Techniques d'optimisation pour la détection et ré-identification de personnes dans un réseau de caméras

Francisco Rodolfo Barbosa-Anda

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ FÉDÉRALE TOULOUSE MIDI-PYRÉNÉES

Délivré par :
l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Présentée et soutenue le *10/12/2018* **par :**

FRANCISCO RODOLFO BARBOSA ANDA

TECHNIQUES D'OPTIMISATION POUR LA DÉTECTION ET RÉ-IDENTIFICATION DE PERSONNES DANS UN RÉSEAU DE CAMÉRAS

JURY

LYNDA TAMINE-LECHANI	Professeur des universités	Président du Jury
MARIE BABEL	Maître de conférences	Membre du Jury
VINCENT T'KINDT	Professeur des universités	Membre du Jury
JORGE MANUEL PEREIRA BATISTA	Professor associado	Membre du Jury
CYRIL BRIAND	Professeur des universités	Membre du Jury
FRÉDÉRIC LERASLE	Professeur des universités	Membre du Jury

École doctorale et spécialité :
EDSYS : Informatique 4200018
Double mention :
EDSYS : Robotique 4200046
Unité de Recherche :
Laboratoire d'analyse et d'architecture des systèmes
Directeur(s) de Thèse :
Cyril BRIAND et *Frédéric LERASLE*
Rapporteurs :
Vincent T'KINDT et *Jorge Manuel M. C. PEREIRA BATISTA*

Résumé : Cette thèse traite de la détection et de la ré-identification de personnes dans un environnement instrumenté par un réseau de caméras à champ disjoint. Elle est à la confluence des communautés Recherche Opérationnelle et Vision car elle s'appuie sur des techniques d'optimisation combinatoire pour formaliser de nouvelles modalités de vision par ordinateur. Dans ce contexte, un détecteur visuel de personnes, basé sur la programmation linéaire en nombres entiers, est tout d'abord proposé. Son originalité est de prendre en compte le coût de traitement et non uniquement les performances de détection. Ce détecteur est évalué et comparé aux détecteurs de la littérature les plus performants. Ces expérimentations menées sur deux bases de données publiques mettent clairement en évidence l'intérêt de notre détecteur en terme de coût de traitement avec garantie de performance de détection. La seconde partie de la thèse porte sur la modalité de ré-identification de personnes. L'originalité de notre approche, dénommée D-NCR (pour *Directed Network Consistent Re-identification*), est de prendre explicitement en compte les temps minimum de transit des personnes dans le réseau de caméras et sa topologie pour améliorer la performance de la ré-identification. On montre que ce problème s'apparente à une recherche de chemins disjoints particuliers à profit maximum dans un graphe orienté. Un programme linéaire en nombres entiers est proposé pour sa modélisation et résolution. Les évaluations réalisées sur une base publique d'images sont prometteuses et montrent le potentiel de cette approche.

Mots clés : optimisation combinatoire, détection de personnes, ré-identification de personnes

Abstract : This thesis deals with people detection and re-identification in an environment instrumented by a network of disjoint-field cameras. It stands at the confluence of the Operational Research and Computer Vision communities as combinatorial optimization techniques are used to formalize new computer vision methods. In this context, a people visual detector, based on mixed-integer programming, is first propose that simultaneously take computation time and detection performances into account. This detector is evaluated and compared to the best detectors of the literature. These experiments, conducted on two public databases, clearly demonstrate the interest of our detector in terms of processing time with detection performance guarantee. The second part of the thesis deals with people re-identification. Our novel approach, called D-NCR (Directed Network Consistent Re-identification), explicitly takes minimum transit times in the camera network into account, as well as the network topology, in order to improve the re-identification performance. This problem is similar to the determination of particular maximum-profitable independent paths in an oriented graph. A mixed-integer program is proposed to model and solve this problem. The experiments made on a public dataset sound promising and tend to prove the potential of the approach.

Keywords : combinatorial optimization, people detection, people re-identification

Remerciements

> « L'homme, sans aucun appui et sans aucun secours, est condamné à chaque instant à inventer l'homme. » (Jean-Paul Sartre, L'Existentialisme est un Humanisme)

Heureusement pour moi, l'homme qui s'est lancé dans la réalisation de cette thèse, n'était pas sans appui. Cette aventure personnelle a ainsi été possible grâce au soutien de plusieurs personnes que je veux ici remercier.

Le premier appui dont je suis reconnaissant est celui fourni par mes directeurs de thèse, Cyril BRIAND et Frédéric LERASLE. Ils m'ont fourni la possibilité de m'embarquer dans cette aventure mais, au-delà, m'ont accompagné pendant toute la durée de mon travail en recommandant les directions et offrant les conseils qui m'ont aidés à arriver ici aujourd'hui.

Je souhaite remercier ensuite mes rapporteurs, Vincent T'KINDT et Jorge Manuel M. C. PEREIRA BATISTA. Ils ont accepté de relire et rapporter ce manuscrit. Je les remercie également pour leur présence lors de ma soutenance. Je remercie Lynda TAMINE-LECHANI pour m'avoir fait l'honneur de présider mon jury de thèse, et Marie BABEL pour avoir accepté d'en faire partie. L'heureux dénouement de cette aventure n'aurait pas été possible sans vous.

Je remercie Jean ARLAT et Liviu NICU, directeurs successifs du LAAS-CNRS, théâtre principal de mon aventure. Je remercie également Michel DEVY, qui a construit les liens entre l'Université de Guanajuato et le LAAS-CNRS, liens sans lesquels rien n'aurait été possible. Je souhaite remercier ensuite Christian ARTIGUES et Marie-José HUGUET, responsables successifs de l'équipe ROC, ainsi que Patrick DANES, responsable de l'équipe RAP, pour m'avoir accueilli dans leurs équipes. Je remercie également tous les membres des équipes ROC et RAP qui m'ont fourni le soutien moral tout au long de cette thèse.

Je souhaite remercier spécialement mes parents, Mercedes Elena ANDA PEÑA et Francisco BARBOSA CASTILLO, et mon épouse, Valeria Irazú RAMÍREZ RANGEL, pour leur soutien et leur patience tout au long de cette période. Je remercie également toute ma famille au Mexique et en Colombie, particulièrement mes grand-parents, Blanca Aurora CASTILLO CUBILLOS et Francisco BARBOSA BERNAL.

Enfin, je remercie le Conseil National de Science et Technologie (CONACyT) du Mexique et le Centre National de la Recherche Scientifique (CNRS) de France pour le soutien et financement des travaux présentés dans cette thèse.

Table des matières

Liste des tableaux

Table des figures

Liste des Algorithmes

Liste des Abréviations

ACF En anglais *Aggregate Channel Features*.

ACP Analyse en composantes principales (En anglais *Principal component analysis*).

BIP Programmation linéaire en nombres binaires (En anglais *Binary integer programming*).

CALTECH En anglais *California Institute of Technology*.

CellLBP En anglais *Cell Structured Local Binary Patterns*.

CIELAB Espace chromatique L*a*b* CIE 1976.

CIELUV Espace chromatique L*u*v* CIE 1976.

CMC En anglais *Cumulative Matching Characteristics*.

CPU Processeur (En anglais *Central processing unit*).

CSS En anglais *Color Self Similarity*.

D-NCR En anglais *Directed Network Consistent Re-identification*

DBP En anglais *Direct Backward Pruning*.

DeepPed En anglais *Deep Convolutional Neural Networks for Pedestrian Detection*.

DET En anglais *Detection Error Trade-off*.

DPM En anglais *Deformable Part-based Model*.

DRLBP En anglais *Discriminative Robust Local Binary Patterns*.

EHPAD Établissement d'hébergement pour personnes âgées dépendantes.

ELF En anglais *Ensemble of localized features*.

EOH Histogramme d'orientation de contours (En anglais *Edge Orientation Histogram*).

FA Fausse alarmes (En anglais *False alarms*).

FN Faux négatifs (En anglais *False negatives*).

FNR Taux de faux négatifs (En anglais *False negatives rate*).

FOV Champ de vue (En anglais *Field of view*).

FP Faux positifs (En anglais *False positives*).

FPPI Faux positifs par image (En anglais *False positives per image*).

FPR Taux de faux positifs (En anglais *False positive rate*).

FPS Images par seconde (En anglais *Frames per second*).

GLS En anglais *Graph Local Search*.

GOG En anglais *Gaussian of Gaussian*.

GPU Processeur graphique (En anglais *Graphics processing unit*).

HDA En anglais *High Definition Analytics*.

HOF En anglais *Histogram of Flow*.

HOG Histogramme de gradient orienté (En anglais *Histogram of oriented gradients*).

HS Teinte, saturation (En anglais *Hue, saturation*).

HSV Teinte, saturation, valeur (En anglais *Hue, saturation, value*).

ICORES En anglais *International Conference on Operations Research and Enterprise Systems*.

ICT En anglais *Implicit Camera Transfer*.

ID Identité (En anglais *Identity*).

IMH En anglais *Integral Motion Histograms*.

INRIA Institut National de Recherche en Informatique et en Automatique.

ISR En anglais *Institute for Systems and Robotics*.

ITML En anglais *Information-theoretic metric learning*.

k-PPV k plus proches voisins (En anglais *k-nearest neighbors*).

KISSME En anglais *Keep It Simple and Straightforward MEtric*.

kMFA En anglais *Kernel Marginal Fisher Analysis*.

LatSvm En anglais *Latent Support Vector Machines*.

LBP Motifs binaires locaux (En anglais *Local Binary Patterns*).

LDA Allocation de Dirichlet latent (En anglais *Latent Dirichlet allocation*).

LDA Analyse Discriminante Linéaire (En anglais *Linear Discriminant Analysis*).

LDCF En anglais *Local Decorrelated Channel Features*.

LDFV En anglais *Local descriptors encoded by Fisher vector*.

LFDA En anglais *Local Fisher discriminant analysis*.

LMNN En anglais *Large margin nearest neighbor*.

LOMO En anglais *Local maxima occurrence*.

MBH En anglais *Motion Boundary Histograms*.

MCL En anglais *Multiple Component Learning.*

MOT En anglais *Multiple Object Tracking.*

MR Taux de défaut (En anglais *Miss rate*).

MS En anglais *Mean-Shift.*

MSCR En anglais *Maximally stable color regions.*

MSCRMP En anglais *Mean-Cascade Response-Time Minimization Problem.*

NCDA En anglais *Network Consistent Data Association.*

NCR En anglais *Network Consistent Re-identification.*

NMS Suppression des non-maxima (En anglais *Non-Maximal Suppression*).

NNNF En anglais *Non-Neighboring and Neighboring Features.*

NP Non déterministe polynomial (En anglais *Non-deterministic polynomial-time*).

NRLBP En anglais *Non-Redundant Local Binary Patterns.*

PCCA En anglais *Pairwise constrained component analysis.*

PHOG En anglais *Pyramid Histogram of Oriented Gradients.*

PLNE Programmation linéaire en nombres entiers.

PM En anglais *Pairwise Max.*

PMT En anglais *Piotr's Computer Vision Matlab Toolbox.*

PTZ Azimut, élévation et zoom (En anglais *pan-tilt-zoom*).

RAM Mémoire vive (En anglais *Random-access memory*).

RCNN En anglais *Region-based Convolutional Neural Network.*

RGB Rouge, vert, bleu (En anglais *Red, green, blue*).

RHSP En anglais *Recurrent high-structured patches.*

ROADEF Société Française de Recherche Opérationnelle et d'Aide à la Décision.

ROC Fonction d'efficacité du récepteur (En anglais *Receiver Operating Characteristic*).

RPN En anglais *Region Proposal Network*

RPN+BF En anglais *Region Proposal Network and Boosted Forests.*

ré-ID Ré-identification.

SCNCD En anglais *Salient color names based color descriptor.*

SDALF En anglais *Symmetry-driven accumulation of local features.*

SIFT En anglais *Scale-invariant feature transform*.

SPRT En anglais *Sequential probability ratio test*.

SSP Problème de la somme de sous-ensembles (En anglais *Subset-Sum Problem*).

SVM Machines à vecteurs de support ou séparateurs à vaste marge (En anglais *Support Vector Machines*).

TN Vrais négatifs (En anglais *True negatives*).

TNR Taux de vrais négatifs (En anglais *True negative rate*).

TP Vrais positifs (En anglais *True positives*).

TPR Taux de vrais positifs (En anglais *True positive rate*).

VOC En anglais *Visual Object Classes*.

WACV En anglais *Winter Conference on Applications of Computer Vision*.

WHOS En anglais *Weighted Histograms of Overlapping Stripes*.

XQDA En anglais *Cross-view Quadratic Discriminant Analysis*.

YCbCr Espace chromatique YCbCr.

YUV Espace chromatique YUV.

Introduction

Contexte des travaux et problématique associée

Ces dernières années, les développements de systèmes de surveillance intelligents se sont multipliés et l'engouement de la communauté scientifique pour développer des systèmes toujours plus performants s'est considérablement accru [Räty 2010, DOrazio 2015]. Räty [Räty 2010] décrit un système de surveillance comme une technologie qui aide les opérateurs humains en offrant une capacité de perception et de raisonnement étendue sur les situations d'intérêt qui se produisent dans des environnements surveillés. Au fil du temps, les systèmes de surveillance ont évolué, passant de systèmes analogiques médiocres à des systèmes de plus en plus complexes capables de détecter automatiquement un événement et, en fonction, d'entreprendre les actions nécessaires. La demande se situe principalement dans les domaines d'application suivants :

La sécurité publique : Pour assurer la sécurité dans des espaces publiques telles que les aéroports [Foucher 2011], les gares [Ronetti 2000], les ports maritimes [Pozzobon 1999], les banques [Zambanini 2009], les centres commerciaux [Bouma 2013] et les parkings [Micheloni 2003] ;

La santé : Pour surveiller les activités des personnes âgées (domicile, EHPAD) et prévenir les accidents et chutes [Zouba 2009] ou surveiller les patients [Pallikonda Rajasekaran 2010] ;

Les services : Pour fournir de l'aide aux personnes nécessitant une assistance automatique avec des bagages lourds [Jayawardena 2010] et une aide à la navigation [Bennewitz 2005] ;

L'inspection : Pour les systèmes automatisés qui inspectent les entrepôts et les sites de stockage, identifient les situations anormales, telles que les inondations et les incendies, détectent les intrus et déterminent le statut des objets inventoriés [Everett 2003] ;

Le militaire : Pour les applications militaires telles que la surveillance des frontières, le suivi d'ennemi, la surveillance de champs de bataille et la classification de cibles [Arampatzis 2005].

Les domaines d'applications sont donc multiples et variés. Les scénarios à interpréter impliquent des cibles telles que des individus, des foules, des véhicules, des objets inanimés, etc.

Notre travail se focalise sur la surveillance de cibles de type personne dans les lieux publics. L'analyse automatique des flux vidéo porte alors sur la détection, le suivi et la reconnaissance d'événements concernant des usagers du lieu issus des caméras instrumentant le lieu. L'enjeu est classiquement la sécurité des lieux publics, c'est-à-dire la reconnaissance d'actions humaines malveillantes, hostiles, ou en détresse.

La figure 1 illustre la séquence des fonctions mises en jeu dans un tel système pour notre contexte applicatif. Tout d'abord, le système perçoit l'environnement à l'aide d'un réseau de capteurs, ici optiques, instrumentant de façon éparse l'environnement. Toutes les personnes dans la zone surveillée sont alors détectées à l'aide du flux vidéo fourni par chaque caméra. La détection alimente un module de suivi qui capture des informations spatio-temporelles relatives à chaque personne cible dans la zone observée. Ces informations inhérentes à chaque capteur sont classiquement fusionnées avec celles provenant des autres capteurs. Les informations spatio-temporelles résultantes caractérisent les activités de chaque personne au sein du réseau. Leur analyse permet alors de vérifier si les comportements, les activités des usagers en transit dans le lieu sont normales ou non. Dans le cas contraire, le système déclenche une alarme destinée à un opérateur humain, une société de surveillance, voire un robot mobile.

Les environnements surveillés sont généralement de grande dimension et relativement structurés (rues, couloirs, passerelles, etc.), présentant donc une certaine topologie. Le défi consiste alors à assurer une couverture éparse de l'environnement via un nombre raisonnable de capteurs de sorte à prendre en compte les exigences en matière de traitement informatique (temps, espace mémoire), de coût financier et de performances attendues. Nous discutons, ci-après, les avantages et inconvénients des principaux types de caméras déployées en vidéo-surveillance (figure 2).

Dans le cas de caméras perspectives classiques (figure 2a), le nombre de caméras requis pour observer l'espace est généralement très élevé, nécessitant une bande passante importante, des unités de traitement spécifiques, induisant un système coûteux. Néanmoins ces capteurs étant fixes, des algorithmes simples et rapides de segmentation peuvent être utilisés pour détecter les personnes en mouvement dans le champ de vision de la caméra. En fonction de la configuration réelle du capteur, leur champ de vision peut couvrir une large zone, fournissant ainsi une perception globale. Ils peuvent ainsi observer des cibles sur une vaste zone pendant une période prolongée. Une difficulté concerne la gestion des angles morts liés à leurs configurations spatiales ou aux éventuelles occultations, etc.

Les caméras *pan-tilt-zoom* (PTZ) (figure 2b) exploitent leurs actionneurs azimut, élévation et zoom pour surveiller des zones plus larges et focaliser sur des individus, des objets ou des événements spécifiques. En conséquence, un nombre plus limité de caméras est a priori nécessaire pour instrumenter la scène. Cependant, l'inconvénient de ces caméras est que le mouvement de la caméra associé à la latence du réseau rend la modélisation de la caméra et la détection / suivi des personnes plus complexes. En effet, l'utilisation de ces caméras impose de planifier leur activité et de gérer cette activité en temps réel, en fonction des détections réalisées. De plus, comme pour les caméras classiques, leur champ de vision reste limité à la zone visible couverte via leurs actionneurs azimut-élévation-zoom.

D'autres solutions plus exotiques existent comme les caméras omnidirectionnelles ou panoramiques (figure 2c). Ces dernières fournissent un champ de vision de 360°. Elles sont certes attrayantes mais relativement coûteuses et exigeantes en termes de calcul du fait que les images panoramiques sont en haute résolution.

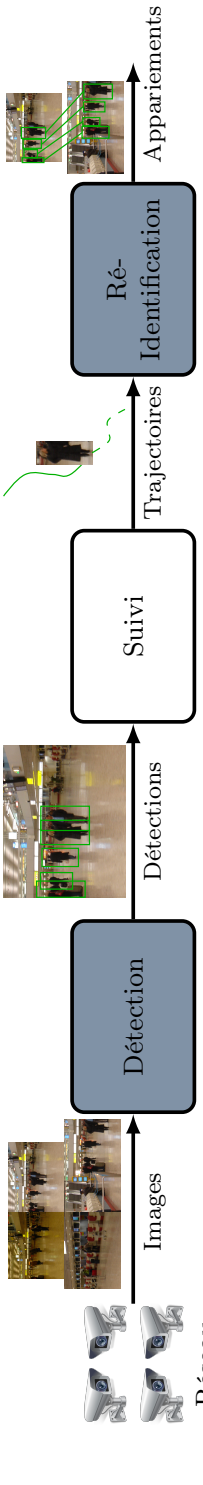


FIGURE 1 – Synoptique général d'un système de vidéo-surveillance.

(a) Caméra fixe. (b) Caméra PTZ. (c) Caméra 360°.

FIGURE 2 – Quelques exemples de caméras déployées en vidéo-surveillance.

Dans notre étude, nous nous plaçons sous l'hypothèse d'un réseau de caméras classiques perspectives qui coopèrent afin d'intégrer les informations produites par chacune, par exemple les détections de cibles propres à chaque champ de vue.

Dans ce contexte, la première partie de nos travaux s'intéresse à un détecteur visuel générique de personnes pouvant être utilisé à partir de caméras perspectives ambiantes. Nous formalisons et évaluons en particulier un détecteur de personnes original, ce type de détecteur n'étant en effet pas étudié dans la littérature (pléthorique) de la communauté Vision, qui considère (optimise) simultanément les performances de détection mais aussi le coût de traitement lié aux détections. L'approche repose sur une modélisation par programme linéaire en nombres entiers (PLNE, modèle très utilisé dans la communauté Recherche Opérationnelle).

La seconde partie de nos travaux propose une réformalisation originale de la modalité de ré-identification (ré-ID) de personnes exploitant : (1) la topologie du réseau de caméras et (2) le temps de transit entre champs de vue de caméras. A notre connaissance, ces contraintes (globales au réseau), pourtant intuitives dès lors que la configuration des caméras est figée dans l'environnement, sont très peu exploitées dans la littérature ; les approches classiques privilégiant une inférence entre paires de caméras prises séparément, sans garantie de cohérence globale. Ici encore, une modélisation PLNE prenant explicitement en compte les caractéristiques du réseau de caméras, donc une inférence globale, est proposée.

Nos travaux se focalisent donc sur la détection et la ré-ID des personnes (blocs bleus dans la figure 1), deux modalités essentielles en vidéo surveillance.

On notera que nos investigations se situent à la confluence des deux communautés Vision et Recherche Opérationnelle.

Contributions et plan du mémoire

Ce mémoire de thèse est structuré en deux parties. Les contributions inhérentes à chaque partie sont détaillées ci-dessous.

La première partie du mémoire porte sur la modalité liée à la détection visuelle de personnes. Tout d'abord, un état de l'art est présenté. Puis, nous développons

une nouvelle formulation mathématique de type PLNE pour régler de façon optimale les paramètres libres du détecteur en considérant simultanément : (1) les temps de calcul et (2) les performances de détection. Des évaluations sur les bases publiques INRIA et CALTECH permettent de valider notre approche et d'étudier ses performances.

La deuxième partie du mémoire, porte sur la modalité ré-identification de personnes. Tout d'abord, un état de l'art est présenté. Nous développons ensuite une nouvelle formulation mathématique de type PLNE pour l'appariement des personnes inter-caméra en considérant simultanément : (1) le temps de transit, et (2) la topologie du réseau de capteurs. Des évaluations préliminaires sur la base publique HDA montrent des résultats prometteurs.

Ce mémoire se termine par une conclusion générale et des perspectives.

Ces travaux ont été valorisés par les publications suivantes :

Publications internationales

Conférences

[**BarbosaAnda 2016a**] F. Barbosa-Anda *et al.* « Mean Response-Time Minimization of a Soft-Cascade Detector ». Dans : *Int. Conf. on Operations Research and Enterprise Systems (ICORES'16)*. Rome, Italy, fév. 2016, p. 252–260

[**BarbosaAnda 2018a**] F. R. Barbosa-Anda *et al.* « Soft-Cascade Learning with Explicit Computation Time Considerations ». Dans : *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Mar. 2018, p. 1234–1243

Publications nationales

Conférences

[**BarbosaAnda 2016b**] Francisco Rodolfo Barbosa-Anda *et al.* « Minimisation du Temps de Réponse moyen d'une Cascade de Détection ». Dans : *Congrès annuel de la société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF)*. Compiègne, France, fév. 2016, 2p.

[**BarbosaAnda 2018b**] Francisco Rodolfo Barbosa-Anda *et al.* « Partitionnement en cliques à profit maximum de graphes orientés avec contraintes de flot ». Dans : *Congrès annuel de la société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF)*. Lorient, France, fév. 2018, 2p.

Plan du mémoire

Le manuscrit est organisé comme suit :

Première partie - Détection de personnes avec prise en compte explicite du temps de réponse

La première partie de ce travail concerne la détection automatisée des personnes à partir d'un capteur visuel. Cette partie est divisée en deux chapitres. Le premier (chapitre 1) présente les tendances, les modes et les différentes considérations de la détection visuelle des personnes dans la littérature. Il traite de l'état de l'art en matière de la détection des personnes en mettant l'accent sur les approches basées sur la vision. Il est utilisé comme point de départ pour mettre en perspective les contributions apportées au chapitre 2.

Le chapitre 2 présente un détecteur de personnes basé sur une *soft-cascade* présentant un nouveau cadre de sélection de paramètres basé sur la programmation linaire en nombres entiers (PLNE). Une formulation détaillée de ce cadre d'optimisation et de son application à un apprentissage par détecteur en *soft-cascade* sont présentées.

Deuxième partie - Ré-identification de personnes dans un réseau de caméras avec des considérations de temps de transit et de topologie du réseau

La deuxième partie de ce travail porte sur la ré-identification des personnes dans une zone surveillée. Similairement à la partie I, il commence par une présentation de l'état de l'art en matière de ré-identification de personnes (chapitre 3) qui permet de mettre en évidence les contributions apportées.

Le chapitre 4 présente une méthodologie de ré-identification de personnes présentant une règle de décision basée sur la programmation linaire en nombres entiers (PLNE). Une formulation détaillée de ce cadre d'optimisation et de son application à une réseau de vidéo-surveillance sont présentées.

Conclusions générales et perspectives

Enfin, le manuscrit finit avec un résumé de nos contributions et discute des améliorations, des limites et des perspectives.

Première partie

Détection de personnes avec prise en compte explicite du temps de réponse

Chapitre 1

État de l'art et positionnement des travaux

Sommaire

La détection visuelle d'objets a une grande importance dans la communauté Vision par Ordinateur avec des retombées en vidéo surveillance [Breitenstein 2011], indexation d'images [Zhang 2009], robotique [Ess 2010], aide à la conduite automobile [Gerónimo 2010a], etc. Cet intérêt est corrélé avec la puissance toujours croissante des ressources informatiques ; citons ici les travaux de Zhang *et al.* [Zhang 2013], Dollár *et al.* [Dollár 2012] et Gerónimo *et al.* [Gerónimo 2010a].

Ceci se vérifie également à travers les différents défis proposés, par exemple ImageNet [Russakovsky 2015], MOT *Challenge* [LealTaixé 2015], the Pascal VOC *Challenge* [Everingham 2010, Everingham 2015]. Nous focalisons notre étude sur la détection visuelle d'objets de type personne qui est une problématique à fort enjeu applicatif et très investiguée dans la communauté [Nguyen 2016] ; nous pourrons ainsi mieux nous positionner et nous comparer aux approches existantes.

Pour rappel, la détection visuelle de personnes vise à détecter l'ensemble des personnes dans le plan image en caractérisant leur position et échelle, donc la boite englobante associée. Nous listons ci-après les principaux verrous associés à cette problématique :

Variations morphologiques des cibles : la morphologie humaine varie beaucoup d'un individu à l'autre, ces variations morphologiques peuvent être induites par le port de vêtements amples ou serrés ;

Variations d'apparence des cibles : l'apparence vestimentaire, la couleur des cheveux et/ou de la peau varient aussi beaucoup ;

Variations de postures des cibles : l'enveloppe corporelle est déformable ;

Variations d'illumination de la scène : les systèmes de vision sont passifs, et donc dépendants des conditions d'illumination de la scène observée ;

Variations de points de vues : les cibles subissent de fortes variations dans le plan image de par le point de vue relatif caméra/scène ;

Encombrement des scènes perçues : l'environnement peut être possiblement encombré et donc induire des fausses détections, voire des occultations ;

Capteurs : les capteurs ont des limitations physiques liées à sa structure optique et électronique. Les images sont par définition bruitées ;

Puissance de calcul : les détecteurs performants induisent des coûts en temps de traitement qui doivent rester compatibles avec l'application. Il est donc important de trouver un compromis entre performances de détection et coût de calcul.

Dans la littérature, la démarche courante de détection visuelle est basée sur une technique de fenêtres glissantes [Viola 2004], surtout lorsqu'il n'y a pas d'information de contexte a priori sur l'image analysée [Dollár 2012]. Le principe est alors d'isoler (segmenter dans l'image) les cibles par un balayage exhaustif des positions et échelles en classant les boîtes englobantes ainsi générées après apprentissage supervisé. Ce principe est évidemment coûteux en temps de calcul.

Ce chapitre vise à positionner et motiver nos choix pour notre modalité de détection de personnes ; celle-ci sera décrite et évaluée au chapitre suivant. Ainsi, la section 1.1 catégorise les principales approches existantes. La section 1.2 focalise sur certaines méthodes clés en lien avec nos travaux. Le chapitre se poursuit par une présentation de différentes bases de données publiques dites de *benchmarks* (section 1.3) et des métriques d'évaluation (section 1.4) qui seront exploitées ultérieurement. Nous discutons alors nos choix (section 1.5). Le chapitre se conclut par un bilan (section 1.6).

1.1 État de l'art

Comme évoqué, la littérature sur la détection de personnes est assez vaste, ainsi que l'atteste les nombreux états de l'art récents [Dollár 2012, Gerónimo 2010a]. Pour résumer, les premiers succès ont été obtenus en utilisant des descripteurs type ondelettes de Haar qui capturent les différences locales d'intensité d'une région, mais ont un pouvoir descriptif limité [Papageorgiou 2000, Viola 2004]. Ceux-ci ont été améliorés de manière significative grâce à l'utilisation de descripteurs HOG basés gradient [Dalal 2005].

De nombreux travaux ont alors exploité ces descripteurs HOG, parfois combinés avec d'autres descripteurs. Citons par exemple : le détecteur HogLBP [Wang 2009] combinant les descripteurs HOG avec les motifs binaires locaux (LBP), et le détecteur MultiFTR [Wojek 2008] combinant HOG avec Haar et des descripteurs de contexte de forme. Des gains substantiels ont ensuite été obtenus par le détecteur *Deformable Parts-based Model* (DPM) qui utilise des descripteurs HOG légèrement modifiées dans une configuration de détection basée sur des pièces qui recherche explicitement différentes parties automatiquement apprises d'une personne (cinq pour être exact) pour la détecter [Felzenszwalb 2010]. Citons ensuite les détecteurs basés sur des *Channel Features* [Dollár 2009a] ou leurs dérivées. Les *Channel Features* combinées avec des processus de *Soft-Cascade* par *Boosting* [Zhang 2007, Bourdev 2005] offrent un bon compromis taux de détection par rapport au temps de calcul [Zhang 2015]. Plus récemment encore, les détecteurs visuels privilégient des paradigmes d'apprentissage approfondi [Hosang 2015, Tian 2015, Ren 2017]. Bien que ces paradigmes se révèlent très prometteurs, des expérimentations récentes montrent que, les détecteurs basés *Soft-Cascade* restent compétitifs en termes de performances de détection et de temps de calcul [Zhang 2016b].

Toutes les méthodes de la littérature respectent plus ou moins le schéma générique de la figure 1.1. Des fenêtres ou boîtes englobantes sont d'abord générées sur l'image à analyser sous l'hypothèse que la cible est présente à l'intérieur. Des descripteurs visuels, supposés discriminants, sont alors extraits de la fenêtre candidate ; ceux-ci sont appris durant une étape hors ligne d'apprentissage supervisé. Enfin, la fenêtre est classée en cible ou non cible (détection = classification binaire) via une règle de décision issue du classificateur. Bien que non mentionné sur la figure, il existe généralement une dernière étape de post-traitement visant à supprimer les détections qui se chevauchent par une suppression des non-maxima (NMS). Son but est de fusionner plusieurs détections associées à la même personne en une seule, comme celles présentes dans la figure 1.2. Deux approches principales utilisées pour cela dans la littérature sont l'estimation du mode *mean-shift* (MS) [Dalal 2006a] et la suppression *Pairwise Max* (PM) [Felzenszwalb 2010]. La seconde (PM) sélectionne la boîte englobante la plus probable parmi les boîtes en recouvrement. La première (MS), comme son nom l'indique, sélectionne la boîte moyenne des boîtes en recouvrement.

