Analyse visuelle sémantique de scènes

Quoc-Cuong Pham

Sorbonne Université

CEA

Unité de recherche **CEA LIST, Laboratoire de Vision et d'Apprentissage pour l'analyse de scène**

Habilitation à diriger les recherches présentée par **Quoc Cuong PHAM**

Soutenue le **22 juin 2021**

Discipline **Sciences de l'ingénieur**

Spécialité **Informatique**

Analyse visuelle sémantique de scènes

De la modélisation des objets à l'analyse de comportement

Composition du jury

Rapporteurs	Stéphane Canu	professeur à l'Institut National des Sciences Appliquées de Rouen
	David Picard	professeur à l'Ecole des Ponts ParisTech
	Jean-Marc Odobez	directeur de recherche à l'Ecole Polytechnique Fédérale de Lausanne
Examinateurs	Mohamed Daoudi	professeur à l'Institut Mines-Télécom de Lille
	Catherine Achard	professeur à Sorbonne Université

Cette HDR a été préparée au

CEA LIST, Laboratoire de Vision et d'Apprentissage pour l'analyse de scène

CEA Saclay NanoInnov
Bâtiment 861-PC 184
91191 Gif-sur-Yvette Cedex France

☎ +33 1 69 08 49 01
Site https://kalisteo.cea.fr/

Analyse visuelle sémantique de scènes
De la modélisation des objets à l'analyse de comportement

Résumé

Dans un monde où l'image est devenue sans conteste un vecteur prééminent d'information, les technologies de reconnaissance visuelle et de compréhension de scène se sont développées ces dernières années à un rythme soutenu dans le but de nous aider à exploiter plus efficacement des quantités toujours plus importantes de données. Portée par les progrès remarquables de l'apprentissage automatique, la disponibilité de corpus d'images annotées et d'outils logiciels, les nouvelles plateformes matérielles de calcul, la vision par ordinateur a connu ces dernières années un essor exceptionnel, permettant aujourd'hui aux acteurs industriels de répondre à de nombreux besoins applicatifs dans beaucoup de domaines : mobilité, sécurité, défense, santé, industrie du futur...
La compréhension automatique des scènes est une tâche complexe dont l'objectif est de produire à partir d'images et de vidéos, à différents niveaux sémantiques, une description précise des éléments d'une scène, une caractérisation de leur évolution dans le temps, et une analyse de leur comportement. Cette vaste problématique recouvre de nombreuses sujets de recherche, chacun d'eux traitant d'un aspect particulier de l'analyse de scène, dont les principaux défis sont la robustesse face à l'immense variété des situations, la précision, la finesse et la pertinence de l'analyse, et la complexité calculatoire.
Dans nos travaux, nous avons cherché à apporter des éléments de réponse aux problèmes de l'analyse sémantique de scènes en explorant, d'une part, de nouvelles méthodes de modélisation des objets et des personnes dans les images, allant de la détection rapide et robuste, à la ré-identification et à la caractérisation des poses 3D à partir d'images 2D, et d'autre part, en proposant des approches directes d'analyse de comportement et de détection d'événements dans les flux vidéos, ne reposant pas sur une individualisation explicite des personnes.

Mots clés : analyse de scène, apprentissage automatique, vision par ordinateur, détection d'objets, ré-identification de personnes, estimation de poses, détection d'événements, analyse de comportements

Abstract

In a world where images have become a preeminent vector of information, visual recognition and scene understanding technologies have developed at a sustained pace over the last few years in order to help us exploit ever larger quantities of data more efficiently. Driven by the remarkable progress of machine learning, the availability of annotated image data sets and reusable software, and new computing hardware platforms, computer vision has experienced exceptional growth in recent years, allowing industrial players to meet many application needs in many areas: mobility, security, defense, health, industry of the future ...
Automatic scene understanding is a complex task whose objective is to produce from images and videos, and at different semantic levels, an accurate description of the elements of a scene, a characterization of their evolution in time, and an analysis of their behavior. This vast problem covers many research topics, each of them dealing with a particular aspect of scene analysis, whose main challenges are robustness in the face of the immense variety of situations, accuracy, finesse and relevance of the analysis, and computational complexity.
In our work, we have sought to provide answers to the problems of semantic scene analysis by exploring, on the one hand, new methods for modeling objects and people in images, ranging from fast and robust detection, to person re-identification and characterization of 3D human poses from 2D images, and on the other hand, by proposing direct approaches to behavior analysis and event detection from video streams, that are not relying on an explicit individualization of people.

Keywords: scene understanding, machine learning, computer vision, object detection, person re-identification, pose estimation, event detection, behavior analysis

CEA LIST, Laboratoire de Vision et d'Apprentissage pour l'analyse de scène
CEA Saclay NanoInnov – Bâtiment 861-PC 184 – 91191 Gif-sur-Yvette Cedex France

Table des matières

Introduction

Les enjeux de l'analyse visuelle sémantique de scènes

Nous sommes aujourd'hui inondés en permanence de flux d'images : images et vidéos provenant des medias, des réseaux sociaux sur internet, des fournisseurs de contenus multimedia pour le loisir ou dans le domaine professionnel. Consommateurs, mais aussi producteurs d'images, nous en générons des quantités phénoménales chaque jour. Les caméras sont partout, dans nos foyers, dans nos voitures, dans les usines, dans la rue... Les chiffres sont vertigineux. Selon GSMA intelligence, on évalue à près de 5,2 milliards le nombre d'utilisateurs de mobile en 2021. La société d'information économique IHS Market projette qu'on comptera plus d'un milliard de caméras de surveillance déployées dans le monde à la fin de l'année.

Ce déluge de données visuelles conforte plus que jamais la nécessité de développer des systèmes automatiques efficaces, capables d'extraire des images les informations utiles et pertinentes, de mettre en place de nouveaux processus de transformation de l'information brute des pixels en une information sémantique concise et mieux exploitable.

L'analyse visuelle sémantique de scènes est un domaine de recherche très actif dont le large spectre couvre la conception, le développement, l'évaluation de méthodes et algorithmes pour la reconnaissance et la description des éléments constitutifs d'une scène, à partir d'images et de vidéos. Par sa complexité et l'étendue des besoins, l'analyse de scènes fait appel à différentes disciplines scientifiques dont les mathématiques (statistiques, optimisation numérique, théorie des graphes, géométrie, etc.), la physique et l'informatique, et plus spécifiquement le traitement de l'image, la vision par ordinateur, et l'intelligence artificielle, notamment l'apprentissage machine.

On ne compte plus les applications qui mettent en jeu des systèmes de vision et des logiciels d'analyse automatique d'images. Les innovations technologiques de la vision et de l'intelligence artificielle irriguent presque tous les secteurs industriels : la mobilité et les transports où les véhicules gagnent en autonomie, l'industrie 5.0 dans laquelle humains et robots vont collaborer étroitement de manière plus fluide et efficace, l'agriculture moderne qui est de plus en plus automatisée et optimisée, la sécurité des biens et des personnes grâce à la vidéo-surveillance intelligente, la défense du futur qui doit se réinventer avec de nouveaux moyens technologiques face à la complexité croissante du monde, la domaine de la santé où la nécessité de fournir des systèmes d'assistance au regard du vieillissement de la population est de plus en plus criante,...

A la croisée des chemins entre les recherches plus fondamentales et théoriques, et les besoins applicatifs, la recherche technologique bénéficie des impulsions venant des deux univers. C'est clairement le cas de la vision pour l'analyse de scène qui, depuis une vingtaine d'années, profite d'une intensification exponentielle des travaux de recherche et voit une accélération des cycles de développement, depuis la formalisation du concept jusqu'au déploiement opérationnel de la technologie. Un meilleur partage de la connaissance, des méthodes et des outils, facilite la progression des chercheurs et des ingénieurs. La mise en place de benchmarks maintenant presque systématique permet la comparaison objective et quantitative des résultats et encourage peut être aussi une forme de "culture de la performance". Il n'en reste pas moins que bien des problèmes

ne sont pas totalement résolus et qu'ils nécessitent encore des efforts de recherche conséquents.

Une tâche complexe et une multiplicité de sous-problèmes

La compréhension de scène par la machine est un problème complexe qui est généralement découpée en sous-problèmes, souvent abordés séparément et parfois de manière conjointe. Ainsi, la détection et la caractérisation des objets dans les images, la segmentation sémantique de scènes, le suivi temporel et la ré-identification d'objets et de personnes, la reconnaissance des actions et des événements sont autant de sujets de recherche qui traitent un aspect particulier de la compréhension de scène.

Le titre d'un article assez ancien de Haritaoglu et al., 1998 qui présentaient un nouveau système d'analyse de scène était formulé avec quatre questions : *W4 : Who ? When ? Where ? What ? A Real Time System for Detecting and Tracking People*. Je reprendrai volontiers ces quatre questions fondamentales en essayant de montrer qu'elles focalisent les objectifs des différentes tâches de perception et de reconnaissance visuelle :

Qui ? Cette question porte sur la reconnaissance des sujets au sens large, objets et personnes, présents dans de la scène et la description de leurs propriétés intrinsèques. Alors que la détection a pour objectif de déterminer la présence d'objets regroupés en catégories, la reconnaissance fine cherche à enrichir la description de ces objets par une sous-catégorisation plus précise et des attributs sémantiques. La ré-identification, quant à elle, réunit plusieurs observations d'une instance/personne sous une même identité. L'importance d'associer entre elles des observations visuelles d'une personne devient évidente lorsque l'on doit interpréter son comportement.

Quand ? L'analyse de scènes dynamiques tente de modéliser les évolutions et les changements qui surviennent. Le suivi des objets permet de les situer dans le temps et dans l'espace de façon continue, l'estimation des déplacements est un premier d'analyse du comportement. Ré-identifier une personne établit une chronologie de sa présence et de son absence en un lieu. La détection des actions et des activités suppose une segmentation temporelle d'une vidéo pour déterminer les intervalles correspondant aux différentes classes d'action/d'activité.

Où ? Outre la dimension temporelle, l'information de localisation spatiale est cruciale. Connaître la position des objets dans leur environnement permet de mieux appréhender leur état et de décrire une situation. Remonter à leur trajectoire et à leur localisation relative donne accès à la compréhension de leurs intentions, de leurs actions et interactions, de leur dynamique. La segmentation sémantique est informative de la structure spatiale de la scène et du contexte.

Quoi ? Cette question porte sur les actions et les activités effectuées par les sujets, les événements observés. On pourrait également poser la question "Comment ?", qui concerne davantage la manière dont les actions ont été effectuées. Ce dernier problème, assez peu souvent traité, implique une précision et une finesse de l'analyse encore plus élevées.

L'analyse de scène embrasse donc de nombreuses problématiques, chacune extrêmement riche en termes de réflexion méthodologique, d'approches innovantes et d'expérimentations. Une partie d'entre elles sont abordées dans les travaux de recherche présentés dans ce mémoire.

Contexte des travaux présentés

Ingénieur chercheur au CEA LIST depuis 2003 dans le domaine de la vision par ordinateur et de l'apprentissage automatique, mon activité de recherche s'inscrit pleinement dans les missions

d'innovation et de transfert technologique du CEA. Elle est partagée entre une recherche appliquée tournée vers la maturation et le transfert de technologies de visions au travers de partenariats industriels, et une recherche plus académique qui se traduit par l'encadrement de chercheurs dans le cadre de projets de ressourcement scientifique (thèses de doctorat, post-doctorats, projets de recherche collaboratifs).

Entre 2003 et 2012, j'ai travaillé, en tant que chercheur et chef de projet, à développer de nouvelles technologies d'analyse vidéo pour les applications de vidéo-surveillance intelligente. Pendant cette période, j'ai exploré au sein de projets de mon laboratoire des approches probabilistes de segmentation fond/forme, des méthodes de détection rapides d'objets fondées sur l'apprentissage de classifieurs de région, et des algorithmes de suivi visuel multi-cibles par filtrage particulaire. Ces travaux ont notamment fait l'objet des premiers transferts technologiques à la société THALES par le biais du laboratoire commun entre le CEA et THALES, le VisionLab, créé en 2008 et toujours actif aujourd'hui.

Depuis 2013, je dirige une équipe de recherche entièrement dédiée aux problématiques de la vision et de l'apprentissage machine pour l'analyse de scène. Cette équipe est formellement devenue un laboratoire en 2019, le Laboratoire de Vision et d'Apprentissage pour l'analyse de scène, qui comprend aujourd'hui une trentaine d'ingénieurs chercheurs. Au titre de responsable de ce laboratoire, j'ai en charge la définition et le suivi de la feuille de route scientifique et la supervision de haut niveau de l'ensemble des projets du laboratoire. Je m'investis néanmoins personnellement de manière plus directe sur plusieurs de ces projets.

Les contributions présentées dans ce mémoire pour l'Habilitation à Diriger les Recherches sont issues d'une sélection de travaux réalisés dans la période de 2012 à 2020, pour lesquels, j'ai encadré trois doctorants, une post-doctorante et un ingénieur chercheur. Les résultats de ces travaux ont donné lieu à 9 publications dont 6 articles de conférence internationale et 3 articles de revue à comité de lecture.

Lignes directrices

Nos recherches sont guidées par l'objectif global de proposer de nouvelles briques de compréhension de scènes qui puissent être implémentées et appliquées dans des cas d'usage variés en assurant :

— une niveau de fiabilité élevé, c'est-à-dire une robustesse et une précision acceptables vis-à-vis des exigences applicatives et une bonne capacité de généralisation des modèles à des contextes et situations variés ;
— une complexité de calcul contenue, compatible avec une exécution en temps réel, ou du moins avec une latence acceptable pour une utilisation interactive.

Bien que le laboratoire possède une expertise sur une large gamme de capteurs de vision (caméras opérant dans le visible, le proche infrarouge et l'infrarouge thermique, caméra stéréoscopiques, caméras 3D, LIDAR), les travaux présentés ici ne concernent que l'exploitation d'images issues de caméras couleur utilisées dans la majorité des cas d'usage de nos projets de recherche.

Nous nous sommes efforcés dans nos travaux de développer des méthodes capables de décrire une scène à différents niveaux sémantiques qui vont de la modélisation des objets et des personnes à l'analyse de leur comportement dans des scènes complexes. Notre démarche a consisté à explorer deux voies en parallèle :

Modélisation élémentaire/individuelle Les objets/les personnes sont détectés et identifiés, leur comportement individuel est ensuite caractérisé et analysé ;

Modélisation directe d'ensemble Les comportements sont reconnus directement à partir de caractéristiques visuelles extraites des images et vidéos, sans passer par une individualisation des personnes.

Les contributions présentées ici donnent quelques pistes de solutions dans les deux directions investiguées. Pour l'approche individuelle, nous présentons des méthodes de détection d'objets, de ré-identification de personnes et d'estimation 3D de leurs poses, qui pourraient aider dans des travaux futurs à analyser leur comportement, tandis que dans l'approche d'ensemble, les méthodes de détection d'événements anormaux et de situation de violence proposées donnent déjà des informations exploitables sur le comportement.

Les résultats obtenus mettent en évidence plusieurs facteurs clés de la performance des ces méthodes : qualité de la représentation visuelle, capacité des méthodes d'apprentissage à produire des modèles discriminants et généraux à partir des données, conception d'algorithmes efficaces et efficients. Replacés dans le contexte d'une recherche mondiale extrêmement compétitive, nos travaux reflètent également des courants de recherche de ces dernières années et s'appuient sur des avancées notables dans le domaine de l'intelligence artificielle, comme pour certaines contributions, celles apportées par les méthodes d'apprentissage profond et les modèles de réseaux de neurones.

Travaux réalisés

La description des travaux et la synthèse des principaux résultats sont regroupées en quatre thématiques, qui constituent les chapitres du mémoire :

Détection rapide multi-classes d'objets Dans la thèse d'Hamidreza Odabai Fard (2012-2015), nous nous sommes intéressés à la détection multi-classes d'objets robuste et rapide (Chapitre 1). Bien l'état de l'art soit aujourd'hui dominé par les méthodes de réseaux de neurones profonds qui ont bénéficié de la mise à disposition de grands ensembles d'images annotées, la plupart des détecteurs existant au moment de ces travaux étaient fondés sur la classification de régions à partir de descripteurs ad-hoc encodés par un modèle spatial. Initialement conçus pour détecter une seule classe d'objets, leur complexité de calcul augmente beaucoup avec le nombre de classes. De plus, il est difficile d'ajouter de nouvelles classes de manière souple. Nous avons introduit un nouveau framework générique d'apprentissage structuré de modèles hiérarchiques optimisés sous des contraintes simultanées de classification et de classement. Un algorithme de parcours rapide de l'arbre permet de réduire les temps d'inférence. Les résultats obtenus sur des ensembles de données de référence montrent un rapport précision sur temps de calcul nettement amélioré par rapport à l'approche un-contre-tous pour 20 classes.

Dans une seconde partie de ce chapitre (Section 1.4), nous décrivons un travail ultérieur mené avec deux autres membres du laboratoire, qui a conduit à l'élaboration d'une nouvelle approche de détection d'objets rapide avec un réseau de neurones profond single-shot. Cette approche pose les fondations des travaux sur l'estimation de poses 3D multi-personnes décrits au Chapitre 4.

Ré-identification visuelle des personnes Alors que mes travaux au sein du VisionLab portaient sur l'apprentissage profond de métriques fondé sur les l'optimisation de modèles de réseaux de neurones par coût contrastif ou triplet, les recherches menées dans le cadre de la thèse de Solène Chan Lang (2014-2017) ont d'abord focalisé sur une approche de ré-identification et de vérification pour traiter les ensembles fermés et ouverts d'identités, puis ont porté sur le codage parcimonieux multi-shot des descripteurs des personnes. La ré-identification en base ouverte, plus difficile que la ré-identification en base fermée, est un problème rarement traité. Pourtant, celui-ci correspond bien mieux au problème réel de ré-identification dans les espaces où les flux de personnes ne sont pas contrôlés. Les performances des méthodes que nous avons proposées ont dépassé de manière significative les approches de l'état de l'art de cette période. Nos contributions en ré-identification de personnes sont décrites au Chapitre 2.

Analyse du comportement et détection d'événements dans les vidéos Cette thématique, abordée au Chapitre 3, concerne la modélisation du comportement humain dans les séquences vidéo, sans recourir à une extraction et une caractérisation explicites des individus. Malgré une précision moindre que les modèles microscopiques, les approches macroscopiques ont l'avantage d'être génériques et de pouvoir fonctionner dans des scènes denses où l'individualisation est difficile. Une première contribution a été apportée avec les travaux de Pedro Ribeiro, dans le cadre du projet FUI Degiv (2011-2015) dédié à la sécurisation des transports ferroviaires. Une nouvelle approche de détection d'événements violents dont les capacités ont été démontrées dans des environnements réels a été proposée. Notre méthode repose sur le principe de l'apprentissage une-classe, et permet la modélisation dynamique de structures de mouvement par des ensembles de descripteurs invariants en rotation. Notre méthode s'est révélée compétitive avec des approches entièrement supervisées de classification binaire. Cette contribution est présentée en Section 3.3).

Le projet FUI FluidTracks (2014-2017) qui visait à développer de nouvelles technologies pour améliorer le confort et la sécurité des personnes circulant dans l'espace public, nous a permis de travailler sur la caractérisation des mouvements de foule (encadrement de Hajer Fradi). Pour analyser le comportement de la foule, nous avons proposé un ensemble de descripteurs spatio-temporels de niveau intermédiaire, liés dans un graphe dynamique qui code les interactions entre les personnes. Ce modèle a été exploité avec succès dans des problèmes de classification de vidéos de foule, de détection d'anomalies et de détection d'événements violents pour lesquels l'approche a dépassé en performances les méthodes concurrentes de l'état de l'art (Section 3.4).

La caractérisation 3D de l'individu pour l'analyse de comportement Le quatrième volet de ce mémoire (Chapitre 4) traite de la caractérisation des personnes par l'estimation de leur pose 3D. La pose humaine est représentée par un graphe structuré des articulations localisées dans l'espace. La connaissance précise de la pose est une étape importante dans la compréhension fine des ses actions et activités. Dans la thèse d'Abdallah Benzine (2017-2020), nous avons abordé le problème d'estimation de poses 3D multi-personnes à partir d'une seule image RGB, en tentant de surmonter les obstacles liés i) à la variété des contextes d'utilisation, des poses des personnes, et des échelles d'observation, ii) aux occultations fréquentes, et iii) à la complexité de calcul engendrée par le traitement d'un grand nombre d'instances dans l'image. Ces travaux ont produit deux nouvelles approches d'estimation de poses 3D multi-personnes à base de réseaux de neurones profonds entraînés avec une supervision en 2D et en 3D, dont les performances surpassent l'état de l'art sur plusieurs ensembles de données de référence.

Chapitre 1

Détection rapide multi-classes d'objets

1.1 Introduction

La détection d'objets est une tâche essentielle de la vision par ordinateur pour l'analyse sémantique de scènes. Elle vise à reconnaître et localiser dans les images des instances d'objets regroupées en classes sémantiques. La détection diffère de la classification d'images par l'ajout de l'information de localisation, et de la segmentation sémantique par l'individualisation des régions correspondant aux instances. Le modèle spatial le plus simple et le plus utilisé pour définir la région contenant un objet est la boîte englobante qui est une zone rectangulaire dans l'image.

Chaque application focalise sur un ensemble particulier de classes d'objets d'intérêt : personnes et véhicules en vidéo-surveillance intelligente, feux et panneaux de signalisation pour les système d'aide à la conduite et les véhicules autonomes, pièces mécaniques pour l'inspection visuelle ou l'aide à l'assemblage dans l'industrie, véhicules militaires dans le domaine de la défense, etc.

Brique de base de nombreuses chaînes d'analyse d'images, un module de détection d'objets fournit des entrées à d'autres algorithmes : caractérisation fine des objets, suivi temporel, réidentification, comptage, analyse de comportement. C'est pour cette raison que la détection d'objets est l'un des sujets de recherches les plus traités. Malgré un nombre impressionnant de travaux et des progrès remarquables, la détection d'objets reste un problème ouvert. Les détecteurs d'objets doivent être robustes à un certain nombre de facteurs :

Variabilité intra-classe Une classe sémantique d'objets recouvre souvent une grande variété de formes, de poses et d'aspects liée aux instances qui la composent. Les conditions de prises de vue et d'environnement amplifient cette variabilité.

Similarité inter-classes Il est difficile de trouver une frontière de décision claire entre des exemples de classes différentes mais d'aspect très proche. Un exemple typique est l'ombre d'un objet qui en possède la forme mais qui doit être classé comme fond.

Occultations et troncatures Les objets en partie masqués par d'autres objets ou des éléments de la scène, tronqués en bord d'image, sont plus difficiles à détecter.

Variations d'échelle Détecter des objets de très petite taille dans les images est un challenge. Détecter les objets malgré une grande variation d'échelle implique une bonne capacité de généralisation sur l'ensemble des échelles observées et soulève par ailleurs des problèmes de temps de calcul.

En plus des défis de la reconnaissance visuelle auxquels doivent faire face les systèmes de détection d'objets, d'autres paramètres critiques sont à prendre en considération lors de leur conception :

Compromis entre performance et complexité de calcul Idéalement, les qualités recherchées sont à la fois la robustesse et la précision de la détection, et une vitesse d'exécution

élevée. Les contraintes sur la complexité de calcul sont d'autant plus fortes dans les applications mettant en œuvre des systèmes embarqués temps réel, où les ressources de calcul sont limitées pour des raisons d'énergie, d'encombrement et de coût, et dans les applications devant traiter un très gros volume d'images en un temps donné. La tendance générale est que les modèles de détection plus précis demandent davantage de ressources de calcul. D'un point de vue pratique, un compromis entre précision et complexité de calcul doit être trouvé.

Extensibilité Lorsque le nombre de classes augmente, un système de détection doit garder de performances satisfaisantes, avec un temps de calcul et une occupation mémoire acceptables en inférence et en apprentissage. Le nombre de classes peut varier de deux (une seule classe d'objets vs le fond) à des dizaines, des centaines voire des milliers. Se pose d'abord la question de la capacité d'un modèle unique de détection à généraliser correctement à l'ensemble des classes à reconnaître. Un autre problème concerne la constitution de l'ensemble des données d'apprentissage qui sont supposées représenter de manière équilibrée et relativement complète toutes les catégories traitées. Des performances élevées de détection peuvent être obtenues dans certains cas avec une spécialisation poussée des modèles et une bonne connaissance des contextes applicatifs dans lesquels ils sont utilisés. A l'inverse, les modèles très génériques, traitant un grand nombre de catégories hétéroclites et appris sur une très grande variété de contextes seront probablement moins précis. Alors que l'occupation mémoire et la complexité de calcul pour la phases d'inférence doivent être contenues pour respecter les contraintes applicatives, comme cela a été évoqué au point précédent, ces limitations existent également pour la phrase d'apprentissage. Elles sont atteintes plus ou moins vite selon la machine utilisée, et vont affecter les performances finales et le temps de mise au point des modèles. D'un point de vue industriel, une durée d'apprentissage de quelques heures à quelques jours est habituellement vu comme une durée raisonnable, au regard des durées globales de développement des technologies.

Evolutivité Cette propriété d'adaptation à de nouveaux contextes et de nouveaux concepts à reconnaître est souhaitée dans le cas de systèmes qui continuent leur apprentissage sur le long terme. Pour un détecteur d'objets, cela se traduit par la prise en compte rapide de nouvelles catégories d'objets ou un nouveau contexte d'utilisation (point de vue de caméra, environnement différents) en capitalisant au maximum les expériences précédentes.

Les travaux présentés dans ce chapitre ont cherché à apporter des réponses au problème de détection d'objets multi-classes en temps réel par l'optimisation du rapport entre la performance de détection et la complexité de calcul.

1.2 Etat de l'art et positionnement

La détection d'objets est sans doute l'une des problématiques qui ont reçu le plus d'attention en vision par ordinateur. Zou et al., 2019 ont publié un panorama synthétique des principaux travaux sur le sujet sur la période des vingt dernières années. Dans ce tour d'horizon, la littérature est segmentée en deux grandes périodes : la première, malicieusement qualifiée par les auteurs de *wisdom of cold weapon era* en raison de l'utilisation de descripteurs conçus à la main, et la période suivante de 2014 à nos jours marquée par le développement intensif et presque exclusif de modèles fondés sur les réseaux de neurones profonds. Ce développement a été grandement facilité par les progrès technologiques des calculateurs de type GPU (Figure 1.2) et la mise à disposition d'ensembles d'images annotées de plus grande taille (Figure 1.1). Dans les méthodes utilisant des réseaux de neurones, les caractéristiques visuelles de l'image et la tâche de détection sont apprises simultanément.

Dataset	train		validation		trainval		test	
	images	objects	images	objects	images	objects	images	objects
VOC-2007	2,501	6,301	2,510	6,307	5,011	12,608	4,952	14,976
VOC-2012	5,717	13,609	5,823	13,841	11,540	27,450	10,991	-
ILSVRC-2014	456,567	478,807	20,121	55,502	476,688	534,309	40,152	-
ILSVRC-2017	456,567	478,807	20,121	55,502	476,688	534,309	65,500	-
MS-COCO-2015	82,783	604,907	40,504	291,875	123,287	896,782	81,434	-
MS-COCO-2018	118,287	860,001	5,000	36,781	123,287	896,782	40,670	-
OID-2018	1,743,042	14,610,229	41,620	204,621	1,784,662	14,814,850	125,436	625,282

FIGURE 1.1 – Caractéristiques des ensembles d'images fréquemment utilisés en détection d'objets (ZOU et al., 2019)

Nous allons présenter les principales approches de détection d'objets de l'état de l'art conçues durant ces deux périodes. Toutes les méthodes citées s'appuient sur le principe de l'apprentissage supervisé.

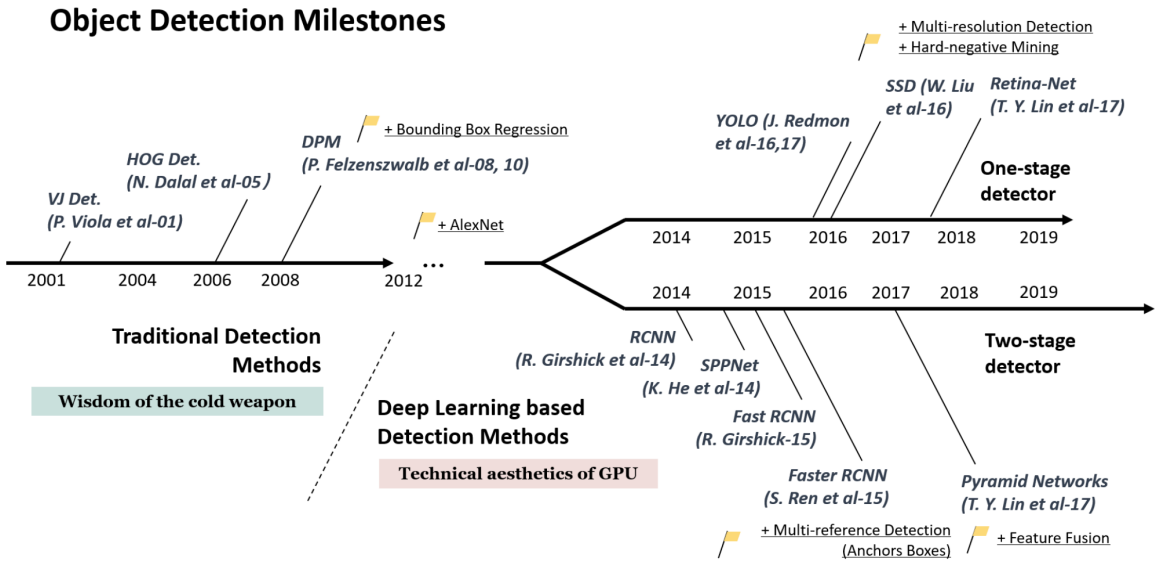


FIGURE 1.2 – Evolution des détecteurs d'objets dans les vingt dernières années. Source : ZOU et al., 2019. NB : Les méthodes les plus récentes (2020) ne sont pas représentées sur cette frise chronologique.

1.2.1 Détecteurs d'objets traditionnels

Le pipeline classique de détection, illustré par la Figure 1.3, comporte plusieurs étapes. L'image originale n'étant pas directement adaptée à la classification des régions qui la composent, elle est transformée en une représentation qui code localement des informations de couleur, de contraste, de contour, de texture, etc. Cette représentation permet de former des caractéristiques visuelles ou descripteurs dans une région donnée. Sur la base de ces descripteurs, un modèle pour chaque classe d'objets, aussi appelé filtre, est appris et appliqué sur toute l'image pour déterminer les régions contenant un objet de cette classe. Afin de détecter des objets à plusieurs échelles, l'image est présentée à plusieurs résolutions sous la forme d'une pyramide d'images. Pour un même objet présent dans l'image, il peut y avoir des réponses multiples spatialement proches. Une étape de suppression des non-maxima locaux (NMS pour *Non Maxima Suppression*) utilise les scores de détection et le recouvrement des boîtes détectées pour produire une seule détection

par objet. Le taux minimal de recouvrement entre boîtes traduit l'hypothèse que deux objets distincts ne peuvent se chevaucher au-dessus d'un certain seuil.

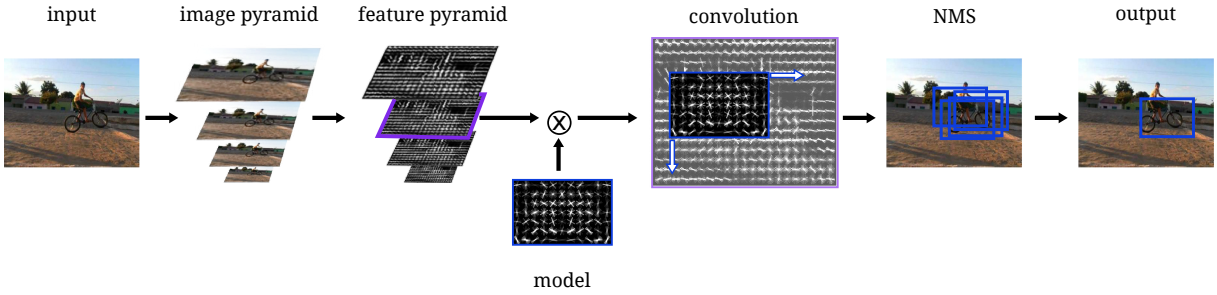


FIGURE 1.3 – Pipeline classique de détection d'objets : l'image est représentée par des caractéristiques calculées à plusieurs échelles (ici les *Histograms of Gradients* (DALAL et TRIGGS, 2005)). En appliquant le modèle de détection comme un filtre appliqué à toutes les positions de l'image, on obtient des réponses positives et négatives, les réponses positives sont regroupées par une stratégie de suppression des non maxima locaux (NMS) pour former les détections finales.

Descripteurs d'objet

Les descripteurs d'objets sont conçus pour projeter l'image dans un espace où la détection est facilitée. Les chercheurs ont rivalisé d'ingéniosité pour concevoir des descripteurs sophistiqués qui soient à la fois rapides à calculer et discriminants. Ainsi, les descripteurs de Haar (VIOLA et JONES, 2004 ; VIOLA et JONES, 2001), les Local Binary Pattern (LBP) (HADID et al., 2004), les Histogram of Gradients (HOG) (DALAL et TRIGGS, 2005) et leur version modifiée ultérieurement par FELZENSZWALB et al., 2008b ont connu un grand succès pour la détection de visages, de personnes et d'autres catégories d'objets.

Modélisation des parties

Pour encoder avec meilleure localité les différentes parties qui composent un objet, les régions sont parfois divisées en cellules (DALAL et TRIGGS, 2005). A la différence des méthodes décrivant les objets dans leur ensemble (holistiques), les approches de modélisation des parties apportent plus de souplesse. Dans une stratégie *bottom-up*, les parties sont détectées et rassemblées pour former des candidats d'objet. Dans le modèle *Implicit Shape Model* (ISM)(LEIBE et al., 2008 ; LEIBE et SCHIELE, 2003), les parties sont associées à des mots visuels et votent pour un centre d'objet. Ces travaux ont inspiré des méthodes ultérieures utilisant un système de vote des parties d'objets via une transformée de Hough (GALL et LEMPITSKY, 2009 ; RAZAVI et al., 2012).

Deformable Part Model

Le *Deformable Part Model* (DPM) a constitué l'état de l'art en détection d'objets au début des années 2010 (FELZENSZWALB et al., 2010a ; FELZENSZWALB et al., 2008b). Le modèle s'inspire du concept des *Pictorial Structures* où un objet est défini par un ensemble de parties dont certaines sont reliées entre elles. Le DPM est constitué d'un modèle de représentation globale et des modèles de parties dont la position relative est apprise de manière latente. La Figure 1.4 illustre les trois modèles de représentation du DPM pour la classe "personne" : un modèle capture l'apparence globale de l'objet à basse résolution (filtre racine), un autre les différentes parties à plus haute résolution (filtres des parties) et un troisième modélise la déformation de l'objet avec un coût quadratique dépendant de la distance des parties à la racine. Les caractéristiques visuelles employées sont une version plus robuste des descripteurs HOG. Une des forces du DPM

est de modéliser plus finement les objets articulés en permettant aux parties de se déplacer dans un certain intervalle de distances. D'autre part, un nombre fixe de composantes sont générées pour encoder différentes vues des objets. Ces composantes sont trouvées par clustering sur les exemples d'apprentissage. Le modèle racine est d'abord optimisé avec un algorithme de SVM, les parties sont ensuite ajoutées dans les zones de forte concentration des poids du SVM. Les modèles d'apparence des parties et de déformation sont optimisés de manière alternée.

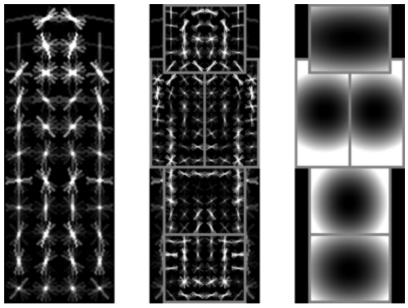


FIGURE 1.4 – Deformable Part Model (FELZENSZWALB et al., 2010a) illustré sur sur la classe "personne" : un objet est représenté par un modèle racine global de basse résolution (a), un modèle de plus haute résolution pour chacune des parties (b) et un modèle de déformation (c) encodant les positions des parties relativement à la racine.

Parcours de l'image par une fenêtre glissante

La technique de la fenêtre glissante est la stratégie de parcours d'image la plus répandue. Elle implémente un parcours exhaustif de l'image. Une fenêtre d'analyse rectangulaire matérialise une région d'intérêt dans laquelle la classification objet/non objet est effectuée. Cette fenêtre est déplacée à toutes les positions d'une grille placée sur l'image et à plusieurs échelles. Les performances de la détection dépendent d'hyper-paramètres qui règlent la finesse de discrétisation de la grille, le nombre d'échelles, le choix des rapports de forme des fenêtres. Ces paramètres doivent être choisis judicieusement afin d'obtenir un compromis satisfaisant entre le temps d'exécution et la précision de la détection.

Méthodes d'accélération

L'analyse exhaustive des régions de l'image pouvant être relativement lourde, différentes stratégies pour réduire le temps de calcul ont été proposées.

L'astuce de l'image intégrale introduite par VIOLA et JONES, 2001 permet de pré-calculer la somme des caractéristiques sur toute l'image et d'évaluer les descripteurs dans une région rectangulaire de manière très rapide : les valeurs à la position des coins du rectangle sont simplement lues dans l'image intégrale et sommées.

Outre l'utilisation des images intégrales pour accélérer le calcul des caractéristiques, les mêmes auteurs introduisent une stratégie hiérarchique de classification qui prend la forme d'une cascade de classifieurs. Le principe de la cascade est d'utiliser des classifieurs de plus en plus précis, le but étant de rejeter le plus rapidement possible, c'est-à-dire dans les premières étapes de la cascade un grand nombre de régions correspondant au fond, et de ne passer l'essentiel du temps que sur les hypothèses pertinentes de régions susceptibles de contenir des objets, comme illustré Figure 1.5.

Le modèle hybride entre les pyramides d'images et les pyramides de classifieurs (DOLLAR et al., 2010 ; DOLLÁR et al., 2014) allie précision et rapidité.

Une tout autre technique de parcours de l'image est adoptée dans (Lampert, 2010 ; Lampert et al., 2009a) : un algorithme de séparation et évaluation (*branch and bound*) recherche le chemin le plus court vers la région de score maximal en réduisant progressivement l'ensemble des régions à évaluer.

L'algorithme *Selective Search* (Uijlings et al., 2013) filtre les régions de l'image par une étape de segmentation (Felzenszwalb et Huttenlocher, 2004) qui fournit une liste restreinte de propositions de régions retenues pour la classification. Cette approche inspirera la méthode R-CNN.

Détecteur de référence en son temps, le DPM a inspiré de nombreux travaux visant à améliorer différents aspects de la méthode, dont les temps de calcul. Ainsi, l'idée de la cascade de détection a été reprise (Felzenszwalb et al., 2010b). Dans (Pedersoli et al., 2011), les parties sont recherchées dans une zone limitée pour gagner en temps, alors que Yan et al., 2014 travaillent plutôt sur divers moyens d'optimiser le calcul des caractéristiques et proposent une stratégie d'élagage pour rejeter des candidats négatifs plus en amont de la cascade. Sadeghi et Forsyth, 2014 utilisent une quantification des vecteurs de caractéristiques pour atteindre le temps réel. Les implémentations optimisées sur GPU (Gadeski et al., 2014) permettent également d'augmenter la vitesse d'exécution du détecteur.

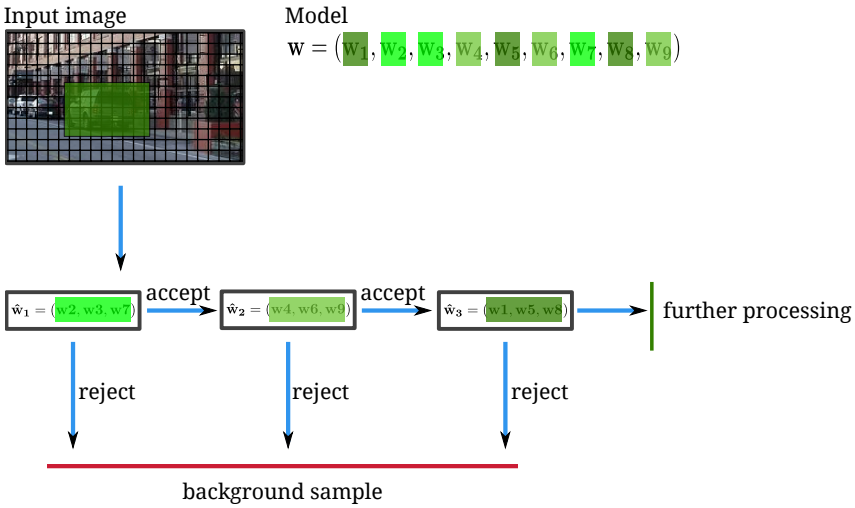


Figure 1.5 – Principe de la cascade de classifieurs. La classification d'une région est effectuée par un ensemble de classifieurs faibles répartis sur les différents étages de la cascade. A chaque niveau de la cascade, les classifieurs faibles de ce niveau sont appliqués pour rejeter la région en tant que fond, ou l'accepter et continuer la classification dans les niveaux suivants.

Méthodes d'apprentissage et d'inférence

Nous avons vu que la détection d'objets peut être comparée à un problème de classification des régions de l'image. Dans ce cadre, l'apprentissage supervisé des classifieurs a été explorée avec différentes familles de méthodes.

Les algorithmes de *boosting* (Freund et Schapire, 1997) reposent sur le principe d'agrégation de classifieurs faibles, qui doivent être meilleurs qu'un classifieur aléatoire, pour construire un classifieur fort. Les classifieurs faibles peuvent être des classifieurs binaires associé à une composante du vecteur de description. Le boosting a rencontré un grand succès depuis les travaux de Viola et Jones, 2001 pour la détection d'une classe d'objets. Dans une version multi-classes proposée par Torralba et al., 2004 et Torralba et al., 2007, une partie des classifieurs faibles est partagée entre plusieurs classes.

Les méthodes de séparateurs à vaste marge ou *Support Vector Machine* (SVM) ont été aussi beaucoup utilisées en apprentissage supervisé pour la détection d'objets (Dalal et Triggs, 2005 ; Felzenszwalb et al., 2010a). Les SVM sont une manière élégante et mathématiquement bien fondée de séparer linéairement les données par des hyperplans en maximisant la marge (la distance) entre les hyperplans. Dans le cas plus général d'une séparation non-linéaire, on utilise des noyaux pour représenter les données dans un espace de dimension supérieure où la séparation linéaire peut s'appliquer.

L'enchaînement d'étapes de classification linéaire à travers une structure d'arbre permet d'obtenir des frontières de décision non linéaires. Les arbres ont été largement exploités avec les algorithmes de boosting et de SVM.

Les forêts aléatoires ou *Random Forests* font intervenir des arbres de décision, chaque nœud représentant en général une fonction de décision simple. Chaque arbre est parcouru et la classification finale résulte de la combinaison de la sortie de tous les arbres. Dans les approches de Gall et Lempitsky, 2009 ; Razavi et al., 2012 une forêt aléatoire établit les correspondances entre les patches des parties et les votes pour le centre de l'objet.

L'usage des réseaux de neurones est devenu presque systématique dans les algorithmes de vision par ordinateur ces dernières années. Les méthodes de détection d'objets par réseaux de neurones profonds seront présentées dans la Section 1.2.2.

Sélection des exemples

Le choix des exemples d'apprentissage, notamment celui des exemples négatifs, influence beaucoup les performances des détecteurs. La sélection d'exemples négatifs dits difficiles favorisent la détermination des frontières de décision plus précises entre les classes. Beaucoup de méthodes emploient la technique du *bootstrapping* qui consistent à prendre des sous-ensembles d'exemples négatifs en ajoutant progressivement des exemples plus difficiles, c'est-à-dire ceux qui n'ont pas été correctement classifiés dans les étapes précédentes (Dalal et Triggs, 2005 ; Felzenszwalb et al., 2008b).

Détection d'objets multi-classes

La plupart des méthodes citées précédemment ne traite nativement qu'une classe d'objets. La généralisation à plusieurs classes rend le problème d'optimisation du rapport performance/complexité de calcul encore plus difficile. En effet le temps de calcul peut exploser avec le nombre de classes. Le partage d'informations entre les classes a été exploré sous différentes formes, d'une part pour économiser du temps de calcul, et d'autre part pour rendre les méthodes plus robustes.

Partage de caractéristiques et structures hiérarchiques Torralba et al., 2007 présentent une approche de boosting nommée *JointBoost* où les classifieurs faibles sont partagés autant que possible pour réduire le nombre total de classifieurs et atteindre une complexité logarithmique. Les auteurs montrent que le partage de classifieurs a aussi un effet sur les performances de reconnaissance grâce à une meilleure généralisation des classifieurs partagés. Zhang et al., 2013a ajoutent à ce framework un mécanisme de partition en sous-catégories quand la variance des exemples est grande.

D'autres approches proposent le partage des modèles d'apparence des parties comme dans (Ott et Everingham, 2011) sur la base du DPM, ou (Razavi et al., 2011) pour l'ISM. Les méthodes compositionnelles utilisent un dictionnaire commun de prototypes élémentaires pour l'ensemble des classes et reconstruisent les objets à partir de ces prototypes (Zhu et al., 2010a).

Le regroupement de classes visuellement proches dans un modèle hiérarchique permet un gain en précision tout en réduisant la complexité de calcul. Cette idée est à l'origine des

travaux en classification d'images sur un très grand nombre de classes où la hiérarchie et les classifieurs sont appris (Bengio et al., 2010 ; Deng et al., 2011). En détection d'objet, Salakhutdinov et al., 2011 prévoient aussi un apprentissage de la structure de l'arbre en plus de celui des filtres, mais se limitent à deux niveaux de profondeur. Cette approche très intéressante de double optimisation de la structure hiérarchique et des filtres a beaucoup inspiré nos travaux.

Dans un cadre de transfert d'apprentissage, Salakhutdinov et al., 2012 reprennent l'idée de la hiérarchie de classes et définissent des super-catégories auxquelles peuvent se rattacher les nouvelles classes qui sont présentées avec un seul exemple. Ces super-catégories sont supposées contenir des informations utiles apprises avec les classes précédentes pour généraliser sur la nouvelle classe.

Le principe du partage de caractéristiques est poussé à un stade supérieur dans les approches de réseaux de neurones (Section 1.2.2).

Stratégies de classification multi-classes L'approche naïve de détection multi-classes consiste à créer autant de classifieurs binaires que de catégories d'objets : c'est l'approche *One-versus-All* (OvA) dont la complexité est linéaire avec le nombre de classes. Chaque classifieur binaire, entraîné séparément, est responsable de discriminer une catégorie d'objets (classe des exemples positifs) par rapport aux autres catégories et au fond (classe des exemples négatifs). Pour une région donnée, la décision est donnée par le classifieur qui donne le meilleur score, une normalisation étant faite au préalable pour pouvoir comparer les scores, par exemple avec une fonction sigmoïde calibrée sur un ensemble de validation dans le cas des SVM (Platt, 1999).

La stratégie *One-versus-One* prend un parti différent (Kressel, 1999). Pour k classes, il existe $k(k-1)/2$ paires de classes différentes pour lesquelles on entraîne un classifieur binaire. Tous les classifieurs binaires sont appris séparément. Un avantage de cette stratégie est de réduire le nombre d'exemples nécessaires pour l'apprentissage d'un classifieur binaire, l'inconvénient majeur est la complexité de calcul. Dans (Platt et al., 2000), les classifieurs sont organisés selon une structure hiérarchique pour réduire la complexité à $\mathcal{O}(k)$.

Toujours avec le formalisme des SVM, Crammer et al., 2001 transforment le problème de classification multi-classes en un problème de classement (rangement) des scores par ordre décroissant, la catégorie la plus probable devant avoir le score le plus élevé.

Techniques d'accélération En plus du partage de caractéristiques, les structures hiérarchiques accélèrent la classification. Le principe de la cascade est étendue au cas de la classification multi-classes : les classes sont regroupées dans les nœuds d'un arbre, plus on avance en profondeur dans la cascade, plus l'ensemble des classes possibles se réduit jusqu'à trouver une seule classe. Dans la méthode de Fidler et al., 2010, le parcours de l'arbre est déterminé par un ensemble de décisions binaires, la classification finale étant donné par la feuille terminale. L'un des problèmes de cette type de parcours est qu'une erreur de décision à un nœud donné ne peut être rattrapée. Un autre inconvénient est que la décision finale ne dépend que du score obtenu à la feuille terminale et non des scores obtenus sur l'ensemble des nœuds parcours.

Dean et al., 2013 s'attaquent au problème de détection avec un très grand nombre de classes (100000). Ils proposent une technique de hachage pour transformer les caractéristiques en un code, et l'évaluation du score par la convolution des filtres est remplacé par un calcul de similarité entre codes.

1.2.2 Réseaux de neurones profonds pour la détection d'objets

Suivant la dynamique des recherches sur la classification d'image avec des réseaux de neurones, le détecteurs d'objets fondés sur l'apprentissage profond ont eux-aussi connu des progrès

importants et rapides depuis 2014. La conception des détecteurs d'objets est plus complexe que celle des modèles de classification d'image car il faut traiter un nombre variable d'objets et les localiser dans l'image.

La plupart des modèles de détection s'appuient sur des architectures de réseau de neurones profonds convolutifs (CNN pour *Convolutional Neural Networks*) dont le réseau-cœur ou *backbone* bénéficie souvent d'un pré-apprentissage sur une tâche de classification d'images sur la base ImageNet.

Contrairement aux détecteurs traditionnels, la plupart des approches de réseaux de neurones permettent l'apprentissage simultané de la représentation de l'image et de la tâche de détection. D'autre part, ces méthodes sont bien adaptées au passage à un grand nombre de classes. Les caractéristiques sont partagées dans une grande partie du réseau, la classification n'étant réalisée qu'à partir des caractéristiques de plus haut niveau sémantique, produites à la fin du réseau.

En contrepartie, les réseaux de neurones profonds demandent des ressources de calcul beaucoup plus importantes. Les opérations pouvant être largement parallélisées, les architectures massivement parallèles des GPU brillent par leur efficacité sur ce type de calcul.

Les détecteurs fondés sur les réseaux de neurones sont souvent classés en deux grandes catégories définies par la stratégie de détection adoptée : les détecteurs en deux étapes, désignés par le terme *two-stage* et les détecteurs en une seule étape, qualifiés de *single-shot*.

Détection d'objets en deux étapes

Le principe de ces approches est de sélectionner dans une première étape un ensemble de régions d'intérêt candidates puis de les classifier séparément en objet/non objet et de trouver les boîtes englobantes précises dans une seconde étape.

R-CNN, SPP-Net, Fast-RCNN, Faster-RCNN Girshick et al., 2014 présentent R-CNN, une des premières méthodes de référence de détection d'objets à deux étapes : les propositions de régions sont générées par l'algorithme *Selective Search* (Felzenszwalb et Huttenlocher, 2004). Chaque région est redimensionnée à une taille fixée et passe dans un CNN qui extrait les caractéristiques pour la classification et la régression des boîtes englobantes. Les régions classées positives sont ensuite fusionnées dans une étape de *Non-maximum suppression* (NMS). Pour obtenir une précision suffisante, il faut générer un très grand nombre de propositions de régions, ce qui occasionne des temps d'apprentissage et d'inférence très longs.

SPP-Net (He et al., 2014) relâche la contrainte de taille fixe de l'entrée grâce à un mécanisme de pyramide spatiale qui sous-échantillonne les cartes de caractéristiques à plusieurs échelles avec un facteur variable de façon à produire un vecteur en sortie de taille fixe, quelle que soit la taille de la région d'entrée.

La méthode R-CNN est améliorée dans Fast-RCNN (Girshick, 2015) où l'étape de sélection des régions d'intérêt est réalisée sur des cartes de caractéristiques issues d'un CNN. L'extraction des caractéristiques et le redimensionnement à une taille fixe sont prises en charge par l'opération de *ROI Pooling*. Dans la seconde partie du réseau, les caractéristiques sont transformées et en fin de réseau, deux branches assurent respectivement la classification et la régression des boîtes.

Faster-RCNN (Ren et al., 2015) améliore davantage la vitesse et la précision de la détection : l'algorithme *Selective Search* est abandonné au profit d'un CNN entraîné pour déterminer des régions d'intérêt (appelé *Region Proposal Network*) et produire des cartes de caractéristiques qui sont traitées de la même manière que dans Fast-RCNN pour la détection. L'ensemble du réseau est entraîné de bout-en-bout. Faster-RCNN est 250x plus rapide que R-CNN et 10x plus rapide que Fast-RCNN.

Cette approche introduit le système d'ancres qui va inspirer beaucoup d'autres méthodes de détection, dont les méthodes à une passe (*single-shot*). Les ancres sont des boîtes prédéfinies, de taille et de rapport de forme variables, placées sur toute l'image selon une grille régulière. Le choix des ancres est déterminant pour la performance de détection car elles doivent être adaptées aux échelles et forme des objets à détecter : ce choix est fait a priori (Ren et al., 2015), ou en faisant une analyse statistique des boîtes de vérité terrain dans la base d'apprentissage pour déterminer des clusters (Redmon et Farhadi, 2017). Faster-RCNN a été étendu ultérieurement à la détection de points d'intérêt et à la segmentation d'instances sous le nom de Mask-RCNN (He et al., 2017).

Features Pyramid Networks Une des difficultés de la détection d'objets est d'arriver à détecter correctement à différentes échelles. Le champ réceptif des caractéristiques de sortie dépendant de l'architecture, l'idée de Lin et al., 2017b est de construire une pyramide de caractéristiques dont les différents étages encodent des informations à plusieurs échelles et niveaux sémantiques. L'information circule dans le sens montant (de la plus grande résolution à la plus petite) qui est le sens de passage habituel d'un CNN, et dans le sens inverse, ce qui permet d'agréger de l'information de plus haut niveau sémantique dans des niveaux de résolution supérieure grâce aux connexions latérales (Figure 1.6). Dans l'article original, le FPN est directement intégré dans le détecteur Faster-RCNN au niveau du RPN.

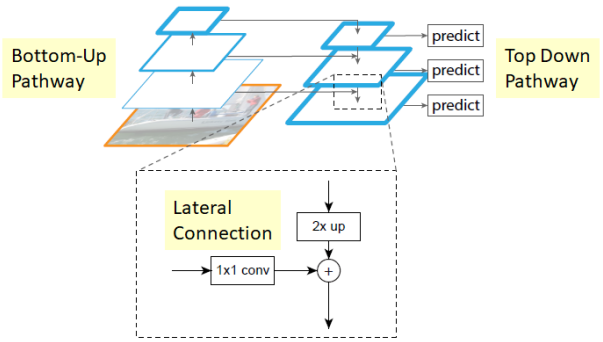


Figure 1.6 – Principe des Feature Pyramid Networks (Lin et al., 2017b)

Détection d'objet en une seule étape

Malgré les améliorations apportées par Faster-RCNN en terme de vitesse d'exécution, les modèles de détection précédents restent trop lents par construction, chaque proposition de région devant passer dans la seconde partie du réseau.

Dans la perspective des applications en temps réel, des détecteurs rapides en une seule étape ou *single-shot*, c'est-à-dire réalisant un seul passage avant du réseau de neurones, ont été créés pour pallier cet inconvénient. Pour les premiers détecteurs *single-shot*, la rapidité est parfois gagnée au détriment de la précision de détection, mais des méthodes de plus en plus performantes ont vu le jour au fil du temps. Dans ce qui suit, chaque méthode présentée obtient des performances supérieures à celles de son prédécesseur (évaluation sur PASCAL VOC et COCO).

Détecteurs utilisant un système d'ancres YOLO (YOu Look only Once)(Redmon et al., 2016) est l'une des premières méthodes de détection *single-shot*. Une grille régulière divise l'image en cellules dans lesquelles la prédiction des objets est effectuée. Les objets ne sont détectés que dans les cellules qui contiennent leur centre. La limitation majeure de la méthode est qu'une cellule ne peut contenir plus d'un objet à la fois. Pour chaque cellule, la méthode prédit un nombre fixe de boîtes avec un score de confiance (IoU avec

la vérité terrain) ainsi que le vecteur des probabilités de toutes les classes. Très rapide, YOLO est plus précis que DPM et R-CNN.

Liu et al., 2016b développent Single-Shot Multibox Detector (SSD), un autre détecteur rapide en une étape très populaire. SSD implémente une grille d'ancres prédéfinies. La prédiction des boîtes (offsets du centre et des dimensions), des scores de classification pour toutes les classes et du score de présence d'objet, est faite de manière dense sur toutes les ancres. Les ancres sont considérés positives si l'IoU avec la vérité terrain est supérieure à 0.5, négative dans le cas contraire. L'architecture de SSD comporte une *backbone* VGG-16 suivi de couches de convolutions de taille de plus en plus petite pour capturer les caractéristiques à différentes échelles et profondeurs du réseau. Les sorties de ces couches sont concaténées dans un vecteur unique de grande dimension qui regroupe toutes les prédictions. Le modèle est entraîné en minimisant le coût de localisation des boîtes (exprimé avec une distance smoothL1) sur toutes les ancres positives et l'entropie croisée sur toutes les classes. Une stratégie d'échantillonnage des ancres sert à équilibrer le nombre d'ancres positives et négatives avec un ratio de 1 3, et permet une meilleure convergence du modèle.

YOLO V2 (Redmon et Farhadi, 2017) apporte plusieurs améliorations à l'algorithme original, dont l'ajout de la Batch Normalization, la gestion de tailles d'image plus grandes et variables, un système d'ancres adaptées et trouvées par clustering des boîtes de vérité terrain sur l'ensemble d'apprentissage avec l'algorithme des K-moyennes. L'architecture, également nouvelle (Darknet-19), est conçue pour réduire la complexité de calcul par rapport à VGG-16.

Le problème du déséquilibre entre les classes rencontré lors de l'apprentissage des modèles *single-shot* est traité dans la méthode RetinaNet (Lin et al., 2017c) différemment de SSD. Le fond souvent prépondérant par rapport aux objets comporte un grand nombre d'exemples faciles à classifier. Dans RetinaNet, une nouvelle fonction de perte, la *Focal Loss*, modifie la fonction d'entropie croisée en ajoutant un terme dont le rôle est d'adapter le coût des exemples difficiles (mal classifiés) par rapport aux exemples bien classifiés. Ainsi, l'influence des classes bien représentées est réduit par rapport aux classes peu représentées. RetinaNet reprend l'architecture des FPN avec une base ResNet à laquelle sont ajoutées des têtes de classification et de régression à chaque étage de la pyramide. RetinaNet bat Faster-RCNN en terme de précision à *backbone* équivalente.

YOLO v3 (Redmon et Farhadi, 2018) innove par rapport à la version précédente par une nouvelle architecture beaucoup plus profonde, le Darknet-53 dont les performances sont similaires à ResNet-152 en étant plus rapide, et une gestion multi-échelles : les caractéristiques sont calculées à trois niveaux de résolution/profondeur du réseau et combinées, le modèle effectue les prédiction à chacun de ces trois niveaux. Plus précis que la version 2, YOLO v3 est cependant plus lent à cause de son architecture plus lourde.

Récemment, l'approche EfficientDet (Tan et al., 2020), inspirée des recherches sur la recherche automatique d'architectures, se distingue par une famille d'architectures dont la taille est paramétrée par un seul facteur appelé *compound scaling*. La fusion des caractéristiques à plusieurs échelles et niveaux sémantiques est gérée par une nouvelle structure, le BiFPN, plus performante que le FPN : reprenant la structure bi-directionnelle de PANet (Liu et al., 2018), le BiFPN offre un meilleur compromis précision/complexité de calcul. EfficientDet fait partie des meilleurs détecteurs d'objets rapides actuels.

Détecteurs n'utilisant pas de système d'ancres D'autres approches s'affranchissent du système d'ancres prédéfinies qui imposent des contraintes sur le choix du nombre d'ancres, de leur taille et forme, et des critères d'association des ancres au boîtes de vérité terrain délicats à mettre en place.

Avec la méthode CornerNet, Law et Deng, 2019 prennent le parti de prédire les boîtes

englobantes à l'aide de deux points-clés, le point haut-gauche et le point bas-droit. La prédiction s'effectue de manière dense à l'aide de cartes de chaleur qui encodent la position de ces points-clés et des vecteurs d'*embedding* permettant d'associer les paires de points d'une même boîte, inspirée de la méthode de Newell et al., 2017. Les performances de CornerNet sont comparées à celles des détecteurs *two-stage* et des détecteurs *single-shot* jusqu'à RetinaNet.

CenterNet (Zhou et al., 2019) étend la modélisation de CornerNet en ajoutant un point-clé supplémentaire qui est le centre de la boîte englobante. En plus des cartes de chaleurs représentant la probabilité de présence des coins, une carte de chaleur pour les centres des boîtes est prédite. Les auteurs montrent que CenterNet est plus précis que CornerNet.

Dans l'approche FCOS (Fully Convolutional One-Stage Object Detection) (Tian et al., 2019), chaque point des cartes de sortie du CNN est associé à un pixel de l'image. Une position est considérée positive si le pixel tombe dans une boîte englobante de vérité terrain, dans ce cas les coordonnées de la boîte sont prédites, ainsi qu'un score de *centerness* représentant la proximité du point au centre de la boîte. Le modèle est donc optimisé sur l'ensemble des pixels appartenant aux boîtes de vérité terrain. L'architecture de FCOS repose sur un FPN qui, outre sa capacité à produire des caractéristiques à différentes échelles et niveaux sémantiques, se montre utile pour gérer les zones de recouvrement entre les objets : dans ces zones où les pixels appartiennent à plusieurs objets, les différents niveaux du FPN peuvent prédire des boîtes différentes.

Une approche très originale présentée récemment, DETR (DEtection TRansformer) (Carion et al., 2020), se débarrasse complètement des hyper-paramètres liées aux ancres, aux critères d'association avec les vérités-terrain, et à l'étape de NMS. Les auteurs proposent l'utilisation d'un modèle encodeur-décodeur de type *transformer* pour prédire directement des ensembles de détections. Le modèle est entraîné de bout-en-bout avec une fonction minimisant le coût d'association un-à-un entre l'ensemble des classes et boîtes prédites aux classes et boîtes de vérité terrain ou une absence d'objet. Un CNN fournit une représentation 2D de l'image, qui est aplatie pour nourrir l'encodeur, les positions étant encodées pour retrouver les informations de localisation. L'encodeur transforme les caractéristiques avec un mécanisme d'attention. Le décodeur sort en parallèle N prédictions à partir de N requêtes d'objet et des caractéristiques issues de l'encodeur. Chaque élément de la séquence de sortie passe dans un FFN (Feed Forward Network) pour produire une prédiction de détection. Les performances de DETR sont supérieures à celles de Faster-RCNN sur l'ensemble d'images COCO, mais pour l'instant inférieures aux meilleures méthodes utilisant des ancres.

1.2.3 Positionnement de nos travaux

Nous présentons dans ce chapitre deux contributions principales.

