méthode de clustering spectral COSC [RH12] respectant toutes les contraintes. « SC2 » est une méthode de clustering spectral [XJC16] respectant moins les contraintes. Reprise de [Voi+16].

coûteuses et respectant peu les contraintes comme les réseaux de neurones et la méthode de clustering spectral [XJC16] profitent plus directement du peuplement accéléré de la connaissance dans le graphe de données. Elles peuvent alors rivaliser avec des méthodes de clustering plus efficaces, mais souvent plus coûteuses comme [RH12]. Nous pouvons ainsi revisiter les comparatifs de méthodes de clustering en intégrant la propagation au coût global de calcul pour identifier la meilleure combinaison.

Également, nous avons donné une garantie de la réduction de la sollicitation de l'expert/oracle, quelle que soit la méthode de partitionnement utilisée. Dans le cas d'un jeu de donnée de taille n à séparer en deux classes, la méthode garantit d'obtenir un partitionnement final en n interventions de l'oracle plutôt que $\frac{n(n-1)}{2}$ sans propagation.

La figure 4.9 illustre ces deux contributions. Elle montre l'évolution de la qualité de partitionnement en fonction du nombre de sollicitations de l'expert sur un problème de bi-partitionnement de données. Pour les 3 méthodes de clustering considérées, la propagation permet d'obtenir la qualité de partitionnement maximale en 208 interventions de l'expert soit la taille du jeu de données. Sans propagation, aucune des méthodes n'atteint cette qualité en si peu d'itérations. En particulier l'approche par réseau de neurones profond souffre d'un manque de données d'apprentissage sans propagation et ne peut converger vers une solution satisfaisante sans augmentation significative du nombre d'expertises.

Ces travaux peuvent être étendus pour optimiser encore la propagation à un stade plus précoce de l'annotation. Cela peut être réalisé par la définition de stratégies plus avancées de sélection des échantillons soumis à l'expertise. En particulier, la détection des n-simplexes permettant le déclenchement de la règle 3 est intéressante, car elle permet, par le nouveau déclenchement des deux premières règles de déduire un nombre très important de nouvelles

contraintes.

Enfin, des travaux plus approfondis sont à mener sur les méthodes d'apprentissage incrémental. La prise en compte de la variabilité de la connaissance nouvelle sur les données apportées par la propagation et les stratégies d'intégration à la connaissance déjà acquise sont à étudier. De telles études sont liées à la problématique typique de l'apprentissage incrémental qui concerne l'effet d'oubli catastrophique révélé par les travaux de Mc Closkey [MC89]. Il s'agit de la perte significative de performances des modèles sur les données liées à des apprentissages réalisés par le passé.

4.2 Architectures de réseaux de neurones profonds compactes

Les réseaux de neurones profonds, gains et risques de la démesure.

Les réseaux de neurones profonds et leur optimisation sont ces dernières années sujets à de très nombreuses propositions. On peut les organiser ainsi :

- les structures et opérateurs des réseaux. Depuis la proposition, en 2012, du réseau AlexNet [KSH12] composé de 8 couches organisées de façon linéaire, des architectures plus profondes et fortement interconnectées ont été proposées pour réaliser des tâches de plus en plus complexes. L'architecture multimodale et multitâche récurrente de Kaiser en est un exemple [Kai+17]. Des opérateurs spécifiques ont également rendu possible l'entrainement de structures neuronales plus profondes, par exemple, les non-linéarités de la famille ReLU et la normalisation de batchs [IS15].
- les méthodes d'apprentissage. Nous avons vu, dans la section précédente 4.1.2, des méthodes classiques. Des méthodes d'apprentissage originales associées à des structures de modèles spécifiques ont également été proposées. Les exemples typiques sont les méthodes d'entrainement de réseaux adverses [Goo+14] et les autoencodeurs variationnels [KW13] qui associent différents objectifs d'apprentissage pour assurer l'optimisation de différentes sous-structures du réseau.
- les techniques de régularisation. L'apprentissage des structures neuronales est soumis à différentes dérives de valeurs des paramètres et de biais de prédiction. Des critères d'optimisation supplémentaires peuvent être ajoutés au critère d'optimisation lié à la tâche afin de contraindre les réseaux à respecter des critères intrinsèques supplémentaires. Au-delà des régularisations de paramètres du type minimisation de leur norme (L^2, L^1 ou élastique), des techniques comme le « Dropout » [Hin+12] permettent par exemple l'entrainement de structures neuronales profondes et très riches en nombres de paramètres. Un usage spécifique du dropout permet également la simplification de réseaux surdéfinis. [Gom+19b]. D'autres techniques, comme l'orthogonalisation des paramètres des neurones au sein de chaque couche, visent à maximiser la complémentarité des neurones, régulariser les paramètres et faciliter l'apprentissage [BCW18].

On constate donc que la structure des réseaux, leur optimisation et leur régularisation sont intimement liées et de nombreuses pistes de recherches sont ouvertes. En revanche des questions « basiques » restent posées et quelques pistes de réponses commencent à se dessiner :

- quel nombre de couches faut-il choisir ? Quel gain réel a-t-on à augmenter ce nombre jusqu'à des valeurs parfois impressionnantes comme les 1000 couches du réseau de classification présenté dans [He+16]. Les travaux de Tishby [ST17] montrent que l'on gagne à utiliser des structures neuronales profondes pour accélérer la phase d'apprentissage. Les travaux de Mallat [Mal16] confirment ce point de vue, augmen-

ter la profondeur facilite la découverte d'invariants à des groupes de transformation complexes.

— quel nombre de neurones par couches ? Quel lien trouve-t-on entre la profondeur et la largeur du réseau ? Des travaux expérimentaux montrent un intérêt à augmenter le nombre de neurones par couche [ZK16]. Élargir un réseau plutôt qu'augmenter sa profondeur permet une meilleure efficacité des calculs grâce à une meilleure parallélisation des opérations. Des réseaux larges réalisent aussi des transformations plus riches des données et peuvent alors réduire la redondance de caractéristiques générées tout au long du réseau, mais au prix d'une augmentation rapide du nombre de paramètres. D'autres travaux comme [Che+16] associent les capacités de mémorisation des réseaux peu profonds, mais larges, aux capacités de généralisation des réseaux profonds.

— quelle structuration ? Le réseau doit-il avoir une forme linéaire ou autoriser plus de communication entre les couches ? Le réseau DenseNet [HLW16] est un exemple intéressant qui montre que l'augmentation de la communication inter couche permet de réduire significativement la profondeur, la largeur et donc le nombre de paramètres d'un réseau pour un niveau de performances similaires.

Les travaux actuels sont donc encore dans une phase exploratoire du domaine. Cet ensemble de questions peut être vu comme un problème d'optimisation multicritères. Il est abordé par deux approches. Une approche experte manuelle qui tend vers un compromis qui associe une augmentation de la profondeur et de la communication au sein des structures. L'alternative consiste à découvrir, automatiquement, de « bonnes structures » et leurs paramètres d'entrainement associés. Le livre récent [HKV19] donne une vision large de ce type d'approche. En particulier, les techniques d'apprentissage par renforcement [ZL17] et évolutionnaires [Rea+17] ont par exemple permis à Google de proposer le système AutoML [5]. Plus récemment, AdaNet [Wei+19] proche des approches par boosting [CMS14] permet un gain supplémentaire en termes de performances et rapidité de convergence des réseaux générés. On peut observer que ces approches favorisent, en général, l'accroissement des structures neuronales. Outre le coût de calcul nécessaire à l'optimisation de ces systèmes, les architectures résultantes sont alors complexes et potentiellement difficiles à justifier.

En revanche, d'une façon générale, les architectures de grande taille présentant un trop grand nombre de paramètres exposent à différentes déconvenues dont voici les principales :

— le sur-apprentissage. Les réseaux de neurones profonds sont, en général, surdéfinis ce qui leur donne une capacité d'apprentissage importante. Ils peuvent alors apprendre directement la base d'entrainement au risque de devenir incapables de prédire correctement sur des données d'exploitation. Sur ce point, les travaux de Hinton [Zha+17] montrent que les structures de réseaux proposés dans la littérature actuelle ont cette propriété et peuvent également apprendre des annotations aléatoires.

— le coût de calcul et les contraintes matérielles. Un grand nombre de paramètres impose des contraintes sur la puissance de calcul à mettre à disposition et sur l'occupation mémoire permettant de stocker les paramètres du réseau. Cela pose rapidement problème sur les systèmes à ressources limitées. Également en informatique connectée, la volumétrie des paramètres d'un réseau est problématique lorsqu'il s'agit de transmettre des mises à jour vers un système déporté.

La réduction de la taille des réseaux est donc tout de même nécessaire. L'approche classique consiste en un élagage en post traitement, après une phase d'entrainement des réseaux complets [HS93 ; HMD15]. Le réseau perd cependant en performances du fait des distorsions alors introduites et un réentrainement d'ajustement devient nécessaire. Cette approche est

5. https://cloud.google.com/automl/

étendue dans la méthode « Dense Sparse Dense » [Han+16] pour régulariser les réseaux et réduire les effets de sur-apprentissage. Cette approche se déroule en trois étapes : un premier entrainement est réalisé sur un réseau complet et est suivi d'un élagage des paramètres contribuant peu. Une poursuite de l'entrainement sur le sous-réseau élagué est appliquée pour consolider ses paramètres. Enfin le réseau complet est reconstitué en réinitialisant aléatoirement les portions préalablement supprimées. Une dernière phase d'entrainement permet la fusion du sous-réseau consolidé avec les portions réintroduites. Ces nouveaux degrés de liberté permettent de corriger les effets de sur-apprentissage. Ce type de méthode est cependant coûteux et ne réduit finalement pas la taille du réseau. Il devient alors préférable d'intégrer directement l'élagage dès la première phase d'entrainement, mais le risque est une réduction trop précoce des capacités d'apprentissage du réseau. Des approches récentes telles que les travaux de Gal [Gom+19b] proposent par exemple l'utilisation du dropout sur les connexions et neurones candidats à l'élagage pour valider ou non leur pertinence. Les réseaux, une fois entrainés, peuvent être directement élagués à un taux voisin de 70% sans perte importante de performance.
L'ensemble de ces méthodes n'intègre cependant pas de connaissances a priori sur les données. Dans ce contexte, nous sous sommes intéressés à la proposition de structures de réseaux légers en s'appuyant sur une telle expertise. Cette étude a pu être réalisée selon deux approches. La première, présentée dans la section 4.2.1, concerne la proposition de structures neuronales originales au travers des travaux de thèse d'Amina Ben Hamida et d'une collaboration avec l'université de Sfax en Tunisie. La seconde, présentée dans la section 4.2.2, concerne la proposition d'opérateurs spécifiquement adaptés aux données. L'objectif général est une réduction des coûts d'entrainement, de validation et de déploiement des réseaux dès la phase de conception par la réduction du nombre de paramètres et l'adéquation des architectures aux sources de données.

4.2.1 Réseaux de neurones compacts

Réduire le nombre de paramètres sur des problèmes de grandes données.

Nous nous sommes intéressés à différentes méthodes visant à réduire la quantité de paramètres dans les réseaux de neurones par des choix de structure.

Nous avons distingué deux approches :

— la réduction du nombre de paramètres pour les réseaux appliqués sur des données de grandes dimensions. Dans ce cadre, les structures neuronales sont en général complexes et à grand nombre de paramètres ce qui pose problème lorsque peu de données d'entrainement sont disponibles. L'approche est d'une façon générale intéressante lorsqu'il est nécessaire de construire un réseau avec une contrainte de faible occupation mémoire dès la phase d'entrainement.

— l'extension de la profondeur de réseau à nombre de paramètres constant. Ce cas de figure est intéressant pour étendre la taille du voisinage à prendre en compte pour réaliser des analyses locales (extension du champ récepteur) tout en maintenant une empreinte réduite en termes de nombre de paramètres.

Ces deux techniques peuvent constituer des règles de construction intégrables dans les méthodes d'exploration de structure automatisées afin de réduire l'espace de recherche vers des propositions d'intérêts connus.