La figure 1.1 illustre un schéma bloc classique de détection visuelle *online*. La nature des descripteurs et du classificateur ainsi que le modèle de personne exact

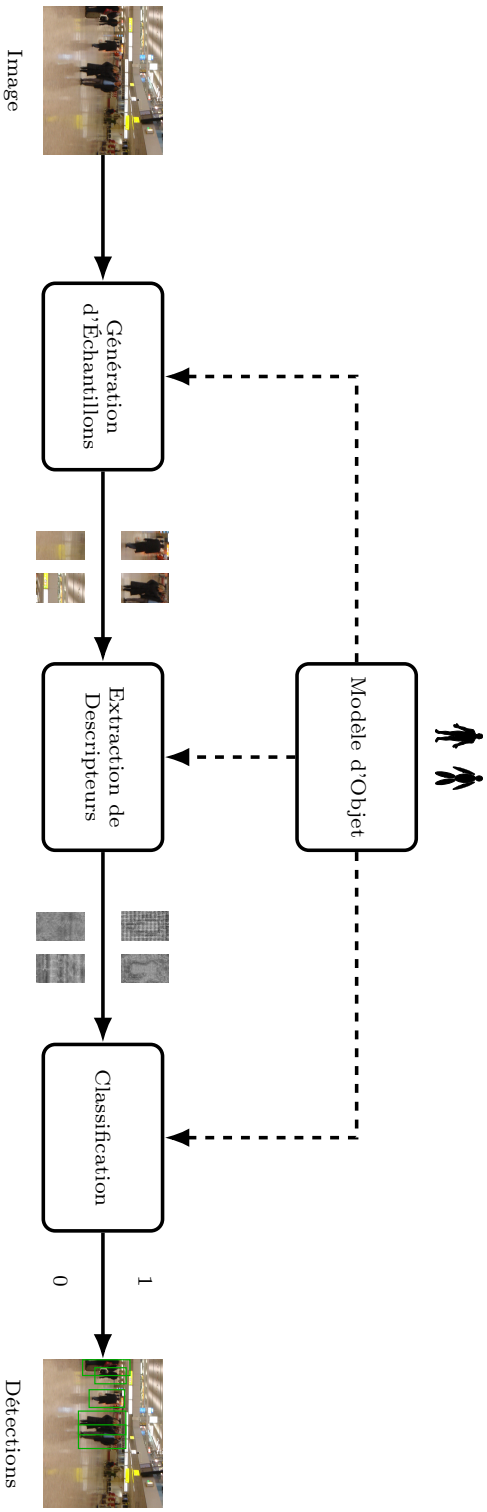


FIGURE 1.1 – Détection visuelle de cibles : classification des fenêtres candidates (processus *online*).

(a) Sans NMS. (b) Avec NMS.

FIGURE 1.2 – Exemple de détection sans / avec suppression des non-maxima (NMS).

utilisé sont à choisir préalablement. Mais, le sous-ensemble (discriminant) des descripteurs à utiliser et le réglage des paramètres du classificateur sont déterminés par un apprentissage hors ligne reposant sur des données d'apprentissage contenant des échantillons ou boîtes englobantes annotées cible, personne dans notre travaux, (positifs) ou sinon (négatifs). La figure 1.3 illustre l'apprentissage du détecteur de personnes.

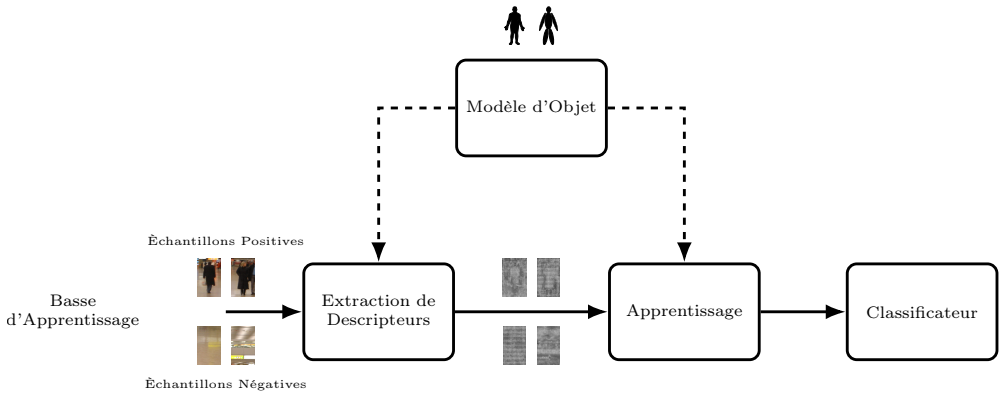


FIGURE 1.3 – Détection visuelle de cibles : processus *offline* d'apprentissage.

1.1.1 Modèles de personne

Classiquement, le détecteur de personnes (figure 1.1) séquence trois étapes : la génération de fenêtres candidates, l'extraction des descripteurs et la classification. Tous ces blocs utilisent généralement un modèle sous-jacent qui encapsule la signature générique de personnes pour rechercher le corps humain complet ou, alternativement, chercher des parties corporelles et les concaténer pour déduire la présence d'une personne. La littérature différencie les méthodes implicites et explicites. Les méthodes implicites n'utilisent pas les indices spécifiques aux humains, préférant

d'autres indicateurs. Par exemple, le mouvement ou l'écart par rapport à la norme. Inversement, les méthodes explicites utilisent des caractéristiques spécifiques aux personnes. La figure 1.4 catégorise les détecteurs existants selon le modèle de cible privilégié.

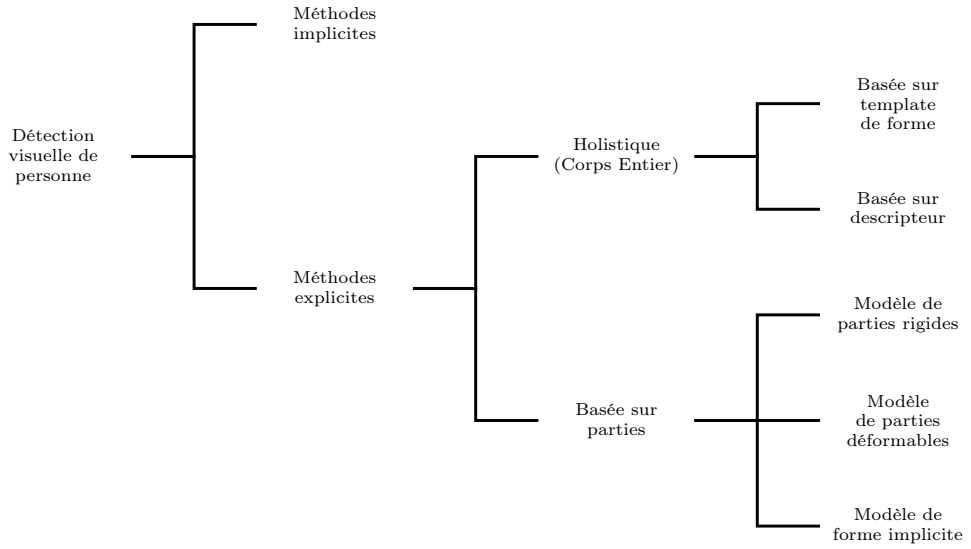


FIGURE 1.4 – Taxonomie des méthodes de détection visuelle de personne.

1.1.1.1 Méthodes implicites

Les méthodes implicites détectent les personnes en considérant leur différence avec l'environnement. Il s'agit de méthodes implicites parce qu'elles ne déduisent pas explicitement la présence de personnes, mais segmentent les objets mobiles au premier plan et les étiquettent en tant que personnes s'ils satisfont le rapport d'aspect d'une personne moyenne. Deux techniques répandues dans cette catégorie sont la soustraction de fond [Stauffer 1999, Piccardi 2004] et l'analyse de flux optique [Beauchemin 1995]. L'inconvénient est que ces techniques ne fonctionnent que pour les caméras statiques et des cibles supposées mobiles ; elles peuvent facilement être trompées par des objets au premier plan ayant un rapport d'aspect comparable aux humains. Bref, ces approches implicites sont moins performantes et les approches explicites sont souvent préférées.

1.1.1.2 Méthodes explicites

Les méthodes explicites utilisent un modèle qui capture des attributs discriminants saillants qui distinguent les personnes des autres informations d'une image. La configuration et les paramètres exacts du modèle sont déterminés à l'aide d'exemples d'apprentissage positif et négatif. La plupart des approches repose sur

ce concept [Dollár 2012, Gerónimo 2010a, Enzweiler 2009]. Les méthodes explicites sont dissociables en deux classes : des approches holistiques (corps entier) et celles par parties. Les approches holistiques considèrent un modèle corps entier de personne. Au contraire, les approches basées parties visent à agréger la détection de parties corporelles pour isoler ou segmenter la personne dans l'image.

Approches holistiques (corps entier) Ces approches considèrent un modèle corps complet pour détecter une personne [Gavrila 2000, Gavrila 1999, Broggi 2000, Broggi 2006]. Le principe est d'apprendre hors ligne un ensemble de templates de forme de personnes à partir d'exemples positifs et représentatifs de la variabilité (apparence, silhouette, etc.). Le processus en ligne vise alors à apparier les fenêtres candidates à l'un des templates appris [Gerónimo 2010a].

Une autre approche, plus répandue, vise à apprendre (hors ligne donc) des descripteurs discriminants sur la boîte englobant le corps entier. Ces descripteurs sont censés séparer au mieux exemples positifs et négatifs via un classificateur. Ce dernier labellise en ligne les régions candidates extraites de l'image à analyser en personne ou non personne. Citons ici les descripteurs usuels type HOG [Dalal 2005] ou, plus récemment, type *Channel Features* [Dollár 2009a].

En général, les approches holistiques sont très attrayantes en raison de leur abstraction simpliste, de l'apprentissage direct de modèles et, en comparaison aux approches basées sur parties, de la réduction du temps de calcul pendant la détection *online*. Cependant, lorsque les modèles sont appris sur des personnes debout, ils sont plutôt inefficaces pour des postures différentes ou dans la cas d'occultations partielles (liées à la scène ou aux bords d'image).

Approches basées parties Ces approches reposent sur la détection de parties corporelles explicites (tête, torse, bras, etc.) ou implicites pour détecter une personne. Le principe repose classiquement sur le *Pictorial Structure Model* [Fischler 1973] représentant la cible comme un ensemble de parties connectées deux à deux [Fischler 1973, Felzenszwalb 2005].

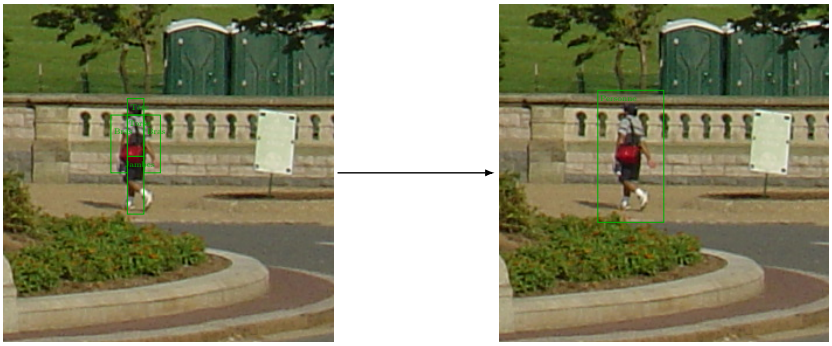


FIGURE 1.5 – Détection visuelle de personnes basées parties.

L'inconvénient avec un modèle de parties associé à l'anatomie est que les parties manquantes, en raison des occultations partielles, affectent la probabilité globale du modèle composite. Pour atténuer cela, Felzenszwalb *et al.* [Felzenszwalb 2010] ont proposé un modèle de parties déformables qui sélectionne des parties en se basant sur leur saillance visuelle – découvrant des parties de modèle de manière non supervisée – plutôt que de s'appuyer sur des informations sémantiques. Les parties réelles ne sont pas spécifiées a priori, mais les boîtes englobantes étiquetées pour le corps complet et le nombre de parties. Leur algorithme sélectionne des parties saillantes à partir des données d'apprentissage grâce à une optimisation itérative avec la carte de déformation associée. Une variante est l'apprentissage de parties basées sur l'efficacité visuelle par des exemples ; elle a été abordée par Dollár *et al.* en utilisant *Multiple Component Learning* (MCL) [Dollár 2008].

En général, les approches basées parties sont mieux adaptées à la détection des personnes car plus robustes : (1) aux occultations partielles, (2) aux changements de points de vue, et (3) aux variations de postures et déformations du corps humain.

Associer des parties à des composants anatomiques humains équivalents facilite l'abstraction, mais est sensible au problème des composants manquants [Mohan 2001, Mikolajczyk 2004]. Des modèles robustes peuvent être obtenus à l'aide de parties génériques qui ne contiennent aucune information sémantique et sont sélectionnées uniquement en fonction de leur visibilité [Felzenszwalb 2010, Leibe 2008, Dollár 2008]. Cela évite également la tâche lourde d'annoter manuellement chaque partie et simplifie la détection d'autres objets. De plus, décider des parties sémantiques peut être ambigu ou subjectif. Cependant, les avantages ci-dessus sont obtenus au détriment du temps de calcul, plus important pour l'apprentissage et la détection. Les méthodes basées parties sont peu efficaces sur des images basse résolution car elles nécessitent un support spatial suffisant pour assurer la robustesse. Une approche hybride pourrait être mise en place pour utiliser une méthode basée parties lorsque la résolution suffisante est possible et une méthode holistique sinon [Park 2010].

1.1.2 Génération de fenêtres candidates

Il existe dans la littérature plusieurs stratégies afin de générer les fenêtres candidates (pour la classification en personne ou non) dans l'image brute. Dans la littérature, trois tendances se dégagent : une approche "force brute", une approche géométrique et une approche par segmentation des points d'intérêts. La plus simple est l'approche force brute, communément appelée stratégie par fenêtres glissantes [Viola 2004]. Ici, les fenêtres candidates avec un rapport d'aspect fixe sont échantillonnées dans toutes les positions et échelles de l'image brute (figure 1.6). Cette stratégie ne requiert aucune connaissance a priori sur la scène et la caméra [Dollár 2012]. Elle est avantageuse car aucune hypothèse (position et échelle) dans l'image n'est écartée, mais sa nature combinatoire augmente le temps de calcul en raison du balayage de zones image incompatibles avec la présence d'humains (plafond, etc.).

FIGURE 1.6 – Exemples de fenêtres glissantes sur une image 769×516 pixels issue de la base INRIA.

Si la relation géométrique entre la caméra et le sol est connue (caméra dite étalonnée), celle-ci permet alors de réduire le nombre total de fenêtres candidates générées. Ainsi, pour un sol plan (classiquement en *indoor*) et une caméra étalonnée, seules les fenêtres image connectées avec la projection du plan du sol sont traitées. Ce type de contraintes géométriques diminue drastiquement le nombre total de possibilités : toutes les hypothèses ou fenêtres sur les hauts murs et plafonds (dans les environnements intérieurs) et sur le ciel (dans les environnements extérieurs) sont rejetées. Pour une caméra embarquée sur véhicule mobile, l'inférence préalable du plan du sol et des déplacements permet également de réduire cette combinatoire [Gerónimo 2010b].

Une alternative est une mécanisme pour identifier et ensuite segmenter une fenêtre candidate à des tests supplémentaires. Par exemple, pour une caméra statique ou ambiantes, les méthodes implicites précitées, typiquement la soustraction de fond, aident à pré-segmenter les zones mobiles de la scène assimilable à un mécanisme d'attention. Bref, tout mécanisme pouvant éliminer une proportion des fenêtres négatives sans manquer de cibles est avantageux afin de réduire le coût de traitement global. Aussi, divers travaux proposent des indices simples qui remplissent les conditions susmentionnées. Citons ici le contraste de la couleur, la densité des bords, les chevauchements superpixel, la symétrie des couleurs et la symétrie des bords comme indiqué dans [Alexe 2010, Paleček 2012] ; Des indices autres tels que la verticalité et l'orientation dominante, sont également proposés dans Paleček *et al.* [Paleček 2012].

Récemment, Liu *et al.* [Liu 2016] ont prototypé un réseau de neurones profond pour extraire des fenêtres candidates de différentes tailles et rapports d'aspect.

1.1.3 Descripteurs

Les défis rencontrés dans la détection visuelle de personnes sont nombreux. L'utilisation de valeurs couleur individuelles de pixels à partir de l'image d'entrée conduit inévitablement à une mauvaise performance de détection et à une généralisation, car les points individuels ne transmettent aucune information globale sur l'apparence des personnes. En regardant simplement un groupe de points voisins, des informations utiles sur l'objet sous-jacent peuvent être obtenues. Par exemple, en regardant les voisins immédiats d'un pixel, on peut déterminer la grandeur et l'orientation de la variation d'aspect spatial de l'objet sous-jacent à ce point. Un élargissement supplémentaire de la région de support peut permettre de capturer le profil structurel de bord de l'objet. Par conséquent, les représentations modifiées de l'image sont essentielles pour améliorer l'interprétation de l'image. Toute information extraite de cette manière est appelée descripteur. Les descripteurs nous permettent de caractériser ou discriminer un type de cible (ici une personne) en extrayant des informations significatives à partir de régions image associées. Ils encapsulent la signature et propriétés de la cible et ont un impact majeur sur les performances du détecteur. Au delà du classificateur, le choix de descripteurs robustes (en terme d'extraction) et discriminants, sont vitaux [Dollár 2012, Gerónimo 2010a].

Initialement, les détecteurs privilégiaient des descripteurs rudimentaires de type Haar issus des ondelettes de Haar [Papageorgiou 2000, Lienhart 2002, Viola 2005]. Ces descripteurs exploitent la somme des différences d'intensité entre contrastes locaux inhérentes à la cible à détecter. Les valeurs des descripteurs sont calculées efficacement à l'aide d'images intégrales [Viola 2004]. Plus précisément, ces descripteurs capturent le changement d'intensité locale selon différentes directions dans l'image. Ce descripteur étant paramétrable (position, orientation, échelle image), les plus discriminants sont sélectionnés par *boosting* [Schapire 2003].

Il existe aussi des descripteurs s'appuyant, non pas sur des contrastes région, mais sur la structure spatiale des pixels de contraste dans la cible. Citons les descripteurs encodant l'histogramme d'orientation de contours (EOH) initialement proposé pour la détection de visage [Levi 2004]. Ces descripteurs représentent des rapports de gradients calculés à partir d'histogrammes d'orientation de contours. Dans une région superposée donnée, les gradients sont d'abord calculés. Ensuite, un histogramme de gradients est construit en quantifiant les orientations de gradient. Ces descripteurs, connus pour être robustes aux changements d'éclairage globaux, sont a priori plus performants que les descripteurs de type Haar [Gerónimo 2007].

Une variante basée gradient image repose sur l'histogramme de gradient orienté (HOG) [Dalal 2005]. Les descripteurs HOG sont extraits d'abord en calculant le gradient, puis en construisant un histogramme pondéré par l'amplitude du gradient dans son intervalle de variation discrétisée appelée cellule. Les histogrammes des cellules voisines sont regroupés en un seul bloc, normalisés en croix et concaténés pour

donner un vecteur descripteur par bloc. Dalal et Triggs [Dalal 2005] concatènent tous les histogrammes de blocs dans une fenêtre candidate pour générer un seul descripteur de grande dimension. Cependant, Zhu *et al.* [Zhu 2006] montrent que des performances de détection comparables peuvent être obtenues en utilisant un sous-ensemble de blocs HOG de taille variable sélectionnés et combinés par *boosting*. Les descripteurs HOG sont combinés via des représentations holistiques [Dollár 2012] ou parties [Felzenszwalb 2010].

Une autre variante de descripteurs repose sur des motifs binaires locaux (LBP). Ceux-ci, intialement, sont assimilés à des descripteurs de texture [Ojala 1996]. L'idée de base est de calculer une étiquette entière modulée à puissance deux pour chaque pixel par un seuillage des pixels voisins par le pixel central passant uniformément. L'étage de seuillage considère les valeurs d'intensité relative rendant le descripteur invariant a l'éclairage et le contraste. Les motifs de texture avec un support spatial différent pourraient être capturés en faisant varier le rayon du voisinage et les points échantillonnés. Le descripteur final pourrait être calculé en construisant simplement un histogramme dans une région rectangulaire, et le faire pour toutes les régions rectangulaires possibles à l'intérieur de la fenêtre candidate, donnant lieu à un ensemble de descripteurs complets ou en créant un descripteur à haute dimension comme HOG [Mu 2008, Satpathy 2013]. Les variantes de descripteurs LBP proposées dans la littérature comprennent les catégories *Non-Redundant Local Binary Patterns* (NRLBP) [Mu 2008], *Discriminative Robust Local Binary Patterns* (DRLBP) [Satpathy 2013] et *Cell Structured Local Binary Patterns* (CellLBP) [Wang 2009].

Les descripteurs de couleur sont rarement utilisés en détection de personnes en raison de la forte variabilité d'apparence induite par les vêtements. Mais la couleur montre une similitude locale même sur les vêtements. Les descripteurs type *Color Self Similarity* (CSS), proposés par Walk *et al.* [Walk 2010], codent les similitudes dans différentes sous-régions. Les descripteurs sont d'abord calculés en subdivisant la fenêtre candidate en blocs non superposés de 8×8 pixels, puis en calculant un histogramme couleur $3 \times 3 \times 3$ dans chaque bloc (avec interpolation). Pour chaque bloc, les similitudes sont calculées comme l'intersection des histogrammes de bloc individuels. Dans [Walk 2010], toutes les valeurs des intersections d'histogramme sont concaténées pour définir un seul vecteur descripteur de haute dimension.

A ce stade, le catalogue focalise sur des descripteurs inférés sur une seule image. Il est également possible d'extraire des descripteurs en considérant des images successives temporellement. Deux descripteurs usuels exploitent ce concept : les descripteurs de mouvement avec des filtres rectangulaires [Viola 2005] et *Histogram of Flow* (HOF) de Dalal *et al.* [Dalal 2006b].

Il est pertinent de combiner les descripteurs pré-cités ; on parle alors de descripteurs hétérogènes. On capture ainsi des informations complémentaires donc a priori plus discriminantes comme l'attestent de nombreux travaux. Gerónimo *et al.* [Gerónimo 2007] l'ont montré avec des descripteurs type Haar et EOH ; Wang *et al.* [Wang 2009] avec HOG et LBP ; Wojek et Schiele [Wojek 2008] avec des descripteurs de type Haar, HOG et descripteurs de contexte de forme ; Walk *et*

al. [Walk 2010] avec une concaténation de HOF, HOF et CSS. Des conclusions similaires sont également émises par Schwartz *et al.* [Schwartz 2009b] et Hussain et Triggs [Hussain 2010] en utilisant – HOG, la fréquence des couleurs et des descripteurs de co-occurrence – et – des variantes HOG et LBP - respectivement. Avec une représentation holistique par image (pas d'informations temporelles), des performances excellentes sont obtenues en associant des descripteurs hétérogènes appelées *Integral Channel Features* [Dollár 2009b]. *Integral Channel Features* désignent une couche de plusieurs canaux d'image calculés en utilisant une transformation d'image unique par canal. Dans leur implémentation, Dollár *et al.* ont utilisé trois classes de descripteurs : image couleur, image gradient et histogrammes de gradient. Chaque composante des descripteurs s'inscrit dans un canal spécifique (3 canaux pour la couleur dans l'espace CIELUV, 1 canal pour la grandeur du gradient et 6 canaux d'orientation pour chaque orientation du gradient), puis des descripteurs spécifiques, telles que les sommes locales, les histogrammes et des descripteurs type Haar sont calculés par chaque canal de manière efficace en utilisant des images intégrales. Dans la littérature on trouve deux variantes remarquables des *Integral Channel Features* : *Aggregate Channel Features* (ACF) [Dollár 2014] et *Local Decorrelated Channel Features* (LDCF) [Nam 2014]. ACF utilise dix canaux - l'amplitude du gradient, HOG (6 canaux) et les canaux de couleur LUV. Chaque canal est agrégé sur des blocs pour créer des canaux de résolution inférieure. LDCF modifie ACF en appliquant des filtres de décorrélation par canal. Les filtres sont déterminés comme les vecteurs propres d'une matrice de covariance spécifique au canal, calculée à partir d'une grande collection d'images naturelles.

Considérant un ensemble de descripteurs hétérogènes, différentes façons peuvent être utilisées pour construire le descripteur composite finale. Quatre stratégies sont ainsi observées : (1) concaténation directe [Walk 2010, Wojek 2008] dans laquelle les différents descripteurs sont concaténés pour créer un vecteur descripteur à haute dimension ; (2) sélection des descripteurs discriminants après sélection par *boosting* [Schapire 2003] ; (3) disposition hiérarchique grossière *ad hoc* [Møgelmose 2012, Pan 2013] où une cascade est construite en utilisant des descripteurs peu coûteux aux étapes initiales et en utilisant des descripteurs complexes aux étapes ultérieures ; et (4) optimisation multi-objectif par rapport au temps de calcul et à la détection [Jourdheuil 2012, Wu 2008].

Les travaux actuels se focalisent désormais sur des techniques d'apprentissage profond. Par exemple, DeepPed [Tomè 2016] emploie une combinaison de LDCF comme algorithme de proposition de régions et d'un réseau de neurones convolutifs profond à réglage fin pour l'extraction de descripteurs. *Region-based Convolutional Neural Network* (RCNN) est basé sur les travaux de Girshick *et al.* [Girshick 2014]. Premièrement, ils génèrent des propositions de régions indépendantes de la catégorie. Ces propositions définissent l'ensemble des détections candidates disponibles et sont basées sur une recherche sélective. Ensuite, ils utilisent un réseau de neurones convolutifs qui extrait un vecteur descripteur de longueur fixe de chaque région (chaque région sélectionnée est transformée en une région rectangulaire prédéfinie et fixe). Ces paradigmes sont très prometteurs mais nécessitent l'utilisation d'archi-

Descripteur	Type	Entrée
Haar [Papageorgiou 2000, Lienhart 2002, Viola 2005]	Homogène	Image
EOH [Levi 2004]	Homogène	Image
HOG [Dalal 2005]	Homogène	Image
LBP [Ojala 1996]	Homogène	Image
CSS [Walk 2010]	Homogène	Image
HOF [Dalal 2006b]	Homogène	Séquence d'images
Integral Channel Features [Dollár 2009b]	Hétérogène	Image
ACF [Dollár 2014]	Hétérogène	Image
LDCF [Nam 2014]	Hétérogène	Image
DeepPed [Tomè 2016]	Apprentissage profond	Image
RCNN [Girshick 2014]	Apprentissage profond	Image

TABLE 1.1 – Synthèse des types de descripteurs.

tectures hardware spécifiques (notablement des cartes graphiques puissantes) ; ceci est limitant pour certaines applications (notamment celles embarquées).

Un résumé des différents types de descripteurs est synthétisé dans la Table 1.1. Les plus notables sont HOG [Dalal 2005] (par sa valeur historique), ACF [Dollár 2014], LDCF [Nam 2014] et les récents descripteurs basés sur un apprentissage profond (DeepPed [Tomè 2016] et RCNN [Girshick 2014]).

1.1.4 Classification

L'étape de classification conduit à l'étiquetage de chaque fenêtre candidate générée, et décrite par les descripteurs sous-jacents, comme personne ou non personne. Le processus génère, pour chaque fenêtre, une étiquette binaire, avec possiblement une vraisemblance quantifiant sa confiance dans l'étiquetage. Ces classificateurs sont entraînés par apprentissage supervisé, c'est-à-dire en exploitant des échantillons (fenêtres) préalablement annotés en positifs (personne) et négatifs (non personne). Les classificateurs discriminants les plus fréquemment utilisés pour la détection des personnes sont des variantes des machines à vecteurs de support (SVM) et des classificateurs boostés. Citons aussi l'analyse discriminante linéaire de Fisher (LDA), par exemple, [Paisitkriangkrai 2008] et réseaux de neurones artificiels, par exemple, [Szarvas 2005, Zhao 1999, Zhao 2000]. Citons, plus récemment, les forêts d'arbres décisionnels [Tang 2012].

Les classificateurs boostés, également appelés classificateurs d'ensemble, construisent un classificateur fort en combinant des classificateurs faibles où chaque classificateur faible consécutif se concentre sur des échantillons précédemment mal classés. La stratégie est efficace car elle sélectionne les descripteurs discriminants automatiquement. Différentes variantes sont proposées dans la littérature ; citons : *AdaBoost* discret, par exemple, [Viola 2004], *Adaboost* réel, par exemple, [Gerónimo 2010a], *Logit-Boost*, par exemple, [Tuzel 2008]. Un intérêt du *boosting* est donc de sélectionner non aléatoirement les classificateurs faibles. Dans la littérature, un classificateur boosté est presque toujours intégré dans une architecture en cascade attentionnelle, également appelée cascade de rejet, qui a la forme d'un arbre déséquilibré [Viola 2004]. Chaque nœud de la cascade est entraîné avec un sous-ensemble

Classificateur	Type
SVM	Classificateur
LDA	Classificateur
Réseau de neurones	Classificateur
Forêt d'arbres décisionnels	Cascade de classificateurs
Hard-cascade	Cascade de classificateurs
Soft-cascade	Cascade de classificateurs

TABLE 1.2 – Synthèse des types de classificateurs.

des échantillons d'apprentissage afin de lui associer les meilleurs classifieurs forts. Les premiers (resp. derniers) nœuds de la cascade filtrent les échantillons les plus simples (resp. difficiles) à discriminer. Ainsi, seuls les échantillons difficiles à classer seront évaluées par tous les nœuds. La cascade boostée est intéressante lorsque la vitesse de détection est déterminante, par exemple [Viola 2004, Dollár 2009a, Felzenszwalb 2010].

Nous citons les *hard-cascades*, où chaque étape en cascade agrège le score pondéré des classificateurs faibles associés et un seuil permet d'étiqueter chaque échantillon comme positif ou négatif. Nous citons aussi les *soft-cascades* : ici chaque étape réalise une évaluation pondérée de classificateurs faibles [Zhang 2007, Bourdev 2005] avec un seuil de classification. Récemment, diverses variantes de *soft-cascade* ont abouti à de meilleures performances de détection [Zhang 2016b, Zhang 2015, Dollár 2014, Dollár 2012]. Elles sont utilisées dans différentes applications [Teutsch 2015, Varga 2016, Han 2017]. Les seuils de la *soft-cascade* sont appris après apprentissage complet des classificateurs faibles dans une sorte de phase d'étalonnage.

Une synthèse des types de classificateurs utilisés pour la détection de personnes est présentée dans la Table 1.2.

1.2 Focus sur quelques détecteurs clés

Comme évoqué, un détecteur est composé par association d'un modèle de personne, un descripteur, un classificateur et souvent un processus de suppression de non-maxima.

Cette section focalise sur des détecteurs visuels usuels et/ou performants ; Ils permettront d'étalonner nos propres détecteurs. Nous focalisons sur les six détecteurs décrits ci-après car ils font consensus dans la littérature et couvrent un large spectre eu égard à notre catégorisation : HOG-SVM [Dalal 2005], DPM [Felzenszwalb 2010], ACF [Dollár 2014], LDCF [Nam 2014], DeepPed [Tomè 2016], et RCNN [Girshick 2014]. Leurs caractéristiques sont synthétisées dans la Table 1.3.

Détecteur	Descripteur	Modèle	Classificateur	NMS
HOG-SVM [Dalal 2005]	HOG	Holistique	SVM linéaire	PM
DPM [Felzenszwalb 2010]	HOG	Basée sur parties	SVM latente	PM
ACF [Dollár 2014]	ACF	Holistique	*Soft-cascade*	PM
LDCF [Nam 2014]	LDCF	Holistique	*Soft-cascade*	PM
DeepPed [Tomè 2016]	Apprentissage profond	Holistique	SVM linéaire	PM
RCNN [Girshick 2014]	Apprentissage profond	Holistique	SVM linéaire	PM

TABLE 1.3 – Focus sur six détecteurs de personnes.

1.2.1 Histogram of Oriented Gradient (HOG-SVM)

Ce détecteur, initié par Dalal et Triggs [Dalal 2005], est un détecteur historique et constitue donc un excellent *benchmark*. Il repose sur des descripteurs type HOG et un classifieur SVM linéaire. Le modèle appris est basé sur une abstraction holistique (corps entier). Les sorties de détection sont filtrées par une technique de suppression des non-maxima (NMS) par *Pairwise Max* (PM) qui supprime les boites englobantes les moins probables pour toute paire de détections en chevauchement.

1.2.2 *Deformable Part-based Model* (DPM)

Le détecteur DPM [Felzenszwalb 2010], aussi connu sous le nom de LatSvm-V2 et LatSvm-L2, repose sur un modèle par parties. Il utilise un mélange de modèles basés sur des parties déformables et une version modifiée des descripteurs HOG. Le modèle comprend un filtre racine (celui qui caractérise le corps entier) et plusieurs filtres de parties ; son score sur une fenêtre candidate est déterminé comme le score du filtre racine plus la somme des scores de chaque filtre de partie, en prenant le maximum sur les emplacements des parties, moins un coût de déformation qui pénalise l'écart des positions image des parties idéales par rapport au filtre racine. Il est appris en utilisant des données partiellement étiquetées avec un SVM latent. La boîte englobante finale de la détection est déterminée avec une fonction de cartographie apprise qui englobe les positions des parties détectées. Enfin, le filtrage des boites englobantes repose sur une technique NMS basée sur PM.

1.2.3 *Aggregate Channel Features* (ACF)

Ce détecteur, à faible coût CPU, repose sur le concept de *channel features* ; il surclasse nombre de détecteurs et ceci sur nombre de bases publiques [Dollár 2014]. Il est basé sur des descripteurs type ACF, un classificateur *soft-cascade*, et une représentation de personne holistique. Le classificateur est appris en utilisant *Ada-Boost* et des arbres de décision de profondeur deux sur ces descripteurs. Enfin, il filtre les détections via une technique NMS basée sur PM.