La première est issue des travaux de thèse de Hamidreza Odabai Fard (2012-2015) (Odabai Fard et al., 2014b,c ; Odabai Fard, 2015). Bien l'état de l'art en détection d'objets soit aujourd'hui dominé par les méthodes de réseaux de neurones profonds qui ont bénéficié de la mise à disposition de grands ensembles d'images annotées et des technologies de calcul parallèle, la plupart des détecteurs existant au moment de nos travaux étaient fondés sur la classification de régions à partir de descripteurs conçus à la main encodés par un modèle spatial. Nous nous intéressons plus particulièrement au problème de la détection multi-classes robuste et rapide. Dans le but d'allier performance et rapidité, nous proposons un nouveau framework générique qui permet l'apprentissage d'une hiérarchie de classes sous forme d'arbre dans laquelle les caractéristiques sont partagées entre classes pour créer des filtres discriminants. Le modèle est appris grâce à une optimisation combinant des contraintes simultanées de classification et de tri, formulée avec

des SVM. Inspirée par Salakhutdinov et al., 2011, notre approche apprend non seulement les filtres pour une hiérarchie de classes donnée, mais également une structure adaptée de l'arbre en fonction des classes, en prenant en compte l'équilibre de l'arbre et les similarités entre classes. La manière de construire le modèle hiérarchique rend possible le transfert d'apprentissage vers de nouvelles classes. Enfin, nous proposons une stratégie de parcours optimal de l'arbre qui limite le nombre d'évaluations par les classifieurs tout en considérant les scores sur l'ensemble du parcours. Nous cherchons aussi à concevoir une méthode permettant de régler facilement le compromis entre la performance de détection et la vitesse d'inférence. Le framework est implémenté avec les descripteurs HOG utilisés dans l'approche DPM. (Felzenszwalb et al., 2010a). Ces travaux sont décrits en Section 1.3.

La seconde contribution résulte d'un projet de recherche réalisé en 2018-2019 avec deux autres chercheurs du laboratoire, Florian Chabot et Mohamed Chaouch (Chabot et al., 2020). Dans ce travail, nous avons cherché à proposer un nouveau modèle de réseau de neurones pour la détection *single-shot* d'objets, qui offre un meilleur compromis entre la performance et la rapidité que les méthodes de l'état de l'art, plus spécifiquement dans la catégorie des détecteurs pouvant tourner en temps réel (vitesse $>$ 10 fps) sur un GPU grand public. Notre détecteur, appelé LapNet, utilise un système d'ancres et intègre plusieurs innovations dans la stratégie d'apprentissage qui sont à l'origine de ses bonnes performances. Plusieurs problèmes rencontrés souvent lors de l'entraînement des modèles de détection utilisant des ancres sont abordés : la qualité de l'association entre les ancres et les objets, le déséquilibre entre classes et le déséquilibre entre tailles d'objets. Les travaux relatifs à la méthode LapNet sont présentés en Section 1.4. Le modèle LapNet a servi de base à la conception de l'algorithme PandaNet d'estimation de poses 3D humaines (Benzine et al., 2020a) qui sera présenté au Chapitre 4.

1.3 Détection efficace d'objets avec une hiérarchie de classes

1.3.1 Description générale de l'approche MCRT

L'approche proposée, appelée *Multi-class Classification and Ranking Tree* (MCRT), a pour objectif d'améliorer la performance de la détection multi-classes par rapport à la stratégie *One-versus-All* (OvA). Cette approche est générique et peut s'appliquer avec n'importe quel choix de descripteurs. Les classes sont regroupées dans une structure d'arbre où les nœuds sont composés de classifieurs traitant des ensembles de classes de plus en plus petits au fur et à mesure que l'on avance dans l'arbre, jusqu'aux feuilles qui correspondent chacune à une classe d'objets. Par construction, le parcours de l'arbre permet d'obtenir une frontière finale de décision non linéaire en appliquant séquentiellement un certain nombre de classifieurs. Pour éviter de créer un nœud supplémentaire pour le fond, les exemples de fond peuvent être rejetés à n'importe quel étape du parcours de l'arbre : les scores sont triés par ordre décroissant et si le score maximal est négatif, alors l'exemple est considéré comme du fond et rejeté. Nous détaillerons ultérieurement comment le modèle est optimisé sous des contraintes de classification et de tri (classement) pour produire des scores triés à chaque nœud. Le parcours est effectué avec l'algorithme A* qui trouve efficacement avec une heuristique le chemin donnant le plus grand score, et donc la bonne classe.

La Figure 1.7 illustre un parcours de l'arbre dans MCRT.

Une région de l'image à la position x doit être catégorisée parmi un ensemble de k classes positives dont les labels sont notés $\mathcal{Y}^+ \equiv \{y_1, \ldots, y_k\}$ et une classe négative (fond) de label $y_{bg} = -1$. On cherche à produire des scores positifs pour chacune des k classes tels que la bonne classe doit avoir le meilleur rang dans la liste triée de ces scores, ou un score négatif si la région est considérée comme du fond :

$$\text{score}(x) = \max_{y \in \{1,\ldots,k\}} \text{score}_y(x) \tag{1.1}$$

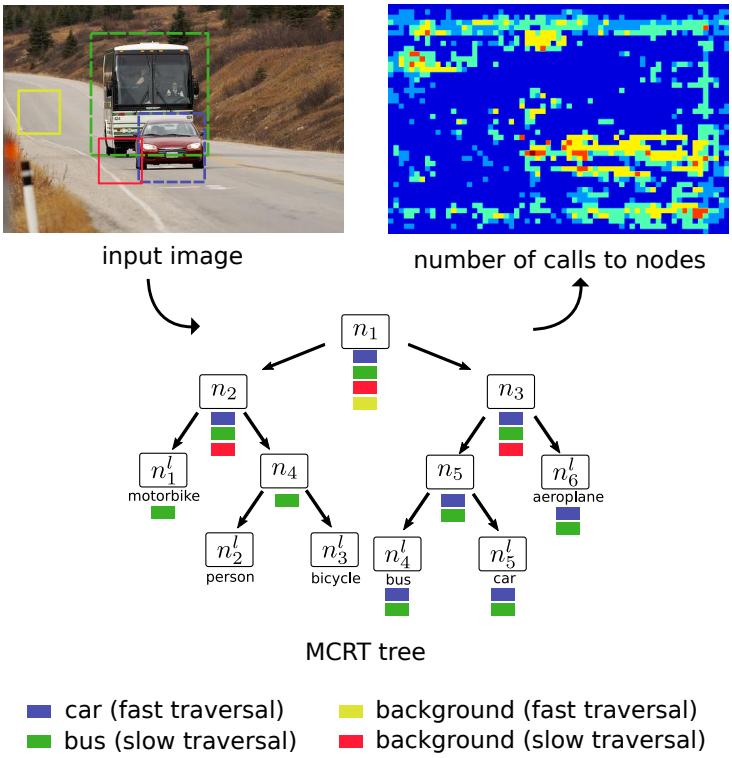


FIGURE 1.7 – Exemple de parcours dans l'arbre dans l'approche MCRT. Les catégories considérées ici sont 'motorbike', 'person', 'bicycle', 'bus', 'car','aeroplane'. Les quatre régions choisies dans l'image (en bleu, en vert en rouge, et en jaune) suivent des chemins différents dans l'arbre. Le nœud initial n_1 contient toutes les classes, la région jaune est rejetée comme fond dès n_1. La région rouge est rejetée dans les nœuds suivants n_2 et n_3, la région verte est correctement classée en 'bus' en arrivant en n_4^l après l'évaluation par 9 filtres, alors que la région bleue est classée en 'car' (n_5^l) après avoir été évaluées par 7 filtres. La carte de chaleur donne un aperçu du nombre de filtres évalués en chaque position dans l'image.

La classe prédite $\hat{y}$ est donnée par la fonction de décision finale h :

$$\hat{y} = h(x) = \begin{cases} -1 & , \text{si score}(x) \leqslant 0 \\ \arg \text{score}(x) & , \text{sinon} \end{cases} \tag{1.2}$$

Le modèle de détection multi-classes de MCRT est un arbre T qui contient |T| nœuds. Chaque nœud correspond à un filtre $w_i \in \{w_1, w_2, \ldots, w_{|T|}\}$ représentant un ensemble de classes. Le nœud racine n_r représente toutes les classes. Les feuilles n_y^l constituent les derniers filtres et donnent la prédiction finale de la classe de label y. On note n_i, $i \in \{1, \ldots, |T|\}$ un nœud intermédiaire, anc(n_i) l'ensemble des ancêtres du nœud n_i (n_i inclus) et desc(n_i) l'ensemble des descendants du nœud n_i (n_i exclu).

L'arbre T est un modèle *coarse-to-fine*. Lorsqu'une région de l'image à la position x est classifiée, elle traverse l'arbre selon un chemin qui dépend du score donné par chaque filtre, et qui doit aboutir à la feuille de la bonne classe où à un rejet précoce s'il s'agit du fond. La Figure 1.8 illustre, pour une hiérarchie de classes donnée, l'application des filtres à chaque nœud de l'arbre.

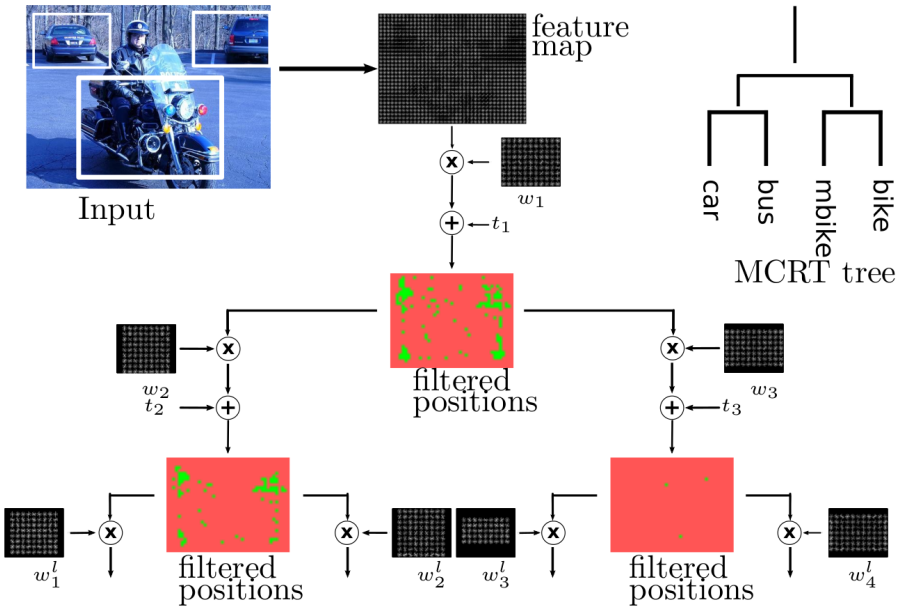


FIGURE 1.8 – Illustration de l'application des filtres w_i aux différents nœuds de l'arbre dans l'approche MCRT, dans le cas d'une classification à quatre classes ('car', 'bus', 'motorbike', 'bike'). Les cartes des positions filtrées montrent les positions évaluées (en vert) et les positions non évaluées (en rouge) par chacun des filtres.

Soit $\phi_j(x)$ le vecteur de caractéristiques extrait à un nœud quelconque n_j. On note $\Phi_i(x)$ le vecteur de caractéristiques résultant de la concaténation des vecteurs $\phi_j(x)$ des nœuds situés sur le chemin de la racine jusqu'au nœud n_i. w^i est le vecteur de poids composé de l'ensemble des paramètres des filtres w_j correspondants aux mêmes nœuds. Le score obtenu à une position x au nœud n_i est alors égal à :

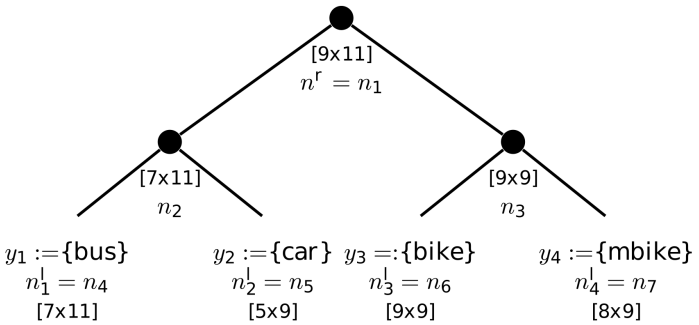
$$\text{score}_{n_i}(x) = w^i \cdot \Phi_i(x) = \sum_{n_j \in \text{anc}(n_i)} w_j^T \cdot \phi_j(x) \tag{1.3}$$

Avec w étant le vecteur global de décision, le score final de prédiction à la position x de la

classe y est calculé par :

$$\text{score}_y(x) = w \cdot \Phi_y^l(x) = \sum_{n_i \in \text{anc}(n_y^l)} w_i^T \cdot \phi_i(x) \tag{1.4}$$

Un exemple de calcul du score est donné Figure 1.9.



$$\begin{aligned}\text{score}_{y_2}(x) &= w \cdot \Phi_{y_2}^l(x) \\ &= w_1 \cdot \phi_1(x) + w_2 \cdot \phi_2(x) + w_5 \cdot \phi_5^l(x)\end{aligned}$$

FIGURE 1.9 – Exemple de calcul du score de prédiction pour la classe 'car' : le score est la somme des scores obtenus aux nœuds traversés. A noter que les vecteurs de caractéristiques sont extraits à chaque nœud sur un support adapté aux classes représentées.

La complexité du modèle égale à $\mathcal{O}(|\text{T}|)$ dépend du nombre de filtres.

1.3.2 Inférence rapide

La stratégie de parcours rapide de l'arbre dans MCRT est inspiré de l'algorithme A*. Cet algorithme permet de trouver dans un graphe le chemin le plus court entre un nœud source et un nœud d'arrivée de manière plus efficace que l'algorithme de Dijkstra, à condition de disposer d'une estimation, à un nœud donné, du coût restant jusqu'au nœud d'arrivée. Cette estimation qui doit être inférieure ou égale au coût réel est appelée heuristique admissible. Plus sa valeur est proche du coût réel, plus le parcours est rapide car moins de nœuds sont visités. Le choix des branches est réalisé en comparant les valeurs de score de A* qui sont la somme des coûts des arcs empruntés et de la valeur de l'heuristique calculées sur chacune des branches.

Dans l'approche MCRT, au lieu de sommer des distances ou des coûts entre les nœuds, nous sommons des scores de classification et nous transformons le problème de calcul du chemin le plus court, c'est-à-dire de coût minimal, en recherche du chemin optimal qui mène à la bonne classe avec le score le plus grand. L'heuristique admissible est donc ici une estimation haute du score jusqu'à la feuille de la classe à prédire. Dans l'algorithme A*, comme les différentes branches possibles sont stockées en mémoire, il est possible à tout moment d'emprunter un autre chemin si son score estimé est supérieur à celui d'une autre branche.

En notant $g(x, n_i)$ la fonction de classification qui est le score au nœud n_i :

$$g(x, n_i) = w^i \cdot \Phi_i(x) \tag{1.5}$$

L'estimation du score final au nœud n_i est donnée par la fonction suivante :

$$f(x, n_i) = g(x, n_i) + t_i \tag{1.6}$$

où t_i est une heuristique admissible qui a une valeur constante définie pendant l'apprentissage.

1.3.3 Apprentissage du modèle hiérarchique

Apprentissage de la structure de l'arbre

Les classes les plus similaires sont regroupées dans l'arbre afin de partager les caractéristiques. Pour déterminer automatiquement la structure de l'arbre et les filtres à chaque nœud, nous calculons la matrice de similarité $\mathcal{S} : k \times k$ qui mesure l'affinité s_{ij} entre les paires de classes $(y_i, y_j) \in \mathcal{Y}^+ \times \mathcal{Y}^+$ sur la base de validation.

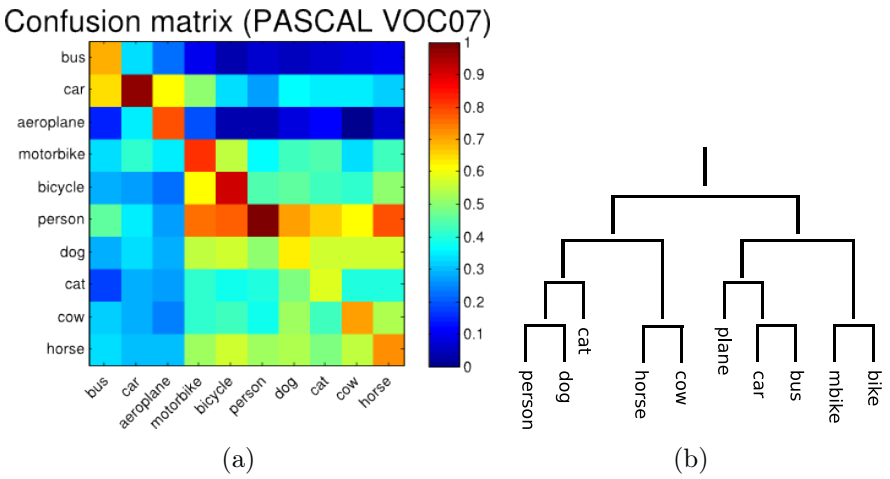
L'affinité s_{ij} est la valeur médiane des scores de classification obtenus en classifiant les exemples de la classe y_i avec un détecteur de la classe y_j. Pour simplifier, le détecteur utilise des descripteurs de type HOG, mais la méthode peut être appliquée avec d'autres types de filtres.

On cherche pour un nœud n_i, ses deux enfants (n_{c_1}, n_{c_2}) tels que la similarité inter-classes soit minimale :

$$(n_{c_1}, n_{c_2}) = \underset{\tilde{n}_{c_1}, \tilde{n}_{c_2}}{\arg\min} \left\{ \texttt{sim}(\tilde{n}_{c_1}, \tilde{n}_{c_2}) \mid \text{toutes combinaisons } (\tilde{n}_{c_1}, \tilde{n}_{c_2}) \right\}. \tag{1.7}$$

où $\texttt{sim}(n_{c_1}, n_{c_2}) = \sum_{y_i \in n_{c_1}} \sum_{y_j \in n_{c_2}} s_{ij}$ est la similarité entre les super-classes.

Ce problème est résolu de manière hiérarchique avec un algorithme de classification spectrale sur la matrice de similarité (Shi et Malik, 2000). L'algorithme regroupe les classes similaires et distribue dans des nœuds différents les classes dissemblables. Un exemple de partition des classes de la base PASCAL VOC 2007 est donné en Figure 1.10.



(a) (b)

Figure 1.10 – (a) Matrice de similarité (matrice de confusion) $\mathcal{S}$ pour $k = 10$ (calculée sur les classes de la base PASCAL VOC 2007). (b) Arbre résultant de la matrice de similarité $\mathcal{S}$ en appliquant l'algorithme de classification spectrale (Shi et Malik, 2000). De manière assez logique, on constate que les classes de véhicules sont regroupées, et que les classes d'animaux sont mises ensemble.

Apprentissage des paramètres du modèle

Les dimensions des filtres des feuilles sont déterminées à partir des boîtes englobantes de vérité terrain sur la base d'apprentissage. Pour chaque classe, le rapport de forme adopté est la moyenne des rapports de forme des boîtes d'annotation et les dimensions des nœuds super-classes

sont la largeur maximum $W(n_i)$ et la hauteur maximum $H(n_i)$ de ses enfants :

$$W(n_i) = \max_{n_j \in \text{desc}(n_i)} W(n_j) \ , \ H(n_i) = \max_{n_j \in \text{desc}(n_i)} H(n_j). \tag{1.8}$$

L'approche choisie pour déterminer les paramètres du modèle est une optimisation combinant des contraintes de classification et de tri, à partir d'un ensemble de n^+ exemples positifs et n^- exemples négatifs. L'optimisation est conduite pour minimiser le risque empirique sur les exemples d'apprentissage.

La formulation du problème d'optimisation s'écrit :

$$\min_{w,\xi_i \geqslant 0, \xi_{ij} \geqslant 0} \frac{1}{2}\|w\|^2 + C \left(\sum_{i=1}^{n^+} (\xi_i + \sum_{j=1}^{n^+} \xi_{ij}) + \sum_{i=1}^{n^-} \xi_i \right) \tag{1.9}$$

avec les contraintes suivantes :

$$\forall y_i \in \mathcal{Y}^+, \forall y_j \in \mathcal{Y}^+ \backslash \{y_i\} : w \cdot \delta\Phi_i(y_j) \geqslant 1 - \xi_{ij} \tag{1.10}$$

$$\forall y_i \in \mathcal{Y}^+ : w \cdot \Phi^l_{y_i}(x_i) \geqslant 1 - \xi_i \tag{1.11}$$

$$\forall y_i \in \mathcal{Y}^-, \forall y_j \in \mathcal{Y}^+ : \ -w \cdot \Phi^l_{y_j}(x_i) \geqslant 1 - \xi_i \tag{1.12}$$

où $\delta\Phi_i(y) = \Phi^l_{y_i}(x_i) - \Phi^l_y(x_i)$ et $w = (w_r, \cdots, w_{|\mathrm{T}|})$ est la concaténation des vecteurs de poids de chaque nœud dans l'arbre T.

Il s'agit d'une formulation identique à celle d'un SVM structuré (S-SVM) (Cai et Hofmann, 2004) où la sortie est une fonction $\Phi_y(x) = \Phi(x, y)$. L'optimisation est conduite sous des contraintes de tri (Equation 1.10) et de classification des exemples positifs (Equation 1.11) et des exemples négatifs de fond (Equation 1.12). La contrainte de tri assure que le score de la bonne classe soit supérieur aux scores des autres classes, et donc que le chemin emprunté par l'exemple dans l'arbre lors de l'inférence aboutisse à la feuille de la classe attendue. Les contraintes de classification servent à apprendre les paramètres de w tels que les exemples positifs et négatifs soient séparés et que les exemples de fond soient rejetés.

En utilisant une formulation avec seule variable ressort ξ pour toutes les contraintes, le problème d'optimisation est réduit à :

$$\min_{w,\xi \geqslant 0} \frac{1}{2}\|w\|^2 + \frac{C}{n}\xi \tag{1.13}$$

avec la contrainte que $\forall(\bar{y}_1, \ldots, \bar{y}_n) \in \mathcal{Y}^{+n}$:

$$\sum_{i=1}^{n^+} w \cdot \Phi^l_{y_i}(x_i) - \max(w \cdot \Phi^l_{\bar{y}_i}(x_i), 0) - \sum_{i=1}^{n^-} w \cdot \Phi^l_{\bar{y}_i}(x_i) \geqslant 1 - \xi \tag{1.14}$$

Le nombre de contraintes augmente exponentiellement avec le nombre d'exemples n car il faut considérer pour chaque exemple tous les labels positifs $\bar{y}_i$. Pour résoudre ce problème de manière efficace, nous utilisons l'algorithme des plans sécants (Joachims et al., 2009). Cet algorithme est rapide car sa complexité est linéaire avec le nombre d'exemples d'apprentissage. Au lieu de traiter toutes les contraintes possibles, une seule contrainte est ajoutée incrémentalement à un sous-ensemble de contraintes. En partant d'un ensemble vide de contraintes, on ajoute à chaque itération la contrainte la plus forte qui correspond, pour chaque exemple x_i de la classe y_i à une autre classe $\bar{y}_i \in \mathcal{Y}^+ \setminus \{y_i\}$ qui lui est la plus proche : $\bar{y}_i = \arg\max_{y \in \mathcal{Y}^+ \setminus \{y_i\}} 1 + w \cdot \Phi^l_y(x_i)$.

Détermination des heuristiques admissibles

Les heuristiques admissibles t_i dans l'algorithme d'inférence doivent donner une estimation optimiste (supérieure) du vrai gain du score pour la classe correcte, à chaque nœud n_i. Plus l'heuristique est proche du vrai gain, plus le parcours sera rapide. Les heuristiques sont déterminées de haut vers le bas et de gauche à droite. Sur une base de validation, on dispose d'un ensemble de boîtes de vérité terrain. MCRT est exécuté dans les régions autour de ces annotations pour produire des détections. Les détections considérées valides, c'est-à-dire recouvrant la boîte de vérité terrain au-dessus d'un certain seuil, sont retenues et constitue un ensemble D. Pour chacune de ces détections on calcule le score du chemin entre chaque nœud n_i et la feuille de la classe de l'objet. L'heuristique admissible est alors choisie comme la borne supérieure des scores pour tous les chemins individuels des éléments de D :

$$\forall (x,y) \in D : t_i \geqslant \max_{\forall D_j \subset D} \min_{\forall (x,y) \in D_j} \sum_{n \in \text{path}(n_i, n_y^l)} \underbrace{w_n \cdot \phi_n(x)}_{\text{individual filter score}} \tag{1.15}$$

Une seule instance de détection par objet de vérité terrain est conservée. Différents choix de l'ensemble D sont possibles. En retirant de D un certain poucentage des meilleures détections, c'est-à-dire les scores les plus grands, on obtient des heuristiques plus strictes (de valeur plus faible), qui ont pour conséquence d'accélérer l'inférence en rejetant plus tôt les exemples négatifs, au détriment de la précision de la détection.

1.3.4 Evaluation expérimentale

Les ensembles de données utilisés pour l'évaluation de MCRT sont PASCAL VOC 2007 et PASCAL VOC 2010 et leur protocole standard (Everingham et al., 2007, 2010b). Ces bases comportent des exemples de 20 classes d'objets et plus de 12000 objets annotés. Les expériences sont menées pour des ensembles de classes de taille différente avec $k = \{2, 4, 6, 8, 10, 20\}$ où les k premières entrées de la liste des classes {'bus', 'bicycle', 'motorbike', 'car', 'aeroplane', 'person', 'cow', 'horse', 'dog', 'cat', 'bird', 'boat', 'bottle', 'chair', 'diningtable', 'pottedplant', 'sheep','sofa', 'train', 'tvmonitor'} sont sélectionnées. Certaines classes ont des similitudes visuelles, d'autres semblent a priori peu semblables.

Les modèles évalués sont :

- OvA : stratégie 'un-contre-tous' : un détecteur par classe est appris, les scores de détection sont calibrés avec une sigmoïde (Platt et al., 2000) ;
- MCR : modèle plat (sans hiérarchie) optimisé avec des contraintes de classification et de tri ;
- MCRT : modèle hiérarchique optimisé avec des contraintes de classification et de tri, les heuristiques admissibles sont apprises ;
- eMCRT : modèle hiérarchique optimisé avec des contraintes de classification et de tri, version exhaustive (tous les nœuds) sont visités ;
- fMCRT : modèle hiérarchique optimisé avec des contraintes de classification et de tri, version rapide (heuristiques admissibles plus strictes) donnant une précision équivalente à la version OvA.

Tous les modèles implémentent les descripteurs HOG tels que décrits dans (Felzenszwalb et al., 2010a). Le Tableau 1.1 résume les propriétés des différents modèles testés ainsi que leur complexité en inférence.

La précision évaluée avec la métrique *mean Average Precision* (mAP) et la vitesse d'inférence pour les modèles MCR, MCRT, eMCRT et fMCRT sont données dans le Tableau 1.2 et comparées aux performances de la stratégie OvA. On constate que les modèles optimisés avec des contraintes de classification et de tri améliorent la précision par rapport à la méthode un-contre-tous (jusqu'à +5 points de mAP sur les deux bases). Les modèles hiérarchiques sont plus précis

Modèle	Classif.	Tri	Hiérarchie	Inférence
OvA	✓			$\|T\|$
MCR	✓	✓		$\|T\|$
MCRT	✓	✓	✓	$< \|T\|$
fMCRT	✓	✓	✓	$\ll \|T\|$
eMCRT	✓	✓	✓	$\|T\|$

TABLEAU 1.1 – Propriétés des modèles de détection

k	2		4		6		8		10		20	
Evaluation	mAP	Speed	mAP	Speed	mAP	Speed	mAP	Speed	mAP	Speed	mAP	Speed
OvA	21.4	1x	22.3	1x	17.8	1x	16.1	1x	12.9	1x	8.8	1x
MCR	23.7	1x	23.0	1x	20.4	1x	17.4	1x	14.7	1x	12.2	1x
MCRT	25.7	1.12x	25.4	0.7x	21.8	0.6x	20.3	0.6x	17.2	0.55x	13.1	0.58x
eMCRT	**25.7**	0.66x	**25.4**	0.55x	**21.8**	0.51x	**20.3**	0.51x	**17.2**	0.51x	**13.1**	0.4x
fMCRT	21.4	**1.5x**	22.3	**2.6x**	17.8	**2.9x**	16.1	**4.6x**	12.9	**6.4x**	8.8	**9.8x**

(a) VOC 2007

k	2		4		6		8		10		20	
Evaluation	mAP	Speed	mAP	Speed	mAP	Speed	mAP	Speed	mAP	Speed	mAP	Speed
OvA	30.5	1x	22.7	1x	21.2	1x	17.0	1x	13.4	1x	9.8	1x
MCR	25.8	1x	19.8	1x	23.2	1x	18.6	1x	16.1	1x	13.0	1x
MCRT	31.4	1.5x	22.8	2.4x	24.5	1.16x	20.0	0.52x	18.0	1.32x	14.4	0.82x
eMCRT	**31.4**	0.67x	**22.8**	0.56x	**24.5**	0.48x	**20.8**	0.47x	**19.0**	0.48x	**14.4**	0.48x
fMCRT	30.5	**1.7x**	22.7	**2.7x**	21.2	**3x**	17.0	**4.2x**	13.4	**5.7x**	9.8	**10.1x**

(b) VOC 2010 *offline*

TABLEAU 1.2 – Précision (mAP) et facteur d'accélération pour les 4 modèles de détection étudiés, comparés avec ceux de la stratégie OvA, et évalués sur (a) la base PASCAL VOC 2007 (b) la base PASCAL VOC 2010 *offline*.

que le modèle MCR, ce qui montre l'intérêt de la structure d'arbre pour le partage des caractéristiques entre classes similaires. La version exhaustive donne les meilleures performances de détection, bien que MCRT produise des résultats similaires. L'algorithme de parcours rapide de l'arbre avec une heuristique admissible apprise est donc efficace puisque la précision n'est presque pas impactée. Toutefois, le gain en vitesse n'est pas toujours évident : le modèle hiérarchique est parfois plus lent que MCR ou la méthode OvA. La version rapide fMCRT, dont l'heuristique est calibrée pour donner le même niveau de performance que la méthode OvA, est en revanche beaucoup plus rapide : le gain en vitesse augmente avec le nombre de classes et atteint 10x pour 20 classes. La Figure 1.11 montre les courbes représentatives du compromis entre la précision de détection et la vitesse d'inférence pour la version la plus précise (eMCRT et la version la plus rapide (fMCRT des modèles de détection hiérarchiques, en fonction du nombre de classes.

1.3.5 Conclusion sur l'approche hiérarchique proposée

Nous avons présenté dans cette section un nouveau framework générique de détection multi-classes fondé sur l'apprentissage d'un modèle hiérarchique des classes, optimisé sous des contraintes de classification et de tri, ainsi qu'un algorithme de parcours efficace de l'arbre pour une inférence rapide (ODABAI FARD et al., 2014b ; ODABAI FARD, 2015). La structure d'arbre, les paramètres des filtres et les heuristiques pour l'inférence sont déterminés automatiquement à l'apprentissage. Un avantage de la hiérarchie de classes est le partage des caractéristiques et des paramètres des

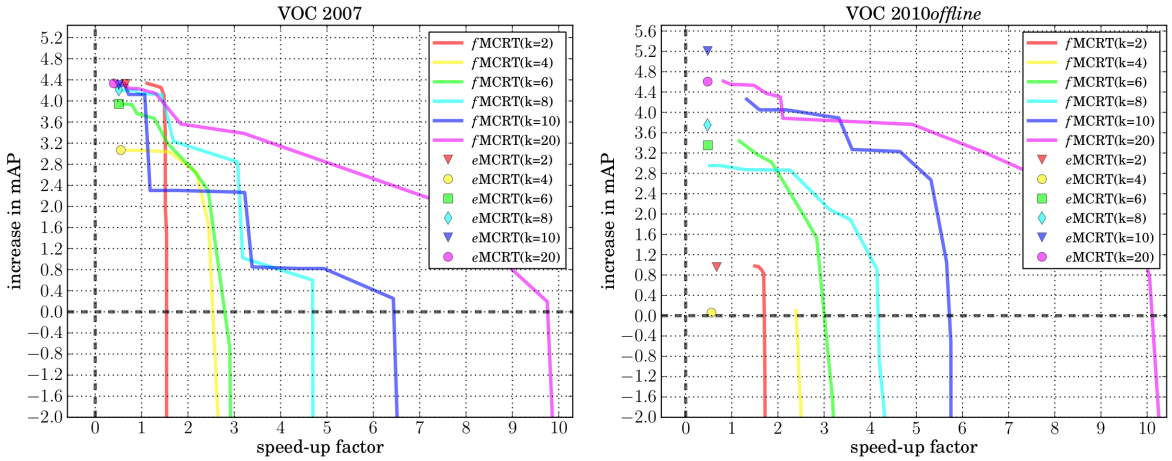


FIGURE 1.11 – Compromis entre précision et facteur d'accélération pour les modèles eMCRT et fMCRT. La référence pour la précision (0 en ordonnée) et la vitesse (1 en abcisse) est donnée par la méthode OvA.

filtres pour les classes visuellement similaires. Les résultats expérimentaux montrent une amélioration de la précision avec ce type de modèles. D'autre part, en jouant sur le seuil de l'heuristique admissible de l'algorithme d'inférence, on peut régler le rapport entre la vitesse de l'inférence et la précision de la détection. C'est une propriété très intéressante d'un point de vue applicatif, car les exigences de précision et de rapidité varient souvent d'un cas d'usage à un autre. Toutefois, l'étude a été limitée à un seul type de descripteurs simples (HOG) et l'évaluation sur des bases de taille modeste par rapport à celle des ensembles d'images apparus ultérieurement. Il aurait été intéressant d'étudier l'influence du choix des descripteurs sur les performances. Bien que la complexité des modèles MCRT puisse être bien inférieure à celle de la méthode OvA, elle reste dépendante du nombre de filtres à implémenter dans l'arbre. Le passage à l'échelle avec un grand nombre de classes constitue de ce fait une difficulté évidente pour construire un arbre qui reste efficace à l'inférence. Une autre limitation est la construction de l'arbre qui n'est pas optimisée simultanément avec les poids des filtres. A condition de disposer d'architectures de calcul adaptées (GPU), les modèles de réseaux de neurones convolutifs pour la détection d'objets multi-classes apportent des réponses sur plusieurs aspects : l'apprentissage de représentations discriminantes pour la détection, le partage d'une très grande partie des paramètres sur toutes les classes, et la gestion d'un grand nombre de classes avec un surcoût calculatoire très limité en comparaison avec les opérations dédiées au calcul des caractéristiques communes à l'ensemble des classes.

1.4 LapNet : un modèle single-shot pour la détection rapide d'objets

Dans cette section, nous décrivons une nouvelle méthode de détection d'objet qui se situe dans la famille des modèles des réseaux de neurones convolutifs *single-shot* tels que YOLO (REDMON et FARHADI, 2018), SSD (LIU et al., 2016b), RetinaNet (LIN et al., 2017c) et FCOS (TIAN et al., 2019). A la différence des approches *two-stage*, les approches *single-shot* effectuent la détection en une seul passage avant du réseau de neurones. Elles sont donc plus adaptées aux applications temps réel, mais souffrent parfois d'une moins bonne précision de détection. Comme dans les travaux précédents, nous cherchons à concevoir un nouveau modèle de détection d'objets ayant le meilleur rapport performance/complexité de calcul possible. Plus spécifiquement, nous visons une vitesse d'inférence supérieure à 10fps sur une carte GPU grand public de type Titan X.

L'approche que nous proposons, appelée LapNet (CHABOT et al., 2020), utilise un système

d'ancres et prédit de manière dense les boîtes englobantes et les labels des objets. Dans ces travaux, nous tentons de résoudre plusieurs problèmes survenant à l'apprentissage, en particulier pour les détecteurs utilisant des ancres :

Qualité de l'association entre les ancres et les objets Avec le critère classique d'*Absolute Overlap*(AO) qui repose sur la comparaison entre l'IoU d'une ancre avec une boîte de vérité terrain par rapport à un seuil, habituellement fixé à 0.5, il peut arriver que, du fait de la discrétisation des ancres, les petits objets et les objets fortement occultés soient écartés à cause de la faible valeur de l'IoU.

Déséquilibres des classes et des tailles d'objets Le déséquilibre des classes est un problème général lié au la représentativité variable des classes dans les bases d'apprentissage. Plus une classe est représentée, plus le modèle va la prendre en compte durant l'entraînement. A l'inverse, pour les classes moins bien représentées, la détection sera généralement moins bonne. Les objets de grande taille ont plus d'influence dans l'optimisation, car davantage d'ancres leur sont associées. Certaines approches traitent ces problèmes de déséquilibre par des techniques d'échantillonnage des exemples (Fu et al., 2017 ; Liu et al., 2016b ; Ren et al., 2015), ou des fonctions de perte spécifiques destinées à renforcer l'influence des exemples difficiles (Lin et al., 2017c ; Tian et al., 2019).

Pour résoudre ces problèmes, nous proposons plusieurs innovations :

— Un nouveau critère de mesure du recouvrement entre une ancre et une vérité terrain, le PONO (*Per Object Normalized Overlap*) permettant d'améliorer la qualité de l'association entre les ancres et les boîtes annotées par rapport au critère classique d'IoU (*Intersection over Union*) en particulier pour les objets de petite taille dans l'image ;

— Une stratégie de filtrage des ancres dont l'assignation aux objets est ambiguë : ce sont les ancres généralement situées à la frontière de plusieurs objets et qui perturbent l'apprentissage du modèle ;

— Une méthode de pondération automatique des fonctions de perte, des classes et des ancres, inspirée des travaux sur l'optimisation multi-tâches (Kendall et al., 2018). Cette méthode vise à pallier les déséquilibres entre classes et tailles d'objets avec une formulation unique.

1.4.1 Description de l'approche

Architecture

L'architecture de LapNet, illustrée dans la Figure 1.12 adopte la structure encodeur/décodeur souvent rencontrée dans les approches de segmentation sémantique. La *backbone* encode l'image à plusieurs résolutions et niveaux sémantiques. Comme dans les Feature Pyramid Networks (FPN), le décodeur est relié à l'encodeur par des connexions latérales à plusieurs niveaux de résolution et les caractéristiques sont agrégées dans le décodeur dans le sens descendant. Mais contrairement aux FPN, au lieu d'effectuer les prédictions à plusieurs échelles, les caractéristiques en sortie du décodeur sont redimensionnées à la plus grande résolution après un passage dans quatre couches de convolution 3x3, puis concaténées. Ce choix est justifié par la recherche d'une plus grande généricité et souplesse par rapport à la taille des objets présents dans la base d'apprentissage. En effet, par souci de simplicité de mise en œuvre, nous voulons éviter de devoir choisir et associer explicitement des ensembles d'ancres en fonction de leur taille aux différents niveaux de la pyramide. Les différentes échelles des objets sont gérées grâce au mécanisme de pondération automatique des ancres (Section 1.4.2). Notons que la contrepartie de ce choix est une plus grande occupation mémoire. Les cartes de caractéristiques concaténées alimentent deux têtes, une pour la classification des ancres, la deuxième pour la localisation des boîtes (régression des offsets des boîtes d'ancres pour les ajuster aux objets). Chaque tête est constituée de quatre couches de convolution 3x3 et d'une couche finale de convolution donnant la sortie du réseau.

Soit N_A le nombre de type d'ancres prédéfinies, N_C le nombre de classes. Les ancres sont déterminées avec la méthode de clustering des K-moyennes sur les boîtes d'annotation de la base d'apprentissage comme dans (Redmon et Farhadi, 2018). Contrairement aux autres méthodes, les ancres sont différenciées par classe. Ainsi, LapNet prédit à une carte de classification de taille $N_A \times N_C$ et une carte de coordonnées de boîtes de taille $N_A \times N_C \times 4$ à chaque position de la grille. Les coordonnées sont obtenues par régressant les offsets (dx, dy, dw, dh) à appliquer aux ancres initiales pour les ajuster aux boîtes de vérité terrain.

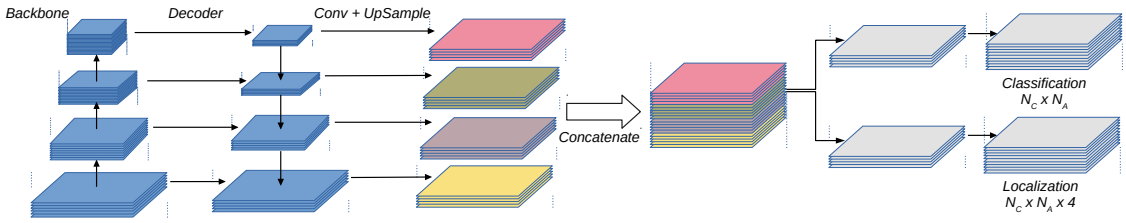


Figure 1.12 – Architecture de LapNet. L'image est encodée puis décodée en caractéristiques à plusieurs niveaux de résolutions. Ces caractéristiques subissent une série de convolutions 3x3 et sont redimensionnées à la plus haute résolution puis concaténées. Une tête de classification et une tête de localisation prennent en entrée cette carte de caractéristiques multi-résolution et génèrent les sorties du réseau.

Association des ancres aux boîtes de vérité terrain

Une grille dense d'ancres de taille $H_f \times W_f \times N_C \times N_A \times 4$ est formée pour prédire les coordonnées des boîtes détectées. H_f et W_f sont les dimensions de la carte de caractéristiques de sortie du réseau avant les têtes.

Le critère généralement employé pour effectuer cette labellisation est le recouvrement absolu (*Absolute Overlap* ou AO) qui est la comparaison de l'IoU entre une ancre et une boîte de vérité terrain et un seuil en général égal à 0.5 (Lin et al., 2017c ; Liu et al., 2016b ; Ren et al., 2015). Le problème de l'AO est que les petits objets et les objets occultés sont parfois écartés, du fait de la discrétisation des ancres et du faible recouvrement avec l'ancre la plus proche.

Afin d'atténuer ce phénomène et rehausser les valeurs du critère de recouvrement, nous introduisons une nouvelle mesure nommée PONO pour *Per-Object Normalized Overlap*. En notant une boîte de vérité terrain B_n et une ancre $A_{c,a,i,j}$ où c désigne la classe, a le type d'ancre, i et j la position de l'ancre sur la grille placée sur la carte de caractéristiques. $\mathcal{C}_{B_n}$ est le cluster des ancres associées à B_n avec le critère d'AO. La mesure du PONO est définie par :

$$O(A_{c,a,i,j}, B_n) = \frac{\text{IoU}(A_{c,a,i,j}, B_n)}{\max\limits_{A_{c,a',i',j'} \in \mathcal{C}_{B_n}} \text{IoU}(A_{c,a',i',j'}, B_n)} \quad \text{with} \quad A_{c,a,i,j} \in \mathcal{C}_{B_n} \tag{1.16}$$

En divisant la valeur de l'IoU par la plus grande valeur d'IoU entre B_n et les ancres associées, on assure d'avoir une valeur de PONO égale à 1 pour au moins une ancre associée à B_n. Les petits objets et les objets occultés, dont l'AO est inférieur au seuil, peuvent avoir un PONO au-dessus du seuil. On désigne par O la carte des valeurs de PONO. La Figure 1.13 illustre le calcul du PONO sur un exemple et compare son effet par rapport à l'AO.

1.4.2 Apprentissage du modèle

Le modèle est entraîné en minimisant deux fonctions de perte, une fonction de classification $\mathcal{L}_{cls}$ servant à prédire si une ancre contient un objet d'une certaine classe, et une fonction de localisation $\mathcal{L}_{loc}$ pour prédire les coordonnées des boîtes de détection.

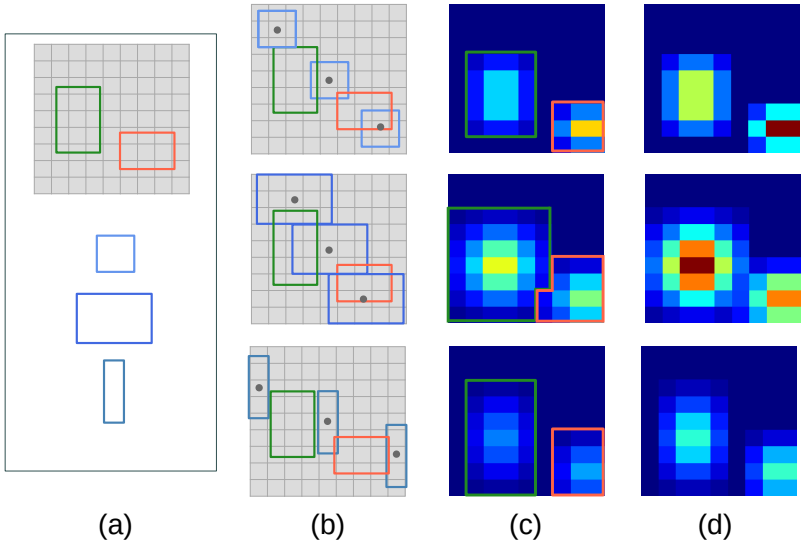


FIGURE 1.13 – Calcul du *Per-Object Normalized Overlap* (PONO). (a) Image d'entrée avec deux objets (haut) et trois types d'ancres prédéfinies (bas). (b) Grille d'ancres. (c) Association des ancres à la vérité terrain avec l'AO. Les clusters d'ancres associées aux deux objets sont représentées en vert et orange. (d) PONO : les deux objets ont au moins une ancre avec un PONO égal à 1 (en rouge).

Fonction de localisation

Pour la fonction $\mathcal{L}_{loc}$, au lieu de régresser directement les décalages des boîtes avec une mesure de distance comme dans (LIN et al., 2017c ; REDMON et FARHADI, 2017 ; REN et al., 2015), nous utilisons la fonction IoU qui est une fonction bornée, pour optimiser les décalages de boîtes de manière latente. Ce choix est justifié expérimentalement par une meilleure stabilité et une meilleure convergence pendant l'entraînement. L'IoU a été utilisée dans (TIAN et al., 2019) pour optimiser des boîtes à partir d'un seul point. En notant $\hat{O}_{c,a,i,j}$ l'IoU entre la boîte prédite et la boîte de vérité terrain associée, la fonction de perte de localisation s'écrit :

$$\mathcal{L}_{loc}(c,a,i,j) = \begin{cases} \|1 - \hat{O}_{c,a,i,j})\|^2, & \text{if } O_{c,a,i,j} > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{1.17}$$

Fonction de classification et filtrage des ancres ambiguës

Pour labelliser les ancres, nous exploitons la carte O de PONO définie précédemment. Afin d'améliorer encore la qualité de cette labellisation, nous introduisons une stratégie de filtrage des ancres dont l'assignement à une vérité terrain est ambiguë. L'ambiguïté provient du fait que ces ancres pourraient être assignées à un objet ou à un autre sans distinction claire. Pour les filtrer, nous utilisons à nouveau la valeur de recouvrement $\hat{O}_{c,a,i,j}$ entre la boîte prédite et la boîte de vérité terrain pour pondérer la valeur de PONO. L'intuition est que le réseau va prédire de faibles valeurs de $\hat{O}_{c,a,i,j}$ pour les ancres ambiguës. Dans ce cas, l'ancre sera considérée négative si le produit du PONO par ce facteur est inférieur ou égale au seuil. La labellisation d'une ancre suit alors la règle suivante :

$$P_{c,a,i,j} = \begin{cases} 1, & \text{if } O_{c,a,i,j} \times \hat{O}_{c,a,i,j} > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{1.18}$$

La fonction de classification est alors donnée par :

$$\mathcal{L}_{cls}(c,a,i,j) = CE(P_{c,a,i,j}, \hat{P}_{c,a,i,j}) \tag{1.19}$$

où CE est la fonction d'entropie croisée binaire et $\hat{P}_{c,a,i,j}$ la probabilité qu'une ancre $A_{c,a,i,j}$ contienne un objet de la classe c. $\hat{P}$ est calculé en appliquant une fonction sigmoïde sur les logits.

La Figure 1.14 donne un aperçu du processus global d'entraînement du modèle LapNet.

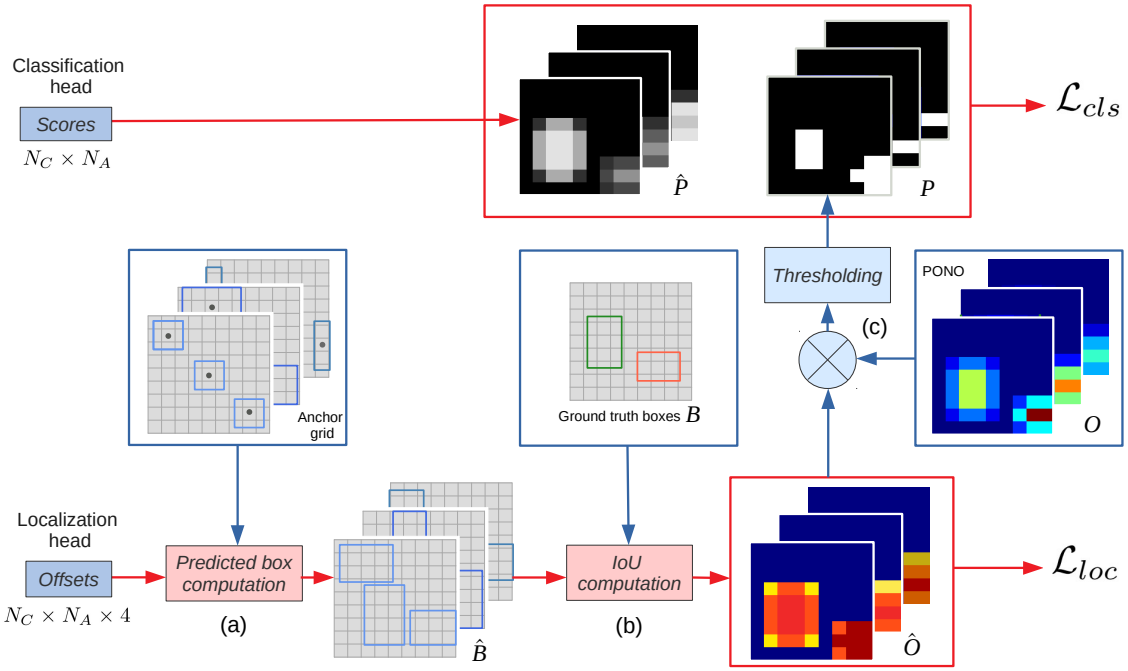


FIGURE 1.14 – Illustration du processus d'apprentissage de LapNet. A partir d'une image, le réseau produit des scores de classification et des décalages de boîtes. (a) Les décalages sont appliquées aux ancres pour prédire les boîtes de détection. (b) La carte d'IoU $\hat{O}$ entre boîtes prédites et vérité terrain est utilisée pour calculer la fonction de localisation $\mathcal{L}_{loc}$. (c) Le produit du PONO O par la carte d'IoU $\hat{O}$ est seuillé pour labelliser les ancres et optimiser la fonction de classification $\mathcal{L}_{cls}$. La rétro-propagation est représentée par le chemin inverse de celui indiqué en rouge.

Pondération automatique des classes et des ancres

Nous étendons l'approche de pondération automatique des fonctions de perte proposée par KENDALL et al., 2018 dans un cadre d'optimisation multi-tâches à la pondération des classes et des ancres. Le principe de cette pondération est la modélisation de l'incertitude homoscédastique de chaque tâche. De cette façon, les tâches les plus difficiles à optimiser sont pondérées plus fortement. Ces poids sont appris en même temps que les paramètres du réseau, ce qui évite une recherche d'hyper-paramètres très coûteuse en temps d'apprentissage. La pondération automatique des classes sert à compenser les déséquilibres de représentativité, alors que celle appliquée sur les ancres permet de corriger l'influence inégale des objets de tailles différentes.

Ainsi les fonctions de perte avec la pondération par tâche, classe et ancre, s'écrivent :

$$\mathcal{L}_{loc} = \lambda_{loc} \frac{1}{N^+} \sum_c \sum_a \lambda_{loc}^{c,a} \sum_{i,j} \mathcal{L}_{loc}(c,a,i,j) \tag{1.20}$$

$$\mathcal{L}_{cls} = \lambda_{cls} \frac{1}{N} \sum_{c} \sum_{a} \lambda_{cls}^{c,a} \sum_{i,j} \mathcal{L}_{cls}(c,a,i,j) \tag{1.21}$$

où $\lambda_{task=loc,cls}$ représente le poids de chaque tâche, et $\lambda_{task=loc,cls}^{c,a}$ le poids de chaque ancre désignée par son type a et la classe c. N^+ est le nombre d'ancres positives pour la tâche de localisation et N le nombre total de positions dans la grille. Un terme de régularisation est ajouté pour empêcher que les poids λ convergent vers 0 :

$$\mathcal{L}_{reg} = log\left(\frac{1}{\lambda_{cls}}\right) + log\left(\frac{1}{\lambda_{loc}}\right) + \frac{1}{N_C N_A} \sum_{c} \sum_{a} log\left(\frac{1}{\lambda_{cls}^{c,a}}\right) + log\left(\frac{1}{\lambda_{loc}^{c,a}}\right) \tag{1.22}$$

1.4.3 Résultats expérimentaux

La méthode est évaluée sur les ensembles de données PASCAL VOC 2007 (Everingham et al., 2010b) et MSCOCO (Lin et al., 2014). Les backbones utilisés sont Darknet-53 (Redmon et Farhadi, 2018) et Inception-Resnet-V2 (Szegedy et al., 2017). Les détails d'implémentation et les hyper-paramètres de l'optimisation sont donnés dans (Chabot et al., 2020). Le Tableau 1.3 compare, sur l'ensemble PASCAL VOC 2007, la performance de LapNet avec d'autres détecteurs de l'état de l'art *two-stage* et *single-shot*, notamment DSSD (Fu et al., 2017). Les résolutions d'image pour les différents modèles sont choisies pour une comparaison équitable des détecteurs. L'évaluation sur PASCAL VOC 2007 montre les performances supérieures de LapNet par rapport à l'état de l'art.

method	backbone	mAP	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Faster(Ren et al., 2015)	VGG-16	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
Faster(Ren et al., 2015)	ResNet-101	76.4	79.8	80.7	76.2	68.3	55.9	85.1	85.3	89.8	56.7	87.8	69.4	88.3	88.9	80.9	78.4	41.7	78.6	79.8	85.3	72.0
MR-CNN(Gidaris et Komodakis, 2015)	VGG-16	78.2	80.3	84.1	78.5	70.8	68.5	88.0	85.9	87.8	60.3	85.2	73.7	87.2	86.5	85.0	76.4	48.5	76.3	75.5	85.0	81.0
R-FCN(Dai et al., 2016)	ResNet-101	80.5	79.9	87.2	81.5	72.0	**69.8**	86.8	88.5	89.8	**67.0**	88.1	74.5	89.8	90.6	79.9	81.2	53.7	81.8	81.5	85.9	79.9
YOLOv2(Redmon et Farhadi, 2017)	DarkNet-19	78.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SSD300(Liu et al., 2016b)	VGG-16	77.5	79.5	83.9	76.0	69.6	50.5	87.0	85.7	88.1	60.3	81.5	77.0	86.1	87.5	83.97	79.4	52.3	77.9	79.5	87.6	76.8
SSD512(Liu et al., 2016b)	VGG-16	79.5	84.8	85.1	81.5	73.0	57.8	87.8	88.3	87.4	63.5	85.4	73.2	86.2	86.7	83.9	82.5	55.6	81.7	79.0	86.6	80.0
SSD513(Fu et al., 2017)	ResNet-101	80.6	84.3	87.6	82.6	71.6	59.0	88.2	88.1	89.3	64.4	85.6	76.2	88.5	88.9	87.5	83.0	53.6	83.9	82.2	87.2	81.3
DSSD513(Fu et al., 2017)	ResNet-101	81.5	86.6	86.2	82.6	74.9	62.5	89.0	88.7	88.8	65.2	87.0	78.7	88.2	89.0	87.5	83.7	51.1	86.3	81.6	85.7	83.7
LapNet512	DarkNet-53	81.7	88.1	88.7	82.0	75.0	65.8	88.1	**91.7**	90.0	65.1	85.9	77.4	88.5	90.8	86.1	**86.2**	51.5	82.7	81.8	89.2	79.4
LapNet512	IncResV2	**83.2**	**89.8**	**89.8**	**83.8**	**76.1**	65.2	**89.8**	90.7	92.0	64.6	**89.8**	**79.0**	**91.8**	**91.8**	**89.5**	84.9	**53.5**	**86.3**	**82.4**	**89.6**	**84.7**

Tableau 1.3 – Comparaison des performances de LapNet avec d'autres détecteurs de l'état de l'art sur PASCAL VOC 2007 `test` set. Les détecteurs *two-stage* et les *single-stage* sont regroupés.

Une étude d'ablation (Tableau 1.4) sur la même base montre :

— l'intérêt du critère PONO par rapport à l'AO ;
— l'impact de la stratégie de filtrage des ancres ambiguës (AMS pour *Ambiguity Management Strategy*) ;
— l'influence de la pondération automatique par tâche, par classe et par ancre.

La combinaison des trois contributions permet d'atteindre les meilleurs résultats. L'analyse de la valeur des poids appris montre que les classes les moins bien représentées et les objets les plus petits bénéficient d'un poids plus important dans l'apprentissage.

Le Tableau 1.5 donne l'évaluation de performances de LapNet sur MSCOCO, et la comparaison avec les approches de l'état de l'art s'évaluant sur la même base. Nous partageons les détecteurs en deux groupes en fixant de manière arbitraire un seuil de vitesse symbolique à 10fps, puisque nous voulons comparer LapNet aux détecteurs les plus rapides. Parmi les détecteurs les plus rapides, LapNet a la meilleure performance de détection (+5 points de mAP par rapport à Yolo v3 (Redmon et Farhadi, 2018)). Les détecteurs plus lents sont plus précis que LapNet, mais l'approche la plus rapide parmi ces détecteurs (FCOS-800 avec Resnet-101 (Tian et al., 2019)) est près de 2.5 fois plus lente que LapNet-608. La Figure 1.15 place les différents modèles de détection en fonction du compromis entre la précision et la vitesse d'exécution.

AMS	no	yes
AO	80.3	78.4
PONO	81.1	**81.7**

(a)

λ_{loc}	λ_{cls}	$\lambda_{loc}^{c,a}$	$\lambda_{cls}^{c,a}$	Loss	mAP
1	R	1	1	CE	79.0
1	R	1	1	FL	71.4
learned	learned	1	1	CE	81.3
learned	learned	learned	learned	CE	**81.7**

(b)

TABLEAU 1.4 – (a) Impact du critère PONO et du filtrage des ancres ambiguës (AMS : Ambiguity Management Strategy) sur les performances de LapNet. AO est le critère de recouvrement absolu (absolute Overlap). (b) Impact de la méthode de pondération automatique par tâche, classe et ancre sur les performance de LapNet. A titre de comparaison, on donne les résultats lorsque le poids des tâches est fixé à la main $R = \frac{N}{N^+}$ et avec la *focal loss* (FL), comme dans (LIN et al., 2017c).

	method	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	ms
Slow Single-stage	SSD-513 (LIU et al., 2016b)	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8	125
	DSSD-513 (FU et al., 2017)	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1	156
	RetinaNet-800 (LIN et al., 2017c)	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2	198
	RetinaNet-800 (LIN et al., 2017c)	ResNeXt-32x8d-101	40.8	61.1	44.1	24.1	44.2	51.2	210
	FCOS-800 (TIAN et al., 2019)	ResNet-101	41.0	60.7	44.1	24.0	44.1	51.0	196
	FCOS-800 (TIAN et al., 2019)	ResNeXt-32x8d-101	42.1	62.1	45.2	**25.6**	44.9	52.0	205
	CornerNet (LAW et DENG, 2019)	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9	244
	CenterNet (ZHOU et al., 2019)	Hourglass-104	**44.9**	**62.4**	**48.1**	**25.6**	**47.4**	**57.4**	300
Fast Single-stage	SSD-321 (LIU et al., 2016b)	ResNet-101-SSD	28.0	45.4	29.3	6.2	28.3	49.3	61
	DSSD-321 (FU et al., 2017)	ResNet-101-DSSD	28.0	46.1	29.2	7.4	28.1	47.6	85
	RetinaNet-500 (LIN et al., 2017c)	ResNet-50	32.5	50.9	34.8	13.9	35.8	46.7	73
	RetinaNet-500 (LIN et al., 2017c)	ResNet-101	34.4	53.1	36.8	14.7	38.5	49.1	90
	YoloV2-544 (REDMON et FARHADI, 2017)	DarkNet-19	21.6	44.0	19.2	5.0	22.4	35.5	25
	YoloV3-608 (REDMON et FARHADI, 2018)	DarkNet-53	33.0	**57.9**	34.4	18.3	35.4	41.9	51
	LapNet-512	DarkNet-53	37.6	55.5	40.4	17.6	40.5	**49.9**	55
	LapNet-608	DarkNet-53	**38.2**	56.6	**41.2**	**20.3**	**41.6**	47.5	80

TABLEAU 1.5 – Evaluation de LapNet sur MSCOCO `test-dev2017` à deux résolutions d'image d'entrée (512x512 et 608x608) et comparaison avec les modèles de l'état de l'art. LapNet dépasse nettement en précision les méthodes les plus rapides tels que Yolo V3. Les détecteurs plus précis que LapNet sont beaucoup plus lents (au moins 2.5x pour FCOS-800).

1.5 Conclusion et perspectives

Dans ce chapitre, nous avons abordé la problème de la détection d'objets multi-classes rapide et robuste, étape clé pour l'interprétation automatique de scènes. Malgré de grandes avancées dans le domaine, la détection d'objets n'est pas un problème entièrement résolu. Comme on peut le constater sur le benchmark construit sur l'ensemble de données MSCOCO, la performance des meilleurs détecteurs d'objets génériques dépasse difficilement les 50-55 points de mAP encore aujourd'hui, et ce au prix d'un temps de calcul sur une plateforme d'exécution standard (GPU grand public) incompatible avec les applications temps réel.

La spécialisation des détecteurs d'objets à un domaine spécifique et restreint (quelques classes d'intérêt, contextes et environnements bien définis) permet souvent d'augmenter nettement la précision des modèles de détection, même si ceux-ci doivent garder une capacité relative de généralisation et ne pas sur-apprendre sur les données d'entraînement. La gestion de l'augmentation du nombre de classes d'objets et de leur variabilité d'aspect avec un maintien de la performance de détection reste un défi à relever : les difficultés résident d'une part dans la collecte et la préparation des données d'apprentissage, de validation et d'évaluation, et d'autre part dans la conception et la mise au point de nouveaux modèles performants sur l'ensemble des classes.