Ces deux approches ont été étudiées dans le cadre de la segmentation sémantique pour laquelle chaque pixel d'une image doit être catégorisé.

Réduction de paramètres pour les structures neuronales profondes

Comment réduire le nombre de paramètres d'un réseau pour analyser les données de grandes dimensions ? Contributions de la thèse d'Amina Ben Hamida.

Dans le cadre de la thèse d'Amina Ben Hamida, nous nous sommes intéressés à la structuration compacte de réseaux pour l'analyse d'images hyperspectrales. Les dimensions spatiales et spectrales des données sont alors grandes. Nous avons vu au chapitre 2, au paragraphe 2.2.2, l'intérêt d'utiliser des convolutions 3D dans les premières couches du réseau afin de traiter l'intégralité de l'information sans imposer de sélection experte des données. Nous avons également proposé une architecture globale pertinente pour répondre au problème de segmentation sémantique. La figure 2.10 présentée au chapitre 2 donne le schéma de principe. Cette structuration a les propriétés suivantes :

- un ensemble de couches de convolutions 3D réalise un prétraitement de l'information spatio-spectrale et réduit progressivement l'information 3D en vecteurs 1D. L'intérêt est de fusionner les informations spatiales et spectrales de façon progressive.
- la partie centrale du réseau est composée de convolutions 1D peu coûteuses, réalisant à différentes échelles des traitements locaux.
- la structure du réseau minimise fortement la quantité de paramètres par l'usage du principe proposé avec l'architecture « SqueezeNet » [For+17]. Elle alterne des couches de convolution comportant de nombreux neurones (quelques dizaines) pour générer des représentations riches avec des couches réduisant la dimension des données en ne comportant que peu de neurones (quelques unités).
- les couches de décision en bout de réseau sont composées de neurones densément connectés à l'ensemble des attributs de la couche précédente, mais ne traitent que des données de dimensions très compactes, réduisant ainsi significativement le nombre de paramètres.

.

Extension de la profondeur à nombre de paramètres constants

Comment augmenter la profondeur tout en limitant l'augmentation du nombre de paramètres global du réseau ? Contributions d'une collaboration avec l'Université de Sfax, Tunisie.

Dans le cadre d'une collaboration avec l'université de Sfax en Tunisie, nos travaux avec Sourour Brahimi se sont focalisés sur la segmentation sémantique d'images pour lesquelles les objets sont de grande taille. La segmentation sémantique locale doit alors s'appuyer sur un voisinage de grande surface. Nous avons pour cela proposé d'augmenter la profondeur des réseaux par l'introduction d'une récurrence entre les couches neuronales :
Pour une opération B et une description initiale x_0 des données, la relation de récurrence proposée suit l'équation 4.4.

$$x_{t+1} = B(x_t) + x_0, x_0 \tag{4.4}$$

L'opérateur B est donc appliqué de façon répétitive avec les mêmes paramètres. Alors l'augmentation de la profondeur du réseau n'entraine pas d'augmentation du nombre de paramètres. Par ailleurs, si ces mêmes opérateurs s'appuient sur des agrégations locales de l'information comme le font les opérateurs de convolution, alors leur enchainement élargit la taille du champ récepteur globale du réseau. La récurrence permet donc d'augmenter la taille de champ récepteur sans augmenter le nombre de paramètres du système. Sur un problème d'analyse d'image par exemple, le traitement d'un pixel est alors influencé par un voisinage plus étendu. La figure 4.10 montre l'architecture récurrente que nous avons proposée. Elle s'intègre dans le réseau global de segmentation sémantique de type encodeur-décodeur décrit dans notre proposition [Bra+18]. Les opérations non linéaires B sont des blocs denses comme proposés dans [Jég+17 ; HLW16]. Ces blocs ont l'avantage de présenter un nombre déjà réduit de paramètres en exploitant la très forte densité de connexions

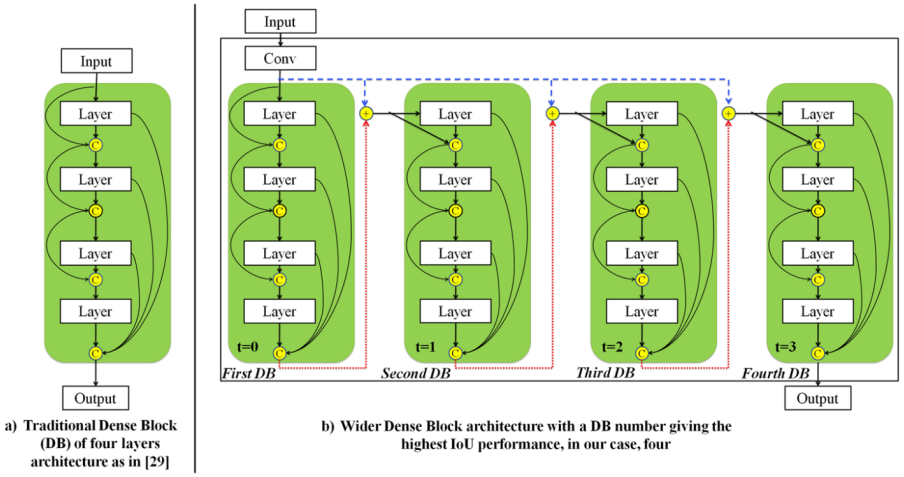


FIGURE 4.10 – À gauche (a), un bloc dense proposé dans [Jég+17 ; HLW16]. À droite (b), la récurrence proposée. Extrait de [Bra+18].

intercouches. La récurrence que nous proposons permet alors d'étendre une architecture parmi les plus compactes de l'état de l'art sans augmenter le nombre de paramètres.

Résultats et limites

Réduire le nombre de paramètres et augmenter la profondeur, une solution économique.

Nos travaux publiés dans [Ben+18b ; Bra+18] montrent l'intérêt des deux approches, chacune étant adaptée à un type de données pour lesquels nous avons pu établir des résultats de l'état de l'art. Il pourra être intéressant d'associer les deux approches dans un cadre applicatif le permettant.

Nous avons montré que les deux approches proposées permettent :

— l'amélioration de la performance de la tâche par l'augmentation de la profondeur du réseau. La recherche d'invariants semble facilitée même en s'appuyant sur un nombre très limité d'exemples et de paramètres. On trouve cependant une limite haute pour laquelle les performances se dégradent et le coût d'entrainement remonte.

— une réduction du besoin en ressource de calcul, les structures neuronales deviennent suffisamment légères pour être optimisées sur des systèmes limités. La méthode de réduction des paramètres est particulièrement intéressante, car elle permet, sur le cas d'étude présenté d'optimiser les meilleurs réseaux identifiés en quelques dizaines de minutes sur un simple CPU ordinateur portable. Le coût de calcul pour la prédiction est alors plus réduit encore et limite l'intérêt des méthodes de compression de réseaux.

Ces travaux exigent en revanche des analyses plus approfondies sur la compréhension des descriptions générées par les architectures proposées. En particulier la fusion des informations spatiales et spectrales locales pour l'analyse de données hyperspectrales est intéressante. Elle ouvre à d'autres applications comme la quantification des différentes catégories recherchées au sein de chaque pixel.

Nous venons de voir que la structuration des réseaux de neurones est un moyen d'adapter efficacement un modèle aux données. Nous allons maintenant nous intéresser à une approche complémentaire qui concerne la proposition d'opérateurs spécifiques.

4.2.2 Convolutions indexées pour l'imagerie non conventionnelle

Comment adapter les outils d'analyse à des images à l'échantillonnage non standard sans surcoût ? Contributions de la thèse de Mikaël Jacquemont.

Les travaux de thèse de Mikaël Jacquemont, en collaboration avec le Laboratoire d'Annecy de Physique des Particules (LAPP) et l'entreprise italienne Orobix [6], ont permis de s'intéresser à des opérateurs spécifiques pour les réseaux de neurones permettant d'appréhender des données issues de capteurs non conventionnels, directement et sans surcoût en termes de calcul et quantité de paramètres.

Insérés dans le projet GammaLearn financé par le projet européen H2020 Asterics et la fondation de l'université Savoie Mont Blanc sur le thème de l'astrophysique, nous nous sommes intéressés à un problème de détection de rayonnement cosmique gamma et de régression des paramètres associés tels que la direction du rayonnement, son énergie et son point d'impact virtuel. Cela est rendu possible par la captation de la gerbe caractéristique que ces rayonnements créent lors de leur entrée dans l'atmosphère terrestre par interaction avec les particules locales. Dans ce cadre, des télescopes terrestres spécifiques sont utilisés. Leurs capteurs présentent des formes et échantillonnages spatiaux variés et nous nous sommes intéressés en particulier aux capteurs à échantillonnage hexagonal qui contrastent avec les matrices classiques de pixels carrés de nos capteurs standards. Cet échantillonnage spécifique se retrouve dans d'autres domaines de l'imagerie non conventionnelle et même dans certains capteurs visant le marché grand public tel que les capteurs plénoptiques [Hah16].

Dans ce contexte, nous avons proposé, développé et diffusé [7] des opérateurs spécifiques de convolution et de sous échantillonnage indexé. Ces opérateurs permettent la création de méthodes de traitement d'image et de réseaux de neurones profonds capables de travailler directement sur des images à l'échantillonnage spécifique, dont le maillage hexagonal. Comparés aux approches classiques, ces nouveaux opérateurs ont les avantages suivants :

- la possibilité de traiter directement les données en sortie de capteur, dans leur échantillonnage initial, sans avoir recours à un sur-échantillonnage pour se ramener à un échantillonnage à maille carrée classique [Hol+17]. Cela permet de ne pas introduire de distorsions liées à cette opération et également de réduire le coût de calcul.
- certains échantillonnages présentent des relations de voisinage plus réduites permettant de réduire la complexité d'opérateurs locaux tels que les convolutions. Par exemple, dans le cas de l'échantillonnage hexagonal, seuls 6 voisins sont à prendre en compte contrairement aux 8 d'un échantillonnage à maille carrée classique. Pour les réseaux de neurones profonds, ceci peut induire une réduction importante du nombre de paramètres.

La figure 4.11 montre le principe de l'approche. Nous nous appuyons sur l'opération *im2col* disponible dans les outils standards qui ramène les convolutions à un produit matriciel optimisé entre les valeurs des pixels à traiter et les paramètres des noyaux de convolution. En adaptant cet opérateur, il devient possible de réaliser tout type de convolution ou sous échantillonnage en définissant préalablement le voisinage local de chaque pixel à intégrer dans l'opération. Cette approche permet d'adapter des convolutions de tout type, par exemple les convolutions à trous [Che+18], à tout type de maillage.

6. https://orobix.com/
7. https://github.com/IndexedConv/IndexedConv

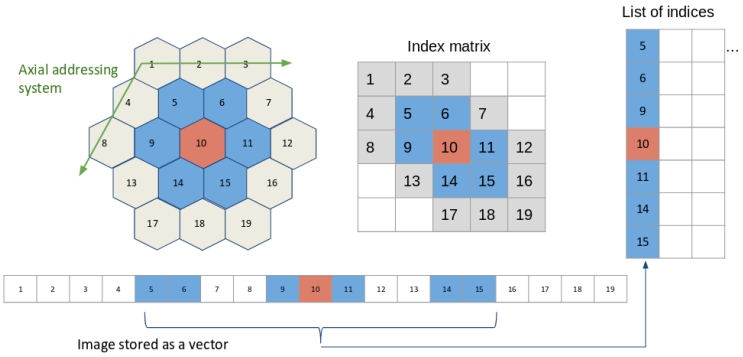


FIGURE 4.11 – Méthode de création des opérateurs indexés appliquée aux images à maillage de pixels hexagonaux (à gauche). Le voisinage du pixel central est identifié selon l'adressage axial des pixels (au centre). Au final, seuls les voisins concernés par l'opération sont conservés et organisés pour l'opération de convolution ramenée à un produit matriciel optimisé (à droite). Extrait de [Jac+19b].