1.2.4 *Local Decorrelated Channel Features* (LDCF)

Le détecteur LDCF [Nam 2014] s'appuie aussi sur des descripteurs *channel features*. Mais, au contraire, il applique une étape préalable de décorrélation des descripteurs ; les arbres de décision sont alors construits sur ces bases orthogonales.

1.2.5 Réseau de neurones convolutifs profonds ou DeepPed

Le détecteur DeepPed [Tomè 2016] repose sur un apprentissage profond i.e. les couches basses du réseau de neurones convolutif extraient les caractéristiques/descripteurs discriminants (ici aucun a priori sur la classe de descripteurs) et la classification finale s'effectue par SVM. Le réseau et le réglage des (nombreux) paramètres libres associés sont explicités dans [Tomè 2016].

1.2.6 *Region-based Convolutional Neural Networks* (RCNN)

Le détecteur de personnes RCNN est basé sur les travaux de Girshick *et al.* [Girshick 2014]. Ce système se compose de trois modules. Le premier génère des propositions de régions indépendantes de la catégorie. Ces propositions définissent l'ensemble des détections de candidats disponibles pour le détecteur et sont basées sur une recherche sélective. Le deuxième module est un grand réseau de neurones convolutif qui extrait un vecteur descripteur de longueur fixe de chaque région. Le troisième module est un SVM linéaire spécifique qui classe chaque vecteur descripteur en tant que personne ou non.

1.2.7 *Faster RCNN*

Le détecteur faster-RCNN de Ren *et al.* [Ren 2015] remplace le premier module du RCNN pour un module *Region Proposal Network* (RPN). qui prédit simultanément les boîtes englobantes des cibles et les scores associés pour chaque position de boîte. Le RPN est entraîné pour générer des propositions de régions pertinentes, et exploitées ensuite par RCNN pour la classification. Les modules RPN et RCNN peuvent partager des descripteurs convolutifs.

1.3 Description des bases de données

1.3.1 Base de données publiques INRIA

La base de données de personnes INRIA, présentée par Dalal et Triggs [Dalal 2005], est largement exploitée dans la littérature afin d'étalonner les détecteurs de personnes. Elle se divise en deux formats : (1) des images brutes (avec et sans cibles) et leurs annotations correspondantes (boîtes englobantes sur l'image entière), et (2) des échantillons (extraction des boîtes englobantes de l'image) positives et négatives de personnes. Chaque sous ensemble est subdivisé en données d'apprentissage et de test.

La base d'apprentissage du premier format inclut 614 images positives (avec présence de personnes) et 1218 images négatives (sans personne). La base de test comprend 288 images positives et 453 images négatives. Quelques exemples d'images sont illustrés figure 1.7.

(a) Image d'apprentissage avec personnes.

(b) Image d'apprentissage sans personnes.

(c) Image de test avec personnes.

(d) Image de test sans personnes.

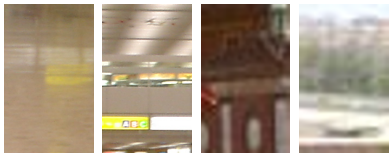
FIGURE 1.7 – Exemples d'images de la base INRIA.

La base d'apprentissage du deuxième format comprend 2416 échantillons positifs (versions originales et en miroir) et 1218 images négatives. Les échantillons ont une résolution de 160×96. Mais, la taille réelle de la boîte englobante des personnes est de 128×64. Le surplus sert à minimiser l'effet de bordure. La base de tests contient 1132 échantillons positifs et 453 images négatives. Les échantillons ont une résolution de 134×70. Quelques échantillons de cette base sont illustrés figure 1.8. Les échantillons positifs dans les bases d'apprentissage et de test sont obtenus à partir d'images de la vie courante et annotées manuellement.

1.3.2 Base de données publiques CALTECH

Cette base [Dollár 2012] englobe environ 250000 images de taille 640×480 acquises à partir d'une conduite de véhicule dans un trafic urbain normal. La base de données est annotée avec 350000 boîtes englobantes incluant 2300 piétons. La base de données est structurée en 11 bases différentes (S0-S10), six bases d'apprentissage (S0-S5) et cinq bases de tests (S6-S10). La figure 1.9 illustre quelques exemples.

(a) Échantillons d'apprentissage positifs.



(b) Échantillons d'apprentissage négatifs.

(c) Echantillons de test positifs.

(d) Echantillons de test négatifs.

FIGURE 1.8 – Exemples d'échantillons de la base de données publique INRIA.

(a) Exemple d'image d'apprentissage.

(b) Exemple d'image de test.

FIGURE 1.9 – Exemples d'images de la base de données publique CALTECH.

Les boîtes englobantes annotées contiennent également des caractéristiques d'aspect piéton qui permettent de catégoriser les évaluations : à faible distance (piéton de plus de 80 pixels de hauteur), à moyenne distance (piétons de 30 à 80 pixels de hauteur), à grande distance (de moins de 30 pixels de haut), sans occultation (piétons non occultés de plus de 50 pixels), occultation partielle (piétons avec occultation partielle 1-35% de superficie occultée), forte occultation (piétons avec occultation d'une surface de 35 à 80%) et raisonnable (50 pixels ou plus grands piétons sans/avec occultation partielle).

Métrique	Type d'évaluation	Sujet d'évaluation
ROC	Par fenêtre	Classificateur
DET	Par fenêtre	Classificateur
Précision et rappel	Par fenêtre	Classificateur
FPPI	Par image	Détecteur

TABLE 1.4 – Synthèse des métriques d'évaluation.

1.4 Métriques d'évaluation

Plusieurs métriques sont considérées pour évaluer/comparer les détecteurs. On dissocie ici les évaluations : (1) par fenêtre et (2) par image. Les évaluations par fenêtre (1) déterminent les performances sur des échantillons (extraites de l'image entière) positifs et négatifs. Cette approche dissocie ainsi les performances du classificateur de celles du système global de détection ; et permet de caractériser les performances du classificateur seul [Dollár 2012]. Au contraire, les évaluations par image (2) considèrent les images brutes en entrée et s'effectuent sur les fenêtres candidates générées automatiquement. La détection est donc multi-échelle, supprime les non-maxima, etc. Elle quantifie les performances du détecteur complet.

De par leurs intérêts respectifs, et à l'instar de nombreux travaux, nos travaux justifient des évaluations par fenêtre et par image. Les évaluations par fenêtre considèrent classiquement trois métriques : (i) la courbe de fonction d'efficacité du récepteur (ROC) [Spackman 1989], (ii) la courbe de *Detection error tradeoff* (DET) [Martin 1997], et (iii) la courbe de précision et rappel [Manning 2008, Hoiem 2012]. Leur synthèse est présentée dans la Table 1.4 puis détaillée ci-après.

1.4.1 Évaluations par fenêtre : courbe ROC

Le classificateur classe en positif tous les échantillons dont le score est supérieur à un seuil fixé S. La courbe ROC représente ses performances en fonction des paramètres libres, en premier lieu le seuil de classification S [Spackman 1989]. Formellement, soient P le nombre total d'échantillons positifs et N le nombre total d'échantillons négatifs. On note :

TP le nombre d'échantillons correctement classés en positifs (vrais positifs),

TN le nombre d'échantillons correctement classés en négatifs (vrai négatifs),

FP le nombre d'échantillons classés incorrectement en positifs (faux positifs),

FN le nombre d'échantillons classés incorrectement en négatifs (faux négatifs).

On définit alors les taux de vrai positifs (1.1), taux de vrais négatifs (1.2), taux de faux positifs (1.3) et taux de faux négatifs (1.4) c'est-à-dire :

$$TPR = \frac{TP}{P} \tag{1.1}$$

$$TNR = \frac{TN}{N} \tag{1.2}$$

$$FPR = \frac{FP}{N} \tag{1.3}$$

$$FNR = \frac{FN}{P} \tag{1.4}$$

avec les propriétés suivantes :

$$TP + FN = P;\ TN + FP = N;\ TPR + FNR = 1;\ TNR + FPR = 1$$

La courbe ROC est obtenue par variation des paramètres libres du détecteur, notamment le seuil S de classification. Elle permet de retrouver les points de fonctionnement optimaux (en terme de performances) du détecteur. L'aire sous la courbe caractérise alors la performance du classificateur. On note quatre variantes de courbe ROC :

tntp : taux de vrai positifs vs. taux de vrai négatifs (variante classique),

tptn : taux de vrai négatifs vs. taux de vrai positifs,

fptp : taux de vrai positifs vs. taux de faux positifs,

fpfn : taux de faux positifs vs. taux de faux négatifs.

1.4.2 Évaluations par fenêtre : *Detection error tradeoff* (DET)

Ces évaluations considèrent une courbe *Detection error tradeoff* (DET) avec taux de défaut (MR) par rapport aux fausses alarmes (FA) sur une échelle log-log [Martin 1997]. La courbe est générée en exécutant le classificateur à différents points de fonctionnement (habituellement obtenus en échantillonnant le seuil du classificateur) et en calculant le taux de défaut, le taux de faux négatif (1.4), et les fausses alarmes (faux positif, équation (1.3)), à chaque point ce qui est similaire à la variante fpfn de la courbe ROC. Le taux de défaut moyen est calculé comme l'aire sous la courbe. Un détecteur performant doit donc exhiber un faible taux de défaut.

A l'instar de la courbe ROC, cette métrique est utilisée pour comparer les performances de détecteurs visuels.

1.4.3 Évaluations par fenêtre : précision et rappel

La courbe de précision et rappel est aussi une métrique usuelle [Manning 2008, Hoiem 2012]. La courbe est obtenue par variation des paramètres libres du détecteur. La précision est définie comme le nombre d'échantillons positifs classés correctement par rapport au nombre d'échantillons classés comme positifs et est calculée via l'équation (1.5). Le rappel ou la sensibilité est définie comme le nombre d'échantillons positifs classés correctement par rapport au nombre total d'échantillons positifs, c'est-à-dire est le taux de vrai positifs calculé comme par l'équation (1.1). On considère une courbe de précision par rapport au rappel, et sa précision moyenne associée correspondant à l'aire sous la courbe. Une meilleure précision moyenne implique un meilleur classificateur.

$$Précision = \frac{TP}{TP + FP} \tag{1.5}$$

1.4.4 Évaluations par image

Pour rappel, ces évaluations sont utiles pour évaluer un détecteur complet. On considère ici des images brutes annotées avec des boîtes englobantes de personnes ; on compare alors les occurrences de TP, FP et FN vérifiées avec la vérité terrain c'est-à-dire les boîtes englobantes annotées. On considère ici la métrique taux de défaut par rapport au taux de faux positifs par image (FPPI). Le FPPI est fondamentalement les occurrences de faux positifs (FP) en moyenne sur le nombre d'images testées. Une détection est considérée comme un vrai positif uniquement si l'intersection entre la boîte englobante inférée par le détecteur et la boîte englobante annotée est supérieur à un seuil fixé, c'est-à-dire si l'intersection des deux boites sur l'union des deux boites est supérieure à 50% [Dollár 2012]. Nous privilégions ce mécanisme à l'instar de la littérature. Pour résumer les performances d'un détecteur, on utilise le taux de défaut moyen logarithmique.

Une autre métrique à considérer est le taux de défaut moyen logarithmique par rapport au nombre d'images par seconde (FPS). Elle schématise les performances du détecteur vs. son temps de calcul par image.

1.5 Expérimentations préliminaires

1.5.1 Implémentation

Il existe différents outils *open source* disponibles pour l'implémentation de détecteurs usuels et pour leurs évaluations. Nous en privilégions deux : la *Piotr's Computer Vision Matlab Toolbox* [Dollár 2016] et le *Caltech Pedestrian Detection Benchmark* [Dollár 2012].

La *Piotr's Computer Vision Matlab Toolbox* (PMT) est une bibliothèque Matlab proposée par Dollár [Dollár 2016]. Elle permet l'apprentissage et l'application des détecteurs ACF [Dollár 2014] et LDCF [Nam 2014] et inclut les détecteurs appris sur les bases INRIA [Dalal 2005] et Caltech [Dollár 2012].

Le *Caltech Pedestrian Detection Benchmark* est un outil d'évaluation et de comparaison de détecteurs proposé par Dollár *et al.* [Dollár 2012]. L'outil est indépendant de l'implémentation des détecteurs. Il reçoit en entrée la vérité terrain de la base de données et les boîtes englobantes inférées par les détecteurs. Il donne en sortie la courbe taux de défaut vs. FPPI. Il englobe les performances de 60 détecteurs sur 5 bases de données dont INRIA [Dalal 2005] et Caltech [Dollár 2012].

1.5.2 Évaluations et discussion

1.5.2.1 Base de données INRIA

Le *Caltech Pedestrian Detection Benchmark* [Dollár 2012] offre des évaluations de 34 détecteurs sur la base de données INRIA [Dalal 2005]. Ces évaluations sont synthétisées sur la Figure 2.15. On note que le détecteur RPN+BF [Zhang 2016a], dérivé du *Faster* RCNN [Ren 2015], offre les meilleures performances, avec un taux de défaut moyen logarithmique de $6,88\%$. LDCF [Nam 2014] offre un taux de défaut moyen logarithmique de $13,79\%$, ACF [Dollár 2014] de $17,28\%$, LatSvm-V2 [Felzenszwalb 2010] de $19,96\%$ et HOG [Dalal 2005] de $45,98\%$.

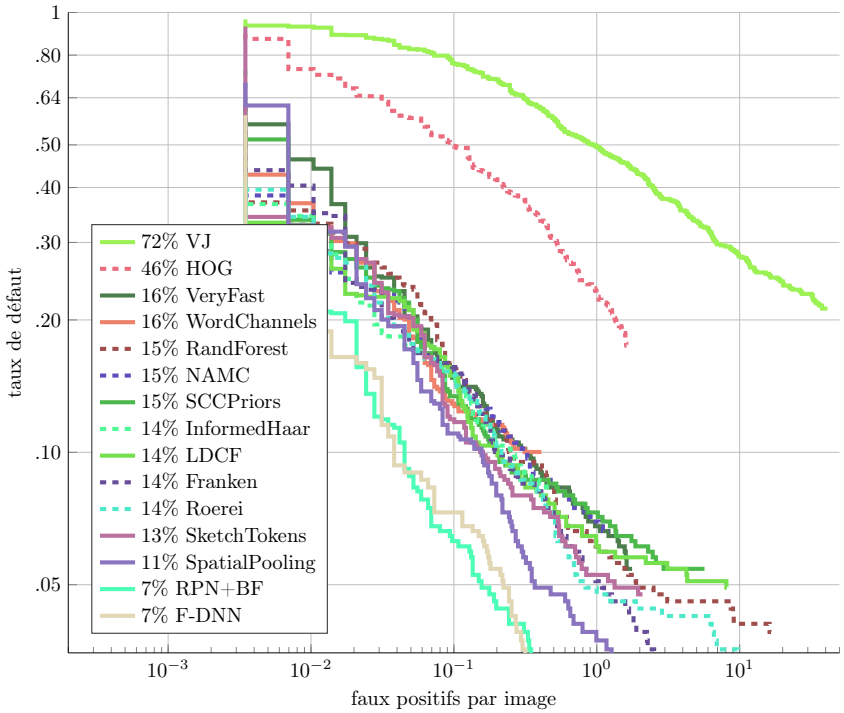


FIGURE 1.10 – Évaluations taux de défaut vs. FPPI sur la base INRIA avec *Caltech Pedestrian Detection Benchmark* [Dollár 2012].

Par rapport à l'indicateur FPS, nous citons les évaluations de 6 détecteurs réalisées par Cao *et al.* [Cao 2016a] et illustrées sur la Figure 1.11. On note que ACF [Dollár 2014] présente un temps de calcul de $31,90$ images par second (FPS), LDCF [Nam 2014] de $4,7$ FPS et LatSvm-L2 [Felzenszwalb 2010] de $0,16$ FPS. Par une analyse de Pareto, nous observons que les détecteurs avec le meilleur rapport performance-temps de calcul sont : Crosstalk, ACF, NNNF et SpatialPooling.

1.5.2.2 Base de données CALTECH

Le *Caltech Pedestrian Detection Benchmark* propose les évaluations de 68 détecteurs sur la base CALTECH [Dollár 2012]. Elles sont illustrées figure 1.12

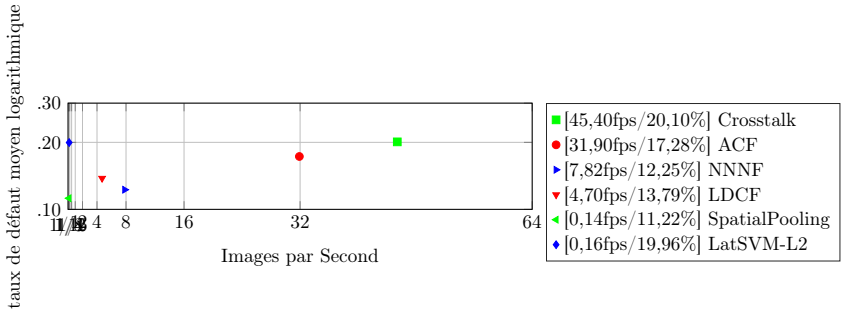


FIGURE 1.11 – Évaluations FPS sur la base INRIA par Cao *et al.* [Cao 2016a].

et détaillées figure 1.13. On observe que le RPN+BF [Zhang 2016a] offre un taux de défaut moyen logarithmique de $9,58\%$, LDCF [Nam 2014] de $24,8\%$, ACF [Dollár 2014] de $44,21\%$, LatSvm-V2 [Felzenszwalb 2010] de $63,26\%$ et HOG [Dalal 2005] de $68,46\%$.

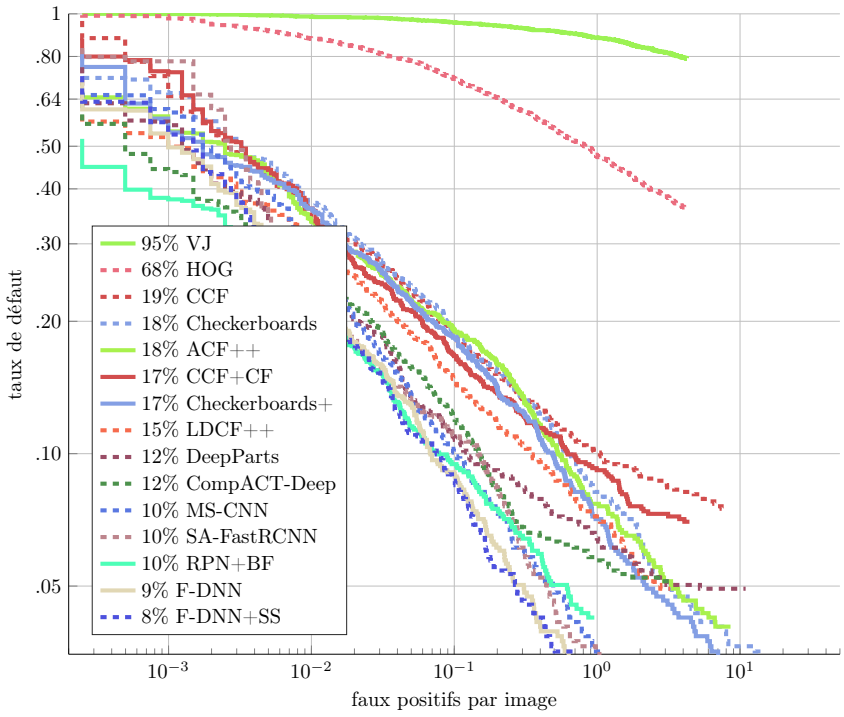
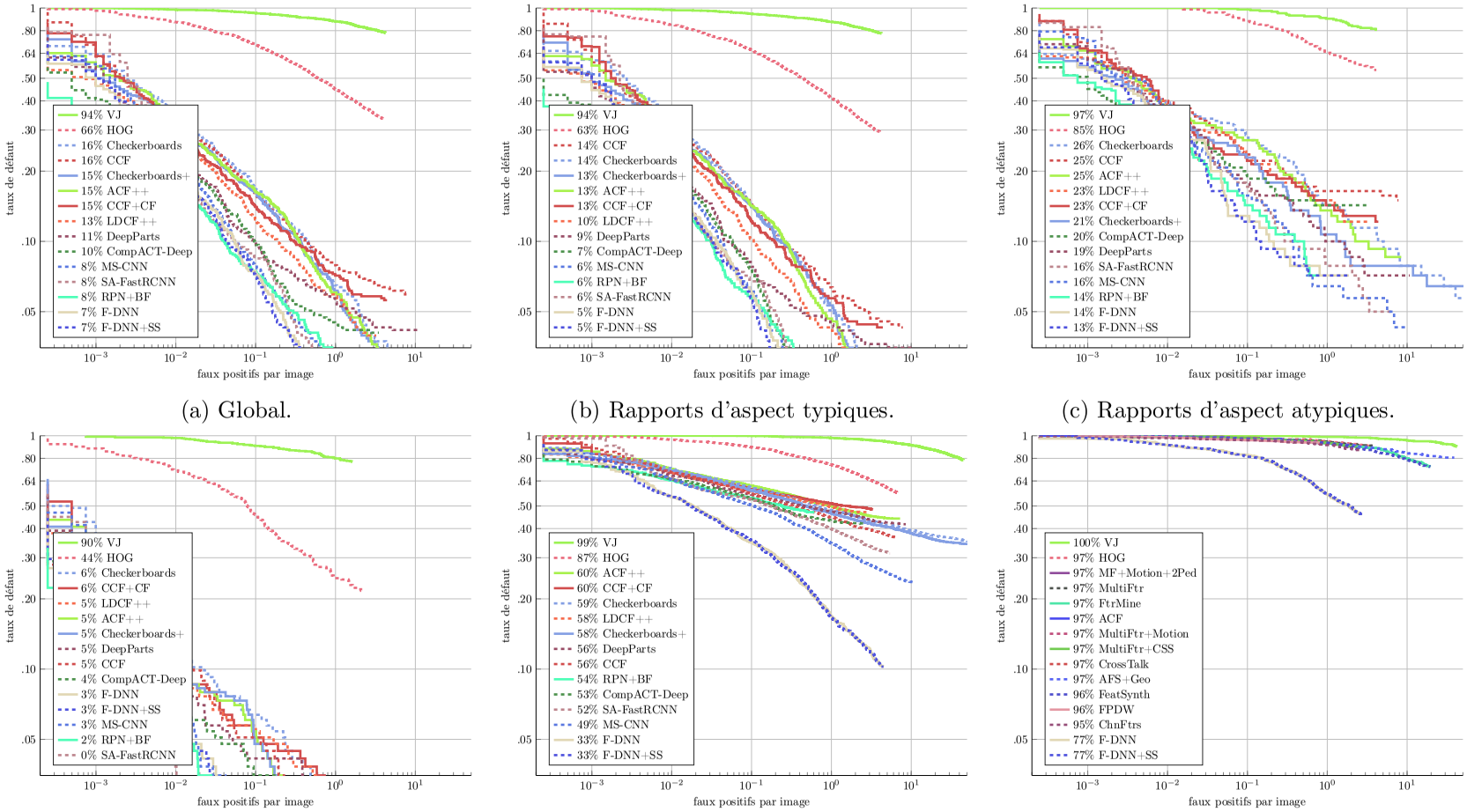


FIGURE 1.12 – Synthèse des évaluations taux de défaut vs. FPPI sur la base CALTECH avec *Caltech Pedestrian Detection Benchmark* [Dollár 2012].

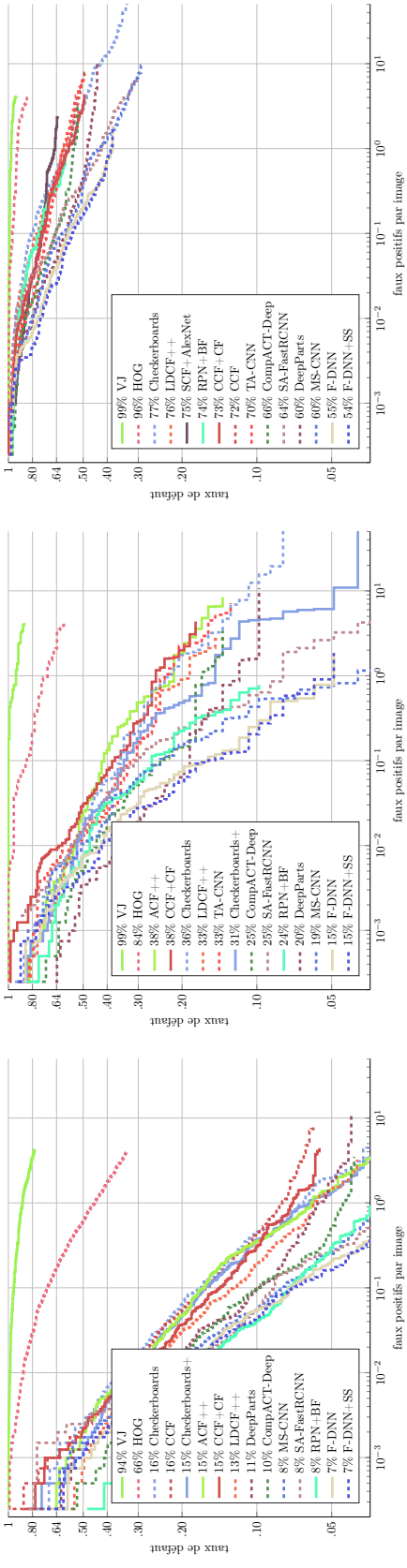
Citons aussi les évaluations de 10 détecteurs réalisées par Cao *et al.* [Cao 2016a] et illustrées sur la Figure 1.14. On note que ACF [Dollár 2014] offre un temps de calcul de $9,49$ images par second (FPS), LDCF [Nam 2014] de $3,62$ FPS et LatSvm-L2 [Felzenszwalb 2010] de $0,16$ FPS. Par une analyse de Pareto, on observe que les



(a) Global. (b) Rapports d'aspect typiques. (c) Rapports d'aspect atypiques.

(d) Faible distance. (e) Moyenne distance. (f) Grande distance.

FIGURE 1.13 – Évaluations détaillées taux de défaut vs. FPPI sur la base CALTECH avec *Caltech Pedestrian Detection Benchmark* [Dollár 2012].



(g) Sans occultation. (h) Occultation partielle. (i) Forte occultation.

FIGURE 1.13 – Évaluations détaillées taux de défaut vs. FPPI sur la base CALTECH avec *Caltech Pedestrian Detection Benchmark* [Dollár 2012] (Cont.).

Détecteur	Hardware	temps/image (s)	MR (%)
LDCF [Nam 2014]	CPU	0,6	24,8
CCF [Yang 2015]	Titan Z GPU	13	17,3
CompACT-Deep [Cai 2015]	Tesla K40 GPU	0,5	11,7
RPN+BF [Zhang 2016a]	Tesla K40 GPU	0,5	9,6

TABLE 1.5 – Comparaison du temps d'exécution sur la base CALTECH [Zhang 2016a].

détecteurs avec le meilleur rapport entre taux de classification et temps de calcul sont : Crosstalk, ACF, LDCF, NNNF-L2 et NNNF-L4.

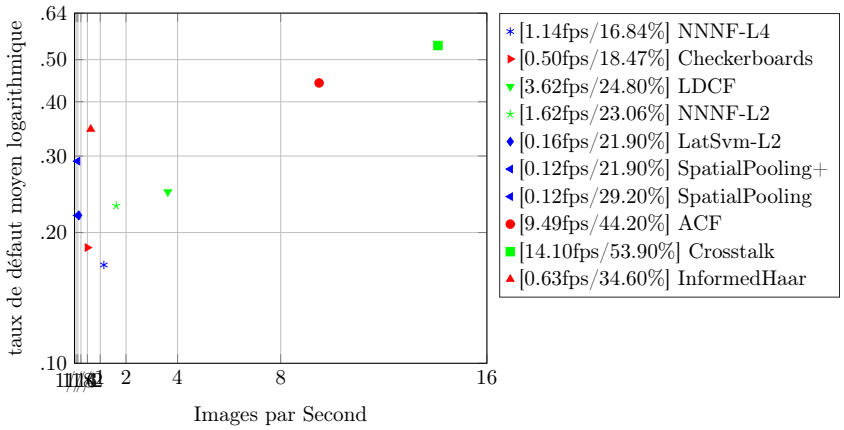


FIGURE 1.14 – Évaluations FPS sur la base CALTECH [Cao 2016a].

Zhang *et al.* [Zhang 2016a] ont fait une comparaison du temps d'exécution de 4 détecteurs sur la base CALTECH [Dollár 2012]. Cette comparaison est présentée dans la Table 1.5. On note que les détecteurs du type apprentissage profond ont besoin de l'utilisation d'un GPU. Néanmoins, ces détecteurs ont de temps de calcul proche du détecteur LDCF [Nam 2014] qui n'utilise pas du GPU mais du CPU.

1.6 Conclusions

Ce chapitre énumère les challenges inhérents à la détection visuelle de cibles type personne, puis catégorise les approches existantes et focalise enfin sur quelques détecteurs clés eu égard à leurs performances, leurs utilisations récurrentes en tant que *benchmark* dans la littérature, etc. Nous présentons ensuite deux bases de données publiques (INRIA, CALTECH), les métriques, et les outils couramment usités dans la communauté Vision pour comparer les performances de ces détecteurs. Les performances de quelques détecteurs clés sur ces deux *datasets* sont alors présentées et discutées.

Parmi les détecteurs présentés, les *soft-cascades* avec des descripteurs du type *Integral Channel Features* (particulièrement ACF) sont une option intéressante,

même face aux dernières approches d'apprentissage profond. L'intérêt de ces détecteurs est lié à sa performance de détection par rapport au coût de traitement, et sa capacité de fonctionner sur des architectures *hardware* classiques (CPU).

Toutes ces considérations sont un préambule nécessaire à la description, les évaluations, et études comparatives, de nos détecteurs visuels qui sont présentés au chapitre suivant.

Chapitre 2

Détecteur *Soft-Cascade* avec considérations de temps de calcul

Sommaire

Comme vu précédemment, la détection visuelle de personnes trouve des applications dans de nombreux domaines, dont l'interaction homme-robot, l'indexation d'images, la vidéo surveillance de lieux publics et, notamment, la ré-identification de personnes qui est traitée en seconde partie du mémoire.

Dans le chapitre précédent, les diverses briques d'un détecteur générique et les classes de détecteurs sont discutées. Pour rappel, pour tout détecteur visuel, deux critères sont à considérer simultanément : (1) les performances de détection et (2) le temps de calcul ou réponse. Les travaux existants proposent certes des contributions significatives mais visent souvent à optimiser l'un ou l'autre de ces facteurs, mais rarement les deux simultanément.

Dans ce chapitre, nous présentons une approche de détection par *soft-cascade* qui intègre les deux critères. La section 2.1 décrit le détecteur étudié et le compromis performances de détection par rapport au temps de réponse. Nous proposons une formalisation mathématique originale de ce problème et plusieurs méthodes de résolution en section 2.2. Les expérimentations, les évaluations et discussions associées sont décrites en section 2.3. La section 2.4 présente les conclusions et perspectives de ces travaux.

2.1 Minimisation du temps de réponse d'un détecteur type *Soft-Cascade*

Dans le chapitre précédent, le concept de détecteur *soft-cascade* et son intérêt ont été introduits. Pour rappel, un détecteur *soft-cascade* utilise une cascade de classificateurs faibles pour étiqueter les fenêtres candidates. Le synoptique de ce type de détecteur est illustré sur la figure 2.1. Dans une *soft-cascade*, l'objectif est de rejeter dès que possible les fenêtres négatives. Soit un classificateur soft-cascade de la forme de l'équation (2.1). x correspond à un échantillon de test qui peut être positif ($y = 1$) ou négatif ($y = -1$), α_l est le poids donné par le *boosting* au classificateur faible de niveau l, $h_l(x)$ est la réponse du l^e classificateur faible qui peut être une valeur dans le domaine $[-1, 1]$ pour un *AdaBoost* réel ou une des valeurs $\{-1, 1\}$ pour un *AdaBoost* discret ; $\text{sign}(\mathcal{H}(x) - \theta_L)$ détermine l'étiquette finale.

$$\mathcal{H}(x) = \sum_{l=1}^{L} \alpha_l h_l(x) \tag{2.1}$$

Le score cumulatif intermédiaire d'un échantillon au l^e classificateur faible est défini par $S_l = \sum_{u=1}^{l} \alpha_u h_u(x)$. Une *soft-cascade* utilise les seuils de classification θ_l qui rejettent un échantillon au niveau l chaque fois que $S_l < \theta_l$ (sinon il passe au classificateur suivant). Les méthodes les plus significatives en termes de stratégies de caractérisation de ces seuils de classification sont le *Direct Backward Pruning* (DBP) de Zhang et Viola [Zhang 2007], le *WaldBoost* de Sochman et Matas [Sochman 2005], la "*soft-cascade*" de Bourdev et Brandt [Bourdev 2005], et la chaîne de boosting de Xiao *et al.* [Xiao 2003].