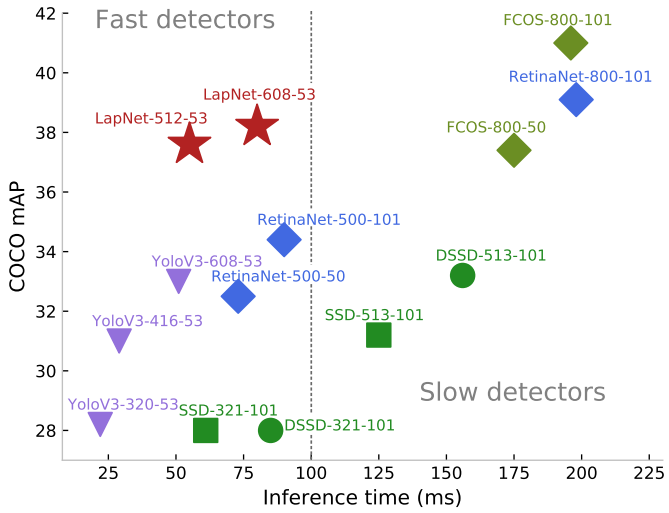


FIGURE 1.15 – Evaluation sur MSCOCO `test-dev` du compromis entre la précision et la vitesse pour les détecteurs *single-shot*, effectuée sur un TITAN X GPU. Seuls les détecteurs tournant à plus de 5fps sont présentés ici. Les modèles sont désignés avec la convention "name-resolution-backbone" (50 : ResNet-50, 101 : ResNet-101, 53 : DarkNet-53). La ligne en pointillés représente le seuil symbolique des 10fps.

Toutefois la performance de détection ne constitue pas à elle-seule le seul critère de choix dans la conception d'une application. Le coût énergétique lié à la complexité de calcul des algorithmes est une deuxième dimension pour laquelle les contraintes matérielles peuvent s'imposer au premier rang. C'est le cas notamment des systèmes autonomes qui doivent embarquer les calculateurs. Les optimisations de code sont utiles pour alléger le calcul sans dégrader les performances, mais ne sont pas une solution définitive. Pour aller plus loin, un compromis entre la performance de détection et le coût de calcul doit être trouvé. C'est précisément cette problématique que nous avons traitée dans les deux contributions présentées dans ce chapitre.

Le premier travail a concerné la proposition et le développement d'un framework générique de détection d'objets multi-classes fondé sur l'apprentissage d'un modèle hiérarchique des classes et utilisant en entrée des descripteurs *ad hoc*. L'approche est inspirée de plusieurs travaux précédents, mais est originale par sa formulation. L'apprentissage du modèle est réalisé sous la forme d'un programme quadratique à optimiser avec des contrainte de classification et de tri. Un algorithme de parcours optimal de l'arbre garantit une inférence rapide, à condition d'avoir une estimation du score du chemin restant. Cette estimation plus ou moins script donne la possibilité de moduler le point de fonctionnement performance de détection/rapidité.

Dans le second travail, l'objectif est toujours de repousser les limites de performances et de rapidité des détecteurs d'objets, cette fois-ci en s'appuyant sur un état de l'art qui a évolué vers les modèles de réseaux de neurones profonds. Ceux-ci ouvrent des possibilités beaucoup plus étendues que les méthodes traditionnelles. L'approche proposée est un nouveau détecteur *single-shot* employant une grille d'ancres et appris de bout en bout. Ce détecteur bénéficie de plusieurs innovations qui portent sur l'association des ancres aux objets et le déséquilibre entre les classes et les tailles d'objet à l'apprentissage. Ces contributions ont permis de placer notre modèle de détection au meilleur niveau de l'état de l'art parmi les algorithmes les plus rapides.

L'enjeu de la réduction du temps de calcul/de la complexité calculatoire pour l'exécution des modèles de détection sur des cibles matérielles limitées en puissance motive de nombreux travaux de recherche. Plusieurs pistes sont examinées : recherche d'architectures de réseaux plus efficaces, automatisée dans certains cas (TAN et al., 2020), méthodes de distillation de

connaissances (Chen et al., 2017), stratégies d'élagage de réseaux (Choudhary et al., 2020), optimisation sur des cibles spécifiques avec des techniques de réduction de précision.

Un autre axe de recherche consiste à rapprocher la détection d'objets d'autres tâches de vision. La recherche de la précision de détourage des objets au niveau pixellique converge avec les objectifs de la segmentation d'instances (Hafiz et Bhat, 2020) et de la segmentation panoptique qui combine la segmentation sémantique et la segmentation d'instances (Kirillov et al., 2019). Au sein de notre laboratoire, une thèse sur le sujet de la segmentation panoptique avec des approches semi-supervisées est en cours de réalisation (thèse d'Adel Redjimi). La détection d'objet peut aussi être associée à d'autres tâches comme la reconnaissance fine et l'estimation de pose 3D (Benzine et al., 2020a ; Chabot et al., 2017), la ré-identification (Loesch et al., 2019) dans un cadre d'optimisation de modèle multi-tâches, ou encore le suivi visuel (Feichtenhofer et al., 2017).

Une des difficultés majeures en apprentissage supervisé est le besoin de constituer de bases annotées représentatives, souvent de très grande taille, pour entraîner les modèles. Toutefois, l'étape d'annotation des objets est coûteuse et les données brutes représentatives du cas d'usage à traiter ne sont pas toujours disponibles en quantité suffisante. D'autre part, le changement de domaine (de contexte) ou l'adaptation rapide à de nouvelles tâches, comme l'ajout de nouvelles classes d'objets, sont des problèmes souvent rencontrés en pratique car on veut éviter de repartir de zéro à chaque fois. Nous explorons au laboratoire, dans différents projets de recherche, des approches d'adaptation de domaine pour la détection d'objets et la segmentation sémantique, soit avec un pré-apprentissage auto-supervisé (Chen et al., 2020 ; Grill et al., 2020), ou exploitant le principe de l'apprentissage semi-supervisé où les données du domaine cible ne sont pas labellisées comme dans (Jeong et al., 2019). S'agissant du problème de frugalité des modèles en données, une thèse sur l'apprentissage à partir de très peu d'exemples pour la détection d'objets a été initiée en 2020 dans notre laboratoire (thèse de Quentin Bouniot).

Chapitre 2

Ré-identification visuelle des personnes

2.1 Contexte et motivations

La ré-identification visuelle d'une instance consiste, à partir d'une ou plusieurs observations, à la retrouver automatiquement dans les images ou vidéos, à des instants et lieux différents, en la discriminant d'autres instances de même nature et potentiellement très similaires en apparence. Les observations utilisées pour définir l'identité de l'instance sont nommées *requête*, et celles qui contiennent les candidats potentiels à la ré-identification constituent la *galerie*. La galerie est dite fermée si l'ensemble des identités candidates est fixe, elle est ouverte lorsque de nouvelles identités peuvent être ajoutées ou retirées.

La ré-identification se différencie de l'identification forte par le fait que la notion d'identité n'est définie que par les données visuelles de la requête, qui ne sont pas forcément pas rattachées à la véritable identité. Dans le cas des personnes, celle-ci est établie par des informations d'état civil, pour un véhicule il s'agit de son numéro d'immatriculation. Ainsi, pour les personnes, on ne se place pas dans le cadre de la biométrie où les caractéristiques physiques et propres des individus sont supposées varier peu et lentement, mais dans celui d'une modélisation globale et temporaire de l'apparence visuelle. Cette apparence inclue l'aspect vestimentaire, qui est en principe suffisante pour ré-identifier les personnes à court terme, tant que leur aspect n'a pas trop changé.

Enjeux applicatifs

Parmi les nombreuses applications de la ré-identification visuelle, c'est la vidéo-surveillance intelligente qui a été le principal moteur des recherches sur la ré-identification des personnes depuis plus d'une vingtaine d'années (Dufour, 2012 ; Leng et al., 2020 ; Ye et al., 2021). Dans le domaine de la sécurité, les technologies d'identification et de ré-identification de personnes sont exploitées dans trois cas d'usage principaux :

Le contrôle d'accès Un dispositif de contrôle d'accès dans une infrastructure critique dont l'accès est restreint, a pour rôle de maîtriser les flux entrants et sortants. Les droits d'accès sont définis par l'autorisation donnée à un ensemble de personnes (*white list* en anglais) ou l'interdiction imposée à d'autres individus (*black list*). Le plus souvent, le contrôle d'accès met plutôt en œuvre des technologies d'identification forte.

Le suivi des personnes dans une infrastructure physique La compréhension du comportement des personnes évoluant dans l'infrastructure pour l'évaluation en temps réel de la situation requiert un suivi fiable de chacune d'elles. L'infrastructure est couverte par un réseau de caméras, soit à champs recouvrants, soit à champs disjoints. La modélisation de l'apparence pour la ré-identification des individus peut servir à renforcer le suivi dans une caméra ; elle est indispensable pour apparier les pistes d'un même individu d'une caméra

à une autre, notamment lorsqu'il y a une rupture temporaire de la couverture visuelle. Pour chaque personne, la galerie est constituée par l'ensemble des personnes suivies à chaque instant.

La recherche automatisée de personnes Désignée en anglais par le terme *person search*, cette tâche a pour but de retrouver dans les flux vidéo du réseau de caméras, le plus rapidement possible, une personne désignée à l'aide d'une ou plusieurs images (requête), afin de reconstruire son parcours dans le passé, d'anticiper ses déplacements futurs, ou pour rechercher des preuves lors d'investigations policières effectuées a posteriori (Figure 2.1).

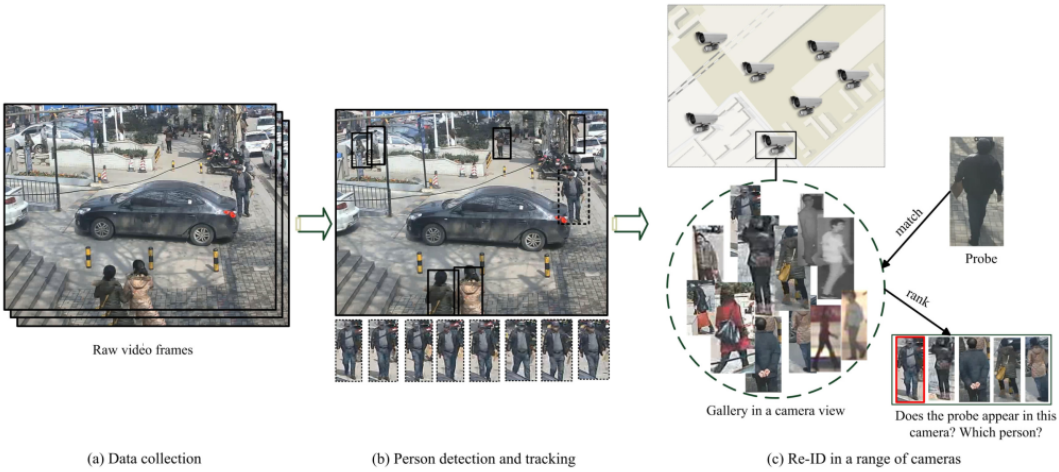


FIGURE 2.1 – Illustration de la ré-identification de personnes dans un scénario d'investigation a posteriori (LENG et al., 2020). Les images des personnes sont extraites automatiquement des vidéos avec des algorithmes de détection et de suivi intra-caméra pour constituer la galerie.

Il est à noter que pour ces deux derniers cas d'usage, lorsqu'il s'agit de la surveillance d'espaces publics (espaces urbains, infrastructures de transport), on se retrouve de façon pratique dans une situation d'ensemble ouvert d'identités, les flux entrants et sortants n'étant pas contrôlés.

Dans d'autres applications, la ré-identification intervient systématiquement pour des scénarios de suivi long terme des personnes et l'analyse de leurs activités, dans une caméra ou un réseau de caméras.

L'accélération de la recherche sur la ré-identification de personnes ces dernières années s'est accompagnée de la mise en libre accès de corpus d'images annotées toujours plus volumineux, et de l'organisation de benchmarks basés sur ces corpus destinés à comparer les performances des méthodes développées (KARANAM et al., 2019).

Le développement de la vidéo-surveillance pose néanmoins des questions éthiques quant à son usage intensif qui peut aller au-delà de la sécurisation des personnes et des biens. La vidéo-surveillance semble parfois être détournée à des fins politiques de limitation des libertés individuelles. C'est ainsi que suite à une enquête du Financial Times (MURGIA et HARLOW, 2019), plusieurs bases d'images pour la reconnaissance faciale et la ré-identification de personnes, dont DukeMTMC4Reid de l'université de Duke (GOU et al., 2017), ont été retirées car soupçonnées d'avoir favorisé le développement aux Etats-Unis et en Chine d'outils de répression et de surveillance systématique des personnes.

Sans aller jusqu'à ces extrémités, la simple exploitation d'images des personnes soulève la problématique des données à caractère personnel. Cette exploitation est aujourd'hui rigoureusement encadrée par le règlement général sur la protection des données (RGPD) dans les pays de l'Union Européenne. Le RGPD, en renforçant la protection des personnes, limite et rend la

collecte et le traitement des images des personnes plus contraignants. En protégeant les citoyens, il constitue un frein au développement technologique.

Autre domaine d'application en plein essor, la gestion de la ville intelligente, et l'optimisation des ses infrastructures et services par la connaissance précise des flux d'usagers a plus récemment incité les chercheurs à transposer la problématique de ré-identification aux véhicules, comme en témoigne la création de nouveaux ensembles de données (Khan et Ullah, 2019) et de challenges (Naphade et al., 2017).

Enfin, de manière plus spécifique, les animaux sont aussi l'objet de travaux sur la ré-identification visuelle dans le cadre de l'étude de leur comportement (Schneider et al., 2020)

Dans ce qui suit, nous nous focaliserons uniquement sur la ré-identification de personnes.

Les défis de la ré-identification de personnes

Martin Handford, dans son célèbre ouvrage "*Where's Wally ?*" nous lance le défi ludique de retrouver Wally (Charlie en français) dans des scènes parfois extraordinairement complexes. Ce jeu donne un avant-goût des difficultés de la tâche de ré-identification, qui sont multiples :

- la variabilité d'apparence pour une même instance : cette variabilité est due aux différents points de vue, à la réponse colorimétrique des caméras, aux variations d'illumination, à la pose des personnes par rapport à la caméra, aux accessoires portés, à la diversité des fonds ;
- la similitude d'apparence entre personnes différentes : deux personnes différentes peuvent se ressembler énormément, car compte tenu de la résolution de leurs images, on ne sera pas en mesure de distinguer les détails qui les discriminent (comme les traits du visage par exemple), il suffit que les couleurs des vêtements soient similaires pour provoquer une confusion ;
- les occultations fréquentes entre personnes, ou par d'autres éléments de la scène ;
- la qualité relative des images : les images proviennent la plupart du temps de caméras de surveillance, qui sont de qualité très variable, du fait de l'héritage historique des installations existantes. D'autre part, pour des raisons de coût ou des raisons techniques d'installation, les angles d'observation et la résolution des personnes dans les images ne sont pas toujours favorables à une ré-identification visuelle ;
- l'imprécision de l'extraction des régions d'intérêt contenant les images des personnes liée à la performance des détecteurs de personnes ;
- la taille de la galerie : plus la galerie est grande, plus il y a des chances de confusion, et plus le temps de calcul augmente ;
- base fermée vs base ouverte : dans des scénarios de base fermée, l'ensemble des identités de la galerie est connue, la personne requête fait forcément partie de la galerie, alors qu'en base ouverte, la personne requête peut être une personne inconnue, aussi appelée imposteur, cette personne doit être rejetée.

La figure 2.2 illustre les difficultés rencontrées en ré-identification, avec des exemples tirés des corpus VIPeR (Gray et al., 2007), ETHZ (Ess et al., 2007), GRID (Chen Change Loy et al., 2009), iLids (Wang et al., 2014b), Shinpuhkan (Kawanishi et al., 2014), PRID2011 (Hirzer et al., 2011), CUHK03 (Li et al., 2014a), Market1501 (Zheng et al., 2015a).

2.2 Etat de l'art et positionnement

Le problème de la ré-identification de personnes a inspiré énormément de recherches depuis une quinzaine d'années. Les deux étapes classiques de la ré-identification sont la modélisation des personnes par un ensemble de caractéristiques visuelles et la recherche de correspondances entre la requête et les candidats de la galerie par le calcul de similarité ou dissimilarité (ou distance) entre

FIGURE 2.2 – Illustration des difficultés rencontrées en ré-identification de personnes. Ligne 1 : variabilité des poses (images des corpus Shinpuhkan, VIPeR, ETHZ). Ligne 2 : variations de couleur liées à l'illumination (image des corpus GRID, iLids, Shinpuhkan). Ligne 3 : problèmes d'alignement liés à l'imprécision de l'extraction des régions d'intérêt (images des corpus Shinpuhkan, Market1501). Ligne 4 : variabilité des fonds (images des corpus VIPeR, PRID2011, CUHK03). Ligne 5 : similitude d'apparence entre personnes différentes (images du corpus VIPeR).

les caractéristiques respectives. La grande majorité des approches s'évaluent dans des scénarios d'ensembles fermés d'identités, le problème plus général et plus complexe de ré-identification en base ouverte n'a été abordé que plus récemment.

2.2.1 Ré-identification en base fermée

Représentation visuelle des personnes

Les premiers travaux en ré-identification de personnes se sont focalisés sur la conception de descripteurs discriminants de l'apparence des personnes. Ces descripteurs, conçus "à la main" de manière intuitive, sont comparés entre eux avec une fonction de distance (distance euclidienne, distance de Bhattacharyya, distance cosinus) ou une combinaison de distances. L'une des premières méthodes de référence est SDALF (Farenzena et al., 2010) qui mélange des informations de couleur et de texture localisées sur les parties du corps en exploitant des propriétés de symétrie et d'asymétrie. La description par parties est exploitée à plusieurs reprises (Bak et al., 2010 ; Cheng et al., 2011a), d'autres approches adoptent la représentation par patches plus simple à obtenir (Zhao et al., 2013a,b, 2014) ou par bandes (Lisanti et al., 2015) qui est moins sujette aux problèmes d'alignement. La modélisation de régions de saillance permet de focaliser sur des détails discriminants des personnes (Zhao et al., 2013a,b). Dans la méthode XQDA (Liao et al., 2015), le descripteur LOMO (*Local Maximal Occurrence Feature*) combine dans des patches recouvrants, à plusieurs échelles, des distributions de couleur et un nouveau descripteur de texture. L'information est réduite par une opération semblable à un *max pooling*.

D'autres travaux ont développé des stratégies de sélection automatique et de pondération de caractéristiques : Gray et Tao, 2008 utilisent un ensemble de classifieurs faibles appris par boosting, tandis que Liu et al., 2012a développent une méthode reposant sur des prototypes de personnes. L'ajout d'information sémantique à la représentation a aussi été exploré : description par couleurs sémantiques (Yang et al., 2014), classification d'attributs de niveau intermédiaire labellisée (Layne et al., 2012a, 2014a ; Shi et al., 2015).

Les approches de description spatio-temporelle exploitent la disponibilité de séquences annotées en tracklets. Wang et al., 2014b modélisent le cycle de la marche avec un descripteur 3D. Ma et al., 2017 établissent des correspondances entre sélectionnant et alignant directement des fragments vidéos.

Apprentissage de métrique

Les approches d'apprentissage de métrique visent à remplacer la distance usuelle (euclidienne) par une distance permettant d'obtenir de meilleures correspondances au sein d'une même identité et d'écarter des identités différentes. La distance de Mahalanobis est souvent employée (Hirzer et al., 2012a ; Koestinger et al., 2012 ; Liao et al., 2015) : cette distance entre deux vecteurs est équivalente à la distance euclidienne entre ces deux éléments projetés par une transformation linéaire. L'apprentissage est généralement supervisé par la contrainte de classement (rangement) relatif des paires positives et négatives, en agissant sur les distances (Dikmen et al., 2010 ; You et al., 2016), uniquement sur le classement (Jose et Fleuret, 2016) ou en estimant des probabilités sur l'ordre des distances pour une paire positive et une paire négative (Zheng et al., 2011b). Les contraintes de minimisation de la variance intra-classe et maximisation de la variance inter-classe sans comparaison directe entre paire positive et paire négatives, ont été étudiées dans diverses approches (Koestinger et al., 2012 ; Liao et al., 2015 ; Mignon et Jurie, 2012 ; Pedagadi et al., 2013 ; Zhang et al., 2016b). Enfin, des combinaisons de plusieurs métriques sont appliquées sur différents descripteurs (Liu et al., 2017) ou paires de caméras (Ma et al., 2017).

Réseaux de neurones profonds pour l'apprentissage de représentations et de métriques

Comme dans les autres tâches de vision, le renouveau des réseaux de neurones a considérablement influencé les recherches en ré-identification de personnes. Depuis 2014 et les travaux de Li et al., 2014a, la littérature sur le sujet de ces dernières années est consacrée presque exclusivement aux méthodes d'apprentissage profond (Ye et al., 2021) qui ont pu se développer en partie grâce à la création d'ensembles de données de grande taille (Figure 2.4). Les modèles de réseaux de neurones produisent des représentations d'une grande expressivité et optimisées pour les tâches cibles (apprentissage de bout-en-bout). Ces méthodes ont quasiment supplanté toutes les approches plus anciennes, compte tenu de leurs performances bien supérieures.

	Image datasets						
Dataset	Time	#ID	#image	#cam.	Label	Res.	Eval.
VIPeR	2007	632	1,264	2	hand	fixed	CMC
iLIDS	2009	119	476	2	hand	vary	CMC
GRID	2009	250	1,275	8	hand	vary	CMC
PRID2011	2011	200	1,134	2	hand	fixed	CMC
CUHK01	2012	971	3,884	2	hand	fixed	CMC
CUHK02	2013	1,816	7,264	10	hand	fixed	CMC
CUHK03	2014	1,467	13,164	2	both	vary	CMC
Market-1501	2015	1,501	32,668	6	both	fixed	C&M
DukeMTMC	2017	1,404	36,411	8	both	fixed	C&M
Airport	2017	9,651	39,902	6	auto	fixed	C&M
MSMT17	2018	4,101	126,441	15	auto	vary	C&M
	Video datasets						
Dataset	time	#ID	#track(#bbox)	#cam.	label	Res.	Eval
PRID-2011	2011	200	400 (40k)	2	hand	fixed	CMC
iLIDS-VID	2014	300	600 (44k)	2	hand	vary	CMC
MARS	2016	1261	20,715 (1M)	6	auto	fixed	C&M
Duke-Video	2018	1,812	4,832 (-)	8	auto	fixed	C&M
Duke-Tracklet	2018	1,788	12,647 (-)	8	auto	C&M	
LPW	2018	2,731	7,694(590K)	4	auto	fixed	C&M
LS-VID	2019	3,772	14,943 (3M)	15	auto	fixed	C&M

Figure 2.3 – Principaux ensembles de données utilisés en ré-identification de personnes (Source : Ye et al., 2021). La taille des jeux de données a considérablement augmenté au fil du temps.

La manière la plus simple d'encoder une image est de générer un vecteur de caractéristiques global à l'image, ou à une paire d'images que l'on souhaite comparer (Wang et al., 2016a). Zheng et al., 2017a proposent d'apprendre des représentations dans un problème de classification multi-classes où chaque classe est une identité distincte. Les modèles d'attention apportent des moyens de mieux focaliser sur des régions discriminantes des personnes et d'être invariant au fond (Chen et al., 2019b ; Li et al., 2018a), dans un groupe de personnes (Chen et al., 2018a), de trouver des relations spatiales entre régions (Zhang et al., 2020b). Les représentations locales sont calculées sur des parties ou des patches pour produire des signatures localisées plus précises. Les représentations locales et globales peuvent être combinées (Suh et al., 2018). Afin de traiter les problèmes d'alignement, d'autres approches exploitent l'information de pose (Su et al., 2017), ou des régions sémantiques (Zhang et al., 2019). Les représentations visuelles sont parfois enrichies par d'autres caractéristiques comme les attributs sémantiques (Lin et al., 2019). Chang et al., 2018 présentent un modèle qui apprend de manière latente différents facteurs discriminants de l'apparence visuelle. Depuis les travaux de Zheng et al., 2017b, plusieurs approches fondées sur les modèles génératifs de type GAN se sont montrées efficaces pour obtenir des représentations plus discriminantes, afin de résoudre le problème d'adaptation de domaine de manière non supervisée (Deng et al., 2018). Quelques travaux portent sur la modélisation de données vidéo pour la ré-identification, avec des réseaux récurrents (McLaughlin et al., 2016 ; Yan et al., 2016). Une méthode de recherche de caractéristiques communes sur plusieurs frames, inspirée des approches de co-segmentation est décrite par Subramaniam et al., 2019.

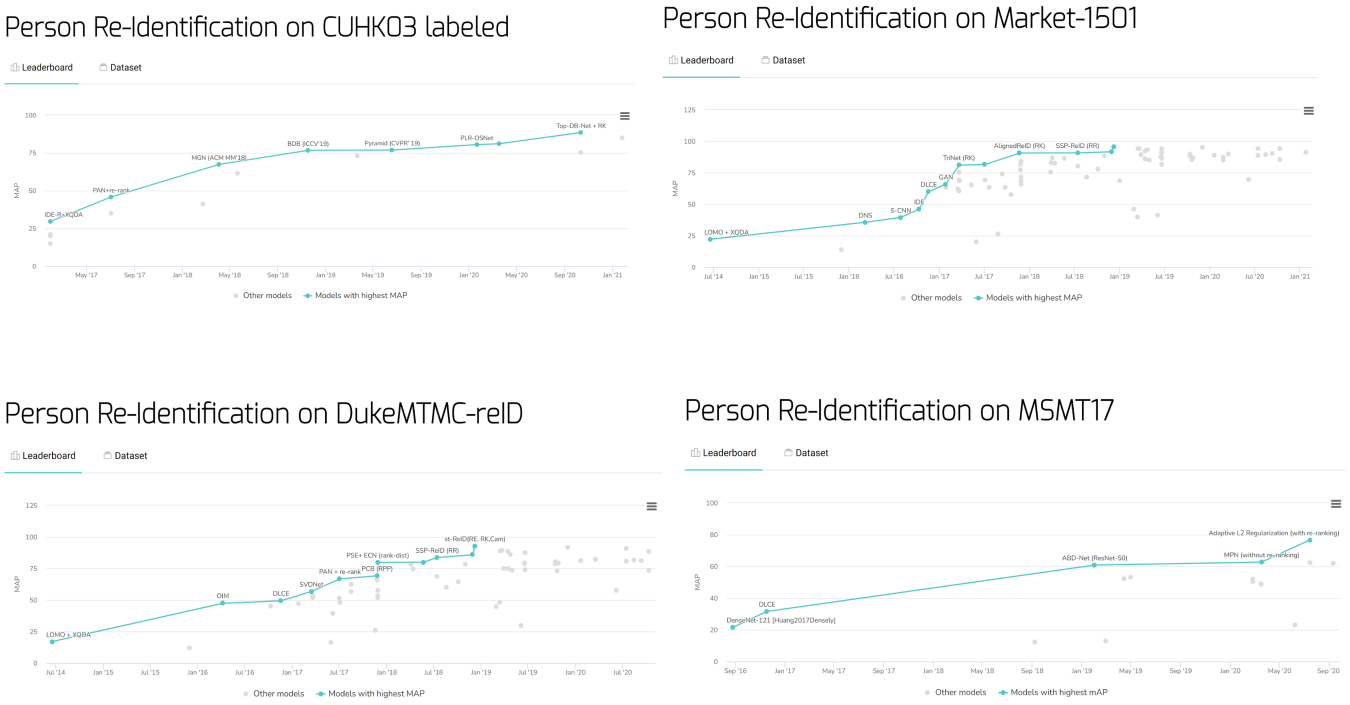


FIGURE 2.4 – Résultats de benchmarks de ré-identification de personnes (graphiques tirés de https://paperswithcode.com/). Les méthodes de ré-identification visuelle en base fermée ont fait un bond en performance entre 2017 et 2020.

L'apprentissage de métrique a été revisité avec les réseaux de neurones profonds (*deep metric learning*) et mis en œuvre sous de nombreuses formes dans les problèmes de ré-identification de personnes.

Lorsque que l'apprentissage du réseau de neurones est réalisé comme dans un problème de classification multi-classes, comme dans (ZHENG et al., 2017a), on utilise des fonctions de perte portant sur l'identité, où les erreurs d'identité doivent être minimisées pendant l'entraînement. En phase de test, le réseau est utilisé comme un extracteur de caractéristiques qui sont comparées avec une distance classique (euclidienne, cosinus). Les travaux de ZHAI et al., 2019 montrent qu'avec les stratégies proposées (multi-branches groupant des ensembles de canaux dans le vecteur de caractéristique), la fonction de perte de classification, réputée moins performante que les autres fonctions de perte liées à l'apprentissage de métrique, peut se révéler très compétitive.

De façon différente, la vérification compare les images par paires : paires positives si les deux images proviennent de la même identité, paires négatives dans le cas contraire. Les paires sont classées par une classification binaire (LI et al., 2014a), en apprenant avec une fonction de coût construite sur la distance cosinus (YI et al., 2014b), ou comparées avec une fonction de perte contrastive. Dans le second cas, on cherche à apprendre une métrique telle que la distance entre images d'une paire positive soit inférieure à celle entre images d'une paire négative, avec une certaine marge (DENG et al., 2018 ; VARIOR et al., 2016). Les fonction de vérification ont été combinées aux fonction de classification pour améliorer les performances de ré-identification (CHEN et al., 2018a ; DENG et al., 2018 ; VARIOR et al., 2016).

La fonction de coût triplet (DING et al., 2015 ; WANG et al., 2016e) étend le concept de comparaison de paires. A partir des représentations de trois images (deux d'une même identité, et une troisième d'une autre identité), on constitue une paire positive et une paire négative qui

ont en commun une image de référence appelée ancre. La distance de la paire négative doit être supérieure à celle de la paire positive, au-delà d'une certaine marge. La fonction de coût triplet a pour effet de ranger correctement les paires. Le choix des négatifs est important pour une bonne performance (Hermans et al., 2017).

Représentations parcimonieuses

Les modèles parcimonieux d'abord rencontrés dans le domaine de la reconnaissance faciale (Wright et al., 2009) ont vu leur apparition en ré-identification de personnes dans plusieurs travaux. Dans (Karanam et al., 2015a, 2017 ; Lisanti et al., 2015), les caractéristiques des images de la galerie constituent le dictionnaire, l'image requête est reconstruite avec une combinaison linéaire des éléments de la galerie les plus représentatifs, et l'erreur de reconstruction sert de score de similarité pour classer ces éléments. Ces approches cherchent à mieux exploiter la disponibilité de plusieurs images par identité (multi-shot). Dans d'autres travaux (An et al., 2016 ; Karanam et al., 2015b ; Kodirov et al., 2015 ; Peng et al., 2016), les codes issus de la représentation parcimonieuse sont utilisés en tant que caractéristiques pour représenter un élément requête ou un élément de la galerie. Plus récemment, He et al., 2018 ont proposé de comparer deux images en reconstruisant l'une par rapport à l'autre à l'aide d'une représentation parcimonieuse des blocs de cartes de caractéristiques de chaque image, issues d'un réseau de neurones convolutif. Les auteurs affirment que la méthode est plus robuste aux problèmes d'alignement du fait de la représentation par blocs non localisés.

Optimisation du classement

Les techniques de re-classement (*re-ranking*) ont pour but d'améliorer davantage la qualité du classement obtenue par les méthodes de ré-identification. Certaines prennent en compte l'interaction utilisateur pour raffiner le classement (Liu et al., 2013 ; Wang et al., 2016c), d'autres sont automatiques. Le principe du *re-ranking* est d'ajouter une étape d'optimisation du classement en se basant les classements relatifs (Li et al., 2012a), la superposition entre les plus proches voisins de la requête et les plus proches voisins des éléments de la galerie en calculant les scores de similarité et de dissimilarité (Ye et al., 2016 ; Zhong et al., 2017). D'autres techniques impliquent une fusion de plusieurs listes triées avec différentes métriques (Bai et al., 2019 ; Paisitkriangkrai et al., 2015).

Métriques utilisées en ré-identification dans les ensembles fermés

En base fermée, on évalue le bon classement (ou rangement) de la liste des identités de la galerie lorsqu'une identité requête est présentée. Les identités les plus similaires à la requête devraient apparaître en tête du classement. Parmi les métriques d'évaluation les plus utilisées, citons :

- la CMC (Cumulative Matching Characteristic) qui est la proportion des bonnes correspondances trouvées dans les r premiers rangs de la liste galerie triée ;
- la mAP (mean Average Precision) : les images ne sont pas groupées par identité, et la précision moyenne (AP) est évaluée pour chaque image requête q d'un ensemble d'images $\mathcal{Q}$. Cette précision moyenne est calculée comme étant $AP(q) = \frac{1}{R}\sum_r Precision_r(q)$ qui moyenne sur les R premiers rangs pertinents le score de précision $Precision_r(q)$ mesurant la proportion d'identités correctes au rang r. Le mAP est alors la moyenne de l'AP pour toutes les requêtes :

$$mAP = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} AP(q) \tag{2.1}$$

2.2.2 Ré-identification en base ouverte

L'hypothèse d'ensemble fermé n'est pas applicable dans la plupart des scénarios réels de ré-identification. En base ouverte, en plus des personnes de la galerie, des identités inconnues sont présentées en requête. Celles-ci ne doivent pas être ré-identifiées avec une identité existante, mais classée à part (rejetée ou ajoutée à la galerie). On peut également considérer une galerie dynamique qui, partant d'un ensemble vide se remplit avec de nouvelles personnes. La ré-identification en base ouverte est donc un problème beaucoup plus complexe qui n'a été que peu abordé dans la littérature en comparaison de la ré-identification en base fermée.

Vérification d'appartenance à un groupe

Il s'agit de déterminer si un individu fait partie d'un groupe de personnes ou non. Le groupe cible (*watch-list*) étant défini, il peut servir à la fois dans la phase d'apprentissage et de test, alors que l'ensemble non-cible diffère. Dans le scénario difficile décrit dans (Zheng et al., 2012, 2016b), une seule image par identité cible et plusieurs images pour les identités de l'ensemble non-cible. Des distances relatives entre éléments dans les ensembles cible et non-cible sont apprises. Les identités de l'ensemble non-cible similaires à celles de la *watch-list* servent à enrichir l'ensemble cible. Li et al., 2018b entraînent des réseaux de manière antagoniste pour distinguer les identités cibles et les imposteurs. Zhu et al., 2018 se posent la question du déploiement à grande échelle et du temps de calcul lorsque l'ensemble requête est très grand, et proposent d'apprendre des représentations binaires entre vues avec des méthodes de hachage pour vérifier l'appartenance de requêtes au groupe cible.

Détection et ré-identification

De façon plus complète que la vérification de groupe, on peut décomposer le problème de ré-identification en base ouverte en deux sous-tâches :

- la détection : cette tâche a pour but de déterminer si l'identité présentée en requête est connue ou inconnue
- la ré-identification : si l'identité requête fait partie de la galerie, elle doit être associée à la bonne identité.

Un nouvel ensemble de données et un benchmark associé, OpenRID (Liao et al., 2014), sont créés afin d'encourager les recherches sur la ré-identification en ensemble ouvert. Dans une configuration multi-caméras, chaque identité n'est observée que dans un sous-ensemble des caméras, et en test, l'ensemble des identités requête n'est pas contenue dans la galerie. De nouvelles métriques (DIR et FAR) sont introduites pour évaluer les performances des sous-tâches de détection et ré-identification effectuées simultanément. Ce benchmark a permis de constater que les approches existantes conçues pour la ré-identification en base fermée ne parviennent pas à traiter correctement ces nouveaux scénarios de détection et ré-identification. Wang et al., 2016d se situent dans un cadre one-shot et présentent une approche d'apprentissage dans un espace latent qui permet de rapprocher les images de caméras différentes. Bien qu'évaluée avec un protocole de ré-identification en base ouverte, leur méthode n'adresse pas explicitement les problèmes de détection et ré-identification simultanées.

Métriques utilisées en ré-identification dans les ensembles ouverts

La ré-identification dans les ensembles ouverts peut être vue comme un problème de classification binaire de paire d'images ou une inférence d'identité (Cancela et al., 2014 ; Karaman et Bagdanov, 2012). Une paire d'images est considérée positive si les deux images correspondent à la même personne, négative si les deux images sont associées à des personne différentes. Ainsi, les métriques de recherche d'information peuvent s'appliquer sur les paires. En notant TP pour *True*

Positive, TN pour *True Negative*, FP pour *False Positive*, FN pour *False Negative*, les métriques de rappel $recall = \frac{TP}{TP+FN}$, de précision $precision = \frac{TP}{TP+FP}$, de spécificité $specificity = \frac{TN}{TN+FP}$ qui est le taux de TN, et de F_1-*score* moyenne harmonique du rappel et de la précision, sont les mesures fréquemment employées.

Les métriques TTR (*True Target recognition Rate*) et FTR (*False Target recognition Rate*) ont été proposées pour évaluer la vérification d'identité dans un groupe de personnes (Zheng et al., 2012). Ces métriques estiment les taux d'identités appartenant ou n'appartenant pas à l'ensemble cible.

Lorsque la ré-identification est vue comme une tâche de détection et d'identification (Liao et al., 2014), d'autres métriques ont été introduites : DIR (*Detection and Identification Rate*) et FAR (*False Acceptance Rate*). On souhaite évaluer pour un ensemble P d'identités requête (*probe*) et un ensemble G d'identités dans la galerie les taux de détection et de fausse acceptation. P est composé de l'ensemble des personnes déjà présentes dans la galerie, noté $P \cap G$, et de l'ensemble des personnes inconnues (aussi appelées "imposteurs") n'appartenant pas à la galerie, noté $P \backslash G$. On souhaite classer les identités en fonction de la distance $dist(j^g, i^p)$ entre l'ensemble des images de la personne j dans la galerie et l'ensemble des images de la personne i dans l'ensemble requête, par rapport à un seuil de décision τ.

Le taux d'identification noté DIR est défini par :

$$DIR(\tau, r) = \frac{|\{i | i \in P \cap G, rank(i) \leqslant r, dist(i^g, i^p) \leqslant \tau\}|}{|P \cap G|} \tag{2.2}$$

$DIR(\tau, r)$ est la proportion d'identités communes à P et G qui sont ré-identifiées jusqu'au rang r avec une distance inférieure ou égale à τ.

Le taux de fausse acceptation noté FAR est donné par :

$$FAR(\tau) = \frac{|\{i | i \in P \backslash G, min_{j \in G} dist(j^g, i^p) \leqslant \tau\}|}{|P \backslash G|} \tag{2.3}$$

$FAR(\tau)$ représente la proportion d'identités "imposteurs" dont la distance aux éléments de la galerie les plus proches est inférieure ou égale à τ. Ces personnes sont ré-identifiées à tort avec des personnes de la galerie.

2.2.3 Positionnement de nos travaux

Les travaux qui sont présentés dans ce chapitre ont été réalisés dans le cadre de la thèse de Solène Chan-Lang (2014-2017) (Chan-Lang, 2017). Ces travaux se situent dans le temps, et par rapport à l'état de l'art, au tout début de l'engouement pour les réseaux de neurones profonds pour l'apprentissage de représentation et de métrique en ré-identification de personnes. Les méthodes concurrentes considérées sont donc celles qui emploient des descripteurs évolués faits à la main, comme LOMO, et l'apprentissage de métriques ne reposant pas sur des réseaux de neurones. Contrairement à de nombreux autres travaux de ré-identification en base fermée, nous cherchons à apporter des solutions au problème plus complexe et moins abordé de la ré-identification en base ouverte qui nécessite à la fois de distinguer des identités inconnues et de ré-identifier celles qui sont connues.

La première contribution, COPReV, tente de répondre aux problèmes de ré-identification en base fermée et en base ouverte, en formulant une tâche de vérification des identités. COPReV cherche à séparer la distribution des distances des paires positives et celle des distances des paires négatives de chaque côté d'un seuil fixé à l'apprentissage et utilisé au test. L'approche est évaluée dans des scénarios de base fermée et de base ouverte sur deux ensembles de données.

La seconde contribution regroupe un ensemble de méthodes de représentation parcimonieuse avec collaboration où les identités encodées dans le dictionnaire sont mises en compétition. Notre intuition est que ce type de représentations est bien adaptée aux tâches de détection (vérification) et de ré-identification, car elles permettent d'une part de sélectionner les éléments pertinents grâce la propriété de parcimonie, et d'autre part de ranger les identités par leur erreur de reconstruction. D'autre part, ces modèles peuvent prendre en compte le multi-shot de façon naturelle. Nous proposons différentes stratégies de représentation parcimonieuse : directe (requête en fonction de la galerie), inverse (galerie en fonction des requêtes) et bi-directionnelle. Enfin, nous comparons leurs performances dans des scénarios de ré-identification en base fermée et en base ouverte.

2.3 Ré-identification dans des ensembles fermés et ouverts

2.3.1 COPReV : *Closed and Open world Person RE-identification and Verification*

Dans les scénarios d'ensembles d'identités ouverts, une contrainte est ajoutée au problème de ré-identification en base fermée : lorsque qu'une identité requête est présentée, il faut déterminer s'il s'agit d'une personne connue, c'est-à-dire faisant partie de la galerie courante, ou d'une nouvelle personne. Cette contrainte se traduit par une prise de décision appelée vérification qui est une classification binaire de chaque paire d'images : paire labellisée positive pour une même identité, paire labellisée négative pour deux identités différentes. En classant toutes les paires possibles formées par l'image requête et les images représentatives des identités de la galerie courante, si aucune correspondance positive n'est trouvée, l'image requête est considérée comme représentant une identité inconnue.

L'approche que nous avons proposée, nommée COPReV pour *Closed and Open world Person RE-identification and Verification*, pour résoudre le problème de ré-identification dans des ensembles fermés ou ouverts d'identités repose sur une formulation de vérification d'identités. Ces travaux ont fait l'objet d'une publication (Chan-Lang et al., 2017).

Nous partons du principe qu'une méthode de classement (ranking) relatif, fondée sur le calcul de scores de dissimilarité entre les descripteurs des paires d'images, peut très bien produire des résultats corrects selon une métrique de classement telle la CMC, sans pour autant respecter le fait que les scores de dissimilarité entre les descripteurs des paires positives soient toujours inférieures aux scores de dissimilarité entre descripteurs de paires négatives, du fait de leur variabilité. D'autre part, il est important de pouvoir fixer un seuil sur les scores de dissimilarité pour discriminer les paires positives et les paires négatives. Les descripteurs d'image sont supposés fixés ; ce sont soit des descripteurs faits à la main, soit des descripteurs issus d'un apprentissage de représentation. Ces descripteurs n'étant pas été optimisés pour respecter une contrainte de vérification sur l'ensemble de la base d'images, nous recherchons une méthode de projection des descripteurs de manière à ranger correctement les paires positives en dessous du seuil de décision, et les paires négatives au-dessus de ce seuil.

Soit x_{il} le vecteur descripteur de l'image l pour la personne $i \in \mathcal{I}$, $\mathcal{I}$ étant l'ensemble des identités, de taille K. Les différences entre descripteurs des paires positives correspondant à l'identité i sont notés $D_{ii} = \{x_{il} - x_{il'}\}_{l,l' \in [1,n_i], l<l'}$ et n_i le nombre d'images de la personne i. L'ensemble de ces différences pour toutes les identités i est de taille m_{ii}.

De la même façon, les différences entre descripteurs des paires négatives formées par des images correspondant à deux personnes différentes i et j sont $D_{ij} = \{x_{il} - x_{jl'}\}_{i \neq j, l \in [1,n_i], l' \in [1,n_j]}$. L'ensemble des différences pour les paires négatives est de taille m_{ij}.

En notant $\tau \in \mathbb{R}$ le seuil arbitrairement fixé, d la dimension de l'espace des caractéristiques de départ, on recherche $L \in \mathbb{R}^{d' \times d}$ la matrice de transformation linéaire qui projette les caractéristiques y dans un espace de dimension d' telle la fonction de coût suivante soit minimisée :

$$E(L) = \sum_{i \in \mathcal{I}} \left[\frac{1}{m_{ii}} \sum_{y \in D_{ii}} \mathcal{L}_+ \left(\|Ly\|_2^2 - \tau \right) + \frac{1}{K-1} \sum_{j \in \mathcal{I} \setminus i} \left(\frac{1}{m_{ij}} \sum_{y \in D_{ij}} \mathcal{L}_- \left(\tau - \|Ly\|_2^2 \right) \right) \right] \quad (2.4)$$

où $\mathcal{L}_+$ et $\mathcal{L}_-$ sont les fonctions de perte associées aux paires positives et négatives respectivement. En prenant des fonctions de Heaviside $H(x)$, ces fonctions de perte s'écrivent $\mathcal{L}_+(z) = H(z)$ et $\mathcal{L}_-(z) = H(-z)$. Calculer la somme sur les paires positives revient à effectuer un comptage des paires mal classées, c'est-à-dire en dessous du seuil τ, et de la même façon des paires négatives qui doivent se situer au dessus de τ. Le facteur $\frac{1}{K-1}$ sert à équilibrer les deux termes, car le nombre de paires négatives est en pratique largement supérieur à celui des paires positives. L'intuition est qu'une fonction de coût fondée sur le comptage des paires mal classées est moins sensible au bruit qu'une fonction prenant en compte directement les scores de dissimilarité, à cause de leur variabilité.

Nous utilisons des fonctions logistiques généralisées en forme de S pour approcher la fonction de Heaviside non différentiable. Les fonctions choisies sont telles qu'elles doivent produire des valeurs entre 0 et 1. Pour obtenir une meilleure robustesse aux valeurs aberrantes, on recherche des fonctions, telle que pour les paires éloignées de la frontière de décision, la fonction de coût aura des valeurs proches de 0 pour les paires bien classées et des valeurs proches de 1 pour les paires mal classées, alors que la variation sera plus grande autour du point d'inflexion. Ceci a pour effet d'influencer surtout les paires qui sont proches du seuil et de les encourager à se déplacer du bon côté de la frontière, les paires déjà très bien classées ou complètement mal classées ne devant pas avoir beaucoup d'influence. Enfin, le point d'inflexion est fixé en 0. Les fonctions vérifiant toutes ces contraintes sont de la forme :

$$\begin{cases} \lambda > 0 \\ S(z) = \frac{1}{(1+\nu e^{-\lambda z})^{\frac{1}{\nu}}} \end{cases} \quad (2.5) \qquad \text{or} \qquad \begin{cases} \lambda < 0 \\ S(z) = 1 - \frac{1}{(1+\nu e^{-\lambda z})^{\frac{1}{\nu}}} \end{cases} \quad (2.6)$$

Une valeur de ν égale à 1 conduit à une sigmoïde symétrique. Pour pénaliser plus fortement des paires classées, on choisit $\nu > 1$ pour le cas $\lambda > 0$ et $\nu < 1$ pour le cas $\lambda < 0$. La figure 2.5 donne quelques exemples de fonctions généralisées ayant les caractéristiques souhaitées.

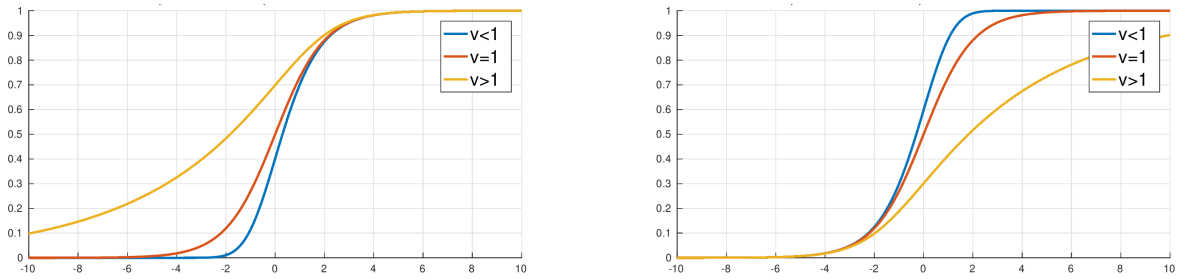


FIGURE 2.5 – Exemples de fonctions logistiques généralisées d'équation 2.6. A gauche : avec $\lambda = 1 > 0$, à droite, avec $\lambda = -1 < 0$

L'optimisation consiste à minimiser par descente de gradient la fonction dérivée $\frac{\partial E}{\partial L}(L)$.

2.3.2 Evaluation de COPReV

La méthode COPReV est évaluée sur les ensembles de données PRID2011 (Hirzer et al., 2011) et iLIDS-VID (Wang et al., 2014b). Pour la ré-identification en ensemble fermé, le protocole de test suivi est celui décrit dans (ibid.), et pour la ré-identification en ensemble ouvert, nous adoptons celui proposé par Ma et al., 2017 qui consiste à réduire la galerie de test pour disposer d'identités requête inconnues. Comme plusieurs images de la même personne sont disponibles, le protocole *multi-shot* est suivi : la distance entre deux identités est la moyenne des distances entre toutes les images de l'identité requête et celles de l'identité galerie. Tous les détails expérimentaux sont donnés dans (Chan-Lang, 2017), seuls les résultats principaux sont rappelés ici.

Les caractéristiques d'entrée sont :

— IR : caractéristiques produites par un réseau de neurones de type Inception-Resnet-v2 entraîné en classification sur ImageNet (Szegedy et al., 2017),

— LOMO (*Local Maximal Occurrence*) : descripteurs faits à la main de l'état de l'art (Liao et al., 2015).

Notre méthode est comparée à l'approche d'apprentissage de métrique XQDA (ibid.), une approche de référence en ré-identification dans les ensembles fermés au moment de nos travaux.

Dataset	PRID2011				iLIDS-VID			
Rank	1	5	10	20	1	5	10	20
MDTS-DTW (Ma et al., 2017)	69.6	89.4	94.3	97.9	49.5	75.7	84.5	91.9
DVR (Wang et al., 2014b)	77.4	93.9	97.0	99.4	51.1	75.7	83.9	90.5
XQDA+IR	41.2	68.9	79.9	90.2	11.1	29.0	39.6	51.5
COPReV+IR	53.0	80.8	91.5	98.1	21.9	51.2	66.9	81.3
XQDA+LOMO(Liao et al., 2015)	**86.4**	**98.3**	**99.6**	**100.0**	**55.9**	**83.4**	90.5	96.1
COPReV+LOMO	82.8	97.8	**99.6**	**100.0**	53.9	**83.4**	**91.6**	**97.9**

Tableau 2.1 – Evaluation de la ré-identification en ensemble fermé. Les valeurs de CMC sont données aux rangs 1, 5, 10, 20 pour les ensembles de données PRID2011 et iLIDS-VID.

Dataset	PRID2011				iLIDS-VID			
FAR(%)	1	10	50	100	1	10	50	100
MDTS-DTW (Ma et al., 2017)	42.7	55.2	70.5	72.8	12.7	**32.6**	**51.8**	57.3
DVR (Wang et al., 2014b)	**46.8**	**58.3**	78.3	79.7	**17.3**	29.1	49.9	57.8
XQDA+IR	3.0	8.7	24.7	47.7	0.6	2.0	8.6	13.7
COPReV+IR	8.3	15.8	40.0	60.5	1.2	5.7	17.4	25.8
XQDA+LOMO (Liao et al., 2015)	21.0	40.5	80.3	**90.3**	5.6	15.4	45.8	**59.9**
COPReV+LOMO	26.5	43.5	**81.0**	87.5	3.9	21.0	47.9	59.1

Tableau 2.2 – Evaluation de la ré-identification en ensemble ouvert. Les valeurs de DIR au rang 1 sont données pour plusieurs valeurs de FAR (1%, 10%,50% and 100%) pour les ensembles de données PRID2011 et iLIDS-VID.

Les résultats obtenus en ré-identification en base fermée sont synthétisés dans le Tableau 2.1 avec la métrique CMC. On constate que sur les deux ensembles PRID2011 et iLIDS-VID, notre méthode de projection COPReV est compétitive avec XQDA (+10%, nettement meilleure avec les caractéristiques IR, un peu moins performante sur les caractéristiques LOMO aux premiers rangs). Les performances varient beaucoup suivant les caractéristiques initiales utilisées.

Le Tableau 2.2 rassemble les résultats en ré-identification en base ouverte, obtenus avec la métrique DIR à différents niveaux de FAR. Sur les deux ensembles de données, COPReV fait généralement mieux que XQDA quelles que soient les caractéristiques d'entrée, mais ses performances restent bien en-deça de l'état de l'art que sont les approches utilisant des caractéristiques spatio-temporelles (Ma et al., 2017 ; Wang et al., 2014b). Une fois de plus, les caractéristiques initiales jouent un grand rôle dans les performances finales.

Nous avons également évalué la capacité de COPReV à réaliser la tâche de vérification avec les métriques de rappel et de spécificité calculés sur le classification des paires. Sur les deux ensembles de données, le rappel est entre 94% et 99%, alors que la spécificité se situe entre 88% et 91% pour les caractéristiques LOMO. Les résultats sont moins bons avec IR (rappel entre 58% et 89% et spécificité entre 88% et 93%). Une analyse des distances moyennes pour les paires positives et paires négatives montre que COPReV appliqué à LOMO parvient à écarter davantage les paires du seuil fixé que COPReV appliqué sur IR.

2.3.3 Limitations de la méthode

COPReV aborde le problème de ré-identification dans un ensemble fermé ou ouvert, en introduisant une contrainte de vérification, c'est-à-dire de classification binaire des paires d'identités, alors que la plupart des méthodes de ré-identification visent plutôt à optimiser le classement des identités et ne s'évaluent en général que dans un contexte de base fermée. Cependant, les résultats des expérimentations effectuées sur les ensembles PRID2011 et iLIDS-VID sont dans l'ensemble mitigés. On peut néanmoins en tirer quelques enseignements :

— COPReV apporte une première réponse au problème pratique de ré-identification en base ouverte, puisque la méthode est capable de classer correctement les paires d'identités dans un problème de vérification ;
— COPReV est compétitive avec les méthodes concurrentes en base fermée. En base ouverte, les méthodes implémentant des caractéristiques spatio-temporelles sont nettement supérieures ;
— Les performances dépendent beaucoup des caractéristiques initiales. Les écarts de performances sont significatifs entre des caractéristiques génériques et des caractéristiques plus spécifiques et plus discriminantes.

2.4 Représentations parcimonieuses pour la ré-identification

Le codage parcimonieux est un outil puissant permettant de représenter un élément en fonction d'un ensemble très réduit d'autres éléments et en pondérant leur contribution. Ce type de représentation présente plusieurs avantages pour le problème de ré-identification. D'abord, il permet de représenter des images requête en fonction de l'ensemble des images de la galerie qui constituent le dictionnaire. Ainsi par l'aspect collaboratif, toutes les identités de la galerie sont mises en compétition, ce qui devrait encourager un classement correct. La contrainte de parcimonie introduit un choix sur les éléments les plus représentatifs et devrait faciliter la tâche de vérification nécessaire en ré-identification dans un ensemble d'identités ouvert. D'autre part, lorsque plusieurs images sont disponibles pour chaque personne, les images les plus pertinentes sont sélectionnées en fonction de la ressemblance visuelle aux images de référence. Par exemple, si on dispose d'images de personnes avec des poses différentes, il sera plus facile de comparer les personnes dans des poses semblables (de face, de profil, de dos). Les représentations parcimonieuses paraissent donc particulièrement adaptées au cas du *multi-shot*.

2.4.1 Représentations parcimonieuses collaboratives

On note P_l la matrice des caractéristiques visuelles de l'identité requête l et G_k la matrice des caractéristiques visuelles l'identité k dans la galerie. L est le nombre d'identités requête, K le nombre d'identités dans la galerie, m_l le nombre d'images de l'identité requête l et m_k le nombre d'images de l'identité galerie k.

Représentation parcimonieuse sans collaboration

Dans le cas d'un codage parcimonieux sans collaboration (noté DNC pour *Direct Non Collaborative*), la représentation de P_l est exprimée en fonction des descripteurs de chaque identité k est résolvant un problème d'optimisation de Lasso. On cherche la représentation A_{P_l,G_k} de P_l telle que :

$$A_{P_l,G_k} = \arg\min_A ||P_l - G_k A||_F^2 + \lambda ||A||_1 \tag{2.7}$$

où $||.||_F$ est la norme de Frobenius, $||.||_1$ est la norme $\mathcal{L}_1$ et λ un coefficient entre 0 et 1 réglant l'équilibre entre l'erreur de reconstruction et le terme de pénalisation de la parcimonie. Les éléments de galerie sont ici considérés de façon indépendante.

Le score de dissimilarité entre l'identité requête l et l'identité k de la galerie est l'erreur résiduelle moyenne de reconstruction :

$$s(l,k) = E_{P_l,G_k} = \frac{||P_l - G_k A_{P_l,G_k}||_F^2}{m_l} \tag{2.8}$$

avec m_l le nombre d'images de l'identité requête. La ré-identification est effectuée en classant les erreurs de reconstruction par ordre croissant.

Représentation parcimonieuse avec collaboration

Dans ce type de représentation, notée DC pour *Direct Collaborative*, toutes les identités de la galerie interviennent. Elles sont compilées en concaténant tous les vecteurs de caractéristiques dans une matrice $G = [G_1, \ldots, G_K]$. Le problème d'optimisation de Lasso s'écrit alors :

$$A_{P_l,G} = \arg\min_A ||P_l - GA||_F^2 + \lambda ||A||_1 \tag{2.9}$$

La représentation parcimonieuse s'écrit sous la forme :

$$A_{P_l,G} = \begin{bmatrix} A_{P_l,G,G_1} \\ A_{P_l,G,G_2} \\ \vdots \\ A_{P_l,G,G_K} \end{bmatrix} \tag{2.10}$$

où A_{P_l,G,G_k} est la sous-matrice éparse de taille $n_k \times m_l$ de $A_{P_l,G}$ qui représente la contribution des éléments G_k du dictionnaire.

Le score de dissimilarité entre l'identité requête l et l'identité k de la galerie est dans ce cas l'erreur résiduelle moyenne de reconstruction en n'utilisant que les éléments de k :

$$s(l,k) = R_{P_l,G,G_k} = \frac{||P_l - G_k A_{P_l,G,G_k}||_F^2}{m_l} \tag{2.11}$$

La contrainte de parcimonie dans une approche avec collaboration se traduit par le fait qu'un petit nombre d'identités de la galerie contribuent à la reconstruction de l'identité requête. Il est

attendu que ce sont celles qui ressemblent le plus à la requête. La ré-identification est effectuée en rangeant les personnes candidates selon leur erreur résiduelle moyenne.

En base fermée, l'identité de la requête fait aussi partie de la galerie. Si le classement des personnes de la galerie est correct, le premier élément trouvé correspondra bien à l'identité de la requête. En revanche, en base ouverte, un imposteur sera ré-identifié à tort avec l'identité de la galerie qui aura la plus petite erreur de reconstruction.

Représentation parcimonieuse avec collaboration élargie

Pour adresser le problème de ré-identification en base ouverte, nous introduisons un codage parcimonieux avec collaboration élargie (DCE pour *Direct Collaborative Enhanced)* où nous modifions le problème d'optimisation de Lasso en ajoutant un dictionnaire D qui a un rôle d'attraction pour un imposteur. En effet, l'effet recherché est que l'imposteur soit plus facilement représenté par des éléments de D que par les personnes de la galerie.

Le problème d'optimisation du Lasso s'écrit maintenant :

$$A_{P_l,[G,D]} = \arg\min_A ||P_l - [G,D]A||_F^2 + \lambda||A||_1 \tag{2.12}$$

et le score de dissimilarité est :

$$s(l,k) = R_{P_l,[G,D],G_k} = \frac{||P_l - G_k A_{P_l,[G,D],G_k}||_F^2}{m_l} \tag{2.13}$$

Le choix du dictionnaire D est déterminant puisqu'il va conditionner la capacité de la méthode à écarter les identités imposteurs sans nuire aux performances de ré-identification. D est construit sur l'ensemble d'apprentissage. Dans une configuration à deux caméras, une pour la requête et l'autre pour la galerie, et pour s'affranchir des problèmes d'écart de domaine, les identités de D sont prises dans la caméra de la galerie. Pour avoir le plus de variété possible, et sans aucune information sur les imposteurs éventuels, on choisit D comme étant l'ensemble de la galerie dans l'ensemble d'apprentissage. L'ensemble du processus est illustré dans la Figure 2.6.

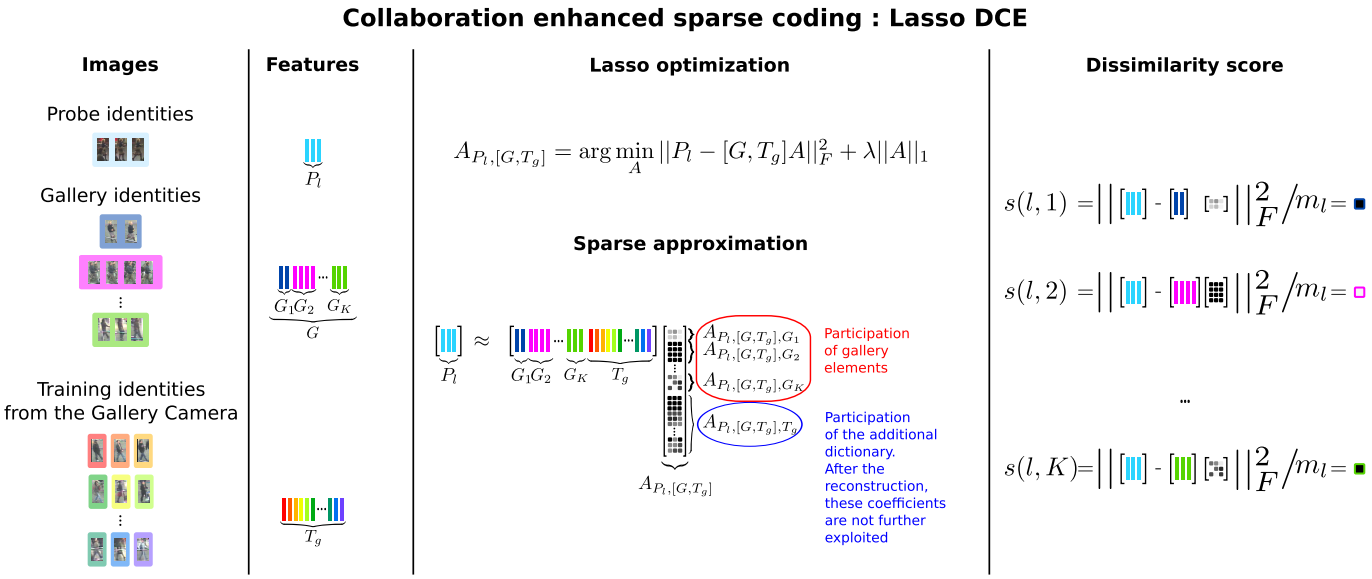


FIGURE 2.6 – Représentation parcimonieuse collaborative élargie (Lasso DCE). Les identités sont représentées par des couleurs différentes, les contributions des éléments du dictionnaire par des petits carrés gris : plus ils sont clairs plus la contribution est importante, plus ils sont sombres, plus les valeurs sont proches de 0.

Evaluation expérimentale

Les méthodes de codage parcimonieux DNC, DC, et DCE sont évaluées sur les ensembles de données iLIDS-VID et PRID2011. Les résultats principaux sont présentés dans cette section. Les caractéristiques utilisées sont les descripteurs LOMO projetés dans un espace de dimension réduite en utilisant la même décomposition que dans (Liao et al., 2015). Les caractéristiques résultantes normalisées avec la norme $\mathcal{L}2$ sont notées LOMO$_{pn}$.

Le Tableau 2.3 compare en base fermée les performances obtenues par les différentes représentations parcimonieuses et les approches de l'état de l'art. L'apport des approches collaboratives est net, le taux de ré-identification au rang est augmenté de +4% et +10% par rapport à XQDA. L'écart est encore plus grand avec les autres méthodes de référence. DC et DCE ont des performances semblables : en base fermée, la collaboration élargie n'a que peu d'impact puisque l'ajout d'éléments supplémentaires ne modifie pas le classement.