FIGURE 4.12 – Exemple de rééchantillonnage d'une image multimédia standard de la base CIFAR vers une matrice hexagonale. Une interpolation sommaire prenant en compte au maximum deux pixels voisins est réalisée. Extrait de [Jac+19b].

Résultats et limites

Les convolutions indexées, un outil générique permettant le traitement direct de données spécifiques.

Les travaux publiés dans [Jac+19b ; Vui+18 ; Jac+19c] montrent l'intérêt de cette approche. Dans [Jac+19b], sur un problème de classification d'images multimédia, un cadre expérimental classique pour les études comparatives standardisées, nous comparons pour une architecture de réseau similaire :

- un réseau à convolution classique appliqué sur des images originales à maillage carré.
- un réseau à convolution indexée appliqué sur la même base de données rééchantillonnée sans interpolation exacte sur un maillage hexagonal comme illustré sur la figure 4.12. Le réseau présentant moins de paramètres que l'approche classique, pour réaliser une comparaison honnête, l'équilibre est rétabli en augmentant marginalement pour chaque couche le nombre de neurones.

Les résultats montrent des niveaux de performances équivalents sur la tâche de classification. Le point fort est lié au fait que le réseau basé sur les convolutions hexagonales comportant moins de paramètres, l'augmentation du nombre de neurones permet d'augmenter la richesse des représentations de l'information traitée. Sur cette expérience, malgré

les distorsions introduites par le rééchantillonnage des images en entrée du réseau, le niveau de performance reste comparable à l'architecture classique appliquée sur les images originales.

Cette approche est désormais déployée sur le problème d'analyse d'images astrophysiques au sein d'un outil unifié permettant la comparaison entre les méthodes standards du domaine et les approches par réseaux de neurones hexagonaux [Vui+18 ; Jac+19c].

La principale limite de cette approche est d'ordre purement technique. Elle est liée au manque d'optimisation des opérateurs indexés que nous avons développés. La comparaison au niveau des temps de calcul n'est bien évidemment pas faite à jeux égaux si l'on compare les solutions de recherche proposées récentes à des solutions optimisées sur le long terme par de nombreuses équipes de recherche et industrielles.

4.3 Synthèse et perspectives sur l'apprentissage et les réseaux de neurones

Des travaux allant vers des méthodes d'entrainement et des structures à coût réduit.

Nous avons vu dans ce chapitre des méthodes d'apprentissages associées à des structures permettant d'estimer et d'utiliser la relation, le modèle, qui lie les données à un objectif. La capacité d'une structure neuronale à estimer un modèle dépend de différents facteurs qui dépassent la simple idée « qu'il faut beaucoup de données ». Certains des plus importants sont :

- la variété des données d'entrainement et l'équilibre de la distribution des différents cas à différencier. Une sélection appropriée des données dans la phase d'entrainement peut permettre, selon le cas d'étude, un entrainement sur une quantité réduite de données.
- la structure du réseau. Des stratégies de réduction de son nombre de paramètres et une augmentation de la profondeur permettent une convergence rapide avec un effet de sur-apprentissage limité.
- les techniques d'annotation. L'apprentissage actif et la propagation automatique de contraintes peuvent permettre à moindre coût d'augmenter la connaissance sur les données et de favoriser l'apprentissage des estimateurs.

Nous avons alors proposé des techniques d'apprentissage et des opérateurs et structurations d'estimateurs permettant l'optimisation de modèles depuis peu d'exemples en limitant les effets de sur-apprentissage.

Ces contributions vont vers l'introduction d'une expertise plutôt que l'automatisation. Cela contre balance l'approche vue aux deux chapitres précédents qui encourageait l'inverse, soit l'introduction de l'apprentissage automatique au détriment de l'expertise. Cependant, le cadre est différent. Dans les chapitres précédents, l'objectif est d'estimer, avec moins d'apriori ou contraintes, un modèle pour compléter ou remplacer une approche experte plus contrainte ou incomplète. La structure de l'estimateur et sa méthode d'entrainement sont supposées déjà choisies. Le choix de ces « hyperparamètres » n'est abordé que dans ce dernier chapitre. Nous avons alors pu voir la variété des approches possibles rapportée par la littérature. L'espace de recherche à explorer par des approches purement automatique est grand et amène à des coûts importants sur différents aspects, notamment :

- un coût lié à la recherche de bonnes structures neuronales et leur technique d'entrainement. La validation d'un modèle requiert une séquence potentiellement importante de phases d'entrainement et de validation elles-mêmes coûteuses en temps et en ressources de calcul.

— un coût potentiel lié à la structure de l'estimateur sélectionné. Les méthodes automatiques actuelles vont plutôt dans la direction de l'expansion que de la compression des structures. La compression est souvent un post traitement introduisant également des coûts supplémentaires.

Une approche intégrant une expertise permet alors de réduire ces différents coûts. Cette expertise génère en contrepartie un coût humain également non négligeable. Une perspective est donc d'intégrer en tant que règles les approches expertes dans les méthodes automatiques pour réduire le coût global, expert compris.

De ces travaux, nous pouvons identifier quelques limites principales :

— les structures d'estimateurs proposées traitent le problème de bout en bout et masquent une possible justification des réponses données.
— nous n'avons abordé ici que des approches par apprentissage principalement supervisé qui imposent l'obtention d'une connaissance coûteuse sur les données.
— les problématiques auxquelles nous nous sommes intéressés sont principalement monotâche.

Ces limites ouvrent à un certain nombre de pistes de recherche.

Pistes de recherche

Vers l'apprentissage multitâches asynchrone et décentralisé.

L'extension de ces travaux est envisagée selon deux axes principaux : l'apprentissage faiblement ou non supervisé et l'apprentissage multi tâche. Ces directions présentent de fortes complémentarités qu'il sera intéressant de valoriser également.

— Sur le thème de l'apprentissage faiblement supervisé, les travaux initiés dans la thèse de Nicolas Voiron sur la prise en compte de contraintes dans l'apprentissage et les travaux préliminaires d'Amina Ben Hamida sur les approches génératives [Ben+18a] et l'apprentissage depuis des données bruitées [Ben+18a] sont à étendre. Ils s'inscrivent dans le sujet actuellement très actif de l'apprentissage faiblement supervisé et de la représentation sous contraintes. L'objectif est de proposer des méthodes de représentation des données pertinentes permettant leur partitionnement. Les méthodes de la famille des autoencodeurs variationnels [KW13] sont des approches typiques et pourraient permettre la désintrication des paramètres contrôlant les données. Cependant, comme nous l'avons vu, des travaux comme [Loc+18] montrent la difficulté du problème sans l'introduction d'apriori. Une solution peut consister à introduire les notions de similarité dans l'apprentissage de représentations intermédiaires. Comme nous l'avons vu dans la thèse de Nicolas Voiron, ceci permet la détection des échantillons ambigus et des situations de conflit. Ceci permet également d'apporter des informations pertinentes permettant d'aller vers la compréhension et la justification des décisions réalisées par les estimateurs et les réseaux de neurones en particulier.
— Sur le thème de l'apprentissage multitâche, la littérature rapporte l'intérêt d'optimiser un système sur de multiples problèmes pour améliorer l'efficacité de l'apprentissage et augmenter le niveau de performance globale. Ces approches ont en général pour but de répondre à des objectifs parallèles et complémentaires, par exemple, catégoriser les pixels d'une image d'une part et estimer leur distance à la caméra d'autre part [Xu+18]. En revanche, le conflit potentiel entre les différentes tâches peut tout de même limiter l'efficacité globale du système. L'identification des fronts de Pareto commence à se dessiner [SK18]. Une extension de nos travaux peut aller dans cette direction en envisageant l'apprentissage multitâche sous un angle différent. Plutôt que de paralléliser en bout de structure, il peut être intéressant d'enchaîner des tâches complémentaires sur des étapes intermédiaires successives d'une même

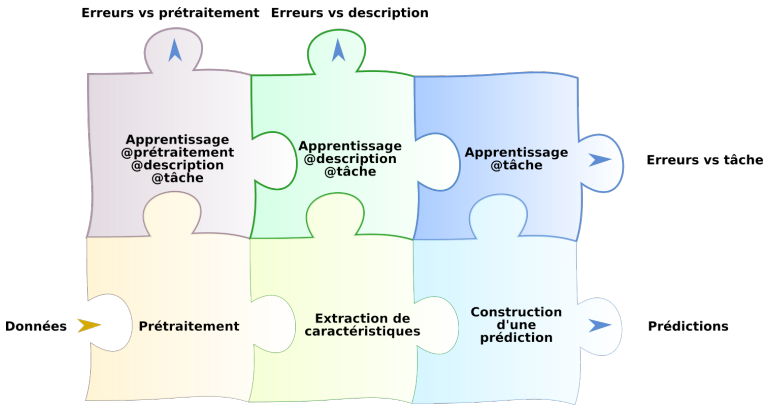


FIGURE 4.13 – Perspectives de travail sur l'apprentissage : apprentissage multitâche sur un enchainement de critères amenant à la prédiction sur la tâche principale. Le prétraitement et l'extraction de caractéristiques peuvent alors intégrer leurs propres critères pour contribuer à l'objectif.

architecture. Si l'on considère le cheminement décrit dans les chapitres précédents, nous avons évoqué l'intérêt de réaliser différentes tâches comme le prétraitement et l'extraction de caractéristiques. Dans un cadre approprié pour lequel une expertise ou des modèles physiques permettent de définir des critères d'optimisation sur ces tâches intermédiaires, il devient alors possible de s'intéresser à leur optimisation associée à un objectif plus global (classification, regression, etc.) dans une approche multitâche. La figure 4.13 illustre ce concept en décomposant une structure neuronale en différents blocs (prétraitement, description, prédiction), chacun étant associé à un critère d'optimisation spécifique. Les critères d'optimisation en aval peuvent impacter les étapes précédentes. On peut alors associer des critères locaux aux objectifs des niveaux supérieurs. Cette approche multitâche permet en outre d'aller vers l'explication et la justification des prédictions par la mise en valeur de données et critères intermédiaires au travers d'une relation de causalité. Un cadre applicatif sur lequel nous travaillons déjà est l'analyse des rayons gamma en astrophysique pour laquelle des grandeurs à estimer dépendent d'autres grandeurs à estimer en amont. Un autre cadre possible est la conduite autonome qui peut associer la décision de commande d'un véhicule à des tâches intermédiaires comme la qualité de la perception apportée par une phase d'amélioration d'image et la détection d'objets environnants. Pour ces cas difficiles, la prise en compte de l'incertitude des différentes estimations est une piste connexe pour estimer la qualité des prédictions voir les renforcer [KGC18].

Enfin, une perspective plus large vise à associer les approches multitâches dans un cadre d'apprentissage asynchrone, décentralisé et à empreinte énergétique faible. Ceci nous amène alors vers l'apprentissage incrémental et multi critères. Cette association de concepts est liée au fort coût énergétique typiquement observable dans l'usage des modèles profonds. D'une façon générale, le cheminement entre la captation des données brutes et la sortie d'un modèle peut être long en termes de flux de calculs, mais aussi en termes de distance physique. Ceci est valable pour la propagation des données de la source vers la sortie pour la prédiction, comme en sens inverse, pour la rétropropagation du gradient dans la phase d'entrainement. Deux inconvénients sont identifiables :

— la spécialisation de la chaîne de traitement à la tâche, prétraitement compris. Dans une phase d'entrainement monotâche classique, toute la chaîne de traitement se spécialise et devient donc moins générique. Les approches multitâches sont des solutions permettant de compenser ce phénomène.
— le coût de transport et de synchronisation des données. Cela est particulièrement visible dans les systèmes décentralisés et l'informatique connectée, mais on le trouve aussi à de plus petites échelles, au coeur d'un système de calcul.