***Direct Backward Pruning* (DBP)** Zhang et Viola [Zhang 2007] proposent une stratégie appelée *Direct Backward Pruning* (DBP) qui utilise le score cumulatif intermédiaire d'un échantillon x_n pour chaque classificateur faible l^e qui est égal à $S_{n,l} = \sum_{u=1}^{l} \alpha_u h_{n,u}$, où α_l est le poids positif appris par *boosting* associé au classificateur faible et $h_{n,l}$ est sa réponse connue. À chaque niveau l, un seuil de classification θ_l est défini, selon (2.2), de sorte que tout échantillon ayant son score $S_{n,l} < \theta_l$ est rejeté. Comme on peut l'observer, la définition de θ_l dépend elle-même de θ_L, qui est le seuil de classification final désiré et qui permet de fournir le nombre minimum d'échantillons vrais positifs.

$$\theta_l = \min_{\{n | S_{n,L} > \theta_L, y_n = 1\}} S_{n,l} \tag{2.2}$$

WaldBoost Sochman et Matas [Sochman 2005] proposent une stratégie appelée *WaldBoost* qui utilise le score cumulatif intermédiaire d'un échantillon x_n pour chaque classificateur faible l^e. Ce score est égal à $S_{n,l} = \sum_{u=1}^{l} \alpha_u h_{n,u}$, où α_l est le poids positif associé au classificateur faible et $h_{n,l}$ est sa réponse connue. À chaque niveau l, deux seuils de classification θ_l^- et θ_l^+ sont définis, en utilisant le *Sequential*

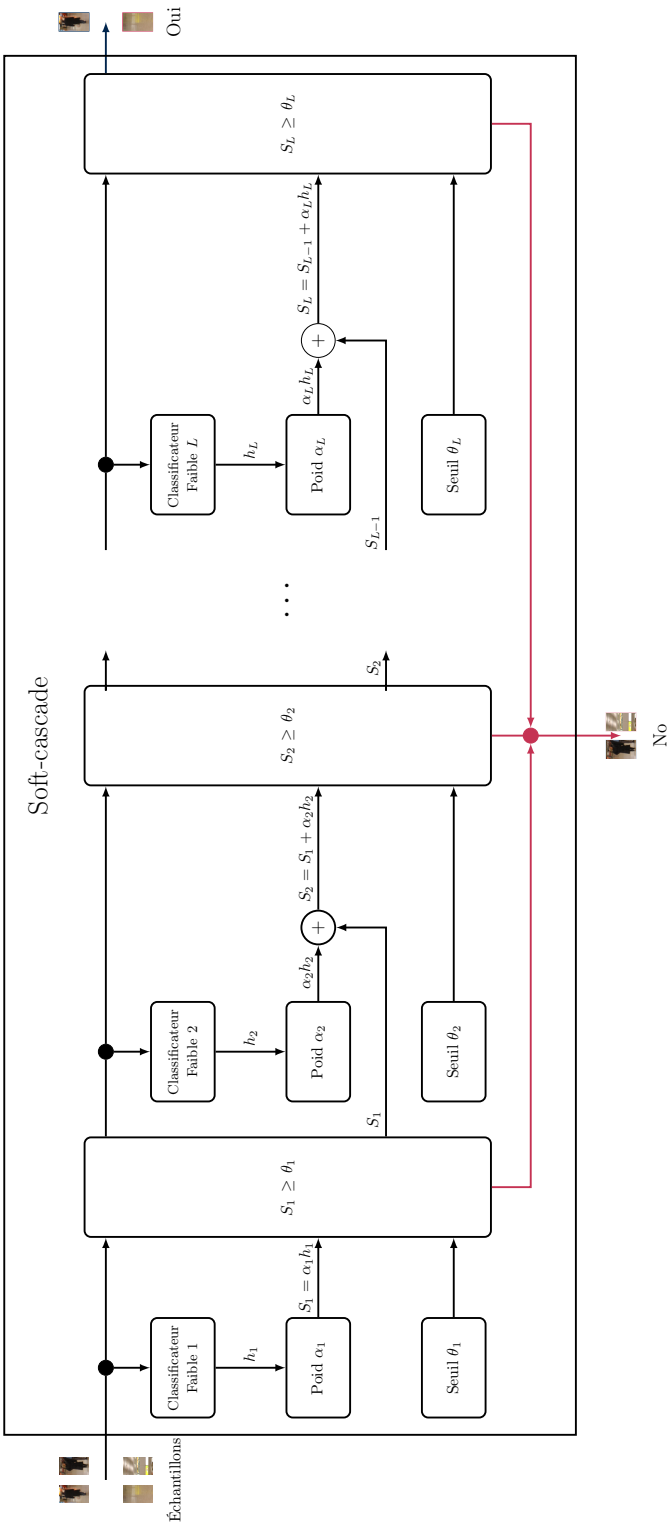


FIGURE 2.1 – Détecteur type *soft-cascade*. Les poids α_l sont calculés par *boosting*. Ils existent plusieurs techniques d'apprentissage pour obtenir les seuils θ_l

probability ratio test (SPRT) de Wald [Wald 1945], de sorte que tout échantillon est classé selon (2.3). La définition de ces seuils dépend de θ_L^+ qui comme précédemment fournis le nombre minimum d'échantillons vrais positifs, mais aussi de θ_L^- du nombre maximum d'échantillons faux positifs FPR.

$$S_{n,l}^* = \begin{cases} 1 & \text{si } S_{n,l} \geq \theta_l^+ \\ -1 & \text{si } S_{n,l} \leq \theta_l^- \\ 0 & \text{si } \theta_l^- < S_{n,l} < \theta_l^+. \end{cases} \tag{2.3}$$

Soft-cascade Bourdev et Brandt [Bourdev 2005] calculent un vecteur de distribution de rejet v composé de la fraction minimale de détections v_l qui peuvent être mal classées au classificateur faible l. Le vecteur v représente un point de fonctionnement du détecteur dans l'espace TPR-FPR-Temps de calcul. La somme de ses éléments $\sum_{l=1}^{L} v_l$ est égal à $FNR = 1 - TPR$. La procédure utilise le score cumulatif intermédiaire d'un échantillon x_n pour chaque classificateur faible l qui est égal à $S_{n,l} = \sum_{u=1}^{l} \alpha_u h_{n,u}$, où α_l est le poids positif associé au classificateur faible et $h_{n,l} \in [0, 1]$ est sa réponse connue. À chaque niveau l, un seuil de classification θ_l est défini, selon (2.4). Comme auparavant, tout échantillon ayant son score $S_{n,l} < \theta_l$ au niveau l sera rejeté.

$$\theta_l = \max_{\{n | \sum_{i=1}^{N} y_i \mathbf{1}_{[0,\infty)}(S_{i,l} - S_{n,l}) \leq \left(\sum_{t=1}^{l} v_t\right)\left(\sum_{i=1}^{N} y_i\right)\}} S_{n,l} \tag{2.4}$$

Chaîne de boosting Xiao *et al.* [Xiao 2003] proposent une stratégie qui utilise le score cumulatif intermédiaire d'un échantillon x_n pour chaque l^e classificateur faible qui est égal à $S_{n,l} = \sum_{u=1}^{l} \alpha_u h_{n,u}$, où α_l est le poids positif associé au classificateur faible et $h_{n,l}$ est sa réponse connue. À chaque niveau l, un seuil de classification θ_l est défini de sorte que tout échantillon ayant son score $S_{n,l} < \theta_l$ est rejeté. Les seuils de classification θ_l sont définis de sorte que pour tout l^e classificateur faible, la meilleure précision soit obtenue.

Détecteur ACF Dollár *et al.* [Dollár 2014] utilisent une procédure qui utilise le score cumulatif intermédiaire d'un échantillon x_n pour chaque classificateur faible l qui est égal à (2.5), où α_l est le poids positif associé au classificateur faible, $h_{n,l}$ est sa réponse connue et γ est un paramètre d'étalonnage de la cascade. Un seuil de classification θ est défini de sorte qu'à chaque niveau l, tout échantillon ayant son score $S_{n,l} < \theta$ est rejeté. Le paramètre d'étalonnage γ et le seuil de classification θ sont hélas définis empiriquement.

$$S_{n,l} = \sum_{u=1}^{l} (\alpha_u h_{n,u} + \gamma) \tag{2.5}$$

Bien que ces stratégies soient utilisées avec succès pour *tuner* les seuils de détecteurs, notamment DBP [Zhang 2007, Dollár 2012], on ne prend nullement en compte

le temps de calcul ou réponse dans la modélisation alors que le coût de traitement est souvent un facteur limitatif vis-à-vis de l'application sous-jacente. Ce constat motive notre nouvelle modélisation de *soft-cascade* basée sur la programmation linéaire en nombres entiers (PLNE) nommé *Mean-Cascade Response-Time Minimization Problem* (MSCRMP) [BarbosaAnda 2016a]. Nos évaluations démontrent que notre cascade, pour un choix donné de descripteurs, est plus rapide tout en préservant les performances de détection à l'identique.

2.1.1 Définition du problème

Cette section définit plus formellement notre stratégie MSCRMP [BarbosaAnda 2016a], mise en oeuvre dans ce travail. Il s'agit d'une *soft-cascade* formée de L classificateurs faibles avec un ensemble de configuration composé de N échantillons, partitionné en J positifs et K négatifs (c'est-à-dire, $\mathbf{N} = \mathbf{J} \cup \mathbf{K}$). Les indices n, j et k désignent respectivement un échantillon quelconque, positif et négatif. Le classificateur faible situé au niveau l a un coût positif c_l qui correspond au temps de calcul nécessaire pour analyser un seul échantillon. Il a aussi un poids positif α_l qui encode son importance dans la cascade. $S_{n,l} = \sum_{u=1}^{l} \alpha_u h_{n,u}$ est le score de l'échantillon n au niveau l où $h_{n,l}$ est la réponse connue du classificateur faible l, qui est égale à 1 (resp. -1) lorsque le classificateur faible voit n comme positif (resp. négatif).

Notre stratégie MSCRMP détermine, à chaque niveau l, un seuil θ_l tel que si $S_{n,l} < \theta_l$ alors l'échantillon n est rejeté au niveau l (il ne passera pas par les niveaux de cascade suivants $u > l$). Inversement, si $S_{n,l} \geq \theta_l$ alors n poursuivra au niveau suivant $l + 1$. Évidemment, lorsqu'un échantillon est rejeté au niveau l, on obtient une économie de temps de calcul égale à $\sum_{u=l+1}^{L} c_u$.

On impose qu'un nombre minimum de TP échantillons positifs ne soient jamais rejetés à aucun niveau de la cascade. L'objectif est de trouver un vecteur de seuil $\Theta = \{\theta_1, \ldots, \theta_L\}$ qui minimise le temps total de calcul (ou de manière équivalente, qui maximise le temps total de calcul économisé). Bien que cet objectif ne s'applique qu'à l'ensemble des échantillons dits de configuration, notons que celui-ci est censé être statistiquement représentatif de tout autre sous-ensemble d'échantillons de la base, comme cela est couramment supposé dans l'apprentissage automatique supervisé. Par conséquent, la minimisation du temps de calcul total (pour l'ensemble de configuration) peut être considérée comme équivalente à la minimisation du temps de réponse moyen (pour tout ensemble d'échantillons inconnus).

En supposant $\alpha_l = 1 \ \forall l$ et considérant une cascade à cinq niveaux ($L = 5$), la figure 2.2 illustre l'évolution du score de six échantillons ; les échantillons positifs (resp. négatifs) sont dessinés avec des cercles bleus (resp. carrés rouges). Les scores, qui prennent leur valeur dans l'ensemble discret $\{-3, -2, -1, 0, 1, 2, 3\}$ pour cet exemple, sont affichés sur l'axe vertical.

Si l'on suppose un taux de vrais positifs minimum désiré $TPR = 50\%$, seulement deux vecteurs de seuil possibles Θ sont réalisables : $\Theta_a = \{1, 0, -1, 0, 1\}$ (figure 2.3) et $\Theta_b = \{-1, 0, 1, 2, 3\}$ (figure 2.4). En rappelant que $c_l = 1 \ \forall l$, l'utilisation de Θ_a

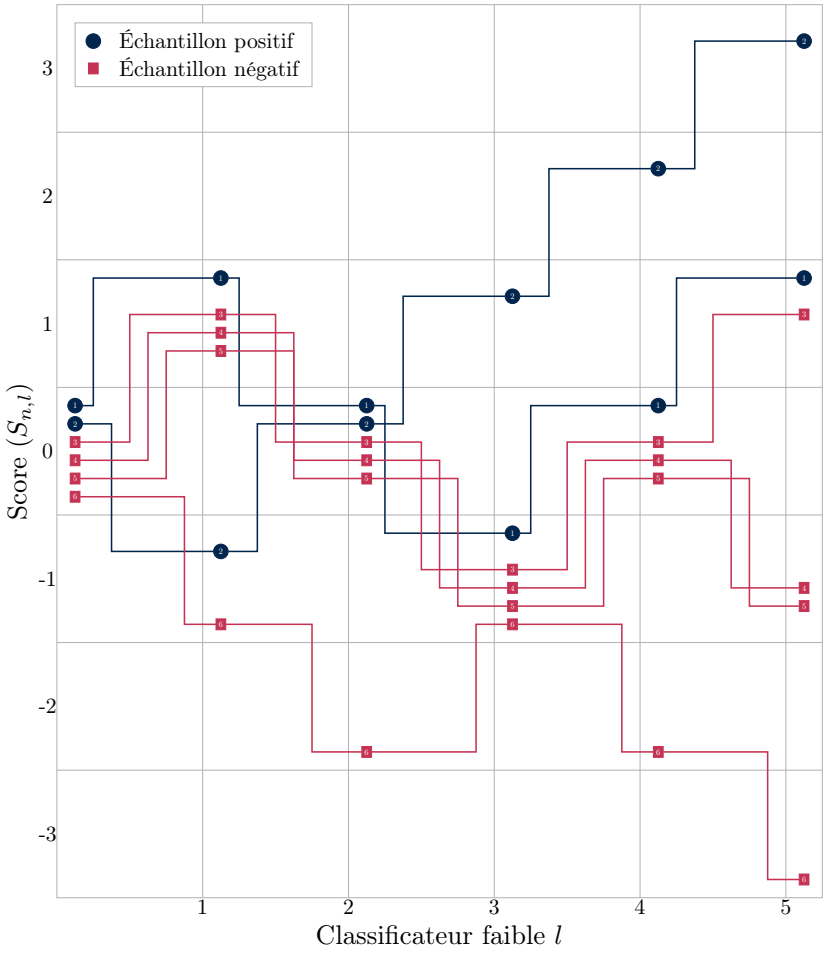


FIGURE 2.2 – Exemple d'évolution du score des échantillons d'une *soft-cascade* avec 5 classificateurs faibles (ayant le même coût $c_l = 1$ et le même poids $\alpha_l = 1$), 2 échantillons positifs et 4 échantillons négatifs.

donne un temps de calcul total égal à $6+4+4+4+4 = 22$, alors que celui associé à Θ_b est $6+6+5+1+1 = 19$, qui est optimal. Cet exemple illustre l'intérêt de rechercher un vecteur de seuil optimal en terme de temps de réponse.

Rappelons que la performance d'une *soft-cascade* n'est pas seulement caractérisée par le taux de vrais positifs TPR mais aussi par le taux de faux positifs FPR, qui donne le pourcentage d'échantillons négatifs jamais rejetés à tout niveau de la cascade. Évidemment, ce taux doit être aussi bas que possible et les échantillons négatifs devraient être rejetés le plus tôt possible de sorte que en pratique $|\mathbf{J}| << |\mathbf{K}|$. Par conséquent, dès que la décision de rejet d'un échantillon positif j ayant un score $S_{j,l}$ est faite à un niveau l, il faut également rejeter au niveau l tout échantillon négatif k ayant un score $S_{k,l} \leq S_{j,l}$. En d'autres termes, le FPR peut être vu comme la conséquence du vecteur de seuil faisable choisi Θ et il n'est pas utile de l'imposer puisque, en pratique, avec $|\mathbf{J}| << |\mathbf{K}|$, l'optimisation tend à le rendre le plus faible possible.

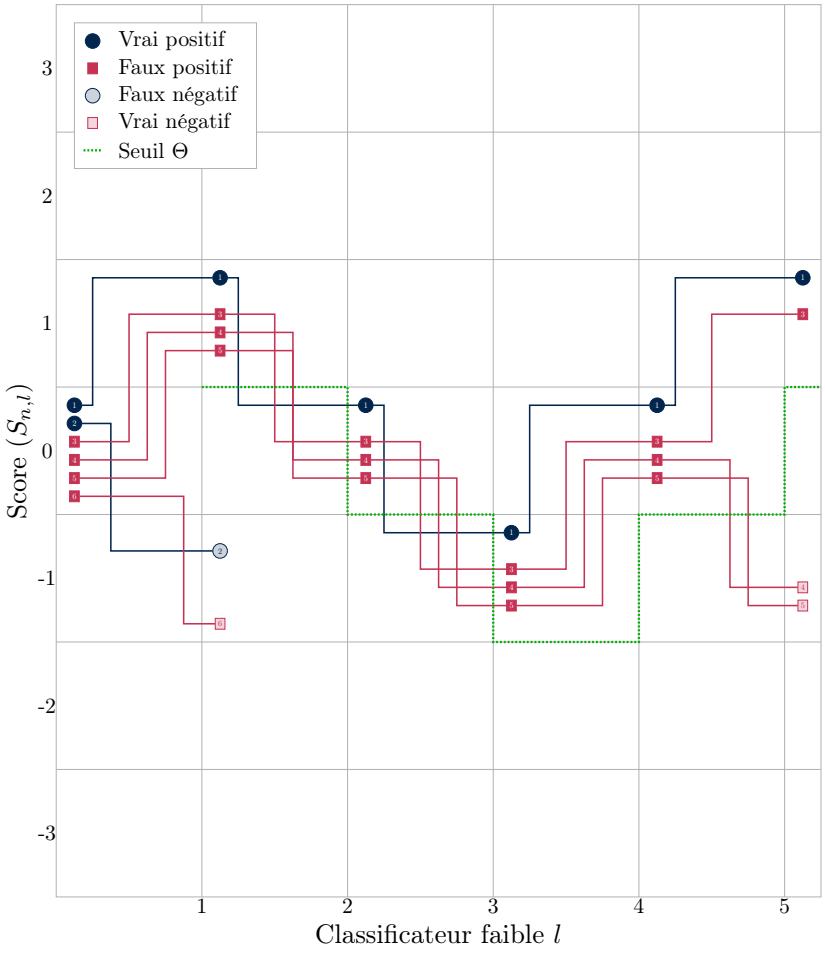


FIGURE 2.3 – Une solution possible Θ pour $TPR = 50\%$ pour l'exemple figure 2.2.

2.1.2 Complexité du problème

Nous prouvons la proposition suivante :

Théorème 2.1. *MSCRMP est NP-difficile.*

Démonstration. La preuve est basée sur une réduction du problème de la somme de sous-ensembles (SSP). Une instance de SSP est une paire (Σ, t), où $\Sigma = \{\sigma_1, \ldots, \sigma_R\}$ est un ensemble de R entiers positifs et t (la cible) est un nombre entier positif. Le problème de décision consiste à déterminer s'il existe un sous-ensemble σ^* de Σ dont la somme est égale à t. SSP est connu pour être NP-complet au sens ordinaire [Garey 1979].

Premièrement, pour tout MSCRMP, étant donné un vecteur de seuil Θ, on peut vérifier en temps polynomial s'il est réalisable vis-à-vis de TPR et déterminer le temps de calcul total qui en résulte. MSCRMP est donc dans NP. Considérant maintenant n'importe quelle instance de SSP, nous construisons un MSCRMP comme suit. La *soft-cascade* est composée de L niveaux avec $L = 2R$ et $R \in \mathbb{N}$. Nous

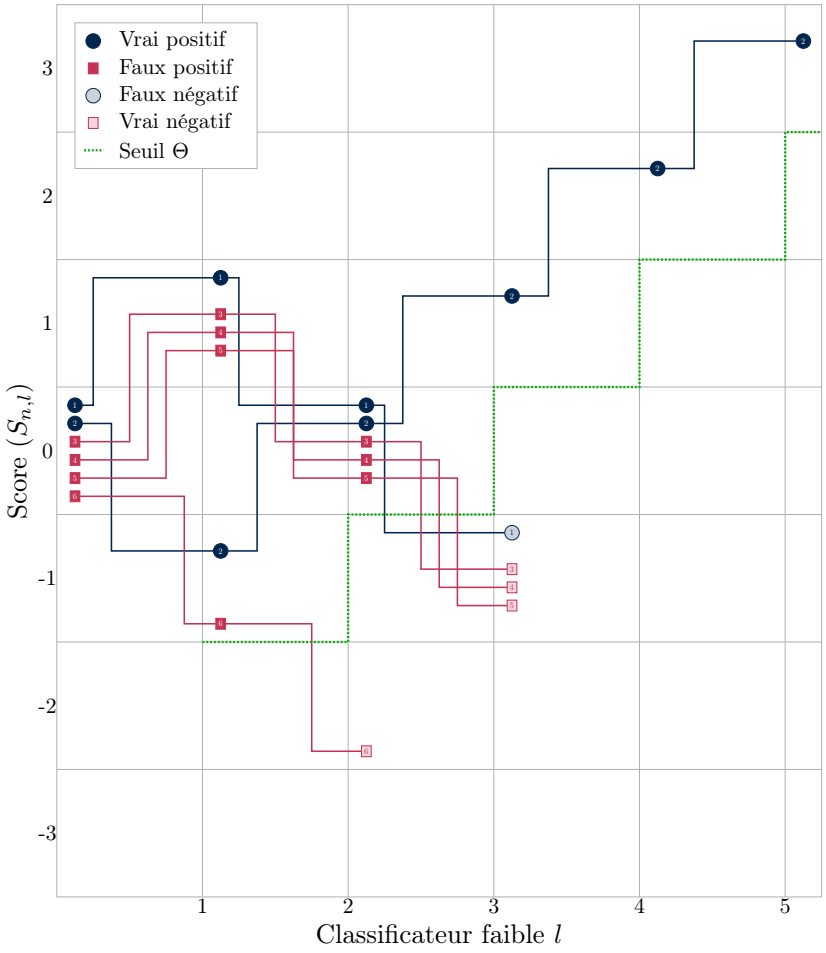


FIGURE 2.4 – Une autre solution possible Θ pour $TPR = 50\%$ pour l'exemple figure 2.2.

définissons $\alpha_{2u} = \alpha_{2u-1} = 1$ avec $u \leq R$ et $u \in \mathbb{N}$ (les scores sont entiers) et $TPR = |\mathbf{J}| - t$. Les coûts sont tels que $c_L = 1$ et $c_l = 0$ $\forall l < L$. L'ensemble $\mathbf{J}$ est partitionné en R sous-ensembles (c'est-à-dire, $\mathbf{J} = \{\mathbf{J}_1, \ldots, \mathbf{J}_R\}$) tels que : i) $|\mathbf{J}_i| = \sigma_i$ et ii) le score d'un échantillon $j \in \mathbf{J}_i$ est égal à $l \div 2 - 1$ quand $l = 2i - 1$ et à $l \div 2$ à tout autre niveau, $\forall l = 1, \ldots, 2R$.

Sous ces hypothèses, faisons quelques observations. Premièrement, à un niveau de cascade pair $l = 2i$, tous les échantillons ont la même valeur de score i, alors que pour un niveau impair $l = 2i - 1$, seulement les échantillons appartenant à $\mathbf{J}_i$ ont score $i - 1$ (i pour tous les autres échantillons). De plus, comme $TP = |\mathbf{J}| - t$, aucun vecteur de seuil réalisable ne peut rejeter un échantillon à un niveau pair, car il rejetterait tout l'ensemble $\mathbf{J}$. De plus, il n'y a que deux décisions possibles à un niveau impair $l = 2i - 1 : \theta_l > i - 1$ et l'ensemble $\mathbf{J}_i$ est rejeté, ou $\theta_l \leq i - 1$ et tous les échantillons restants sont conservés. Enfin, si $\mathbf{J}_i$ est rejeté au niveau $l = 2i - 1$, l'économie de temps sera égale à $|\mathbf{J}_i| \sum_{u=l+1}^{L} c_u = \sigma_i$. Comme l'économie totale doit être maximisée, le MSCRMP réduit vise à maximiser le nombre total d'échantillons

positifs rejetés, à condition qu'il ne reste pas supérieur à t. Nous montrons ci-dessous qu'il existe un vecteur de seuil Θ^* pour cette instance du MSCRMP si et seulement si l'instance SSP est réalisable.

($\Leftarrow$) Considérons qu'une solution réalisable $\Sigma^A \subseteq \Sigma$ de notre instance du SSP existe telle que $\sum_{\sigma_i \in \Sigma^A} \sigma_i = t$. Clairement, dans l'instance réduite du MSCRMP correspondante, si l'on considère un vecteur de seuil Θ^A tel que i) $\theta^A_{2i-1} = \theta^A_{2i} = i$ quand $\sigma_i \in \Sigma^A$ et ii) $\theta^A_{2i-1} = \theta^A_{2i} = i - 1$ quand $\sigma_i \notin \Sigma^A$, alors seulement les échantillons positifs dans l'ensemble $\cup_{i|\sigma_i \in \Sigma^*} \mathbf{J}_i$ seront rejetés, qui compte exactement t membres. Par conséquent, Θ^A est une solution réalisable (et optimale) de l'instance du MSCRMP. ($\Rightarrow$) Considérons maintenant une solution optimale (réalisable) Θ^* d'un MSCRMP réduit et une solution Σ^* de l'instance du SSP initiale telle que $\sigma_i \in \Sigma^*$ si et seulement si $\mathbf{J}_i$ est rejeté (c'est-à-dire, $\theta^*_{2i-1} < i - 1$). Si le nombre total z^* d'échantillons positifs rejetés est égal à t alors Σ^* sera évidemment réalisable pour le SSP initial. Inversement, si le nombre z^* d'échantillons positifs rejetés est strictement inférieur à t alors Σ^* ne sera évidemment pas réalisable pour le SSP. De plus, comme le nombre d'échantillons positifs est maximisé, il n'y a aucun moyen d'augmenter z^* en préservant le taux de vrais positifs $TPR = |\mathbf{J}| - t$, ce qui signifie que le SSP initial n'est pas réalisable. □

Le MSCRMP étant défini et sa complexité ayant été analysée, nous proposons ci-après une formulation mathématique de ce problème.

2.2 Formalisation

2.2.1 Modèle par discrimination d'échantillons

Proposition 2.1. *L'ensemble des solutions d'une instance de MSCRMP peut être limité aux vecteurs de seuil Θ tels que i) à tout niveau l, $\exists n : \theta_l = S_{n,l}$ et ii) $\theta_l - \alpha_{l+1} \leq \theta_{l+1}$.*

Démonstration. Supposons que le seuil θ_l ne respecte pas la propriété i). Considérons maintenant un échantillon n atteignant le niveau l et ayant le score le plus bas, tel que $\theta_l < S_{n,l}$, il est clair que θ_l peut être augmenté jusqu'à la valeur $S_{n,l}$ sans que cela ne provoque aucun changement. De plus, si la propriété ii) n'est pas satisfaite (c'est-à-dire, $\theta_{l+1} < \theta_l - \alpha_{l+1}$), alors θ_{l+1} peut être augmenté jusqu'à $\theta_l - \alpha_{l+1}$ toujours sans conséquence sur les échantillons rejetés. □

À partir de cette propriété directe, il est possible de proposer un PLNE qui n'utilise aucune variable θ_l dans sa formulation. Les variables binaires $x_{n,l}$ sont introduites telles que : $x_{n,l} = 1$ si $\theta_l > S_{n,l}$ pour la première fois. La fonction objectif (2.6) maximise le gain de temps de calcul, qui s'exprime linéairement en fonction des variables $x_{n,l}$. Les contraintes de type (2.7) assurent qu'un échantillon n'est rejeté qu'une seule fois au plus. Le TPR désiré est pris en compte grâce à la contrainte (2.8). Finalement, les contraintes (2.9) décrivent les relations entre les échantillons et les scores : à chaque fois que $x_{n,l} = 1$, ils indiquent que tous les

échantillons ayant un score inférieur ou égal à $S_{n,l}$ au niveau l doivent avoir été rejeté à un niveau $u \leq l$. À partir d'une solution optimale, le vecteur des seuils Θ peut être facilement calculé puisque : $\theta_l = \min_{n \in N : \sum_{u=1}^{l} x_{n,u}=0} S_{n,l}\ \forall l$.

Première Formulation PLNE (BIP1)

Maximiser

$$\sum_{l=1}^{L}\sum_{n=1}^{N}\left[x_{n,l}\left(\sum_{u=l+1}^{L} c_u\right)\right] \tag{2.6}$$

Sous contrainte que

$$\sum_{l=1}^{L} x_{n,l} \leq 1\ \forall n \tag{2.7}$$

$$|\mathbf{J}| - \sum_{l=1}^{L}\sum_{n\in\mathbf{J}} x_{n,l} \geq TP \tag{2.8}$$

$$x_{n,l} \leq \sum_{u=1}^{l} x_{v,u}\ \forall (n,l)\ \text{et}\ \forall v | S_{v,l} \leq S_{n,l} \tag{2.9}$$

$$x_{n,l} \in \{0,1\}\ \forall (n,l) \tag{2.10}$$

Soulignons que l'ensemble des contraintes de type (2.9) est très important (c'est-à-dire, $\frac{N(N-1)}{2} \times L$). Par conséquent, seulement des instances de problème de petite taille peuvent être résolues. Dans la sous-section suivante, une autre formulation PLNE est proposée pour pallier ce problème. Celle-ci est davantage performante à la fois en termes de temps de calcul et de capacité de taille d'instance.

2.2.2 Analyse de l'espace de recherche et reformulation du modèle

À partir de l'analyse des poids α_l de chaque classificateur faible, un arbre de score avec 2^L nœuds peut être construit de sorte que tout nœud (l, s), correspondant à un score s accessible au niveau l, ait deux nœuds adjacents $(l+1, s-\alpha_{l+1})$ et $(l+1, s+\alpha_{l+1})$. Un exemple d'un tel arbre est donné figure 2.5 pour une instance de *soft-cascade* avec quatre classificateurs faibles appris par AdaBoost ayant des poids $A = \{0.5253, 0.4443, 0.3753, 0.4081\}$. Tout chemin allant du nœud $(0,0)$ à un nœud de niveau L correspond à une évolution possible du score d'un échantillon. Nous notons en outre $\mathcal{C}_{(l,s)}$ l'ensemble des échantillons n tels que $S_{n,l} = s$. Remarquons que, selon l'ensemble d'échantillons considéré, il reste possible qu'aucun échantillon n'obtienne le score s au niveau l (c'est-à-dire, $\mathcal{C}_{(l,s)} = \emptyset$).

À partir d'un tel arbre de score, un graphe de seuil $\mathcal{T}(V, E)$ peut être construit à son tour ayant le même ensemble V de nœuds que l'arbre de score et tel que n'importe quel nœud $(l, s) \in V$ de niveau l est connecté par un arc $e \in E$ à un nœud $(l+1, s') \in V$ si et seulement si $s' \geq s - \alpha_{l+1}$, ainsi qu'illustré sur la figure 2.6 pour l'exemple de la figure 2.5. Selon la Proposition 2.1, toute solution du MSCRMP

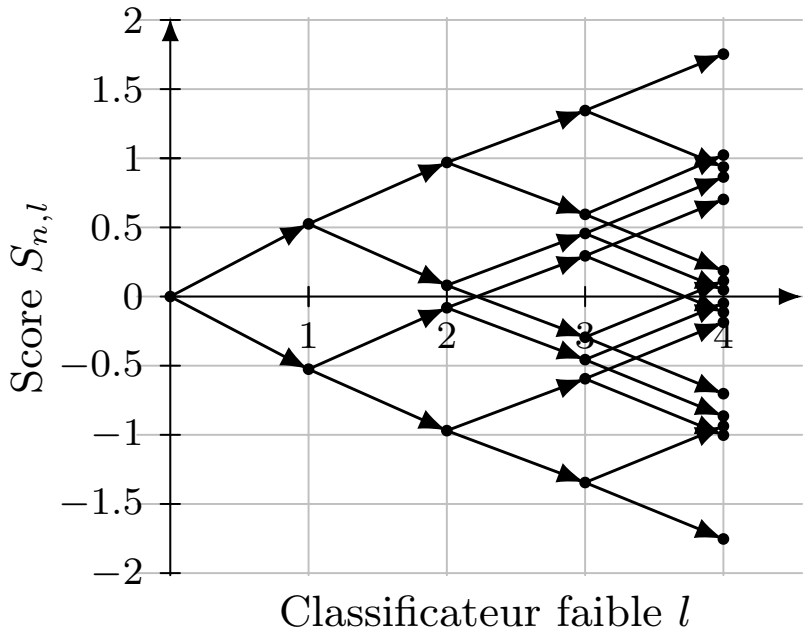


FIGURE 2.5 – Exemple d'un arbre de scores.