Dataset	iLIDS-VID				PRID2011			
Rank	1	5	10	20	1	5	10	20
MDTS-DTW (Ma et al., 2017)	49.5	75.7	84.5	91.9	69.6	89.4	94.3	97.9
DVR (Wang et al., 2014b)	51.1	75.7	83.9	90.5	77.4	93.9	97.0	99.4
LOMO + XQDA (Liao et al., 2015)	55.3	83.1	90.3	**96.3**	86.3	**98.3**	**99.6**	**100.0**
LOMO$_{pn}$ + Lasso DNC	56.1	81.9	88.5	94.5	87.3	98.2	**99.6**	**100.0**
LOMO$_{pn}$ + Lasso DC	64.9	**87.1**	**92.5**	96.1	90.2	98.0	99.3	**100.0**
LOMO$_{pn}$ + Lasso DCE	**65.1**	86.6	92.4	96.1	**90.6**	97.9	99.2	**100.0**

Tableau 2.3 – Evaluation des approches fondées sur le codage parcimonieux en base fermée sur iLIDS-VID et PRID2011. Les valeurs de CMC sont données aux rangs $1, 5, 10$ et 20.

En base ouverte, l'approche avec collaboration élargie DCE dépasse l'approche DC, qui est elle-même supérieure à DNC (Tableau 2.4). Ceci confirme la pertinence de la collaboration des éléments de la galerie dans la modélisation. La méthode DC atteint un gain de performance de +28% du DIR à FAR=1% par rapport à XQDA sur PRID2011. L'ajout d'un dictionnaire générique dans DCE pour rejeter les imposteurs améliore également bien les résultats, et permet de gagner encore +5% de DIR à FAR=1%. Les performances sont comparables ou légèrement meilleures que les approches utilisant des descripteurs spatio-temporels (Ma et al., 2017 ; Wang et al., 2014b).

Dataset	iLIDS-VID				PRID2011			
FAR(%)	1	10	50	100	1	10	50	100
MDTS-DTW (Ma et al., 2017)	12.7	32.6	51.8	57.3	42.7	55.2	70.5	72.8
DVR (Wang et al., 2014b)	**17.3**	29.1	49.9	57.8	46.8	58.3	78.3	79.7
LOMO + XQDA(Liao et al., 2015)	5.1	15.2	45.3	59.1	21.2	40.7	78.8	90.5
LOMO$_{pn}$ + Lasso DNC	4.5	16.6	45.7	61.5	21.2	39.2	74.5	90.2
LOMO$_{pn}$ + Lasso DC	12.9	35.1	58.8	68.5	49.8	69.3	88.2	**93.8**
LOMO$_{pn}$ + Lasso DCE	17.2	**37.5**	**62.8**	**69.0**	**55.7**	**71.0**	**90.2**	93.2

Tableau 2.4 – Evaluation des approches fondées sur le codage parcimonieux en base ouverte sur iLIDS-VID et PRID2011. La valeur du DIR au premier rang et donnée pour plusieurs valeurs de FAR ($1\%, 10\%, 50\%$ et 100%).

Des expériences complémentaires décrites dans (Chan-Lang, 2017) indiquent que le choix du dictionnaire D dans la galerie de l'ensemble d'apprentissage ou dans l'ensemble des identités

requête de l'ensemble d'apprentissage ne semble pas beaucoup influencer la performance de la méthode Lasso DCE. Afin de valider l'apport des méthodes proposées pour la tâche de vérification, nous avons étudié la variation du rappel et de la spécificité selon le seuil de décision. Les approches avec collaboration présentent une variation plus rapide entre 0 et 1, DCE a la variation la plus rapide. Pour des taux de faux positifs faibles, les seuils de décision s'approchent de 1, ce qui paraît logique car les erreurs de reconstruction pour les éléments qui n'interviennent pas dans la représentation parcimonieuse sont proches de 1.

2.4.2 Représentations parcimonieuses bi-directionnelles

Les approches directes de représentation parcimonieuse avec collaboration codent une personne requête avec le dictionnaire de la galerie, ce qui permet à la fois de ranger les personnes selon leur ressemblance à la requête et de rejeter les imposteurs. Nous proposons de renforcer davantage la robustesse de ces approches en rendant la relation réciproque (Chan-Lang et al., 2016). Une représentation parcimonieuse inverse exprime chaque personne de la galerie à l'aide d'un dictionnaire des personnes requête. De la même façon que dans le sens direct, on peut calculer dans le sens inverse un score de dissimilarité entre une identité de la galerie et une identité requête en utilisant la représentation parcimonieuse de l'identité galerie. L'approche bi-directionnelle considère le score de dissimilarité final comme étant la somme des scores de dissimilarité dans le sens direct et dans le sens inverse.

Même si la représentation parcimonieuse inverse d'un élément de la galerie avec les éléments requête peut fournir un classement des identités requête, l'objectif final reste le classement des identités de la galerie en fonction de leur similarité à la personne requête.

En pratique, toutes les personnes requête ne sont pas toujours disponibles au même moment. L désigne le nombre d'identités requête à ré-identifier, K le nombre de personnes dans la galerie. Lorsque les identités requête sont présentées une par une, la méthode est nommée "inverse 1 personne" ($L = 1$) ; quand les toutes les personnes requête sont présentées simultanément, la méthode est appelée "inverse toutes personnes" ($L > 1$). Les méthodes inverses sont dénommées RCE pour *Reverse Collaboration Enhanced*.

$P = [P_{l_1}, \ldots, P_{l_L}]$ est la concaténation des caractéristiques des images requête.

La représentation parcimonieuse inverse obtenue par une optimisation de Lasso pour une identité k de la galerie s'écrit :

$$A_{G_k,[P,D_k]} = \arg\min_A ||G_k - [P, D_k]A||_F^2 + \lambda||A||_1 \tag{2.14}$$

où D_k est un dictionnaire additionnel qui est choisi pour chaque identité de la galerie.

Le score de dissimilarité entre l'identité requête l et l'identité galerie k est dans ce cas égal à :

$$s(l,k) = R_{G_k,[P,D_k],P_l} = \frac{||G_k - P_l A_{G_k,[P,D_k],P_l}||_F^2}{n_k} \tag{2.15}$$

Comme pour la méthode Lasso DCE, les dictionnaires additionnels sont pris dans l'ensemble d'apprentissage, mais cette fois-ci dans la caméra des requêtes.

Les Figures 2.7 et 2.8 présentent sous forme schématique les processus de représentation parcimonieuse inverse RCE pour les deux cas "1 personne" et "toutes personnes" :

— Dans le cas d'une seule identité requête, un dictionnaire est formé avec les caractéristiques de l'identité requête et de dictionnaires additionnels qui servent à apporter de la variété. Pour chaque nouvelle identité requête, le dictionnaire est différent, mais l'important est qu'il soit unique afin de ranger les éléments de la galerie par rapport à identité requête donnée. Ainsi, dans le sens direct, on range les éléments de la galerie et les mettant en

compétition dans la représentation parcimonieuse de la personne requête. Dans le sens inverse, on range les éléments de la galerie par rapport à l'importance de la contribution de la personne requête dans les reconstructions des éléments de la galerie.

— Dans le cas d'une requête avec toutes les personnes présentées simultanément, la représentation inverse met en compétition toutes les personnes de la requête. En reconstruisant chaque identité de la galerie par rapport à une identité requête requête, on peut ranger les personnes de la galerie. Cette double comparaison est particulièrement intéressante pour renforcer le classement, la seule contrainte étant de disposer de toutes les personnes requête en même temps.

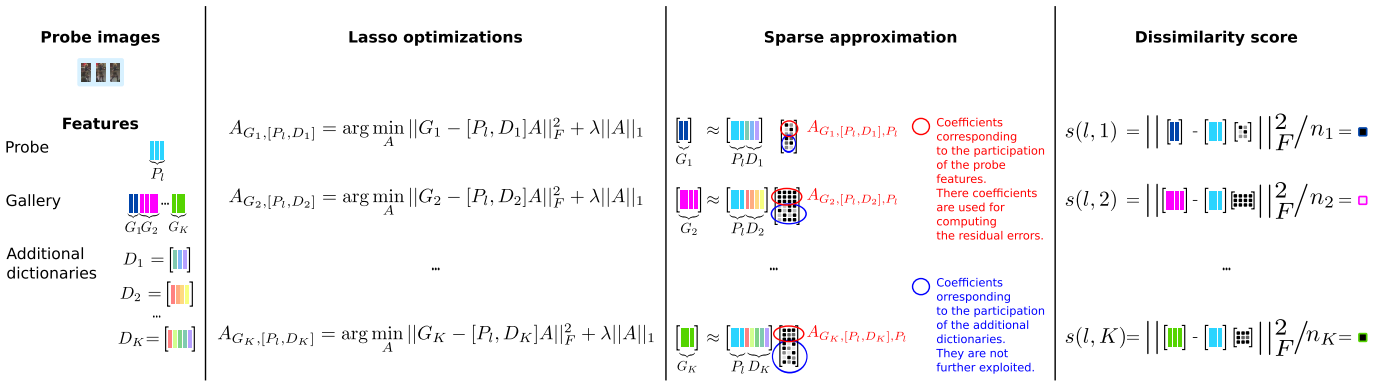


FIGURE 2.7 – Représentation parcimonieuse inverse (Lasso RCE) "1 personne". Une seule personne est présentée à la fois dans la requête de ré-identification. Les identités sont représentées par des couleurs différentes, les contributions des éléments du dictionnaire par des petits carrés gris : plus ils sont clairs plus la contribution est importante, plus ils sont sombres, plus les valeurs sont proches de 0. Dans le sens inverse, on peut ranger les personnes de la galerie en comparant les scores de dissimilarité avec l'identité requête, c'est-à-dire les contributions de l'identité requête dans les reconstructions des éléments de la galerie.

La méthode inverse a une complexité calculatoire plus élevée que la méthode directe car chaque élément de la galerie a sa propre représentation parcimonieuse. Pour garder un temps de calcul acceptable, la taille des dictionnaires doit rester contenue. Plusieurs stratégies sont décrites dans (CHAN-LANG, 2017).

Evaluation expérimentale

Les approches de représentation parcimonieuse inverse et bi-directionnelle sont testées en base fermée sur les ensembles iLIDS-VID et PRID2011. Les résultats donnés dans le Tableau 2.5 montrent que :

— les méthodes inverses (RCE) sont plus performantes que la méthode directe avec collaboration élargie (DCE). Lorsque toutes les personnes requêtes sont présentées simultanément, RCE dépasse DCE d'environ 4% de CMC au premier rang sur les deux datasets ;

— l'approche bi-directionnelle "1 personne" est plus performante que l'approche inverse "1 personne" (+4% de CMC au premier rang) ;

— l'approche bi-directionnelle "toutes personnes" a des performances similaires à l'approche inverse "toutes personnes".

— les meilleures approches (RCE "toutes personnes" ou bi-directionnelle "toutes personnes") dépassent XQDA (LIAO et al., 2015) d'environ 14% de CMC au premier rang sur iLIDS-VID et de 12% au premier rang sur PRID2011.

L'évaluation en base ouverte sur les deux mêmes ensembles de données fait apparaître (Tableau 2.6) que :

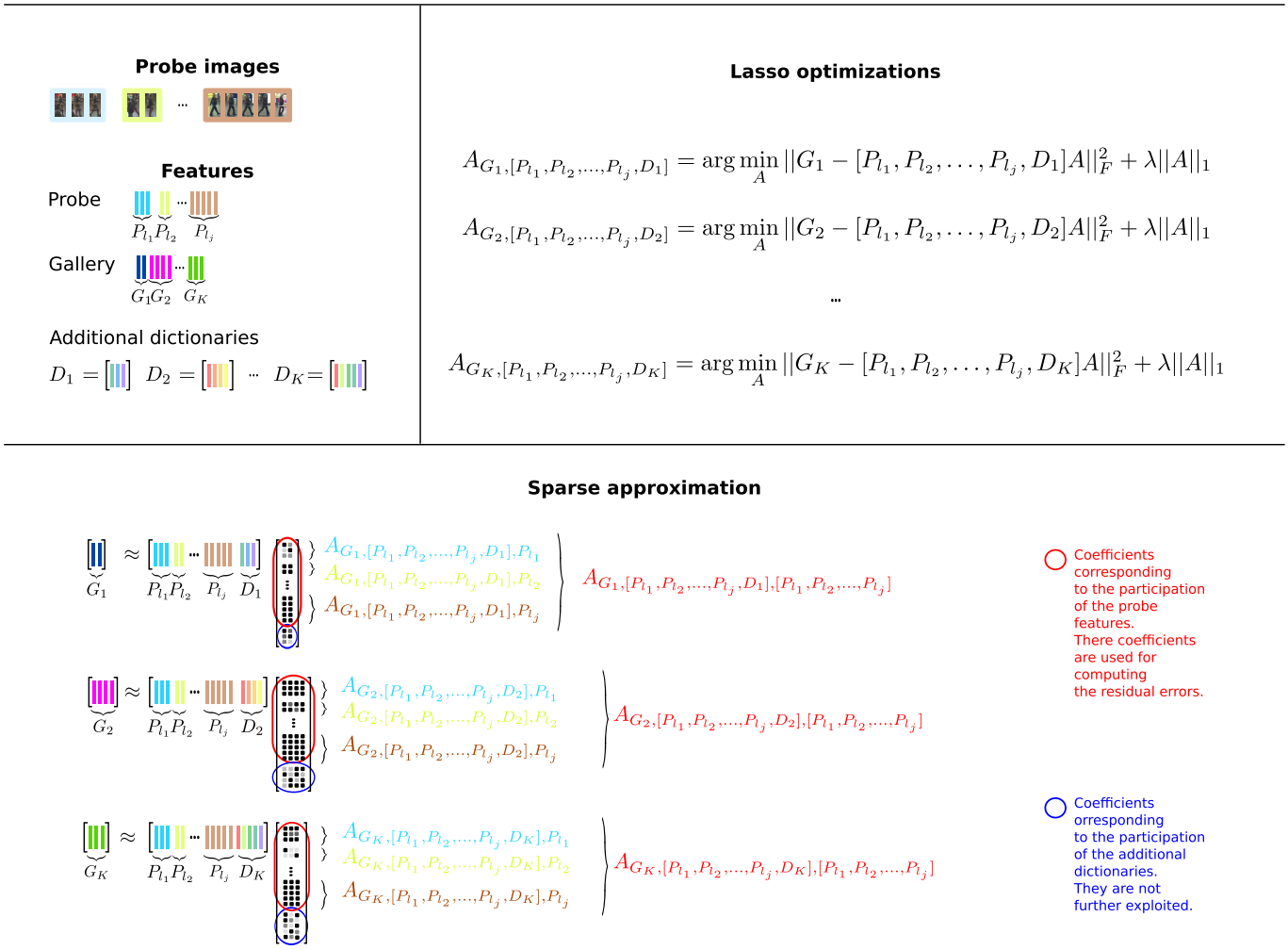


FIGURE 2.8 – Représentation parcimonieuse inverse (Lasso RCE) “toutes personnes”. Plusieurs personnes sont présentées simultanément dans la requête. Les identités sont représentées par des couleurs différentes, les contributions des éléments du dictionnaire à la représentation par des petits carrés : plus clairs, la contribution est importante, plus sombres, les contributions sont proches de 0. On considère et on range les éléments de la galerie pour lesquels l'erreur de reconstruction est faible, quand aucun élément n'est bien reconstruit par la requête, celle-ci est rejetée.

— les méthodes inverses sont plus performantes que la méthode directe avec collaboration élargie présentée précédemment (RCE "toutes personnes" dépasse DCE d'environ 5% en terme de DIR pour FAR=1% sur iLIDS-VID et de 17% de DIR pour la même valeur de FAR sur PRID2011) ;
— les approches bi-directionnelles augmentent encore les performances de ré-identification par rapport aux approches inverses ;
— par rapport à l'état de l'art (Wang et al., 2014b), notre approche DCE+RCE "toutes personnes" a un DIR supérieur de 6% pour FAR=1% sur iLIDS-VID et de 26% de DIR pour FAR=1% sur PRID2011.

Dataset	iLIDS-VID				PRID2011			
Rank	1	5	10	20	1	5	10	20
MDTS-DTW (Ma et al., 2017)	49.5	75.7	84.5	91.9	69.6	89.4	94.3	97.9
DVR (Wang et al., 2014b)	51.1	75.7	83.9	90.5	77.4	93.9	97.0	99.4
XQDA(Liao et al., 2015)	55.3	83.1	90.3	96.3	86.3	98.3	99.6	**100.0**
Lasso DC	64.9	87.1	92.5	96.1	90.2	98.0	99.3	**100.0**
Lasso DCE	65.1	86.6	92.4	96.1	90.6	97.9	99.2	**100.0**
Lasso RCE 1 pers.	65.4	88.3	93.9	96.8	89.8	98.2	99.2	**100.0**
Lasso RCE toutes pers.	**69.9**	**89.8**	**94.2**	**96.9**	**94.2**	98.5	99.2	**100.0**
Lasso DCE+RCE 1 pers.	68.1	88.9	93.7	96.7	91.2	98.4	99.4	**100.0**
Lasso DCE+RCE toutes pers.	69.8	89.6	93.5	96.8	93.8	**98.3**	**99.6**	**100.0**

Tableau 2.5 – Evaluation en base fermée des méthodes avec collaboration élargie inverses (Lasso RCE) et bi-directionnelles (Lasso DCE+RCE) sur iLIDS-VID et PRID2011. Les valeurs de CMC sont fournies pour les rangs 1, 5, 10 et 20.

Dataset	iLIDS-VID				PRID2011			
FAR(%)	1	10	50	100	1	10	50	100
MDTS-DTW (Ma et al., 2017)	12.7	32.6	51.8	57.3	42.7	55.2	70.5	72.8
DVR (Wang et al., 2014b)	17.3	29.1	49.9	57.8	46.8	58.3	78.3	79.7
XQDA(Liao et al., 2015)	5.1	15.2	45.3	59.1	21.2	40.7	78.8	90.5
Lasso DC	12.9	35.1	58.8	68.5	49.8	69.3	88.2	93.8
Lasso DCE	17.2	37.5	62.8	69.0	55.7	71.0	90.2	93.2
Lasso RCE 1 pers.	13.9	35.6	61.5	69.2	58.0	69.7	91.0	93.0
Lasso RCE toutes pers.	22.0	44.7	67.4	72.7	73.2	**87.3**	**95.3**	**96.0**
Lasso DCE+RCE 1 pers.	18.0	40.5	66.5	71.6	60.8	76.0	91.7	93.3
Lasso DCE+RCE toutes pers.	**23.8**	**45.8**	**68.1**	**72.9**	**73.3**	83.2	94.0	95.2

Tableau 2.6 – Evaluation des méthodes avec collaboration élargie inverses (Lasso RCE) et bidirectionnelles (Lasso DCE+RCE) en base ouverte sur iLIDS-VID et PRID2011. Les valeurs de DIR au rang 1 sont fournies pour les valeurs de FAR : 1%, 10%,50% et 100%.

2.5 Conclusion et perspectives

Dans ce chapitre, nous avons abordé le problème de la ré-identification visuelle des personnes et avons présenté deux contributions principales.

La première contribution porte sur le formalisme COPReV qui aborde sous l'angle de la

vérification les problèmes de ré-identification dans des scénarios d'ensembles fermés et ouverts d'identités. Dans l'approche COPReV, on cherche à apprendre une transformation linéaire des caractéristiques initiales des images telle que les paires d'images positives (même identité) et négatives (identité différente) soient bien classées par rapport à un seuil sur la distance entre les caractéristiques projetées, les paires positives devant avoir des distances inférieures au seuil, et les paires négatives des distances supérieures au seuil. L'optimisation est réalisée en minimisant un critère de comptage de paires mal classées. Bien que compétitive avec les méthodes de l'état de l'art de l'époque pour la ré-identification en base fermée sur les ensembles PRID2011 et iLIDS-VID, notre méthode est distancée dans les scénarios de base ouverte par les approches concurrentes qui exploitent des descripteurs spatio-temporels. Nous analysons ces faiblesses par le fait que d'une part, aucune contrainte de classement des identités n'est utilisée, et que d'autre part, le choix des caractéristiques initiales est de première importance pour atteindre de bonnes performances, la transformation linéaire optimisée par COPReV n'étant sans doute pas suffisante pour obtenir des représentations discriminantes. Ce dernier argument est validé par le constat que les nombreuses approches d'apprentissage profond qui ont vu le jour ces cinq dernières années ont permis d'apprendre de bien meilleures représentations pour la tâche de ré-identification. L'amélioration des performances apportées par les approches récentes de ré-identification en ensemble fermé a été spectaculaire, et vérifiée sur plusieur ensembles de données. A titre d'exemple, sur l'ensemble Market1501 (Zheng et al., 2015a), on est passé d'une mAP de 25% avec XQDA+LOMO (Liao et al., 2015) à plus de 80% (Hermans et al., 2017) voire 90% (Quispe et Pedrini, 2019 ; Zhang et al., 2018) ! Ces progrès sur l'apprentissage de représentations et de métrique laissent penser qu'il serait intéressant de revisiter une approche comme COPReV dans le cadre de l'apprentissage profond, en intégrant dans le même processus d'apprentissage la contrainte de vérification et de classement.

La seconde contribution prend une direction différente. Nous nous intéressons aux modèles de représentation parcimonieuse pour plusieurs raisons. Dans les approches directes de représentation parcimonieuse avec collaboration, les caractéristiques d'une personne requête sont exprimées à l'aide un dictionnaire formé par les caractéristiques des personnes de la galerie, ce qui permet la compétition des éléments de la galerie. La parcimonie impose que seuls les éléments les plus représentatifs soient sélectionnés dans la représentation. D'autre part, l'erreur de reconstruction sert à trier les candidats de la galerie par leur degré de ressemblance avec la requête. Enfin, le multi-shot est géré de manière transparente avec ce formalisme. L'apport des approches parcimonieuses avec collaboration est démontré dans une évaluation expérimentale sur PRID2011 et iLIDS-VID en base fermée et en base ouverte. Nous montrons également qu'une collaboration élargie avec l'ajout d'un dictionnaire générique augmente les performances à la fois en base fermée et en base ouverte en ajoutant de la diversité et en favorisant le rejet des imposteurs par une reconstruction à partir d'éléments extérieurs à la galerie. Dans un second temps, les représentations inverses sont présentées. Cette fois-ci, chaque élément de la galerie est représenté par les éléments de la requête lorsque ceux-ci sont présentés en même temps. La représentation inverse s'avère très efficace car elle permet à la fois de comparer les élements de la galerie en utilisant le même dictionnaire et de mettre en compétition les identités de la requête. La version bi-directionnelle combine l'approche directe et l'approche inverse. Les résultats expérimentaux sur PRID2011 et iLIDS-VID montrent la pertinence des représentations inverses et bi-directionnelles pour la ré-identification en base fermée et en base ouverte, qui dépassent assez largement celles de approches directes et des méthodes concurrentes de l'état de l'art. Dans cette seconde contribution, les expériences ont une nouvelle fois été réalisées en utilisant des caractéristiques en entrée assez faibles (LOMO) comparées à celles qui seraient apprises par un réseau de neurones profond dans les méthodes plus récentes. Il serait intéressant d'apprendre des descripteurs adaptés à un appariement avec des représentations parcimonieuses. D'autres perspectives de recherche pourraient concerner l'apprentissage de dictionnaires plus optimaux, la réduction des temps de calcul in-

duits par les représentations inverses et bi-directionnelles, ou encore la modélisation d'une galerie dynamique qui s'enrichit au fil du temps avec de nouvelles identités.

Ces dernières années, les méthodes fondées sur les réseaux de neurones ont saturé l'espace des publications sur la ré-identification visuelle. Grâce à un effort de recherche exceptionnel de la communauté dans ce domaine, un grand saut de performance a été constaté, faisant passer les technologies de ré-identification à un état de maturité qui laisse envisager leur déploiement à grande échelle. Il subsiste néanmoins beaucoup de défis à relever. En effet, la ré-identification en base ouverte est encore aujourd'hui un problème peu abordé en comparaison de la ré-identification en base fermée. Pourtant, les scénarios de base fermée ne sont pas très réalistes d'un point de vue applicatif. A ce titre, nos travaux ont permis d'ébaucher quelques pistes de réflexion et éléments de réponse.

La modélisation d'apparence des personnes avec plusieurs observations (multi-shot) mérite également d'être étudiée plus avant, car c'est certainement l'une des clés qui permettra gagner de façon substantielle en robustesse. On pourra se poser la question du choix des images à prendre en compte pour représenter les personnes (en fonction de la variation d'apparence, de pose, etc.), et de la complexité des méthodes vis à vis du nombre d'images utilisées, comme dans (Zhou et al., 2018).

D'autres voies sont à explorer. Le pipeline classique de ré-identification de personnes fait intervenir un détecteur externe. L'apprentissage de bout-en-bout de la détection et ré-identification simultanées permettrait de combiner en un seul modèle les deux tâches. Les avantages seraient de tirer parti du partage de caractéristiques et de réduire les temps de calcul des deux étapes. Un premier travail engagé par Loesch et al., 2019 constitue une base de travail intéressante. Suivi visuel multi-cibles et ré-identification sont deux tâches qui partagent beaucoup de similitudes, en particulier le besoin de modéliser de façon discriminante l'apparence des individus. Les travaux de thèse de Loïc Fagot-Bouquet (Fagot-Bouquet et al., 2016) menés dans notre laboratoire sur le suivi multi-cibles, ont en partie inspiré nos contributions sur la ré-identification avec des modèles de représentation parcimonieuse collaborative. Une approche récente et très prometteuse, FairMOT (Zhang et al., 2020a), combine détection et ré-identification dans un seul réseau de neurones pour effectuer le suivi visuel multi-cibles, la signature de ré-identification servant à associer les détections au cours du temps. Dans ce modèle *single-shot*, les caractéristiques de ré-identification sont néanmoins dépendantes du contexte local des scènes où se trouvent les personnes dans les données d'apprentissage. La tâche de ré-identification reste très difficile dans les environnements denses, où dans les scènes dynamiques comportant de nombreuses interactions entre les personnes dont l'apparence est similaire.

Le problème du défaut de généralisation des modèles apparaît dans toutes les tâches de vision, et en particulier en ré-identification. La plupart du temps, il n'est pas envisageable de créer pour chaque nouveau contexte une base d'images annotées de taille suffisante et de ré-entraîner le modèle complètement. Ces raisons motivent les travaux actuels sur l'apprentissage non supervisé pour l'adaptation de domaine avec des méthodes de pseudo-labelling (Dubourvieux et al., 2020 ; Ge et al., 2020) ou l'apprentissage de représentations désenchevêtrées (Li et al., 2019b) où l'information d'identité est séparée de la pose et du domaine. Une autre manière d'obtenir une meilleure invariance au domaine consiste à utiliser des informations auxiliaires comme les attributs sémantiques (Wang et al., 2018a), très utiles pour faciliter la recherche dans des grands volumes d'images. Plusieurs projets de recherche, dont la thèse de Fabian Duvourvieux, sont en cours de réalisation au laboratoire sur ces sujets.

Chapitre 3

Analyse du comportement et détection d'événements dans les vidéos

3.1 Contexte et motivations

Dans ce chapitre, nous abordons la thématique de l'analyse du comportement humain à partir de données de vidéo-surveillance. Nous nous intéressons plus particulièrement au problème de la détection de comportements anormaux, en portant notre attention sur deux problématiques : la reconnaissance de situations de violence et l'analyse de comportement dans les foules. Nous allons d'abord situer plus précisément le cadre de nos travaux.

Détection de comportements anormaux Un comportement anormal n'est défini que par opposition à un comportement normal. Le terme "anormal" peut être estimé impropre car il se réfère souvent dans le langage commun à un jugement de valeur de ce qui n'est pas acceptable. Au sens strict, l'anormalité constitue ce qui sort de la norme, c'est-à-dire ce qui inhabituel, rare. En pratique, les observations des exemples des comportements à reconnaître auront tendance à être difficiles à trouver, alors que les exemples des comportements habituels seront disponibles en grand nombre. Prenons deux exemples. Un système d'analyse de trafic routier installé au bord d'une autoroute n'a que très peu de chances d'observer un accident. Un mouvement de panique dans une foule n'est -fort heureusement- vu par une caméra de surveillance que de manière exceptionnelle. Un problème analogue auquel il faut faire face dans les applications de contrôle industriel est la détection d'anomalies : il s'agit dans ce cas de reconnaître des défauts qui apparaissent de manière très sporadique sur des équipements ou des pièces manufacturées.

Evénements violents Dans le domaine de la sécurité, les collectivités locales, les forces de l'ordre, les opérateurs de transports, regardent avec un grand intérêt les innovations technologiques en matière de vidéo-surveillance qui pourraient augmenter leur réactivité, leur capacité d'analyse et d'anticipation. Les caméras de surveillance génèrent un flot continu et immense de données vidéo qu'un nombre limité d'observateurs humains est incapable de scruter avec attention. Les systèmes de détection automatique d'événements d'intérêt, à condition que leur robustesse et leur sélectivité soient suffisantes, constituent des outils précieux pour alerter de situations dangereuses, filtrer l'information présentée aux analystes pour faciliter leur travail d'investigation dans le cadre de la lutte contre la criminalité et la prévention des délits. Parmi les événements anormaux d'intérêt, les situations de violence occupent une place de choix. Un des principaux problèmes lorsqu'on s'attaque à ce sujet est qu'il est très difficile de délimiter avec précision le périmètre de la violence car celle-ci englobe des situations et des contextes très hétérogènes, où le nombre de personnes impliquées, la nature et le style d'exécution des actions varient fortement.

Les agressions physiques, les bousculades, les rixes, le vol, le vandalisme sont autant de manifestations de comportements violents ou agressifs.

Analyse des comportements dans la foule La sécurité et le confort des usagers sont des enjeux majeurs de la gestion des lieux de forte fréquentation que sont les espaces urbains, les infrastructures de transport, ou au cours d'événements culturels, religieux, politiques, sportifs qui rassemblent un grand nombre de personnes. Les événements passés comme l'incident de la Love Parade à Duisburg en 2010 et la bousculade de Phnom Penh survenue la même année, illustrent tragiquement le besoin de mesures et d'instruments adaptés pour mieux anticiper les situations critiques lors de grands rassemblements. Une problématique plus légère mais néanmoins importante concerne l'aménagement des villes de forte densité de population où la circulation des personnes doit être fluidifiée. D'autres domaines d'application, comme l'étude des comportements humains en sciences sociales, ou le développement de modèles de simulation pour le jeu vidéo et le cinéma s'appuient sur la caractérisation de la foule à partir d'images. La modélisation des comportements des personnes dans une foule est un problème complexe. Une foule est composée d'individus qui ont chacun un comportement propre, or celui-ci est dans le même temps largement influencé par la présence et le comportement des personnes proches. On peut observer des comportements collectifs qui, en fonction du lieu et de la situation, sont considérés normaux (un flux important de personnes traversant dans un couloir de métro aux heures de pointe), ou sont au contraire totalement inattendus (un mouvement de panique dans un stade suite à des actes de violence, une densité de personnes inhabituelle).

Approches macroscopiques vs approches microscopiques Alors que les travaux présentés dans les chapitres 1 et 2 avaient pour vocation d'établir des briques de base pour la modélisation individuelle, nous nous penchons ici sur les possibilités de caractériser un comportement sans nécessiter une telle individualisation. De telles approches sont qualifiées d'approches macroscopiques, par opposition aux approches microscopiques, puisqu'elles considèrent la scène comme un ensemble sans chercher à distinguer les éléments (objets, personnes) qui la composent. Plusieurs raisons justifient l'intérêt d'explorer ce type d'approches. La modélisation individuelle n'est pas toujours réalisable avec suffisamment de fiabilité à cause de la complexité de la scène (scènes très denses, très faible résolution des personnes dans les images). En effet, elle nécessite de détecter, suivre et interpréter les actions de chacun des individus, de reconnaître leurs interactions avec une précision acceptable. Par ailleurs, l'approche ascendante qui consiste à partir de comportements individuels à inférer un comportement collectif n'est absolument pas triviale et requiert une modélisation sophistiquée des interactions. Un autre défi pour les approches microscopiques est la complexité calculatoire des chaînes de traitement engendrée par la prise en compte d'un grand nombre d'individus. A l'inverse, les approches macroscopiques abordent de manière plus directe l'analyse du comportement au sein d'un ensemble.

Quelles difficultés ?

L'objectif d'analyser le comportement humain à partir de vidéos, notamment pour la détection d'événements anormaux ou violents, nous confronte à un certain nombre de difficultés.

Définition du domaine sémantique La premier obstacle est sans doute la difficulté de définir de façon précise et non ambiguë les types de comportements d'intérêt. Il existe une telle variété de comportements qu'il est en général plus simple de les regrouper dans de grandes catégories plutôt que d'essayer de les distinguer plus spécifiquement. Sous le vocable de "violence", on rassemble les agressions, le vandalisme, l'altercation musclée, la panique dans un groupe ou une foule qui dégénèrent en rixe, etc. Un autre marqueur de la difficulté de délimiter les classes d'événements d'intérêt est la notion même d'anormalité

qui est une définition par la négation. L'intensité des événements est également difficile à évaluer de manière objective.

Multiplicité des comportements Dans un groupe ou une foule, les personnes ont à la fois un comportement propre et interagissent avec les autres. Même si on peut considérer à un niveau global que les interactions mènent à des actions collectives, il n'en reste pas moins qu'on peut observer dans le même temps des comportements distincts et localisés dans la scène. La localisation est une information importante dont l'estimation est rarement précise.

Complexité des scènes Les scènes de vidéo-surveillance comportent des facteurs multiples de complexité. La densité élevée de personnes dans les scènes de foule, ou dans un espace confiné, complique la tâche, en raison de la faible résolution dans les images et des phénomènes d'occultation fréquents. Beaucoup d'approches d'analyse de scène, conçues et testées dans des contextes simples de faible densité de personnes, deviennent caduques. Modéliser la grande variabilité des environnements, des poses, des gestes et de l'apparence des personnes est une tâche ardue. Les environnements mobiles ajoutent les difficultés de changements de luminosité brutaux et de mouvements apparents du monde extérieur.

Conditions d'acquisition vidéo Il faut faire face sur le terrain aux contraintes techniques de captation des flux vidéo : angles de vue de caméra défavorables, vision trop rapprochée ou trop globale, hétérogénéité de la qualité d'image et de la fréquence d'acquisition des caméras de surveillance héritée des isntallations existantes... La couverture de l'espace à bord d'un train par des caméras est à elle-seule un véritable casse-tête car on dispose de très peu de recul en hauteur pour observer un espace relativement long, occulté de toute part par les sièges. Un cas assez différent est celui d'une prise de vue zénithale et très distante d'un grand espace, dans lequel les personnes sont observées avec une très faible résolution spatiale mais ne s'occultent que peu.

Moyens d'évaluation L'évaluation quantitative des performances d'un système de détection d'événements dans les séquences vidéo et la comparaison objective des méthodes s'avèrent particulièrement délicates. Les règles de labellisation varient : comment annoter avec précision le début et la fin d'un événement, comment déterminer sa localisation dans l'image ? L'évaluation par image est simple à réaliser : on compte le nombre d'images où la vérité terrain et la détection correspondent ou pas. Cependant, ce protocole met plus de poids sur les événements de longue durée, au détriment des événements plus fugaces. Une autre manière consiste à évaluer en comptant les événements, mais dans ce cas, comment fixer les seuils sur le nombre minimal d'images avant la première détection, indicatif de la latence de détection, le nombre minimal d'images pour lesquelles on a une bonne détection sur la durée totale de l'événement, le nombre d'images où l'événement est détecté après la fin réelle de cet événement ?

3.2 Etat de l'art et positionnement

3.2.1 Détection d'événements violents

Au moment de nos travaux, la littérature portant précisément sur la détection d'événements de violence dans les environnements complexes de vidéo-surveillance était assez peu abondante. Plus généralement, on trouve un grand nombre de travaux qui traitent de l'analyse de comportement humain dans les vidéos. Ces travaux peuvent être regroupés en deux grandes thématiques : la reconnaissance d'actions humaines, et la détection d'événements anormaux.

Reconnaissance d'actions humaines

La littérature fourmille de travaux sur la reconnaissance d'actions humaines dans les vidéos (Pareek et Thakkar, 2021). Sans entrer dans une présentation détaillée, nous faisons référence à quelques travaux qui mettent en perspective ceux que nous avons réalisés sur la détection de violence. Des approches de classification multi-classes par apprentissage supervisé ont été proposées. Parmi les travaux publiés peu de temps avant les nôtres, citons (Gaidon et al., 2014 ; Guo et al., 2013 ; Marín-Jiménez et al., 2013 ; Wang et al., 2013 ; Wang et Schmid, 2013). Dans les tâches de classification d'actions, les vidéos sont présentées sous la forme de courts clips faisant apparaître une action principale. Dans ces vidéos, le point de vue de la caméra est quelconque et la caméra souvent mobile. Les classes d'actions diffèrent selon le jeu de données utilisé : actions sportives, actions de la vie quotidienne, etc. Parmi les jeux de données fréquemment utilisés pour évaluer la tâche de classification d'actions, on retrouve Hollywood2 (Marszalek et al., 2009), HMDB51 (Kuehne et al., 2011), Olympic Sports (Niebles et al., 2010), UCF50 (Reddy et Shah, 2013) et son extension UCF101 (Soomro et al., 2012), et des ensembles de données vidéo plus récents et de taille beaucoup plus grande tels que Youtube-8M (Abu-El-Haija et al., 2016) et Kinetics (Smaira et al., 2020). Certaines classes d'actions de ces jeux de données telles que ('fighting', 'punching', 'kicking', 'wrestling', 'boxing') pourraient être regroupées pour représenter les actions agressives. Néanmoins, il n'est pas certain que l'union des ces classes d'actions, ainsi que celle des autres classes couvrent correctement l'ensemble des situations de violence et de situations normales pour un contexte donné.

Les approches de classification d'actions reposent sur des représentations visuelles spatio-temporelles :

Descripteurs faits à la main Ces descripteurs ne nécessitent pas de détecter des personnes ou des parties de corps. L'extension des points d'intérêt et des descripteurs locaux au domaine spatio-temporel (STIP pour *Space-Time Interest Points*) a été souvent explorée (Laptev et al., 2008 ; Liu et al., 2009), encore récemment (Nazir et al., 2018), et souvent en combinaison avec des méthodes de sacs de mots visuels (Laptev et al., 2008 ; Marín-Jiménez et al., 2013). La performance de ces approches dépend grandement de la qualité d'extraction des points d'intérêt. Les LBP sont étendus au domaine spatio-temporel dans (Yeffet et Wolf, 2009), alors que les SIFT inspirent Chen et Hauptmann, 2009 pour créer les descripteurs MoSIFT. Guo et al., 2013 proposent une représentation dense et concise du mouvement sous la forme de matrices de covariance de vecteurs de caractéristiques. Cette modélisation est néanmoins appliquée dans des cas assez simples.

Trajectoires La modélisation par trajectoires de points denses (Ballas et al., 2013 ; Wang et al., 2013) a elle aussi connu un grand succès dans le domaine de la reconnaissance d'actions. Cette approche combine plusieurs types de descripteurs (HOG, HOF, HoMB) pour capturer la forme, l'apparence, et les mouvements locaux au cours du temps, tout en s'affranchissant du mouvement global de la caméra. Gaidon et al., 2014 développent une méthode d'apprentissage non supervisée d'une représentation hiérarchique des trajectoires denses pour caractériser l'activité des personnes. L'estimation robuste des trajectoires s'avère délicate lorsque les mouvements observés sont chaotiques et très rapides, comme c'est le cas des événements agressifs.

Apprentissage de représentation L'apprentissage profond et les réseaux de neurones ont permis de revoir les représentations utilisées pour la reconnaissance d'actions. Ainsi, l'architecture *Two-stream ConvNet* (Simonyan et Zisserman, 2014a) combine un réseau spatial et un réseau temporel qui encode le mouvement à partir de plusieurs images consécutives et du flot optique. Dans (Wang et al., 2015) les caractéristiques issues de l'apprentissage sont combinées à des descripteurs faits à la main, en les extrayant le long

de trajectoires par une opération de *pooling*. Les architectures de type Conv3D étendent la convolution dans le domaine spatio-temporel (Ji et al., 2013 ; Qiu et al., 2017). Les réseaux récurrents LTSM ont été mis en œuvre pour renforcer la modélisation temporelle de longue durée (Ng, Joe Yue-Hei et al., 2015).

Classification d'événements violents

Le problème plus spécifique de détection de violence a été beaucoup moins souvent abordé que celui de la reconnaissance d'actions. Certains travaux considèrent la détection de violence dans les vidéos comme un problème de classification binaire où le modèle de classification est appris par apprentissage supervisé sur des données de sport dans des flux multimedia (Bermejo Nievas et al., 2011 ; De Souza et al., 2010) ou de vidéo-surveillance (Hassner et al., 2012). La détection de scènes de violence dans les données multimedia a donné lieu à des benchmarks spécialisés dans le cadre de MediaEval (Sjöberg et al., 2014). Dans ces données multimedia, l'information audio et le contenu textuel jouent un rôle prépondérant dans les tâches de classification. Cependant des indices visuels tels que la présence de flammes, de sang, les niveaux d'intensité des mouvements, etc. peuvent être extraits des images. Giannakopoulos et al., 2006 relient l'activité observée dans la vidéo à une situation de violence par l'intermédiaire d'une métrique sur le mouvement (mouvement moyen et variance de la direction du mouvement). Un autre travail intéressant concerne l'estimation du flou dans les images, conséquence de l'acquisition de mouvements dans le cas de fortes accélérations, typiques des actions de violence (Déniz et al., 2014). Pour les applications multimedia, il existe un biais important dans les données qui provient du fait que l'action principale bénéficie d'un cadrage centré spatialement et temporellement. De plus, les données audio (bruits d'explosions, cris, sons de coups portés, d'objets brisés, etc.) apportent des informations précieuses pour comprendre la situation. En vidéo-surveillance, l'information sonore are rarement disponible et l'événement peut survenir à n'importe quel endroit de la scène observée. Toutefois, Lefter et al., 2013 ont étudié la fusion multimodale pour la surveillance de l'intérieur d'un train. Pour reconnaître les situations de violence dans les scènes denses, Hassner et al., 2012 proposent un descripteur de texture dynamique, *violent flow*, spécialement adapté à l'observation de foules à grande distance pour classifier les clips vidéos. Ce descripteur exploite la variation d'amplitude du mouvement estimé par une méthode de flux optique. Les auteurs fournissent un ensemble de clips vidéo collectés sur Youtube, appelé *Violent Flows*.

Plus récemment, de nouveaux ensembles de données vidéo contenant des scènes de violence, de plus grande ampleur ont été mis à disposition, notamment pour entraîner des modèles de deep learning (Soliman et al., 2019). RWF-2000 est à ce jour une des plus grandes bases disponibles, avec 2000 vidéos de vidéo-surveillance (Cheng et al., 2020).

3.2.2 Détection d'événements anormaux

Si l'accès à des bases de vidéos de grande taille pour l'apprentissage de modèles de détection de violence est possible depuis peu, il n'en reste pas moins que les situations de violence sont rares dans un contexte donné. L'opportunité nous est donnée d'attaquer le problème de reconnaissance d'événements violents sous l'angle de la détection d'événements anormaux. Au niveau sémantique, la définition de l'anormalité est certes moins précise, mais cette généralité peut se tourner en avantage si l'on trouve des façons de transposer le cadre des approches développées à d'autres applications.

La détection d'événements anormaux, de comportements suspects a préoccupé de nombreux chercheurs (Dhiman et Vishwakarma, 2019 ; Popoola et Wang, 2012). Certains travaux se concentrent sur les scènes de foule, comme dans Kratz et Nishino, 2009 ; Li et al., 2014b ; Luvison et al., 2011 ; Mehran et al., 2009 ; Mousavi et al., 2015, d'autres étudient le cas plus général des comportements anormaux sans a priori sur le type de scène observée (Bertini et

al., 2012 ; Boiman et Irani, 2005 ; Roshtkhari et Levine, 2013b ; Roshtkhari et Levine, 2013c). Les travaux pionniers de Boiman et Irani, 2005 établissent un cadre de détection d'événements anormaux localisés dans les vidéos où la distribution statistique d'ensemble de patches 3D est modélisée de manière non paramétrique à partir des exemples d'apprentissage. Les patches sont représentés par des descripteurs locaux et leur position relative au centre de l'ensemble. Les "irrégularités" sont détectées en comparant les ensembles de patches à évaluer avec ceux du modèle grâce à une méthode d'inférence qui recherche l'ensemble le plus probable dans la base maximisant la vraisemblance de l'ensemble requête. Roshtkhari et Levine, 2013c partent du même principe d'ensemble avec les STV (*Spatio-Temporal Volumes*) qui codent sur une grille dense 3D l'apparence et le mouvement dans des cubes élémentaires, ainsi que leurs relations par le moyen d'un graphe. Une différence importante avec (Boiman et Irani, 2005 ; Boiman et Irani, 2007) est qu'au lieu de conserver toutes les données, les auteurs se servent d'un dictionnaire de mots visuels pour réduire la taille du modèle. Le dictionnaire peut être appris en ligne, et l'algorithme est adapté à un fonctionnement en temps réel. La méthode est appliquée pour détecter et localiser des événements anormaux beaucoup plus simples que les événements de violence : ce sont pour la plupart la présence inattendue d'objets dans une zone de l'image, ou une trajectoire spatiale d'objets déviant de celles qui sont vues à l'apprentissage.

Le cadre d'apprentissage "une-classe", plus souple d'emploi que l'apprentissage supervisé multi-classes, ne requiert pour l'entraînement des modèles que des données des situations normales. La labellisation des données est donc implicite et gratuite. La généricité de ce type d'approches reporte quelque peu le problème de la discrimination des catégories d'événements spécifiques sur la conception de représentations visuelles adaptées. On trouve souvent dans la littérature des descripteurs s'appuyant sur l'estimation du mouvement par flux optique (Cong et al., 2011 ; Kim et Grauman, 2009 ; Mehran et al., 2009 ; Zhao et al., 2011 ; Zhu et al., 2014b), le gradient (Bertini et al., 2012 ; Kratz et Nishino, 2009 ; Zhao et al., 2011), la modélisation de la texture au cours du temps (Li et al., 2014b ; Ma et Cisar, 2009 ; Zaharescu et Wildes, 2010), ou encore la variation de texture dans le domaine fréquentiel (Kaltsa et al., 2012 ; Wang et al., 2012a). Des approches plus récentes utilisent des réseaux de neurones pour créer les descripteurs, comme dans (Wang et al., 2019) avec PCANet. Une autre idée intéressante est celle qui est exploitée dans (Ravanbakhsh et al., 2017) : l'entraînement adverse d'un modèle génératif conditionnel à générer des images et des cartes de flux optique de données normales permet à l'inférence de déterminer des différences avec le modèle lorsque la reconstruction est mauvaise.

3.2.3 Analyse du comportement de la foule

Les recherches sur l'analyse des scènes de forte densité de personnes se divisent en plusieurs thématiques (Bendali-Braham et al., 2021 ; Khan et al., 2020) :

- La détection, le comptage et l'estimation de densité (Sindagi et al., 2019 ; Sindagi et Patel, 2018 ; Thanasutives et al., 2020),
- Le suivi d'individus dans la foule (Dendorfer et al., 2020 ; Franchi et al., 2020),
- L'analyse de la dynamique des flux des personnes (Ali et Shah, 2007 ; Shao et al., 2014 ; Zhu et al., 2014a),
- La détection d'événements anormaux (Afiq et al., 2019 ; Kratz et Nishino, 2009 ; Mehran et al., 2009),
- La classification de comportements spécifiques. (Dupont et al., 2017 ; Wei et al., 2020),

Nous limitons le périmètre de notre étude aux deux dernières problématiques. Les méthodes microscopiques (centrées sur les objets) réalisent une détection et un suivi visuel des individus, puis infèrent des comportements collectifs à partir de l'ensemble des trajectoires individuelles (Choi et Savarese, 2012). Ces approches se heurtent aux nombreuses difficultés inhérentes aux scènes de densité élevée : la faible résolution des personnes dans les images, les occultations et les interactions fréquentes entre les personnes qui mettent déjà assez facilement en échec les

étapes de détection et de suivi. Les méthodes récentes de détection reposant sur l'apprentissage profond commencent à délivrer des performances intéressantes dans les scènes de foule, mais n'adressent pour la plupart que le problème du comptage et de l'estimation de densité (Khan et al., 2020). Les approches macroscopiques ou holistiques traitent la foule comme un ensemble et cherchent à en extraire des propriétés le plus souvent associées aux mouvements des individus qui le composent (Li et al., 2015c).

Modélisation physique de la dynamique de la foule

Les mouvements au sein d'une foule sont dits structurés lorsque les personnes semblent agir vers un même but collectif et se déplacer de manière cohérente, et non structurés lorsqu'on peut distinguer des comportements localement différents attribuables à des groupes de personnes. Plusieurs travaux font appel aux principes de la mécanique des fluides pour étudier la dynamique de la foule dont les individus sont vus comme des particules en mouvement. Ali et Shah, 2007 adoptent la description lagrangienne pour représenter le déplacement de points comme l'écoulement d'un fluide. La méthode est appliquée pour la segmentation de régions dont les flux majoritaires sont différents. Le *Social Force Model* (SFM) présenté pour la première fois par Helbing et Molnár, 1995 décompose le comportement des personnes en plusieurs termes, la motivation intrinsèque des personnes, les forces d'attraction et les forces de répulsion qui les maintiennent à une certaine distance. Les modèles physiques sont à l'origine de plusieurs approches de détection de comportement anormales dont celles de Mehran et al., 2009, Raghavendra et al., 2011 et Wu et al., 2010. Dupont et al., 2017 proposent une ontologie de 11 comportements élémentaires (Gas Free, Gas Jammed, Laminar Flow, Turbulent Flow, Crossing Flows, Merging Flow, Diverging Flow, Static Calm, Static Agitated, Interacting Crowd, No Crowd) inspirée de la mécanique des fluides.

Descripteurs visuels 2D+t

Le mouvement apparent dans la vidéo est extrait à l'aide de méthodes d'estimation de flux optique et sert à calculer des descripteurs de bas niveau construits les gradients spatio-temporels (Cong et al., 2013 ; Kratz et Nishino, 2009 ; Krausz et Bauckhage, 2012 ; Luvison et al., 2011 ; Varadarajan et Odobez, 2009).

Dans des travaux ultérieurs, l'information de mouvement se retrouve dans des modèles de trajectoires de points d'intérêt suivis au cours du temps (*tracklets*) (Fradi et Dugelay, 2014 ; Mousavi et al., 2015 ; Shao et al., 2014), ou en utilisant la technique d'advection de particules pour propager temporellement la position d'un point en suivant la direction et l'amplitude du flux optique (Mehran et al., 2010 ; Wu et al., 2010). L'avantage des modèles de *tracklets* par rapport aux descripteurs de flux optique est de capturer les relations spatio-temporelles à des échelles de temps plus grandes et donc d'apporter des informations plus globales et de plus haut niveau sémantique.Zhou et al., 2012 regroupe les trajectoires cohérentes en analysant le mouvement des points voisins. Mousavi et al., 2015 utilisent les *tracklets* pour calculer des descripteurs d'histogrammes de *tracklets* orientés et détecter des événements anormaux dans la foule. Bien que majoritairement employées pour les tâches de détection, de comptage et d'estimation de densité dans les foules, les représentations issues de réseaux de neurones convolutifs profonds sont étudiées depuis peu pour résoudre le problème de classification de vidéos de foule (Dupont et al., 2017 ; Shri et Jothilakshmi, 2019).

Modèles d'interactions

Certaines approches séparent explicitement les comportements individuels des interactions entre les personnes. Les individus ne sont pas réellement détectés, mais simplement représentés

par des points d'intérêt. Le modèle SFM de Mehran et al., 2009 mesure les forces d'interactions en se basant sur la différence entre la vitesse attendue et la vitesse observée. Cui et al., 2011 définissent une énergie potentielle d'interaction qui tient compte de la distance entre deux points et de l'angle entre les directions de déplacement. Dans (Cho et Kang, 2014), la modélisation des actions individuelles et collectives prend la forme d'un système hybride où des agents statiques codent les variations d'amplitude et de direction de mouvement, et des agents dynamiques se spécialisent sur les interactions sociales. Un agent dynamique associé à un objet évalue les mouvements relatifs de ses voisins en utilisant les critères de force sociale et d'énergie potentielle d'interaction introduits dans les travaux précédents. Le comportement de la foule est ensuite codé sous la forme d'un histogramme de mots visuels trouvés par une méthode de clustering et la détection de comportements anormaux s'effectue en optimisant de manière supervisée un SVM pour la classification binaire des classes normal et anormal.

Les travaux les plus proches des nôtres sont ceux de Shao et al., 2014 : pour mieux capturer les différentes propriétés individuelles et collectives, les auteurs développent un ensemble étendu de descripteurs spatio-temporels calculés dans des groupes détectés et segmentés au préalable. Les trajectoires des points d'intérêt (*tracklets*) sont extraites avec un algorithme de suivi de type KLT (Shi et Tomasi, 1994) et servent à la fois à détecter les groupes et à calculer les descripteurs. Les *tracklets* sont reliées dans un graphe par la méthode des plus proches voisins. Sur ce graphe, quatre types de descripteurs intra-groupe et inter-groupes sont calculés : le comportement collectif qui mesure la cohérence des mouvements au sein d'un groupe, la stabilité qui représente la variation de topologie dans un groupe au sens des voisins d'un élément, l'uniformité qui évalue la répartition spatiale des individus, et le conflit qui mesure les interactions (frictions) entre les groupes (Figure 3.1. Cette description riche repose néanmoins sur la qualité de l'étape de segmentation des groupes. Les propriétés étant évaluées au niveau des groupes, l'information plus locale est quelque peu effacée.

Figure 3.1 – Illustration des descripteurs de groupe proposés par Shao et al., 2014.

Modèles d'apprentissage

Les modèles probabilistes sont les modèles les plus répandus pour l'apprentissage des motifs de mouvement et d'activité dans la foule. Les relations temporelles entre motifs de mouvement dans des blocs de l'image ont été modélisées par des chaînes de Markov cachées pour la détection d'anomalies (Kratz et Nishino, 2009) : la probabilité qu'un motif dans un bloc appartienne à une distribution d'observations prototype est évaluée dans un tube spatio-temporel. Le couplage entre chaînes de Markov cachées permet d'ajouter des relations entre blocs voisins. Les modèles de langage naturel ont également inspiré des recherches sur l'analyse de foule et la détection d'anomalies. L'apprentissage non supervisé de modèles thématiques, comme dans l'approche pLSA (*probabilistic Latent Semantic Analysis*) qui est appliquée sur un jeu de descripteurs locaux de position, taille et mouvement (Varadarajan et Odobez, 2009), fait émerger des motifs d'activités récurrentes. La détection d'événements anormaux est alors effectuée en calculant la

vraisemblance du modèle thématique par rapport aux données. Mehran et al., 2009, dans leur approche de Social Force Model, utilisent la LDA (*Latent Dirichlet Allocation*) avec des sacs de mots visuels représentant les flots de force, pour trouver la distribution des thèmes (topics) du comportement normal dans la foule. L'utilisation exclusive de données de situations habituelles pendant l'apprentissage pour la tâche de détection d'événements anormaux dans la foule fait aussi appel à des méthodes de reconstruction avec un modèle parcimonieux (Cong et al., 2011 ; Huo et al., 2012) et des méthodes d'apprentissage une-classe comme le *one-class* SVM (Wang et Snoussi, 2015 ; Yang et al., 2019a).

3.2.4 Positionnement de nos travaux

Nous présentons dans la suite de ce chapitre deux contributions principales.

Détection d'événements violents La première contribution est le fruit de travaux que nous avons menés sur la problématique de la détection de situations de violence dans les flux vidéos. Ces travaux s'inscrivent dans le cadre des activités du laboratoire commun VisionLab entre le CEA et THALES dans le domaine de la vidéo-surveillance intelligente, et s'appuient en particulier sur les projets de recherche DéGIV (Détection et Gestion d'Incidents dans un Véhicule ferroviaire, 2011-2015, programme du FUI) et Secur-ED (Secure Urban Transportation - An European Demonstration, 2011-2014, programme FP7). Ces deux projets avaient pour but de développer de nouvelles solutions de sécurisation des usagers des transports publics, urbains et ferroviaires.

Nous avons pris le parti d'aborder la détection automatique d'événements violents dans des scènes complexes de vidéo-surveillance en la considérant comme un problème de détection d'événements anormaux. L'apprentissage ne s'effectue que sur les données de situations normales (apprentissage une-classe). Nous justifions ce choix par le constat qu'il est beaucoup plus difficile de collecter de grandes quantités de données vidéo de violence que des données de situations habituelles, pour un contexte donné, par exemple à l'intérieur d'un véhicule ferroviaire. Contrairement aux travaux de Lefter et al., 2013, seules les données visuelles sont exploitées. Un objectif secondaire est que l'apprentissage puisse être effectué en ligne pour adapter de manière flexible la modélisation à la scène, une fois le système installé.

Il est alors capital de disposer d'une représentation visuelle suffisamment discriminante pour la détection de la violence. C'est la ligne directrice que nous avons suivie lors de la conception du descripteur RIMOC (*Rotation-Invariant* feature modeling *MOtion Coherence*) spécialement adapté à la représentation de la variation de structures de mouvement. Nous avons cherché à mettre au point un modèle de description visuelle concis et rapide à calculer, capable de séparer les mouvements lisses et cohérents des mouvements destructurés. Cette représentation repose sur le principe des ensembles spatio-temporels calculés à plusieurs échelles introduits par Roshtkhari et Levine, 2013b ainsi que sur la stratégie de compression du modèle en un dictionnaire appris en ligne, et trouve son originalité dans l'idée d'exploiter les valeurs propres de la matrice de covariance des histogrammes de flux optique pour caractériser le mouvement. L'approche est particulièrement intéressante lorsqu'elle est utilisée comme un premier module de filtrage et de remontée d'alertes d'un système plus complet de vidéo-surveillance automatique embarqué.

Nos travaux ont également permis de créer un nouveau jeu de données vidéo entièrement dédié à la surveillance à bord des trains, qui est l'application ciblée dans le projet DéGIV. Ce ensemble comporte des séquences acquises i) dans une maquette de laboratoire à l'échelle 1 :1 d'une voiture de train et ii) dans un vrai véhicule ferroviaire roulant, pour un total de 18h de vidéo et 276 événements de violence. Ces séquences mettent en scène des scénarios variés d'agressions, d'altercations et de rixes à différents niveaux d'intensité.

Ce corpus de vidéos a servi à évaluer les performances de l'approche proposée, et à les comparer à celles des méthodes de l'état de l'art.

Les résultats de ces travaux ont fait l'objet d'une publication dans une revue (Ribeiro et al., 2016).

Analyse du comportement dans la foule La seconde contribution porte sur la proposition d'un nouvel ensemble de descripteurs pour la caractérisation des comportements dans des scènes de forte densité. Comme pour la première contribution, ces résultats ont été générés au sein du laboratoire commun CEA-THALES VisionLab. Nos travaux sont portés par le projet FluidTracks (2014-2017, programme du FUI) dédié au développement de solutions innovantes pour l'amélioration du confort et de la sécurité des personnes circulant dans l'espace public. Un des enjeux du projet consistait à trouver des moyens d'évaluer en temps réel la fluidité des déplacements des usagers dans les lieux de forte fréquentation et de détecter des événements anormaux tels que des situations de congestion ou de panique. FluidTracks a aussi été l'opportunité de coupler des outils de simulation multi-agents de la foule développés par THALES avec l'analyse vidéo. La simulation a servi à tester les systèmes de détection d'événements ; réciproquement, le simulateur a été nourri avec des données extraites par les algorithmes d'analyse vidéo pour augmenter son niveau de réalisme.

Nous partons des concepts de description introduits par Shao et al., 2014 (comportement collectif, uniformité, stabilité et conflit) qui ont montré des résultats prometteurs de compréhension des comportements dans la foule. Au lieu de procéder à une segmentation de la foule en groupes, nous préférons conserver une modélisation globale tout en estimant localement les différentes propriétés de mouvement, pour plusieurs raisons. Dans cette méthode de référence, la détection des groupes doit être faite hors-ligne, préalablement au calcul des descripteurs de comportement, ce qui n'est pas compatible avec un fonctionnement en temps réel. D'autre part, la qualité de la description est largement dépendante de la performance de la segmentation des groupes. Enfin, la notion même de groupes peut être remise en question puisque ces groupes peuvent varier continûment au cours du temps. Afin de modéliser les interactions, nous basons notre représentation sur un graphe issu d'une triangulation de Delaunay pour connecter les trajectoires de points (les nœuds du graphe). Cette triangulation assure une certaine régularité spatiale du maillage, tout en permettant de relier des points proches comme des points distants, sans identifier explicitement les groupes. Une modification de l'algorithme de suivi de points fait intervenir un mécanisme de ré-identification après une perte de suivi afin d'obtenir des *trackets* sur une plus longue durée. Les descripteurs de niveau intermédiaire qui seront détaillés dans la section 3.4 expriment différents composantes de la déformation du graphe au cours du temps.

Nous montrons ensuite expérimentalement comment cet ensemble de descripteurs combiné dans un seul vecteur de caractéristiques est appliqué avec succès dans trois applications différentes : la classification de vidéos de foule, la détection d'événements anormaux, et la détection de violence dans les scènes de foule. L'ensemble des résultats est détaillé dans l'article de revue (Fradi et al., 2017).

3.3 Modèle spatio-temporel de cohérence de mouvement pour la détection d'événements violents

3.3.1 Description générale de l'approche

Nous présentons dans cette section une nouvelle approche de modélisation de la structure du mouvement dans les vidéos fondée sur un nouvel espace de représentation (Ribeiro et al.,

2016). Dans cet espace, les valeurs propres des histogrammes de flux optiques (HOF : *Histograms of Optical Flow*) fournissent une caractérisation locale du mouvement qui est invariante à la rotation dans l'espace de l'image et permet de discriminer les types de mouvement d'intérêt des autres mouvements observés dans l'image. Le descripteur proposé, nommé RIMOC pour *Rotation-Invariant* feature modeling of *MOtion Coherence*, encode des caractéristiques locales du mouvement. Nous faisons l'hypothèse que les mouvements observés dans les situations de violence sont des mouvements déstructurés présentant de fortes variations de direction et d'amplitude au cours du temps. L'information d'apparence n'est pas prise en compte dans cette approche, seule l'information de mouvement est utilisée. Pour la reconnaissance des situations de violence, RIMOC est implémenté dans un cadre de détection d'événements anormaux, où seule la classe des événements normaux est apprise dans un cadre probabiliste (apprentissage une-classe). Les événements anormaux sont ceux qui dévient du modèle appris.