Les systèmes biologiques ont trouvé une solution intéressante à ce problème par l'usage de processus d'apprentissages asynchrones décentralisés [MWK16]. Partant d'une structure surparamétrée et évolutive, des tâches d'optimisation locales permettent de répondre à des problématiques génériques (amélioration de signal, description, etc.) et sont complétées par des apprentissages plus spécifiques (classification, détection, etc.). Une modulation dynamique de ces différents processus d'apprentissage permet au cours du temps de converger, à coût réduit vers un système robuste. Les processus proches de capteurs deviennent génériques par la confrontation à une grande richesse de données et l'influence d'une grande variété de tâches, génériques et spécialisées. Les processus de plus haut niveau sont plus spécialisés par une variété de données et tâches plus réduites. Pour aller dans cette direction, la définition d'une architecture multitâche générique et d'un processus de régulation des apprentissages est à identifier. Nous pourrons nous inspirer des neurosciences, [MWK16] mais aussi des travaux des communautés systèmes distribués et apprentissage [Zhe18]. Enfin, des propositions innovantes comme les « capsules » de Hinton proposées dans [SFH18] permettant la description de l'information sous une forme hiérarchique complètent l'éventail des pistes d'intérêt.

Enfin, cette approche s'appuie sur la notion d'apprentissage continu ou incrémental et devra donc s'intéresser à l'effet d'oublie catastrophique révélé par les travaux de Mc Closkey dès 1989 [MC89]. Il s'agit de la perte significative de performances des modèles sur les représentations des données apprises par le passé. Des solutions telles que les récentes approches par « distillation » [HVD15] appliquées sur des problèmes multitâches [SSA17] ou par l'usage de modèles génératifs pour « raviver les souvenirs » [Dat+19] sont des pistes d'intérêt pour contrer ce type de difficulté. L'apprentissage distribué asynchrone est donc à associer à cette problématique connexe.

Chapitre 5

Synthèse et Perspectives

Les travaux présentés s'organisent autour d'une chaîne de traitement de l'information spécialisée sur l'analyse d'images. Elle couvre les trois étapes traditionnelles : le prétraitement, l'extraction de descriptions et la prédiction. Au travers d'un ensemble de thèses et de projets, nous avons exploré différentes méthodologies pour aborder ces étapes. Dans la variété de cadres applicatifs qui en découlent, nous avons considéré des approches « expertes » et des approches par apprentissage :

- les méthodes expertes peuvent introduire une connaissance sur les processus qui se présente alors sous la forme de contraintes de visualisation, de concordance avec un modèle de référence ou de description. En revanche, des limites sont observées. Elles sont principalement liées à notre connaissance partielle du lien entre les données et l'objectif et induisent des contraintes non optimales.
- les méthodes par apprentissage permettent de faire abstraction de ces connaissances partielles et d'apprendre directement un lien entre les données et l'objectif. Cela se fait par la confrontation de modèles entrainables à des données dans un processus d'apprentissage. En s'appuyant sur des structures à multiples niveaux de représentations telles que les réseaux de neurones profonds, il est possible d'apprendre des modèles très élaborés et de dépasser le niveau de performances des approches « expertes ». Cependant, des limites telles que le manque de compréhension des processus appris et le besoin d'introduire une expertise pour guider et faciliter l'apprentissage ont pu être montrés.
- l'apprentissage a été exploré dans le cadre des approches par réseaux de neurones profonds. Nous avons montré l'intérêt d'aller vers des processus d'optimisation associés à des structures neuronales permettant d'apprendre efficacement tout en présentant un coût réduit aussi bien dans la phase d'optimisation que dans la phase de prédiction.

D'un point de vue général, ces travaux se situent dans le contexte de la progression fulgurante de l'apprentissage profond qui tendraient à remettre en cause l'usage de méthodes expertes et à remplacer la chaîne de traitement composée de multiples étages par une unique fonction paramétrique optimisée de bout en bout. Nous avons vu, au travers d'un panel de cadres applicatifs, les limites de cette approche aussi bien sur des problématiques d'entrainabilité des modèles que pour la justification de leurs prédictions. Nous avons alors montré l'intérêt d'introduire une structuration multi niveaux et multitâches, pour lesquelles une expertise viendrait assister l'apprentissage.

Dans les trois principaux chapitres de ce document, nous avons identifié un ensemble de perspectives que nous pouvons synthétiser ainsi :

- **introduire des contraintes** liées à une expertise ou à des modèles physiques. Que l'on considère la fonction paramétrique dans son ensemble ou seulement certaines de ses composantes (prétraitement, extraction de descriptions ou autre), introduire une connaissance même partielle permet de :
  - simplifier le processus d'apprentissage en réduisant l'espace de recherche vers un ensemble de solutions plausibles.
  - renforcer nos connaissances initiales en comparant un modèle expert au modèle obtenu par apprentissage afin d'identifier les propriétés et lois non prises en compte par le modèle expert.
  - permettre de contrôler l'évolution du modèle. La comparaison avec un modèle physique même partiel permet également de suivre la dérive d'un modèle sur des situations de référence dans sa phase d'apprentissage.
- **décomposer la fonction et les critères**. Il s'agit de structurer la fonction monobloc paramétrique en une composition de fonctions ayant chacune ses critères d'optimisation. Cela revient à décomposer le problème en étapes avec les objectifs suivants :
  - permettre une justification des réponses des processus en différents points de la structure en se ramenant à des considérations intelligibles. Les contraintes de visualisation de prétraitements et de descriptions pour lesquels des critères d'optimisation spécifiques sont imposés en complément de la tâche principale sont une façon d'aborder le problème. Elles peuvent être complétées par des approches comme la détection d'attributs spécifiques des données permettant de justifier la prédiction [Has+19].
  - faciliter la décentralisation du processus en permettant de localiser certains sous-processus comme le prétraitement ou la description, plus proche de la source de données. Les effets attendus sont alors une réduction des flux d'information vers le processus central et un encodage des données facilitant la confidentialité. La synthèse de descriptions compactes par autoencodeur est alors une approche classique pour aborder le problème.
  - apprendre de façon asynchrone et multitâches. La décomposition du système en une composition de fonctions permet d'intégrer des critères d'optimisation pouvant être considérés de façon synchrone et asynchrone. Les critères faisant appel à une expertise ou annotation ne sont en général applicables que sur des données spécifiques maîtrisées alors que les tâches d'apprentissage non supervisées peuvent être optimisées en continu. Une gestion dynamique de l'ensemble des critères devient intéressante pour maximiser l'efficacité de l'apprentissage en s'adaptant au contexte et à la nature des données. Les systèmes biologiques ont su construire une dynamique efficace [MWK16] et sa modélisation en apprentissage profond est d'un intérêt certain. Cette perspective nous amène à l'union de différentes problématiques que sont la compréhension des processus biologiques en neurosciences, l'apprentissage optimal avec par exemple l'identification des fronts de Pareto en apprentissage multitâches, l'apprentissage continu ou incrémental pour apprendre depuis les données récentes tout en limitant l'effet d'oubli des anciens acquis.

Ces perspectives sont donc nombreuses. Leur mise en oeuvre dépend bien sûr du contexte, des collaborations existantes et à venir, de la construction de projets académiques et industriels.

— Au niveau local, la collaboration avec les confrères au sein de l'université pour le montage de projets est déjà étendue. L'imagerie satellitaire, l'apprentissage statistique et le calcul distribués au LISTIC, la physique au LAPP, l'analyse de séries temporelles pour l'habitat au LOCIE constituent des pistes déjà ouvertes dans lesquelles ces perspectives peuvent se développer.
— Au niveau national et international, les échanges et travaux communs avec les confrères de la communauté multimédia (en particulier au LABRI, LIG, GIPSA), mais aussi dans des domaines plus spécifiques comme la physique (LAPP et NTNU) et l'imagerie satellite (IRISA Vannes, ONERA) permettent de construire des projets et répondre à des appels comme l'ANR. Par ailleurs, la participation à des concours internationaux comme TRECVid a montré nos capacités à étendre encore les collaborations et à se confronter en équipe à des problématiques scientifiques difficiles. Certaines des perspectives énoncées ci-dessus comme l'explicabilité des réseaux de neurones sont déjà initiées au travers de réponses dans un appel à projets ANR.
— Les partenaires industriels ouvrent enfin à la définition et la mise en oeuvre de méthodes répondant directement à un besoin pratique. Des collaborations déjà établies permettent d'identifier des sujets d'intérêt. On trouve des projets avec les petites et moyennes entreprises comme AboutGoods et Noveltis, mais aussi avec de grands groupes comme Total et Renault. Les applications industrielles sont directement concernées par des verrous technologiques comme la justification des prédictions, la modélisation de phénomènes physiques et l'optimisation des processus discutés dans nos perspectives.
— Enfin, les filières d'enseignement comme l'IUT et Polytech sur les sites de l'université Savoie Mont Blanc sont propices à l'exploration de méthodologies avec une certaine liberté d'action. Les stages et les financements sur projets locaux comme les AAP ou la fondation Savoie Mont Blanc permettent alors de développer des méthodologies et de consolider des pistes de recherche avant de les impliquer dans des projets de plus grande ampleur. C'est dans ce cadre qu'une collaboration avec des entreprises comme AboutGoods a pu être montée, tout comme le montage d'une thèse avec le LAPP sur des problèmes d'astrophysique.

Pour conclure, les perspectives proposées donnent une orientation sur les directions de recherche envisagées pour les années à venir. Elles aideront à construire et définir les projets académiques et industriels qui permettront de les développer.

Bibliographie

[CBA19] Souad Chaabouni, Jenny Benois-Pineau et Chokri Ben Amar. « ChaboNet : Design of a deep CNN for prediction of visual saliency in natural video ». In : *Journal of Visual Communication and Image Representation* 60 (2019), p. 79-93.

[Che+19] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu et Martial Hebert. « Image Deformation Meta-Networks for One-Shot Learning ». In : *CVPR*. 2019.

[CBP19] Veronika Cheplygina, Marleen de Bruijne et Josien PW Pluim. « Not-so-supervised : a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis ». In : *Medical image analysis* 54 (2019), p. 280-296.

[Dat+19] Pham Thanh Dat, Anuvabh Dutt, Denis Pellerin et Georges QuÃ©not. « Classifier Training from a Generative Model ». In : *Content-Based Multimedia Indexing (CBMI) 2019*. Dublin, Ireland, sept. 2019.

[Gom+19a] Aidan N. Gomez, Ivan Zhang, Kevin Swersky, Yarin Gal et Geoffrey E. Hinton. « Learning Sparse Networks Using Targeted Dropout ». In : *CoRR* abs/1905.13678 (2019). arXiv : 1905.13678. URL : http://arxiv.org/abs/1905.13678.

[Gom+19b] Aidan N. Gomez, Ivan Zhang, Kevin Swersky, Yarin Gal et Geoffrey E. Hinton. « Learning Sparse Networks Using Targeted Dropout ». In : *CoRR* abs/1905.13678 (2019). arXiv : 1905.13678. URL : http://arxiv.org/abs/1905.13678.

[Has+19] Muneeb ul Hassan, Philippe Mulhem, Denis Pellerin et Georges QuÃ©not. « Explaining Visual Classification using Attributes ». In : *Content-Based Multimedia Indexing (CBMI) 2019*. Dublin, Ireland, sept. 2019.

[HKV19] Frank Hutter, Lars Kotthoff et Joaquin Vanschoren. *Automated Machine Learning-Methods, Systems, Challenges*. 2019.

[Jac+19a] Kevin Jacq, Yves Perrette, Bernard Fanget, Pierre Sabatier, Didier Coquin, Ruth Martinez Lamas, Maxime Debret et Fabien Arnaud. « High-resolution prediction of organic matter concentration with hyperspectral imaging on a sediment core ». In : *Science of the Total Environment* 663 (jan. 2019), p. 236-244. DOI : 10.1016/j.scitotenv.2019.01.320. URL : https://hal-sde.archives-ouvertes.fr/hal-01995416.

[Sch+19] Eli Schwartz, Leonid Karlinsky, Rogerio Feris, Raja Giryes et Alex M. Bronstein. « Baby steps towards few-shot learning with multiple semantics ». In : *arXiv preprint arXiv :1906.01905* (2019).