(c'est-à-dire, tout vecteur de seuil Θ) correspond à un chemin spécifique du nœud $(0,0)$ à un nœud de niveau L dans $\mathcal{T}$ tel que, si le nœud (l,s) est traversé, alors $\theta_l = s$. Le nombre de chemins différents est clairement exponentiel. Cependant, il est possible de le réduire radicalement en utilisant les propriétés de dominance suivantes.

Proposition 2.2. *Tout arc $e \in E$ du graphe de seuil $\mathcal{T}$ reliant un nœud (l,s) avec un nœud $(l+1,s')$ peut être supprimé si $\forall n \in \mathcal{C}_{(l,s)}$ $s' > S_{n,l+1}$.*

Démonstration. Considérons un chemin passant par le nœud (l,s). Il est clair que tous les échantillons (restants) appartenant à $\mathcal{C}_{(l,s)}$ ont un score $S_{n,l} = \theta_l$. Si le chemin continue de (l,s) à $(l+1,s')$ avec $s' > S_{n,l+1}$, $\forall n \in \mathcal{C}_{(l,s)}$, tous les échantillons restants appartenant à $\mathcal{C}_{(l,s)}$ seront rejetés au niveau $l+1$. Par conséquent, il aurait été plus rentable par rapport à la fonction objectif de les rejeter au niveau l. En d'autres termes, tout chemin passant par arc $e = ((l,s),(l+1,s'))$ avec $s' > S_{n,l+1}$, $\forall n \in \mathcal{C}_{(l,s)}$, est dominé par au moins un chemin passant par un nœud (l,s'') avec $s'' > s$. □

Proposition 2.3. *Tout arc $e \in E$ du graphe de seuil $\mathcal{T}$ reliant un nœud (l,s) avec un nœud $(l+1,s')$ peut être supprimé si $\forall n \in \mathcal{C}_{(l+1,s')}$ $s > S_{n,l}$.*

Démonstration. Considérons un chemin passant par le nœud (l,s). Si le chemin continue de (l,s) à $(l+1,s')$ avec $s > S_{n,l}$, $\forall n \in \mathcal{C}_{(l+1,s')}$, tous les échantillons appartenant à $\mathcal{C}_{(l+1,s')}$ ont été rejetés au niveau l. Par conséquent, tout chemin

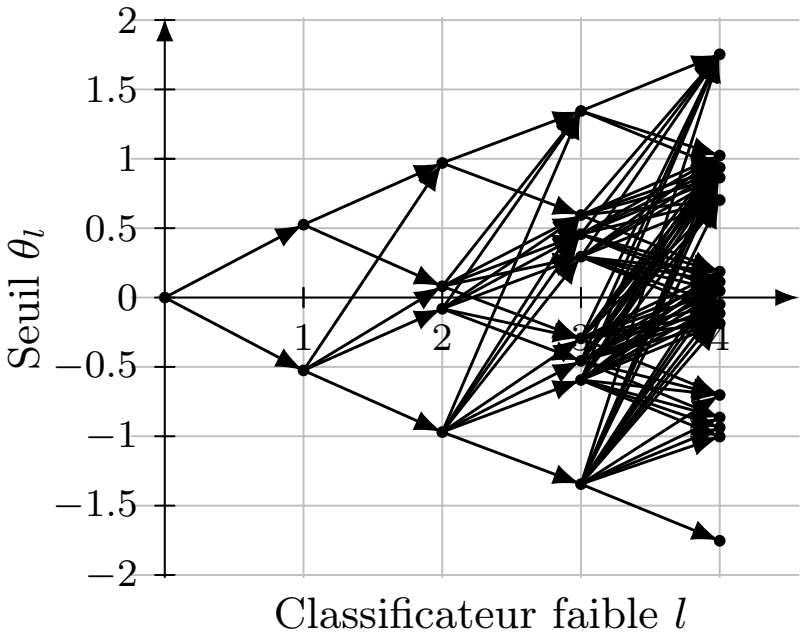


FIGURE 2.6 – Le graphe de seuil pour l'arbre de scores de la Figure 2.5.

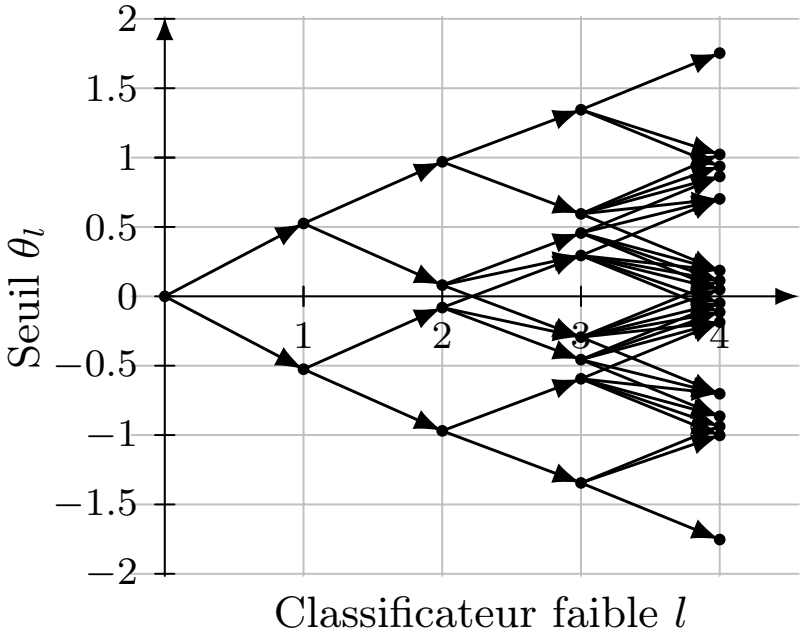


FIGURE 2.7 – Le graphe de seuil de la Figure 2.6 après avoir appliqué la proposition 2.2.

passant par arc $e = ((l, s), (l + 1, s'))$ avec $s > S_{n,l}$, $\forall n \in \mathcal{C}_{(l+1,s')}$, est dominé par au moins un chemin passant par un nœud $(l + 1, s'')$ avec $s'' > s'$. □

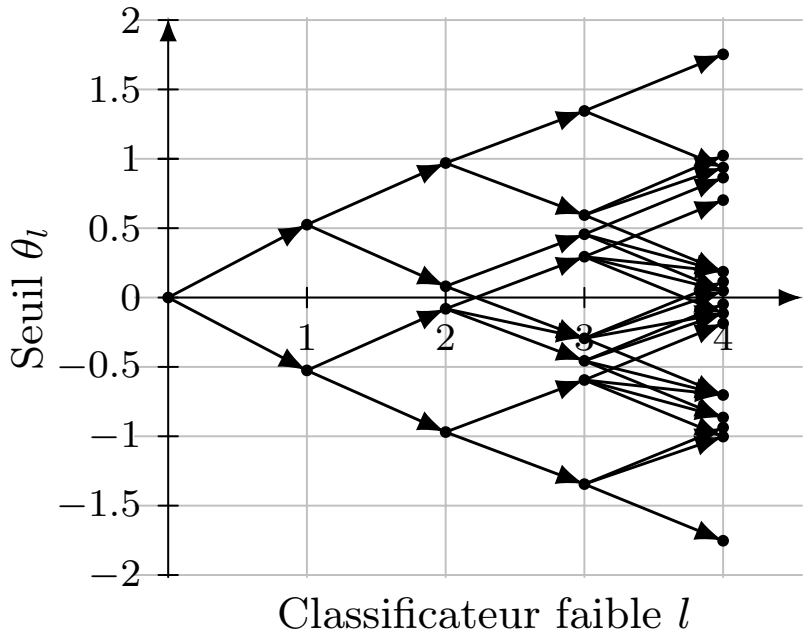


FIGURE 2.8 – Le graphe de seuil de la Figure 2.6 après appliquer les Propositions 2.2 et 2.3.

En plus des propriétés de dominance précédentes, si l'on considère la cible TPR, on peut réduire $\mathcal{T}$ en supprimant les chemins qui mènent nécessairement à de mauvaises $100\% - TPR$-solutions ou à des valeurs de TPR trop basses. Une mauvaise $100\% - TPR$-solution est un vecteur de seuil offrant un $TPR = 100\%$ tel qu'il existe encore des échantillons négatifs k qui auraient pu être rejetés en modifiant le vecteur de seuil, tout en conservant la valeur $TPR = 100\%$. Il est clair que les rejeter améliore la valeur FPR et le temps total de calcul. Des mauvaises $100\% - TPR$-solutions peuvent être facilement filtrées en enlevant de $\mathcal{T}$ tout nœud $(l, s) \in V$ tel que $s < \min_{j \in \mathbb{J}} S_{j,l}$. D'un autre côté, on peut également supprimer tout nœud $(l, s) \in V$ de sorte que la somme des échantillons positifs ayant un score $S_{j,l} \geq s$ est inférieure à TPR. Nous notons $\mathcal{T}_p(V_p, E_p)$ le graphe de seuil réduit obtenu après application des règles de réduction précédentes.

Afin de tirer profit de ce graphe de seuil réduit, nous proposons ci-dessous une formulation PLNE basée flot (BIP2). Comme dans le premier PLNE, nous considérons toujours les variables binaire s$x_{n,l}$ qui valent 1 si l'échantillon n est tel que $S_{n,l} < \theta_l$ pour la première fois. De plus, nous introduisons maintenant les variables binaires $\varphi_{n,l}$ et $\psi_{s,t,l}$. D'une part, $\varphi_{n,l} = 1$ s'il existe un classificateur faible au niveau $u \leq l$ tel que $S_{n,u} < \theta_u$. Clairement, les variables $x_{n,l}$ et $\varphi_{n,l}$ sont liées ensemble par la relation $\varphi_{n,l} = \varphi_{n,l-1} + x_{n,l}$, qui est valide pour tout échantillon n et niveau l. D'autre part, $\psi_{s,t,l}$ est une variable de flot : $\psi_{s,t,l} = 1$ si l'arc $e \in E_p$ entre le nœud $(l, s) \in V_p$ et $(l+1, t) \in V_p$ de $\mathcal{T}_p$ est sélectionné dans la solution. Les variables $\psi_{s,t,l}$ doivent être choisies de telle sorte qu'elles définissent un chemin dans $\mathcal{T}_p$, c'est-à-

dire, $\sum_{(l-1,t)\in\sigma^{-1}_{(l,s)}} \psi_{t,s,l-1} = \sum_{(l+1,t)\in\sigma_{(l,s)}} \psi_{s,t,l}$, $\forall(l,s)$. Les trois types de variables sont liés par la relation suivante : $x_{n,l} \leq \sum_{t|S_{n,l}<t} \sum_{(l+1,t)\in\sigma_{(l,s)}} \psi_{s,t,l} \leq \varphi_{n,l}$. Elle modélise qu'à tout niveau l, un échantillon n ne peut être rejeté que si la classe de score sélectionnée (l,s) est telle que $s > S_{n,l}$. Comme on peut l'observer, ce deuxième PLNE présente une quantité moindre de contraintes, bien que de nouvelles variables aient été introduites.

Deuxième Formulation PLNE (BIP2)

Maximiser

$$\sum_{l=1}^{L}\sum_{n=1}^{N}\left[x_{n,l}\left(\sum_{u=l+1}^{L} c_u\right)\right] \tag{2.11}$$

Sous contrainte que

$$\sum_{n\in\mathbf{J}} x_{n,L+1} \geq TP \tag{2.12}$$

$$\varphi_{n,l} = \varphi_{n,l-1} + x_{n,l}\ \forall(n,l) \tag{2.13}$$

$$\psi_{0,0,1} + \psi_{0,1,1} = 1 \tag{2.14}$$

$$\sum_{(l-1,t)\in\sigma^{-1}_{(l,s)}} \psi_{t,s,l-1} = \sum_{(l+1,t)\in\sigma_{(l,s)}} \psi_{s,t,l}\ \forall(s,l) \tag{2.15}$$

$$x_{n,l} \leq \sum_{t|S_{n,l}<t}\ \sum_{s|(l+1,t)\in\sigma_{(l,s)}} \psi_{s,t,l} \leq \varphi_{n,l}\ \forall(n,l) \tag{2.16}$$

$$x_{n,l} \in \{0,1\}, \varphi_{n,l} \in \{0,1\}\ \forall(n,l) \tag{2.17}$$

$$\psi_{s,t,l} \in \{0,1\}\ \forall(s,t,l) \tag{2.18}$$

On note que, du fait de la relation de récurrence (2.13), les variables $\varphi_{n,l}$ peuvent être substituées puisque $\varphi_{n,l} = \sum_{u=1}^{l} x_{n,l}$. Nous avons choisi de les conserver dans la description pour plus de clarté.

2.2.3 Méthodes d'approximation par recherche locale

Comme le nombre de contraintes (2.16) dépend du produit de la taille de cascade L et de la taille de l'ensemble de configuration N, l'approche précédente demeure difficile à appliquer à des *soft-cascades* de grande échelle car la taille du modèle explose et devient trop grande pour permettre une résolution en temps fini, même en utilisant les meilleurs solveurs commerciaux.

Nous détaillons ci-après une procédure de recherche itérative permettant de surmonter les inconvénients précédents [BarbosaAnda 2018a]. Même si elle ne garantit plus l'optimalité, elle permet de résoudre des instances de grande taille et fournit des solutions de bonne qualité en ce qui concerne les critères de performance de détection et de temps de calcul. Deux techniques complémentaires sont utilisées : (1) une procédure de recherche locale itérative qui considère une structure de voisinage

spécifique définie comme une enveloppe d'un chemin dans $\mathcal{T}$, et (2) une procédure de réduction de la cascade.

2.2.3.1 Une méthode de recherche locale

Considérant un chemin solution Θ_α de $\mathcal{T}$ trouvé à l'itération α, nous définissons une enveloppe autour de ce chemin. Cette enveloppe est un un sous-graphe $\mathcal{T}_{\Theta_\alpha,\Delta}(V_{\Theta_\alpha,\Delta}, E_{\Theta_\alpha,\Delta}) \subset \mathcal{T}(V,E)$ définit grâce à un vecteur d'écart $\Delta = \{\delta_1, \ldots, \delta_L\}$. Nous notons (l, s^0) le nœud de $\mathcal{T}$ sélectionné dans Θ_α au niveau l. Les nœuds de $V_{\Theta_\alpha,\Delta}$ sont tels que $V_{\Theta_\alpha,\Delta} = \bigcup_{l=0}^{L}\{(l, s^{-\delta_l}), (l, s^{-\delta_l+1}), \ldots, (l, s^{-1}), (l, s^0), (l, s^1), \ldots, (l, s^{\delta_l-1}), (l, s^{\delta_l})\}\}$, où (l, s^u) ((l, s^{-u}), respectivement) est le u^{e} score supérieur (inférieur, respectivement) à (l, s^0). $E_{\Theta_\alpha,\Delta} \subseteq E$ sont tous les arcs appartenant à E qui connectent les nœuds de $V_{\Theta_\alpha,\Delta}$, de sorte que $\mathcal{T}_{\Theta_\alpha,\Delta}(V_{\Theta_\alpha,\Delta}, E_{\Theta_\alpha,\Delta}) \subseteq \mathcal{T}(V,E)$. Ceci est illustré sur la figure 2.9 où la solution initiale $\Theta_\alpha = \{-0.5253, -0.9693, -1.3449, -1.753\}$ est présentée en rouge, le voisinage $\mathcal{T}_{\Theta_\alpha,\Delta}(V_{\Theta_\alpha,\Delta}, E_{\Theta_\alpha,\Delta})$ avec $\delta_l = 1$ est présenté en noir et l'espace de seuil non-exploré $\mathcal{T}(V,E) \setminus \mathcal{T}_{\Theta_\alpha,\Delta}(V_{\Theta_\alpha,\Delta}, E_{\Theta_\alpha,\Delta})$ est en gris. Notez que Δ permet un contrôle précis de la forme de l'enveloppe. Évidemment, $\mathcal{T}_{\Theta_\alpha,\Delta}(V_{\Theta_\alpha,\Delta}, E_{\Theta_\alpha,\Delta}) = \mathcal{T}(V,E)$ pour les grandes valeurs de δ_l.

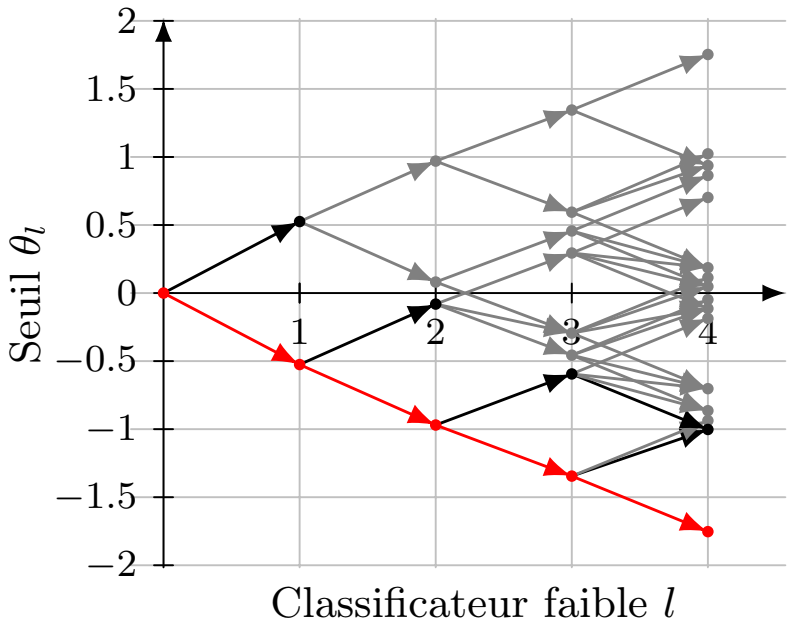


FIGURE 2.9 – Exemple d'enveloppe de taille $\delta_l = 1$ dans le graphe de seuil de la figure 2.8.

Le modèle PLNE peut être appliqué efficacement sur le sous-graphe $\mathcal{T}_{\Theta_\alpha,\Delta}(V_{\Theta_\alpha,\Delta}, E_{\Theta_\alpha,\Delta})$ pour trouver un autre chemin $\Theta_{\alpha+1}$ strictement meilleur que Θ_α. Dans ce cas, une nouvelle enveloppe peut être définie à partir de $\Theta_{\alpha+1}$ et le processus de résolution peut être itéré jusqu'à ce qu'il n'y ait plus d'amélioration.

La figure 2.10 présente l'itération $\alpha + 1$ pour la recherche locale effectuée à partir de l'itération α dans la figure 2.9. Lorsque plus aucune amélioration n'est trouvée, l'enveloppe peut être étendue progressivement autour de Θ_α jusqu'à ce qu'une meilleure solution soit trouvée ou qu'une enveloppe maximale donnée soit atteinte. La figure 2.11 présente le voisinage $\mathcal{T}_{\Theta_\alpha,\Delta}(V_{\Theta_\alpha,\Delta}, E_{\Theta_\alpha,\Delta})$ avec $\delta_l = 2$ pour l'itération $\alpha + 1$ de la figure 2.10. Cette procédure, baptisée *GLS*, est présentée dans l'algorithme 2.1.

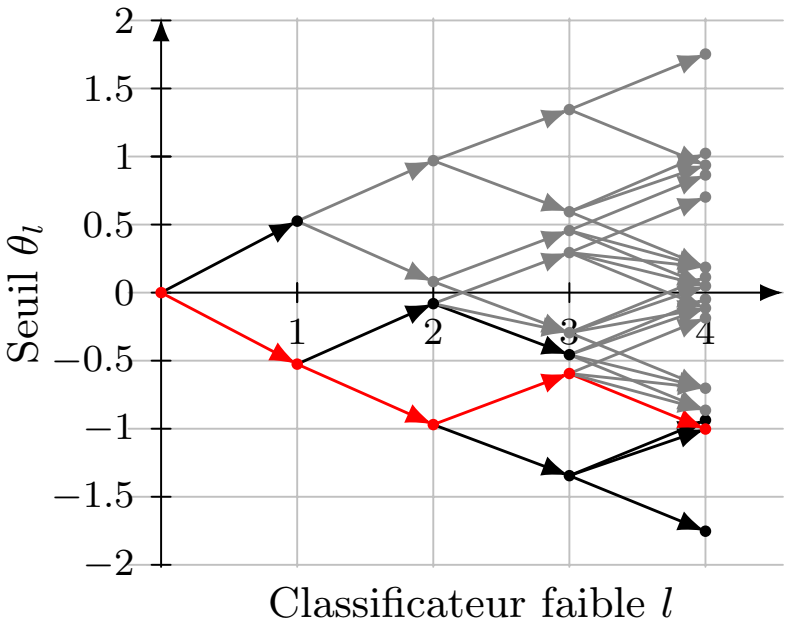


FIGURE 2.10 – Exemple de l'itération $\alpha + 1$ de la recherche locale pour l'itération α de la figure 2.9.

2.2.3.2 Une procédure de réduction de la cascade

Étant donnée une solution réalisable Θ, si la cible TPR obligatoire est atteinte pour la première fois à l'étape $l_{TPR} < L$, alors les seuils des étapes $l_{TPR} + 1$ à L sont imposés : ils sont égaux au score minimum des échantillons vrais positifs dans l'étape l_{TPR}. En effet, comme la cible TPR obligatoire est atteinte à l'étape l_{TPR}, le détecteur n'est plus autorisé à classer comme négatif un quelconque échantillon vrai positif restant, cela jusqu'à la fin de la cascade.

Sur la base de cette propriété, étant donnée une solution initiale Θ_0, nous nous concentrons uniquement sur la détermination d'une solution partielle $\Theta^{L'}$ pour les premières étapes $L' = l_{TPR}$. Nous définissons la solution complète Θ en fonction de l'équation (2.19) - c'est-à-dire, pour toutes les étapes $l \leq L'$ les valeurs de seuil θ_l sont celles de la solution partielle $\Theta^{L'}$, alors que pour toutes les autres étapes, les valeurs de seuil θ_l sont fixées au score minimum $S_{n,l}$ enregistré par les échantillons positifs qui ont un score partiel final $S_{n,L'}$ supérieur à $\theta^{L'}_{L'}$.

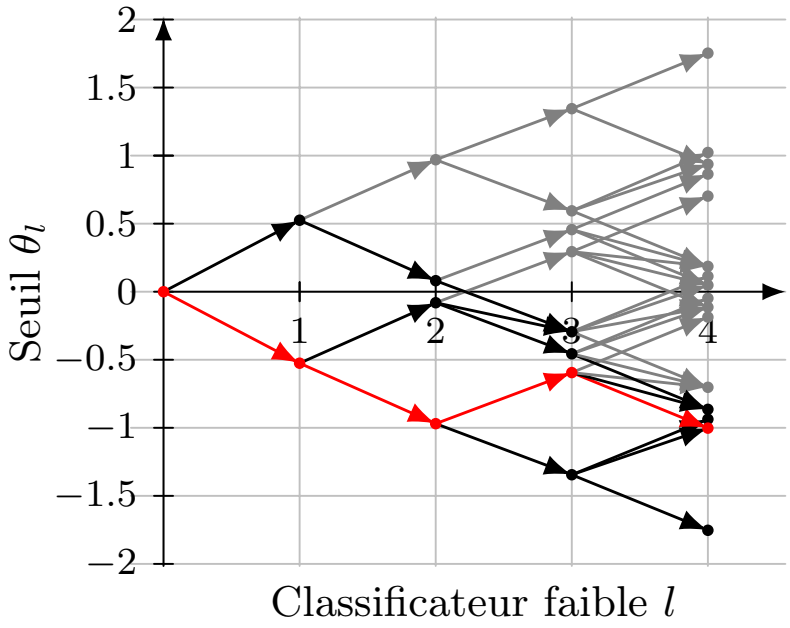


FIGURE 2.11 – Exemple d'enveloppe de taille $\delta_l = 2$ dans le graphe de seuil de la Figure 2.10.

$$\theta_l = \begin{cases} \theta_l^{L'} & l \leq L' \\ \min\limits_{\{n|S_{n,L'} > \theta_{L'}^{L'}, y_n = 1\}} S_{n,l} & \text{sinon} \end{cases} \tag{2.19}$$

Un exemple de cette réduction est illustré sur la figure 2.12. La solution initiale Θ_0 est présentée en rouge, les étapes $1 \leq l \leq l_{TPR}$ sont présentées en noir et les étapes $l > l_{TPR}$ sont en gris.

La solution complète Θ peut être utilisée comme nouvelle solution initiale Θ_β pour répéter la procédure jusqu'à ce que le nouveau l_{TPR} soit égal à la longueur de la cascade réduite L', comme présenté dans l'algorithme 2.2.

2.2.3.3 Méthode composée

Une procédure composée est construite à partir des deux méthodes précédentes. La *soft-cascade* est réduite comme indiqué dans la section 2.2.3.2. Au lieu d'utiliser le modèle PLNE comme dans l'algorithme 2.2, la recherche locale est appliquée selon l'algorithme 2.1, en utilisant comme entrée la solution initiale Θ_β (ayant la performance souhaitée TPR). La procédure finale est décrite dans l'algorithme 2.3.

2.2.4 Partitionnement de la base de calibration

La base de calibration peut être constituée d'un très grand nombre d'échantillons. Rappelons que la taille du PLNE proposé dépend non seulement de la longueur de la *soft-cascade*, et par extension de la taille de l'espace de recherche, mais

Algorithme 2.1 GLS

Nécessite: $\text{tpr}(\Theta_0) \geq TPR$ **et** $\delta_{\max} \geq 1$
  $\alpha \leftarrow 1$
  $better1 \leftarrow$ **vrai**
  tant que better1 **faire**
    $\delta \leftarrow 1$
    $better2 \leftarrow$ **faux**
    $\Theta_{\alpha,0} \leftarrow \Theta_{\alpha-1}$
    tant que non $better2$ **et** $\delta \leq \delta_{\max}$ **faire**
      $\Theta_{\alpha,\delta} \leftarrow \text{BIP}\left(\mathcal{T}_{\Theta_\alpha,\delta-1}, TPR\right)$
      $better2 \leftarrow \text{objfunc}(\Theta_{\alpha,\delta}) < \text{objfunc}(\Theta_{\alpha,\delta-1})$
      $\delta \leftarrow \delta + 1$
    fin tant que
    $\Theta_\alpha \leftarrow \Theta_{\alpha,\delta-1}$
    $better1 \leftarrow \text{objfunc}(\Theta_\alpha) < \text{objfunc}(\Theta_{\alpha-1})$
    $\alpha \leftarrow \alpha + 1$
  fin tant que
  retourne $\Theta_{\alpha-1}$

Algorithme 2.2 Procédure de réduction de la cascade

Nécessite: $\text{tpr}(\Theta_0) \geq TPR$
  $better \leftarrow$ **vrai**
  $\beta \leftarrow 1$
  $L' \leftarrow L$
  tant que $better$ **et** $l_{TPR} < L'$ **faire**
    $L' \leftarrow l_{TPR}$
    $\Theta^{L'}_{\beta-1} \leftarrow \{\theta_1, \ldots, \theta_{L'}\} \in \Theta_{\beta-1}$
    $\Theta^{L'}_{\beta} \leftarrow \text{BIP}\left(\mathcal{T}_{\Theta^{L'}_{\beta-1}}, TPR\right)$
    pour tout $l | 1 \leq l \leq L$ **faire**
      si $l \leq L'$ **alors**
        $\theta_l \in \Theta_\beta \leftarrow \theta^{L'}_l \in \Theta^{L'}_\beta$
      sinon
        $\theta_l \in \Theta_\beta \leftarrow \min_{\{n | S_{n,L'} > \theta^{L'}_{L'} \in \Theta^{L'}_\beta, y_n = 1\}} S_{n,l}$
      fin si
    fin pour
    $better \leftarrow \text{objfunc}(\Theta_\beta) < \text{objfunc}(\Theta_{\beta-1})$
    $\beta \leftarrow \beta + 1$
  fin tant que
  retourne $\Theta_{\beta-1}$

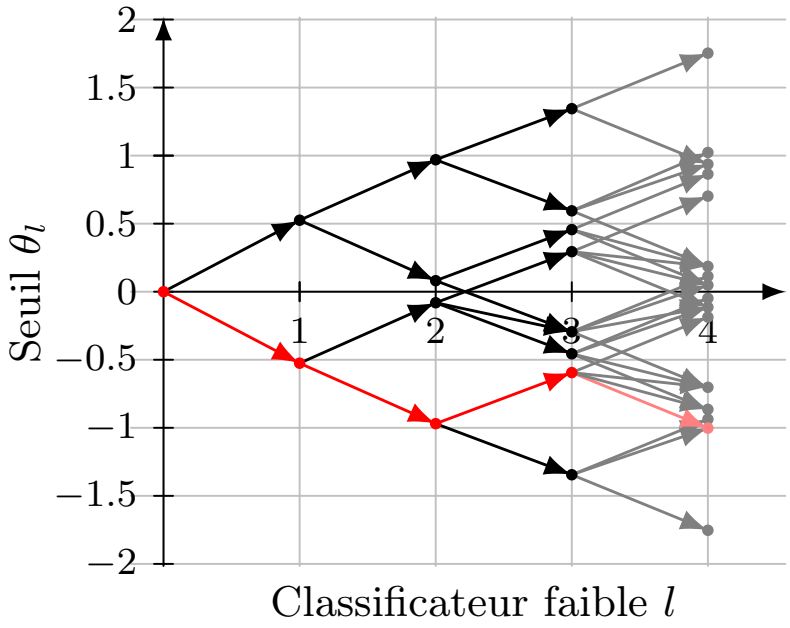


FIGURE 2.12 – Exemple de réduction d'un graphe de seuil d'une *soft-cascade* quand le TPR désiré est atteint à l'étape $l_{TPR} = 3$.

aussi de la taille de la base de calibration. Certains échantillons de la base sont redondants et induisent les mêmes contraintes dans le processus d'optimisation. Il est donc pertinent de les filtrer a priori.

Ainsi, les échantillons ayant une évolution de score similaire peuvent être remplacés par un unique échantillon ayant cette évolution. On affecte alors à l'échantillon conservé une pondération qui sera prise en compte ultérieurement dans le temps de calcul et les performances de détection désirés. Les échantillons négatifs, ayant une forte pondération, seront ainsi rejetés plus tôt dans la cascade pour réduire le coût de traitement. De plus, les échantillons positifs ayant une forte pondération contribuent davantage à atteindre le TPR désiré.

Une deuxième idée consiste à partitionner les échantillons positifs ou négatifs dans la base de calibration. Le but est aussi d'éliminer les échantillons redondants. Nous proposons d'utiliser un partitionnement en k-moyennes (*k-mean clustering*) sur l'ensemble des échantillons à réduire. L'entrée du partitionnement est le vecteur de score de chaque échantillon. Le partitionnement infère k centroïdes, i.e. k échantillons désirés pour notre base. Ces centroïdes étant le vecteur moyen des échantillons assignés à chaque classe k, il n'y a aucune garantie qu'ils seront des vecteurs de scores possibles dans l'espace des scores. Chaque centroïde est ainsi remplacé par l'échantillon réel le plus proche. Nous obtenons alors un sous-ensemble de notre base avec k échantillons qui sont représentatifs de la base entière. Naturellement, il est important de ne pas mélanger des échantillons positifs et négatifs dans le partitionnement, cela pouvant créer des erreurs.