La séquence vidéo est découpée en volumes spatio-temporels élémentaires centrés sur une position spatiale dans l'image et un instant donnés. Chaque image du volume est représentée par un HOF dont on calcule les valeurs propres. L'étape d'entraînement consiste à apprendre un dictionnaire de mots c dans cet espace, à partir de données vidéo de situations normales. A l'étape d'apprentissage, les configurations spatio-temporelles des mots sont modélisées comme des distributions δ des positions relatives de ces mots dans un voisinage plus grand autour d'une position centrale. A l'inférence, on estime si une instance requête $\mathcal{E}$ est générée par ce modèle en calculant le maximum de vraisemblance sur toutes les configurations possibles, sachant $\mathcal{E}$, à différentes échelles. La similarité entre la requête et le modèle est alors prise comme le minimum des simarités calculées à chaque échelle. Ce principe d'inférence est inspiré de celui décrit par Roshtkhari et Levine, 2013b. Les données correspondant aux situations normales représentées par le modèle sont supposées plus probables que celles relatives aux situations anormales. La figure 3.2 donne une vue générale de l'approche.

3.3.2 RIMOC : une représentation concise du mouvement pour la détection de mouvements déstructurés

Notre objectif est de discriminer deux catégories de mouvement : les mouvements observés dans des situations normales, et les mouvements liés aux événements violents. Nous supposons que dans les situations normales, les mouvements possèdent une certaine régularité, continuité, et cohérence dans le temps. Dans les mouvements normaux, nous considérons l'absence de mouvement, les mouvements constants, les changements légers de direction, de vitesse ou des deux à la fois. A l'opposé, les mouvements agressifs sont désordonnés, chaotiques, destructurés : ces types de mouvements présentent de plus grandes variations de direction et d'amplitude. L'amplitude du mouvement dans l'image ne donne pas directement l'amplitude du mouvement réel, puisqu'elle dépend de la distance d'observation. Nous choisissons de raisonner uniquement dans l'espace image et de représenter le mouvement à différentes échelles pour gérer à la fois la localité du mouvement et l'adaptation de la description à la distance d'observation, en considérant que le mouvement est provoqué par des personnes et que donc, l'amplitude du mouvement doit être en quelque sorte rapporté à la taille des personnes dans l'image.

Malgré le fait que les deux catégories de mouvement (normal et agressif) soient très générales et comportent de nombreuses variations en leur sein, nous cherchons une représentation bas-niveau du mouvement qui soit indépendante des spécificités des mouvements. Les mouvements constants doivent avoir la même description quelle que soit leur direction, et les petites variations de direction et de vitesse représentées par des vecteurs proches dans l'espace de représentation choisi. En revanche, les descripteurs doivent être suffisamment différents dans le cas de mouvements agressifs pour les discriminer.

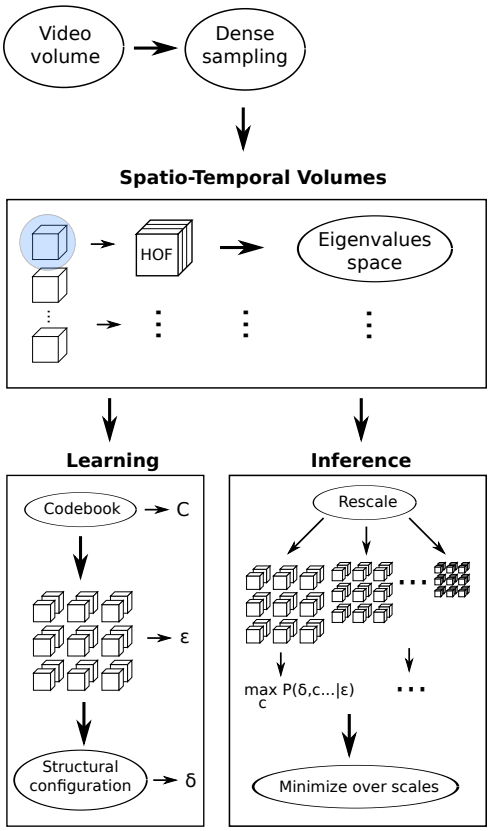
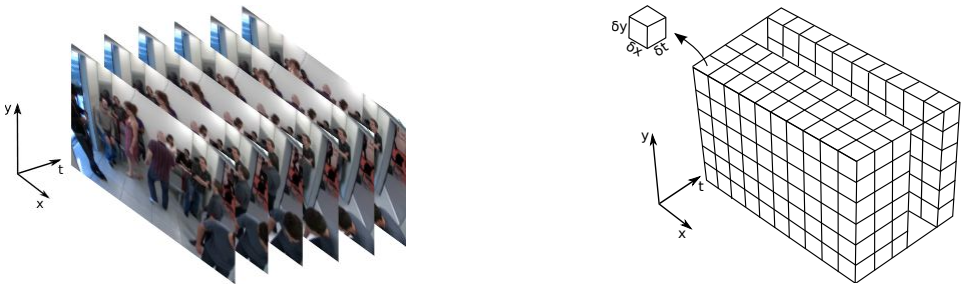


FIGURE 3.2 – Illustration de l'approche de détection d'événements anormaux : le mouvement est encodé sous la forme des valeurs propres des histogrammes de flux optique, dans des volumes spatio-temporels élémentaires. Lors de l'apprentissage sur des situations normales, un dictionnaire de mouvements de référence est formé et un modèle probabiliste des configurations de mouvement est appris en utilisant les mots du dictionnaire. A l'inférence, on estime si la structure observée est similaire au modèle ou non, et ce à plusieurs échelles d'observation.

Calcul des statistiques du mouvement sur une grille spatio-temporelle dense

Le calcul des statistiques du mouvement local est réalisé comme suit. Le mouvement est extrait de manière dense dans l'image avec un algorithme d'estimation de flux optique, comme celui de Farnebäck, 2003 qui offre un bon compromis entre la précision et la rapidité (Wang et al., 2013). Un échantillonnage régulier dans l'espace image et dans le temps permet d'obtenir sur les données vidéos brutes des volumes vidéo où les trois dimensions sont les dimensions de l'image x, y, et le temps t, comme représenté par la Figure 3.3. L'estimation dense est préférée aux stratégies de points d'intérêt comme proposé par Shi et al., 2013.



(a) Séquence d'images formant un volume vidéo. (b) Volumes spatio-temporels.

Figure 3.3 – Grille régulière et dense de volumes spatio-temporels encodant une information de mouvement local et court-terme.

Chaque élément de la grille 3D est un volume spatio-temporel $(\delta x, \delta y, \delta t)$ (STV pour *Spatio-Temporal Volume*). Pour chaque STV, nous calculons un ensemble d'histogrammes de flux optiques (HOF) non normalisés $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n\}$, n étant le nombre d'images dans la fenêtre temporelle δt, et $\mathbf{X}_i \in \mathbb{R}^p$ le vecteur HOF pour la frame i (Figure 3.4). Un histogramme de flux optique compte les vecteurs de flux optique dans chaque intervalle de direction (d'angle). La valeur de δt est importante, car si elle est fixée à une valeur trop petite, on va toujours se ramener au cas de mouvements avec des changements de direction et d'amplitude faibles ou nuls. Cette valeur est déterminée expérimentalement pour pouvoir distinguer les deux catégories de mouvement. Afin de rendre la description du mouvement invariante à la distance d'observation, les vecteurs HOF sont normalisés en utilisant une norme L_2. Les vecteurs normalisés pour chaque image i du volume sont notés $\mathbf{X}_i^{\bar{2}} = \mathbf{X}_i / ||\mathbf{X}_i||_2$. Pour capturer les mouvements à différentes échelles, on considère des volumes spatio-temporels de plusieurs tailles. Enfin, la moyenne des vecteurs et la matrice de covariance sont calculés et constituent les statistiques du mouvement dans le volume. Les différentes étapes du calcul de ces statistiques dans un volume spatio-temporel sont représentées dans la Figure 3.4.

Descripteur invariant à la rotation

Les histogrammes de flux optique sont des discrétisations des distributions continues du champ de mouvement selon la direction θ. Le produit scalaire de deux distributions f et g du champ de mouvement, 2π périodiques, ne change pas lorsqu'on leur applique un même décalage en angle η. Ainsi, $\langle f, g \rangle = \langle f', g' \rangle$, pour $f'(\theta) = f(\theta - \eta)$ et $g'(\theta) = g(\theta - \eta)$, $\forall \eta$. Une rotation dans le plan image se traduit par un décalage en angle. Dans le cas discret, les vecteurs HOF sont des approximations des vraies distributions $\mathbf{X}_i \simeq f, \mathbf{Y}_i \simeq g$. Par conséquent, on considère qu'un décalage en angle ne change que peu le produit scalaire entre deux vecteurs HOF : $\langle \mathbf{X}_i, \mathbf{Y}_i \rangle \simeq \langle f, g \rangle$ et le produit scalaire entre les vecteurs auxquels on applique un décalage d'angle η est similaire, $\langle \mathbf{X}_i, \mathbf{Y}_i \rangle \simeq \langle \mathbf{X}_i + \eta, \mathbf{Y}_i + \eta \rangle$, la précision de l'approximation dépendant de la discrétisation employée.

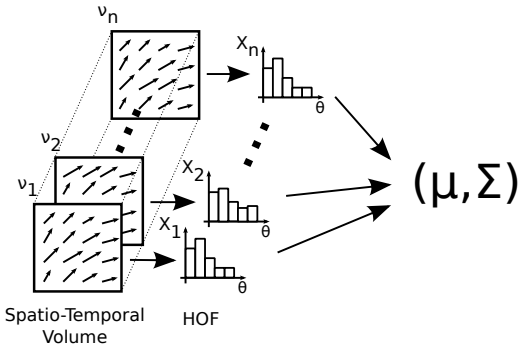


FIGURE 3.4 – Calcul des statistiques du mouvement dans un volume spatio-temporel. Un histograme de flux optique (HOF) est calculé à chaque image du volume. Le vecteur moyenne μ et la matrice de covariance Σ de ces histogrammes sont ensuite estimés.

Le produit scalaire entre deux vecteurs unitaires étant une mesure de l'angle formé par ces vecteurs, une rotation dans l'image provoque un décalage en angle sur les vecteurs HOF qui ne modifie pas l'angle relatif entre eux. L'espace des vecteurs HOF normalisées avec la norme $\mathcal{L}_2$ de dimension $\mathbb{R}^p$ est une sous-variété riemanienne représentée par l'ensemble des points d'une sphère unitaire $\mathbb{S}_+^{p-1} = \{x \in \mathbb{R}_+^p : x^T x = 1\}$ sur laquelle la distance entre deux points est proportionnelle à l'angle entre les deux vecteurs correspondants. Dans cette variété, les valeurs propres de la matrice de covariance des vecteurs HOF normalisés $\mathbf{\Sigma}_{\bar{\mathbf{X}}^2} = cov(\bar{\mathbf{X}}^2)$ sont conservés si on applique un décalage en angle sur tous les vecteurs.

Le descripteur RIMOC est formé par les valeurs propres de la matrice de covariance $\mathbf{\Sigma}_{\bar{\mathbf{X}}^2}$:

$$\boldsymbol{\lambda} = [\lambda_1, \lambda_2, ..., \lambda_{p-1}] \quad \text{avec } \lambda_1 > \lambda_2 > ... > \lambda_{p-1} \tag{3.1}$$

Cette représentation est beaucoup plus concise que le descripteur cov3D (SANIN et al., 2013) qui est la covariance des vecteurs de gradient et de flux optique.

Le calcul des covariances dans la variété $\mathbb{S}_+^{p-1}$ autour d'un point x_0 peut être fait dans un espace linéaire de même dimension qui est le plan tangent à $\mathbb{S}_+^{p-1}$ en x_0. En pratique, dans chaque STV, le calcul est réalisé dans le plan tangent en un point qui est la moyenne des vecteurs HOF.

Interprétation de l'espace des descripteurs RIMOC

Les $p-1$ valeurs propres de la matrice de covariance des vecteurs HOF normalisés permettent de distinguer différentes types de mouvement. Lorsque le mouvement ne varie pas dans le temps (absence de mouvement ou mouvement constant), la covariance sera essentiellement liée au bruit aléatoire dans l'image et au bruit d'approximation de la distribution du champ de mouvement. Les variations faibles du mouvement, correspondant a des trajectoires lisses et cohérentes des objets, seront observées sur peu de dimensions dans l'espace de représentation. En revanche, pour des mouvements erratiques et désordonnées, il y aura des variations importantes sur davantage de dimensions.

On distingue donc les cas suivants :

$$\begin{aligned} \lambda_1, ..., \lambda_{p-1} \sim 0 &\quad \Rightarrow \quad \text{Absence de mouvement ou mouvement constant.} \\ \lambda_1 \text{ grand, } \lambda_2, ..., \lambda_{p-1} \sim 0 &\quad \Rightarrow \quad \text{Variation légère du mouvement.} \\ \lambda_1, ..., \lambda_{m-1} \text{ grand, } \lambda_m, ..., \lambda_{p-1} \sim 0 &\quad \Rightarrow \quad \text{Variation brutale du mouvement.} \end{aligned}$$

A titre d'illustration, nous avons représenté dans la Figure 3.5, pour trois situations différentes (déplacement régulier dans deux directions et situation de violence) la distribution des valeurs

propres de $\mathbf{\Sigma}_{\mathbf{X}^{\bar{2}}}$ correspondantes dans un STV. Cet exemple montre bien la différence de la distribution des valeurs propres selon les types de mouvement, et la pertinence du descripteur RIMOC pour distinguer les mouvement chaotiques des mouvements réguliers.

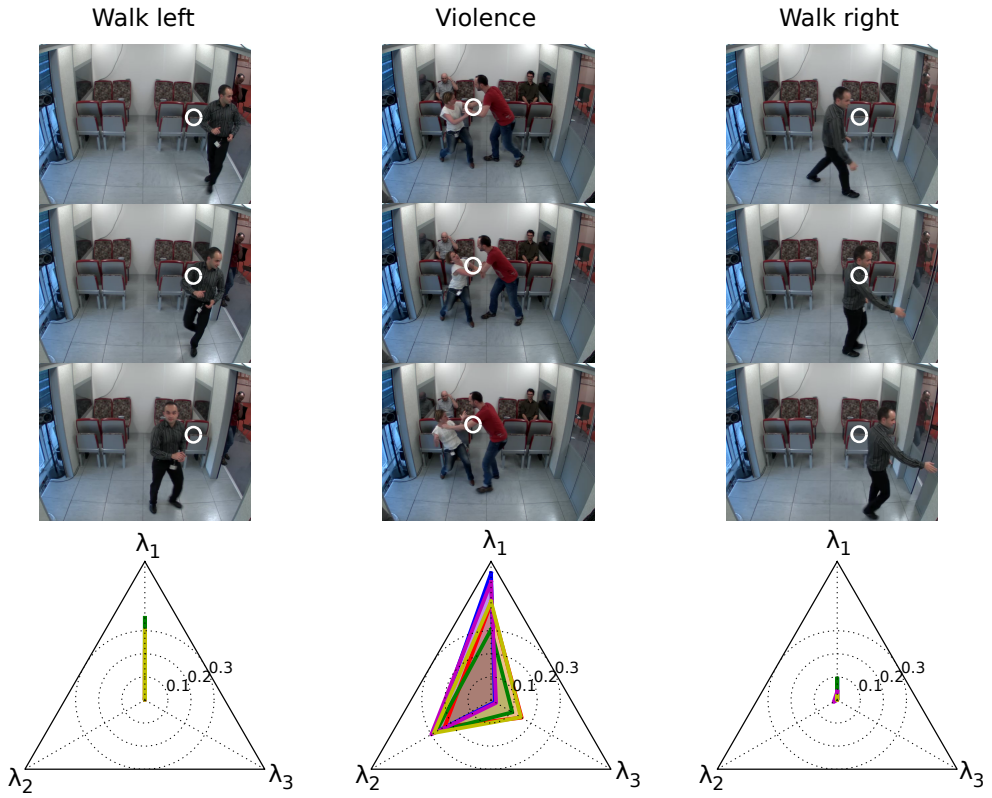


FIGURE 3.5 – Visualisation des 3 premières valeurs propres de la matrice de covariance des vectors HOF dans un STV, pour 3 situations différentes : i) personne marchant vers la gauche, ii) scénario d'agression, iii) personne marchant vers la droite. Le cercle blanc indique la position du STV dans l'image, où les valeurs propres sont calculées. On constate que pour les déplacements réguliers, les valeurs propres sont faibles ou peu dispersées, alors que pour le scénario d'agression, l'énergie du mouvement est beaucoup plus dispersée dans plusieurs directions.

3.3.3 Détection des événements anormaux

Modèle d'ensemble de volumes spatio-temporels et méthode de détection

L'information de mouvement contenue dans un volume spatio-temporel est cependant très locale. Afin de mieux représenter les interactions entre les personnes, nous étendons la modélisation du mouvement dans des régions spatio-temporelles plus grandes, en utilisant des compositions de STV, comme dans (BOIMAN et IRANI, 2005) et en les comparant entre elles en calculant des erreurs de reconstruction. Notre approche est basée sur celle de ROSHTKHARI et LEVINE, 2013b qui propose une manière de stocker l'information efficacement et d'effectuer des inférences rapides sur les nouvelles données.

Tout d'abord, les volumes spatio-temporels similaires sont regroupés pour former un dictionnaire à l'aide d'une version de calcul en ligne de l'algorithme de clustering des K-moyennes appliqué sur les descripteurs RIMOC $\boldsymbol{\lambda}$. Chaque mot (ou code) du dictionnaire est le représentant d'un cluster. La probabilité a posteriori d'affecter un nouveau volume représenté par $\boldsymbol{\lambda}$ à

un code c du dictionnaire est égale à :

$$P(c|\boldsymbol{\lambda}) = \frac{1}{\sum_i \frac{1}{D(c_i,\boldsymbol{\lambda})}} \times \frac{1}{D(c,\boldsymbol{\lambda})}, \tag{3.2}$$

où $D(c,\boldsymbol{\lambda})$ est la distance euclidienne entre un code et le descripteur $\boldsymbol{\lambda}$.

On considère à présent des ensembles de STV comme des configurations de mots placés dans une configuration spatio-temporelle spécifique. Chaque STV est caractérisé par un descripteur $\boldsymbol{\lambda}$ et représenté par le mot le plus proche dans le dictionnaire. Les positions relatives des mots de l'ensemble de STV par rapport au STV central sont calculées et stockées dans la variable δ. La configuration d'un ensemble de STV s'apparente à un graphe en étoile où le nœud central α est le STV de référence (Figure 3.6). Dans ce modèle, les nœuds sont caractérisés par un terme de similitude visuelle $P(c|\boldsymbol{\lambda})$ avec les mots du dictionnaire alors que la distribution des positions relatives pour chaque paire de mots $\{c, c_\alpha\}$ définit la structure de l'ensemble.

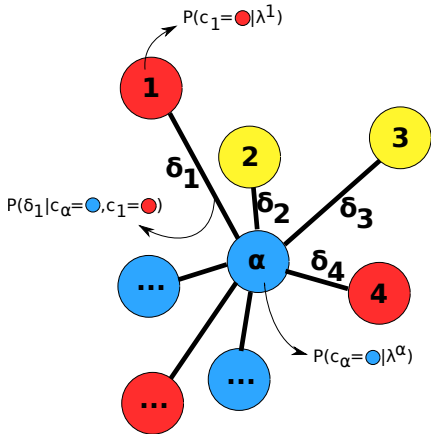


FIGURE 3.6 – Ensemble de STV autour du STV central α. δ code la position relative des STV. La couleur de chaque STV représente le mot associé au descripteur $\boldsymbol{\lambda}$. Le modèle d'ensemble se compose d'un modèle d'apparence $P(c|\boldsymbol{\lambda})$ pour chaque STV et d'un modèle structurel $P(\delta|c_\alpha, c)$ pour chaque paire STV central/STV voisin.

Pendant l'apprentissage sur des données de situations normales, on crée un ensemble d'hypothèses de configurations d'ensembles de STV sous la forme :

$$\mathcal{H} = \bigcup_{\substack{c\in\mathcal{C}\\ c_\alpha\in\mathcal{C}}} \{(\delta, c_\alpha, c)\},$$

A l'inférence, on cherche à évaluer la similarité de nouveaux ensembles observés (requête) avec les ensembles observés à l'apprentissage et modélisés par $\mathcal{H}$. La requête est représentée par un ensemble $\mathcal{E} = \{\boldsymbol{\Delta}^k, \boldsymbol{\lambda}^\alpha, \boldsymbol{\lambda}^k\}_{k=1}^K$, où $\boldsymbol{\Delta}^k$ est la position relative du k^{ime} STV. La similarité de la requête avec les hypothèses $\mathcal{H}$ du modèle appris est obtenue en maximisant la probabilité a posteriori d'affecter la requête à une hypothèse sur l'ensemble des hypothèses :

$$\max_{\mathcal{H}} P(\mathcal{H}|\mathcal{E}) = \max_{\substack{c\in\mathcal{C}\\ c_\alpha\in\mathcal{C}}} P(\delta, c_\alpha, c|\mathcal{E}), \tag{3.3}$$

En supposant l'indépendance statistique des STV, cette recherche s'écrit aussi sous la forme :

$$\max_{\substack{c\in\mathcal{C}\\ c_\alpha\in\mathcal{C}}} P(\delta, c_\alpha, c|\mathcal{E}) = \max_{\substack{c\in\mathcal{C}\\ c_\alpha\in\mathcal{C}}} \prod_{k=1}^{K} P(\delta_k|c_\alpha, c) P(c|\boldsymbol{\lambda}^k) P(c_\alpha|\boldsymbol{\lambda}^\alpha). \tag{3.4}$$

En effectuant l'inférence à chaque pixel de l'image, on obtient une carte de similarité spatio-temporelle qui est seuillée pour détecter les régions où la structure de mouvement correspond à un évènement anormal (similarité avec le modèle faible).

Analyse multi-échelles

A la manière des méthodes de détection d'objets, les ensembles de STV sont appris à plusieurs échelles pour adresser différentes distances d'observation, en profitant du fait que le flux optique est déjà calculé à plusieurs résolutions d'image (Farnebäck, 2003). De manière plus simple que dans la méthode de Roshtkhari et Levine, 2013b, un ensemble de STV n'inclut pas d'éléments de différentes tailles, ce qui permet de bien différencier les échelles d'observation et de ne pas diluer les faibles scores de similarité dans un ensemble avec des volumes de taille différente. L'échelle la plus grande est fixée par rapport à la taille de la personne de plus grande taille dans l'image, les échelles inférieures réduisent progressivement cette taille. Un avantage de l'approche multi-échelles est qu'un modèle pré-entraîné pourrait être exploité en inférence dans des contextes un peu différents, même si les distances d'observations varient légèrement. La stratégie proposée consiste à effectuer la détection de manière indépendante à plusieurs échelles et de réaliser une fusion tardive des scores de détection en minimisant la réponse sur l'ensemble des échelles :

$$\mathrm{Sim}_{\mathcal{E}} = \min_{Sc} \left(\max_{\mathcal{H}} P\left(\mathcal{H}|\mathcal{E}_{Sc}\right) \right), \tag{3.5}$$

avec $\mathcal{E}_{Sc}$ l'ensemble requête calculé à l'échelle Sc. Le score le plus bas est retenu, c'est-à-dire celui qui correspond au plus grand écart avec le modèle. Un exemple de détection multi-échelles est montré dans la Figure 3.7.

3.3.4 Evaluation expérimentale de la méthode

Comparaison avec une méthode de classification binaire par apprentissage supervisé

Afin de juger des performances atteintes avec notre approche probabiliste d'apprentissage "une-classe", nous la comparons à une classification binaire par apprentissage supervisé en suivant la méthode de référence de Wang et Schmid, 2013 conçue originellement pour la reconnaissance d'actions. Cette approche met en œuvre des descripteurs de trajectoires denses estimées sur une grille à différentes échelles. La classification est effectuée ensuite par un algorithme de SVM non linéaire avec une fonction noyau de type RBF :

$$K(x_i, x_j) = \exp\left(-\sum_c \frac{1}{A^c} D(x_i^c, x_j^c) \right),$$

où $D(x_i^c, x_j^c)$ est la distance entre les descripteurs pour les vidéos i et j par rapport à la composante c du descripteur et A^c une constante de normalisation. Nous transformons le problème de classification en détection en ajoutant une localisation temporelle dans des fenêtre d'analyse de différentes tailles, comme dans (Oneata et al., 2013). La meilleure réponse à travers les échelles temporelles est retenue.

Ensemble de données d'évaluation

Au moment de ces travaux, il existait plusieurs ensembles de données vidéo pour la reconnaissance d'actions, et certains d'entre, eux assez généraux, incluent quelques scénarios de com-

FIGURE 3.7 – Exemple de détection à 3 échelles différentes (images du bas) et résultat de la fusion (image du haut), sur un scénario d'agression observé à une distance relativement grande de la caméra. Les valeurs de similarité sont représentées avec des couleurs allant du vert vers le rouge (ordre décroissant de similarité). A noter que les valeurs de similarité les plus faibles, correspondant à une plus grande probabilité d'événement anormal, sont obtenues à l'échelle la plus petite, la plus adaptée à la taille dans l'image des personnes impliquées dans le scénario d'agression.

portement agressifs (Behave [1], CAVIAR [2], Hollywood2 [3], SDHA(2010) [4] and UCF101 [5]), d'autres proposent des corpus de scènes violentes dans des contextes spécifiques de cinéma (MediaEval Violent Scenes [6]), ou de sport (Hockey fights [7]). La base ViolentFlows [8] comporte des scènes de violence dans la foule, mais les clips vidéo de quelques secondes sont pré-segmentés pour la tâche de classification. D'autres ensembles comportent des séquences vidéo pour la détection d'événements anormaux dans des scénarios très variés qui ne sont pas forcément liés aux événements violents (ZAHARESCU et WILDES, 2010).

Nos travaux sont motivés d'un point de vue applicatif par le développement d'une solution de sécurisation des trains (projet FUI Degiv). Aucune base existante ne répondant complètement au besoin, nous avons créé un grand ensemble de données à partir : i) de vidéos collectées sur Youtube, ii) des vidéos du dataset de l'université d'Amsterdam (Train Station), et iii) des vidéos que que nous avons acquises à bord d'une fausse voiture de train en laboratoire, et dans des conditions réelles d'un métro parisien en fonctionnement, en mettant en scène des situations d'hostilité, d'agression, de rixe, avec des intensités et des densités de personnes variées. La Figure 3.8 donne les caractéristiques de l'ensemble de données vidéo créé. Les détails sur le jeu de

1. http ://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/
2. http ://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/
3. http ://www.di.ens.fr/ laptev/actions/hollywood2/
4. http ://cvrc.ece.utexas.edu/SDHA2010/
5. http ://crcv.ucf.edu/data/UCF101.php
6. http ://www.multimediaeval.org/mediaeval2014/violence2014/
7. http ://visilab.etsii.uclm.es/personas/oscar/FightDetection/index.html
8. http ://www.openu.ac.il/home/hassner/data/violentflows/

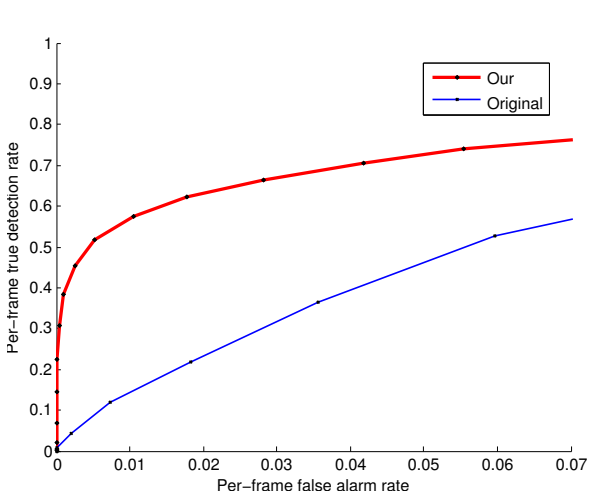
données sont donnés dans la publication (Ribeiro et al., 2016). Pour compléter la base d'événements normaux, des séquences de l'ensemble i-LIDS [9] prises dans un aéroport et dans le métro londonien ont été sélectionnées. Ces séquences montrent des situations relativement complexes.

	Fake train		Real train		Train station	Real life context
	Normal	Violence	Normal + Violence			
	A1	A2	Validation	Test		Qualitative
#Frames	70290	495930	542790	327600	25629	45550
Duration (hours)	0.66	5.025	5	7	0.35	0.4
#Aggressive events	0	116	116	31	33	13

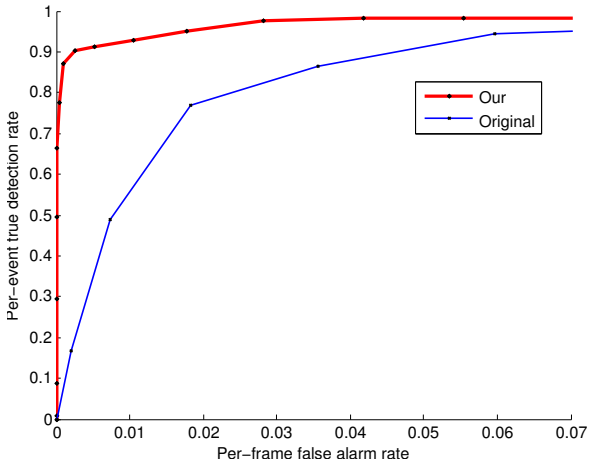
Figure 3.8 – Caractéristiques de l'ensemble de données créé pour les besoins de l'étude, et partitions utilisées pour les différentes expériences.

Résultats expérimentaux

Les hyper-paramètres de la méthode, tels que le nombre de bins des HOF, la taille des volumes spatio-temporels, la taille du dictionnaire, la taille des ensembles de STV, sont déterminés expérimentalement sur une base de validation de l'ensemble *Fake train*. Avec ces hyper-paramètres fixés, nous comparons sur le même ensemble de validation la performance du descripteur RIMOC à celle obtenue avec le descripteur utilisé dans (Roshtkhari et Levine, 2013b) dans un cadre de détection d'événements anormaux (Figure 3.9). Les courbes ROC sont tracées avec deux protocoles distincts : une évaluation de la détection par image, et une évaluation de la détection par événement. Pour le deuxième protocole, un événement est détecté si le pourcentage d'images où la détection a lieu parmi toutes les images annotées avec un événement anormal (violent) est supérieur à un certain seuil de recouvrement.



(a) Courbe ROC évaluée par image.



(b) Courbe ROC évaluée par événement (10% de recouvrement).

Figure 3.9 – Comparaison sur l'ensemble de validation de notre approche RIMOC avec l'approche de Roshtkhari et Levine, 2013b désignée ici par "original" pour la détection d'événements anormaux (violents). L'apprentissage est réalisée sur l'ensemble A1.

Nous comparons ensuite les performances de notre méthode par rapport à l'approche de classification binaire supervisée (notée S) décrite en 3.3.4 sur l'ensemble de validation (Figure 3.10),

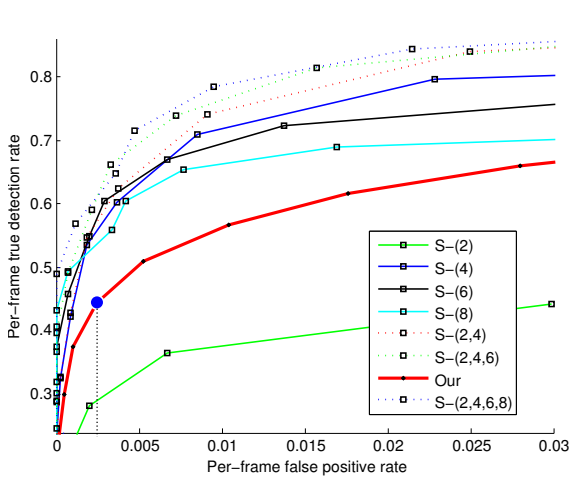
9. i-LIDS Multiple Camera Tracking and Abandoned Baggage Detection scenarios, CAST, United Kingdom's Home Office : www.ilids.co.uk and www.eecs.qmul.ac.uk/ andrea/avss2007_d.html

puis sur l'ensemble de test (Figure 3.11). Pour entraîner le classifieur binaire, nous utilisons l'ensemble d'apprentissage A1+A2. Le but de cette comparaison est d'une part de valider l'espace de description RIMOC et d'autre part de comparer la méthode d'apprentissage "une-classe" pour la détection d'événements anormaux (violents), avec un apprentissage complètement supervisé sur deux classes (normal et violence).

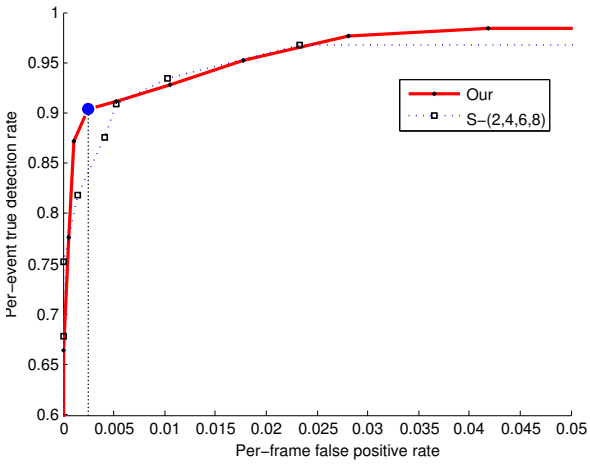
Sur l'ensemble de validation, on remarque que la méthode supervisée sur deux classes a des performances variables selon la fenêtre temporelle utilisée, les meilleures performances étant obtenues avec une version combinant toutes les échelles temporelles ($S-(2,4,6,8)$). C'est cette version qui est retenue pour les comparaisons ultérieures. Notre approche RIMOC obtient de moins bonnes performances que S avec le protocole d'évaluation de la détection par image, mais des performances équivalentes lorsque l'on compte par événement. Pour des taux de fausses alarmes bas, elle est même légèrement meilleure.

Sur l'ensemble de test, l'approche supervisée S a de meilleures performances quand on évalue image par image. L'évaluation par événement montre en revanche que l'approche RIMOC pour la détection d'événements anormaux est globalement meilleure que S, sauf pour les taux de faux positifs très bas.

Ces résultats confirment que la modélisation du mouvement avec le descripteur RIMOC combinée à un apprentissage sur la seule classe des situations normales, permettent de concurrencer une méthode supervisée de l'état de l'art nécessite des données annotées de situations de violence à l'apprentissage.



(a) Courbe ROC évaluée par image.



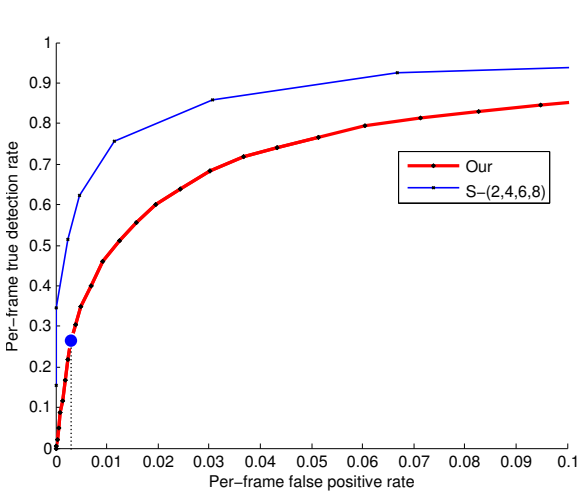
(b) Courbe ROC évaluée par événement (10% de recouvrement).

FIGURE 3.10 – Comparaison sur l'ensemble de validation de notre approche RIMOC de détection d'événements anormaux avec la méthode de classification binaire par apprentissage supervisé sur A1+A2 suivant la méthode de WANG et SCHMID, 2013 (notée S). La méthode S est testée pour différentes fenêtres temporelles de {2,4,6,8} secondes.

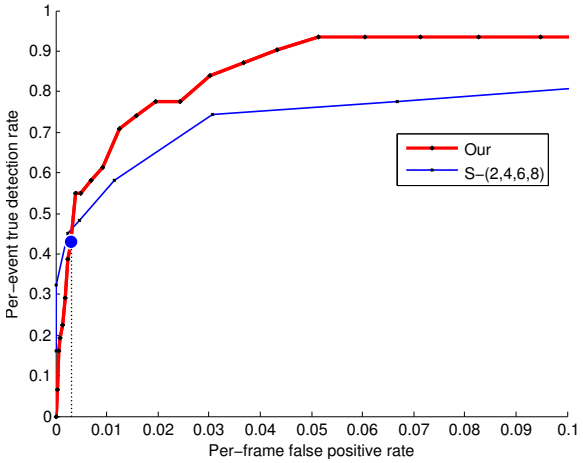
Nous montrons quelques résultats qualitatifs de détection sur des séquence de *Real train* (Figure 3.12) et de *Real life context* (Figure 3.13). Pour les séquences de train roulant, on constate que la méthode résiste aux mouvements apparents liés au déplacement du train et aux changements de luminosité, et qu'elle détecte bien les interactions violentes entre les personnes.

Temps de calcul

L'algorithme implémenté en C++ s'exécute sur un CPU i7 quad-core à 2.50GHz à 15 images par seconde, ce qui est largement suffisant pour les applications temps réel.



(a) Courbe ROC évaluée par image



(b) Courbe ROC évaluée par événement (10% de recouvrement).

FIGURE 3.11 – Comparaison sur l'ensemble de test de notre approche RIMOC de détection d'événements anormaux avec la méthode de classification binaire par apprentissage supervisé sur A1+A2 suivant la méthode de WANG et SCHMID, 2013 (notée S).

FIGURE 3.12 – Quelques résultats qualitatifs de détection d'événements violents avec notre approche, sur les données *Real train*. L'ellipse blanche indique une détection d'événement violent.

3.3.5 Conclusion sur la détection d'événements violents

Nous nous sommes attachés dans ces travaux à répondre au problème de détection des situations de violence dans des scènes complexes de vidéosurveillance. Ce problème est abordé sous l'angle de l'apprentissage "une-classe" : l'apprentissage est réalisé sur des données de situations normales uniquement, du fait de la faible occurrence des situations de violence. Un descripteur bas-niveau invariant à la rotation dans l'image et codant de manière concise et locale le mouvement dans l'image, RIMOC, a été proposé. Il est intégré dans un modèle spatio-temporel de régions qui permet de discriminer les mouvements réguliers et cohérents des mouvements déstructurés accompagnant généralement les situations de violence. L'apprentissage et la détection sont formulés dans un cadre probabiliste où les événements violents sont assimilés aux événe-

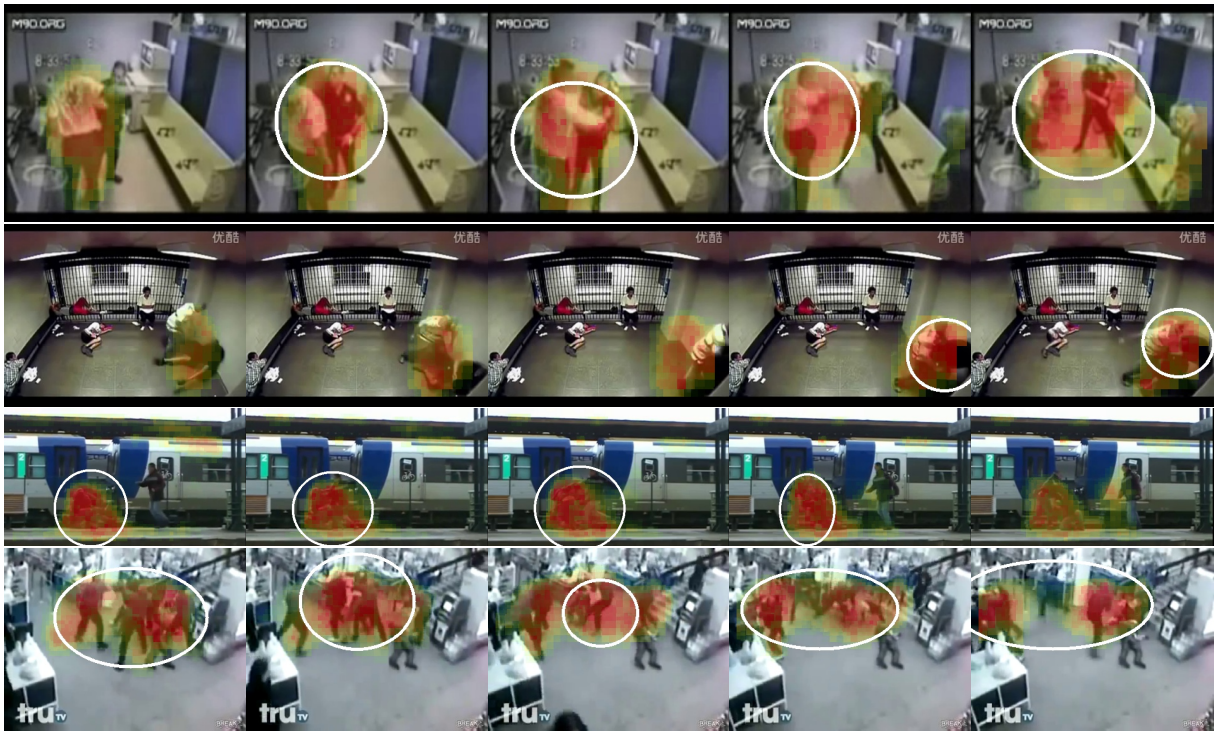


FIGURE 3.13 – Quelques résultats qualitatifs de la détection d'événements d'agression avec notre méthode sur les données *Real life context*. L'ellipse blanche indique une détection d'événement violent.

ments anormaux qui s'écartent du modèle. Les évaluations quantitatives et qualitatives dans des contextes divers valident la pertinence de l'approche proposée. De plus, la rapidité de l'inférence est compatible avec un fonctionnement en temps réel.

Du fait de la modélisation bas niveau et de l'absence d'information explicite sur la localisation des personnes dans l'image et leurs interactions, la robustesse de la méthode pourrait être mise à l'épreuve dans les situations rares où on observe des mouvements désordonnés, sans pour autant qu'elles soient associées à un événement violent impliquant des personnes. Une manière d'implémenter un système de surveillance automatique serait d'utiliser notre approche comme un premier module attentionnel rapide et léger pour filtrer les événements d'intérêt, et de les classifier avec plus de précision en utilisant des informations de plus haut niveau sémantique.

3.4 Descripteurs de niveau intermédiaire pour l'analyse de scènes de forte densité

Cette section aborde le problème de la caractérisation des comportements de personnes dans des contextes de forte densité. Nous présentons ici des travaux réalisés sur la conception d'un ensemble de descripteurs visuels de niveau sémantique intermédiaire destinés l'estimation des propriétés d'une foule. Ces descripteurs s'appuient sur un modèle de dynamique locale de la foule évalué à partir de descripteurs bas niveau de mouvement (trajectoires courtes de points d'intérêt dans la scène, encore appelées *tracklets*) et d'un modèle spatial graphique qui relie les *tracklets* pour représenter les interactions entre les personnes. L'intérêt des descripteurs de niveau intermédiaire proposés est validée dans deux applications : la classification de vidéo de foule et la détection de situations anormales dans les foules.

3.4.1 Modélisation spatio-temporelle de la foule

Notre modèle matérialise les mouvements des individus au sein d'une foule sous la forme d'un graphe évoluant au cours du temps. Les nœuds du graphe sont un ensemble de descripteurs locaux connectés grâce à une triangulation de Delaunay, et associés dans le temps par un suivi visuel de ces descripteurs. Nous utilisons un détecteur de points d'intérêt tel que FAST (Rosten et al., 2010), déjà éprouvé dans des travaux d'analyse de foule (Butenuth et al., 2011), pour extraire des points sur les personnes. Une étude comparative (Fradi et Dugelay, 2015b) montre l'efficacité des points extraits avec FAST par rapport à d'autres méthodes dans le contexte de l'analyse de foule. Une étude plus récente (Bojanic et al., 2019) montre même que les détecteurs de points et descripteurs locaux "historiques" restent très compétitifs dans des tâches d'appariement d'images face à leurs concurrents "deep".

Naturellement, ce type de détecteurs ne garantissent pas de maintenir ni les mêmes points physiques, ni le même nombre de points sur les personnes, mais nous supposons que, étant la grande distance d'observation des scènes de foule, la densité des points extraits dans l'image est représentative de la densité de personnes. Notons qu'il est possible de produire un nombre de points plus cohérent en adaptant le seuil de détection avec l'échelle d'observation grâce à l'usage d'un modèle géométrique simple reliant la distance à la caméra à la position dans l'image.

Les points détectés sont appariés et suivis au cours du temps avec une stratégie améliorée par rapport à la méthode de KLT (Shi et al., 2010) et (Fradi et Dugelay, 2015a) : non seulement les points sont suivis dans les directions directes et inverses pour obtenir des trajectoires plus cohérentes, mais les pertes de suivi peuvent être rattrapés en appariant de nouveaux points détectés par ré-identification. Les trajectoires qui ne sont pas mises à jour depuis un certain temps, et les points statiques sont filtrés, afin de ne garder que l'information de mouvement de la foule. Ces améliorations permettent d'obtenir des trajectoires plus longues et moins sujettes au bruit. Ces trajectoires de points sont appelées *tracklets* dans ce qui suit.

On note pour chaque instant k les m_k trajectoires filtrées parmi les n_k trajectoires initiales :

$$\mathcal{T}_k = \{T_1^k, ..., T_{m_k}^k | T_i^k = \{\mathbf{x_i}(k - \Delta t_i^k), ..., \mathbf{x_i}(k)\}\} \tag{3.6}$$

où Δt_i^k est l'intervalle de temps depuis la créaton de la trajectoire T_i^k et l'instant courant, cet intervalle est borné à L images consécutives. $\mathbf{x_i}(k)$ est la position dans l'image du point suivi au temps k.

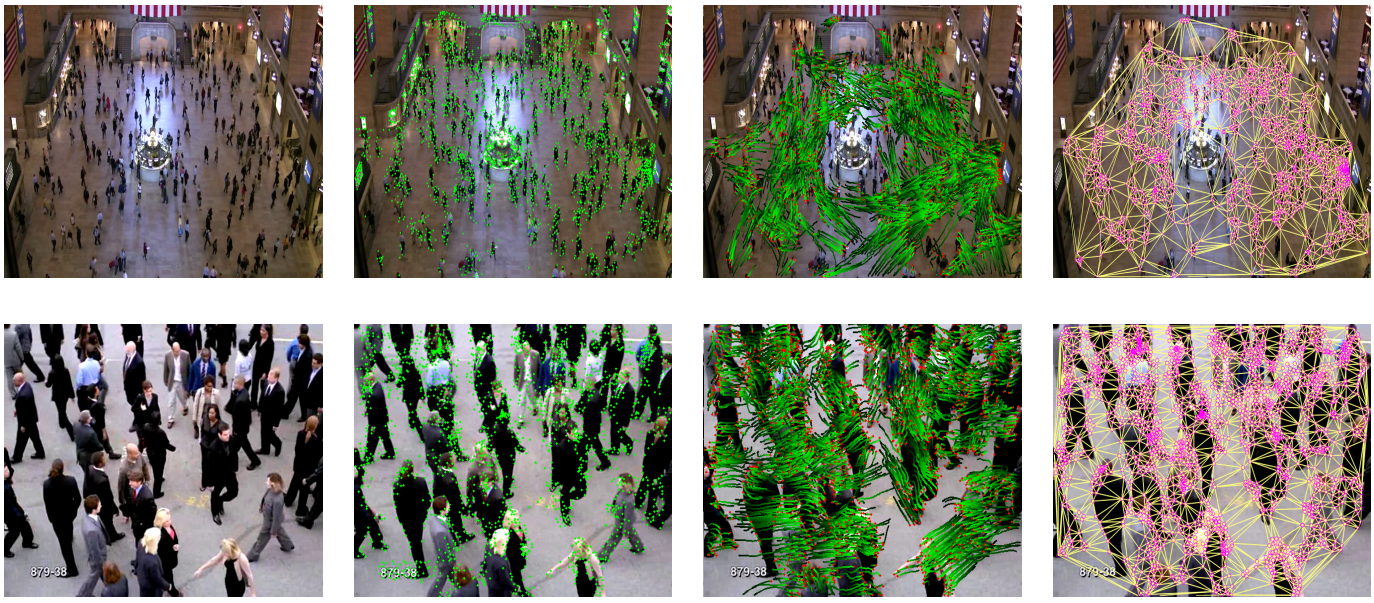
Un graphe est défini par la triangulation de Delaunay permettant de relier les points suivis au cours du temps. On note ce graphe $\mathcal{G}^k(\mathcal{V}^k, \mathcal{E}^k, \mathcal{F}^k)$, où $\mathcal{V}^k = \{V_1^k, ..., V_{m_k}^k | V_i^k = \mathbf{x_i}(k)\}$ sont les nœuds représentant les points d'intérêt, $\mathcal{E}^k$ l'ensemble des arêtes reliant ces nœuds et $\mathcal{F}^k$ l'ensemble des faces issues de la triangulation. La triangulation de Delaunay est préférée à la méthode de génération de graphes reliant les k plus proches voisins dans un groupe, comme dans (Shao et al., 2014 ; Zhang et al., 2013b), car elle produit un modèle global d'interaction (des nœuds distants pouvant être connectés) et une partition de l'espace relativement régulière. Les points étant suivis au cours du temps, l'étude des déformations locales du graphe nous permet d'estimer des propriétés de la foule relatives aux mouvements des individus et à leurs interactions.

Les différentes étapes d'extraction, de suivi des points d'intérêt, et de calcul du graphe sont illustrées dans la Figure 3.14 sur deux exemples.

Dans le graphe $\mathcal{G}$ ainsi défini, on considère pour modéliser des interactions plus ou moins locales, des relations de voisinage entre nœuds. Le voisinage d'ordre 1 pour un point V_i^k est défini par l'ensemble des nœuds voisins directs de V_i^k :

$$\mathcal{N}(V_i^k) = \{V_i^k\} \cup \{V_j^k, \forall (V_i^k, V_j^k) \in \mathcal{E}^k\} \tag{3.7}$$

Le voisinage d'ordre n de V_i^k est l'ensemble des voisins indirects définis de la manière suivante :



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

FIGURE 3.14 – Illustration de la représentation spatio-temporelle proposée, combinant des trajectoires de points d'intérêt et un modèle d'interaction qui est un graphe reliant ces points, sur deux exemples de l'ensemble CUHK (SHAO et al., 2014) (a) Image originale, (b) Points d'intérêt extraits avec le détecteur FAST (c) Trajectoires des points suivis avec une version modifiée de l'algorithme KLT (d) Modèle spatial d'interaction représenté par un graphe issu de la triangulation de Delaunay sur l'ensemble des points d'intérêt.

$$\mathcal{N}_n(V_i^k) = \mathcal{N}_{n-1}(V_i^k) \cup \{\mathcal{N}_1(V_j^k), \forall V_j^k \in \mathcal{N}_{n-1} \setminus \mathcal{N}_{n-2}\} \tag{3.8}$$

avec $\mathcal{N}_0(V_i^k) = \{V_i^k\}$ et $\mathcal{N}_1(V_i^k) = \mathcal{N}(V_i^k)$.

On note alors $C(V_i^k)$ le sous-graphe de voisinage d'ordre 1 du nœud V_i^k formé par les nœuds $\mathcal{N}(V_i^k)$ et les arêtes appartenant à $\mathcal{E}^k$ qui les relient entre eux. De façon similaire, $C_n(V_i^k)$ est le sous-graphe de voisinage d'ordre n du nœud V_i^k formé par les les nœuds $\mathcal{N}_n(V_i^k)$ et les arêtes appartenant à $\mathcal{E}^k$ qui les relient entre eux.

3.4.2 Ensemble de descripteurs locaux de niveau intermédiaire

Nous décrivons ici un ensemble complet de descripteurs destinés à représenter les différentes caractéristiques du comportement des personnes dans une foule. Ces descripteurs sont calculés sur la base du modèle spatio-temporel décrit précédemment. Nous distinguons les descripteurs de comportements individuels et les descripteurs d'interactions entre personnes.

Modélisation des comportements individuels

Les descripteurs de comportements individuels sont calculés sur les nœuds V_i^k sans tenir compte de leurs voisins.

Vitesse de déplacement La vitesse de déplacement locale est estimée en calculant le déplacement d'un nœud V_i^k dans une fenêtre temporelle τ_1 en suivant la *tracklet* associée.

Elle est donnée par l'expression suivante :

$$D^{veloc}(V_i^k) = \frac{1}{\tau_1}.\|\overrightarrow{V_i^{k-\tau_1}V_i^k}\|_2 \tag{3.9}$$

Seuls les déplacements au-dessus d'un seuil sont retenus et ils sont normalisés en tenant compte de la variation d'échelle liée à la perspective de la scène dans l'image. La constante τ_1 est adaptée à la fréquence d'acquisition de la vidéo.

Variation de la direction du flux Le second descripteur de comportement individuel concerne la modélisation de la direction du mouvement. La direction des mouvements dans la foule étant dépendante des scènes, nous proposons plutôt de capturer la variation de cette direction dans les *tracklets*, dans le but de séparer des prototypes de mouvements (mouvements lisses ou chaotiques). En découpant les trajectoires en F segments $\{S_i^k, S_i^{k-\tau_2}, ..., S_i^{k-(F-1)\tau_2}\}$ de durée τ_2 adaptée à la fréquence d'acquisition de la vidéo, nous définissons le descripteur suivant :

$$D^{var}(V_i^k) = \frac{1}{F}.\sum_{f=0}^{F-2} d_\theta(S_i^{k-f\tau_2}, S_i^{k-(f+1)\tau_2}) \tag{3.10}$$

où $S_i^j = \overrightarrow{V_i^{j-\tau_2}V_i^j}$.
$d_\theta(\mathbf{a},\mathbf{b}) = \Delta_\theta(|\theta(\mathbf{a}) - \theta(\mathbf{b})|)$ est une mesure de l'angle entre deux vecteurs $\mathbf{a}$ et $\mathbf{b}$, $\theta(\mathbf{v})$ l'angle entre le vecteur $\mathbf{v}$ et l'axe horizontal. Par convention,

$$\Delta_\theta(\alpha) = \begin{cases} \alpha, & \text{si } \alpha < \pi \\ 2\pi - \alpha, & \text{sinon} \end{cases}$$

La Figure 3.15 illustre le segmentation des trajectoires de points suivis en fragments et l'estimation de la direction du mouvement dans ces fragments.

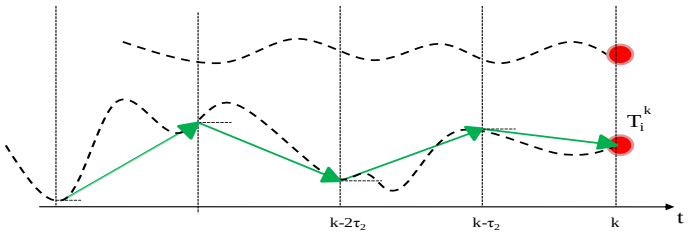


FIGURE 3.15 – Découpage des trajectoires de points en segments pour le calcul de la variation de direction du mouvement. Le descripteur $D^{var}(V_i^k)$ accumule les différences d'angles entre les vecteurs (en vert) de segments consécutifs.

Modélisation globale et modélisation des interactions entre individus

Aux deux descripteurs de modélisation individuelle, nous ajoutons trois descripteurs caractérisant des propriétés spatio-temporelles et deux descripteurs codant des propriétés spatiales dans la scène. Ces descripteurs sont assez largement inspirés de ceux de SHAO et al., 2014, mais diffèrent par leur formulation fondée sur le graphe de Delaunay qui permet une estimation locale.

Stabilité La stabilité dans une foule est caractérisée par le fait que le graphe qui la représente varie peu au cours du temps, autrement dit que sa topologie reste similaire et que la distance entre nœuds voisins varie peu. Pour exprimer cette propriété, nous définissons le

descripteur suivant qui estime la déformation du sous-graphe de voisinage d'un nœud à deux instants différents k et $k-\tau_2$ en calculant la distance :

$$D^{stab}(V_i^k) = dist_g(C_n(V_i^k), C_n(V_i^{k-\tau_2})) \tag{3.11}$$

L'appariement des sous-graphes de voisinage d'un point est assuré par le suivi des points dans le temps, en exploitant les *tracklets*, comme représenté dans la Figure 3.16.

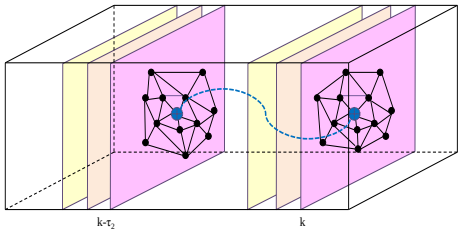


FIGURE 3.16 – Mise en correspondance de sous-graphes de voisinage de points appariés au cours du temps au moyen des *tracklets*.

Pour calculer la distance entre les sous-graphes de voisinage d'un point à deux instants différents, on considère l'ensemble des triangles orientés $R_n(V_i^{t_1})$ et $R_n(V_i^{t_2})$ qui les composent. La distance entre deux triangles $r_{i\alpha}$ et $r_{i\beta}$ est définie par le produit entre la différence d'aire et la différence du birapport entre les deux triangles :

$$g(r_{i\alpha}, r_{i\beta}) = \|a_{i\alpha} - a_{i\beta}\|.\|c_{i\alpha} - c_{i\beta}\| \tag{3.12}$$

où $a_{i\alpha}$ et $c_{i\alpha}$ sont l'aire et le birapport défini sur le triangle $i\alpha$ tel que défini dans (SHIN et TJAHJADI, 2008). La distance entre les sous-graphes se calcule alors en moyennant les distances entre triangles appariés grâce aux *tracklets* et les minima des distances entre triangles non appariés :

$$\begin{aligned} dist_g(C_n(V_i^{t_1}), C_n(V_i^{t_2})) = \frac{1}{|C_n(V_i^{t_1})|} \sum_{\substack{r_{i\alpha_1} \in R_n(V_i^{t_1}) \\ r_{i\beta_1} \in R_n(V_i^{t_2})}} \\ g(r_{i\alpha_1}, r_{i\beta_1}) + \sum_{r_{i\alpha_2} \in R_n(V_i^{t_1})} \min_{r_{i\beta_2} \in R_n(V_i^{t_2})} g(r_{i\alpha_2}, r_{i\beta_2}) \end{aligned} \tag{3.13}$$

Comportement collectif Ce comportement se caractérise par un déplacement cohérent, c'est-à-dire dans la même direction, d'un groupe d'individus. Le descripteur de comportement collectif est calculé localement dans un sous-graphe de voisinage d'un point, en prenant en compte l'angle entre les vecteurs de déplacement des nœuds voisins :

$$D^{collec}(V_i^k) = \frac{1}{|C_n(V_i^k)|} \sum_{V_j^k \in C_n(V_i^k)} h(\overrightarrow{V_i^{k-\tau_1} V_i^k}, \overrightarrow{V_j^{k-\tau_1} V_j^k})$$

avec

$$h(\mathbf{a}, \mathbf{b}) = \begin{cases} d_\theta(\mathbf{a}, \mathbf{b}), & \text{if } d_\theta(\mathbf{a}, \mathbf{b}) < T_1 \\ 0, & \text{otherwise} \end{cases} \tag{3.14}$$

Conflit Le descripteur de conflit évalue au sein d'un sous-graphe de voisinage les interactions potentielles entre personnes proches. Ainsi, en retenant uniquement les voisins dont

le déplacement converge vers le point central V_i^k (ensemble noté $C'(V_i^k)$), on peut estimer le niveau de conflit en V_i^k en calculant la moyenne des différences angulaires de ces déplacements.

$$D^{conf}(V_i^k) = \frac{1}{|C(V_i^k)|} \sum_{V_j^k \in C'(V_i^k)} \frac{d_\theta(\overrightarrow{V_i^{k-\tau_1} V_i^k}, \overrightarrow{V_j^{k-\tau_1} V_j^k})}{\|V_i^k V_j^k\|_2} \tag{3.15}$$

Densité spatiale La densité locale est estimée dans les sous-graphes de voisinage des points, avec un noyau gaussien 2D dont l'écart-type est réglée de façon à englober suffisamment de points et en tenant compte de la distance d'observation (calibrage faible de la perspective de la scène).

Uniformité Le critère d'uniformité donne une caractérisation de la répartition spatiale des personnes dans l'espace. Dans une foule uniforme, les distances entre personnes sont semblables, alors que dans une foule non uniforme, des petits groupes sont formés, se traduisant par une répartition non régulière des points du graphe. Les groupes de personnes sont trouvés en appliquant une méthode de clustering à densité sur les points (Figure 3.17). L'uniformité dans chacun des cluster N_i se calcule alors avec l'expression :

$$D^{unif}(N_i) = \frac{\mathcal{A}(N_i, N_i)}{\mathcal{A}(N, N)} - \left(\frac{\mathcal{A}(N_i, N)}{\mathcal{A}(N, N)}\right)^2 \tag{3.16}$$

$$\text{avec} \begin{cases} \mathcal{A}(N_i, N_i) = \sum\limits_{p \in N_i} \sum\limits_{\substack{q \in C_1(p) \\ q \in N_i}} \frac{1}{\|pq\|_2} \\ \mathcal{A}(N_i, N) = \sum\limits_{p \in N_i} \sum\limits_{\substack{q \in C_1(p) \\ q \notin N_i}} \frac{1}{\|pq\|_2} \\ \mathcal{A}(N, N) = \sum\limits_{i \in \mathcal{N}} \sum\limits_{p \in N_i} \sum\limits_{q \in C_1(p)} \frac{1}{\|pq\|_2} \end{cases}$$

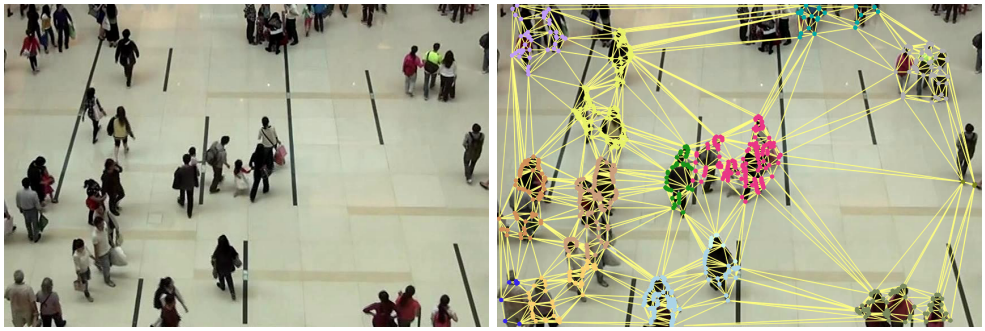


FIGURE 3.17 – Clustering sur les nœuds du graphe de Delaunay pour retrouver les groupes de personnes.

La valeur d'uniformité $D^{unif}(N_i)$ est grande lorsque les points sont espacés de manière régulière dans le cluster, plus faible lorsque des sous-groupes sont formés.

Agrégation des descripteurs

On stocke pour chacun des descripteurs de modélisation individuelle et collective décrits en 3.4.2 et 3.4.2 leur distribution sur des régions de l'image, ou sur l'image entière, sous la forme d'un

histogramme 1-D. Tous les histogrammes sont ensuite concaténés dans un vecteur de description unique.