[Wei+19] Charles Weill, Javier Gonzalvo, Vitaly Kuznetsov, Scott Yang, Scott Yak, Hanna Mazzawi, Eugen Hotaj, Ghassen Jerfel, Vladimir Macko, Ben Adlam, Mehryar Mohri et Corinna Cortes. *AdaNet : A Scalable and Flexible Framework for Automatically Learning Ensembles*. 2019. URL : https://arxiv.org/abs/1905.00080.

[Anc+18a] Codruta O. Ancuti, Cosmin Ancuti, Radu Timofte et Christophe De Vleeschouwer. « I-HAZE : a dehazing benchmark with real hazy and haze-free indoor images ». In : *arXiv :1804.05091v1*. 2018.

[Anc+18b] Codruta O. Ancuti, Cosmin Ancuti, Radu Timofte et Christophe De Vleeschouwer. « O-HAZE : a dehazing benchmark with real hazy and haze-free outdoor images ». In : *IEEE Conference on Computer Vision and Pattern Recognition, NTIRE Workshop*. NTIRE CVPR'18. Salt Lake City, Utah, USA, 2018.

[Bb18] Randall Balestriero et richard baraniuk. « A Spline Theory of Deep Learning ». In : *Proceedings of the 35th International Conference on Machine Learning*. Sous la dir. de Jennifer Dy et Andreas Krause. T. 80. Proceedings of Machine Learning Research. StockholmsmÃ¤ssan, Stockholm Sweden : PMLR, juil. 2018, p. 374-383. url : http://proceedings.mlr.press/v80/balestriero18b.html.

[BCW18] Nitin Bansal, Xiaohan Chen et Zhangyang Wang. « Can we gain more from orthogonality regularizations in training deep networks ? » In : *Advances in Neural Information Processing Systems*. 2018, p. 4261-4271.

[Che+18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff et Hartwig Adam. « Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation ». In : *ECCV*. 2018.

[EGE18] Deniz Engin, Anil Genc et Hazim Kemal Ekenel. « Cycle-Dehaze : Enhanced CycleGAN for Single Image Dehazing ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2018.

[Hab18] Mateus Habermann. « Band selection in hyperspectral images using artificial neural networks ». Theses. Université de Technologie de Compiègne, sept. 2018. url : https://tel.archives-ouvertes.fr/tel-02094282.

[Loc+18] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf et Olivier Bachem. « Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations ». In : *CoRR* abs/1811.12359 (2018). arXiv : 1811.12359. url : http://arxiv.org/abs/1811.12359.

[SFH18] Sara Sabour, Nicholas Frosst et G Hinton. « Matrix capsules with EM routing ». In : *6th International Conference on Learning Representations, ICLR*. 2018.

[SK18] Ozan Sener et Vladlen Koltun. « Multi-task learning as multi-objective optimization ». In : *Advances in Neural Information Processing Systems*. 2018, p. 527-538.

[WIC18] Emily J. Ward, Leyla Isik et Marvin M. Chun. « General Transformations of Object Representations in Human Visual Cortex ». In : *Journal of Neuroscience* 38.40 (2018), p. 8526-8537. issn : 0270-6474. doi : 10.1523/JNEUROSCI.2800-17.2018. eprint : https://www.jneurosci.org/content/38/40/8526.full.pdf. url : https://www.jneurosci.org/content/38/40/8526.

[Xu+18] Dan Xu, Wanli Ouyang, Xiaogang Wang et Nicu Sebe. « PAD-Net : Multi-Tasks Guided Prediction-and-Distillation Network for Simultaneous Depth Estimation and Scene Parsing ». In : *CoRR* abs/1805.04409 (2018). arXiv : 1805.04409. url : http://arxiv.org/abs/1805.04409.

[Zhe18] Wenjie Zheng. « A distributed Frank-Wolfe framework for trace norm minimization via the bulk synchronous parallel model ». Theses. Sorbonne Université, juin 2018. URL : https://tel.archives-ouvertes.fr/tel-02134166.

[DG17] Dheeru Dua et Casey Graff. *UCI Machine Learning Repository*. 2017. URL : http://archive.ics.uci.edu/ml.

[EMU17] Gabriel Eilertsen, Rafal K Mantiuk et Jonas Unger. « A comparative review of tone-mapping algorithms for high dynamic range video ». In : *Computer Graphics Forum*. T. 36. 2. Wiley Online Library. 2017, p. 565-592.

[For+17] N Iandola Forrest, Han Song, W Matthew, Ashraf Khalid et J William Dally. « SqueezeNet : AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size ». In : *ICLR'17 conference proceedings*. 2017, p. 207-212.

[Hol+17] Tim Lukas Holch, Idan Shilon, Matthias Büchele, Tobias Fischer, Stefan Funk, Nils Groeger, David Jankowsky, Thomas Lohse, Ullrich Schwanke et Philipp Wagner. « Probing Convolutional Neural Networks for Event Reconstruction in {\gamma}-Ray Astronomy with Cherenkov Telescopes ». In : *arXiv preprint arXiv :1711.06298* (2017).

[Jég+17] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero et Yoshua Bengio. « The one hundred layers tiramisu : Fully convolutional densenets for semantic segmentation ». In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, p. 11-19.

[Kai+17] Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones et Jakob Uszkoreit. « One Model To Learn Them All ». In : *CoRR* abs/1706.05137 (2017). arXiv : 1706.05137. URL : http://arxiv.org/abs/1706.05137.

[LTG17] Pierre-Jean Lapray, Jean-Baptiste Thomas et Pierre Gouton. « High Dynamic Range Spectral Imaging Pipeline For Multispectral Filter Array Cameras ». In : *Sensors* 17.6 (juil. 2017), p. 1281. ISSN : 1424-8220. DOI : 10.3390/s17061281. URL : http://dx.doi.org/10.3390/s17061281.

[Li+17] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu et Dan Feng. « An All-in-One Network for Dehazing and Beyond ». In : *CoRR* abs/1707.06543 (2017).

[SSA17] Konstantin Shmelkov, Cordelia Schmid et Karteek Alahari. « Incremental learning of object detectors without catastrophic forgetting ». In : *Proceedings of the IEEE International Conference on Computer Vision*. 2017, p. 3400-3409.

[ST17] Ravid Shwartz-Ziv et Naftali Tishby. « Opening the Black Box of Deep Neural Networks via Information ». In : *CoRR* abs/1703.00810 (2017). arXiv : 1703.00810. URL : http://arxiv.org/abs/1703.00810.

[Zha+17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht et Oriol Vinyals. « Understanding deep learning requires rethinking generalization ». In : *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 2017. URL : https://openreview.net/forum?id=Sy8gdB9xx.

[ZL17] Barret Zoph et Quoc V. Le. « Neural Architecture Search with Reinforcement Learning ». In : *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 2017. URL : https://arxiv.org/abs/1611.01578.

[Cai+16] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing et Dacheng Tao. « DehazeNet : An End-to-End System for Single Image Haze Removal ». In : *CoRR* abs/1601.07661 (2016).

[Che+16] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir et al. « Wide & deep learning for recommender systems ». In : *Proceedings of the 1st workshop on deep learning for recommender systems*. ACM. 2016, p. 7-10.

[El 16] Jessica El Khoury. « Model and quality assessment of single image dehazing ». Thèse de doct. Déc. 2016.

[ETM16] Jessica El Khoury, Jean-Baptiste Thomas et Alamin Mansouri. « A Color Image Database for Haze Model and Dehazing Methods Evaluation ». In : *Image and Signal Processing*. Sous la dir. d'Alamin Mansouri, Fathallah Nouboud, Alain Chalifour, Driss Mammass, Jean Meunier et Abderrahim Elmoataz. Cham : Springer International Publishing, 2016, p. 109-117. ISBN : 978-3-319-33618-3.

[GBC16] Ian Goodfellow, Yoshua Bengio et Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[Hah16] Christopher Hahne. « The Standard Plenoptic Camera : Applications of a Geometrical Light Field Model ». Thèse de doct. Univ. of Bedfordshire, jan. 2016.

[Han+16] Song Han, Jeff Pool, Sharan Narang, Huizi Mao, Shijian Tang, Erich Elsen, Bryan Catanzaro, John Tran et William J Dally. « DSD : regularizing deep neural networks with dense-sparse-dense training flow ». In : *arXiv preprint arXiv :1607.04381* 3.6 (2016).

[He+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun. « Identity Mappings in Deep Residual Networks ». In : *arXiv preprint arXiv :1603.05027* (2016).

[Hou+16] Xianxu Hou, LinLin Shen, Ke Sun et Guoping Qiu. « Deep Feature Consistent Variational Autoencoder ». In : *CoRR* abs/1610.00291 (2016). arXiv : 1610.00291. URL : http://arxiv.org/abs/1610.00291.

[HLW16] Gao Huang, Zhuang Liu et Kilian Q. Weinberger. « Densely Connected Convolutional Networks ». In : *CoRR* abs/1608.06993 (2016).

[LS16] Zhongping Lee et Shaoling Shang. « Visibility : How applicable is the century-old Koschmieder model ? » In : *Journal of the Atmospheric Sciences* 73.11 (2016), p. 4573-4581.

[Mal16] Stéphane Mallat. « Understanding deep convolutional networks ». In : *Philosophical Transactions of the Royal Society of London Series A* 374.2065 (avr. 2016), p. 20150203. DOI : 10.1098/rsta.2015.0203. arXiv : 1601.04920 [stat.ML].

[MWK16] Adam H. Marblestone, Greg Wayne et Konrad P. Kording. « Toward an Integration of Deep Learning and Neuroscience ». In : *Frontiers in Computational Neuroscience* 10 (2016), p. 94. ISSN : 1662-5188. DOI : 10.3389/fncom.2016.00094. URL : https://www.frontiersin.org/article/10.3389/fncom.2016.00094.

[Sel+16] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh et Dhruv Batra. « Grad-CAM : Why did you say that ? Visual Explanations from Deep Networks via Gradient-based Localization ». In : *CoRR* abs/1610.02391 (2016). arXiv : 1610.02391. URL : http://arxiv.org/abs/1610.02391.

[XJC16] Caiming Xiong, David M Johnson et Jason J Corso. « Active clustering with model-based uncertainty reduction ». In : *IEEE transactions on pattern analysis and machine intelligence* 39.1 (2016), p. 5-17.

[ZK16] Sergey Zagoruyko et Nikos Komodakis. « Wide Residual Networks ». In : *British Machine Vision Conference 2016*. York, France : British Machine Vision Association, jan. 2016. DOI : 10.5244/C.30.87. URL : https://hal-enpc.archives-ouvertes.fr/hal-01832503.

[Cho+15] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous et Yann LeCun. « The loss surfaces of multilayer networks ». In : *Artificial Intelligence and Statistics*. 2015, p. 192-204.

[HMD15] Song Han, Huizi Mao et William J Dally. « Deep compression : Compressing deep neural networks with pruning, trained quantization and huffman coding ». In : *arXiv preprint arXiv :1510.00149* (2015).

[HVD15] Geoffrey Hinton, Oriol Vinyals et Jeff Dean. « Distilling the knowledge in a neural network ». In : *arXiv preprint arXiv :1503.02531* (2015).

[Hu+15] Wei Hu, Yangyu Huang, Wei Li, Fan Zhang et Heng-Chao Li. « Deep Convolutional Neural Networks for Hyperspectral Image Classification ». In : *J. Sensors* 2015 (2015), 258619 :1-258619 :12. DOI : 10.1155/2015/258619. URL : https://doi.org/10.1155/2015/258619.

[IS15] Sergey Ioffe et Christian Szegedy. « Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift ». In : *International Conference on Machine Learning*. 2015, p. 448-456.

[KZS15] Gregory Koch, Richard Zemel et Ruslan Salakhutdinov. « Siamese neural networks for one-shot image recognition ». In : *ICML deep learning workshop*. T. 2. 2015.

[LSW15] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby et Ole Winther. « Autoencoding beyond pixels using a learned similarity metric ». In : *CoRR* abs/1512.09300 (2015). arXiv : 1512.09300. URL : http://arxiv.org/abs/1512.09300.

[Mak+15] K. Makantasis, K. Karantzalos, D. Anastasios et D. Nikolaos. « Deep supervised learning for hyperspectral data classification through convolutional neural networks ». In : *International Geoscience and Remote Sensing Symposium, Milan, Italy*. 2015.