Algorithme 2.3 Procédure itérative de recherche

Nécessite: $\mathrm{tpr}\,(\Theta_0) \geq TPR$ **et** $\delta_{\max} \geq 1$
  $better \leftarrow$ **vrai**
  $\beta \leftarrow 1$
  $L' \leftarrow L$
  tant que $better$ **et** $l_{TPR} < L'$ **faire**
    $L' \leftarrow l_{TPR}$
    $\Theta^{L'}_{\beta-1} \leftarrow \{\theta_1, \ldots, \theta_{L'}\} \in \Theta_{\beta-1}$
    $\Theta^{L'}_{\beta} \leftarrow \mathrm{GLS}\left(\Theta^{L'}_{\beta-1}, TPR, \delta_{\max}\right)$
    pour tout $l | 1 \leq l \leq L$ **faire**
      si $l \leq L'$ **alors**
        $\theta_l \in \Theta_\beta \leftarrow \theta^{L'}_l \in \Theta^{L'}_\beta$
      sinon
        $\theta_l \in \Theta_\beta \leftarrow \min_{\{n | S_{n,L'} > \theta^{L'}_{L'} \in \Theta^{L'}_\beta, y_n = 1\}} S_{n,l}$
      fin si
    fin pour
    $better \leftarrow \mathrm{objfunc}\,(\Theta_\beta) < \mathrm{objfunc}\,(\Theta_{\beta-1})$
    $\beta \leftarrow \beta + 1$
  fin tant que
  retourne $\Theta_{\beta-1}$

2.3 Expérimentations

2.3.1 Implémentation

Pour tester notre approche et ses variantes, deux *soft-cascades* sont formées et testées sur des images issues : (1) de la base de données publique INRIA [Dalal 2005] ; et (2) de la base de données publique Caltech [Dollár 2009a, Dollár 2012]. L'apprentissage est réalisé à l'aide de la boîte à outils *Piotr's Computer Vision Matlab Toolbox* [Dollár 2016] qui fournit une implémentation de la *soft-cascade* ACF issue de [Dollár 2014]. Ici, nous choisissons d'utiliser ACF comme référence, car il est l'un des meilleurs détecteurs de la littérature eu égard à son coût de traitement et ses performances de détection (voir chapitre 1). Il constitue aussi le détecteur de référence dans le MOT *Challenge* [LealTaixé 2015].

L'implémentation de la *soft-cascade* ACF de [Dollár 2014] définit le vecteur de seuils Θ sur un ensemble linéaire fixe de valeurs, c'est-à-dire $\theta_l = -1 + l\gamma$ où γ est un paramètre d'étalonnage fixé empiriquement pour chaque base de données. Nous définissons c_l comme une constante à tout niveau de cascade l et utilisons c_l comme unité de traitement pour le reste de ce travail. En utilisant l'algorithme DBP [Zhang 2007, Dollár 2012] et la *soft-cascade* ACF, une autre variante de la cascade a été construite. Nous avons construit deux variantes supplémentaires de la cascade en utilisant notre procédure composée appliquée sur les deux variantes précédentes, en utilisant $\delta_{\max} = 4$. Cela donne un total de quatre variantes de cas-

cade pour chaque base de données. Nous listons chaque variante avec son acronyme respectif :

ACF Le détecteur ACF comme décrit dans [Dollár 2014],

ACF+GLS Le détecteur ACF réglé par notre approche et le vecteur de seuils par défaut comme solution initiale,

ACF+DBP Le détecteur ACF réglé avec l'algorithme DBP,

ACF+DBP+GLS Le détecteur ACF réglé par notre approche et la sortie de DBP comme solution initiale.

Tous les tests expérimentaux ont été effectués sur un processeur Intel® Core™ i5-4670 CPU de 3,4 GHz avec 16 Go de mémoire RAM DDR3 1600 MHz. Aucun GPU n'est utilisé.

2.3.2 Évaluations et discussions associées

2.3.2.1 Évaluations sur la base de données INRIA

Pour la base INRIA, la *soft-cascade* ACF comptent 2048 étages. La base d'apprentissage pour apprendre les classificateurs faibles et leurs poids associés est composée de 2474 échantillons positifs et de 15710 échantillons négatifs extraits des images d'apprentissage. Notre base de calibration inclut 1178 échantillons positifs et 10947 échantillons négatifs extraits d'images de test. Cette base est utilisée pour exécuter l'algorithme DBP et la méthode composée pour générer les quatre variantes de *soft-cascade* : ACF, ACF+GLS, ACF+DBP et ACF+DBP+GLS. Dans nos expériences, pour la détermination des seuils de classification, nous avons expérimenté un temps de calcul moyen de la méthode composée qui équivaut à $1,5$ heures.

Une fois le réglage des détecteurs réalisé, des évaluations sont effectuées sur les échantillons de la base de calibration. La courbe ROC (figure 2.13) illustre les performances des quatre variantes de *soft-cascade* sur cette base de calibration. Ces performances sont proches ; on note une légère amélioration pour le détecteur ACF.

La Figure 2.14 illustre les évaluations sur le temps moyen de calcul par fenêtre. Nous observons que les variantes de la méthode composée (ACF+GLS et ACF+DBP+GLS) ont un meilleur temps moyen de calcul que leurs variantes initiales (ACF et ACF+DBP), soit environ $3,06\%$.

De façon similaire, des évaluations par image [Dollár 2012] sont réalisées sur les 288 images de test de la base INRIA. La courbe des évaluations FPPI dans la figure 2.15 montre les performances des quatre variantes de la *soft-cascade* sur cette base. Ces évaluations sont cohérentes avec celles de la figure 2.13, montrant que nos variantes de *soft-cascade* ont des performances de détection similaires à celles de la *soft-cascade* ACF modulo un léger gain pour cette dernière.

La table 2.1 synthétise les fréquences de traitement par variante de la *soft-cascade*. Ces résultats corroborent ceux illustrés sur la figure 2.14. Nous observons que notre variante ACF+GLS traite 1.15 fois plus d'images par seconde que la *soft-cascade* ACF d'origine.

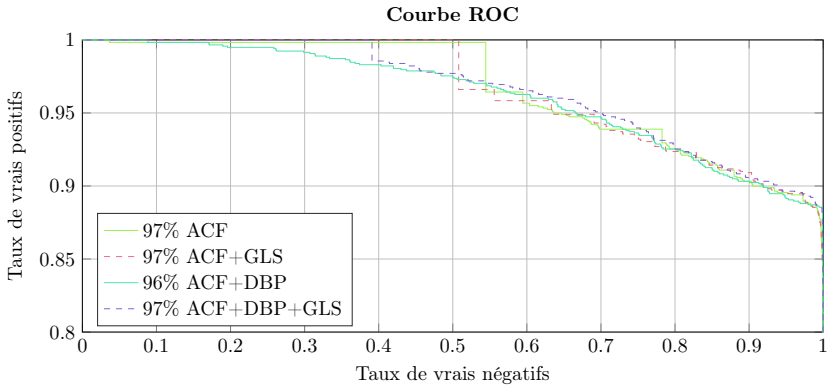


FIGURE 2.13 – Courbe ROC des évaluations par échantillon sur la base de calibration de la base INRIA.

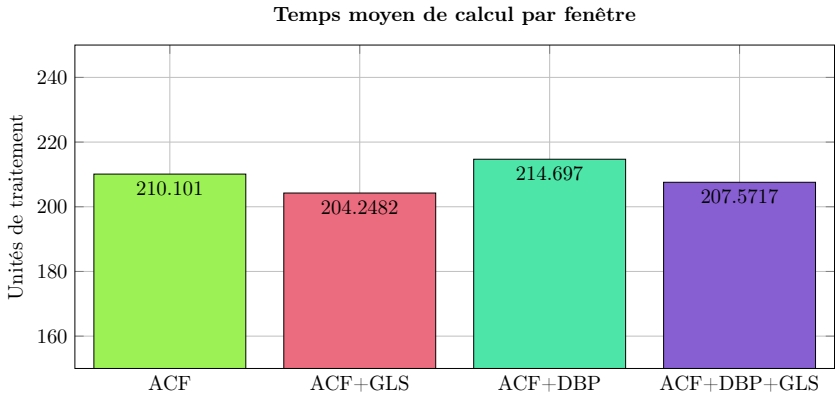


FIGURE 2.14 – Valeurs de la fonction objectif c'est-à-dire temps moyens de calcul des évaluations par fenêtre sur la base INRIA.

Détecteur	Images par seconde	Taux de défaut moyen logarithmique
ACF	30.71	**16.83**%
ACF+GLS	**31.86**	17.03%
ACF+DBP	27.10	17.28%
ACF+DBP+GLS	29.42	17.06%

TABLE 2.1 – Comparaison des performances de détection par rapport aux images par seconde sur la base INRIA.

Sur la figure 2.16, ces performances sont comparées avec celles rapportées dans [Cao 2016a]. Notons que les deux séries d'évaluations sont réalisées avec différentes configurations de *hardware*, comme indiqué dans les différences entre leurs résultats et ceux du détecteur ACF. Nous observons que nos résultats sont comparables à ceux rapportés par Cao *et al.* [Cao 2016a]. Le détecteur Crosstalk a un meilleur temps de calcul de $45,40$ images par seconde suivi de notre ACF+GLS

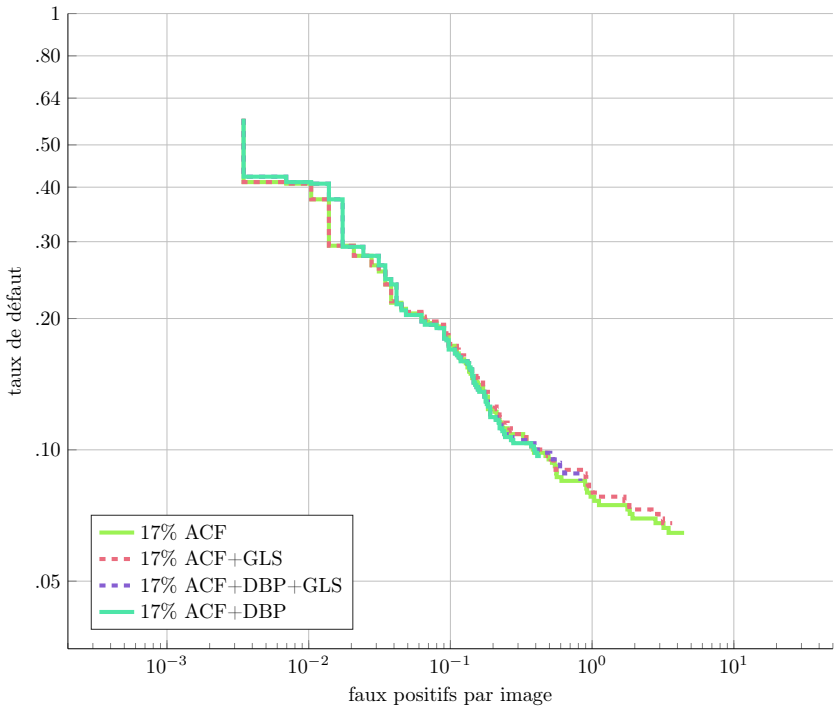


FIGURE 2.15 – Résumé des évaluations FPPI sur la base INRIA.

avec 31, 86 images par seconde et du détecteur ACF avec une moyenne de 31, 305 images par seconde. Les détecteurs offrant le meilleur compromis coût de traitement et performances de détection sont : Crosstalk, ACF+GLS, ACF, NNNF et SpatialPooling.

2.3.2.2 Évaluations sur la base de données Caltech

Pour la base Caltech, les *soft-cascades* comptent 4096 étapes. La variante de la soft-cascade ACF pour la base Caltech utilise Adaboost continue dans son apprentissage. Les classificateurs faibles ne renvoient pas de réponse binaire mais continue. Par conséquent, comme les méthodes proposées nécessitent une réponse du classificateur faible binaire. Ils ne peuvent pas être appliqués directement. Nous proposons donc une variante de la *soft-cascade* pour Caltech en utilisant plutôt AdaBoost discret, appelé ACF_DISC. La base d'apprentissage pour les classificateurs faibles est composée de 24498 échantillons positifs et de 100000 échantillons négatifs, extraits des images d'apprentissage. La base de calibration pour l'ajustement des seuils est composé de 1996 échantillons positifs et de 54413 échantillons négatifs, extraits des images de test. Le nombre d'échantillons étant trop important pour appliquer directement notre approche, nous effectuons au préalable un partitionnement des échantillons négatifs comme décrit dans la section 2.2.4 pour retenir 16120 échantillons négatifs, ce qui a permis d'appliquer notre approche. L'algorithme DBP a été exécuté sur la base de calibration initiale, tandis que la base de calibration partitionnée

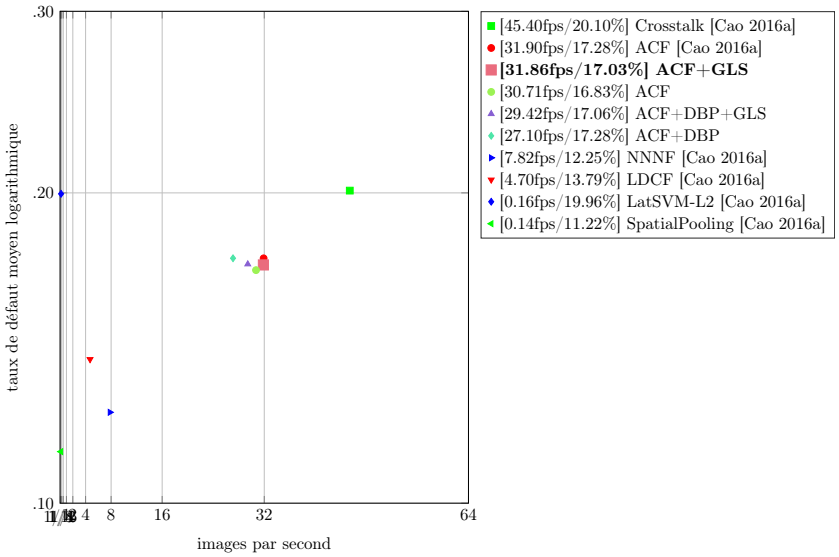


FIGURE 2.16 – Comparaison des performances de détection par rapport aux images par seconde sur la base INRIA avec [Cao 2016a].

est utilisée pour notre approche composée. Ainsi, quatre variantes de *soft-cascades* sont testées, appelées ACF_DISC, ACF_DISC+CLUS+GLS, ACF_DISC+DBP et ACF_DISC+CLUS+DBP+GLS. Dans ce cas, nous relevons un temps de calcul moyen de la méthode composée qui équivaut approximativement à 10 heures pour l'ajustement des seuils après apprentissage.

Les évaluations par fenêtre sont effectuées sur la base de calibration complète. La courbe ROC, figure 2.17, montre les performances des quatre variantes de *soft-cascade* dans la base de calibration ; celles-ci sont similaires pour les quatre variantes.

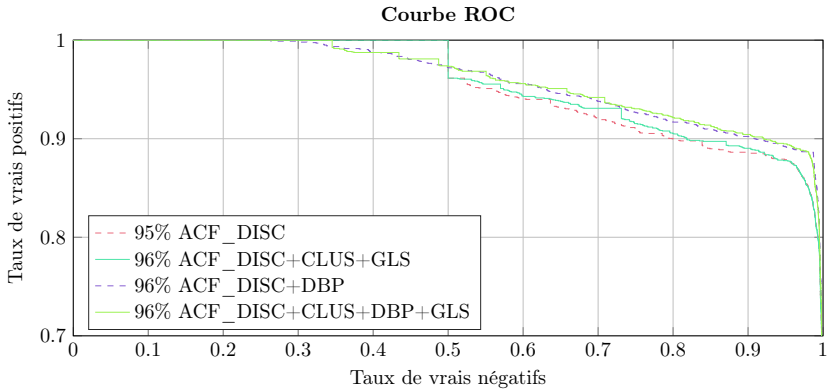


FIGURE 2.17 – La courbe ROC des évaluations par fenêtre sur la base Caltech.

La figure 2.18 détaille les évaluations du temps moyen de détection par fenêtre. Nous observons que les variantes de la méthode composée (ACF_DISC+CLUS+GLS et ACF_DISC+CLUS+DBP+GLS) ont un temps de

réponse moyen plus faibles que leurs variantes originales respectives (ACF_DISC et ACF_DISC+DBP), environ 13.54%.

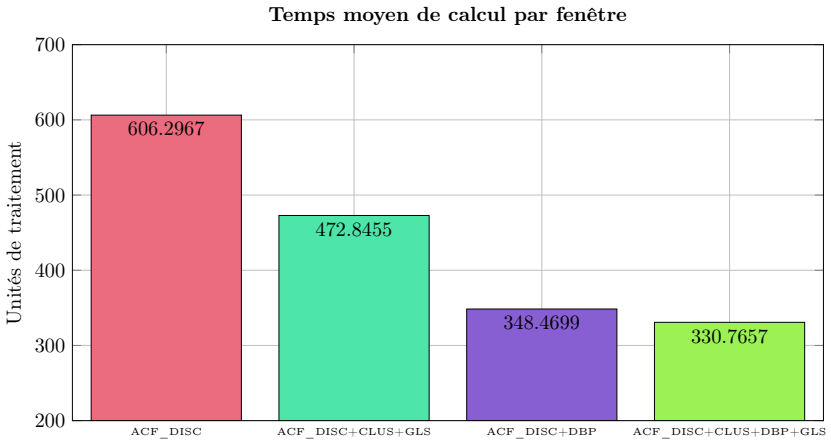


FIGURE 2.18 – Valeur de la fonction objectif, c'est-à-dire le temps moyen de calcul des évaluations par fenêtre sur la base Caltech.

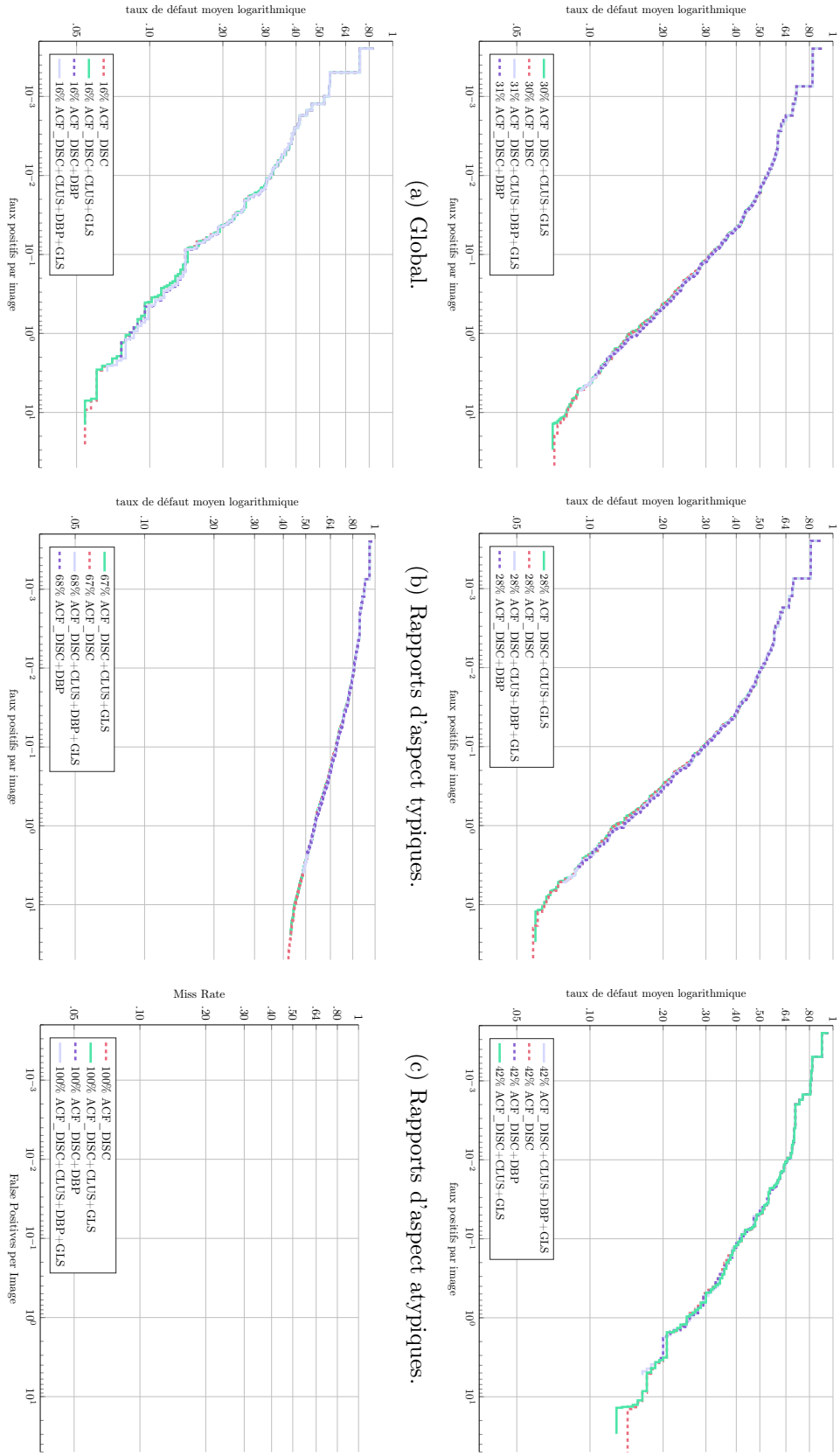
De même, des évaluations par images sont réalisées à partir des 4024 images de test de la base Caltech. Les courbes des évaluations FPPI, figure 2.19, illustrent les performances des quatre variantes de *soft-cascades* dans les images de la base de test. Ces évaluations sont cohérentes avec celles de la figure 2.17, montrant que nos variantes de la *soft-cascade* ont les mêmes performances que la *soft-cascades* ACF_DISC. On note que la *soft-cascade* ACF originale [Dollár 2012], qui est entraîné en utilisant AdaBoost continue, a une performance ici légèrement meilleure.

La table 2.2 détaille le taux d'images par seconde par variante de la *soft-cascade*. Ici encore, nous observons que la *soft-cascade* ACF_DISC+CLUS+GLS traite 3.03 fois plus d'images par seconde que la *soft-cascade* ACF_DISC. Le même constat est observé pour la *soft-cascade* ACF_DISC+CLUS+DBP+GLS qui traite 1.93 fois plus d'images par seconde que la *soft-cascade* ACF_DISC+DBP.

Détecteur	Images par seconde	Taux de défaut moyen logarithmique
ACF_DISC	6.85	**32.50**%
ACF_DISC+CLUS+GLS	**9.88**	32.52%
ACF_DISC+DBP	7.92	32.99%
ACF_DISC+CLUS+DBP+GLS	9.85	32.84%

TABLE 2.2 – Comparaison des performances de détection par rapport aux images par seconde sur la base Caltech.

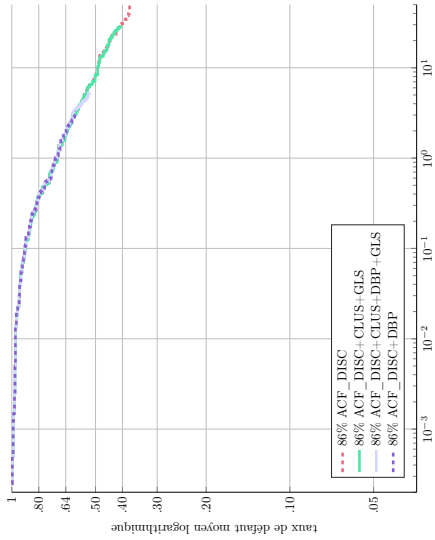
La figure 2.20 confronte nos résultats avec ceux rapportés par Cao *et al.* [Cao 2016a] et Zhang *et al.* [Zhang 2016a]. Soulignons que ces évaluations sont menées sur des architectures matérielles plus puissantes que les nôtres (CCF, CompACT-Deep et RPN+BF [Zhang 2016a]) même en utilisant des architectures GPU. Néanmoins, nous observons que nos détecteurs sont compétitifs avec les dé-



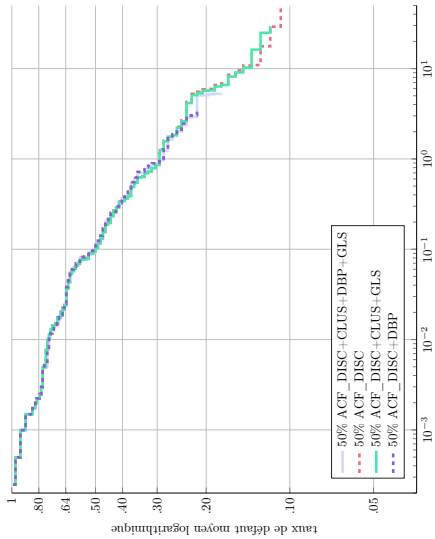
(a) Global. (b) Rapports d'aspect typiques. (c) Rapports d'aspect atypiques.

(d) Faible distance. (e) Moyenne distance. (f) Grande distance.

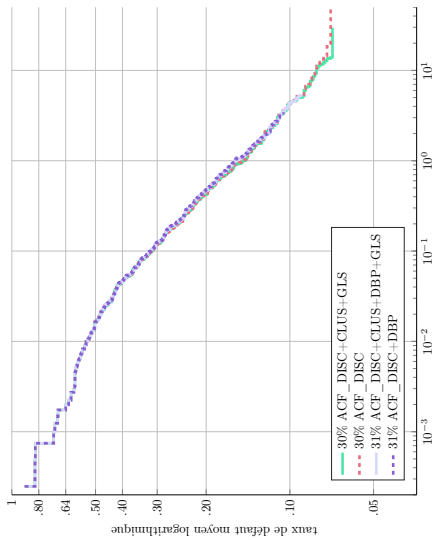
FIGURE 2.19 – Évaluations FPPI détaillées sur la base Caltech.



(i) Forte occultation.



(h) Occultation partielle.



(g) Sans occultation.

FIGURE 2.19 – Évaluations FPPI détaillées sur la base Caltech (Cont.).

tecteurs les plus rapides de la littérature. Le détecteur Crosstalk a un meilleur temps de calcul de 14.10 images par seconde suivi de notre ACF_DISC+CLUS+GLS avec 9.88 images par seconde et ACF avec une moyenne de 9.49 images par seconde. Les détecteurs offrant le meilleur compromis entre temps de traitement ou réponse et performances de détection sont : Crosstalk, ACF_DISC+CLUS+GLS, LDCF et RPN+BF.

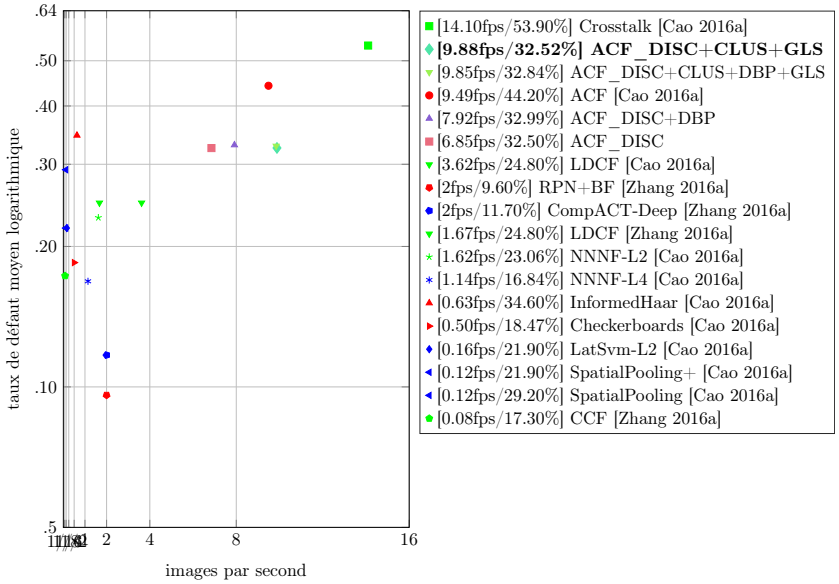


FIGURE 2.20 – Comparaison des performances de détection par rapport aux images par seconde sur la base Caltech avec [Cao 2016a] et [Zhang 2016a].

Les évaluations sur la base Caltech montrent également que le partitionnement effectué sur la base de calibration ne réduit pas les performances de détection finales et permet d'appliquer notre algorithme sans perte d'information.

2.3.2.3 Discussions associées

Les évaluations présentées dans les sections 2.3.2.1 et 2.3.2.2 démontrent que l'on peut améliorer le temps de réponse d'une *soft-cascade* sans diminuer les performances de détection. L'amélioration moyenne des images par seconde est de $6,15\%$ pour la base INRIA et de $34,32\%$ pour la base Caltech. Les différences de performances s'expliquent par les raisons suivantes. Les fenêtres glissantes dans la base INRIA ont une taille de 128×64 et la *soft-cascade* une longueur de 2048 étapes. Dans la base Caltech, la taille de la fenêtre est égale à 64×32 et la longueur de la *soft-cascade* est de 4096 étapes. Par conséquent, le mécanisme de fenêtres glissantes génère beaucoup plus d'échantillons pour la base Caltech que pour INRIA, ce qui, combiné à la longueur de la *soft-cascade*, tend à diminuer les images par seconde. Par conséquent, dans le cas de *soft-cascades* de grande taille, la *soft-cascade* a un espace de recherche plus riche permettant d'explorer les meilleures solutions ; notre approche est donc particulièrement efficace. Nos détecteurs sont

une bonne option pour garder le temps de calcul acceptable, surtout parce que nous dépassons la *soft-cascade* ACF classique pourtant performante en temps de calcul [Cao 2016a, Cao 2016b]. Malgré l'absence d'architecture GPU, nous restons compétitifs avec certains détecteurs d'apprentissage profond de la littérature qui proposent une moyenne de 2 images par seconde sur la base Caltech [Zhang 2016a] avec une architecture GPU Tesla K40.

2.4 Conclusions

Nous avons développé et évalué une nouvelle approche (notée MSCRMP). Celle-ci s'est avérée pertinente pour l'apprentissage des seuils de détecteurs *soft-cascade*. Nous avons donné une définition formelle de ce problème d'optimisation et prouvé qu'il était NP-difficile. Deux variantes PLNE sont proposées ; elles permettent de trouver un vecteur de seuils de détection optimal en utilisant des solveurs PLNE. Le problème consiste à trouver un chemin optimal à l'intérieur d'un graphe de seuils. Nous avons également fourni des propriétés de dominance qui permettent de réduire drastiquement ce graphe et l'espace de recherche.

L'originalité de notre approche est de chercher à optimiser le coût de traitement pour des performances de détection données. La démarche repose sur une recherche locale pour l'optimisation des seuils de détection qui considère une structure de voisinage définie comme une enveloppe autour d'un chemin. Cette approche permet d'envisager un détecteur de grande longueur et de traiter une base de calibration de grande taille. Elle peut également être utilisée comme une phase de post-traitement, lors de la conception de détecteurs *soft-cascade* traditionnels, dans le but d'améliorer les performances CPU. De plus, il est indépendant du type de descripteur ou du contexte de l'application, la seule restriction étant d'utiliser un classificateur *soft-cascade* avec des classificateurs faibles binaires. Sur les bases de données considérées, nos détecteurs offrent de meilleures performances en comparaison avec d'autres détecteurs *soft-cascade* de la littérature, en terme de temps de réponse moyen.

Notre approche est extensible à des descripteurs hétérogènes ; il conviendrait alors de considérer le coût CPU propre à chaque descripteur dans notre modélisation. Celle-ci pourrait également gérer l'ordre des descripteurs sélectionnés et ainsi améliorer encore les performances.

Les travaux présentés ici ont donné lieu à trois publications : *International Conference on Operations Research and Enterprise Systems* (ICORES) [BarbosaAnda 2016a], congrès annuel de la société Française de Recherche Opérationelle et d'Aide à la Décision (ROADEF) [BarbosaAnda 2016b] et *IEEE Winter Conference on Applications of Computer Vision* (WACV) [BarbosaAnda 2018a].

Les détecteurs développés peuvent constituer l'une des briques de base dans un système plus complet de ré-identification de personnes dans un réseau de caméras. Cette problématique est traitée dans la seconde partie du mémoire.

Deuxième partie

Ré-identification de personnes dans un réseau de caméras avec des considérations de temps de transit et de topologie de réseau

Chapitre 3

État de l'art et positionnement des travaux

Sommaire

La ré-identification (ré-ID) de personnes vise à reconnaître les piétons en transit dans un environnement humain à travers leurs détections et appariements dans les vues des caméras ambiantes instrumentant cet environnement. Ces caméras en réseau ont en général des champs de vues (FOV) disjoints afin de limiter l'instrumentation des lieux en nombre de caméras. Cette tâche de vidéo surveillance est motivée par : (1) la demande croissante de sécurité publique eu égard au contexte actuel, et (2) le déploiement massif de systèmes de vidéo surveillance dans les maisons de retraite, les campus universitaires, les rues, etc.