3.4.3 Résultats expérimentaux

Pour démontrer l'efficacité des descripteurs visuels de foule proposés, nous avons mené des expériences pour trois applications : la classification de vidéos de foule, la détection d'anomalies et la détection de situations de violence dans les scènes denses. Afin de comparer les résultats à ceux de l'état de l'art, nous utilisons trois bases de données académiques :

— la base CUHK (Shao et al., 2014) qui regroupe les événements dans la foule en 8 catégories (voir Tableau 3.1),
— l'ensemble de données *Crowd Abnormal Behaviour* de l'Université du Minnesota (UMN) [10] qui consiste en 11 séquences vidéo mettant en scène des situations normales et des situations anormales (fuite avec dispersion, foule courant dans la même direction),
— Violence in Crowds (Hassner et al., 2012) qui un ensemble de vidéos diverses collectées sur Youtube présentant des événements de violence dans des environnements denses.

La Figure 3.18 donne un aperçu du contenu de ces trois bases de vidéos.

	Crowd events
1	Highly mixed pedestrian walking
2	Crowd following a mainstream and well organized
3	Crowd following a mainstream but poorly organized
4	Crowd merge
5	Crowd split
6	Crowd crossing in opposite directions
7	Intervened escalator traffic
8	Smooth escalator traffic

Tableau 3.1 – Catégories d'événements représentés dans la base CUHK crowd (Shao et al., 2014).

Les valeurs des hyper-paramètres (seuil de détection des points d'intérêt, taille des fenêtres temporelles) utilisés et les détails sur l'évaluation sont donnés dans la publication (Fradi et al., 2017). Nous rappelons ici les principaux résultats.

Classification de vidéos de foule

La tâche de classification de vidéos sur l'ensemble de données CUHK consiste à catégoriser des séquences vidéo parmi les 8 catégories décrites dans le Tableau 3.1 en adoptant le protocole d'évaluation *one-leave-out* décrit dans (Shao et al., 2014). La classification est réalisée avec un SVM multi-classes avec RBF en tirant plusieurs fois des partitions apprentissage/test et en prenant soin de ne pas utiliser les mêmes scènes dans l'apprentissage et le test. Notre méthode donne une précision de classification au rang 1 de plus de 85%, dépassant largement les résultats obtenus avec les descripteurs de Kratz et Nishino, 2009 implémentés pour la classification (44%) et la méthode de Shao et al., 2014 (70%). Les matrices de confusion entre les 8 catégories pour les 3 méthodes sont données dans la Figure 3.19. Une étude d'ablation montre que la modification de l'algorithme de suivi de points KLT apporte 5% de précision en plus, et la modélisation par triangulation de Delaunay ajoute 7% de précision par rapport à un graphe de type k-NN.

10. http ://mha.cs.umn.edu/proj_events.shtml#crowd

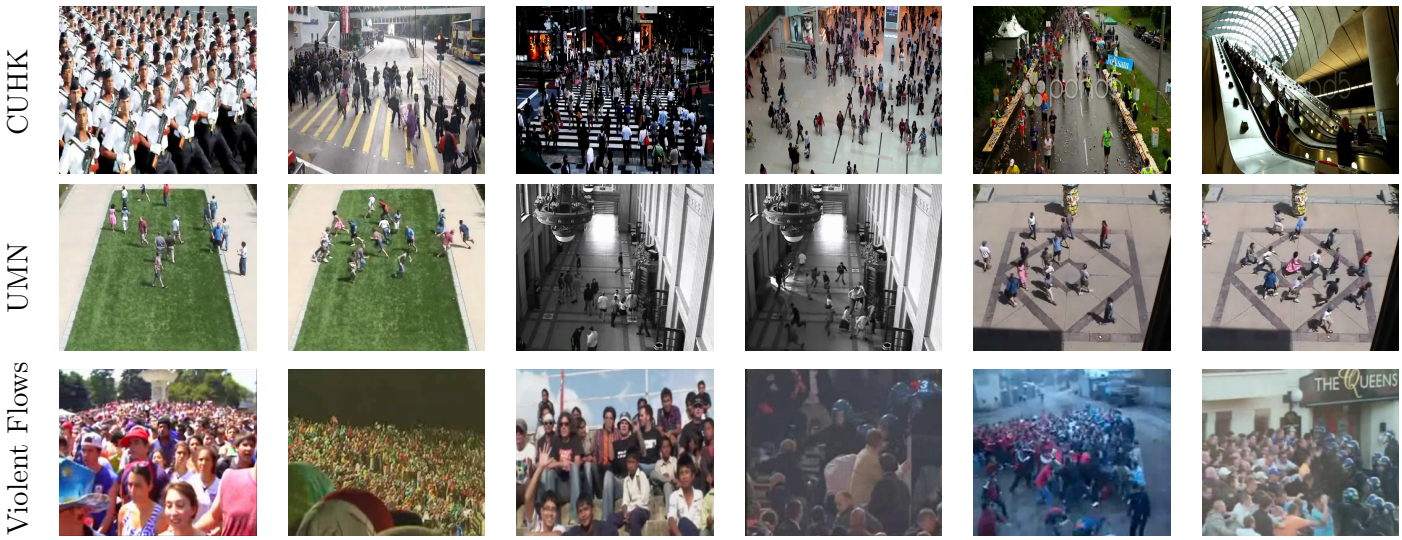


FIGURE 3.18 – Exemples d'images tirées de vidéos des bases utilisées dans nos expérimentations.

	1	2	3	4	5	6	7	8
1	0.32	0.30	0.02	0.02	0.04	0.06	0.20	0.04
2	0.02	0.90	0.02	0	0	0.02	0.02	0.02
3	0.02	0.50	0.24	0.02	0.02	0.12	0	0.08
4	0.06	0.46	0.08	0	0.22	0.08	0.04	0.06
5	0.02	0.50	0	0	0	0.10	0	0.38
6	0	0.22	0.10	0	0.02	0.58	0	0.08
7	0.06	0.14	0	0	0.04	0.04	0.56	0.16
8	0	0.02	0	0	0	0	0.02	0.96

(A)

	1	2	3	4	5	6	7	8
1	0.76	0.06	0.04	0.02	0	0.06	0.06	0
2	0	0.92	0	0	0	0.02	0	0.06
3	0.02	0.26	0.52	0	0	0.14	0	0.06
4	0	0.14	0.04	0.74	0	0.06	0.02	0
5	0.02	0.06	0.12	0	0.58	0.14	0	0.08
6	0	0.18	0.06	0	0	0.72	0	0.04
7	0	0.14	0.02	0.08	0	0.02	0.66	0.08
8	0.02	0.20	0	0	0 .04	0.02	0.02	0.70

(B)

	1	2	3	4	5	6	7	8
1	0.98	0	0	0	0	0.0200	0	0
2	0	0.88	0.10	0	0	0	0	0.02
3	0	0	0.94	0.02	0.02	0	0.02	0
4	0.08	0	0.08	0.76	0	0.08	0	0
5	0	0.08	0.46	0	0.42	0.04	0	0
6	0.02	0	0.04	0	0	0.92	0.02	0
7	0	0	0	0	0	0.02	0.96	0.02
8	0	0.02	0	0	0	0	0.02	0.96

(C)

FIGURE 3.19 – Matrices de confusion pour la classification des vidéos de l'ensemble CUHK (A) avec les descripteurs de KRATZ et NISHINO, 2009 (précision de 44%), (B) avec les descripteurs de groupe de SHAO et al., 2014 (70% de précision) et (C) avec nos descripteurs (précision de 85%)

Détection d'événements anormaux

Il s'agit dans cette application de détecter dans des séquences vidéos les événements anormaux. Contrairement à l'application précédente, le type précis de comportement n'est pas recherché, mais la segmentation temporelle est requise car les événements à détecter ne surviennent qu'à certains moments. Le jeu de données UMN Crowd abnormal behaviour est utilisé ici. Malgré sa simplicité (les événements sont globaux et assez évidents à voir), ce jeu de données est intéressant pour évaluer la généricité de notre approche. Comme dans (KALTSA et al., 2015), le critère d'évaluation utilisé est l'aire sous la courbe ROC de détection (AUC), les détections étant évaluées par image. Le Tableau 3.2 indique les performances de détection obtenues par notre approche et différentes méthodes de l'état de l'art. On constate des performances élevées pour toutes les approches, du fait de la simplicité du jeu de données. La Figure 3.20 montre un exemple de détection au cours du temps sous la forme d'un chronogramme et la localisation dans l'image lorsque la méthode est appliquée dans des cellules de l'image.

Détection de violence

La robustesse et la généricité de notre approche sont éprouvées dans une troisième application qui est la détection de violence dans une foule. Celle-ci est évaluée sur la base de vidéos *Violence in crowds*. L'évaluation consiste à effectuer une validation croisée avec 5 partitions constituées chacune de 4 ensembles d'apprentissage et un ensemble de test. Comme pour la classification

Method	SFM	Chaotic invariants	Sparse reconstruction	Local statistics	MDT	HOG+HOS	Ours
AUC	94.9	99.4	99.6	99.5	99.5	97.02	98.6

TABLEAU 3.2 – Performance de détection d'événements anormaux sur le jeu de données UMN Crowd abnormal behaviour pour notre approche, comparée à celles obtenues par les approches de l'état de l'art : SFM (MEHRAN et al., 2009), Chaotic invariants (WU et al., 2010), Sparse reconstruction (CONG et al., 2011), Local statistics (SALIGRAMA et CHEN, 2012), MDT (LI et al., 2014b), HOG+HOS (KALTSA et al., 2015).

FIGURE 3.20 – Illustration des résultats de détection d'anomalies sur une séquence du jeu de données UMN Crowd Abnormal Behaviour. Le chronogramme (en bas) montre en vert les images correspondant aux situations normales, et en rouge les images relatives à l'événement anormal (mouvement de fuite). A titre de comparaison, nous montrons les résultats de SFM (MEHRAN et al., 2009). L'image à droite présente la localisation de la détection lorsque celle-ci est effectuée sur une grille régulière de l'image.

de vidéos sur CUHK, la classification binaire violence/non violence est réalisée avec un SVM avec un noyau de type RBF. Les métriques d'évaluation employées sont la précision moyenne (ACC pour *accuracy*) et l'aire sous la courbe ROC (AUC), comme dans (HASSNER et al., 2012). Les résultats reportés dans le Tableau 3.3 montrent que les descripteurs proposés permettent d'atteindre des performances supérieures à celles des méthodes de l'état de l'art.

3.5 Conclusion et perspectives

Ce chapitre a exposé les problématiques de détection de comportement anormaux et d'analyse des comportement dans la foule dans le contexte de la vidéo-surveillance. Nous avons d'abord insisté sur la complexité de ces tâches, leur définition, et leur réalisation et leur évaluation. L'état de l'art précédant nos travaux montre que parallèlement aux recherches sur la détection et le suivi d'objets, pré-requis pour construire une approche microscopique d'analyse de comportement humain, la communauté scientifique a cherché à progresser sur ces sujets en proposant des modélisations macroscopiques. D'une part, les approches microscopiques n'étaient pas envisageables pour un bon nombre de cas d'usage, en raison de leur manque de précision et de robustesse, d'autre part, les approches macroscopiques présentent plusieurs avantages dont la généricité de représentation des mouvements et des actions et la possibilité de modéliser les interactions avec le même formalisme.

Nos travaux s'inscrivent dans ce courant. Dans les contributions présentées, que ce soit pour

Method	ACC (%)	AUC
Local Trinary Patterns (Yeffet et Wolf, 2009)	71.53	79.86
Histogram of Oriented Gradients (Laptev et al., 2008)	57.43	61.82
Histogram of Oriented Optical flow (Wang et Snoussi, 2012)	58.53	57.60
HNF (Laptev et al., 2008)	56.52	59.94
ViF (Hassner et al., 2012)	81.30	85.00
HOT (Mousavi et al., 2015)	78.30	n/a
HOT (no orientation) (Mousavi et al., 2015)	82.30	n/a
Our proposed approach	**84.44**	**88.00**

Tableau 3.3 – Comparaison des performances de détection de violence avec des méthodes de l'état de l'art, sur la base *Violence in Crowds*.

la détection d'événements violents ou pour l'analyse de la foule, nous avons travaillé à développer de nouvelles représentations visuelles, plus riches et plus adaptées aux types de comportements à reconnaître. Les deux approches proposées exploitent l'information locale de mouvement et un modèle graphique permettant d'expliciter les relations spatiales et temporelles entre les régions de l'image. La conception de ces approches est menée avec la préoccupation de pouvoir déployer celles-ci dans des applications temps réel. La méthode RIMOC implémente un nouveau descripteur codant la cohérence du mouvement dans des volumes spatio-temporels. Les descripteurs et les relations avec les volumes voisins sont appris en ligne sur des données vidéo de situations normales et comparés avec ceux des nouvelles données, à l'inférence, pour détecter les mouvements qui diffèrent du modèle appris. Les performances de détection scènes de violence obtenues avec cette méthode nous ont conduit à en développer une version plus optimisée, qui a été transférée à THALES et intégrée dans ses solutions de vidéo analytique. La fonctionnalité a été déployée dans de grandes gares françaises et testée avec succès lors d'expérimentations à grande échelle à l'occasion de l'Euro 2016 de football. La modélisation des mouvements dans une foule par un ensemble de descripteurs de niveau intermédiaire apporte également de outils intéressants pour adresser des applications telles que la classification des comportements, la détection de comportements inhabituels, et la détection de violence dans les milieux denses. Le graphe de Delaunay défini à partir des trajectoires de points permet de modéliser les interactions entre les personnes.

Il faut néanmoins avoir conscience des limitations des systèmes d'analyse automatique de vidéos. Ces outils sont aujourd'hui assez puissants pour filtrer efficacement les flux vidéo et attirer l'attention des opérateurs sur des événements d'intérêt, mais ne permettent pas encore une détection totalement fiable et précise, les taux de faux positifs étant trop élevés pour se reposer entièrement sur les algorithmes. La mise en œuvre trop précoce et sans précaution de technologies insuffisamment matures peut s'avérer contre-productive car elle va susciter la défiance de la part des utilisateurs et des citoyens, qui sont déjà pour diverses raisons peu enclins à l'adoption des systèmes de vidéo-surveillance : sensibilité aux questions des libertés individuelles, retour sur investissement mal évalué, crainte d'un remplacement des personnes par les machines, modifications profondes des habitudes et des processus métier,... L'usage le plus pertinent et le mieux accepté aujourd'hui reste un emploi de ces technologies avec "l'humain dans la boucle", qui garde ainsi le pouvoir de jugement et de décision finale.

Malgré les verrous scientifiques, les obstacles techniques et les débats sociétaux, il nous semble important de pouvoir continuer la recherche sur ces thématiques. Beaucoup de chemins restent à explorer pour atteindre des niveaux bien supérieurs de précision, de robustesse et de richesse de la description sémantique.

Les progrès remarquables en détection et suivi d'objets ces dernières années, grâce à de nouveaux modèles d'apprentissage, rebattent les cartes et font penser que les approches microscopiques apporteront d'ici peu de nouveaux moyens d'accéder à la caractérisation du comportement dans des scènes complexes et denses. Des ensembles de données pour la détection des personnes

dans les scènes de foule ont vu le jour (Shao et al., 2018) et l'estimation de leur pose (Li et al., 2019a). Le challenge MOT (*Multiple Object Tracking*) a ajouté récemment à la compétition de suivi multi-cibles 8 séquences vidéo de scènes de foule (Dendorfer et al., 2020). Un autre jeu de données dédié au suivi des têtes des personnes dans la foule, CroHD, vient d'être annoncé (Sundararaman et al., 2021). Une voie intéressante de recherche serait d'examiner comment hybrider des approches microscopiques et approches macroscopiques pour tirer parti de leurs avantages respectifs.

S'agissant de la détection d'événements violents, la mise à disposition de grandes bases annotées (Cheng et al., 2020) facilitera l'exploration de méthodes d'apprentissage profond de représentations spatio-temporelles de l'apparence et du mouvement dans les séquences vidéo.

Afin d'éviter de recourir à la labellisation des données vidéo, l'intérêt des méthodes d'apprentissage auto-supervisé paraît évident. Le pré-apprentissage de modèles temporels avec des approches contrastives a montré des résultats prometteurs pour la reconnaissance d'actions (Lorre et al., 2020). Il serait intéressant d'évaluer la pertinence de ces représentations dans un cadre d'apprentissage une-classe pour des contextes spécifiques.

La détection d'anomalies reste un problème ouvert. Sans préciser la nature des événements anormaux, Sultani et al., 2018 proposent un ensemble de clips de vidéo-surveillance divisées en deux catégories normal/anormal, et formulent la détection comme un problème de régression d'un score d'anormalité dans un cadre d'apprentissage d'instances multiples. Une piste de réflexion serait d'imaginer une méthode d'apprentissage exploitant très peu d'exemples de vidéos d'anomalies, en tirant parti d'une étape de pré-apprentissage ou d'un apprentissage semi-supervisé. Dans un autre domaine applicatif, Fang et al., 2019 se sont intéressé à l'attention du conducteur dans la scène pour concevoir une méthode de reconnaissance des situations d'accidents. Les mécanismes d'attention visuelle seront sans doute des objets d'étude à investiguer davantage pour la détection d'anomalies.

Chapitre 4

Vers une caractérisation 3D des personnes pour l'analyse de comportement

4.1 Contexte et motivations

Approches macroscopiques vs approches microscopiques

Les recherches sur l'analyse de comportement dans les vidéos, présentées au chapitre 3, sont fondées sur une hypothèse de travail forte qui est l'absence de modélisation individuelle. Elles se justifiaient par les considérations suivantes :

- Il est difficile de détecter les personnes de manière robuste et de décrire précisément leur comportement dans des scènes complexes (densité de personnes élevées, occultations fréquentes, grandes variations d'échelle).
- Sans modélisation individuelle, il est néanmoins possible de procéder à une analyse globale de l'image et des séquences d'images pour reconnaître des événements ou des types de comportements. Les approches macroscopiques présentent plusieurs avantages qui sont la simplicité de la modélisation, la généricité et une complexité de calcul relativement indépendante du nombre de personnes.
- Dans le cas de foules très denses, on peut se poser la question de la pertinence de l'analyse du comportement individuel, tant les interactions entre personnes sont fortes, un groupe de personnes pouvant être vu comme un ensemble cohérent agissant d'une même façon.

Malgré leur intérêt, les approches macroscopiques n'apportent qu'une réponse partielle aux problématiques d'analyse automatique de comportement. Plusieurs raisons nous amènent à réfléchir de nouveau à l'opportunité et à la pertinence de consacrer des efforts de recherche sur les approches de modélisation individuelle.

La première est l'objectif permanent d'élever le niveau sémantique et la finesse d'interprétation des scènes. L'analyse du comportement des personnes est d'autant plus précise et complète quand elle s'enrichit de connaissances spécifiques à chaque individu : localisation, déplacement, posture, gestes, actions, interactions avec des objets ou avec d'autres personnes, activités, intentions, émotions, etc. L'accès à ces informations laisse entrevoir de nombreuses possibilités d'applications (Figure 4.2).

L'évolution des capteurs est aussi un facteur important de la montée en performances des modèles de reconnaissance visuelle. La résolution des caméras actuelles et leur capacité à produire des images de qualité malgré des conditions de lumière difficiles et changeantes, va faciliter la capture d'images détaillées des personnes à de plus grandes distances, dans des environnements extérieurs et potentiellement denses. Les dispositifs d'acquisition d'images opérant à très haute

fréquence (>100 Hz), employés par exemple pour capter les événements sportifs avec une grande précision temporelle, ouvrent également la voie à l'exploration et la mise au point de nouvelles fonctionnalités d'analyse du comportement.

Troisièmement, grâce aux progrès spectaculaires réalisés ces dernières années en détection d'objets (Zhao et al., 2019), on peut envisager le développement de nouvelles méthodes d'analyse de comportement individuel, plus précises et plus fiables. Comme pour d'autres domaines de l'intelligence artificielle, les avancées en détection d'objets sont à la fois méthodologiques, liés aux évolutions technologiques, et dopées par l'exploitation de grands ensembles de données annotés partagés publiquement, comme cela a été évoqué au Chapitre 1. Suivant les mêmes tendances, les autres tâches de caractérisation des objets et des personnes, comme la prédiction de points d'intérêt, l'estimation de pose 2D, la reconnaissance d'actions, le suivi temporel, etc. concentrent un grand nombre de travaux et connaissent elles aussi une évolution rapide.

Intérêt de l'estimation de poses 3D humaines

Nous nous plaçons dans le cadre du problème de l'estimation de poses 3D à partir d'images RGB, sans l'utilisation de dispositifs spécifiques d'acquisition de données 3D (caméras stéréoscopiques, caméras RGBD, capteurs *Light Detection and Ranging* (LiDAR), systèmes de *Motion Capture* (MoCap)). Même si les nouvelles technologies de capteurs 3D commencent à atteindre le marché du grand public avec des compromis performance/coût/facilité d'intégration toujours plus attrayants (on peut citer par exemple le LIDAR équipant certains smartphones haut de gamme, ou ceux destinés au marché des systèmes d'assistance automatique des véhicules autonomes), la caméra couleur 2D reste un dispositif d'acquisition d'images peu onéreux en général, très facile à déployer à grande échelle, et déjà présent sur plusieurs milliards d'appareils mobiles dans le monde.

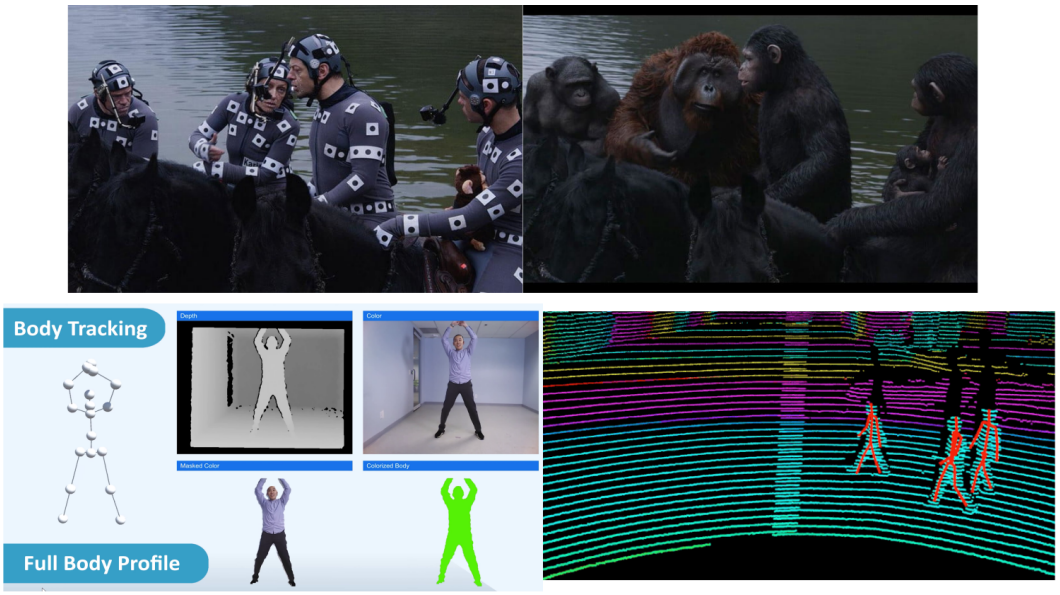


Figure 4.1 – Estimation de pose 3D avec des dispositifs d'acquisition de données 3D. En haut : système de MoCap utilisé pour le tournage du film *Dawn of the Planet of The Apes* (2014), en bas à gauche : suivi du squelette avec la caméra RGBD ORBECC Astra, en bas à droite : estimation de poses 3D dans un nuage de points issus du LIDAR Luminar.

L'information de géométrie donnée par la pose 3D des personnes est très informative de leur posture et de leurs gestes. La pose 3D est exploitée par le monde de l'animation pour la création de contenus virtuels dans les jeux vidéo et le cinéma. C'est aussi une information utile

à exploiter à plus haut niveau pour la reconnaissance des actions (Luvizon et al., 2020) et la compréhension des interactions. Les enjeux applicatifs sont importants car cette information peut être utilisée à des fins diverses : l'analyse de comportements pour la vidéosurveillance, l'anticipation de l'intention des piétons pour le véhicule autonome (Ranga et al., 2020), l'analyse d'activité pour la ville ou le bâtiment intelligents (Vaquette et al., 2019).

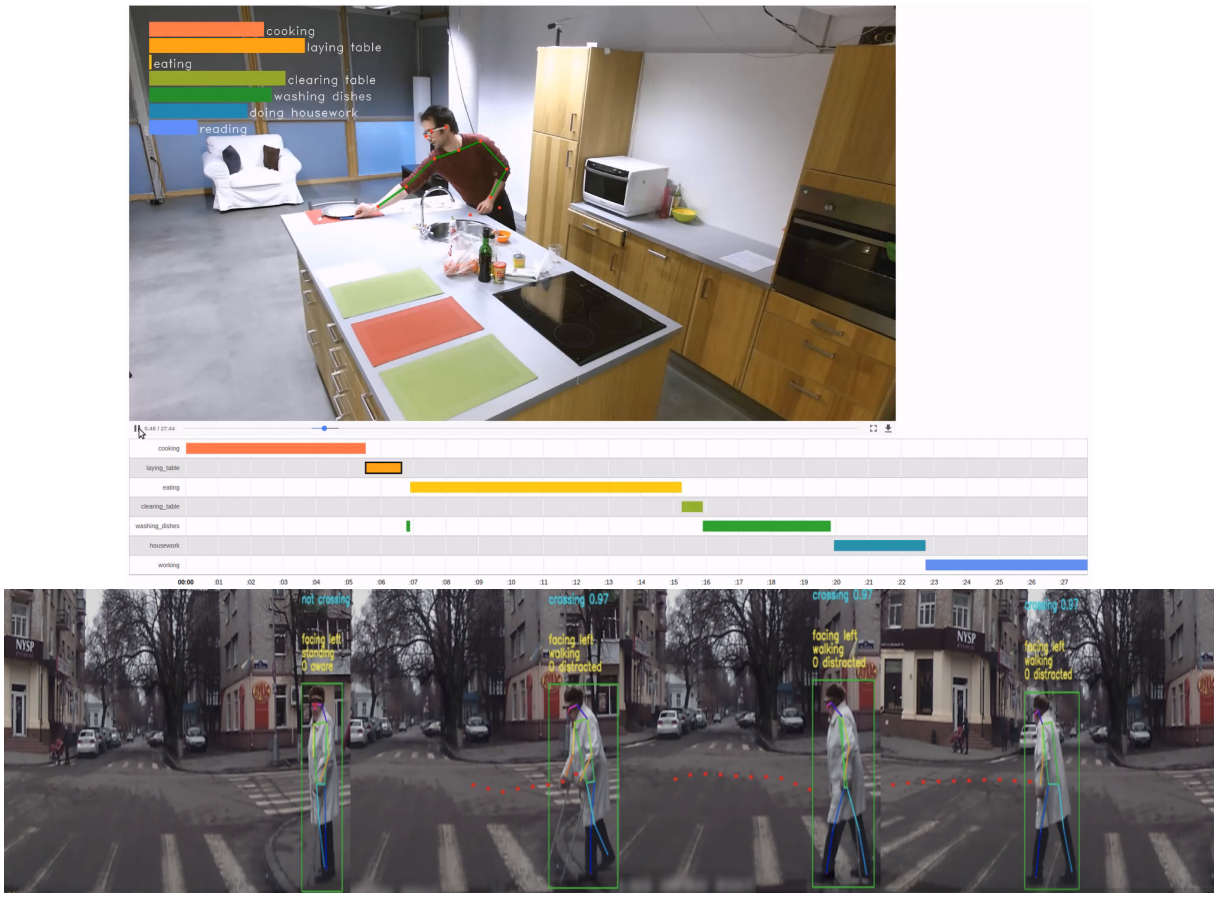


Figure 4.2 – Exploitation de la pose humaine pour différentes tâches de reconnaissances, en haut : pour la reconnaissance d'activités dans une application de maison intelligente (Vaquette et al., 2019), en bas : pour la reconnaissance de l'intention des piétons par un véhicule autonome (Ranga et al., 2020).

Les challenges de l'estimation de poses 3D multi-personnes

Les qualités principales souhaitées d'un estimateur de poses 3D multi-personnes sont d'abord la robustesse face à la variété de situations réelles et la précision de l'estimation 3D. Dans les applications réelles, la difficulté de la tâche provient de différents facteurs :

- la variabilité morphologique des humains,
- la variété des poses, certaines poses étant jugées "difficiles" car rarement observées, telles que celles constatées lors de l'exécution de gestes sportifs ou de gestes techniques,
- la variété et la complexité des scènes notamment en environnement extérieur,
- le nombre et la densité variables de personnes (qui vont de quelques personnes isolées à des groupes denses),
- les occultations fréquentes entre personnes lorsqu'elles interagissent, les occultations par d'autres éléments de la scène, l'absence de visibilité par troncature,

— une grande dynamique dans les échelles d'observation, lorsque le caméra couvre un espace large comme c'est le cas pour la vidéosurveillance urbaine par exemple,
— des points de vue variés et des points de vue défavorables à l'estimation non ambiguë de certaines poses.

Autre point important, le temps de calcul de l'estimation de pose est un facteur dimensionnant dans beaucoup d'applications requérant un traitement en temps réel ou du moins en temps acceptable pour l'utilisateur. Cette contrainte est d'autant plus difficile à respecter dans le cas multi-personnes. Enfin, une difficulté vient s'ajouter aux précédentes, c'est le manque de données représentatives annotées en poses 3D dans des conditions variées : les ensembles de données disponibles sont issus d'acquisitions en laboratoire, mettent en scène un nombre très limité d'intervenants, et présentent peu de variété d'environnements (Ionescu et al., 2014 ; Joo et al., 2017 ; Mehta et al., 2018).

4.2 Etat de l'art et positionnement

4.2.1 Estimation de poses 2D

La littérature sur l'estimation de poses 2D comporte de nombreux travaux qui ont notamment tiré profit de la création d'ensembles d'images dont les annotations contiennent les points-clés localisés sur les personnes, comme COCO KeyPoints (Lin et al., 2014). Les méthodes proposées pour l'estimation de pose 2D mono-personne sont fondées soit sur la régression directe des articulations en 2D (Carreira et al., 2016 ; Toshev et al., 2014), soit sur la détection des articulations au moyen de cartes de chaleur prédites par un réseau de neurones convolutif (Newell et al., 2016 ; Tompson et al., 2015 ; Wei et al., 2016). L'étape d'extraction des positions des articulations est externe au réseau de neurones, car c'est une opération non différentiable, et ne peut être entraîné dans un apprentissage de bout-en-bout. Les cartes de chaleur sont habituellement calculées à des résolutions beaucoup plus basses que celle de l'image d'entrée pour des raisons d'occupation mémoire, ce qui limite la précision de la localisation des articulations. L'architecture *stacked hourglass network* de (Newell et al., 2016) est particulièrement intéressante car elle permet de fusionner des caractéristiques à différents niveaux de résolution et sémantiques pour encoder des informations locales et des informations plus globales de contexte.

L'extension au cas multi-personnes est traitée de deux manières différentes :
— Approches *top-down* : un détecteur localise les régions d'intérêt contenant les personnes, dans lesquelles la pose 2D de chaque personne détectée est estimée (Chen et al., 2018b ; He et al., 2017 ; Papandreou et al., 2017).
— Approches *bottom-up* : les articulations sont détectées dans des cartes de chaleur, elles sont ensuite regroupées en squelettes entiers. L'association des articulations s'appuie sur différentes modélisations (*Part Affinity Fields* (Cao et al., 2017), *associative embeddings* (Newell et al., 2017), *Part Associative Fields* (Kreiss et al., 2019)).

Plus précises, les méthodes *top-down* sont plus généralement plus lentes à cause de la succession des étapes de détection et d'estimation de poses. A l'inverse, les méthodes *bottom-up* plus rapides, souffrent des limitations relatives à la précision des cartes de chaleur et de la complexité de l'étape d'association qui est sujette aux erreurs.

4.2.2 Estimation de poses 3D

De la même manière, les travaux sur l'estimation de poses 3D ont d'abord porté sur le cas mono-personne.

Les approches par reconstruction prédisent la pose 2D dans l'image, puis infèrent la pose 3D à partir de la pose 2D. La reconstruction est réalisée par un petit réseau de neurones (Martinez et al., 2017), en établissant des contraintes géométriques entre la 2D et la 3D (Li et Lee, 2019 ;

Moreno-Noguer, 2017), des contraintes cinématiques et de poses 3D anatomiquement correctes (Chen et al., 2019a ; Fang et al., 2018 ; Nie et al., 2017), ou en utilisant l'ordonnancement des profondeurs des articulations 2D (Wang et al., 2018b).

Les approches directes prédisent les poses 3D sans passer par l'étape intermédiaire d'estimation des poses 2D. Les approches de Pavlakos et al., 2017 et Sun et al., 2018 sont fondées sur une représentation volumétrique de l'espace, alors que Tekin et al., 2016 proposent d'apprendre une représentation latente des poses 3D avec un auto-encodeur. Afin d'augmenter les performances et les capacités de généralisation de l'estimation de poses 3D, d'autres travaux ont exploité des tâches supplémentaires en 2D (Li et Chan, 2014), le transfert d'apprentissage (Mehta et al., 2018), la modélisation de contraintes géométriques sur les articulations (Zhou et al., 2017), des modèles d'apprentissage antagoniste (Yang et al., 2018), la supervision faible de l'ordonnancement de la profondeur (Pavlakos et al., 2018), le respect de contraintes anatomiques (Dabral et al., 2018). Le cas plus complexe multi-personnes n'est traité que depuis peu de temps. Rogez et al., 2017 proposent une modélisation des poses 3D à partir d'ancres de poses qui sont raffinées. Moon et al., 2019a développent une méthode *top-down* complexe mettant en œuvre trois réseaux de neurones dédiés respectivement à la détection des personnes, la régression des poses 3D relatives, la localisation 3D de l'articulation racine. Zanfir et al., 2018a retrouvent la pose 3D des personnes dans des séquences d'images en formulant un problème d'optimisation global sur les poses 3D des personnes et les trajectoires, et en intégrant des contraintes géométriques et sémantiques. Parmi les méthodes rapides *single-shot*, Mehta et al., 2018 adressent le problème des occultations avec un modèle de stockage redondant des poses 3D dans des *Occlusion-Robust Pose Maps* (ORPM), ce qui permet de récupérer des positions d'articulations occultées à d'autres endroits du squelette. Fabbri et al., 2020 introduisent une représentation volumétrique de toutes les personnes qui est compressée pour réduire sa dimension.

Métriques d'évaluation

En estimation de poses 3D, les deux métriques d'évaluation de la précision les plus utilisées sont :

- Mean Per Joint Position Error (MPJPE) : c'est la distance euclidienne entre les positions estimées des articulations et leurs positions réelles, moyennées pour toutes les articulations de la personne (en mm).
- 3D Percentage of Correct Keypoints (3DPCK) (Mehta et al., 2017a) : Pourcentage des articulations correctement localisées. Cette métrique considère une articulation comme correctement estimée si l'erreur de l'estimation est inférieure à 150mm.

Dans le cas de l'estimation de poses 3D relatives, les positions 3D des articulations sont exprimées par rapport à une origine placée sur l'articulation racine qui est le pelvis. Dans ce qui suit, les métriques MPJPE et 3DPCK seront calculées sur des poses 3D relatives.

4.2.3 Positionnement de nos travaux

Alors que les méthodes les plus précises de l'état de l'art mettent en œuvre des pipelines complexes dont le temps d'exécution explose avec un nombre important de personnes, les approches les plus rapides manquent toujours de précision et de robustesse. Nos travaux ont pour but de proposer des nouvelles solutions d'estimation de poses 3D multi-personnes qui concilient précision, robustesse aux occultations et aux variations d'échelle, et rapidité, même pour des scènes avec un grand nombre de personnes. Ces travaux ont été effectués dans le cadre de la thèse d'Abdallah Benzine (2017-2020) (Benzine, 2020) et ont donné lieu à plusieurs contributions : la proposition d'une première méthode d'estimation de pose 3D relative rapide *bottom-up single-shot* (Benzine et al., 2019, 2020b) aux performances très compétitives avec l'état de l'art, puis

une seconde méthode *top-down* et single-shot encore plus performante (Benzine et al., 2020b). Une extension à l'estimation de poses 3D absolues a également été étudiée.

4.3 Approche bottom-up single-shot pour l'estimation de poses 3D humaines

4.3.1 Présentation de l'approche

Nous avons cherché à mettre au point une méthode d'estimation de poses 2D et 3D multi-personnes à la fois précise, robuste aux occultations, et rapide, c'est-à-dire dont la complexité n'augmente pas trop avec le nombre de personnes présentes dans l'image. Les approches top-down procèdent généralement en deux temps, comme celle de Moon et al., 2019a : la première étape est dédiée à la détection des personnes, la seconde à l'estimation de la pose 3D à partir des caractéristiques extraites dans les régions d'intérêt contenant chaque personne détectée. De ce fait, elles sont particulièrement lentes pour un grand nombre de personnes. A l'inverse, les approches *bottom-up*, plus rapides, réalisent la détection des articulations dans l'image en une passe, puis le regroupement de ces articulations dans une étape de post-traitement (Mehta et al., 2018 ; Zanfir et al., 2018a).

L'approche que nous avons proposée (Benzine et al., 2019, 2020b) est un réseau de neurones profond entraîné de bout-en-bout pour effectuer conjointement :

- la localisation 2D des articulations,
- l'estimation des coordonnées 3D des articulations,
- le regroupement des prédictions d'articulations en squelettes complets,

pour un nombre inconnu et variable de personnes dans l'image.

Notre approche est inspirée de celle de Mehta et al., 2018, et reprend le principe des cartes de chaleur pour prédire la position des articulations en 2D et les modèles *Occlusion-Robust Pose-Maps* (ORPM) pour le stockage des coordonnées 3D des articulations (Figure 4.3). En revanche, elle s'en différencie sur deux aspects principaux :

- Le choix de l'architecture de réseau de type *stacked hourglass* à la place de l'architecture *Resnet50* : les architectures *stacked hourglass* ont démontré leur efficacité pour l'estimation de poses 2D (Newell et al., 2016). Dans cette architecture, chaque module *hourglass* capture et combine les caractéristiques visuelles à plusieurs échelles et niveaux sémantiques, et leur empilement permet de raffiner les prédictions en passant successivement dans chacun des modules (Figure 4.4).
- L'utilisation des *Associative Embedding* (AE)(Newell et al., 2017) à la place des *Part Affinity Fields* : c'est une méthode de regroupement implicite des articulations par le biais d'une tâche de prédiction de tags (ou *embeddings*) qui doivent être proches pour toutes les articulations d'une même personne, et éloignées pour des personnes différentes. Les résultats expérimentaux donnés par ibid. montrent la supériorité des *Associative Embedding* sur les *Part Affinity Fields* pour l'estimation de poses 2D.

Le réseau de neurones prend en entrée une image RGB $\mathbf{I}$ de taille W x H. Après plusieurs couches de convolution et de pooling, les cartes de caractéristiques de plus basse résolution W'xH' passent par le réseau *stacked hourglass*. Le réseau prédit en sortie une carte de chaleur et une carte d'*associative embedding* toutes deux de dimension W' x H' x K, et une carte d'ORPM de dimension W' x H' x 3 K qui stockent aux coordonnées 2D des articulations leurs coordonnées 3D, K étant le nombre d'articulations d'un squelette. Une illustration de l'approche proposée est donnée Figure 4.3.

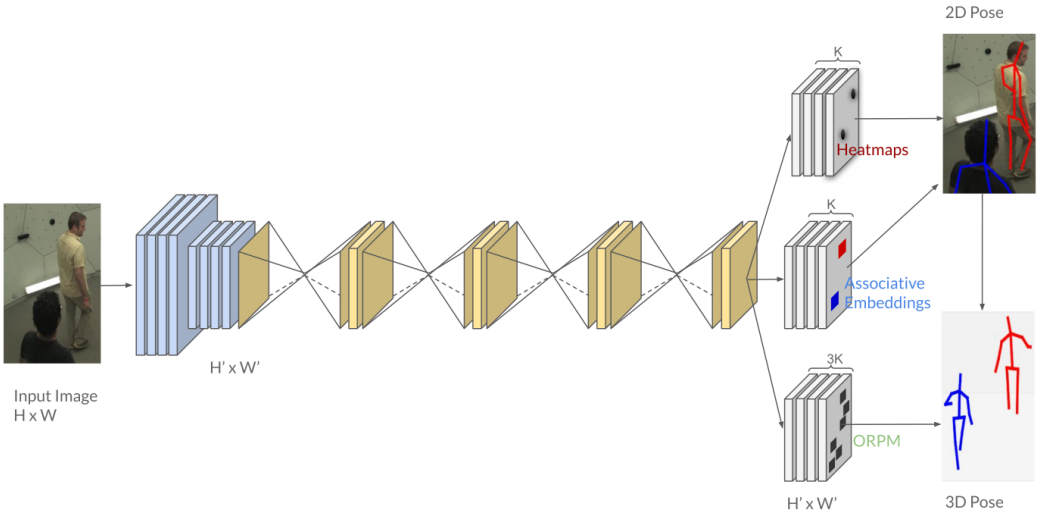


FIGURE 4.3 – Illustration générale de l'approche *bottom-up* d'estimation de poses 3D multi-personnes proposée : l'approche est fondée sur une architecture de type *stacked hourglass network* entraînée à prédire des cartes de chaleur pour la position 2D des articulations, des *associative embeddings* servant à regrouper les articulations d'une même personne, et des ORPM permettant un stockage des coordonnées 3D des articulations robuste aux occultations partielles.

Regroupement des articulations par Associative Embeddings

Le regroupement par les *Associative Embeddings* (AE) est formulé de la façon suivante. Soit $E_k \in \mathbb{R}^{W' \times H'}$ une carte d'*associative embedding* prédite par le réseau pour la $k^{\text{ième}}$ articulation et $e_k(\boldsymbol{x})$ la valeur contenue dans cette carte à la position 2D $\boldsymbol{x}$. N étant le nombre de personnes dans l'image, $\boldsymbol{x_{k,n}}$ est la position 2D de vérité de terrain de la $k^{\text{ième}}$ articulation de la personne n. L'*embedding* de référence d'une personne est alors l'*embedding* prédit moyen de toutes ses articulations :

$$\overline{e}_n = \frac{1}{K} \sum_k e_k(\boldsymbol{x_{k,n}}) \tag{4.1}$$

Pour entraîner le modèle à prédire les *embeddings*, on définit la fonction de perte à minimiser suivante :

$$\mathcal{L}_{AE} = \frac{1}{NK} \sum_n \sum_k (\overline{e}_n - e_k(x_{k,n}))^2 + \frac{1}{N^2} \sum_n \sum_{n' \neq n} \exp\left(-\frac{1}{2}(\overline{e}_n - \overline{e}_{n'})^2\right) \tag{4.2}$$

Le premier terme a pour effet de rapprocher les *embeddings* d'une même personne (fonction d'attraction ou *pull loss*), alors que le second terme a pour but d'éloigner les *embeddings* de personnes différentes (fonction de répulsion ou *push loss*).

Stockage des poses 3D robuste aux occultations

Le principe de l'ORPM consiste à stocker de manière redondante, à plusieurs positions 2D de la carte, les coordonnées 3D de chaque articulation, afin d'en assurer une estimation en cas d'occultation. Pour chaque articulation, on définit un graphe qui connecte de proche en proche cette articulation à un ensemble d'articulation parentes en remontant l'arbre cinématique jusqu'à l'articulation racine (pelvis ou cou si le pelvis est occulté, ces deux articulations sont considérées comme moins mobiles que les autres et moins sujettes aux occultations). Par exemple, les coordonnées 3D du poignet droit seront stockées au positions 2D du poignet droit, du coude droit, de

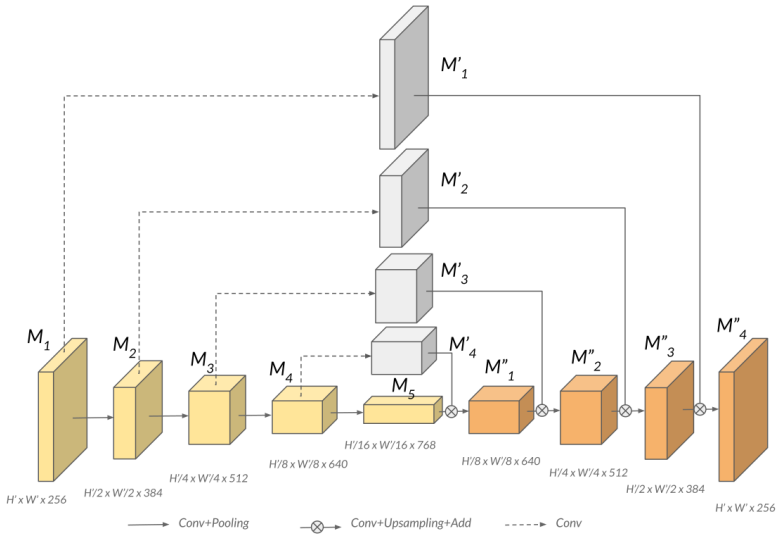


FIGURE 4.4 – Un module *hourglass* transforme et fusionne les caractéristiques à différentes résolutions et niveaux sémantiques.

l'épaule droite, du pelvis et du cou. La lecture est alors effectuée dans l'ordre suivant : poignet droit → coude droit → épaule droit → pelvis → cou en s'arrêtant à la première articulation valide. Une articulation est considérée valide si son score de détection en 2D est supérieur à un certain seuil et si les contraintes géométriques (i) de distance de l'articulation aux autres articulations inférieure à une distance maximum en 2D, et (ii) de taille du membre anatomiquement plausible, sont respectées. La stratégie de lecture de la coordonnée 3D d'une articulation est illustrée dans la Figure 4.5.

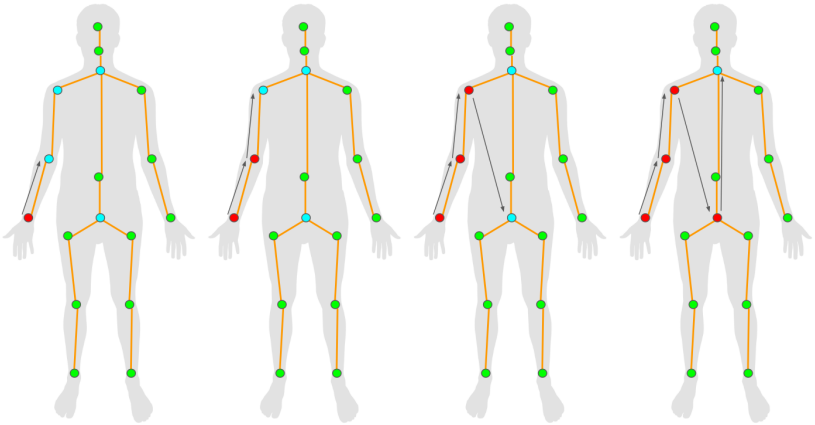


FIGURE 4.5 – Stratégie de lecture des coordonnées 3D des articulations dans les ORPM. Dans cet exemple, on cherche à lire les coordonnées 3D du poignet droit. S'il y a occultation ou que la position n'est pas géométriquement valide (en rouge), on remonte à l'articulation parente, jusqu'à rencontrer une position valide. La dernière articulation possible est le cou.

Pendant l'apprentissage, le modèle est entraîné à prédire les ORPM en minimisant la fonction de perte $\mathcal{L}_{\text{ORPM}}$ qui correspond à la distance euclidienne entre les ORPM prédites et celles de vérité de terrain. Il est à noter que la localisation 3D d'une articulation estimée à ses coordonnées 2D est potentiellement plus précise que celle estimée à l'emplacement d'une articulation maître.

Fonction de perte totale

Le modèle est entraîné en minimisant la fonction de perte totale suivante :

$$\mathcal{L}_{\text{3DMP}} = \mathcal{L}_{\text{2D}} + \mathcal{L}_{\text{ORPM}} + \lambda_{\text{AE}} \mathcal{L}_{\text{AE}} \tag{4.3}$$

Où $\mathcal{L}_{\text{2D}}$ est la distance euclidienne entre les cartes de chaleur prédites et celles de vérité de terrain, $\mathcal{L}_{\text{ORPM}}$ la fonction de perte associée aux ORPM et $\mathcal{L}_{\text{AE}}$ est la fonction de perte définie par l'équation (4.2). $\lambda_{\text{AE}} = 0.001$ est la pondération associée à la fonction de perte des *associative embeddings*.

Inférence et prédiction des poses 2D et 3D

Les trois types de cartes en sortie du réseau de neurones sont utilisées pour prédire les poses 2D et 3D des personnes. Tout d'abord, les articulations sont extraites en 2D en appliquant un algorithme de suppression des non-maxima aux cartes de chaleur. Pour regrouper les articulations en squelettes complets, on part des articulations de cou qui matérialisent les personnes détectées, et on recherche de manière itérative les articulations dont l'*embedding* est le plus proche de l'*embedding* moyen des articulations déjà associées, la distance entre *embeddings* devant être toujours inférieure à un seuil donné. A l'issue de cette étape, les poses 2D sont estimées. Ensuite, les poses 3D sont lues dans les cartes d'ORPM aux positions 2D des articulations détectées et regroupées, en suivant la stratégie décrite plus haut.

Afin d'améliorer la robustesse de la méthode lorsque la variabilité des échelles des personnes est grande, une stratégie d'inférence multi-échelles peut être mise en œuvre. L'image est donnée au réseau de neurones à différentes échelles, les prédictions en sortie sont redimensionnées à la haute résolution et fusionnées :

- les cartes de chaleur multi-échelles sont la moyenne des cartes de chaleur estimées aux différentes échelles,
- les cartes d'AE multi-échelles sont la concaténation des cartes d'AE aux différentes échelles, ce qui revient à définir des tags de plus grande dimension concaténant les tags de toutes les échelles traitées,
- les cartes d'ORPM multi-échelles sont obtenues en calculant la moyenne pondérée par les cartes de chaleurs prédites des cartes d'ORPM aux différentes échelles, et en moyennant pour chaque articulation sur tous les emplacements de stockage associés. Cette moyenne pondérée permet de faire intervenir la confiance accordée à l'estimation de la coordonnée 3D à chaque niveau de résolution, ce qui a pour effet de diminuer la sensibilité aux erreurs d'estimation à une échelle donnée.

4.3.2 Résultats expérimentaux

L'évaluation de la méthode a été réalisée sur trois jeux de données complémentaires, présentant des caractéristiques différentes en termes d'environnement, de nombre de personnes et de scénarios :

- CMU-Panoptic (Joo et al., 2017) : images réelles, quelques personnes (4-6), en environnement contrôlé et points de vue variés, ayant des interactions complexes ;
- MuPoTS-3D (Mehta et al., 2018) : images réelles, jusqu'à 3 personnes en intérieur et en extérieur, effectuant des actions variées ;
- JTA (Fabbri et al., 2018) : images de synthèse réalistes générées avec le moteur de jeu vidéo GTA, environnements urbains variés avec conditions lumineuses et points de vue variés, jusqu'à 60 personnes ayant des poses variées, grande variabilité d'échelles.

Les expérimentations réalisées sur l'ensemble de données CMU-Panoptic valident d'abord

l'intérêt de l'architecture du *stacked hourglass network* : la précision de l'estimation de pose 3D augmente à chaque fois qu'un module hourglass est ajouté, un empilement de quatre modules constituant un bon compromis entre la précision obtenue et la complexité du réseau. D'autre part, l'utilisation des ORPM augmente significativement les performances par rapport à une approche de lecture naïve des coordonnées 3D des articulations. Notre approche est bien plus précise que celle de Zanfir et al., 2018a et dépasse Zanfir et al., 2018b (Tableau 4.1). Des exemples qualitatifs de poses 3D estimées sur des images de CMU Panoptic sont données Figure 4.6 et montrent la capacité de notre méthode bottom-up single-shot à gérer efficacement un nombre variable de personnes et des situations d'occultation.

Méthode	MPJPE
Zanfir et al., 2018a	153.4
Zanfir et al., 2018b	72.1
Notre approche	**68.5**

Tableau 4.1 – MPJPE en mm sur l'ensemble de données CMU Panoptic en suivant le protocole Panoptic-1

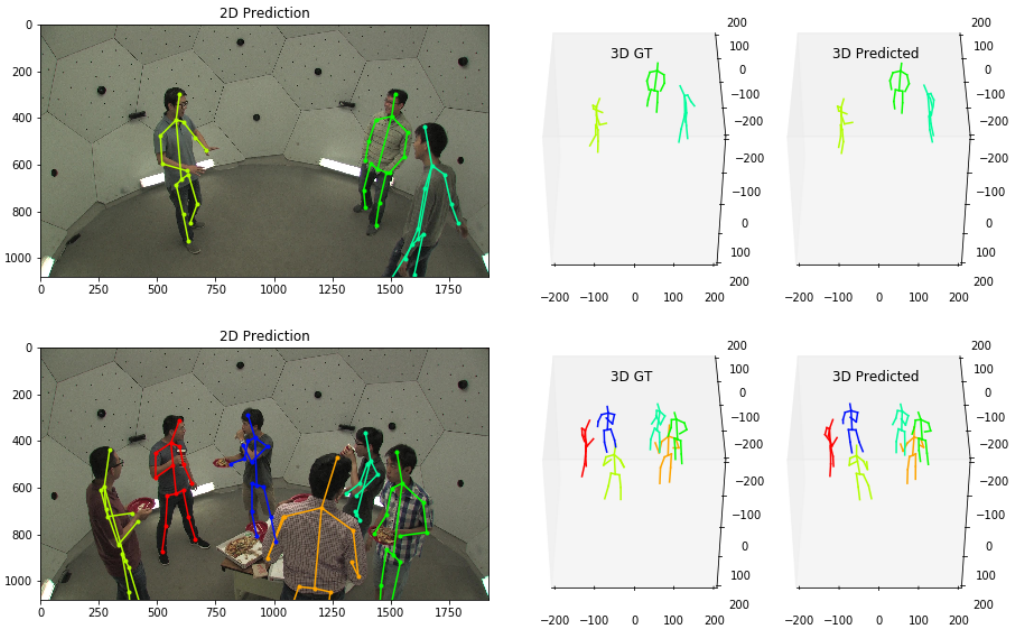


Figure 4.6 – Poses 3D multi-personnes prédites par l'approche proposée sur deux images de l'ensemble de données CMU-Panoptic. Pour les besoins de la visualisation, les poses 3D relatives estimées sont transformées en poses 3D absolues en utilisant les translations de la vérité terrain.

Contrairement à CMU-Panoptic, l'ensemble de données MuPoTS-3D présente l'avantage de contenir des scènes en environnement non contrôlé, et des fonds plus variés. En contrepartie, les scénarios ne font intervenir que trois personnes au maximum. Le protocole d'évaluation introduit dans (Mehta et al., 2018) pour MuPoTS-3D implique de réaliser l'apprentissage sur un ensemble constitué pour moitié d'images de MuCo-3DHP, pour l'autre moitié d'images du dataset COCO KeyPoints utilisé pour forcer la généralisation des modèles à des images *in the wild*. Pour les images de COCO KeyPoints qui ne sont pas annotées en 3D, seule la supervision 2D est maintenue dans l'optimisation. La précision obtenue par notre modèle est reportée sous forme de 3DPCK dans le Tableau 4.2, et comparée avec celle des approches concurrentes. Moon et al., 2019a

donne des résultats supérieurs à notre méthode au prix de la combinaison de deux modèles, un détecteur de personnes (Ren et al., 2015) suivi d'un modèle performant d'estimation de poses 3D appliquée sur chaque personne détectée (Sun et al., 2018). En revanche, notre méthode surpasse la méthode single-shot de référence de Mehta et al., 2018 sur le dataset MuPoTS-3D.

Méthode		3DPCK
Top-Down	Moon et al., 2019a	81.8
Single-Shot	Mehta et al., 2018	65.0
	Notre approche	**67.5**

Tableau 4.2 – 3DPCK sur l'ensemble de données MuPoTS-3D : bien que moins précise que l'approche *top-down* en deux étapes de Moon et al., 2019a, notre méthode dépasse l'approche *single-shot* concurrente de Mehta et al., 2018.

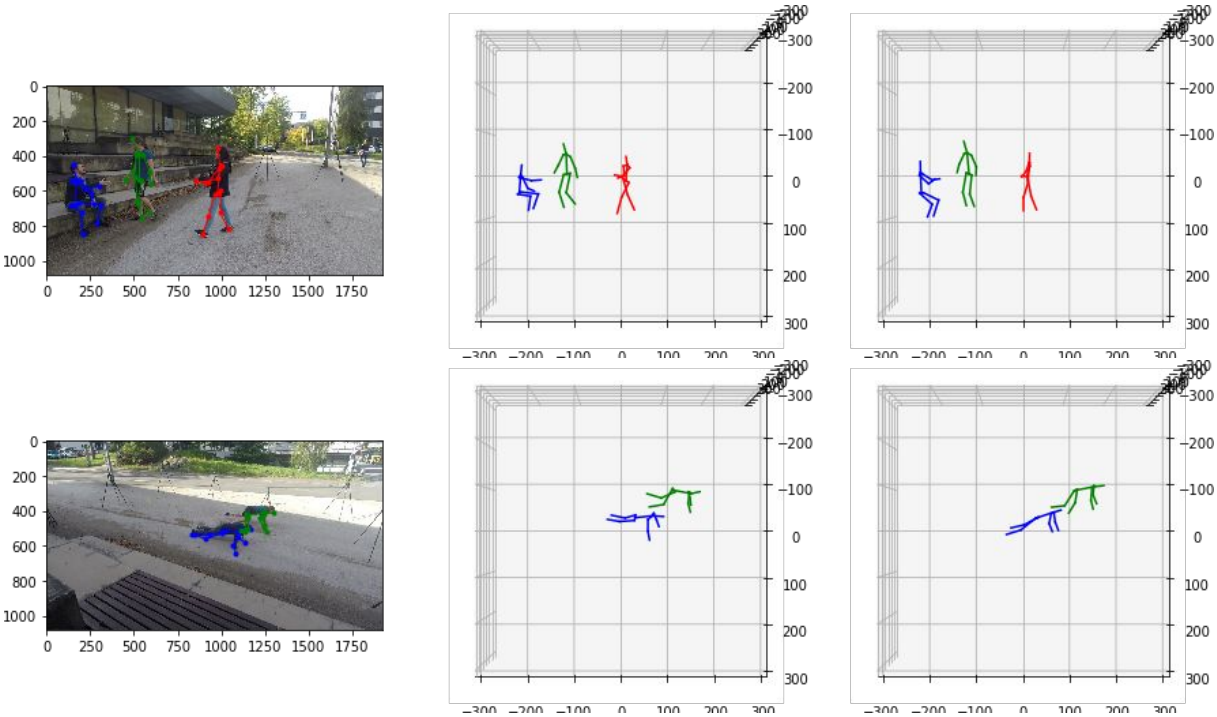


Figure 4.7 – Poses 3D prédites par l'approche proposée sur deux images de MuPoTS-3D. Première colonne : image d'entrée, deuxième colonne : poses 3D de vérité terrain, troisième colonne : poses 3D estimées. Les tailles des personnes et les translations de vérité de terrain ont été utilisées pour la visualisation des poses.

L'ensemble JTA contient des images synthétiques mais néanmoins réalistes de scènes avec des fonds, des conditions de lumière et météorologiques, des angles de vues et des poses variés, et des dynamiques d'échelle des personnes dans l'image beaucoup plus étendues et des densités de personnes beaucoup plus importantes que celles représentées dans les ensembles CMU-Panoptic et MuPoTS-3D. Ainsi l'évaluation de la performance de notre approche sur JTA donne une idée de sa faculté à détecter et estimer les poses 3D de personnes très grandes et très petites dans l'image, et à gérer les situations d'occultations. Une étude de la précision de localisation 3D relative de chaque articulation, en fonction des distances d'observation, donne les résultats présentés dans le Tableau 4.3. L'inférence multi-échelles est appliquée avec 3 niveaux de résolution. On constate des performances assez homogènes sur une grande gamme de distances, entre 0 et plus de 40m, avec une 3DPCK variant entre 41 et 68. De façon assez logique et intuitive, les articulations les

plus stables sont mieux localisées (hanches, colonne vertébrale, alors que les articulations les plus mobiles et plus sujettes aux occultations (poignets, chevilles) sont moins bien localisées.

Distance à la caméra	tête	cou	clavicules	épaules	coudes	poignets	colonne vertébrale	hanches	genoux	chevilles	toutes
>0	41.1	44.6	44.9	33.8	27.2	19.0	74.4	73.9	25.7	8.9	43.9
<10m	68.1	48.1	48.5	37.5	39.5	30.6	94.2	94.0	29.0	7.3	55.8
>10m et <20m	76.2	115.4	75.9	62.9	55.2	40.4	98.5	93.1	46.8	17.8	68.4
>20m et <30m	61.0	70.6	67.8	46.5	33.5	20.8	97.5	85.8	30.2	11.1	57.8
>30m et <40m	48.0	60.9	56.0	30.3	19.8	11.0	95.7	79.8	20.8	6.8	49.3
>40m	39.2	50.1	45.3	18.0	11.1	6.0	89.7	72.1	13.1	4.5	41.7

TABLEAU 4.3 – 3DPCK par articulation, sur l'ensemble de données JTA. L'inférence multi-échelles est appliquée pour gérer la grande dynamique des échelles des personnes. Les articulations les mieux localisées sont les articulations les plus stables (en vert) et celles qui sont estimées avec une précision plus faibles sont les articulations les plus mobiles (en rouge).

La visualisation des résultats qualitatifs (Figure 4.8) montre d'une part, des estimations de poses cohérentes pour un grand nombre de personnes et dans des situations variées, d'autre part, des erreurs d'estimation de poses 3D plus marquées pour les personnes les plus éloignées et occultées.

4.3.3 Limitations de l'approche *bottom-up*

Les évaluations expérimentales présentées en 4.3.2 montrent que notre méthode se place par ses performances en tête des meilleures approches *single-shot*. Pourtant, nous entrevoyons les limites de la modélisation par cartes de chaleur qui, si elles ne sont pas suffisamment résolues, ne peuvent garantir une discrimination claire des articulations entre personnes proches, surtout pour des personnes de faible résolution dans l'image. La stratégie d'inférence multi-échelles compense quelque peu la perte de résolution, mais ajoute un coût calculatoire non négligeable. De plus, le regroupement des articulations n'est pas parfait, et conduit à des erreurs d'association dans le cas de proximité forte entre deux personnes.

4.4 PandaNet : une approche *top-down single-shot* pour la prédiction de poses 3D

L'argument principal qui nous a incité à préférer une approche *bottom-up* est la rapidité d'exécution de ce type de modèle qui ne nécessite qu'un seul passage dans le réseau de neurones (*single-shot*), quelque que soit le nombre de personnes à détecter et à caractériser en 3D dans l'image. Cependant, une stratégie *top-down* présente l'avantage de garantir une pose 3D pour chaque détection, et évite le problème de regroupement des articulations détectées en squelettes entiers. Donc, dès lors que les personnes sont correctement détectées, on cherchera à estimer la position de leur articulations, même en cas de faible résolution, de chevauchement, d'occultation. Est-il alors possible de concilier les avantages des deux catégories d'approches, c'est-à-dire un estimateur de poses 3D *top-down* rapide ?