[Mel+15] Jaime Melendez, Bram van Ginneken, Pragnya Maduskar, Rick HHM Philipsen, Helen Ayles et Clara I Sánchez. « On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis ». In : *Ieee transactions on medical imaging* 35.4 (2015), p. 1013-1024.

[Su+15] Hang Su, Zhaozheng Yin, Seungil Huh, Takeo Kanade et Jun Zhu. « Interactive cell segmentation based on active and semi-supervised learning ». In : *IEEE transactions on medical imaging* 35.3 (2015), p. 762-777.

[Voi15] Nicolas Voiron. « Structuration de bases multimédia pour une exploration visuelle ». ThÃ¨se de doctorat dirigÃ©e par Lambert, Patrick et BenoÃ®t, Alexandre Sciences et techniques de l'information et de la communication, traitement de lâinformation Grenoble Alpes 2015. Thèse de doct. 2015. URL : http://www.theses.fr/2015GREAA036/document.

[CMS14] Corinna Cortes, Mehryar Mohri et Umar Syed. « Deep Boosting ». In : *Proceedings of the 31st International Conference on Machine Learning*. Sous la dir. d'Eric P. Xing et Tony Jebara. T. 32. Proceedings of Machine Learning Research 2. Bejing, China : PMLR, juil. 2014, p. 1179-1187. URL : http://proceedings.mlr.press/v32/cortesb14.html.

[Goo+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville et Yoshua Bengio. « Generative Adversarial Nets ». In : *Advances in Neural Information Processing Systems 27*. Sous la dir. de Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence et K. Q. Weinberger. Curran Associates, Inc., 2014, p. 2672-2680. URL : http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[Kav+14] K. Kavitha, P. Nivedha, S. Arivazhagan et P. Palniladevi. « Wavelet transform based land cover classification of hyperspectral images ». In : *2014 International Conference on Communication and Network Technologies*. Déc. 2014, p. 109-112. DOI : 10.1109/CNT.2014.7062735.

[Liu+14] Ningning Liu, Emmanuel Dellandréa, Bruno Tellez et Liming Chen. « A Selective Weighted Late Fusion for Visual Concept Recognition ». In : *Fusion in Computer Vision : Understanding Complex Visual Content*. Sous la dir. de Bogdan Ionescu, Jenny Benois-Pineau, Tomas Piatrik et Georges Quénot. Cham : Springer International Publishing, 2014, p. 1-28. ISBN : 978-3-319-05696-8. DOI : 10.1007/978-3-319-05696-8_1. URL : https://doi.org/10.1007/978-3-319-05696-8_1.

[LSD14] Jonathan Long, Evan Shelhamer et Trevor Darrell. « Fully Convolutional Networks for Semantic Segmentation ». In : *CoRR* abs/1411.4038 (2014). arXiv : 1411.4038. URL : http://arxiv.org/abs/1411.4038.

[Nar+14] Manish Narwaria, Rafal Mantiuk, Matthieu Perreira Da Silva et Patrick Le Callet. « HDR-VDP-2.2 : a calibrated method for objective quality prediction of high-dynamic range and standard images : a calibrated method for objective quality prediction of high-dynamic range and standard images ». In : *Journal of Electronic Imaging* 24.1 (déc. 2014), p. 010501. DOI : 10.1117/1.JEI.24.1.010501. URL : https://hal.archives-ouvertes.fr/hal-01149491.

[Yos+14] Jason Yosinski, Jeff Clune, Yoshua Bengio et Hod Lipson. « How Transferable Are Features in Deep Neural Networks ? » In : *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada : MIT Press, 2014, p. 3320-3328. URL : http://dl.acm.org/citation.cfm?id=2969033.2969197.

[ZF14] Matthew D. Zeiler et Rob Fergus. « Visualizing and Understanding Convolutional Networks ». In : *Computer Vision – ECCV 2014*. Sous la dir. de David Fleet, Tomas Pajdla, Bernt Schiele et Tinne Tuytelaars. Cham : Springer International Publishing, 2014, p. 818-833. ISBN : 978-3-319-10590-1.

[KW13] Diederik P Kingma et Max Welling. « Auto-Encoding Variational Bayes ». In : *arXiv e-prints*, arXiv :1312.6114 (déc. 2013), arXiv :1312.6114. arXiv : 1312.6114 [stat.ML].

[LTC13] Marc T Law, Nicolas Thome et Matthieu Cord. « Quadruplet-Wise Image Similarity Learning ». In : *IEEE International Conference on Computer Vision (ICCV)*. Sydney, Australia, déc. 2013, p. 249-256. DOI : 10.1109/ICCV.2013.38. URL : https://hal.archives-ouvertes.fr/hal-01094069.

[Str13] Sabin Tiberius Strat. « Analyse et interprÃ©tation de scÃ¨nes visuelles par approches collaboratives ». 2013GRENA026. Thèse de doct. 2013. URL : http://www.theses.fr/2013GRENA026/document.

[Hin+12] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever et Ruslan Salakhutdinov. « Improving neural networks by preventing co-adaptation of feature detectors ». In : *CoRR* abs/1207.0580 (2012). arXiv : 1207.0580. URL : http://arxiv.org/abs/1207.0580.

[KSH12] Alex Krizhevsky, Ilya Sutskever et Geoffrey E. Hinton. « ImageNet Classification with Deep Convolutional Neural Networks ». In : *Advances in Neural Information Processing Systems 25 : 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.* 2012, p. 1106-1114. URL : http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

[Kum+12] Uttam Kumar, Kumar S. Raja, Chiranjit Mukhopadhyay et T.V. Ramachandra. « A Neural Network Based Hybrid Mixture Model to Extract Information from Non-linear Mixed Pixels ». In : *Information* 3.3 (2012), p. 420-441. ISSN : 2078-2489. DOI : 10.3390/info3030420. URL : http://www.mdpi.com/2078-2489/3/3/420.

[NOF12] Sobhan Naderi Parizi, J.G. Oberlin et P.F. Felzenszwalb. « Reconfigurable models for scene recognition ». In : juin 2012, p. 2775-2782. ISBN : 978-1-4673-1226-4. DOI : 10.1109/CVPR.2012.6248001.

[Pin+12] Julien Pinquier, Svebor Karaman, Laetitia Letoupin, Patrice Guyot, Rémi Mégret, Jenny Benois-Pineau, Yann Gaestel et Jean-François Dartigues. « Strategies for multiple feature fusion with Hierarchical HMM : application to activity recognition from wearable audiovisual sensors ». In : *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE. 2012, p. 3192-3195.

[RH12] Syama Sundar Rangapuram et Matthias Hein. « Constrained 1-Spectral Clustering ». In : *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Sous la dir. de Neil D. Lawrence et Mark Girolami. T. 22. Proceedings of Machine Learning Research. La Palma, Canary Islands : PMLR, avr. 2012, p. 1143-1151. url : http://proceedings.mlr.press/v22/sundar12.html.

[Tar+12] J.-P. Tarel, N. Hautière, L. Caraffa, A. Cord, H. Halmaoui et D. Gruyer. « Vision Enhancement in Homogeneous and Heterogeneous Fog ». In : *IEEE Intelligent Transportation Systems Magazine* 4.2 (2012), p. 6-20.

[BDP11] Nicolas Ballas, Betrand Delezoide et Françoise Prêteux. « Trajectories based descriptor for dynamic events annotation ». In : *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*. ACM. 2011, p. 13-18.

[HST11] K. He, J. Sun et X. Tang. « Single Image Haze Removal Using Dark Channel Prior ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.12 (déc. 2011), p. 2341-2353. issn : 0162-8828. doi : 10.1109/TPAMI.2010.168.

[Jai10] Anil K. Jain. « Data clustering : 50 years beyond K-means ». In : *Pattern Recognition Letters* 31.8 (2010). Award winning papers from the 19th International Conference on Pattern Recognition (ICPR), p. 651-666. issn : 0167-8655. doi : https://doi.org/10.1016/j.patrec.2009.09.011. url : http://www.sciencedirect.com/science/article/pii/S0167865509002323.

[PY10] S. J. Pan et Q. Yang. « A Survey on Transfer Learning ». In : *IEEE Transactions on Knowledge and Data Engineering* 22.10 (oct. 2010), p. 1345-1359. issn : 1041-4347. doi : 10.1109/TKDE.2009.191.

[Tar+10] J.-P. Tarel, N. Hautière, A. Cord, D. Gruyer et H. Halmaoui. « Improved Visibility of Road Scene Images under Heterogeneous Fog ». In : *Proceedings of IEEE Intelligent Vehicle Symposium (IV'2010)*. San Diego, California, USA, 2010, p. 478-485.

[BFC09] Gabriel J. Brostow, Julien Fauqueur et Roberto Cipolla. « Semantic object classes in video : A high-definition ground truth database ». In : *Pattern Recognition Letters* 30.2 (2009). Video-based Object and Event Analysis, p. 88-97. issn : 0167-8655. doi : https://doi.org/10.1016/j.patrec.2008.04.005. url : http://www.sciencedirect.com/science/article/pii/S0167865508001220.

[CH09] Ming-Yu Chen et Alexander Hauptmann. « MoSIFT : Recognizing Human Actions in Surveillance Videos ». In : (2009).

[XLJ09] Z. Xu, X. Liu et N. Ji. « Fog Removal from Color Images using Contrast Limited Adaptive Histogram Equalization ». In : *2009 2nd International Congress on Image and Signal Processing*. Oct. 2009, p. 1-5. doi : 10.1109/CISP.2009.5301485.

[MS08] RafaÅ Mantiuk et Hans-Peter Seidel. « Modeling a Generic Tone-mapping Operator ». In : *Computer Graphics Forum* 27.2 (2008), p. 699-708. doi : 10.1111/j.1467-8659.2008.01168.x. eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2008.01168.x. url : https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2008.01168.x.

[YKA08] Emine Yilmaz, Evangelos Kanoulas et Javed A. Aslam. « A simple and efficient sampling method for estimating AP and NDCG ». In : *In SIGIR â08 : Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 603â610*. 2008.

[Ben+07] Yoshua Bengio, Pascal Lamblin, Dan Popovici et Hugo Larochelle. « Greedy layer-wise training of deep networks ». In : *Advances in neural information processing systems*. 2007, p. 153-160.

[Ert+07] Seyda Ertekin, Jian Huang, Leon Bottou et Lee Giles. « Learning on the border : active learning in imbalanced data classification ». In : *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM. 2007, p. 127-136.

[LSP06] Svetlana Lazebnik, Cordelia Schmid et Jean Ponce. « Beyond bags of features : spatial pyramid matching for recognizing natural scene categories ». In : *IEEE Conference on Computer Vision & Pattern Recognition (CPRV '06)*. New York, United States : IEEE Computer Society, juin 2006, p. 2169-2178. doi : 10.1109/CVPR.2006.68. url : https://hal.inria.fr/inria-00548585.

[CHL+05] Sumit Chopra, Raia Hadsell, Yann LeCun et al. « Learning a similarity metric discriminatively, with application to face verification ». In : *CVPR (1)*. 2005, p. 539-546.

[Siv+05] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman et William T. Freeman. « Discovering objects and their location in images ». In : *In ICCV*. 2005.

[Low04] David G. Lowe. « Distinctive Image Features from Scale-Invariant Keypoints ». In : *International Journal of Computer Vision* 60.2 (nov. 2004), p. 91-110. issn : 1573-1405. doi : 10.1023/B:VISI.0000029664.99615.94. url : https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[LL03] Ivan Laptev et Tony Lindeberg. « Space-time Interest Points ». In : *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*. ICCV '03. Washington, DC, USA : IEEE Computer Society, 2003, p. 432-. isbn : 0-7695-1950-4. url : http://dl.acm.org/citation.cfm?id=946247.946605.

[DV02] Minh N Do et Martin Vetterli. « Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance ». In : *IEEE transactions on image processing* 11.2 (2002), p. 146-158.