Classiquement, la plupart des approches de ré-ID [Zheng 2016b, Karanam 2018] de la communauté Vision par Ordinateur raisonnent sur des paires de caméras. Le principe est alors de déduire si deux détections (issues d'un détection générique de personnes, *Cf.* travaux décrits dans la partie I de ce manuscrit), c'est-à-dire deux sous-images segmentant les personnes et provenant de deux caméras en réseau, correspondent ou non à la même identité (ID) de personnes. La ré-ID globale est obtenue par la relation directe entre les ré-ID locales pour chaque personne. Ce raisonnement entre paires de caméras induit irrémédiablement des incohérences de ré-ID au niveau réseau : par exemple, une même personne ré-identifiée dans plusieurs caméras simultanément, ou avec des personnes différentes dans chaque caméra. Par conséquent, les méthodes de ré-ID par paires ne permettent pas d'obtenir des résultats d'appariements d'identités optimaux (cohérents) pour l'ensemble du réseau de caméras.

À partir de ce constat, certaines approches [Das 2014, Chakraborty 2016, Lin 2017] proposent de raisonner simultanément sur l'ensemble du réseau de caméras afin de renforcer la cohérence globale des identités ; cette stratégie permet ainsi de lever des ambiguïtés. Le principe classique de la ré-ID cohérente en réseau (*Network Consistent Re-identification*, notée NCR) vise alors à optimiser un coût global d'association de paires, compte tenu de contraintes de cohérence d'identités entre l'ensemble des caméras du réseau.

Ce chapitre vise à positionner et motiver nos choix pour notre approche de ré-ID (notée D-NCR pour Directed Network Consistent Re-identification) de personnes ; celle-ci sera formalisée et évaluée dans le chapitre suivant. Ainsi, la section 3.1 catégorise les principales approches existantes, puis présente les différentes bases de données publiques (*benchmarks*) et les métriques d'évaluation qui seront exploitées ultérieurement. Nous discutons alors nos choix (section 3.2). Le chapitre se conclut par une petite synthèse (section 3.3).

3.1 État de l'art

Initialement, les techniques de ré-ID de personnes, à l'instar du suivi multi-caméras, reposent sur des modèles d'apparence des personnes cibles et un étalonnage géométrique préalable des caméras disjointes. Huang et Russell [Huang 1997] ont proposé une formulation bayésienne pour estimer la prédiction de l'apparence des cibles dans une caméra, eu égard aux observations dans les caméras précédentes. Le modèle d'apparence inclut des caractéristiques telles que la couleur, la taille, la vélocité et l'horodatage de l'observation. Le concept de ré-ID de personnes a été initié par Zajdel *et al.* [Zajdel 2005]. Le principe est de "ré-identifier une personne quand elle quitte le champ de vue puis rentre". Gheissari *et al.* [Gheissari 2006] dissocient plus nettement ré-ID de personnes et suivi multi-caméra, la ré-ID devenant une problématique spécifique. Initialement prévu pour le suivi de personnes dans les vidéos, la plupart des travaux de ré-ID se concentrent à présent sur l'appariement d'images.

Plus tard, la problématique de ré-ID a été étendue au traitement vidéo donc au *multi-shot*, c'est-à-dire avec plusieurs images par personne ; citons ici les travaux clés de [Bazzani 2010, Farenzena 2010]. Il est alors montré que l'utilisation de plusieurs images par cible améliore la stratégie dite *one shot* (une seule image analysée) même si les performances de ré-ID plafonnent à mesure que le nombre d'images considérées augmente. Le succès de l'apprentissage profond en Vision par Ordinateur s'applique également à la ré-ID [Yi 2014, Li 2014]. Bien que les performances obtenues ne soient pas encore stables sur les petites bases de données, les méthodes d'apprentissage profond deviennent une alternative pour la ré-ID.

Nombreux travaux, afin de découpler les problèmes, privilégient une segmentation manuelle et non automatique (via un détecteur) des cibles. Néanmoins, Xu *et al.* [Xu 2014] montrent qu'il est souhaitable de caractériser les performances du système complet (détection, suivi et ré-ID). Il est ainsi montré que la prise en

compte/caractérisation simultanée des briques séquencées détection (voire suivi) et ré-ID améliore les performances globales de ré-ID car le système est alors optimisée sur les vraies données (boîtes englobantes de détection, etc.).

Les stratégies courantes de ré-ID peuvent donc se classifier en deux catégories : image et vidéo [Zheng 2016b]. Les méthodes basées sur la vidéo [Karanam 2015] focalisent sur une analyse *multi-shot* afin de réduire l'impact des occultations, des variations d'illumination, et des artefacts de détecteur. Nous décrivons quelques travaux basés sur chacune de ces stratégies.

3.1.1 Ré-ID de personnes par paires d'images

Classiquement, la ré-ID considère des paires d'images. Considérons une banque $\mathcal{G}$ de N images, dénotées g_i pour $i \in \{1, \ldots, N\}$. A chaque image est associée une ID préalable de personne. Étant donnée une nouvelle image, son ID est trouvée par l'équation (3.1), où i^* est l'identité de l'image non-identifiée q et sim est une métrique de similarité. La modalité de ré-ID se caractérise donc par un descripteur (signature) de chaque boite englobante détectée et une métrique de distance/similarité entre les deux sous-images q et g_i. La figure 3.1 illustre le principe.

$$i^* = \arg \max_{i \in \{1,\ldots,N\}} \text{sim}\,(q, g_i) \tag{3.1}$$

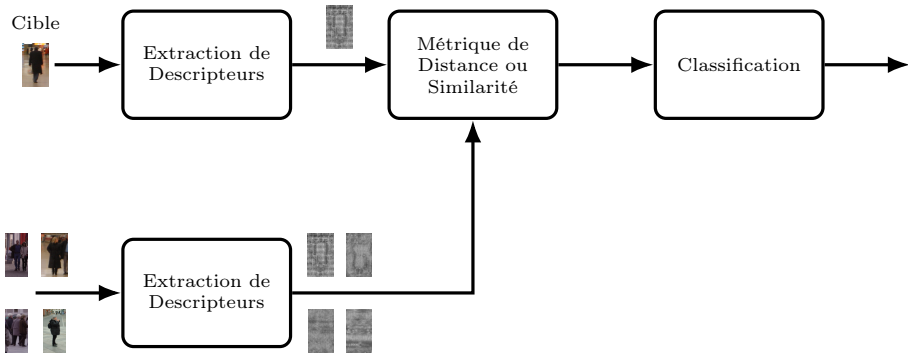


FIGURE 3.1 – Synoptique pour la ré-ID de personnes sur paires d'images.

3.1.1.1 Descripteurs

La signature courante repose davantage sur la couleur des cibles que sur la texture. Gheissari *et al.* [Gheissari 2006] proposent une méthode de segmentation spatio-temporelle pour dissocier le premier plan (incluant a priori la personne) de la scène. Pour chaque région, un histogramme de couleur HS (pour *Hue-Saturation*) et un histogramme de segments sont calculés. Gray et Tao [Gray 2008] utilisent 8 canaux de couleur (RGB, HS et YCbCr) et 21 filtres de texture sur le canal de luminance, et la région d'intérêt (extraite par détection) est partitionnée en bandes horizontales. Ce descripteur est appelé *ensemble of localized features* (ELF).

Farenzena *et al.* [Farenzena 2010] dissocient le premier plan de l'arrière-plan afin de limiter l'influence de ce dernier dans le descripteur. Ils calculent alors un axe symétrique pour chaque partie corporelle. Puis, l'histogramme de couleur HSV, les *maximally stable color regions* (MSCR), et les *recurrent high-structured patches* (RHSP) sont calculés. Ce vecteur de caractéristiques est dénommé *symmetry-driven accumulation of local features* (SDALF).

De même, Mignon et Jurie [Mignon 2012] construisent un vecteur de caractéristiques à partir des histogrammes des canaux couleur RGB, YUV et HSV et le descripteur de texture LBP (pour *Local Binary Pattern*) en bandes horizontales. Ma *et al.* [Ma 2012] codent des descripteurs de pixels locaux comprenant l'information spatiale de pixel, l'intensité et des informations de gradient dans une représentation de vecteur de Fisher. Ce descripteur est appelé LDFV.

Les descripteurs ont évolué ces dernières années. Zhao *et al.* [Zhao 2013b, Zhao 2013a, Zhao 2014] extraient l'histogramme de couleurs CIELAB à 32 dimensions et le descripteur *scale-invariant feature transform* (SIFT) [Lowe 1999, Lowe 2004] à 128 dimensions. Ce descripteur est appelé DenseColorSIFT. Li *et al.* [Li 2013c] extraient également des descripteurs de couleur locaux. Pedagadi *et al.* [Pedagadi 2013] extraient les histogrammes et les moments de couleur des espaces couleur HSV et YUV. Das *et al.* [Das 2014] appliquent l'histogramme de couleur HSV sur la tête, le torse et les jambes de la silhouette proposée en [Bazzani 2010]. Liu *et al.* [Liu 2014] extraient l'histogramme de couleur HSV, l'histogramme de gradient et l'histogramme LBP. Yang *et al.* [Yang 2014] introduisent un descripteur noté *salient color names based color descriptor* (SCNCD) pour une description couleur globale de personne. Xiong *et al.* [Xiong 2014] calculent des histogrammes de couleurs dans les espaces de couleur RGB, YCbCr et HS et des descripteurs de texture LBP. Ce descripteur est appelé HistLBP. Liao *et al.* [Liao 2015] proposent le descripteur *local maxima occurrence* (LOMO). Matsukawa *et al.* [Matsukawa 2016] proposent un descripteur gaussien hiérarchique pour décrire les indices de couleur et de texture. Ce descripteur est appelé *Gaussian of Gaussian* (GOG).

En marge des descripteurs bas niveau couleur et/ou texture, une alternative est de privilégier des descripteurs utilisant des attributs dits de niveau intermédiaire, supposés plus robustes à l'analyse d'image. Ainsi, Layne *et al.* [Layne 2012] annotent 15 attributs binaires sur la base de données VIPeR liés à la tenue vestimentaire et à la biométrie douce (shorts, jupes, sandales, sac à dos, jeans, logo, encolure en V, vêtements de plein air, rayures, lunettes de soleil, écouteurs, cheveux longs, cheveux courts, sexe, objet de transport). Liu *et al.* [Liu 2012b] améliorent le modèle d'allocation de Dirichlet latent (LDA) en utilisant des attributs annotés pour filtrer les sujets LDA bruyants. Liu *et al.* [Liu 2012a] proposent de caractériser des prototypes de personnes avec des attributs communs de manière non supervisée et de déterminer de manière adaptative les poids des caractéristiques de différentes personnes non-identifiés en fonction des prototypes. Ma *et al.* [Ma 2014] codent des descripteurs biologiquement inspirés à échelles multiples en utilisant des descripteurs de covariance. Ce descripteur est appelé gBiCov. Su *et al.* [Su 2015] incorporent les attributs sémantiques binaires de chaque cible et issus de caméras

Descripteur	Publication
ELF	ECCV 2008 [Gray 2008]
SDALF	CVPR 2010 [Farenzena 2010], CVIU 2013 [Bazzani 2013]
WACN	CVPR 2012 [Martinel 2012]
ICT	ECCV 2012 [Avraham 2012]
LDFV	ECCV Workshops 2012 [Ma 2012]
DenseColorSIFT	CVPR 2013 [Zhao 2013b], ICCV 2013 [Zhao 2013a], CVPR 2014 [Zhao 2014]
SCNCD	ECCV 2014 [Yang 2014]
HistLBP	ECCV 2014 [Xiong 2014]
gBiCov	IVC 2014 [Ma 2014]
LOMO	CVPR 2015 [Liao 2015]
WHOS	T-PAMI 2015 [Lisanti 2015]
GOG	CVPR 2016 [Matsukawa 2016]

TABLE 3.1 – Résumé de descripteurs pour la ré-ID de personnes par image [Zheng 2016b].

différentes dans un espace d'attribut continu de rang inférieur. Shi *et al.* [Shi 2015] proposent d'apprendre un certain nombre d'attributs, y compris la couleur, la texture et des étiquettes de catégorie.

Une synthèse, tirée de [Zheng 2016b], des types de descripteurs est présentée Table 3.1

3.1.1.2 Métriques de distance/similarité

Considérant les descripteurs décrits précédemment, la distance euclidienne (l_2) est souvent utilisée pour classer les individus. La règle de décision consiste alors à apparier la cible avec l'identité ayant la distance à elle minimale. Intégrer l'apprentissage supervisée en utilisant des données d'apprentissage conduit à des performances supérieures, ce qui est l'objectif de l'apprentissage de métriques. La métrique privilégiée doit, autant que possible, minimiser la distance intra-classe entre vecteurs de descripteurs associée à un même ID et maximiser la distance inter-classe de ces vecteurs associée à des IDs différents. Citons ici la distance de Mahalanobis, qui généralise la distance euclidienne et permet de prendre en considération la distribution des données. La distance entre les vecteurs x_i et x_j s'écrit alors comme dans l'équation (3.2), où $\mathbf{M}$ est la matrice de covariance sur les vecteurs de descripteurs, matrice semi-définie positive.

$$d(x_i, x_j) = (x_i - x_j)^{\mathrm{T}} \mathbf{M} (x_i - x_j) \tag{3.2}$$

Sur la base de l'équation (3.2), d'autres méthodes d'apprentissage de métriques ont été déclinées. Certaines métriques reposent sur une analyse des distances des k plus proches voisins (k-PPV) de chaque vecteur descripteur. Weinberger *et al.* [Weinberger 2006] proposent la méthode *large margin nearest neighbor* (LMNN) qui établit un périmètre pour le voisin cible (paires appariées) et pénalisent ceux qui envahissent le périmètre (imposteurs). Davis *et al.* [Davis 2007] proposent la méthode nommée *information-theoretic metric learning* (ITML) comme compromis

Métrique	Publication
l_2	
LMNN	NIPS 2006 [Weinberger 2006], JMLR 2009 [Weinberger 2009]
ITML	ICML 2007 [Davis 2007]
kMFA	T-PAMI 2007 [Yan 2007], IJCAI 2015 [Chen 2015]
PCCA	CVPR 2012 [Mignon 2012]
kPCCA	CVPR 2012 [Mignon 2012]
KISSME	CVPR 2012 [Köstinger 2012]
LFDA	CVPR 2013 [Pedagadi 2013]
rPCCA	ECCV 2014 [Xiong 2014]
kLFDA	ECCV 2014 [Xiong 2014]
XQDA	CVPR 2015 [Liao 2015]

TABLE 3.2 – Résumé de métriques de similarité pour la ré-ID de personnes par paire d'images [Zheng 2016b].

entre la satisfaction des contraintes de similarité des vecteurs descripteurs et l'assurance que la métrique apprise est proche de la fonction de distance initiale. Une métrique populaire, notée *Keep It Simple and Straightforward MEtric* (KISSME) [Köstinger 2012] ; également basée sur l'équation (3.2), utilise un test de rapport de vraisemblance. Il est ainsi montré que la distance de Mahalanobis peut être dérivée naturellement du test du rapport de vraisemblance logarithmique.

Outre le choix de la métrique, certains travaux se concentrent sur l'apprentissage de sous-espaces discriminants. Mignon et Jurie [Mignon 2012] proposent le *pairwise constrained component analysis* (PCCA) qui apprend une fonction de transformation linéaire pour pouvoir travailler directement dans des espaces de grande dimension. Pedagadi *et al.* [Pedagadi 2013] combinent séquentiellement l'analyse en composantes principales (ACP) non supervisée et l'analyse discriminante locale de Fisher (LFDA) supervisée qui préserve la structure de voisinage locale. Xiong *et al.* [Xiong 2014] proposent une version améliorée de deux méthodes de projection sous-spatiales existantes, c'est-à-dire rPCCA et kLFDA. Liao *et al.* [Liao 2015] proposent d'apprendre la projection à un sous-espace de faible dimension avec des données croisées similaires à l'analyse discriminante linéaire (LDA). Puis, une fonction de distance est apprise dans le sous-espace résultant en utilisant KISSME. Cet métrique est appelée XQDA.

Un résumé des types de métriques, issu de [Zheng 2016b] est présenté dans la Table 3.2.

3.1.1.3 Méthodes par apprentissage profond

Récemment, les méthodes d'apprentissage profond ont été appliquées à cette problématique. Le principal verrou de l'apprentissage profond pour la ré-ID (et plus largement) est le manque de données d'apprentissage. Pour cette raison, la plupart des méthodes d'apprentissage profond pour la ré-ID se concentrent sur le

modèle siamois. La structure de ce réseau de neurones convolutionnels permet de comparer deux entrées. Par exemple, dans Yi *et al.* [Yi 2014], une image d'entrée est partitionnée en trois parties horizontales qui se chevauchent, et les parties traversent deux couches convolutives plus une couche entièrement connectée qui les fusionne et produit un vecteur descripteur de cette image. L'architecture proposée par Li *et al.* [Li 2014] diffère par le fait qu'une couche supplémentaire de correspondance de paires d'images est ajoutée.

Une limitation du modèle siamois est de ne pas utiliser pleinement les annotations de ré-ID. En effet, le modèle siamois ne considère que les identités entre paires d'images, c'est-à-dire il prend en compte une identité a chaque fois. Une autre stratégie potentiellement efficace consiste à utiliser un mode de classification utilisant pleinement tous les identités. Xiao *et al.* [Xiao 2016] rassemblent les identités provenant de plusieurs bases de données pour constituer une base large échelle d'apprentissage et utilisent une fonction coût type softmax dans le réseau. Cette intégration conduit à de bonnes performances de ré-ID.

Les réseaux précités encodent le système complet, c'est-à-dire descripteur, métrique et règle de décision. Citons des alternatives qui prennent des descripteurs bas niveau en entrée. Ainsi, Wu et Hengel [Wu 2017] regroupent des descripteurs bas niveau (par exemple SIFT, histogrammes de couleurs, etc.) en un seul vecteur de Fisher pour chaque imagette. Le réseau hybride construit des couches entièrement connectées sur les vecteurs Fisher d'entrée et utilise LDA comme fonction objectif afin de favoriser une faible variance intra-classes et une forte variance inter-classes.

Certes, ces modèles sont très prometteurs mais nécessitent l'utilisation d'architectures matérielles spécifiques, notablement des cartes graphiques puissantes. Ils ont besoin aussi de bases de données de grande taille pour l'apprentissage, ce qui n'est pas toujours réaliste pour le cas de la ré-ID de personnes.

3.1.2 Ré-ID de personnes par vidéo

Au cours des dernières années, la ré-ID par vidéo est devenue populaire en raison de la richesse de l'information sous-jacente. Elle favorise un appariement *multi-shot* et une gestion d'informations temporelles. La formulation peut alors s'apparenter à celle de la ré-ID par paires d'images. Ainsi, la ré-ID par vidéo remplace les images q et g_i par deux ensembles d'images $\mathbf{q}$ et $\mathbf{g}_i$, possiblement de tailles différentes. Cette formulation est exprimée par l'équation (3.3), où N est le nombre de séquences de suivi d'un personne (*tracklets*) dans la galerie.

$$i^* = \arg \max_{i \in \{1,\dots,N\}} \text{sim}\left(\mathbf{q}, \mathbf{g}_i\right) \tag{3.3}$$

[Bazzani 2010, Farenzena 2010] ont ainsi été les précurseurs en 2010. Ils utilisent essentiellement des descripteurs de couleur et segmentent le premier plan pour isoler la personne de l'image. Ils utilisent des descripteurs de cibles inspirés des méthodes de ré-ID par paire d'images ; la différence majeure réside donc dans la fonction d'appariement. Les deux méthodes calculent généralement la distance euclidienne

minimale entre deux ensembles de boîtes englobantes en tant que similarité de l'ensemble. Karanam *et al.* [Karanam 2015] utilisent les séquences d'images pour caractériser une personne comme une combinaison linéaire des instances de la même personne.

Les méthodes précitées créent des modèles d'apparence basés sur du *multi-shot*. La tendance récente consiste à incorporer des descripteurs spatio-temporels dans le modèle comme proposé par Wang *et al.* [Wang 2014] incluant HOG3D [Klaeser 2008] pour l'information spatial et *gait energy image* [Han 2006] pour l'information temporelle.

Le choix de la métrique de distance est également vital lors de l'appariement de vidéos. Zheng *et al.* [Zheng 2012] proposent une méthode de ré-ID basée réseau de neurones qui infère si la cible correspond à l'une des images appartenant à la même identité. Ce réseau de neurones subit un premier apprentissage sur une base d'apprentissage générique puis un ré-apprentissage sur la base d'apprentissage d'intérêt spécifique, cet approche d'apprentissage est connu comme *transfer learning*. Li *et al.* [Li 2013c] proposent une extension au *multi-shot* du modèle de correspondance local qui minimise la distance des paires les mieux adaptées et réduit le nombre de transformations croisées. Zhu *et al.* [Zhu 2016] proposent l'apprentissage simultané de métriques de distance intra-vidéo et inter-vidéo pour rendre la représentation vidéo plus compacte et discriminer les vidéos de différents IDs. You *et al.* [You 2016] proposent la méthode d'apprentissage à distance *top-push* qui optimise l'appariement de premier rang dans la ré-ID par vidéo en sélectionnant des descripteurs discriminants.

Les méthodes d'apprentissage profond sont aussi considérées pour la ré-ID par vidéo. Le volume de données est alors généralement plus grand que dans la ré-ID par image, car chaque tracklet inclut un ensemble d'images. Une différence fondamentale entre la ré-ID par vidéo et par image est qu'avec plusieurs images pour chaque identité, une stratégie multi-appariements ou une stratégie à correspondance unique, c'est-à-dire après concaténation de la vidéo en un seul vecteur de descripteurs, doit être utilisée. La stratégie multi-appariements est utilisée dans des travaux anciens [Bazzani 2010, Farenzena 2010] et induit des coûts de calcul élevés, ce qui peut poser problème sur des bases de données volumineuses car la combinatoire devient trop importante. Au contraire, le regroupement des descripteurs par image en un vecteur global favorise un meilleur passage à l'échelle. Par conséquent, les méthodes actuelles de ré-ID par vidéo impliquent généralement une étape de regroupement. Cette étape peut être un regroupement maximal ou moyen des descripteurs [Zheng 2016a, McLaughlin 2016], ou apprise par une couche entièrement connectée [Yan 2016]. Une autre bonne pratique consiste à injecter des informations temporelles dans la représentation finale [McLaughlin 2016, Yan 2016, Wu 2016] pour aider le réseau de neurones à différencier des personnes d'apparences visuelles similaires.

3.1.3 Méthodes d'appariement

Comme évoqué, dans le cas d'appariement par paires, la règle de décision pour la ré-ID classique communément utilisée repose sur l'équation (3.1) pour la ré-ID par image, et l'équation (3.3) pour la ré-ID par vidéo. Afin de résoudre le problème global d'appariement, une heuristique glouton est classiquement utilisée. Elle consiste à apparier une à une chaque cible à l'identité la plus similaire (ou celle avec la plus petite distance de similarité) et de itérer pour chaque paire de caméras. Les personnes non-identifiées de la première caméra sont ainsi les cibles q et les personnes identifiées de la deuxième caméra sont la galerie $\mathcal{G}$.

3.1.3.1 Ré-identification de personnes avec contraintes inhérentes au réseau

L'un des principaux problèmes liés à l'algorithme glouton précédent vient du fait que l'appariement par paire prise séparément peut être incompatible avec l'ensemble du réseau de caméras. Un exemple est décrit sur la figure 3.2. Nous y observons un appariement par paires dans un réseau à 3 caméras, les correspondances étant représentées en trait plein. Comme on peut le voir, les correspondances entre les caméras 1-2 et 1-3 sont correctes, mais il y a une erreur entre les caméras 2-3 pour la même personne.

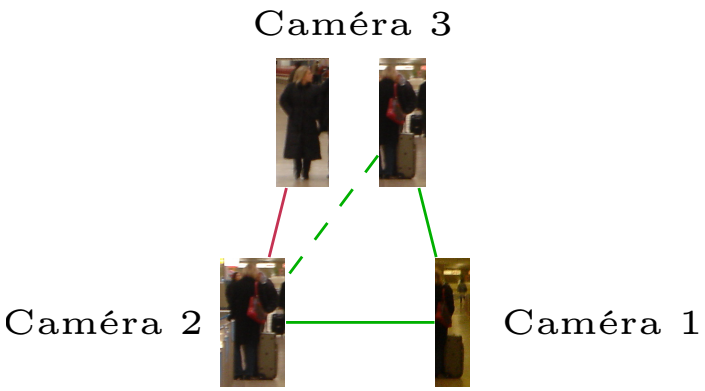


FIGURE 3.2 – Exemple de cohérence du réseau dans la ré-identification. Une correspondance est représentée par une ligne continue. La couleur verte indique une correspondance correcte et la couleur rouge indique une mauvaise correspondance. La ligne pointillée représente la correspondance cohérente.

Pour résoudre ce problème, Das *et al.* [Das 2014] proposent une stratégie globale appelée *Network Consistent Re-identification* (NCR), également connue sous le nom de *Network Consistent Data Association* (NCDA) [Chakraborty 2016]. Les auteurs proposent de solutionner l'appariement au niveau réseau donc simultanément sur l'ensemble des paires d'images, via la résolution d'un programme linéaire en nombres entiers (PLNE). Ce PLNE maximise la similarité globale dans le réseau, et non la similarité individuelle par paire de caméras, et ajoute des contraintes de cohérence

de boucle entre chaque triplet de caméras du réseau, à l'instar de l'exemple présenté. Ce modèle est détaillé au chapitre 4.

Plus récemment, Lin *et al.* [Lin 2017] combinent ce concept de cohérence réseau avec un approche par apprentissage profond. Un réseau de neurones profond est utilisé pour caractériser les descripteurs. Au lieu d'utiliser un PLNE pour l'optimisation globale, une optimisation par descente de gradient est proposée. Cela permet de rétro-propager l'erreur de ré-ID finale dans le réseau de neurones profond de sorte que les descripteurs calculés prennent en compte la cohérence du réseau. Les auteurs affirment que leur modèle d'optimisation peut être utilisé avec n'importe quel descripteur en entrée, à l'instar de la stratégie NCR.

Ces optimisations globales au niveau réseau garantissent des gains en performances de ré-ID par rapport à la stratégie classique de Ré-ID successives sur paire d'images.

3.1.4 Bases de données pour la ré-identification

Nous listons brièvement dans cette partie les différentes bases publiques d'images pour la ré-ID de personnes. La table 3.3 en fait une synthèse. Nous décrivons chacune de ces bases de données via les caractéristiques suivantes : multi-shot (MS), séquences de suivi (SS), disponibilité des images complètes (IC), variation du point de vue (VP), variations d'illumination (VI),erreurs de détection (ED), occultations (OCC), complexité arrière plan (EF), images basse résolution (RES), faux positifs (FP) et topologie de la réseau (TR). Nous indiquons également le nombre d'images ou de boîtes englobantes et le nombre de caméras dans chaque base d'images, ainsi que les stratégies d'annotation des boîtes englobantes pour constituer la vérité terrain : annotation manuelle ou via un détecteur. Cette vérité terrain est cruciale pour évaluer les performances de ré-ID.

La base VIPeR [Gray 2008] se compose de 632 personnes observées à partir de deux vues disjointes. Chaque personne n'a qu'une image par vue. VIPeR inclut des variations de point de vue et des variations d'illumination. GRID [Loy 2010, Loy 2013] offre 250 paires d'images issues de 8 caméras disjointes. Pour accroître le réalisme du scénario, 775 personnes sans correspondances sont incluses dans la galerie, ce qui rend cette base de données extrêmement difficile. GRID inclut des variations de point de vue, d'encombrement de fond, d'occultations et des images de faible résolution.

CAVIAR4ReID [Cheng 2011] considère deux caméras dans un centre commercial. CAVIAR4ReID inclut des variations de point de vue et d'images à basse résolution. Dans le cas de 3DPeS [Baltieri 2011], la communauté de ré-ID utilise un ensemble d'instantanés sélectionnés au lieu de la vidéo d'origine, qui comprend 192 personnes et 1011 images. 3DPeS inclut des variations du point de vue et des variations d'illumination. PRID2011 [Hirzer 2011] est construite à partir de deux caméras extérieures, avec 385 séquences de suivi d'une caméra et 749 séquences de suivi de l'autre caméra. PRID2011 inclut des variations de point de vue et de variations d'illumination. V47 [Wang 2011] contient 47 personnes en situation de marche ob-

Base de données	Publication	Nb de personnes	Nb de caméras	Nb d'images	Annotation	Taille d'image	Caractéristiques
VIPeR	ECCV 2008 [Gray 2008]	632	2	1264	manuelle	128x48	VP,VI
ETH1,2,3	SIBGRAPI 2009 [Schwartz 2009a]	85,35,28	1	8580	manuelle	varie	MS,SS,IC
QMUL iLIDS	BMVC 2009 [Zheng 2009]	119	2	476	manuelle	varie	MS
GRID	IJCV 2010 [Loy 2010],ICIP 2013 [Loy 2013]	1025	8	1275	manuelle	varie	VP,EF,OCC,RES
CAVIAR4ReID	BMVC 2011 [Cheng 2011]	72	2	1220	manuelle	varie	MS,VP,RES
3DPeS	J-HGBU 2011 [Baltieri 2011]	192	8	1011	manuelle	varie	MS,IC,VP,VI
PRID2011	SCIA 2011 [Hirzer 2011]	934	2	24541	manuelle	128x64	MS,SS,IC,VP,VI
V47	ICCVW 2011 [Wang 2011]	47	2	752	manuelle	varie	MS,IC
WARD	CVPR 2012 [Martinel 2012]	70	3	4786	manuelle	128x48	MS,SS,VI
SAIVT-Softbio	DICTA 2012 [Bialkowski 2012]	152	8	64472	manuelle	varie	Ms,SS,IC,VP,VE,EF
CUHK01	ACCV 2012 [Li 2013b]	971	2	3884	manuelle	160x60	MS,VP,OCC
CUHK02	CVPR 2013 [Li 2013a]	1816	10 (5 paires)	7264	manuelle	160x60	MS,VP,OCC
CUHK03	CVPR 2014 [Li 2014]	1467	10 (5 paires)	13164	manuelle et DPM	varie	MS,VP,ED,OCC
RAiD	ECCV 2014 [Das 2014]	43	4	6920	manuelle	128x64	MS,VP,VI
iLIDS-VID	ECCV 2014 [Wang 2014],T-PAMI 2016 [Wang 2016]	300	2	42495	manuelle	varie	MS,SS,VP,VI,EF,OCC
MPR Drone	ECCVW 2014 [Layne 2015]	84	1		ACF	varie	MS,IC
HDA Person Dataset	IJMISSP 2014 [Nambiar 2014],ECCVW 2015 [Figueira 2015]	53	12	2976	manuelle et ACF	varie	MS,SS,IC,VP,VI,ED,TR
Shinpuhkan Dataset	FCV2014 [Kawanishi 2014]	24	16		manuelle	128x48	MS,SS
Market1501	ICCV 2015 [Zheng 2015]	1501	6	32217	manuelle et DPM	128x64	MS,VP,ED,RES
MARS	ECCV 2016 [Zheng 2016a]	1261	6	1191003	DPM+GMMCP	256x128	MS,SS
DukeMTMC4reID	CVPRW 2017 [Gou 2017]	1852	8	46261	Doppia	varie	MS,IC,VP,VI,ED,EF,OCC
Airport	T-PAMI 2018 [Karanam 2018]	9651	6	39902	ACF	128x64	MS,VP,VI,ED,EF,OCC
MSMT17	CVPR 2018 [Wei 2018]	4101	15	126441	Faster RCNN	varie	MS

TABLE 3.3 – Base publiques d'images pour la ré-ID de personnes [Karanam 2018].

servées par deux caméras d'intérieur. WARD [Martinel 2012] recueille 4786 images de 70 personnes dans 3 caméras disjointes. SAIVT-Softbio [Bialkowski 2012] se compose de 152 personnes vues depuis un réseau de caméras de surveillance comportant 8 caméras. SAIVT-Softbio inclut des variations de point de vue et d'illumination. CUHK01 [Li 2013b] se compose de 971 personnes et 3884 images capturées à partir de 2 vues de caméra à champs de vues disjoints sur un campus universitaire. CUHK02 [Li 2013a] intègre 1816 personnes et 7264 images capturées à partir de 5 paires de caméras à champs de vues disjoints sur un campus universitaire. CUHK01 et CUHK02 sont annotées manuellement. CUHK03 [Li 2014] offre une observation de 1360 personnes avec 13164 images et 5 paires de caméras à champs de vues disjoints. CUHK03 est annotée manuellement mais également avec le détecteur DPM. CUHK03 inclut des variations de point de vue, erreurs de détection et des occultations. RAiD [Das 2014] comprend 43 personnes observées depuis deux caméras en intérieur et deux caméras en extérieur et inclut des variations de points de vue et d'illumination.