4.4.1 Une approche *top-down single-shot* rapide fondée sur une représentation par ancres

PandaNet (BENZINE et al., 2020a) repose sur le principe des détecteurs d'objets *single-shot* utilisant une représentation d'ancres tels que SSD (LIU et al., 2016b), Yolo (REDMON et FARHADI, 2017, 2018) et RetinaNet (LIN et al., 2020). Plus précisément, PandaNet reprend les bases de la méthode LapNet (CHABOT et al., 2020) que nous avons développée au laboratoire, et qui a été présentée en 1.4, et l'étend en ajoutant les tâches d'estimation de poses 2D et 3D, et en généralisant les mécanismes de pondération automatique.

FIGURE 4.8 – Résultats qualitatifs d'estimation de poses 3D, re-projetées dans les images, par notre approche *single-shot bottom-up* sur l'ensemble de test de JTA. Les translations et les tailles des personnes de la vérité terrain sont utilisées pour la visualisation. En regardant de plus près les personnes les plus éloignées, on constate des incohérences sur les poses 3D estimées (dans les zones encadrées en rouge).

Les ancres sont des boîtes rectangulaires prédéfinies, de différentes tailles et rapports de forme, disposées sur toute l'image sous forme de grille. Les prédictions pour la détection (classification de l'ancre et régression de ces coordonnées) et l'estimation des poses 2D et 3D sont réalisées systématiquement pour toutes les ancres, de manière dense. De cette façon, chaque ancre stocke une pose 2D et une pose 3D complètes, comme illustré Figure 4.9.

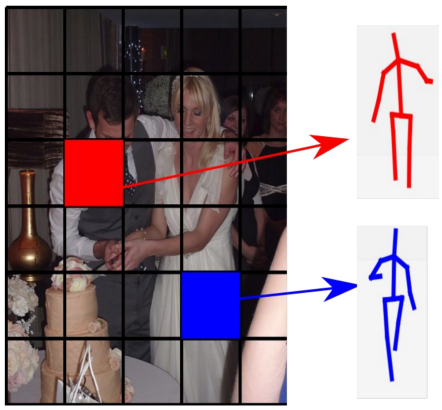


FIGURE 4.9 – Représentation sous forme de grille d'ancres : chaque ancre stocke une pose 2D et une pose 3D complètes associées.

L'architecture de PandaNet est présentée Figure 4.10. De manière similaire à LapNet, elle s'appuie sur une transformation de l'image par une *backbone* d'un CNN constitué d'un encodeur et d'un décodeur qui produisent des cartes de caractéristiques à plusieurs résolutions et niveaux sémantiques et qui sont assemblées en une pyramide de caractéristiques, à la manière des *Feature Pyramid Networks* (FPN) (LIN et al., 2017b). Ces caractéristiques sont ensuite transformées par quatre convolutions 3x3, puis redimensionnées à la plus haute résolution et concaténées, avant de nourrir les quatre têtes du réseau chargées de la prédiction des scores de détection, boîtes englobantes, poses 2D et poses 3D. Chaque tête est formée de quatre convolutions, une convolution sous-pixellique (SHI et al., 2016b) ayant pour but d'améliorer la précision spatiale, et une convolution 1x1 pour obtenir les cartes de prédiction finales.

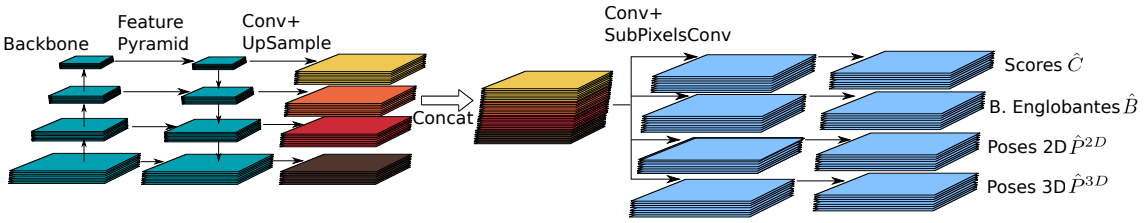


FIGURE 4.10 – Architecture de PandaNet

On désigne par $\mathcal{B} = \{b_n \in \mathbb{R}^4\}$ avec $n \in [1, \ldots, N]$ les boîtes englobantes de vérité terrain correspondant aux N personnes présentes dans l'image I.

L'ensemble des poses 2D et 3D associées à ces personnes sont notées respectivement $\mathcal{P}^{2D} = \{p_n^{2D} \in \mathbb{R}^{2\times N_K}\}$ et $\mathcal{P}^{3D} = \{p_n^{3D} \in \mathbb{R}^{3\times N_K}\}$ avec N_k le nombre d'articulations.

Chaque élément de la grille d'ancres $A \in \mathbb{R}^{H\times W\times N_A\times 4}$, noté $A_{i,j,a}$, est défini par un type d'ancre a et sa position (i, j) dans l'image. La boîte de vérité terrain associée à une ancre donnée est celle qui recouvre le mieux à l'ancre au sens de la métrique Intersection-Over-Union (IoU) :

$$B_{i,j,a} = \arg\max_{b_n \in \mathcal{B}} \text{IoU}(b_n, A_{i,j,a}) \tag{4.4}$$

Les poses 2D et 3D relatives à une ancre $A_{i,j,a}$ associée à la boîte $B_{i,j,a}$ de vérité terrain sont

alors notées :

$$P^{2D}_{i,j,a} = p^{2D}_n \tag{4.5}$$

$$P^{3D}_{i,j,a} = p^{3D}_n \tag{4.6}$$

Pour mesurer la qualité de recouvrement d'une ancre avec sa boîte de vérité terrain associée, on utilise le critère du PONO (*Per-Object Normalized Overlap*) comme dans la méthode LapNet (Chabot et al., 2020) qui, pour une ancre donnée, permet de réhausser la valeur du critère d'IoU avec la vérité terrain, en la divisant par l'IoU maximal entre la boîte de vérité terrain et toutes les ancres associées. On note le PONO entre une ancre $A_{i,j,a}$ et une boîte de vérité terrain associée $B_{i,j,a}$:

$$O_{i,j,a} = \text{PONO}(A_{i,j,a}, B_{i,j,a}) \tag{4.7}$$

et $\mathcal{A}^+$ l'ensemble des ancres positives associées à une vérité terrain dont la valeur de PONO est supérieure à 0.5.

4.4.2 Apprentissage supervisé de PandaNet

Supervision de la prédiction des boîtes englobantes

Comme dans d'autres travaux (Chabot et al., 2020 ; Tian et al., 2019) sur la détection d'objets, la fonction de perte utilisée pour superviser la prédiction des boîtes englobantes repose sur la fonction IoU, qui est une alternative à la fonction SmoothL1 (Lin et al., 2020 ; Redmon et Farhadi, 2017 ; Ren et al., 2015). On cherche à prédire la boîte $\hat{B}_{i,j,a}$ par la déformation d'une ancre $A_{i,j,a} \in \mathcal{A}^+$ associée à $B_{i,j,a}$, de manière à ce qu'elle recouvre au mieux la boîte de vérité terrain $B_{i,j,a}$ au sens du critère de l'IoU :

$$\hat{O}_{i,j,a} = \text{IoU}(B_{i,j,a}, \hat{B}_{i,j,a}) \tag{4.8}$$

La fonction de perte de localisation à minimiser est alors définie de la manière suivante :

$$\mathcal{L}_{loc}(i,j,a) = \begin{cases} \left\| 1 - \hat{O}_{i,j,a} \right\|^2, & \text{si } A_{i,j,a} \in \mathcal{A}^+ \\ 0, & \text{sinon} \end{cases} \tag{4.9}$$

Supervision de la prédiction des poses 2D

Le réseau apprend à prédire les poses 2D en maximisant, pour chaque articulation k le recouvrement entre des boîtes carrées unitaires dans le référentiel de l'ancre $S_{i,j,a,k}$ et $\hat{S}_{i,i,j,a,k}$ centrées respectivement sur la position 2D de l'articulation de vérité terrain et celle de l'articulation prédite :

$$\hat{O}^{2D}_{i,j,a,k} = \text{IoU}(S_{i,j,a,k}, \hat{S}_{i,j,a,k}) \tag{4.10}$$

Ce critère est une alternative à la distance euclidienne normalisée par la taille de l'ancre, et a été choisi car il s'est avéré plus stable dans nos expériences d'apprentissage.

Ainsi, la fonction de perte à minimiser relative à la supervision des poses 2D s'écrit pour toutes les articulations k de chaque ancre $A_{i,j,a}$:

$$\mathcal{L}_{2D}(i,j,a,k) = \begin{cases} \left\| 1 - \hat{O}^{2D}_{i,j,a,k} \right\|^2, & \text{si } A_{i,j,a} \in \mathcal{A}^+ \\ 0, & \text{sinon} \end{cases} \tag{4.11}$$

Supervision de la prédiction de pose 3D

Enfin, le modèle est entraîné à prédire les poses 3D en minimisant la distance euclidienne entre les positions 3D des articulations prédites et celles des articulations de vérité terrain. La fonction de perte associée est formulée par :

$$\mathcal{L}_{3D}(i,j,a,k) = \begin{cases} \left\| P^{3D}_{i,j,a,k} - \hat{P}^{3D}_{i,j,a,k} \right\|^2, & \text{si } A_{i,j,a} \in \mathcal{A}^+ \\ 0, & \text{sinon} \end{cases} \tag{4.12}$$

où $P^{3D}_{i,j,a,k}$ sont les coordonnées de l'articulation k pour la pose 3D associée à l'ancre $A_{i,j,a}$ et $\hat{P}^{3D}_{i,j,a,k}$ les coordonnées 3D prédites pour cette même articulation.

Sélection des ancres par la qualité de l'estimation de pose pour la classification

Lorsque les personnes sont proches dans l'image, on observe souvent des zones d'ambiguïté entre les personnes, c'est-à-dire qui ne correspondent pas de manière claire à l'une ou l'autre des personnes à cause du chevauchement entre les boîtes englobantes. Dans ces zones, les ancres responsables de la détection et de l'estimation de pose sont associées à une seule boîte de vérité terrain mais peuvent contenir des parties de l'image englobées par les autres boîtes de vérité terrain proches. Par conséquent, l'estimation de pose ne sera pas fiable dans ces configurations. Pour résoudre ce problème, nous proposons un mécanisme de sélection des ancres qui écarte les ancres ambiguës en utilisant une information de la qualité de prédiction des poses 2D par le réseau de neurones. Cette qualité de prédiction des poses 2D est mesurée par le chevauchement des poses 2D $\hat{O}^{2D}_{i,j,a}$ qui est la moyenne des prédictions $\hat{O}^{2D}_{i,j,a,k}$ définies pour toutes les articulations.

Les labels $C_{i,j,a}$ de classification des ancres sont alors définis comme suit :

$$C_{i,j,a} = \begin{cases} 1 & \text{si } O_{i,j,a} \times \hat{O}^{2D}_{i,j,a} > 0.5 \\ 0 & \text{sinon} \end{cases} \tag{4.13}$$

L'effet souhaité est que des valeurs faibles du critère $\hat{O}^{2D}_{i,j,a}$ favorisent le classement de l'ancre en ancre négative.

La fonction de perte de classification est alors égale, pour chaque position (i,j) et chaque ancre a, à :

$$\mathcal{L}_{cls}(i,j,a) = \mathcal{H}(C_{i,j,a}, \hat{C}_{i,j,a}) \tag{4.14}$$

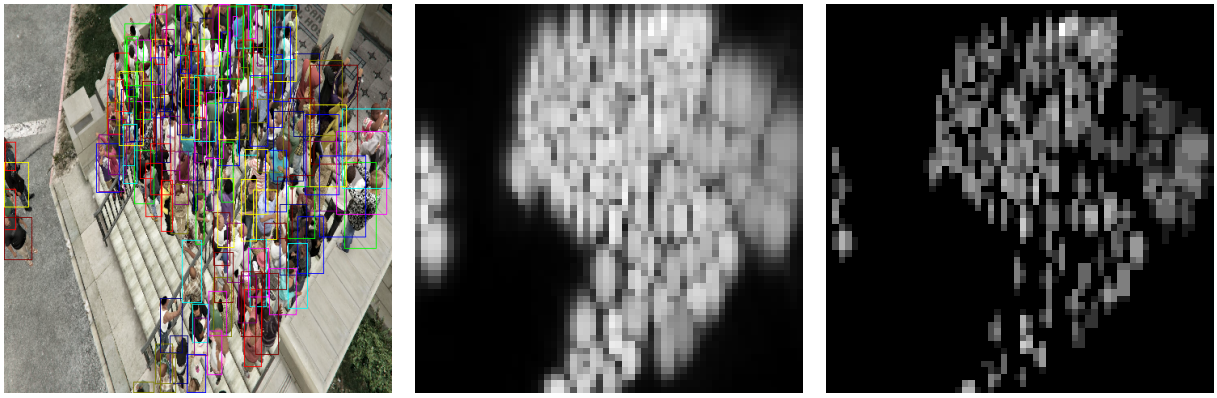
où $\mathcal{H}$ est l'entropie croisée.

Fonction de perte totale et stratégie de pondération automatique

La fonction de perte totale est la somme pondérée des fonction de pertes associées aux tâches de classification (notée cls), de localisation (notée loc), de prédiction des poses 2D (notée $2D$) et des poses 3D (notée $3D$). Reprenant les principes de pondération automatique issus de la formulation de l'incertitude homoscédastique utilisée pour les modèles multi-tâches (Chabot et al., 2020 ; Kendall et al., 2018), nous introduisons des paramètres de pondération pour :

- chaque tâche (λ_{task}),
- chaque ancre (λ^a_{task}),
- chaque articulation ($\lambda^{a,k}_{task}$).

avec $task = \{cls, loc, 2D, 3D\}$. Des termes de régularisation en $log\left(\frac{1}{\lambda_{task}}\right)$, $log\left(\frac{1}{\lambda^a_{task}}\right)$, $log\left(\frac{1}{\lambda^{a,k}_{task}}\right)$ sont ajoutés pour éviter de converger vers des valeurs de pondération nulles. Tous les paramètres



(a) Image de JTA Ext avec boîtes de vérité terrain (b) PONO O (c) $O \times \hat{O}^{2D}$

FIGURE 4.11 – Effet de la sélection d'ancres utilisant la qualité de prédiction des poses 2D sur une scène dense de JTA Ext. La densité de personnes dans l'image 4.11a est très élevée, provoquant d'importants chevauchements entre boîtes englobantes. La carte de PONO O calculée (Figure 4.11b) s'active dans toutes les zones de chevauchement où les ancres sont ambiguës. La stratégie de sélection d'ancres proposée (Figure 4.11c) filtre les ancres ambiguës avec une carte de supervision plus éparse.

λ sont optimisés durant l'entraînement par rétropropagation en même temps que les autres paramètres du modèle. On peut interpréter les λ comme l'inverse de la variance des bruits de prédiction en supposant un modèle de bruit gaussien. Cette stratégie de pondération automatique permet d'éviter une étape explicite de recherche d'hyper-paramètres coûteuse en temps d'apprentissage à cause de leur combinatoire. Par rapport à CHABOT et al., 2020, l'intérêt de généraliser la pondération automatique aux articulations est de modéliser de façon plus spécifique l'incertitude de prédiction de chaque type d'articulation, les articulations mobiles étant en général plus difficiles à estimer avec précision que les articulations plus stables.

Ainsi les différentes fonctions de perte pondérées s'écrivent :

$$\mathcal{L}_{cls} = \frac{\lambda_{cls}}{HWN_A} \sum_a \lambda_{cls}^a \sum_{i,j} \mathcal{L}_{cls}(i,j,a) + \log\left(\frac{1}{\lambda_{cls}}\right) + \frac{1}{N_A}\sum_a \log\left(\frac{1}{\lambda_{cls}^a}\right) \tag{4.15}$$

$$\mathcal{L}_{loc} = \frac{\lambda_{loc}}{N^+} \sum_a \lambda_{loc}^a \sum_{i,j} \mathcal{L}_{loc}(i,j,a) + \log\left(\frac{1}{\lambda_{loc}}\right) + \frac{1}{N_A}\sum_a \log\left(\frac{1}{\lambda_{loc}^a}\right) \tag{4.16}$$

$$\mathcal{L}_{2D} = \frac{\lambda_{2D}}{N_K N^+} \sum_{i,j,a,k} \lambda_{2D}^{a,k} \mathcal{L}_{2D}(i,j,a,k) + \log\left(\frac{1}{\lambda_{2D}}\right) + \frac{1}{N_K N_A}\sum_{a,k} \log\left(\frac{1}{\lambda_{2D}^{k,a}}\right) \tag{4.17}$$

$$\mathcal{L}_{3D} = \frac{\lambda_{3D}}{N_K N^+} \sum_{i,j,a,k} \lambda_{3D}^{a,k} \mathcal{L}_{3D}(i,j,a,k) + \log\left(\frac{1}{\lambda_{3D}}\right) + \frac{1}{N_K N_A}\sum_{a,k} \log\left(\frac{1}{\lambda_{3D}^{k,a}}\right) \tag{4.18}$$

La fonction de perte totale est finalement définie par :

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{loc} + \mathcal{L}_{2D} + \mathcal{L}_{3D} \tag{4.19}$$

4.4.3 Résultats expérimentaux

Comme pour la méthode précédente, PandaNet est évalué sur les ensembles de données JTA (Fabbri et al., 2018), CMU-Panoptic (Joo et al., 2017) et MuPoTS-3D (Mehta et al., 2018). Les détails de ces expériences sont donnés dans le mémoire de thèse d'Abdallah Benzine (Benzine, 2020). Les résultats principaux sont décrits ci-après.

Les expériences effectuées sur l'ensemble JTA montrent :

— l'intérêt de la stratégie de sélection des ancres reposant sur les poses 2D (résultats de l'étude ablative donnés dans le Tableau 4.4),

— l'impact positif de la méthode de pondération automatique sur les tâches, les ancres et les articulations (Tableau 4.5),

— une performance d'estimation de pose 3D de PandaNet bien supérieure à celle de la méthode *bottom-up* présentée en 4.3 sur cet ensemble de données, quelles que soient les distances d'observation (Tableau 4.6) et les articulations considérées (Tableau 4.7), établissant un nouvel état de l'art en estimation de poses 3D sur l'ensemble de données JTA.

Ces résultats sur JTA indiquent que l'approche *top-down* et les différentes contributions implémentées dans PandaNet apportent une meilleure robustesse et une meilleure précision d'estimation de poses 3D pour des scènes complexes, avec un nombre et des densités de personnes très variables, et une grande dynamique d'échelle.

Type de sélection d'ancres	AP	3DPCK
Aucune	84.1	80.7
Reposant sur les boîtes (Chabot et al., 2020)	85.1	81.9
Reposant sur les poses 2D (PandaNet)	**85.3**	**83.2**

Tableau 4.4 – Influence de la stratégie de sélection d'ancres reposant sur les poses 2D évaluée sur JTA. Les modèles sont entraînés avec la méthode de pondération automatique des fonctions de pertes. On reporte ici la précision moyenne de la détection (AP) et la précision de l'estimation de pose 3D (3DPCK).

tâche	ancre	articulation	AP	3DPCK
appris	1	1	84.1	80.8
appris	appris	1	85.2	81.7
appris	appris	appris	**85.3**	**83.2**

Tableau 4.5 – Influence de la pondération automatique de fonctions de perte évaluée sur JTA. 'tâche', 'ancre' et 'articulation' représentent le type de pondération λ. Les modèles sont entraînés avec la stratégie de sélection d'ancres reposant sur la pose 2D. On reporte ici la précision moyenne de la détection (AP) et la précision de l'estimation de pose 3D (3DPCK).

Distance	<10	10-20	20-30	30-40	>40	All
Bottom-up	55.8	68.4	57.8	49.3	41.7	43.9
PandaNet	**95.6**	**93.7**	**87.3**	**80.5**	**71.2**	**83.2**

Tableau 4.6 – 3DPCK sur l'ensemble de données JTA en fonction de la distance à la caméra (exprimée en mètres) pour PandaNet et comparaison avec la méthode *bottom-up* présentée dans la section 4.3 (Benzine et al., 2020b). La précision de l'estimation de pose 3D est clairement que celle obtenue avec l'approche précédente, pour toutes les gammes de distances.

Méthode	tête	cou	clavicules	épaules	coudes	poignets	colonne vertébrale	hanches	genoux	chevilles	toutes
Bottom-up	41.1	44.6	44.9	33.8	27.2	19.0	74.4	73.9	25.7	8.9	43.9
PandaNet	**92.7**	**99.1**	**97.0**	**78.4**	**72.1**	**60.1**	**99.9**	**87.8**	**71.8**	**58.0**	**83.2**

TABLEAU 4.7 – 3DPCK par articulation sur l'ensemble de données JTA pour PandaNet, comparée à celle obtenue avec la méthode bottom-up présenté dans la section 4.3 (BENZINE et al., 2020b). La précision est bien améliorée pour l'ensemble des articulations, que ce soit pour les articulations les plus mobiles (en rouge) que pour les articulations plus stables (en vert).

L'évaluation de PandaNet sur l'ensemble CMU-Panoptic confirme sa capacité à traiter efficacement des scènes avec des interactions complexes entre les personnes. Sur cet ensemble, PandaNet atteint des performances nettement supérieures aux approches concurrentes (ZANFIR et al., 2018a,b) y compris notre approche *bottom-up* précédente (BENZINE et al., 2020b) (Tableau 4.8).

Méthode	Mean
ZANFIR et al., 2018a	153.4
ZANFIR et al., 2018b	72.1
Bottom-up (BENZINE et al., 2020b)	68.5
PandaNet	**42.7**

TABLEAU 4.8 – MPJPE (en mm) pour l'ensemble de données CMU-Panoptic. PandaNet bat les approches concurrentes, y compris notre approche précédente (BENZINE et al., 2020b)

Méthode		Avg 3DPCK
Two-Stage	LCR-Net (ROGEZ et al., 2017)	53.8
	LCR-Net++ (ROGEZ et al., 2019)	70.6
	MOON et al., 2019a	**81.8**
Single-Shot	MEHTA et al., 2018	65.0
	XNect (MEHTA et al., 2019)	72.1
	Bottom-up (BENZINE et al., 2020b)	67.5
	PandaNet	**72.6**

TABLEAU 4.9 – 3DPCK sur l'ensemble MuPoTS-3D. La méthode la plus performante est celle de MOON et al., 2019a qui est une approche *two-stage*, mais PandaNet est la meilleure méthode *single-shot*.

PandaNet se révèle lors de l'évaluation sur MuPoTS-3D la meilleure méthode *single-shot* alors que l'approche *top-down* de MOON et al., 2019a se classe première. Cependant, cette dernière est beaucoup plus lente à cause de l'enchaînement des étapes de détection et d'estimation des poses 3D, surtout quand le nombre de personnes à traiter est grand. Une évaluation du temps de calcul sur JTA, ensemble choisi en raison du grand nombre de personnes présentes dans certaines scènes (jusqu'à 60 personnes), confirme la supériorité de PandaNet sur la méthode de ibid. en terme de rapidité d'exécution (Figure 4.12). Le temps de calcul de PandaNet augmente très peu avec le nombre de personnes, contrairement à l'autre approche.

La figure 4.13 montre des résultats qualitatifs de poses 3D prédites par PandaNet sur des images de MuPoTS-3D, COCO et JTA, et illustre ses facultés de généralisation à des poses diverses.

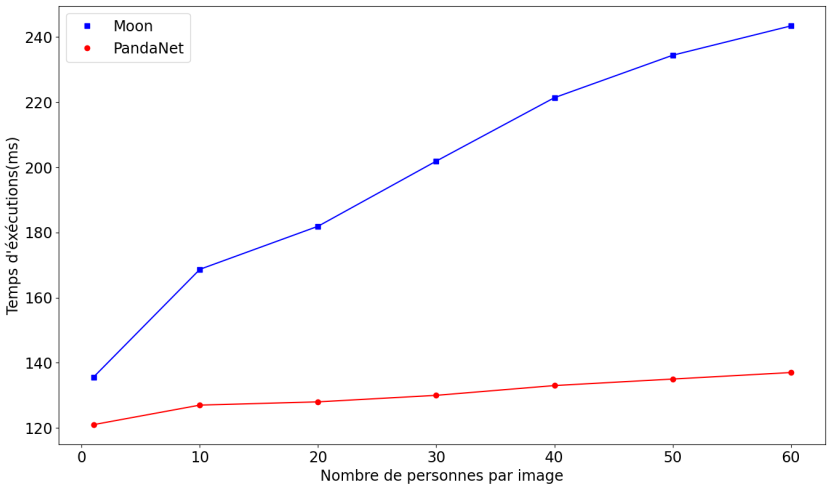


FIGURE 4.12 – Comparaison du temps d'exécution de PandaNet et de celui de l'approche de MOON et al., 2019a en fonction du nombre de personnes par image. Les temps sont évalués sur JTA, les modèles sont exécutés sur une carte GPU Nvidia Titan X.

4.5 Vers l'estimation de poses 3D absolues

PandaNet n'estime que des poses 3D relatives, le référentiel local étant pris à la position d'une articulation racine, habituellement le pelvis. Or, dans le cas multi-personnes, il est important de pouvoir replacer ces poses 3D dans un référentiel commun global associé à la scène. Cette information est indispensable si la localisation des personnes dans l'environnement est recherchée en plus de leur posture et pour analyser d'éventuelles interactions entre elles.

Si la matrice de calibration intrinsèque K de la caméra et la taille des personnes sont connues ou fixées, on peut estimer la pose 3D absolue à partir de la pose relative en cherchant la translation qui minimise l'erreur de reprojection dans l'image (DABRAL et al., 2019 ; ROGEZ et al., 2017). Pour ces méthodes, les erreurs sur la taille des personnes et la calibration intrinsèques se répercutent sur la localisation 3D absolue. Un autre moyen consiste à estimer un plan du sol et d'utiliser la position des pieds sur le sol (MOON et al., 2019a). Malheureusement les pieds ne sont pas toujours visibles.

Contrairement aux méthodes géométriques, FABBRI et al., 2020 proposent avec la méthode LoCO d'apprendre à prédire directement les positions 3D absolues des articulations avec un réseau de neurones produisant une représentation volumétrique des poses. Ainsi l'aspect visuel des personnes et du contexte est pris en compte pour estimer la localisation 3D. Cependant, la méthode dépend des calibrations des caméras utilisées durant l'apprentissage, à travers les points de vue dans les images.

Nous choisissons plutôt d'apprendre les poses 3D absolues à partir des poses 2D et des poses 3D relatives fournies par PandaNet dans une extension de la méthode que nous appelons Absolute PandaNet. La méthode reprend le principe de l'approche MonoLoco (BERTONI et al., 2019) qui apprend à prédire les poses 3D absolues à partir des poses 2D dans l'image avec un réseau de neurones léger, mais en diffère par l'ajout, en entrée du réseau prédicteur de poses 3D absolues, l'information des poses 3D relatives. L'architecture d'Absolute PandaNet est présentée dans la Figure 4.14. Pour rendre les poses 2D indépendantes de la calibration de la caméra, on les normalise par la matrice intrinsèque. L'architecture de Absolute PandaNet est décrite dans la Figure 4.14. $\hat{d}$ étant la distance prédite et d la distance vérité terrain, la fonction de perte à minimiser est la même que celle proposée dans l'approche MonoLoco :

FIGURE 4.13 – Résultats qualitatifs de poses 3D prédites par PandaNet. Première ligne : sur deux images de MuPoTS-3D, deuxième ligne : sur deux images de JTA, troisième ligne : sur une image de JTA. Pour les besoins de la visualisations, les translations de la vérité terrain sont utilisées pour passer des poses relatives aux poses absolues.

$$L_{ADEM} = \frac{|1 - \frac{\hat{d}}{d}|}{\alpha} + \log \alpha \tag{4.20}$$

En pratique, on prédit plutôt $b = \log \alpha$ pour des raisons de stabilité numérique.

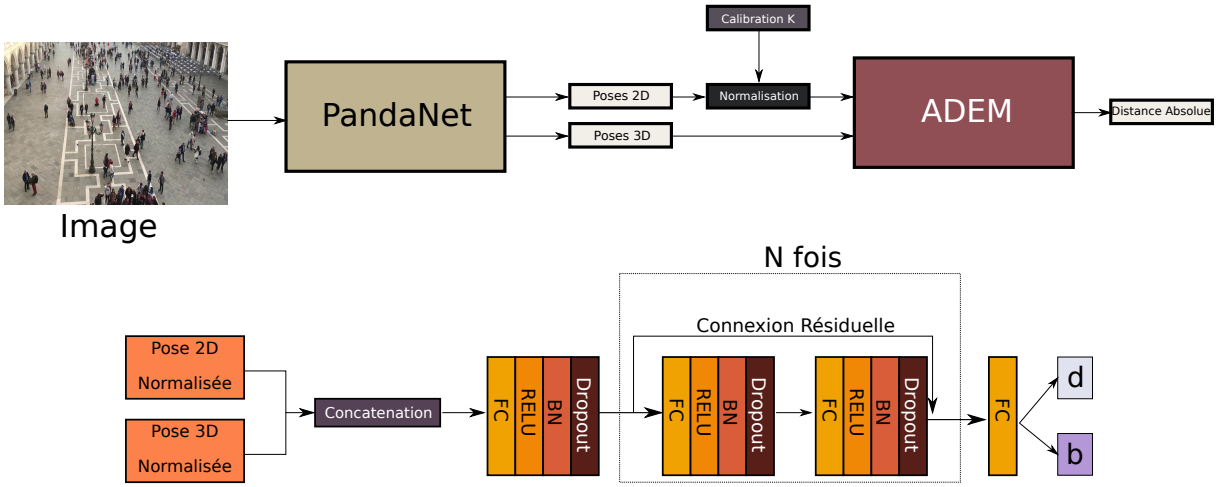


FIGURE 4.14 – Architecture de Absolute PandaNet. Les poses 2D sont normalisées par la calibration intrinsèque de la caméra. Le module ADEM (*Absolute Distances Estimation Module*) prend en entrée les poses 2D normalisées par la calibration intrinsèque de la caméra et les poses 3D relatives ; c'est un réseau entièrement connecté, composé d'une succession de modules, avec une connexion résiduelle permettant de passer directement de l'entrée d'un module au module suivant. En sortie, ADEM prédit la distance absolue d et un terme b de régularisation qui modélise l'incertitude de la prédiction. Le nombre de modules et la taille des couches sont déterminés expérimentalement.

Les résultats d'évaluation sur JTA donnés dans (BENZINE, 2020) indiquent une meilleure précision d'estimation des distances absolues quand l'algorithme prend en entrée à la fois les poses 2D et 3D, que quand seules les poses 2D sont utilisées. D'autre part, contrairement à l'approche de FABBRI et al., 2020, les performances sont maintenues quand on applique des changements de résolution sur les images. Sur MuPoTS-3D, notre approche Absolute PandaNet dépasse celle de MOON et al., 2019a.

4.6 Conclusion et perspectives

Dans ce chapitre, nous avons abordé la modélisation 3D des personnes sous l'angle de de l'estimation de poses 3D à partir d'images RGB. Après avoir évoqué les enjeux applicatifs de l'estimation de pose pour l'analyse de comportement, nous justifions les orientations prises dans ces travaux de recherche et présentons trois contributions principales. Ces contributions sont de nouvelles approches de réseaux de neurones profonds optimisés par apprentissage supervisé, qui tentent de répondre aux défis de l'estimation de poses 3D multi-personnes robuste, précise et rapide *in-the-wild*, d'abord en limitant la problématique à la prédiction de poses relatives, puis en élargissant aux poses absolues. Les résultats obtenus sur différents ensembles de données sont très encourageants et montrent un positionnement très compétitif de nos approches par rapport à la concurrence.

Ces travaux font aussi apparaître que même si des premières réponses ont été apportées, beaucoup reste à faire. Le premier problème est le manque de généralisation dans les situations réelles variées, les capacités de ce type d'approches étant intimement liées aux données d'appren-

tissage disponibles. Contrairement à d'autres tâches de vision qui bénéficient déjà de très grands ensembles de données généralistes, il est toujours difficile de collecter de grandes quantités de données réelles annotées en poses 3D et présentant des situations riches et variées. La démocratisation des capteurs 3D changera peut-être la donne dans les années à venir. Mais d'autres pistes de recherche méthodologiques intéressantes pourraient être envisagées.

Nous avons vu avec l'ensemble JTA que la simulation permet de générer des images de synthèse très variées, annotées *by-design*, en changeant les paramètres contrôlant l'environnement, les points de vue, les densités de personnes, leur apparence, leur pose, etc. Beaucoup de données de synthèse issues du monde de l'animation sont aujourd'hui disponibles, et présentent de ce point de vue un grand intérêt pour l'entraînement de modèles de reconnaissance visuelle. Mais l'écart de domaine entre les images synthétiques, même avec un bon niveau de réalisme, et les images réelles est tel que les performances obtenues par application directe d'un modèle appris uniquement sur des données de synthèse sur des données réelles sont presque toujours décevantes. Dès lors, peut-on trouver des moyens de transférer efficacement un apprentissage réalisé sur des données virtuelles vers des données réelles ? Cette question touche de nombreux sujets, parmi lesquels l'adaptation de domaine, l'apprentissage avec peu d'exemples, l'apprentissage de représentations des poses 3D invariantes à l'apparence, l'utilisation de modèles génératifs... Des informations supplémentaires de géométrie telles que l'orientation des articulations et le contact pourront également être exploitées de façon utile. C'est dans ce cadre qu'en 2021 débute au laboratoire la thèse de Mohamed Tahiri qui portera sur l'apprentissage par transfert du virtuel au réel pour l'estimation de poses 3D et que j'aurai plaisir de co-encadrer.

Un autre aspect important est que l'information temporelle a été jusqu'ici totalement ignorée dans nos travaux sur l'estimation de pose, l'analyse étant faite image par image. Or, lors de l'exécution des gestes, les poses 3D varient de façon continue. Il semble donc tout à fait opportun d'imaginer des méthodes qui puissent tirer parti de la continuité des poses afin d'augmenter à la fois la précision et la robustesse de la prédiction. Dans le cas multi-personnes, cela implique d'apparier au cours du temps les différentes détections, ce qui reviendrait en quelque sorte à travailler sur des approches de suivi multi-cibles en 3D avec l'ajout de l'information de la posture.

La prédiction de poses 3D absolues pourrait également bénéficier d'une meilleure exploitation d'indices visuels dans l'image, la méthode de localisation 3D d'Absolute PandaNet utilisant pour le moment uniquement des informations géométriques de poses. En effet, l'utilisation du contexte permettrait d'estimer des échelles et un positionnement 3D relatif entre les personnes de façon plus cohérente.

Alors qu'une précision d'une quinzaine de centimètres sur les articulations est suffisante pour un certain nombre d'applications, elle s'avère tout à fait insuffisante pour caractériser et qualifier un geste sportif ou un geste technique. Ainsi dans le contexte de l'analyse de la performance sportive, nous nous intéressons au développement de nouveaux outils permettant d'aider les entraîneurs à analyser les postures des athlètes dans un objectif de réduire les blessures et améliorer la performance. Partant des approches développées dans la thèse d'Abdallah Benzine, le laboratoire a commencé à investir des efforts de recherche sur ces questions, dans le cadre du projet de recherche ANR FULGUR (19-STHP-0003) qui porte sur la planification individualisée de la charge d'entraînement adaptée aux propriétés musculaires pour réduire l'incidence des blessures en sprint. Dans ce projet, nous recherchons des moyens d'augmenter la précision de l'estimation de la posture du coureur pendant les différentes phases de sa course.

Dans le même programme de recherche "Sport de très haute performance" consacré au développement de solutions innovantes pour l'optimisation de la préparation des athlètes français en vue des Jeux Olympiques de Paris en 2024, notre laboratoire intervient dans le projet ANR TEAM-sports (19-STHP-0006). Ce projet met l'accent sur l'analyse comportementale des joueurs dans les sports collectifs, en particulier les interactions sociales et l'attitude des joueurs sur le terrain. Là encore, l'information de poses 3D est fondamentale pour reconnaître les différentes

situations, mais les difficultés résident plus ici dans la gestion des interactions complexes avec de nombreux contacts, comme dans le rugby.

D'autres chercheurs du laboratoire, dans le cadre d'un projet de transfert technologique dans le domaine de la sécurité au sein du VisionLab, laboratoire commun entre le CEA et Thales, travaillent sur la reconnaissance efficace des interactions entre les personnes et les objets (Chafik et al., 2020), et entre les personnes. Une voie de recherche intéressante serait d'intégrer aux approches en cours de développement une modélisation de la pose 3D des personnes afin d'en améliorer les performances.

Conclusion générale et perspectives

Dans le vaste champ d'exploration de l'analyse de scènes visuelles, nous avons présenté quelques-unes de nos contributions qui ont concerné des thématiques de recherche aussi variées que la détection d'objets, la ré-identification des personnes, la détection d'événements anormaux, l'analyse de scènes de foule, et la caractérisation des personnes par l'estimation de la pose 3D. Du plus bas niveau - la modélisation des objets - vers le plus haut niveau d'interprétation sémantique - l'analyse de comportement -, nous avons conçu dans ce domaine plusieurs approches nouvelles, développé des briques de base et esquissé des pistes de réflexion, que nous espérons utiles. Nos travaux témoignent d'une recherche toujours tournée vers les applications concrètes et qui s'appuie sur les progrès scientifiques, les innovations technologiques toujours plus nombreuses et plus rapides. Plusieurs de nos contributions sont exploitées aujourd'hui dans des projets de maturation et de transfert technologique auprès de partenaires industriels du CEA.

Les avancées extraordinaires de l'intelligence artificielle,notamment en vision par ordinateur, sont à l'origine de technologies logicielles qui sont aujourd'hui présentes dans les produits et les services que nous utilisons quotidiennement. Cela était encore difficilement imaginable il y a quelques années et pourrait faire penser que la fin de l'histoire est proche. Les recherches en analyse de scène ont certes quelque peu évolué dans la forme avec un courant actuel dominant qui consiste à revisiter la quasi-totalité des problématiques avec des méthodes d'apprentissage profond. Les algorithmes de reconnaissance visuelle sont maintenant capables de répondre à un certain nombre de cas d'usage.

Alors qu'ils peuvent dépasser la performance humaine dans certaines applications comme la reconnaissance faciale à grande échelle, leurs capacités d'analyse de scènes complexes, changeantes et inconnues restent limitées. Les modèles manquent cruellement de fiabilité lorsqu'ils sont appliqués dans un domaine assez différent du domaine des données d'entraînement, même pour les fonctions de bas niveau, comme la détection ou la segmentation sémantique. Tandis que certaines applications, comme celles du web, se contentent d'une robustesse et d'une précision approximatives, d'autres comme les systèmes autonomes et les systèmes critiques exigent des performances proches du zero faute. D'abord promis pour une introduction dès 2020, la date de mise en service des véhicules autonomes comme les robo-taxis est sans cesse reportée par les différents constructeurs, en partie pour des raisons réglementaires mais surtout pour des raisons techniques. En vidéo-surveillance, le suivi temporel et la ré-identification des personnes dans des conditions réelles et sur le long terme restent un défi majeur et un objectif à atteindre pour les systèmes automatiques.

Un autre problème fondamental est que la plupart des modèles de perception sont incapables d'évaluer avec précision le niveau de confiance qu'on pourrait accorder à leur prédiction et de se rendre compte d'eux-mêmes d'erreurs commises afin de corriger leur exécution. Ils sont aussi inaptes à s'adapter rapidement et efficacement à de nouvelles situations. L'apprentissage est presque toujours réalisé une seule fois hors-ligne, conduit et validé par des experts. Dans bien des cas, il serait trop risqué de laisser un modèle s'optimiser et s'adapter de manière autonome, sans compter les difficultés de mise en œuvre liées aux ressources de calcul nécessaires à un auto-apprentissage.

L'interprétation de plus haut niveau sémantique, comme l'analyse du comportement, manque encore de maturité. Les approches macroscopiques ne donnent qu'une information partielle et peu précise des comportements observés. Il faudra encore faire progresser les méthodes de caractérisation fine des comportements individuels, de reconnaissance des interactions entre les personnes, et des interactions entre les personnes et les objets. Ces méthodes dépendent largement de la qualité des fonctions de bas niveau. Pour mieux tenir compte des éléments de contexte, l'analyse de haut niveau devra aussi intégrer des logiques de raisonnement indispensables à une compréhension cohérente et juste des situations et des événements.

Directions de recherche

De nombreux défis sont à donc relever en analyse de scène. Les pistes multiples d'exploration augurent de travaux de recherche futurs passionnants. Le laboratoire de Vision et d'Apprentissage pour l'analyse de scène s'est engagé sur une feuille de route scientifique de plusieurs années, qui s'articule sur plusieurs axes :

Compréhension de haut niveau A mesure que les modèles de reconnaissance visuelle gagnent en performance, on peut ambitionner d'élever le niveau sémantique d'interprétation des scènes pour aller progressivement de la description simple restreinte à la détermination de la nature des objets, à une caractérisation plus riche, telle que la reconnaissance d'attributs sémantiques. S'agissant des personnes, cela concernera des attributs d'apparence, et des aspects relatifs au comportement tels que la gestuelle, l'attitude, et l'intention. Il faudra aussi trouver des moyens d'élargir l'analyse afin de tendre vers une compréhension plus globale qui incluera les relations entre les éléments de la scène et les relations temporelles entre les événements, par l'exploration de nouveaux modèles de représentation de la dynamique des scènes. L'ajout de la dimension temporelle constitue toujours une difficulté qui est généralement traitée soit en transposant des approches 2D en 3D, ce qui alourdit considérablement le calcul et les stratégies d'apprentissage, soit en travaillant sur des méthodes d'agrégation temporelle. La modélisation de la géométrie 3D des objets et de la scène, qui peut être inférée à partir de données 2D, ouvrira de grandes perspectives de compréhension de scène plus riche et plus complète, à condition de trouver des solutions pour superviser ou auto-superviser l'apprentissage avec des informations 3D.

Unification des approches Les réseaux de neurones profonds offrent un cadre méthodologique générique et flexible pour traiter la majorité des tâches de reconnaissance visuelle. Des liens naturels ont été trouvés entre ces différentes tâches. La détection d'objets et la segmentation sémantique d'images peuvent être traitées avec des appproches similaires, ce qui a donné lieu à la proposition de plusieurs méthodes de segmentation d'instances. A leur tour, segmentation d'instances et segmentation sémantique sont unifiées sous la forme de la segmentation panoptique qui est le niveau le plus complet et le plus précis de la description de la nature des éléments de la scène. Les approches multi-tâches permettent de mettre au point des modèles capables d'effecter des tâches distinctes tout en favorisant le partage de caractéristiques communes, bénéfique pour chacune des tâches. L'intérêt de ces approches paraît évident pour adresser à la fois la reconnaissance des objets et leur caractérisation fine. Pourtant, l'optimisation des modèles multi-tâches n'est pas toujours chose aisée, car il n'existe pas de méthodologie générale pour définir l'importance des tâches, les architectures adaptées, le séquencement des étapes d'apprentissage, l'équilibre des données d'apprentissage. Il sera également intéressant de voir à quel point des tâches apparemment plus éloignées pourront être combinées de manière étroite. L'une des difficultés est que les pré-requis sur les données d'apprentissage peuvent être divergents : généralité vs spécificité, information locale vs information de contexte, information spatiale vs information temporelle.

Robustesse et adaptation des modèles La confiance que l'on accorde aux systèmes d'intelligence artificielle est en premier lieu fondée sur leur fiabilité et leur robustesse dans les environnements et les contextes où ils sont déployés. Il existe tellement de facteurs de variabilité dans les images qu'il est toujours difficile de disposer de données d'apprentissage suffisamment représentatives de l'ensemble des situations observables et de concevoir des modèles qui généralisent correctement à cet ensemble. Une façon de pallier le problème de manque de données réelles serait de trouver des solutions de génération de données synthétiques et d'étudier des moyens de transposer la connaissance apprise de ces données synthétiques pour améliorer les capacités des modèles. Une deuxième difficulté de taille est que la spécification du domaine d'application n'est pas forcément connue lors de la phase de conception, ou que les données annotées correspondant à ce domaine ne sont pas disponibles. Dans les deux cas se pose le problème d'adaptation de domaine, dont les contraintes sont les écarts de domaine importants (apparence très hétérogène, liée à des modalités, des capteurs, des contextes différents) et l'absence d'informations de label. Pour améliorer la capacité de généralisation et d'adaptation des modèles, plusieurs approches sont à étudier. L'adaptation de domaine non supervisé apportera des éléments de réponse au problème de spécialisation à un domaine cible à partir d'un apprentissage initial dans un domaine source, sans recourir à une nouvelle labellisation. Le pré-apprentissage auto-supervisé pourrait fournir des points de départ plus proche de la solution, ce qui permettrait une adaptation plus facile et plus rapide à un domaine ou une tâche spécifique. L'apprentissage semi-supervisé, en composant entre optimisation sur des données labellisées et des données non labellisées, présente aussi beaucoup d'intérêt pour travailler sur la robustesse et l'adaptation des modèles. Dans ces deux familles d'approches, l'enjeu est de trouver des moyens d'apprendre des modèles de représentation plus expressifs et plus généraux avec créant des méthodes astucieuses d'auto-supervision.

Frugalité des données et des ressources Sur presque toutes les tâches de reconnaissance visuelle, les meilleures performances de l'état de l'art sont obtenues de nos jours avec des méthodes de deep learning et l'apprentissage supervisé. Toutefois, il n'est pas toujours possible de constituer au préalable d'immenses bases d'images ou de vidéos annotées. L'économie des annotations nous oriente vers les paradigmes d'apprentissage évoqués plus haut (auto-supervision, semi-supervision). Dans d'autres cas, les données brutes ne peuvent être disponibles en quantité suffisante pour entraîner les modèles avec les stratégies d'apprentissage classiques. Avec le problème de détection d'anomalies, nous avons vu que les exemples d'anomalies sont rares et difficiles à collecter. Le caractère sensible de certaines données, comme celles du monde de la défense, impose d'imaginer des méthodes capables de traiter ce type de données sans en avoir vu un échantillon représentatif au préalable. La proposition et la maîtrise de nouveaux formalismes d'apprentissage à partir de très peu d'exemples seront des atouts déterminants et stratégiques dans ces domaines. Enfin, les besoins exponentiels de consommation de données et de ressources de calcul soulèvent des questions cruciales autour de l'impact énergétique et écologique du développement des technologies d'intelligence artificielle. Alors, peut-on faire aussi bien et mieux avec moins de données et moins de ressources matérielles ? L'enjeu énergétique nous oblige à intensifier nos efforts pour limiter la complexité des algorithmes et rendre les systèmes plus efficients, par la mise en commun de ressources pour des tâches diverses, par l'hybridation de modèles d'apprentissage profond avec des algorithmes beaucoup moins gourmands, par des stratégies d'optimisation algorithmique.

Maîtrise des workflows d'apprentissage et d'évaluation Le dernier axe concerne les méthodologies de conception et d'évaluation des modèles. La mise au point et l'évolution d'un système d'intelligence artificielle au cours du temps ne se résument pas à une phase d'apprentissage, de validation et de test. Elles correspondent plutôt à un développement

itératif, cyclique, qui met en jeu plusieurs étapes dont aucune n'est à négliger : la collecte et la sélection pertinente de données, en fonction du domaine opérationnel de conception spécifié par l'application et les manques identifiés, la caractérisation des données en termes de diversité, de représentativité et d'éventuels biais, la labellisation des données à automatiser le plus possible en y intégrant des algorithmes de pré-annotation ou d'annotation interactive, l'optimisation des modèles et l'évaluation quantitative de leurs performances par rapport au domaine opérationnel défini. On trouve assez peu de travaux concernant la qualification des données. Les bases de données sont la plupart du temps considérées comme des entrées qui sont peu remises en question. Pourtout, comment savoir si un jeu de données est suffisant et complet pour résoudre la tâche que nous nous sommes fixée ? Est-il possible de se doter de méthodes et d'outils pour en avoir une idée plus précise, de relier les incertitudes des modèles et les caractéristiques intrinsèques des bases déjà exploitées à la nature des données à ajouter dans l'apprentissage et l'évaluation ? Toutes ces questions doivent être considérées avec attention dans nos recherches futures.

Projets de recherche

Pour terminer, je donne ici une description succincte de deux projets de recherche qui viennent de commencer et pour lesquels je vais m'investir plus personnellement pendant les trois ou quatre années à venir.

Apprentissage par transfert du synthétique au réel pour l'estimation de poses 3D humaines

L'analyse de la posture humaine dans les images et les vidéos est une tâche clé pour la compréhension des actions, des interactions et des activités des personnes. Les travaux récents sur l'estimation des poses 3D à partir d'images 2D ont ouvert une voie alternative à l'utilisation de caméras 3D ou d'équipements de capture de mouvements à base de marqueurs. L'estimation de la pose 3D a d'abord été étudiée dans le cas mono-personne (Pavlakos et al., 2017 ; Zhou et al., 2017), puis étendue au multi-personnes, avec une recherche constante de précision améliorée, de robustesse aux occultations et aux variations d'échelles, et d'inférence rapide (Benzine et al., 2020a). L'estimation de la pose 3D absolue est également traitée depuis peu (Moon et al., 2019a).

Ce projet de recherche fait suite aux travaux décrits dans le Chapitre 4 et fait l'objet de la thèse de Mohamed Ayoub Tahiri qui vient de débuter. Un des enseignements importants de l'étude précédente est que beaucoup des progrès réalisés en estimation de poses 3D à partir d'images 2D, ont été rendus possibles notamment grâce à la mise à disposition d'ensembles de données tels que Human3.6M (Ionescu et al., 2014), CMU Panoptic (Joo et al., 2019), MPI-INF-3DHP (Mehta et al., 2017a), MuCo-3D-HP (Mehta et al., 2017b), JTA (Fabbri et al., 2018). Cependant, ces jeux de données sont assez peu représentatifs de la diversité du monde réel. D'ailleurs, l'évaluation de la précision de l'estimation de la pose 3D sur un jeu de données comme Human3.6M semble perdre progressivement de son intérêt : la plupart des méthodes monoculaires de l'état de l'art atteignent une MPJPE de 50mm ou moins en moyenne, faut-il mettre toute son énergie pour spécifiquement diminuer encore la valeur de cette erreur, étant donné la relative simplicité de cette base ?

Il reste difficile d'annoter les poses en 3D pour un ensemble de données de très grande taille. Par conséquent, les approches actuelles sont limitées par leur mauvaise gestion des postures "difficiles" (rarement observées et assez éloignées de celles qui ont été vues pendant l'apprentissage) et des occultations, et leur faible capacité de généralisation. L'idée directrice de ce projet est d'améliorer la généralisation des modèles d'estimation de poses 3D en contournant le problème

du manque de données réelles annotées en 3D. La première idée consiste à exploiter des données synthétiques issues du monde de l'animation 3D et du jeu vidéo, en tirant parti de la possibilité de créer facilement un grand nombre d'exemples variés, et des annotations gratuites car disponibles *by-design*, précises et riches (informations sur l'orientation et la cinématique des articulations, information de contact...) Un travail très intéressant a consisté à utiliser un jeu de données de synthèse (AMASS), généré avec des techniques de capture de mouvement, pour entraîner dans un cadre d'apprentissage antagoniste un modèle capable d'inférer la pose et l'enveloppe 3D des personnes sur des images réelles (Kocabas et al., 2020).

En revanche, les modèles appris sur des données synthétiques se transfèrent mal aux données réelles. L'objectif du projet est donc de proposer des approches capables d'exploiter pleinement les données virtuelles et de résoudre le problème de l'écart de domaine entre le virtuel et le réel, pour améliorer l'estimation de poses 3D dans les images réelles *in the wild*. Les pistes pour l'adaptation de domaine pourront être l'utilisation de modèles génératifs, l'apprentissage semi-supervisé, l'apprentissage de représentations séparant l'apparence des informations de géométrie (désenchevêtrement des représentations). Une autre approche séduisante ne fait intervenir que les annotations de poses 2D en entrée, et met en jeu une auto-supervision géométrique où les poses 3D sont générées, transformées géométriquement, et reprojetées en 2D : ces poses 2D doivent être réalistes et proches de la pose de départ (Chen et al., 2019a). Cette méthode astucieuse pourrait aussi être une belle source d'inspiration pour nos travaux.

Apprentissage auto-supervisé de représentations visuelles pour la segmentation d'images

Le succès de l'apprentissage profond pour effectuer les tâches de vision par ordinateur est en grande partie dû à l'utilisation de grands ensembles de données annotées tels que Imagenet (un million d'images, 1000 classes) (Russakovsky et al., 2015). Non seulement les réseaux de neurones ont permis d'avoir les meilleures performances en classification d'images, mais ces réseaux entraînés à catégoriser les images peuvent être transférés sur d'autres tâches de reconnaissance telles que la segmentation sémantique d'image et la détection d'objets grâce à des jeux de données annotés pour ces tâches (Cordts et al., 2016 ; Lin et al., 2014). Le pré-apprentissage sur Imagenet est devenu de-facto une étape préliminaire presque systématique. Il n'est pas certain que l'utilisation de la base Imagenet soit la solution unique et idéale pour résoudre les problèmes spécifiques dans des domaines variés. Certains se posent la question de l'existence de biais induits par l'entraînement de modèles sur une même base d'images, d'autres s'interrogent sur l'indépendance technologique des intelligences artificielles qui sont produites à partir de ces données. Dans tous les cas, construire une grande base de données annotées requiert des moyens importants.

Dans ce projet de recherche, nous souhaitons proposer des solutions alternatives de pré-apprentissage de réseaux de neurones sans recourir à une grande base d'images annotées. Le principe est d'optimiser le réseau par apprentissage auto-supervisé. Les signaux de supervision sont extraits automatiquement soit en comparant les différentes vues d'un exemple, soit en prédisant une partie manquante de l'information. Des résultats très récents de l'état de l'art ont montré que les méthodes d'auto-supervision de l'apprentissage, telles que SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020) ou encore SwAV (Caron et al., 2020), arrivent à rivaliser avec les méthodes complètement supervisées dans la tâche de classification d'image.

Encore peu explorées sur d'autres tâches de vision, ces approches produisent des représentations globales de l'image en comparant des projections de versions augmentées d'une même image ou d'images différentes. Notre idée est de voir comment ce ce principe pourrait être appliqué aux problèmes de détection d'objets et de segmentation sémantique d'images.

Par ailleurs, un verrou technique réside dans le fait de devoir constituer des mini-batches de grande taille pour obtenir un bon modèle, ce qui implique de maîtriser les techniques d'apprentissage distribué. Nous avons pour objectif de proposer de nouvelles approches d'appren-

tissage auto-supervisé de représentations adaptées à la segmentation sémantique d'images en pré-apprentissage, notamment en exploitant la localité spatiale de l'information visuelle. Les performances obtenues avec un tel pré-apprentissage seront comparées à celles des méthodes utilisant un pré-apprentissage supervisé en classification sur Imagenet. Nous investiguerons également comment ces mécanismes d'auto-supervision pourront être utilisés dans un cadre d'adaptation de domaine avec des données non labellisées. Enfin, nous chercherons à proposer des algorithmes moins gourmands en ressources de calcul et en mémoire que les approches existantes.

Bibliographie

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B. & Vijayanarasimhan, S. (2016). YouTube-8M : A Large-Scale Video Classification Benchmark. *arXiv :1609.08675*.

Afiq, A. A., Zakariya, M. A., Saad, M. N., Nurfarzana, A. A., Khir, M. H. M., Fadzil, A. F., Jale, A., Gunawan, W., Izuddin, Z. A. A. & Faizari, M. (2019). A review on classifying abnormal behavior in crowd scene. *Journal of Visual Communication and Image Representation*, *58*, 285-303.

Ali, S. & Shah, M. (2007). A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 1-6.

An, L., Chen, X., Yang, S. & Bhanu, B. (2016). Sparse representation matching for person re-identification. *Information Sciences*, *355-356*, 74-89.

Bai, S., Tang, P., Torr, P. H. S. & Latecki, L. J. (2019). Re-Ranking via Metric Fusion for Object Retrieval and Person Re-Identification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 740-749.

Bak, S., Corvee, E., Brémond, F. & Thonnat, M. (2010). Person re-identification using spatial covariance regions of human body parts. *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, 435-440.

Ballas, N., Yang, Y., Lan, Z.-Z., Delezoide, B., Preteux, F. & Hauptmann, A. (2013). Space-Time Robust Representation for Action Recognition. *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2704-2711.

Bendali-Braham, M., Weber, J., Forestier, G., Idoumghar, L. & Muller, P.-A. (2021). Recent trends in crowd analysis : A review. *Machine Learning with Applications*, *4*, 100023.

Bengio, S., Weston, J. & Grangier, D. (2010). Label Embedding Trees for Large Multi-Class Tasks. *Conference on Neural Information Processing Systems (NIPS)*.

Benzine, A. (2020). *Estimation de poses 3D multi-personnes à partir d'images RGB* (These de doctorat). Sorbonne Université.

Benzine, A., Chabot, F., Luvison, B., Pham, Q. C. & Achard, C. (2020a). PandaNet : Anchor-Based Single-Shot Multi-Person 3D Pose Estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6856-6865.

Benzine, A., Luvison, B., Pham, Q. C. & Achard, C. (2019). Deep, Robust and Single Shot 3D Multi-Person Human Pose Estimation from Monocular Images. *2019 IEEE International Conference on Image Processing (ICIP)*, 584-588.

Benzine, A., Luvison, B., Pham, Q. C. & Achard, C. (2020b). Single-shot 3D multi-person pose estimation in complex images. *Pattern Recognition*, 107534.

Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G. & Sukthankar, R. (2011). Violence Detection in Video Using Computer Vision Techniques. In P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano & W. Kropatsch (Éd.), *Computer Analysis of Images and Patterns* (p. 332-339). Springer Berlin Heidelberg.

Bertini, M., Bimbo, A. D. & Seidenari, L. (2012). Multi-scale and real-time non-parametric approach for anomaly detection and localization. *Computer Vision and Image Understanding*, *116*(3), 320-329.

Bertoni, L., Kreiss, S. & Alahi, A. (2019). Monoloco : Monocular 3d pedestrian localization and uncertainty estimation. *Proceedings of the IEEE International Conference on Computer Vision*, 6861-6871.

Boiman, O. & Irani, M. (2005). Detecting irregularities in images and in video. *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, *1*, 462-469 Vol. 1.

Boiman, O. & Irani, M. (2007). Detecting Irregularities in Images and in Video. *International Journal of Computer Vision*, *74*(1), 17-31.

Bojanic, D., Bartol, K., Pribanic, T., Petkovic, T., Donoso, Y. D. & Salvi, J. (2019). On the Comparison of Classic and Deep Keypoint Detector and Descriptor Methods. In S. Loncaric, R. Bregovic, M. Carli & M. Subasic (Éd.), *11th International Symposium on Image and Signal Processing and Analysis, ISPA 2019, Dubrovnik, Croatia, September 23-25, 2019* (p. 64-69). IEEE.

Butenuth, M., Burkert, F., Schmidt, F., Hinz, S., Hartmann, D., Kneidl, A., Borrmann, A. & Sirmacek, B. (2011). Integrating pedestrian simulation, tracking and event detection for crowd analysis. *ICCV Workshops*, 150-157.

Cai, L. & Hofmann, T. (2004). Hierarchical document categorization with support vector machines. *ACM International Conference on Information and Knowledge Management (CIKM)*.

Cancela, B., Hospedales, T. & Gong, S. (2014). Open-world Person Re-Identification by Multi-Label Assignment Inference. *Proceedings of the British Machine Vision Conference*.

Cao, Z., Simon, T., Wei, S.-E. & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. In A. Vedaldi, H. Bischof, T. Brox & J.-M. Frahm (Éd.), *Computer Vision – ECCV 2020* (p. 213-229). Springer International Publishing.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. & Joulin, A. (2020). Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *34th Conference on Neural Information Processing Systems, NeurIPS'20*, *33*, 9912-9924.

Carreira, J., Agrawal, P., Fragkiadaki, K. & Malik, J. (2016). Human pose estimation with iterative error feedback. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4733-4742.

Chabot, F., Chaouch, M., Rabarisoa, J., Teulière, C. & Chateau, T. (2017). Deep MANTA : A Coarse-to-Fine Many-Task Network for Joint 2D and 3D Vehicle Analysis from Monocular Image. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1827-1836.

Chabot, F., Pham, Q.-C. & Chaouch, M. (2020). LapNet : Automatic Balanced Loss and Optimal Assignment for Real-Time Dense Object Detection. *arXiv :1911.01149*.

Chafik, S., Orcesi, A., Audigier, R. & Luvison, B. (2020). Classifying All Interacting Pairs in a Single Shot. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2881-2890.

Chang, X., Hospedales, T. M. & Xiang, T. (2018). Multi-level Factorisation Net for Person Re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chan-Lang, S. (2017). *Closed and Open World Multi-shot Person Re-identification* (These de doctorat). Paris 6.

Chan-Lang, S., Pham, Q.-C. & Achard, C. (2016). Bidirectional sparse representations for multi-shot person re-identification. *13th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2016, Colorado Springs, CO, USA, August 23-26, 2016*, 263-270.

Chan-Lang, S., Pham, Q.-C. & Achard, C. (2017). Closed and Open-World Person Re-Identification and Verification. *2017 International Conference on Digital Image Computing : Techniques and Applications, DICTA 2017, Sydney, Australia, November 29 - December 1, 2017*, 1-8.

Chen, C.-H., Tyagi, A., Agrawal, A., Drover, D., Mv, R., Stojanov, S. & Rehg, J. M. (2019a). Unsupervised 3D Pose Estimation With Geometric Self-Supervision. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5707-5717.

Chen, D., Xu, D., Li, H., Sebe, N. & Wang, X. (2018a). Group Consistent Similarity Learning via Deep CRF for Person Re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8649-8658.

Chen, G., Lin, C., Ren, L., Lu, J. & Zhou, J. (2019b). Self-Critical Attention Learning for Person Re-Identification. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9636-9645.

Chen, G., Choi, W., Yu, X., Han, T. & Chandraker, M. (2017). Learning efficient object detection models with knowledge distillation. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 742-751.

Chen, M.-Y. & Hauptmann, A. (2009). *MoSIFT : recognizing human actions in surveillance videos* (T. CMU-CS-09-161). Technical report, Carnegie Mellon University.

Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. In H. D. III & A. Singh (Éd.), *Proceedings of the 37th International Conference on Machine Learning* (p. 1597-1607). PMLR.

Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G. & Sun, J. (2018b). Cascaded pyramid network for multi-person pose estimation. *CVPR*.

Chen Change Loy, Xiang, T. & Gong, S. (2009). Multi-camera activity correlation analysis. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 1988-1995.

Cheng, D. S., Bazzani, L., Cristani, M., Stoppa, M. & Murino, V. (2011a). Custom Pictorial Structures for Re-identification. *Proceedings of the British Machine Vision Conference*, 68.1-68.11.

Cheng, M., Cai, K. & Li, M. (2020). RWF-2000 : An Open Large Scale Video Database for Violence Detection. *arXiv :1911.05913*.

Cho, S.-H. & Kang, H.-B. (2014). Abnormal behavior detection using hybrid agents in crowded scenes. *Pattern Recognition Letters*, *44*, 64-70.

Choi, W. & Savarese, S. (2012). A Unified Framework for Multi-Target Tracking and Collective Activity Recognition [event-place : Florence, Italy]. *Proceedings of the 12th European Conference on Computer Vision - Volume Part IV*, 215-230.

Choudhary, T., Mishra, V., Goswami, A. & Sarangapani, J. (2020). A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, *53*(7), 5113-5155.