[FLW02] Raanan Fattal, Dani Lischinski et Michael Werman. « Gradient Domain High Dynamic Range Compression ». In : *ACM Trans. Graph.* 21.3 (juil. 2002), p. 249-256. issn : 0730-0301. doi : 10.1145/566654.566573. url : http://doi.acm.org/10.1145/566654.566573.

[NK00] Kwong Bor Ng et Paul B. Kantor. « Predicting the effectiveness of naïve data fusion on the basis of system characteristics ». In : *JASIS* 51 (2000), p. 1177-1189.

[BM98] Avrim Blum et Tom Mitchell. « Combining Labeled and Unlabeled Data with Co-training ». In : *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. COLT' 98. Madison, Wisconsin, USA : ACM, 1998, p. 92-100. isbn : 1-58113-057-0. doi : 10.1145/279943.279962. url : http://doi.acm.org/10.1145/279943.279962.

[IKN98] L. Itti, C. Koch et E. Niebur. « A model of saliency-based visual attention for rapid scene analysis ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.11 (nov. 1998), p. 1254-1259. issn : 0162-8828. doi : 10.1109/34.730558.

[Kit+98] Josef Kittler, Mohamad Hatef, Robert P. W. Duin et Jiri Matas. « On combining classifiers ». In : *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (1998), p. 226-239.

[SB+98] Richard S Sutton, Andrew G Barto et al. *Introduction to reinforcement learning*. T. 2. 4. MIT press Cambridge, 1998.

[Pod97] J. Podani. « A MEASURE OF DISCORDANCE FOR PARTIALLY RANKED DATA WHEN PRESENCE/ABSENCE IS ALSO MEANINGFUL ». In : *Coenoses* 12.2/3 (1997), p. 127-130. issn : 03939154. url : http://www.jstor.org/stable/43461201.

[OS96] David W Opitz et Jude W Shavlik. « Actively Searching for an Effective Neural Network Ensemble ». In : *Connection Science* 8.3-4 (1996), p. 337-354. doi : 10.1080/095400996116802. eprint : https://doi.org/10.1080/095400996116802. url : https://doi.org/10.1080/095400996116802.

[JT94] Jianbo Shi et Tomasi. « Good features to track ». In : *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Juin 1994, p. 593-600. doi : 10.1109/CVPR.1994.323794.

[HS93] Babak Hassibi et David G Stork. « Second order derivatives for network pruning : Optimal brain surgeon ». In : *Advances in neural information processing systems*. 1993, p. 164-171.

[MC89] Michael McCloskey et Neal J. Cohen. « Catastrophic Interference in Connectionist Networks : The Sequential Learning Problem ». In : sous la dir. de Gordon H. Bower. T. 24. Psychology of Learning and Motivation. Academic Press, 1989, p. 109-165. doi : https://doi.org/10.1016/S0079-7421(08)60536-8. url : http://www.sciencedirect.com/science/article/pii/S0079742108605368.

[Can86] J. Canny. « A Computational Approach to Edge Detection ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8.6 (nov. 1986), p. 679-698. issn : 0162-8828. doi : 10.1109/TPAMI.1986.4767851.

[WV85] Michael A. Webster et Russell L. De Valois. « Relationship between spatialfrequency and orientation tuning of striate-cortex cells ». In : *J. Opt. Soc. Am. A* (1985).

[FB81] Martin A. Fischler et Robert C. Bolles. « Random Sample Consensus : A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography ». In : *Commun. ACM* 24.6 (juin 1981), p. 381-395. issn : 0001-0782. doi : 10.1145/358669.358692. url : http://doi.acm.org/10.1145/358669.358692.

[LK81] Bruce D. Lucas et Takeo Kanade. « An Iterative Image Registration Technique with an Application to Stereo Vision ». In : *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'81. Vancouver, BC, Canada : Morgan Kaufmann Publishers Inc., 1981, p. 674-679. url : http://dl.acm.org/citation.cfm?id=1623264.1623280.

[Sha76] Glenn Shafer. *A mathematical theory of evidence*. T. 42. Princeton university press, 1976.

[HW62] D. H. Hubel et T. N. Wiesel. « Receptive fields, binocular interaction and functional architecture in the cat's visual cortex ». In : *The Journal of Physiology* 160.1 (1962), p. 106-154. doi : 10.1113/jphysiol.1962.sp006837. eprint : https://physoc.onlinelibrary.wiley.com/doi/pdf/10.1113/jphysiol.1962.sp006837. url : https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1962.sp006837.

[Bar53] H. B. Barlow. « Summation and inhibition in the frog's retina ». In : *The Journal of Physiology* 119.1 (1953), p. 69-88. doi : 10.1113/jphysiol.1953.sp004829. eprint : https://physoc.onlinelibrary.wiley.com/doi/pdf/10.1113/jphysiol.1953.sp004829. url : https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1953.sp004829.

[MP43] Warren S. McCulloch et Walter Pitts. « A logical calculus of the ideas immanent in nervous activity ». In : *The bulletin of mathematical biophysics* 5.4 (déc. 1943), p. 115-133. issn : 1522-9602. doi : 10.1007/BF02478259. url : https://doi.org/10.1007/BF02478259.

[Tur37] Alan Mathison Turing. « On computable numbers, with an application to the Entscheidungsproblem ». In : *Proceedings of the London mathematical society* 2.1 (1937), p. 230-265.

[Kos24] Harald Koschmieder. « Theorie der horizontalen Sichtweite ». In : *Beitrage zur Physik der freien Atmosphare* (1924), p. 33-53.

Chapitre 6

Liste des publications de l'auteur (post-thèse)

Contribution à ouvrage

[Str+12a] Tiberius Strat, **Benoit Alexandre**, Hervé Bredin, Georges Quenot et Patrick Lambert. « Hierarchical Late Fusion for Concept Detection in Videos ». In : *ECCV 2012 - 12th European Conference on Computer Vision*. Sous la dir. d'Andrea Fusiello, Vittorio Murino et Rita Cucchiara. T. 7585. Lecture Notes in Computer Science (LNCS). Oral session 1 : WS21 - Workshop on Information Fusion in Computer Vision for Concept Recognition. Firenze, Italy : Springer Berlin, oct. 2012, p. 335-344. doi : 10.1007/978-3-642-33885-4_34. url : https://hal.archives-ouvertes.fr/hal-00732740.

Revues avec comité de lecture

[Bra+19] Sourour Brahimi, **Benoit Alexandre**, Patrick Lambert, Chokri Ben Amar et Najib Ben Aoun. « Semantic Segmentation using Reinforced Fully Convolutional DenseNet with Multiscale Kernel ». In : *Multimedia Tools and Applications* (2019). url : https://hal.archives-ouvertes.fr/hal-02063128.

[Ben+18b] Amina Ben Hamida, **Benoit Alexandre**, Patrick Lambert et Chokri Ben Amar. « Three dimensional Deep Learning approach for remote sensing image classification ». In : *IEEE Transactions on Geoscience and Remote Sensing* 56.8 (avr. 2018), p. 4420-4434. doi : 10.1109/TGRS.2018.2818945. url : https://hal.archives-ouvertes.fr/hal-01814542.

[Str+14] Tiberius Strat, **Benoit Alexandre**, P. Lambert et Alice Caplier. « Retina enhanced SURF descriptors for spatio-temporal concept detection ». In : *Multimedia Tools and Applications* 69.2 (2014), pp. 443-469. doi : 10.1007/s11042-012-1280-0. url : https://hal.archives-ouvertes.fr/hal-00760192.

[Bet+12] Salim Bettahar, Amin Boudghene Stambouli, Patrick Lambert et **Benoit Alexandre**. « PDE Based Enhancement of Color Images in RGB Space ». In : *IEEE Transactions on Image Processing* 21.5 (mai 2012), p. 2500-2512. DOI : 10.1109/TIP.2011.2177844. URL : https://hal.archives-ouvertes.fr/hal-00732708.

[Ben+10] **Benoit Alexandre**, Alice Caplier, Barthélémy Durette et Jeanny Hérault. « Using Human Visual System Modeling for Bio-inspired Low Level Image Processing ». In : *Computer Vision and Image Understanding* 114.7 (2010), p. 758-773. DOI : 10.1016/j.cviu.2010.01.011. URL : https://hal.archives-ouvertes.fr/hal-00377222.

[Ben+08a] **Benoit Alexandre**, Patrick Le Callet, Patrizio Campisi et Romain Cousseau. « Quality Assessment of Stereoscopic Images ». In : *EURASIP Journal on Image and Video Processing* 2008 (déc. 2008), Article ID 659024, 13 pages. DOI : 10.1155/2008/659024. URL : https://hal.archives-ouvertes.fr/hal-00362891.

Communications à des congrès internationaux avec comité de lecture

[Jac+19b] Mikaël Jacquemont, Luca Antiga, Thomas Vuillaume, Giorgia Silvestri, **Benoit Alexandre**, Patrick Lambert et Gilles Maurin. « Indexed Operations for Non-rectangular Lattices Applied to Convolutional Neural Networks ». In : *VISAPP, 14th International Conference on Computer Vision Theory and Applications*. Prague, Czech Republic, fév. 2019. URL : https://hal.archives-ouvertes.fr/hal-02087157.

[Mas+19] Olga Maslova, Louis Klein, Damien **Benoit, Alexandre** Dabernat et Patrick Lambert. « Receipt automatic reader ». In : *Content-Based Multimedia Indexing (CBMI) 2019*. Dublin, Ireland, sept. 2019. URL : https://hal.archives-ouvertes.fr/hal-02196644.

[Bra+18] Sourour Brahimi, Najib Ben Aoun, Chokri Ben Amar, **Benoit Alexandre** et Patrick Lambert. « Multiscale Fully Convolutional DenseNet for Semantic Segmentation ». In : *WSCG 2018, International Conference on Computer Graphics, Visualization and Computer Vision*. Pilsen, Czech Republic, mai 2018. URL : https://hal.archives-ouvertes.fr/hal-01786688.

[CBT18] Leonel Cuevas, **Benoit Alexandre** et Jean-Baptiste Thomas. « Deep learning for dehazing : Comparison and analysis ». In : *Colour and Visual Computing Symposium (CVCS)*. Gjøvik, Norway, sept. 2018. URL : https://hal.archives-ouvertes.fr/hal-01823911.

[Ben+17] Amina Ben Hamida, **Benoit Alexandre**, P. Lambert, L Klein, Chokri Ben Amar, N. Audebert et S. Lefèvre. « Deep Learning For Semantic Segmentation Of Remote Sensing Images With Rich Spectral Content ». In : *IEEE International Geoscience and Remote Sensing Symposium*. Fort Worth, United States, juil. 2017. URL : https://hal.archives-ouvertes.fr/hal-01654187.

[Rao+17] Rizlène Raoui-Outach, Cécile Million-Rousseau, **Benoit Alexandre** et Patrick Lambert. « Deep Learning for automatic sale receipt understanding ». In : *International Conference on Image Processing Theory, Tools and Applications*. Montreal, Canada, nov. 2017. URL : https://hal.archives-ouvertes.fr/hal-01654191.

[Ben+16] Amina Ben Hamida, **Benoit Alexandre**, Patrick Lambert et Chokri Ben-Amar. « Deep Learning Approach For Remote Sensing Image Analysis ». In : *Big Data from Space (BiDS'16)*. Santa Cruz de Tenerife, Spain : Publications Office of the European Union, mar. 2016, p. 133. DOI : 10.2788/854791. URL : https://hal.archives-ouvertes.fr/hal-01370161.

[Voi+16] Nicolas Voiron, **Benoit Alexandre**, Patrick Lambert et Bogdan Ionescu. « Deep Learning vs Spectral Clustering into an active clustering with pairwise constraints propagation ». In : *14th International Workshop on Content-based Multimedia Indexing (CBMI 2016)*. Sous la dir. d'IEEE. 2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI). Bucarest, Romania, juin 2016. URL : https://hal.archives-ouvertes.fr/hal-01319969.

[Ben+15] Amina Ben Hamida, **Benoit Alexandre**, Patrick Lambert et Chokri Ben-Amar. « Could Multimedia approaches help Remote Sensing Analysis ? » In : *IIM 2015 Conference Image Information Mining : Earth Observation meets Multimedia*. CEOSpaceTech, Bucharest, Romania. Bucharest, Romania, oct. 2015. URL : https://hal.archives-ouvertes.fr/hal-01238247.

[Voi+15b] Nicolas Voiron, **Benoit Alexandre**, Andrei Filip, Patrick Lambert et Bogdan Ionescu. « Semi-supervised Spectral Clustering with automatic propagation of pairwise constraints ». In : *13th International Workshop on Content-Based Multimedia Indexing (CBMI 2015)*. Prague, Czech Republic, juin 2015, p. 1-6. DOI : 10.1109/CBMI.2015.7153608. URL : https://hal.archives-ouvertes.fr/hal-01174875.

[SBL14b] Tiberius Strat, **Benoit Alexandre** et Patrick Lambert. « Bags of Trajectory Words for video indexing ». In : *CBMI 2014*. Klagenfuhrt, Austria, juin 2014. DOI : 10.1109/CBMI.2014.6849820. URL : https://hal.archives-ouvertes.fr/hal-01096109.

[SBL14c] Tiberius Strat, **Benoit Alexandre** et Patrick Lambert. « Retina enhanced bag of words descriptors for video classification ». In : *EUSIPCO 2014*. Lisbon, Portugal, sept. 2014. URL : https://hal.archives-ouvertes.fr/hal-01096114.

[SBL13] Tiberius Strat, **Benoit Alexandre** et Patrick Lambert. « Retina enhanced SIFT descriptors for video indexing ». In : *11th International Workshop on Content-Based Multimedia Indexing (CBMI 2013)*. Veszprem, Hungary, juin 2013, p. 201-206. URL : https://hal.archives-ouvertes.fr/hal-00875044.

[Str+12b] Tiberius Strat, **Benoit Alexandre**, Patrick Lambert et Alice Caplier. « Retina-Enhanced SURF Descriptors for Semantic Concept Detection in Videos ». In : *3rd International Conference on Image Processing Theory, Tools and Applications (IPTA 2012)*. Istanbul, Turkey, oct. 2012, p. 1-6. URL : https://hal.archives-ouvertes.fr/hal-00732736.

[VBL12] Nicolas Voiron, **Benoit Alexandre** et Patrick Lambert. « Automatic difference measure between movies using dissimilarity measure fusion and rank correlation coefficients ». In : *10th International Workshop on Content-Based Multimedia Indexing (CBMI 2012)*. Annecy, France, juin 2012, p. 1-6. URL : https://hal.archives-ouvertes.fr/hal-00732734.

[Bet+11] Salim Bettahar, Amin Boudghene Stambouli, P. Lambert et **Benoit Alexandre**. « Enhancement of Multi-Valued Images Using PDE Coupling ». In : *European Signal Processing Conference (EUSIPCO-2011)*. Spain, août 2011, CDROM. URL : https://hal.archives-ouvertes.fr/hal-00623702.

[Ion+11] Bogdan Ionescu, C. Vertan, P. Lambert et **Benoit Alexandre**. « A color-action perceptual approach to the classification of animated movies ». In : *ACM International Conference on Multimedia Retrieval (ICMR 2011)*. Trento, Italy, avr. 2011. DOI : 10.1145/1991996.1992006. URL : https://hal.archives-ouvertes.fr/hal-00623707.

[Ben+09] **Benoit Alexandre**, David Alleysson, Patrick Le Callet et Jeanny Hérault. « Spatio-Temporal Tone Mapping Operator based on a Retina model ». In : *Computational Color Imaging Workshop (CCIW'09)*. LNCS. Saint Etienne, France : Springer, mar. 2009, p. 12-22. URL : https://hal.archives-ouvertes.fr/hal-00366126.

[Ben+08b] **Benoit, Alexandre**, Patrick Le Callet, Patrizio Campisi et Romain Cousseau. « Using disparity for quality assessment of stereoscopic images ». In : *IEEE International Conference on Image Processing, ICIP 2008*. San Diego, United States, oct. 2008. URL : https://hal.archives-ouvertes.fr/hal-00324052.

Communications à des congrès nationaux avec comité de lecture

[Ben+18a] Amina Ben Hamida, **Benoit Alexandre**, Patrick Lambert et Chokri Ben Amar. « Generative Adversarial Network (GAN) for Remote Sensing Images unsupervised Learning ». In : *RFIAP 2018*. AFRIF, SFPT, IEEE GRSS. Marne-la-Vallée, France, juin 2018. URL : https://hal.archives-ouvertes.fr/hal-01970319.

[Rao+16] Rizlène Raoui-Outach, Cécile Million-Rousseau, **Benoit Alexandre** et Patrick Lambert. « Lecture automatique d'un ticket de caisse par vision embarquée sur un téléphone mobile ». In : *RFIA 2016*. Travaux réalisés dans le cadre d'une thèse CIFRE. Clermont-Ferrand, France, juin 2016. URL : https://hal.archives-ouvertes.fr/hal-01319960.

[Voi+15a] Nicolas Voiron, **Benoit Alexandre**, Andrei Filip, Patrick Lambert et Bogdan Ionescu. « Clustering Spectral semi-supervisé avec propagation des contraintes par paires ». In : *12ème COnference en Recherche d'Information et Applications - CORIA*. Paris, France, mar. 2015. URL : https://hal.archives-ouvertes.fr/hal-01143664.

[SBL14a] Sabin Tiberius Strat, **Benoit Alexandre** et Patrick Lambert. « Analyse de trajectoires pour l'indexation sémantique des vidéos à grande échelle ». In : *Reconnaissance de Formes et Intelligence Artificielle (RFIA) 2014*. France, juin 2014. URL : https://hal.archives-ouvertes.fr/hal-00989034.

Communications à des congrès internationaux sans comité de lecture

[Jac+19c] Mikaël Jacquemont, Thomas Vuillaume, **Benoit Alexandre**, Patrick Lambert, Gilles Maurin, Giovanni Lamanna et Ari Brill. « GammaLearn : a Deep Learning framework for IACT data ». In : *ICRC 2019 - 36th International Cosmic Ray Conference*. 2019.

[CTB18] Leonel Cuevas Valeriano, Jean-Baptiste Thomas et **Benoit Alexandre**. « Deep learning for dehazing : Benchmark and analysis ». In : *NOBIM 2018*. Hafjell, Øyer, Norway, mar. 2018. URL : https://hal.archives-ouvertes.fr/hal-01786653.

[Vui+18] Thomas Vuillaume, Mikaël Jacquemont, Luca Antiga, **Benoit Alexandre**, Patrick Lambert, Gilles Maurin, Giorgia Silvestri et the CTA consortium. « GammaLearn - first steps to apply Deep Learning to the Cherenkov Telescope Array data ». In : *CHEP 2018 - 23rd International Conference on Computing in High Energy and Nuclear Physics*. 2018.

[Man+17] Boris Mansencal, Jenny Benois-Pineau, Hervé Bredin, **Benoit Alexandre**, Nicolas Voiron, Patrick Lambert, Hervé Le Borgne, Adrian Popescu, Alexandru Ginsca et Georges Quénot. « IRIM at TRECVID 2017 : Instance Search ». In : *TRECVid workshop 2017*. Gaithersburg, Maryland, United States, nov. 2017. URL : https://hal.archives-ouvertes.fr/hal-01834223.

[Man+16] Boris Mansencal, Jenny Benois-Pineau, Hervé Bredin, **Benoit Alexandre**, Nicolas Voiron, Patrick Lambert et Georges Quénot. « IRIM at TRECVID 2016 : Instance Search ». In : *TRECVid workshop 2016*. TRECVid workshop proceedings. National Institute of Standards and Technology (NIST). Gaithersburg, Maryland, United States, nov. 2016. URL : http://hal.univ-smb.fr/hal-01416953.

[Le +15] Hervé Le Borgne, Philippe Gosselin, David Picard, Miriam Redi, Bernard Mérialdo, Boris Mansencal, Jenny Benois-Pineau, Stéphane Ayache, Abdelkader Hamadi, Bahjat Safadi, Nadia Derbas, Mateusz Budnik, Georges Quénot, Boyang Gao, Chao Zhu, Yuxing Tang, Emmanuel Dellandrea, Charles-Edmond Bichot, Liming Chen, **Benoit Alexandre**, Patrick Lambert et Tiberius Strat. « IRIM at TRECVID 2015 : Semantic Indexing ». In : *Proceedings of TRECVid*. Gaithersburg, MD, United States, nov. 2015. URL : https://hal.archives-ouvertes.fr/hal-01233398.

[Bal+14] Nicolas Ballas, Benjamin Labbé, Hervé Le Borgne, Philippe Gosselin, David Picard, Miriam Redi, Bernard Mérialdo, Boris Mansencal, Jenny Benois-Pineau, Stéphane Ayache, Abdelkader Hamadi, Bahjat Safadi, Nadia Derbas, Mateusz Budnik, Georges Quénot, Boyang Gao, Chao Zhu, Yuxing Tang, Emmanuel Dellandrea, Charles-Edmond Bichot,

Liming Chen, **Benoit Alexandre**, Patrick Lambert et Tiberius Strat. « IRIM at TRECVID 2014 : Semantic Indexing and Instance Search ». In : *Proceedings of TRECVID*. Orlando, United States, nov. 2014. URL : https://hal.archives-ouvertes.fr/hal-01132491.

[Bal+13] Nicolas Ballas, Benjamin Labbé, Hervé Le Borgne, Philippe Gosselin, Miriam Redi, Bernard Merialdo, Rémi Vieux, Boris Mansencal, Jenny Benois-Pineau, Stéphane Ayache, Abdelkader Hamadi, Bahjat Safadi, Thi-Thu-Thuy Vuong, Han Dong, Nadia Derbas, Georges Quénot, Boyang Gao, Chao Zhu, Yuxing Tang, Emmanuel Dellandrea, Charles-Edmond Bichot, Liming Chen, **Benoit Alexandre**, Patrick Lambert et Tiberius Strat. « IRIM at TRECVID 2013 : Semantic Indexing and Instance Search ». In : *Proc. TRECVID Workshop*. Gaithersburg, MD, United States, 2013. URL : https://hal.inria.fr/hal-00953092.

[Bal+12] Nicolas Ballas, Benjamin Labbé, Aymen Shabou, Hervé Le Borgne, Philippe-Henri Gosselin, Miriam Redi, Bernard Merialdo, Hervé Jégou, Jonathan Delhumeau, Rémi Vieux, Boris Mansencal, Jenny Benois-Pineau, Stéphane Ayache, Abdelkader Hamadi, Bahjat Safadi, Franck Thollard, Nadia Derbas, Georges Quénot, Hervé Bredin, Matthieu Cord, Boyang Gao, Chao Zhu, Yuxing Tang, Emmanuel Dellandrea, Charles-Edmond Bichot, Liming Chen, **Benoit Alexandre**, Patrick Lambert, Tiberius Strat, Joseph Razik, Sébastien Paris, Hervé Glotin, Tran Ngoc Trung, Dijana Petrovska-Delacrétaz, Gérard Chollet, Andrei Stoian et Michel Crucianu. « IRIM at TRECVID 2012 : Semantic Indexing and Instance Search ». In : *TRECVID - TREC Video Retrieval Evaluation workshop*. Gaithersburg, MD, United States, nov. 2012, 12p. URL : https://hal.archives-ouvertes.fr/hal-00770258.

Note de recherche

[ABL18] Abdourrahmane Atto, **Benoit Alexandre** et Patrick Lambert. « Hilbert Based Video To Timed-Image Filling and Learning To Recognize Visual Violence ». working paper or preprint. Nov. 2018. URL : https://hal.archives-ouvertes.fr/hal-01920608.

Thèse

[Ben07] **Benoit Alexandre**. « The human visual system as a complete solution for image processing ». Thèse de doct. Institut National Polytechnique de Grenoble - INPG, fév. 2007. URL : https://tel.archives-ouvertes.fr/tel-00193715.