Cong, Y., Yuan, J. & Liu, J. (2011). Sparse reconstruction cost for abnormal event detection. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 3449-3456.

Cong, Y., Yuan, J. & Liu, J. (2013). Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, *46*(7), 1851-1864.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. & Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene

Understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213-3223.

Crammer, K., Singer, Y., Cristianini, N., Shawe-taylor, J. & Williamson, B. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research (JMLR)*.

Cui, X., Liu, Q., Gao, M. & Metaxas, D. N. (2011). Abnormal detection using interaction energy potentials. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 3161-3167.

Dabral, R., Gundavarapu, N. B., Mitra, R., Sharma, A., Ramakrishnan, G. & Jain, A. (2019). Multi-person 3d human pose estimation from monocular images. *2019 International Conference on 3D Vision (3DV)*, 405-414.

Dabral, R., Mundhada, A., Kusupati, U., Afaque, S., Sharma, A. & Jain, A. (2018). Learning 3d human pose from structure and motion. *Proceedings of the European Conference on Computer Vision (ECCV)*, 668-683.

Dai, J., Li, Y., He, K. & Sun, J. (2016). R-FCN : Object Detection via Region-based Fully Convolutional Networks. *NIPS*.

Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *1*, 886-893 vol. 1.

De Souza, F., Chávez, G., Do Valle, E. & De A Araujo, A. (2010). Violence Detection in Video Using Spatio-Temporal Features. *Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on*, 224-230.

Dean, T., Ruzon, M. A., Segal, M., Shlens, J., Vijayanarasimhan, S. & Yagnik, J. (2013). Fast Accurate Detection of 100000 Object Classes on a Single Machine. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K. & Leal-Taixé, L. (2020). MOT20 : A benchmark for multi object tracking in crowded scenes. *arXiv :2003.09003*.

Deng, J., Satheesh, S., Berg, A. & Fei-Fei, L. (2011). Fast and Balanced : Efficient Label Tree Learning for Large Scale Object Recognition. *Conference on Neural Information Processing Systems (NIPS)*.

Deng, W., Zheng, L., Kang, G., Yang, Y., Ye, Q. & Jiao, J. (2018). Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Déniz, O., Serrano, I., Bueno, G. & Kim, T.-K. (2014). Fast Violence Detection in Video. *VISAPP*, *2*, 478-485.

Dhiman, C. & Vishwakarma, D. K. (2019). A review of state-of-the-art techniques for abnormal human activity recognition. *Engineering Applications of Artificial Intelligence*, *77*, 21-45.

Dikmen, M., Akbas, E., Huang, T. S. & Ahuja, N. (2010). Pedestrian recognition with a learned metric. *Asian conference on Computer vision*.

Ding, S., Lin, L., Wang, G. & Chao, H. (2015). Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*.

Dollar, P., Belongie, S. & Perona, P. (2010). The Fastest Pedestrian Detector in the West. *British Machine Vision Conference (BMVC)*, 68.1-68.11.

Dollár, P., Appel, R., Belongie, S. & Perona, P. (2014). Fast Feature Pyramids for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Dubourvieux, F., Audigier, R., Loesch, A., Ainouz, S. & Canu, S. (2020). Unsupervised Domain Adaptation for Person Re-Identification through Source-Guided Pseudo-Labeling. *arXiv :2009.09445*.

Dufour, J.-Y. (Éd.). (2012). *Intelligent Video Surveillance Systems* (1st edition). Wiley-ISTE.

Dupont, C., Tobías, L. & Luvison, B. (2017). Crowd-11 : A Dataset for Fine Grained Crowd Behaviour Analysis. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2184-2191.

Ess, A., Leibe, B. & Gool, L. V. (2007). Depth and Appearance for Mobile Scene Analysis. *2007 IEEE 11th International Conference on Computer Vision*, 1-8.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. & Zisserman, A. (2007). *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.*

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. & Zisserman, A. (2010b). *The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results.*

Fabbri, M., Lanzi, F., Calderara, S., Alletto, S. & Cucchiara, R. (2020). Compressed Volumetric Heatmaps for Multi-Person 3D Pose Estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7204-7213.

Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R. & Cucchiara, R. (2018). Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World.

Fagot-Bouquet, L., Audigier, R., Dhome, Y. & Lerasle, F. (2016). Improving Multi-frame Data Association with Sparse Representations for Robust Near-online Multi-object Tracking. In B. Leibe, J. Matas, N. Sebe & M. Welling (Éd.), *Computer Vision – ECCV 2016* (p. 774-790). Springer International Publishing.

Fang, H.-S., Xu, Y., Wang, W., Liu, X. & Zhu, S.-C. (2018). Learning Pose Grammar to Encode Human Body Configuration for 3D Pose Estimation. *AAAI Conference on Artificial Intelligence.*

Fang, J., Yan, D., Qiao, J. & Xue, J. (2019). DADA : A Large-scale Benchmark and Model for Driver Attention Prediction in Accidental Scenarios. *arXiv :1912.12148.*

Farenzena, M., Bazzani, L., Perina, A., Murino, V. & Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. *The IEEE Conference on Computer Vision and Pattern Recognition.*

Farnebäck, G. (2003). Two-Frame Motion Estimation Based on Polynomial Expansion. In J. Bigun & T. Gustavsson (Éd.), *Image Analysis* (p. 363-370). Springer Berlin Heidelberg.

Feichtenhofer, C., Pinz, A. & Zisserman, A. (2017). Detect to Track and Track to Detect. *2017 IEEE International Conference on Computer Vision (ICCV)*, 3057-3065.

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

Felzenszwalb, P., McAllester, D. & Ramanan, D. (2008b). A Discriminatively Trained, Multiscale, Deformable Part Model. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Felzenszwalb, P. F., Girshick, R. B. & McAllester, D. A. (2010b). Cascade object detection with deformable part models. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Felzenszwalb, P. F. & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International journal of computer vision*, *59*(2), 167-181.

Fidler, S., Boben, M. & Leonardis, A. (2010). A Coarse-to-Fine Taxonomy of Constellations for Fast Multi-class Object Detection. *European Conference on Computer Vision (ECCV).*

Fradi, H., Luvison, B. & Pham, Q. C. (2017). Crowd Behavior Analysis Using Local Mid-Level Visual Descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, *27*(3), 589-602.

Fradi, H. & Dugelay, J.-L. (2014). Sparse Feature Tracking for Crowd Change Detection and Event Recognition. *22nd International Conference on Pattern Recognition, ICPR*, 4116-4121.

Fradi, H. & Dugelay, J.-L. (2015a). Spatial and temporal variations of feature tracks for crowd behavior analysis. *Journal on Multimodal User Interfaces, Springer*.

Fradi, H. & Dugelay, J.-L. (2015b). Towards crowd density-aware video surveillance applications. *Information Fusion*, *24*, 3-15.

Franchi, G., Aldea, E., Dubuisson, S. & Bloch, I. (2020). Tracking Hundreds of People in Densely Crowded Scenes With Particle Filtering Supervising Deep Convolutional Neural Networks. *2020 IEEE International Conference on Image Processing (ICIP)*, 2071-2075.

Freund, Y. & Schapire, R. E. (1997). A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. *Journal of Computer and System Sciences*.

Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A. & Berg, A. C. (2017). DSSD : Deconvolutional single shot detector. *arXiv :1701.06659*.

Gadeski, E., Fard, H. O. & Borgne, H. L. (2014). GPU deformable part model for object recognition. *Journal of Real-Time Image Processing (JRTIP)*, 1-13.

Gaidon, A., Harchaoui, Z. & Schmid, C. (2014). Activity representation with motion hierarchies. *International Journal of Computer Vision*, *107*(3), 219-238.

Gall, J. & Lempitsky, V. S. (2009). Class-specific Hough forests for object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ge, Y., Chen, D. & Li, H. (2020). Mutual Mean-Teaching : Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification. *arXiv :2001.01526*.

Giannakopoulos, T., Kosmopoulos, D., Aristidou, A. & Theodoridis, S. (2006). Violence Content Classification Using Audio Features. In G. Antoniou, G. Potamias, C. Spyropoulos & D. Plexousakis (Éd.), *Advances in Artificial Intelligence* (p. 502-507). Springer Berlin Heidelberg.

Gidaris, S. & Komodakis, N. (2015). Object Detection via a Multi-region and Semantic Segmentation-Aware CNN Model. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1134-1142.

Girshick, R. (2015). Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*, 1440-1448.

Girshick, R., Donahue, J., Darrell, T. & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580-587.

Gou, M., Karanam, S., Liu, W., Camps, O. & Radke, R. J. (2017). DukeMTMC4ReID : A Large-Scale Multi-camera Person Re-identification Dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1425-1434.

Gray, D., Brennan, S. & Tao, H. (2007). Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. *10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*.

Gray, D. & Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. *Computer Vision–ECCV 2008* (p. 262-275). Springer.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R. & Valko, M. (2020). Bootstrap your own latent : A new approach to self-supervised Learning. *arXiv :2006.07733*.

Guo, K., Ishwar, P. & Konrad, J. (2013). Action Recognition From Video Using Feature Covariance Matrices. *Image Processing, IEEE Transactions on*, *22*(6), 2479-2494.

Hadid, A., Pietikainen, M. & Ahonen, T. (2004). A discriminative feature space for detecting and recognizing faces. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, *2*, II-II.

Hafiz, A. M. & Bhat, G. M. (2020). A survey on instance segmentation : state of the art. *International Journal of Multimedia Information Retrieval*, *9*(3), 171-189.

Haritaoglu, I., Harwood, D. & Davis, L. S. (1998). W4S : A real-time system for detecting and tracking people in 2 1/2D. In H. Burkhardt & B. Neumann (Éd.), *Computer Vision — ECCV'98* (p. 877-892). Springer Berlin Heidelberg.

Hassner, T., Itcher, Y. & Kliper-Gross, O. (2012). Violent flows : Real-time detection of violent crowd behavior. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, 1-6.

He, K., Gkioxari, G., Dollár, P. & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*, 2961-2969.

He, K., Zhang, X., Ren, S. & Sun, J. (2014). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In D. Fleet, T. Pajdla, B. Schiele & T. Tuytelaars (Éd.), *Computer Vision – ECCV 2014* (p. 346-361). Springer International Publishing.

He, L., Liang, J., Li, H. & Sun, Z. (2018). Deep Spatial Feature Reconstruction for Partial Person Re-identification : Alignment-free Approach. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7073-7082.

Helbing, D. & Molnár, P. (1995). Social force model for pedestrian dynamics. *Physical Review E*, 4282-4286.

Hermans, A., Beyer, L. & Leibe, B. (2017). In Defense of the Triplet Loss for Person Re-Identification. *arXiv :1703.07737*.

Hirzer, M., Beleznai, C., Roth, P. M. & Bischof, H. (2011). Person Re-identification by Descriptive and Discriminative Classification. In A. Heyden & F. Kahl (Éd.), *Image Analysis* (p. 91-102). Springer.

Hirzer, M., Roth, P. M. & Bischof, H. (2012a). Person re-identification by efficient impostor-based metric learning. *The IEEE International Conference on Advanced Video and Signal-Based Surveillance*.

Huo, J., Gao, Y., Yang, W. & Yin, H. (2012). Abnormal Event Detection via Multi-Instance Dictionary Learning. In H. Yin, J. A. F. Costa & G. Barreto (Éd.), *Intelligent Data Engineering and Automated Learning - IDEAL 2012* (p. 76-83). Springer Berlin Heidelberg.

Ionescu, C., Papava, D., Olaru, V. & Sminchisescu, C. (2014). Human3.6M : Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1325-1339.

Jeong, J., Lee, S., Kim, J. & Kwak, N. (2019). Consistency-based Semi-supervised Learning for Object detection. *NeurIPS*.

Ji, S., Xu, W., Yang, M. & Yu, K. (2013). 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221-231.

Joachims, T., Finley, T. & Yu, C.-N. (2009). Cutting-Plane Training of Structural SVMs. *Machine Learning*.

Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T., Nabbe, B., Matthews, I. et al. (2019). Panoptic studio : A massively multiview system for social interaction capture. *IEEE transactions on pattern analysis and machine intelligence*.

Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T. S., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S. & Sheikh, Y. (2017). Panoptic Studio :

A Massively Multiview System for Social Interaction Capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jose, C. & Fleuret, F. (2016). Scalable metric learning via weighted approximate rank component analysis. *European Conference on Computer Vision*, 875-890.

Kaltsa, V., Briassouli, A., Kompatsiaris, I., Hadjileontiadis, L. J. & Strintzis, M. G. (2015). Swarm Intelligence for Detecting Interesting Events in Crowded Environments. *IEEE Transactions on Image Processing*, *24*(7), 2153-2166.

Kaltsa, V., Briassouli, A., Kompatsiaris, I. & Strintzis, M. G. (2012). Timely, robust crowd event characterization. *IEEE International Conference on Image Processing, ICIP*, 2697-2700.

Karaman, S. & Bagdanov, A. D. (2012). Identity inference : generalizing person re-identification scenarios. *Computer Vision–ECCV 2012. Workshops and demonstrations*, 443-452.

karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O. & Radke, R. J. (2019). A Systematic Evaluation and Benchmark for Person Re-Identification : Features, Metrics, and Datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(3), 523-536.

Karanam, S., Li, Y. & Radke, R. (2015a). Sparse re-id : Block sparsity for person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 33-40.

Karanam, S., Li, Y. & Radke, R. J. (2015b). Person Re-Identification With Discriminatively Trained Viewpoint Invariant Dictionaries. *The IEEE International Conference on Computer Vision (ICCV)*.

Karanam, S., Li, Y. & Radke, R. J. (2017). Person re-identification with block sparse recovery. *Image and Vision Computing*, *60*, 75-90.

Kawanishi, Y., Wu, Y., Mukunoki, M. & Minoh, M. (2014). Shinpuhkan2014 : A multi-camera pedestrian dataset for tracking people across multiple cameras [Issue : 7]. *20th Korea-Japan joint workshop on frontiers of computer vision*, *5*.

Kendall, A., Gal, Y. & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *CVPR*.

Khan, K., Albattah, W., Khan, R. U., Qamar, A. M. & Nayab, D. (2020). Advances and Trends in Real Time Visual Crowd Analysis. *Sensors (Basel, Switzerland)*, *20*(18), 5073.

Khan, S. D. & Ullah, H. (2019). A survey of advances in vision-based vehicle re-identification. *Computer Vision and Image Understanding*, *182*, 50-63.

Kim, J. & Grauman, K. (2009). Observe locally, infer globally : A space-time MRF for detecting abnormal activities with incremental updates. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2921-2928.

Kirillov, A., He, K., Girshick, R., Rother, C. & Dollar, P. (2019). Panoptic Segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9396-9405.

Kocabas, M., Athanasiou, N. & Black, M. J. (2020). VIBE : Video Inference for Human Body Pose and Shape Estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kodirov, E., Xiang, T. & Gong, S. (2015). Dictionary Learning with Iterative Laplacian Regularisation for Unsupervised Person Re-identification. In M. W. J. Xianghua Xie & G. K. L. Tam (Éd.), *Proceedings of the British Machine Vision Conference (BMVC)* (p. 44.1-44.12). BMVA Press.

Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M. & Bischof, H. (2012). Large scale metric learning from equivalence constraints. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2288-2295.

Kratz, L. & Nishino, K. (2009). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1446-1453.

Krausz, B. & Bauckhage, C. (2012). Loveparade 2010 : Automatic Video Analysis of a Crowd Disaster. *Computer Vision and Image Understanding*, *116*(3), 307-319.

Kreiss, S., Bertoni, L. & Alahi, A. (2019). Pifpaf : Composite fields for human pose estimation. *CVPR*.

Kressel, U. H.-G. (1999). Pairwise Classification and Support Vector Machines. In B. Schölkopf, C. J. C. Burges & A. J. Smola (Éd.), *Advances in Kernel Methods* (p. 255-268). MIT Press.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. & Serre, T. (2011). HMDB : A large video database for human motion recognition. *2011 International Conference on Computer Vision*, 2556-2563.

Lampert, C. (2010). An efficient divide-and-conquer cascade for nonlinear object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1022-1029.

Lampert, C., Blaschko, M. & Hofmann, T. (2009a). Efficient Subwindow Search : A Branch and Bound Framework for Object Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *31*(12), 2129-2142.

Laptev, I., Marszalek, M., Schmid, C. & Rozenfeld, B. (2008). Learning realistic human actions from movies. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1-8.

Law, H. & Deng, J. (2019). CornerNet : Detecting Objects as Paired Keypoints. *International Journal of Computer Vision*.

Layne, R., Hospedales, T. & Gong, S. (2012a). Person Re-identification by Attributes. *Proceedings of the British Machine Vision Conference*, 24.1-24.11.

Layne, R., Hospedales, T. & Gong, S. (2014a). Re-id : Hunting Attributes in the Wild. *Proceedings of the British Machine Vision Conference*.

Lefter, I., Rothkrantz, L. & Burghouts, G. (2013). A comparative study on automatic audio–visual fusion for aggression detection using meta-information. *Pattern Recognition Letters*, *34*(15), 1953-1963.

Leibe, B., Leonardis, A. & Schiele, B. (2008). Robust Object Detection with Interleaved Categorization and Segmentation. *International Journal of Computer Vision (IJCV)*, *77*(1-3), 259-289.

Leibe, B. & Schiele, B. (2003). Interleaved Object Categorization and Segmentation. *British Machine Vision Conference (BMVC)*, 78.1-78.10.

Leng, Q., Ye, M. & Tian, Q. (2020). A Survey of Open-World Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(4), 1092-1108.

Li, C. & Lee, G. H. (2019). Generating multiple hypotheses for 3d human pose estimation with mixture density network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9887-9895.

Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H. & Lu, C. (2019a). CrowdPose : Efficient Crowded Scenes Pose Estimation and a New Benchmark. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10855-10864.

Li, S. & Chan, A. B. (2014). 3d human pose estimation from monocular images with deep convolutional neural network. *Asian Conference on Computer Vision*, 332-347.

Li, T., Chang, H., Wang, M., Ni, B., Hong, R. & Yan, S. (2015c). Crowded Scene Analysis : A Survey. *IEEE Transactions on Circuits and Systems for Video Technology*, *25*, 367-386.

Li, W., Zhao, R., Xiao, T. & Wang, X. (2014a). DeepReID : Deep Filter Pairing Neural Network for Person Re-identification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 152-159.

Li, W., Zhu, X. & Gong, S. (2018a). Harmonious Attention Network for Person Re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2285-2294.

Li, W., Wu, Y., Mukunoki, M. & Minoh, M. (2012a). Common-near-neighbor analysis for person re-identification. *Image Processing (ICIP), 2012 19th IEEE International Conference on*, 1621-1624.

Li, W., Mahadevan, V. & Vasconcelos, N. (2014b). Anomaly Detection and Localization in Crowded Scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *36*(1), 18-32.

Li, X., Wu, A. & Zheng, W. (2018b). Adversarial Open-World Person Re-Identification. *ECCV*.

Li, Y.-J., Lin, C.-S., Lin, Y.-B. & Wang, Y.-C. F. (2019b). Cross-Dataset Person Re-Identification via Unsupervised Pose Disentanglement and Adaptation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7918-7928.

Liao, S., Hu, Y., Zhu, X. & Li, S. Z. (2015). Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liao, S., Mo, Z., Zhu, J., Hu, Y. & Li, S. Z. (2014). Open-set person re-identification. *arXiv :1408.0872*.

Lin, T., Goyal, P., Girshick, R., He, K. & Dollár, P. (2020). Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 318-327.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. & Belongie, S. (2017b). Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117-2125.

Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. (2017c). Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, 2980-2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014). Microsoft COCO : Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele & T. Tuytelaars (Éd.), *Computer Vision – ECCV 2014* (p. 740-755). Springer International Publishing.

Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C. & Yang, Y. (2019). Improving person re-identification by attribute and identity learning. *Pattern Recognition*, *95*, 151-161.

Lisanti, G., Masi, I., Bagdanov, A. D. & Del Bimbo, A. (2015). Person re-identification by iterative re-weighted sparse ranking. *IEEE transactions on pattern analysis and machine intelligence*, *37*(8), 1629-1642.

Liu, C., Change Loy, C., Gong, S. & Wang, G. (2013). POP : Person Re-identification Post-rank Optimisation. *The IEEE International Conference on Computer Vision (ICCV)*.

Liu, C., Gong, S., Loy, C. C. & Lin, X. (2012a). Person re-identification : what features are important ? *European Conference on Computer Vision Workshops and Demonstrations*.

Liu, J., Luo, J. & Shah, M. (2009). Recognizing realistic actions from videos "in the wild". *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1996-2003.

Liu, S., Qi, L., Qin, H., Shi, J. & Jia, J. (2018). Path Aggregation Network for Instance Segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8759-8768.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. & Berg, A. C. (2016b). Ssd : Single shot multibox detector. *European conference on computer vision*, 21-37.

Liu, X., Wang, H., Wang, J. & Ma, X. (2017). Person re-identification by multiple instance metric learning with impostor rejection. *Pattern Recognition*, *67*, 287-298.

Loesch, A., Rabarisoa, J. & Audigier, R. (2019). End-To-End Person Search Sequentially Trained On Aggregated Dataset. *2019 IEEE International Conference on Image Processing (ICIP)*, 4574-4578.

Lorre, G., Rabarisoa, J., Orcesi, A., Ainouz, S. & Canu, S. (2020). Temporal Contrastive Pretraining for Video Action Recognition. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 651-659.

Luvison, B., Chateau, T., Sayd, P., Pham, Q. C. & Lapresté, J. (2011). Automatic detection of unexpected events in dense areas for videosurveillance applications. *Video Surveillance*. IntechOpen.

Luvizon, D., Picard, D. & Tabia, H. (2020). Multi-task Deep Learning for Real-Time 3D Human Pose Estimation and Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-1.

Ma, X., Zhu, X., Gong, S., Xie, X., Hu, J., Lam, K.-M. & Zhong, Y. (2017). Person re-identification by unsupervised video matching. *Pattern Recognition*, *65*, 197-210.

Ma, Y. & Cisar, P. (2009). Event detection using local binary pattern based dynamic textures. *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, 38-44.

Marín-Jiménez, M. J., Yeguas, E. & Pérez De La Blanca, N. (2013). Exploring STIP-based Models for Recognizing Human Interactions in TV Videos. *Pattern Recogn. Lett.*, *34*(15), 1819-1828.

Marszalek, M., Laptev, I. & Schmid, C. (2009). Actions in context. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2929-2936.

Martinez, J., Hossain, R., Romero, J. & Little, J. J. (2017). A simple yet effective baseline for 3d human pose estimation. *Proceedings of the IEEE International Conference on Computer Vision*, 2640-2649.

McLaughlin, N., Rincón, J. & Miller, P. (2016). Recurrent Convolutional Network for Video-Based Person Re-identification. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mehran, R., Oyama, A. & Shah, M. (2009). Abnormal crowd behavior detection using social force model. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 935-942.

Mehran, R., Moore, B. E. & Shah, M. (2010). A Streakline Representation of Flow in Crowded Scenes. *European Conference on Computer Vision, ECCV*, *6313*, 439-452.

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W. & Theobalt, C. (2017a). Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. *3D Vision (3DV), 2017 Fifth International Conference on*.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.-P., Rhodin, H., Pons-Moll, G. & Theobalt, C. (2019). Xnect : Real-time multi-person 3d human pose estimation with a single rgb camera. *arXiv :1907.00837*.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G. & Theobalt, C. (2017b). Single-shot multi-person 3d body pose estimation from monocular rgb input. *arXiv :1712.03453*.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G. & Theobalt, C. (2018). Single-shot multi-person 3d pose estimation from monocular rgb. *2018 International Conference on 3D Vision (3DV)*, 120-130.

Mignon, A. & Jurie, F. (2012). Pcca : A new approach for distance learning from sparse pairwise constraints. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2666-2672.

Moon, G., Chang, J. Y. & Lee, K. M. (2019a). Camera Distance-aware Top-down Approach for 3D Multi-person Pose Estimation from a Single RGB Image. *arXiv :1907.11346*.

Moreno-Noguer, F. (2017). 3d human pose estimation from a single image via distance matrix regression. *CVPR*.

Mousavi, H., Mohammadi, S., Perina, A., Chellali, R. & Murino, V. (2015). Analyzing Tracklets for the Detection of Abnormal Crowd Behavior. *IEEE Winter Conference on Applications of Computer Vision*, 148-155.

Murgia, M. & Harlow, M. (2019). Who's using your face ? The ugly truth about facial recognition.

Naphade, M., Anastasiu, D. C., Sharma, A., Jagrlamudi, V., Jeon, H., Liu, K., Chang, M., Lyu, S. & Gao, Z. (2017). The NVIDIA AI City Challenge. *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation*, 1-6.

Nazir, S., Yousaf, M. H. & Velastin, S. A. (2018). Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition. *Computers & Electrical Engineering*, *72*, 660-669.

Newell, A., Huang, Z. & Deng, J. (2017). Associative Embedding : End-to-End Learning for Joint Detection and Grouping. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Éd.), *Advances in Neural Information Processing Systems 30* (p. 2277-2287). Curran Associates, Inc.

Newell, A., Yang, K. & Deng, J. (2016). Stacked hourglass networks for human pose estimation. *European conference on computer vision*, 483-499.

Ng, Joe Yue-Hei, Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R. & Toderici, G. (2015). Beyond short snippets : Deep networks for video classification. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4694-4702.

Nie, B. X., Wei, P. & Zhu, S.-C. (2017). Monocular 3d human pose estimation by predicting depth on joints. *2017 IEEE International Conference on Computer Vision (ICCV)*, 3467-3475.

Niebles, J. C., Chen, C.-W. & Fei-Fei, L. (2010). Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In K. Daniilidis, P. Maragos & N. Paragios (Éd.), *Computer Vision – ECCV 2010* (p. 392-405). Springer Berlin Heidelberg.

Odabai Fard, H., Chaouch, M., Pham, Q.-C., Vacavant, A. & Chateau, T. (2014b). Joint Hierarchical Learning for Efficient Multi-class Object Detection. *IEEE Winter Conference on applications of computer vision (WACV)*.

Odabai Fard, H., Chaouch, M., Pham, Q.-C., Vacavant, A. & Chateau, T. (2014c). Joint Learning for Multi-class Object Detection. *International conference on computer vision theory and applications (VISAPP)*.

Odabai Fard, S. H. (2015). *Efficient multi-class objet detection with a hierarchy of classes* (These de doctorat). Clermont-Ferrand 2.

Oneata, D., Verbeek, J. & Schmid, C. (2013). Action and Event Recognition with Fisher Vectors on a Compact Feature Set. *Computer Vision (ICCV), 2013 IEEE International Conference on*, 1817-1824.

Ott, P. & Everingham, M. (2011). Shared parts for deformable part-based models. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Paisitkriangkrai, S., Shen, C. & van den Hengel, A. (2015). Learning to Rank in Person Re-Identification With Metric Ensembles. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C. & Murphy, K. (2017). Towards accurate multi-person pose estimation in the wild. *CVPR*.

Pareek, P. & Thakkar, A. (2021). A survey on video-based Human Action Recognition : recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, *54*(3), 2259-2322.

Pavlakos, G., Zhou, X. & Daniilidis, K. (2018). Ordinal depth supervision for 3d human pose estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7307-7316.

Pavlakos, G., Zhou, X., Derpanis, K. G. & Daniilidis, K. (2017). Coarse-to-fine volumetric prediction for single-image 3D human pose. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7025-7034.

Pedagadi, S., Orwell, J., Velastin, S. & Boghossian, B. (2013). Local Fisher Discriminant Analysis for Pedestrian Re-identification. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pedersoli, M., Vedaldi, A. & Gonzàlez, J. (2011). A coarse-to-fine approach for fast deformable object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T. & Tian, Y. (2016). Unsupervised Cross-Dataset Transfer Learning for Person Re-Identification. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Platt, J. C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*.

Platt, J. C., Cristianini, N. & Shawe-taylor, J. (2000). Large Margin DAGs for Multiclass Classification. *Conference on Neural Information Processing Systems (NIPS)*.

Popoola, O. & Wang, K. (2012). Video-Based Abnormal Human Behavior Recognition - A Review. *Systems, Man, and Cybernetics, Part C : Applications and Reviews, IEEE Transactions on*, *42*(6), 865-878.

Qiu, Z., Yao, T. & Mei, T. (2017). Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 5534-5542.

Quispe, R. & Pedrini, H. (2019). Improved person re-identification based on saliency and semantic parsing with deep neural network models. *Image and Vision Computing*, *92*, 103809.

Raghavendra, R., Del Bue, A., Cristani, M. & Murino, V. (2011). Abnormal Crowd Behavior Detection by Social Force Optimization. In A. A. Salah & B. Lepri (Éd.), *Human Behavior Understanding* (p. 134-145). Springer Berlin Heidelberg.

Ranga, A., Giruzzi, F., Bhanushali, J., Wirbel, E., Pérez, P., Vu, T.-H. & Perotton, X. (2020). VRUNet : Multi-Task Learning Model for Intent Prediction of Vulnerable Road Users. *Electronic Imaging*, *2020*(16), 109–1–109-10.

Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C. & Sebe, N. (2017). Abnormal event detection in videos using generative adversarial nets. *2017 IEEE International Conference on Image Processing (ICIP)*, 1577-1581.

Razavi, N., Gall, J. & Gool, L. J. V. (2011). Scalable multi-class object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Razavi, N., Gall, J., Kohli, P. & Gool, L. J. V. (2012). Latent Hough Transform for Object Detection. *European Conference on Computer Vision (ECCV)*.

Reddy, K. K. & Shah, M. (2013). Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, *24*(5), 971-981.

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. (2016). You only look once : Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779-788.

Redmon, J. & Farhadi, A. (2017). YOLO9000 : better, faster, stronger. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263-7271.

Redmon, J. & Farhadi, A. (2018). Yolov3 : An incremental improvement. *arXiv :1804.02767*.

Ren, S., He, K., Girshick, R. & Sun, J. (2015). Faster R-CNN : towards real-time object detection with region proposal networks. *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, 91-99.

Ribeiro, P. C., Audigier, R. & Pham, Q. C. (2016). RIMOC, a feature to discriminate unstructured motions : Application to violence detection for video-surveillance. *Computer Vision and Image Understanding*, *144*, 121-143.

Rogez, G., Weinzaepfel, P. & Schmid, C. (2017). Lcr-net : Localization-classification-regression for human pose. *CVPR*.

Rogez, G., Weinzaepfel, P. & Schmid, C. (2019). Lcr-net++ : Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*.

Roshtkhari, M. J. & Levine, M. D. (2013b). An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Computer Vision and Image Understanding*, *117*(10), 1436-1452.

Roshtkhari, M. & Levine, M. (2013c). Online Dominant and Anomalous Behavior Detection in Videos. *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2611-2618.

Rosten, E., Porter, R. & Drummond, T. (2010). Faster and Better : A machine learning approach to corner detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *32*, 105-119.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*.

Sadeghi, M. A. & Forsyth, D. A. (2014). 30Hz Object Detection with DPM V5. *European Conference on Computer Vision (ECCV)*, 65-79.

Salakhutdinov, R., Tenenbaum, J. B. & Torralba, A. (2012). One-Shot Learning with a Hierarchical Nonparametric Bayesian Model. *Unsupervised and Transfer Learning - Workshop held at ICML*, 195-206.

Salakhutdinov, R., Torralba, A. & Tenenbaum, J. B. (2011). Learning to share visual appearance for multiclass object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Saligrama, V. & Chen, Z. (2012). Video anomaly detection based on local statistical aggregates. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2112-2119.

Sanin, A., Sanderson, C., Harandi, M. & Lovell, B. (2013). Spatio-temporal covariance descriptors for action and gesture recognition. *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, 103-110.

Schneider, S., Beery, S. & Parham, J. (2020). WACV2020 AI for Animal Re-ID.

Shao, J., Loy, C. C. & Wang, X. (2014). Scene-Independent Group Profiling in Crowd. *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2227-2234.

Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X. & Sun, J. (2018). CrowdHuman : A Benchmark for Detecting Human in a Crowd. *arXiv :1805.00123*.

Shi, F., Petriu, E. & Laganiere, R. (2013). Sampling Strategies for Real-Time Action Recognition. *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2595-2602.

Shi, J. & Malik, J. (2000). Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Shi, J. & Tomasi, C. (1994). Good features to track. *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 593-600.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D. & Wang, Z. (2016b). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874-1883.

Shi, Y., Gao, Y. & Wang, R. (2010). Real-Time Abnormal Event Detection in Complicated Scenes. *Proceedings of the 2010 20th International Conference on Pattern Recognition*, 3653-3656.

Shi, Z., Hospedales, T. M. & Xiang, T. (2015). Transferring a Semantic Representation for Person Re-Identification and Search. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shin, D. & Tjahjadi, T. (2008). Similarity Invariant Delaunay Graph Matching. *Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshop*, 25-34.

Shri, S. J. & Jothilakshmi, S. (2019). Crowd Video Event Classification using Convolutional Neural Network. *Computer Communications*, *147*, 35-39.

Simonyan, K. & Zisserman, A. (2014a). Two-Stream Convolutional Networks for Action Recognition in Videos [event-place : Montreal, Canada]. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, 568-576.

Sindagi, V., Yasarla, R. & Patel, V. (2019). Pushing the Frontiers of Unconstrained Crowd Counting : New Dataset and Benchmark Method. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1221-1231.

Sindagi, V. A. & Patel, V. M. (2018). A survey of recent advances in CNN-based single image crowd counting and density estimation. *Pattern Recognition Letters*, *107*, 3-16.

Sjöberg, M., Ionescu, B., Jiang, Y.-G., Quang, V. L., Schedl, M. & Demarty, C.-H. (2014). The MediaEval 2014 Affect Task : Violent Scenes Detection. *Working Notes Proceedings of the MediaEval 2014 Workshop*.

Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A. & Zisserman, A. (2020). A Short Note on the Kinetics-700-2020 Human Action Dataset. *arXiv :2010.10864*.

Soliman, M. M., Kamal, M. H., Nashed, M. A. E.-M., Mostafa, Y. M., Chawky, B. S. & Khattab, D. (2019). Violence Recognition from Videos using Deep Learning Techniques. *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, 80-85.

Soomro, K., Zamir, A. R. & Shah, M. (2012). UCF101 : A dataset of 101 human actions classes from videos in the wild. *CoRR*.

Su, C., Li, J., Zhang, S., Xing, J., Gao, W. & Tian, Q. (2017). Pose-Driven Deep Convolutional Model for Person Re-identification. *2017 IEEE International Conference on Computer Vision (ICCV)*, 3980-3989.

Subramaniam, A., Nambiar, A. & Mittal, A. (2019). Co-Segmentation Inspired Attention Networks for Video-Based Person Re-Identification. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Suh, Y., Wang, J., Tang, S., Mei, T. & Lee, K. M. (2018). Part-Aligned Bilinear Representations for Person Re-identification. *European Conference on Computer Vision (ECCV)*, *11218*, 418-437.

Sultani, W., Chen, C. & Shah, M. (2018). Real-World Anomaly Detection in Surveillance Videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6479-6488.

Sun, X., Xiao, B., Wei, F., Liang, S. & Wei, Y. (2018). Integral human pose regression. *Proceedings of the European Conference on Computer Vision (ECCV)*, 529-545.

Sundararaman, R., Braga, C. D. A., Marchand, E. & Pettre, J. (2021). Tracking Pedestrian Heads in Dense Crowd. *arXiv :2103.13516*.

Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. *Thirty-first AAAI conference on artificial intelligence*.

Tan, M., Pang, R. & Le, Q. V. (2020). EfficientDet : Scalable and Efficient Object Detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V. & Fua, P. (2016). Structured prediction of 3d human pose with deep neural networks. *Proceedings of the British Machine Vision Conference (BMVC)*.

Thanasutives, P., Fukui, K., Numao, M. & Kijsirikul, B. (2020). Encoder-Decoder Based Convolutional Neural Networks with Multi-Scale-Aware Modules for Crowd Counting. *ArXiv, abs/2003.05586*.

Tian, Z., Shen, C., Chen, H. & He, T. (2019). Fcos : Fully convolutional one-stage object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 9627-9636.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y. & Bregler, C. (2015). Efficient object localization using convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 648-656.

Torralba, A., Murphy, K. P. & Freeman, W. T. (2004). Sharing features : efficient boosting procedures for multiclass object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Torralba, A., Murphy, K. P. & Freeman, W. T. (2007). Sharing Visual Features for Multiclass and Multiview Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Toshev, A., Szegedy, C. & DeepPose, G. (2014). Human pose estimation via deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA*, 24-27.

Uijlings, J. R. R., Sande, K. E. A. v. d., Gevers, T. & Smeulders, A. W. M. (2013). Selective Search for Object Recognition. *International Journal of Computer Vision (IJCV)*.

Vaquette, G., Achard, C. & Lucat, L. (2019). Robust information fusion in the DOHT paradigm for real-time action detection. *Journal of Real-Time Image Processing*, *16*(5), 1511-1524.

Varadarajan, J. & Odobez, J. (2009). Topic models for scene analysis and abnormality detection. *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, 1338-1345.

Varior, R. R., Shuai, B., Lu, J., Xu, D. & Wang, G. (2016). A siamese long short-term memory architecture for human re-identification. *European Conference on Computer Vision*, 135-153.

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

Viola, P. A. & Jones, M. J. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, B., Ye, M., Li, X., Zhao, F. & Ding, J. (2012a). Abnormal crowd behavior detection using high-frequency and spatio-temporal features. *Machine Vision and Applications*, *23*(3), 501-511.

Wang, F., Zuo, W., Lin, L., Zhang, D. & Zhang, L. (2016a). Joint Learning of Single-Image and Cross-Image Representations for Person Re-Identification. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, H., Gong, S., Zhu, X. & Xiang, T. (2016c). Human-in-the-loop person re-identification. *European Conference on Computer Vision*, 405-422.

Wang, H., Zhu, X., Xiang, T. & Gong, S. (2016d). Towards unsupervised open-set person re-identification. *Image Processing (ICIP), 2016 IEEE International Conference on*, 769-773.

Wang, H., Kläser, A., Schmid, C. & Liu, C.-L. (2013). Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *International Journal of Computer Vision*, *103*(1), 60-79.

Wang, H. & Schmid, C. (2013). Action Recognition with Improved Trajectories. *Computer Vision (ICCV), 2013 IEEE International Conference on*, 3551-3558.

Wang, J., Zhu, X., Gong, S. & Li, W. (2018a). Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2275-2284.

Wang, L., Qiao, Y. & Tang, X. (2015). Action recognition with trajectory-pooled deep-convolutional descriptors. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4305-4314.

Wang, M., Chen, X., Liu, W., Qian, C., Lin, L. & Ma, L. (2018b). Drpose3d : Depth ranking in 3d human pose estimation. *arXiv :1805.08973*.

Wang, T., Gong, S., Zhu, X. & Wang, S. (2014b). Person Re-identification by Video Ranking. In D. Fleet, T. Pajdla, B. Schiele & T. Tuytelaars (Éd.), *Computer Vision – ECCV 2014* (p. 688-703). Springer International Publishing.

Wang, T., Gong, S., Zhu, X. & Wang, S. (2016e). Person Re-Identification by Discriminative Selection in Video Ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, T., Miao, Z., Chen, Y., Zhou, Y., Shan, G. & Snoussi, H. (2019). AED-Net : An Abnormal Event Detection Network. *Engineering*, *5*(5), 930-939.

Wang, T. & Snoussi, H. (2012). Histograms of Optical Flow Orientation for Visual Abnormal Events Detection. *IEEE Advanced Video and Signal-based Surveillance, AVSS*, 13-18.

Wang, T. & Snoussi, H. (2015). Detection of abnormal events via optical flow feature analysis. *Sensors (Basel, Switzerland)*, *15*(4), 7156-7171.

Wei, S.-E., Ramakrishna, V., Kanade, T. & Sheikh, Y. (2016). Convolutional pose machines. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 4724-4732.

Wei, X., Du, J., Xue, Z., Liang, M., Geng, Y., Xu, X. & Lee, J. (2020). A very deep two-stream network for crowd type recognition. *Neurocomputing*, *396*, 522-533.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S. & Ma, Y. (2009). Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(2), 210-227.

Wu, S., Moore, B. E. & Shah, M. (2010). Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2054-2060.

Yan, J., Lei, Z., Wen, L. & Li, S. Z. (2014). The Fastest Deformable Part Model for Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2497-2504.

Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y. & Yang, X. (2016). Person re-identification via recurrent feature aggregation. *European Conference on Computer Vision*, 701-716.

Yang, M., Rajasegarar, S., Erfani, S. M. & Leckie, C. (2019a). Deep Learning and One-class SVM based Anomalous Crowd Detection. *2019 International Joint Conference on Neural Networks (IJCNN)*, 1-8.

Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H. & Wang, X. (2018). 3d human pose estimation in the wild by adversarial learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5255-5264.

Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D. & Li, S. Z. (2014). Salient Color Names for Person Re-identification. *European Conference on Computer Vision*.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L. & Hoi, S. C. H. (2021). Deep Learning for Person Re-identification : A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-1.

Ye, M., Liang, C., Yu, Y., Wang, Z., Leng, Q., Xiao, C., Chen, J. & Hu, R. (2016). Person Reidentification via Ranking Aggregation of Similarity Pulling and Dissimilarity Pushing. *IEEE Transactions on Multimedia*, *18*(12), 2553-2566.

Yeffet, L. & Wolf, L. (2009). Local Trinary Patterns for human action recognition. *Computer Vision, 2009 IEEE 12th International Conference on*, 492-497.

Yi, D., Lei, Z., Liao, S. & Li, S. Z. (2014b). Deep metric learning for person re-identification. *Pattern Recognition (ICPR), 2014 22nd International Conference on*, 34-39.

You, J., Wu, A., Li, X. & Zheng, W.-S. (2016). Top-Push Video-Based Person Re-Identification. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zaharescu, A. & Wildes, R. (2010). Anomalous Behaviour Detection Using Spatiotemporal Oriented Energies, Subset Inclusion Histogram Comparison and Event-Driven Processing. In K. Daniilidis, P. Maragos & N. Paragios (Éd.), *Computer Vision – ECCV 2010* (p. 563-576). Springer Berlin Heidelberg.

Zanfir, A., Marinoiu, E. & Sminchisescu, C. (2018a). Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes–The Importance of Multiple Scene Constraints. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zanfir, A., Marinoiu, E., Zanfir, M., Popa, A.-I. & Sminchisescu, C. (2018b). Deep network for the integrated 3d sensing of multiple people in natural images. *Advances in Neural Information Processing Systems*, 8410-8419.

Zhai, Y., Guo, X., Lu, Y. & Li, H. (2019). In Defense of the Classification Loss for Person Re-Identification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Zhang, L., Xiang, T. & Gong, S. (2016b). Learning a Discriminative Null Space for Person Re-Identification. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, S., Li, N., Cheng, X. & Wu, Z. (2013a). Adaptive object detection by implicit sub-class sharing features. *Signal Processing*, *93*(6), 1458-1470.

Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C. & Sun, J. (2018). AlignedReID : Surpassing Human-Level Performance in Person Re-Identification. *arXiv :1711.08184*.

Zhang, Y., Qin, L., Yao, H., Xu, P. & Huang, Q. (2013b). Beyond particle flow : Bag of Trajectory Graphs for dense crowd event recognition. *IEEE International Conference on Image Processing, ICIP*, 3572-3576.

Zhang, Y., Wang, C., Wang, X., Zeng, W. & Liu, W. (2020a). FairMOT : On the Fairness of Detection and Re-Identification in Multiple Object Tracking. *arXiv :2004.01888*.

Zhang, Z., Lan, C., Zeng, W. & Chen, Z. (2019). Densely Semantically Aligned Person Re-Identification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 667-676.

Zhang, Z., Lan, C., Zeng, W., Jin, X. & Chen, Z. (2020b). Relation-Aware Global Attention for Person Re-Identification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3183-3192.

Zhao, B., Fei-Fei, L. & Xing, E. (2011). Online detection of unusual events in videos via dynamic sparse coding. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 3313-3320.

Zhao, R., Ouyang, W. & Wang, X. (2013a). Person Re-identification by Salience Matching. *The IEEE International Conference on Computer Vision (ICCV)*.

Zhao, R., Ouyang, W. & Wang, X. (2013b). Unsupervised Salience Learning for Person Re-identification. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, R., Ouyang, W. & Wang, X. (2014). Learning Mid-level Filters for Person Re-identification. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, Z., Zheng, P., Xu, S. & Wu, X. (2019). Object Detection With Deep Learning : A Review. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(11), 3212-3232.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. & Tian, Q. (2015a). Scalable Person Re-identification : A Benchmark. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1116-1124.

Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y. & Tian, Q. (2017a). Person Re-identification in the Wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3346-3355.

Zheng, W.-S., Gong, S. & Xiang, T. (2011b). Person re-identification by probabilistic relative distance comparison. *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, 649-656.

Zheng, W.-S., Gong, S. & Xiang, T. (2012). Transfer re-identification : From person to set-based verification. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2650-2657.

Zheng, W.-S., Gong, S. & Xiang, T. (2016b). Towards open-world person re-identification by one-shot group-based verification. *IEEE transactions on pattern analysis and machine intelligence*, *38*(3), 591-606.

Zheng, Z., Zheng, L. & Yang, Y. (2017b). Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. *2017 IEEE International Conference on Computer Vision (ICCV)*.

Zhong, Z., Zheng, L., Cao, D. & Li, S. (2017). Re-ranking Person Re-identification with k-Reciprocal Encoding. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3652-3661.

Zhou, B., Tang, X. & Wang, X. (2012). Coherent Filtering : Detecting Coherent Motions from Crowd Clutters. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato & C. Schmid (Éd.), *Computer Vision – ECCV 2012* (p. 857-871). Springer Berlin Heidelberg.

Zhou, J., Su, B. & Wu, Y. (2018). Easy Identification from Better Constraints : Multi-shot Person Re-identification from Reference Constraints. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5373-5381.

Zhou, X., Huang, Q., Sun, X., Xue, X. & Wei, Y. (2017). Towards 3d human pose estimation in the wild : a weakly-supervised approach. *Proceedings of the IEEE International Conference on Computer Vision*, 398-407.

Zhou, X., Wang, D. & Krähenbühl, P. (2019). Objects as Points. *arXiv :1904.07850*.

Zhu, F., Wang, X. & Yu, N. (2014a). Crowd Tracking with Dynamic Evolution of Group Structures. In D. Fleet, T. Pajdla, B. Schiele & T. Tuytelaars (Éd.), *Computer Vision – ECCV 2014* (p. 139-154). Springer International Publishing.

Zhu, L., Chen, Y., Torralba, A., Freeman, W. T. & Yuille, A. L. (2010a). Part and appearance sharing : Recursive Compositional Models for multi-view. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhu, X., Wu, B., Huang, D. & Zheng, W. (2018). Fast Open-World Person Re-Identification. *IEEE Transactions on Image Processing*, *27*(5), 2286-2300.

Zhu, X., Liu, J., Wang, J., Li, C. & Lu, H. (2014b). Sparse representation for robust abnormality detection in crowded scenes. *Pattern Recognition*, *47*(5), 1791-1799.

Zou, Z., Shi, Z., Guo, Y. & Ye, J. (2019). Object Detection in 20 Years : A Survey. *arXiv :1905.05055*.

Publications et brevets

Publications associées aux travaux présentés

Benzine, A., Chabot, F., Luvison, B., Pham, Q. C. & Achard, C. (2020a). PandaNet : Anchor-Based Single-Shot Multi-Person 3D Pose Estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6856-6865.

Benzine, A., Luvison, B., Pham, Q. C. & Achard, C. (2019). Deep, Robust and Single Shot 3D Multi-Person Human Pose Estimation from Monocular Images. *2019 IEEE International Conference on Image Processing (ICIP)*, 584-588.

Benzine, A., Luvison, B., Pham, Q. C. & Achard, C. (2020b). Single-shot 3D multi-person pose estimation in complex images. *Pattern Recognition*, 107534.

Chan-Lang, S., Pham, Q.-C. & Achard, C. (2016). Bidirectional sparse representations for multi-shot person re-identification. *13th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2016, Colorado Springs, CO, USA, August 23-26, 2016*, 263-270.

Chan-Lang, S., Pham, Q.-C. & Achard, C. (2017). Closed and Open-World Person Re-Identification and Verification. *2017 International Conference on Digital Image Computing : Techniques and Applications, DICTA 2017, Sydney, Australia, November 29 - December 1, 2017*, 1-8.

Fradi, H., Luvison, B. & Pham, Q. C. (2017). Crowd Behavior Analysis Using Local Mid-Level Visual Descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, *27*(3), 589-602.

Odabai Fard, H., Chaouch, M., Pham, Q.-C., Vacavant, A. & Chateau, T. (2014a). Apprentissage hiérarchique simultané pour la détection efficace d'objets. *Reconnaissance de Formes et Intelligence Artificielle (RFIA) 2014*.

Odabai Fard, H., Chaouch, M., Pham, Q.-C., Vacavant, A. & Chateau, T. (2014b). Joint Hierarchical Learning for Efficient Multi-class Object Detection. *IEEE Winter Conference on applications of computer vision (WACV)*.

Odabai Fard, H., Chaouch, M., Pham, Q.-C., Vacavant, A. & Chateau, T. (2014c). Joint Learning for Multi-class Object Detection. *International conference on computer vision theory and applications (VISAPP)*.

Ribeiro, P. C., Audigier, R. & Pham, Q. C. (2016). RIMOC, a feature to discriminate unstructured motions : Application to violence detection for video-surveillance. *Computer Vision and Image Understanding*, *144*, 121-143.

Liste des publications

Badri, J., Tilmant, C., Lavest, J., Pham, Q. & Sayd, P. (2007a). Hybrid sensors calibration : Application to pattern recognition and tracking. *2007 IEEE International Symposium on Intelligent Signal Processing*, 1-5.

Badri, J., Tilmant, C., Lavest, J.-M., Pham, Q.-C. & Sayd, P. (2007b). Camera-to-Camera Mapping for Hybrid Pan-Tilt-Zoom Sensors Calibration. In B. K. Ersbøll & K. S. Pedersen (Éd.), *Image Analysis, 15th Scandinavian Conference, SCIA 2007, Aalborg, Denmark, June 10-14, 2007, Proceedings* (p. 132-141). Springer.

Badri, J., Tilmant, C., Lavest, J.-M., Pham, Q.-C. & Sayd, P. (2007c). Hybrid dynamic sensors calibration from camera-to-camera mapping : An automatic approach. In A. Ranchordas, H. Araújo & J. Vitrià (Éd.), *VISAPP 2007 : Proceedings of the Second International Conference on Computer Vision Theory and Applications, Barcelona, Spain, March 8-11, 2007 - Volume 2* (p. 498-506). INSTICC - Institute for Systems ; Technologies of Information, Control ; Communication.

Badri, J., Tilmant, C., Lavest, J.-M., Sayd, P. & Pham, Q. C. (2008). Automatic Calibration of Hybrid Dynamic Vision System for High Resolution Object Tracking. *Pattern Recognition Techniques, Technology and Applications*. IntechOpen.

Benzine, A., Chabot, F., Luvison, B., Pham, Q. C. & Achard, C. (2020a). PandaNet : Anchor-Based Single-Shot Multi-Person 3D Pose Estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6856-6865.

Benzine, A., Luvison, B., Pham, Q. C. & Achard, C. (2019). Deep, Robust and Single Shot 3D Multi-Person Human Pose Estimation from Monocular Images. *2019 IEEE International Conference on Image Processing (ICIP)*, 584-588.

Benzine, A., Luvison, B., Pham, Q. C. & Achard, C. (2020b). Single-shot 3D multi-person pose estimation in complex images. *Pattern Recognition*, 107534.

Bourgeois, S., Martinsson, H., Pham, Q.-C. & Naudet, S. (2005). A Practical Guide to Marker Based and Hybrid Visual Registration for AR Industrial Applications. In A. Gagalowicz & W. Philips (Éd.), *Computer Analysis of Images and Patterns* (p. 669-676). Springer Berlin Heidelberg.

Carton, F., Filliat, D., Rabarisoa, J. & Pham, Q. C. (2021). Using Semantic Information to Improve Generalization of Reinforcement Learning Policies for Autonomous Driving. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 144-151.

Chabot, F., Pham, Q.-C. & Chaouch, M. (2020). LapNet : Automatic Balanced Loss and Optimal Assignment for Real-Time Dense Object Detection. *arXiv :1911.01149*.

Chague, S., D'Hose, J., Goudou, J.-F., Dorizzi, B., Giulieri, L., Pham, Q.-C., Sèdes, F., Brut, M., Nicholson, D. & Pietquin, O. (2011). METHODEO : Méthodologie d'évaluation des algorithmes d'exploitation des enregistrements de la vidéoprotection. *Workshop Interdisciplinaire sur la Sécurité Globale (WISG 2011)*.

Chan-Lang, S., Pham, Q.-C. & Achard, C. (2016). Bidirectional sparse representations for multi-shot person re-identification. *13th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2016, Colorado Springs, CO, USA, August 23-26, 2016*, 263-270.

Chan-Lang, S., Pham, Q.-C. & Achard, C. (2017). Closed and Open-World Person Re-Identification and Verification. *2017 International Conference on Digital Image Computing : Techniques and Applications, DICTA 2017, Sydney, Australia, November 29 - December 1, 2017*, 1-8.

Dhome, Y., Luvison, B. & Pham, Q. C. (2013). *Method for locating objects by resolution in the three-dimensional space of the scene* (WO/2013/057030).

Didier, J.-Y., Roussel, D., Mallem, M., Otmane, S., Naudet, S., Pham, Q.-C., Bourgeois, S., Mégard, C., Leroux, C. & Hocquard, A. (2005). AMRA : Augmented Reality Assistance for Train Maintenance Tasks. *Workshop Industrial Augmented Reality, 4th ACM/IEEE International Symposium on Mixed and Augmented Reality (ISMAR 2005)*, (Elect. Proc.)

Fagot-Bouquet, L., Rabarisoa, J. & Pham, Q. C. (2014). Fast and accurate video annotation using dense motion hypotheses. *2014 IEEE International Conference on Image Processing (ICIP)*, 3122-3126.

Fradi, H., Luvison, B. & Pham, Q. C. (2017). Crowd Behavior Analysis Using Local Mid-Level Visual Descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, *27*(3), 589-602.

Guillot, C., Pham, Q.-C., Patrick, S., Tilmant, C. & Lavest, J.-M. (2012a). Détection et localisation d'objets stationnaires par une paire de caméras PTZ. *RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle)*, 978–2–9539515–2-3.

Guillot, C., Pham, Q.-C., Sayd, P., Tilmant, C. & Lavest, J.-M. (2012b). Detection and Localisation of Stationary Objects with a Pair of PTZ Cameras. In G. Csurka & J. Braz (Éd.), *VISAPP 2012 - Proceedings of the International Conference on Computer Vision Theory and Applications, Volume 1, Rome, Italy, 24-26 February, 2012* (p. 591-596). SciTePress.

Guillot, C., Pham, Q.-C., Sayd, P., Tilmant, C. & Lavest, J.-M. (2012c). Détection et localisation d'objets stationnaires par une paire de caméras PTZ. *Traitement du Signal*, *29*(3-5), 307-332.

Guillot, C., Taron, M., Sayd, P., Pham, Q.-C., Tilmant, C. & Lavest, J.-M. (2010). Background subtraction adapted to PTZ cameras by keypoint density estimation. In F. Labrosse, R. Zwiggelaar, Y. Liu & B. Tiddeman (Éd.), *British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010. Proceedings* (p. 1-10). British Machine Vision Association.

Luvison, B., Chateau, T., Sayd, P. & Pham, Q. C. (2009a). Méthode d'apprentissage non supervisée pour la détection d'évènements inattendus. *CORESA (COmpression et REprésentation des Signaux Audiovisuels) 2009*, 6.

Luvison, B., Chateau, T., Sayd, P., Pham, Q. C. & Lapresté, J. T. (2010). Estimation Parcimonieuse de Densité par Fonctions Noyaux : Application à la Détection Temps Réel d'Evènements Rares. *Reconnaissance de Formes et Intelligence Artificielle (RFIA) 2010*, 8.

Luvison, B., Chateau, T., Sayd, P., Pham, Q. C. & Lapresté, J. (2011). Automatic detection of unexpected events in dense areas for videosurveillance applications. *Video Surveillance*. IntechOpen.

Luvison, B., Chateau, T., Sayd, P., Pham, Q.-C. & Lapresté, J.-T. (2009b). An Unsupervised Learning based Approach for Unexpected Event Detection. In A. Ranchordas & H. Araújo (Éd.), *VISAPP 2009 - Proceedings of the Fourth International Conference on Computer Vision Theory and Applications, Lisboa, Portugal, February 5-8, 2009 - Volume 1* (p. 506-513). INSTICC Press.

Maggio, S., Haugeard, J.-E., Meden, B., Luvison, B., Audigier, R., Burger, B. & Pham, Q. (2013). Tracking of Objects of Interest in a Sequence of Images. *Intelligent Video Surveillance Systems* (p. 123-146). John Wiley ; Sons.

Maggio, S., Luvison, B. & Pham, Q. C. (2014). *Method for tracking a target in an image sequence, taking the dynamics of the target into consideration* (WO/2014/135404).

Makela, T., Clarysse, P., Sipila, O., Pauna, N., Quoc Cuong Pham, Katila, T. & Magnin, I. E. (2002). A review of cardiac image registration methods. *IEEE Transactions on Medical Imaging*, *21*(9), 1011-1021.

Mäkelä, T., Pham, Q. C., Clarysse, P., Nenonen, J., Lötjönen, J., Sipilä, O., Hänninen, H., Lauerma, K., Knuuti, J., Katila, T. & Magnin, I. E. (2003). A 3-D model-based registration approach for the PET, MR and MCG cardiac data fusion. *Medical Image Analysis*, *7*(3), 377-389.

Mäkelä, T., Pham, Q.-C., Clarysse, P., Lötjönen, J., Lauerma, K., Hänninen, H., Knuuti, J., Katila, T. & Magnin, I. E. (2001). A 3-D Model-Based Approach for the PET-Functional and MR-Anatomical Cardiac Imaging Data Fusion. In T. Katila, J. Nenonen, I. E. Magnin, P. Clarysse & J. Montagnat (Éd.), *Functional Imaging and Modeling of the Heart* (p. 83-90). Springer Berlin Heidelberg.

Odabai Fard, H., Chaouch, M., Pham, Q.-C., Vacavant, A. & Chateau, T. (2014a). Apprentissage hiérarchique simultané pour la détection efficace d'objets. *Reconnaissance de Formes et Intelligence Artificielle (RFIA) 2014*.

Odabai Fard, H., Chaouch, M., Pham, Q.-C., Vacavant, A. & Chateau, T. (2014b). Joint Hierarchical Learning for Efficient Multi-class Object Detection. *IEEE Winter Conference on applications of computer vision (WACV)*.

Odabai Fard, H., Chaouch, M., Pham, Q.-C., Vacavant, A. & Chateau, T. (2014c). Joint Learning for Multi-class Object Detection. *International conference on computer vision theory and applications (VISAPP)*.

Oumsis, M., Pham, Q.-C., Sdigui, A. D., Neyran, B. & Magnin, I. E. (2003). Modeling and Tracking of the Cardiac Left Ventricular Motion by a State Space Harmonic Model in MRI Sequence. In I. E. Magnin, J. Montagnat, P. Clarysse, J. Nenonen & T. Katila (Éd.), *Functional Imaging and Modeling of the Heart* (p. 184-193). Springer Berlin Heidelberg.

Paillet, P., Audigier, R., Lerasle, F. & Pham, Q.-C. (2013). IMM-Based Tracking and Latency Control with Off-the-Shelf IP PTZ Camera. In J. Blanc-Talon, A. J. Kasinski, W. Philips, D. C. Popescu & P. Scheunders (Éd.), *Advanced Concepts for Intelligent Vision Systems - 15th International Conference, ACIVS 2013, Poznań, Poland, October 28-31, 2013. Proceedings* (p. 564-575). Springer.

Paillet, P., Audigier, R., Lerasle, F. & Pham, Q.-C. (2014). Perception-prediction-control Architecture for IP Pan-Tilt-Zoom Camera through Interacting Multiple Models. In S. Battiato & J. Braz (Éd.), *VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications, Volume 3, Lisbon, Portugal, 5-8 January, 2014* (p. 314-324). SciTePress.

Paillet, P., Audigier, R., Lerasle, F. & Pham, Q.-C. (2015). Collaborative Tracking and Distributed Control for an IP-PTZ Camera Network. In M. D. Marsico, M. A. T. Figueiredo & A. L. N. Fred (Éd.), *ICPRAM 2015 - Proceedings of the International Conference on Pattern Recognition Applications and Methods, Volume 2, Lisbon, Portugal, 10-12 January, 2015* (p. 219-226). SciTePress.

Pham, Q., Gond, L., Begard, J., Allezard, N. & Sayd, P. (2007). Real-Time Posture Analysis in a Crowd using Thermal Imaging. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1-8.

Pham, Q. C., Vincent, F., Clarysse, P., Croisille, P. & Magnin, I. E. (2001). A FEM-based deformable model for the 3D segmentation and tracking of the heart in cardiac MRI. *ISPA 2001. Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis. In conjunction with 23rd International Conference on Information Technology Interfaces (IEEE Cat.*, 250-254.

Pham, Q.-C., Lapeyronnie, A., Baudry, C., Lucat, L., Sayd, P., Ambellouis, S., Sodoyer, D., Flancquart, A., Barcelo, A.-C., Heer, F., Ganansia, F. & Delcourt, V. (2010). Audio-video surveillance system for public transportation. *2010 2nd International Conference on Image Processing Theory, Tools and Applications*, 47-53.

Pham, Q. C. (2002). *Segmentation et mise en correspondance en imagerie cardiaque multimodale conduites par un modèle anatomique bi-cavités du coeur* (PhD Thesis).

Pham, Q.-C., Dhome, Y., Gond, L. & Sayd, P. (2008). Video Monitoring of Vulnerable People in Home Environment [event-place : Ames, IA, USA]. *Proceedings of the 6th International Conference on Smart Homes and Health Telematics*, 90-98.

Ribeiro, P. C., Audigier, R. & Pham, Q. C. (2016). RIMOC, a feature to discriminate unstructured motions : Application to violence detection for video-surveillance. *Computer Vision and Image Understanding*, *144*, 121-143.

Vray, D., Discher, A., Lefloc'h, J., Mai, W., Clarysse, P., Pham, Q. C., Montagnat, J. & Janier, M. (2002). 3D quantification of ultrasound images : Application to mouse embryo imaging in vivo. *2002 IEEE Ultrasonics Symposium, 2002. Proceedings.*, *2*, 1597-1600 vol.2.

Vu, V., Bremond, F., Davini, G., Thonnat, M., Pham, Q., Allezard, N., Sayd, P., Rouas, J., Ambellouis, S. & Flancquart, A. (2006). Audio-Video Event Recognition System for Public Transport Security. *2006 IET Conference on Crime and Security*, 414-419.

Brevets

Dhome, Y., Luvison, B. & Pham, Q. C. (2013). *Method for locating objects by resolution in the three-dimensional space of the scene* (WO/2013/057030).

Maggio, S., Luvison, B. & Pham, Q. C. (2014). *Method for tracking a target in an image sequence, taking the dynamics of the target into consideration* (WO/2014/135404).