La base iLIDS-VID [Wang 2014, Wang 2016] comprend 600 séquences de suivi de 300 personnes issues de deux caméras disjointes dans un aéroport et inclut des variations de points de vue, des variations d'illumination, d'encombrements de fond et des occultations. HDA Person Dataset [Nambiar 2014, Figueira 2015] représente un banc d'essai pour un système de ré-ID automatique. Des images entièrement étiquetées pour des vidéos de 30 minutes de 13 caméras disjointes sont fournies. Market1501 [Zheng 2015] offre une observation de 1501 personnes avec 32643 images et 2793 fausses alarmes provenant du détecteur DPM. Market1501 inclut des variations de point de vue, des erreurs de détection et des images basse résolution. DukeMTMC4reID [Gou 2017] offre 1852 identités avec 46261 images et 21551 fausses alarmes provenant du détecteur Doppia [Benenson 2015]. Les images sont capturées à partir d'un réseau de 8 caméras disjointes situé sur le campus de l'Université de Duke. Airport [Karanam 2018] est créé en utilisant des vidéos de 6 caméras d'un réseau de surveillance intérieur dans un aéroport. À partir de ces vidéos, des tracklets correspondant à 9651 personnes sont extraites. Le nombre d'images dans la base de données est 39902.

Parmi les 23 bases de données pré-citées, huit seulement incluent des images complètes, les autres offrant les boîtes englobantes donc une segmentation préalable. Neuf bases de données ont des images issues de une ou deux caméras, n'offrant donc pas de réel topologie de réseau. De plus, rares sont les bases de données incluant des horodatages et une information topologique du réseau. La base de données HDA répond à ces critères. Elle inclut des images complètes, des annotations de deux types (manuelle et par détecteur ACF), des horodatages et l'information topologique du réseau. Elle constitue donc une base idéale pour évaluer notre approche. Nous décrivons plus précisément ci-après son contenu et les descripteurs pré-cités.

3.1.5 Métriques d'évaluation

Les courbes *Cumulative Matching Characteristics* (CMC) sont les mesures d'évaluation les plus usuelles pour les méthodes de ré-ID de personnes. Supposons que chaque identité de la galerie n'a qu'une seule image. Pour chaque personne cible, un algorithme va classer tous les identités de la galerie en fonction de leurs distances à la cible de petite à grande. La précision de rang k est définie par l'équation (3.4). La courbe CMC est calculée comme la moyenne pour toutes les requêtes de la précision de rang k.

$$Précision_k = \begin{cases} 1 & \text{si les top-}k\text{ échantillons les plus similaires de la} \\ & \text{galerie contiennent l'identité de la cible,} \\ 0 & \text{sinon.} \end{cases} \tag{3.4}$$

Pour les galeries *multi-shots*, c'est-à-dire incluant plusieurs images de la même identité, deux stratégies sont envisagées : (1) choisir (souvent aléatoirement) une seule instance censée représenter la galerie, ou au contraire (2) utiliser toutes les instances.

3.2 Choix et méthodologies pour notre approche

Comme évoqué en préambule, un système de ré-ID de personnes est composé classiquement de 3 briques : un/des descripteur(s), une métrique et une règle de décision. Rappelons que nos travaux se focalisent sur la brique de décision.

La littérature met en évidence qu'une inférence globale, c'est-à-dire au niveau réseau, doit améliorer intuitivement les performances de ré-ID. Fort de ce constat, deux approches nous semblent à approfondir/amender : celles basées sur la PLNE (NCR [Das 2014, Chakraborty 2016]) et celles utilisant l'apprentissage profond qui ne séparent pas les briques/éléments mentionnés mais les intègrent au sein du réseau.

Il nous semble que la connaissance de la topologie du réseau peut être exploitée plus largement ; elle offre des connaissances supplémentaires et intuitives, exploitables dans le processus global d'optimisation du réseau, en particulier : le temps de transit entre deux champs de vue de caméras (pour une personne qui marche normalement), la configuration topologique du réseau. À notre connaissance, aucune méthode n'intègre simultanément ces deux informations dans la modélisation.

Le chapitre 4 décrit notre nouvelle approche, baptisée D-NCR (*Directed Network Consistent Re-identification*), pour la décision de ré-ID. Elle est inspirée de la stratégie NCR mais reformalise cette dernière afin d'intégrer l'information topologique et temporelle. À l'instar de NCR, notre stratégie de décision/classification est indépendante du descripteur et de la métrique utilisés. Néanmoins, cette stratégie doit être accompagnée d'un choix de descripteur et de métrique de similarité pour constituer une chaîne de ré-ID complète et ainsi évaluer les performances. Pour nos expérimentations, une base de données est aussi à privilégier.

Nous avons choisi d'utiliser les descripteurs SDALF [Farenzena 2010] et WACN [Martinel 2012], avec leurs métriques de distance respectives. Deux raisons, au moins, ont motivé ce choix. Ces descripteurs ont été privilégiés dans les travaux de Das *et al.* [Das 2014] qui constitue notre approche de référence. La comparaison sera donc immédiate. De plus, ils sont faciles à implémenter grâce à une bibliothèque de leurs auteurs.

Des base de données présentés précédemment, nous avons choisi d'utiliser la base de donnés HDA car elle est une des plus complètes et est la seule à offrir l'information de la topologie de la réseau de caméras, ce qui est de grande importance pour nos contributions.

3.2.1 Focus sur les descripteurs SDALF et WACN

Le descripteur SDALF a été initialement proposé par Farenzena *et al.* [Farenzena 2010] et Bazzani *et al.* [Bazzani 2013]. Pour son calcul, une segmentation préalable de l'arrière plan est effectuée. Le première plan, c'est-à-dire a priori la personne, est divisé en parties grâce au calcul des axes de symétrie et d'asymétrie. Pour chaque partie, trois descripteurs sont calculés : un histogramme couleur HSV et les *Maximally Stable Color Regions* (MSRC) pour représenter la couleur et les *Recurrent Highly Structured Patches* (RHSP) pour la texture. Deux exemples sont présentés figure 3.3. Une distance de Bhattacharyya est calculée pour comparer les histogrammes HSV. Pour le descripteur MSCR, la distance minimale entre éléments MSCR est considérée. Ces distances sont une combinaison linéaire des distances entre centroïdes et la distance couleur. Les descripteurs RHSP sont comparés aussi par une distance de Bhattacharyya. La métrique de distance finale est une combinaison linéaire de ces trois distances. Ce descripteur est défini pour une ré-ID par image et par vidéo, dans notre travail nous privilégions l'implementation par image.

Le descripteur WACN a été initié par Martinel et Micheloni [Martinel 2012]. Pour son calcul, une segmentation préalable de l'arrière plan est aussi effectuée. Trois descripteurs sont calculés : un histogramme couleur HSV à partir des points d'intérêt SIFT, des descripteurs PHOG pour chaque canal couleur HSV et un vecteur descripteur Haralick (de texture) pour le corps et pour les jambes. Un exemple est illustré sur la figure 3.4. Les points d'intérêt SIFT sont divisés en parties du corps et une distance χ^2 est utilisée pour les comparer. La distance χ^2 est utilisée aussi pour la comparaison des descripteurs PHOG. Les vecteurs Haralick sont comparés via une distance euclidienne (norme l_2). La métrique de distance finale est une combinaison linéaire de ces trois distances.

3.2.2 Focus sur la base publique HDA

La base HDA Person Dataset [Nambiar 2014, Figueira 2015] inclut des séquences d'images haute résolution (640x480, 1280x800 et 2560x1600) multi-caméras. 18 caméras ont été enregistrées simultanément pendant 30 minutes dans un scénario de

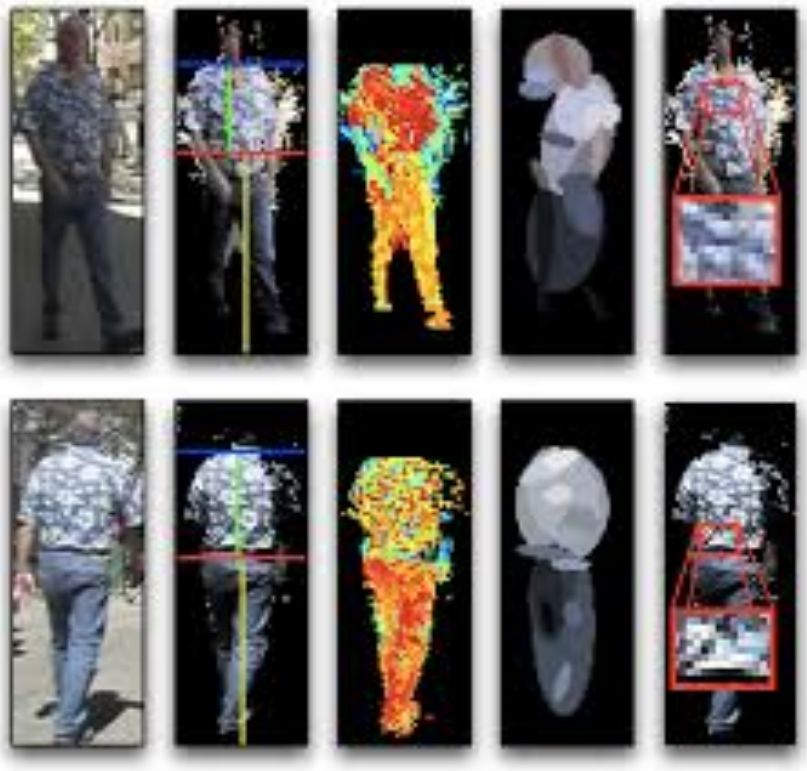


FIGURE 3.3 – Exemples du descripteur SDALF [Farenzena 2010]. Les colonnes présentent respectivement : deux images de la même personne, les axes de symétrie et asymétrie, la projection de l'histogramme couleur HSV, les *Maximally Stable Color Regions* (MSRC) et les *Recurrent Highly Structured Patches* (RHSP).

scène d'intérieur typique à une heure de pointe (heure du déjeuner) impliquant plus de 80 personnes. Dans la version actuelle, 12 caméras ont été entièrement annotées. Des exemples d'images acquises par ces caméras sont montrées figure 3.5.

L'environnement est constitué de trois étages de l'*Institute for Systems and Robotics (ISR-Lisbon)*. La figure 3.6 illustre l'emplacement des caméras. Les 18 caméras sont matérialisées par un petit cercle rouge. Les images issues des 12 caméras avec un champ de vue coloré ont été entièrement annotées.

Chaque image est annotée avec des boîtes englobantes ajustées sur les personnes présentes, leurs IDs, et des champs de bits indiquant si ces personnes cibles sont occultées et/ou immergées dans une foule.

3.3 Conclusions

Ce chapitre a permis de dresser un panorama des travaux existants sur la ré-ID de personnes et ainsi motiver notre approche D-NCR et justifier les choix sous-jacents. Notre contribution, présentée au chapitre suivant, est une reformulation originale du problème de ré-ID en exploitant plus largement des contraintes réseau

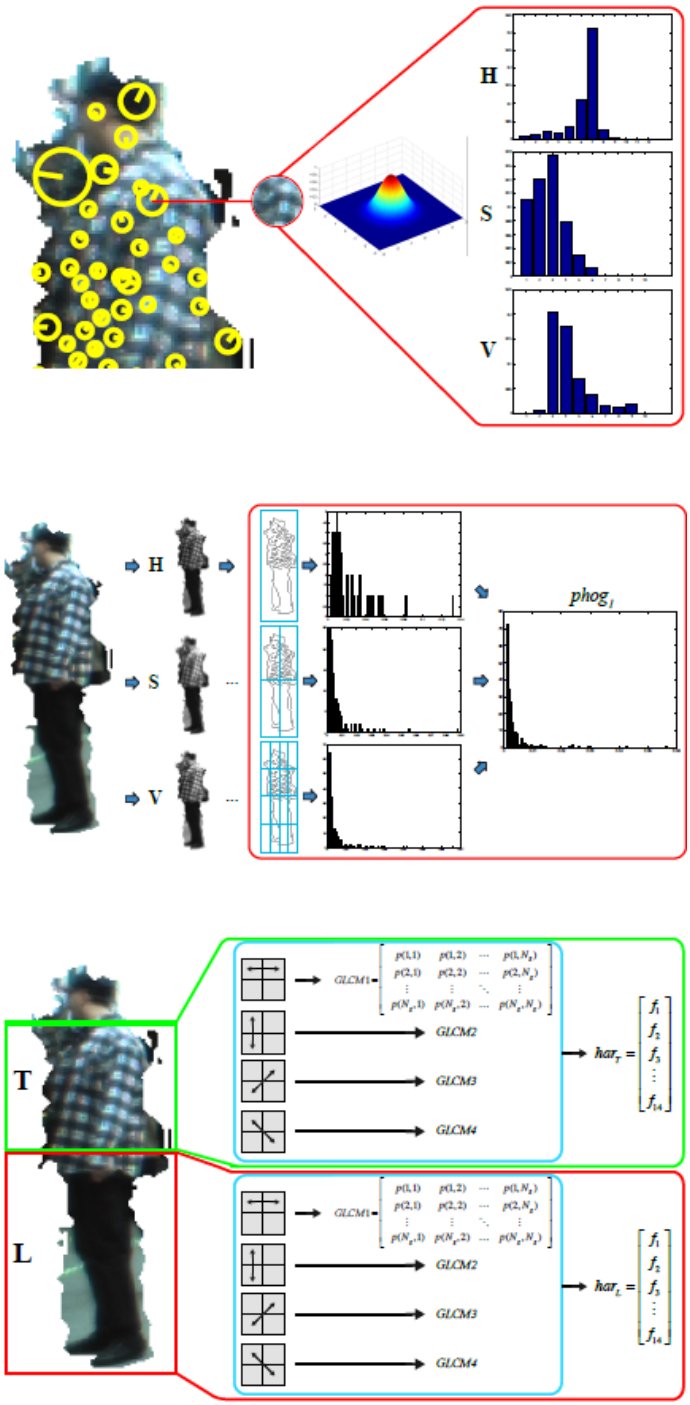


FIGURE 3.4 – Exemple du descripteur WACN [Martinel 2012]. Les descripteurs calculés sont : un histogramme couleur HSV à partir des points d'intérêt SIFT première ligne), des descripteurs PHOG pour chaque canal couleur HSV (seconde ligne), et un vecteur descripteur Haralick pour le corps et pour les jambes (dernière ligne.

(a) Cam 17. (b) Cam 18. (c) Cam 19.

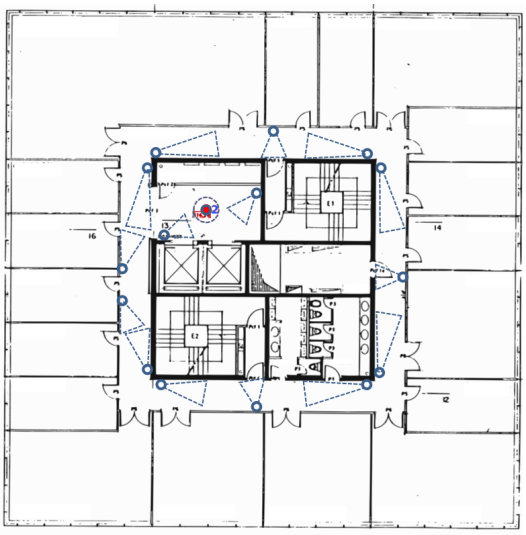
(d) Cam 40. (e) Cam 50. (f) Cam 53.

(g) Cam 54. (h) Cam 56. (i) Cam 57.

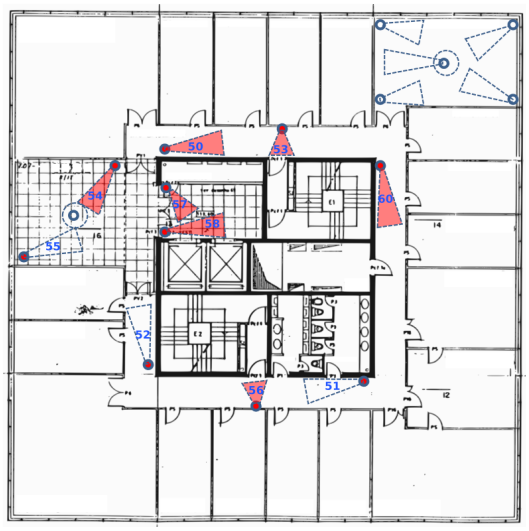
(j) Cam 58. (k) Cam 59. (l) Cam 60.

FIGURE 3.5 – Exemples d'images annotées de la base HDA.

c'est-à-dire sa topologie globale et la cohérence temporelle des images acquises. Nos évaluations préliminaires sur la base de HDA montrent que cette reformulation induit des gains en performances comparativement aux approches classiques par paires d'images ou globales types NCR.



(a) Étage 6.



(b) Étage 7.

(c) Étage 8.

FIGURE 3.6 – Emplacement des caméras de la base HDA.

Chapitre 4

Ré-identification de personnes avec prise en compte de contraintes temporelles et topologiques inhérentes au réseau de caméras

Sommaire

Les approches conventionnelles de ré-ID, y compris NCR et [Lin 2017] ignorent généralement les informations liées à la topologie du réseau de caméras et aux temps de transit entre deux caméras voisines, alors que la configuration spatiale de ces réseaux est souvent figée dans l'environnement. Comme évoqué au chapitre précédent, intégrer ces connaissances a priori dans la modélisation peut potentiellement induire des gains de ré-ID. En effet, exploiter la topologie du réseau permet d'éviter les appariements de détections entre deux caméras non reliées par une trajectoire dans le réseau. De plus, la connaissance des temps de transit minimum/moyen des personnes entre paires de caméras peut renforcer le filtrage des appariements en évitant

l'association de détections non cohérentes temporellement. Remarquons cependant qu'un tel filtrage temporel impose un horodatage préalable des images acquises et donc une synchronisation temporelle (même grossière) des caméras du réseau.

Dans ce chapitre, nous présentons notre approche de ré-ID de personnes avec prise en compte de la topologie réseau et des contraintes de temps. Cette approche, inspirée de la stratégie NCR formalisée en section 4.1, sera notée D-NCR. La section 4.2 décrit l'approche D-NCR avec prise en compte explicite des contraintes de topologie et de temps de transit. Elle constitue une formalisation mathématique originale du problème de ré-ID. Les expérimentations, les évaluations et discussions associées à l'implémentation de notre approche sont décrites en section 4.3. Les méthodes de résolution proposées sont présentées en section 4.4.2. Enfin, la section 4.5 présente les conclusions et perspectives de nos travaux.

4.1 Focus sur la stratégie NCR

Das *et al.* [Das 2014] proposent la stratégie NCR pour résoudre le problème de ré-ID. Elle repose sur une utilisation des scores de similarité par paires de caméras et paires de cibles. Nous précisons tout d'abord les notations puis la terminologie associée à cette méthode avant d'analyser plus en profondeur la méthode de résolution.

Soit m le nombre de caméras du réseau. La $i^{\text{ème}}$ personne présente sur la caméra p est notée $\mathcal{P}_i^p$ et est associée à un nœud d'un graphe. Soit $\mathbf{C}^{(p,q)}$ la matrice de similarité entre les personnes détectées sur les caméras p et q. Ainsi, le $(i,j)^{\text{ème}}$ élément de $\mathbf{C}^{(p,q)}$ indique le score de similarité entre les personnes $\mathcal{P}_i^p$ et $\mathcal{P}_j^q$. Résoudre le problème de ré-ID consiste à proposer une association entre les personnes $\mathcal{P}_i^p$ et $\mathcal{P}_j^q$. Cette association peut être représentée à l'aide de matrices d'association, une pour chaque paire de caméras. Chaque variable $x_{i,j}^{p,q}$ de la matrice d'association $\mathbf{X}^{p,q}$ entre la paire de caméras (p,q) est égale à 1 si $\mathcal{P}_i^p$ et $\mathcal{P}_j^q$ sont associés à la même cible, 0 sinon.

En sommant toutes les scores de similarité associés aux appariements réalisés, un score de similarité globale peut être défini comme décrit par l'équation (4.1), où k est un paramètre empirique dans le domaine des scores de similarité qui aide a gérer le diffèrent nombre de détections dans chaque caméra de la réseau. Ce concept de score de similarité est influencé par le constat que les scores associés pour la plupart des vrais appariements sont plus importants que la majorité des faux appariements. Une personne de n'importe quelle caméra p peut être associée au plus une fois avec une autre personne d'une caméra q. Ceci est mathématiquement exprimé par l'ensemble des équations (4.2)-(4.3), cette contrainte est vérifiée pour toutes les paires de caméras possibles. La contrainte (4.4) provient de l'exigence de cohérence, formulée comme une contrainte de boucle ou transitivité. Elle exprime que si les associations entre les personnes $\mathcal{P}_i^p$ et $\mathcal{P}_k^r$ et entre les personnes $\mathcal{P}_k^r$ et $\mathcal{P}_j^q$ ont été réalisées, alors l'association entre les personnes $\mathcal{P}_i^p$ et $\mathcal{P}_j^q$ doit être aussi réalisé.

Maximiser

$$\sum_{p=1}^{m}\sum_{q=p+1}^{m}\sum_{i=1}^{n_p}\sum_{j=1}^{n_q}\left(c_{i,j}^{p,q}-k\right)x_{i,j}^{p,q} \tag{4.1}$$

Sous contraintes que

$$\sum_{j=1}^{n_q} x_{i,j}^{p,q} \leq 1 \quad \forall i \text{ and } \forall\,(p,q)\,|p<q \tag{4.2}$$

$$\sum_{i=1}^{n_p} x_{i,j}^{p,q} \leq 1 \quad \forall j \text{ and } \forall\,(p,q)\,|p<q \tag{4.3}$$

$$x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \quad \forall\,(i,j) \text{ and } \forall\,(p,q,r)\,|p<r<q \tag{4.4}$$

$$\forall x_{i,j}^{p,q} \in \{0,1\}$$

Cette méthodologie de ré-ID a été utilisée dans d'autres applications de ré-ID, notamment le suivi de cellules [Chakraborty 2016].

4.1.1 Expérimentations préliminaires associées

Nous avons évalué l'approche NCR sur la base de données HDA [Nambiar 2014, Figueira 2015]. Pour rappel, cette base publique est composée de flux vidéo de 30 minutes acquis par un réseau de 18 caméras réparties sur trois étages d'un bâtiment (étages 6, 7 et 8). Cependant, seulement 12 flux vidéo sur 30 sont entièrement annotés.

Pour nos expérimentations préliminaires, nous focalisons sur 10 minutes de vidéo d'un seul étage possédant 5 caméras (étage 8). Comme cet étage est instrumenté de paires de caméras avec des FOVs qui se chevauchent, deux caméras sont omises pour finalement se focaliser un réseau de trois caméras avec des FOVs complètement disjoints. Ce choix est justifié par le fait que la ré-ID des personnes dans des caméras à FOVs joints est une problématique spécifique, hors du spectre de cet travail. Ainsi, nous nous focalisons sur les 10 premières minutes des vidéos des caméras 17, 18 et 40 du 8$^{\text{ème}}$ étage. Ces séquences vidéo comportent respectivement 11, 38 et 54 *tracklets* de personnes annotées. Chaque *tracklet* est représenté temporellement par l'image de la détection médiane. Nous évaluons toutes les combinaisons possibles de paires de caméras (17-18, 17-40, 18-17, 18-40, 40-17 et 40-18). Pour chaque paire de caméras, nous comparons chaque détection de la première caméra avec chaque échantillon de la seconde. Nous supposons qu'il n'existe pas de galerie de personnes connues a priori, ce qui est le cas pour une première ré-ID. Pour cette raison, pour chaque paire de caméras, toutes les *tracklets* de la deuxième caméra sont utilisées comme galerie et les *tracklets* de la première sont les cibles à ré-identifier.

Comme évoqué au chapitre précédent, deux descripteurs sont privilégiés dans nos expérimentations : SDALF [Farenzena 2010] et WACN [Martinel 2012]. Les boîtes à outils MATLAB fournies par les auteurs sont utilisées pour leur implé-

mentation. Dans le cas du descripteur SDALF, seul l'histogramme HSV est pris en compte à l'instar de nombreux travaux. Pour le descripteur WACN, la composante relative au SIFT est omise, c'est-à-dire seules les composantes relatives aux pHOG et Haralick (texture) sont utilisées. Ce choix est imposé du fait que la méthode NCR suppose des matrices de similarité symétriques. Pour chacun des descripteurs SDALF et WACN, la méthode NCR [Das 2014, Chakraborty 2016] est évaluée.

Le modèle PLNE NCR a été implémenté en C++ et résolu en utilisant le solveur PLNE commercial Gurobi 7.5.1. Tous les tests expérimentaux sont effectués sur un processeur Intel® Core™ i5-4670 CPU de 3,4 GHz avec 16 Go de mémoire RAM DDR3 1600 MHz. Aucun GPU n'est utilisé.

4.1.1.1 Évaluation des performances de ré-ID

Pour rappel, nous privilégions la métrique d'évaluation CMC et focalisons ici sur les taux de reconnaissance au premier rang. Ceux-ci sont présentés dans la table 4.1. Nous observons que le raisonnement au niveau réseau de la méthode NCR améliore logiquement les performances de ré-ID. Ainsi, la méthode NCR induit des gains en performances, respectivement de 1.46% et de 0.49% par rapport aux techniques classiques de ré-ID par paires d'images avec descripteurs SDALF et WACN.

Les courbes CMC par paires de caméras sont présentées sur la figure 4.1. Ces courbes sont cohérentes avec les résultats de la table 4.1 : les gains en ré-ID au premier rang intersectent les performances globales CMC.

Les courbes CMC globales sont présentées figure 4.2. Elles sont cohérentes avec les résultats de la table 4.1 et les courbes CMC de la figure 4.1. Eu égard à la métrique CMC, l'approche NCR conduit à des gains de performance, relativement faibles par rapport aux techniques classiques de ré-ID par paires d'images car seul le premier rang est impacté en termes d'inférence.

Évoquons les performances en absolu ; celles-ci sont secondaires ici car on s'intéresse aux gains induits par la stratégie NCR. Ces performances sont fortement dépendantes du choix des descripteurs et métriques. Or, les descripteurs SDALF et WACN n'étant pas les meilleurs, les performances CMC globales restent donc faibles en absolu.

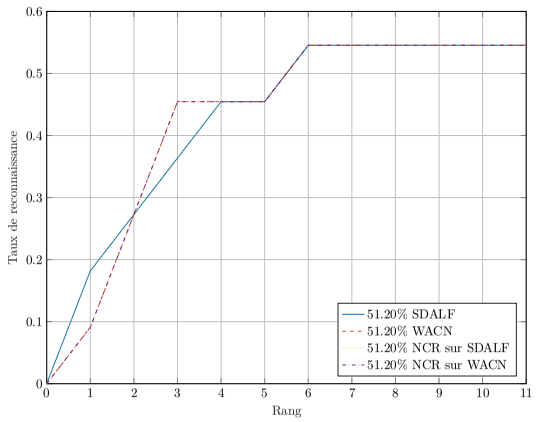
4.1.1.2 Exemple de scénario

Un exemple de ré-ID inféré par descripteur SDALF est présenté sur la figure 4.3. Ici, l'identité 10 de la caméra 17 est appariée avec l'identité 2 de la caméra 18, l'identité 10 de la caméra 17 est appariée avec l'identité 1 de la caméra 40 et l'identité 2 de la caméra 18 est appariée avec l'identité 46 de la caméra 40. Nous pouvons observer que l'appariement des caméras 18 et 40 n'est pas cohérent avec les appariements des caméras 17 et 18 et des caméras 17 et 40, car l'identité 2 de la caméra 18 est appariée avec l'identité 46 de la caméra 40 alors qu'elle devrait être appariée avec l'identité 1 de la caméra 40.

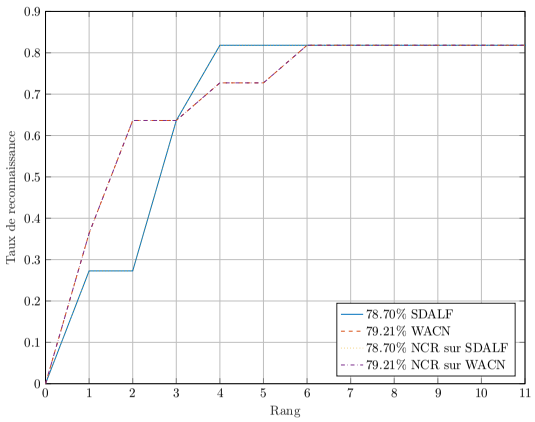
Au contraire, nous observons figure 4.4 que la méthode NCR permet de lever cette ambiguïté.

Méthode	17-18	18-17	17-40	40-17	18-40	40-18	Global
SDALF	18.18% (2/11)	21.05% (8/38)	27.27% (3/11)	22.22% (12/54)	34.21% (13/38)	35.19% (19/54)	27.67% (57/206)
NCR sur SDALF	18.18% (2/11)	21.05% (8/38)	27.27% (3/11)	24.07% (13/54)	36.84% (14/38)	37.04% (20/54)	29.13% (60/206)
WACN	9.09% (1/11)	15.79% (6/38)	36.36% (4/11)	11.11% (6/54)	39.47% (15/38)	29.63% (16/54)	23, 30% (48/206)
NCR sur WACN	9.09% (1/11)	15.79% (6/38)	36.36% (4/11)	11.11% (6/54)	39.47% (15/38)	31.48% (17/54)	23, 79% (49/206)

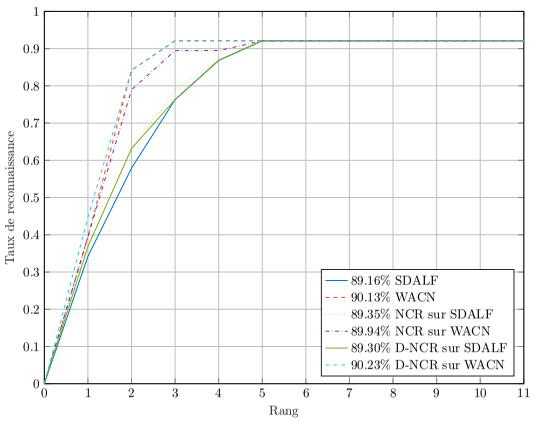
TABLE 4.1 – Taux de reconnaissance de premier rang par paires de caméras pour les caméras 17, 18 et 40 de la base de données HDA.



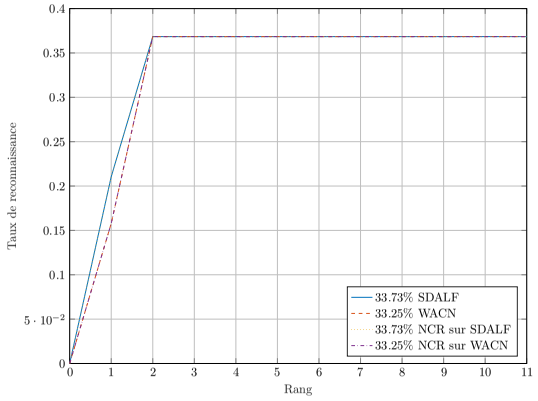
(a) Caméras 17-18.



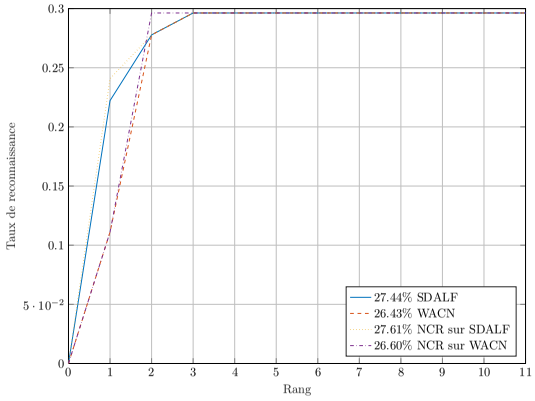
(b) Caméras 17-40.



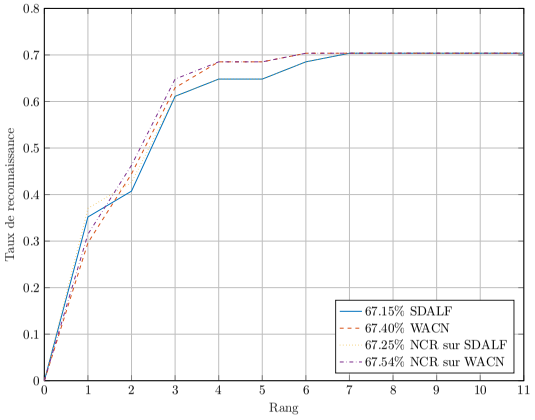
(c) Caméras 18-40.



(d) Caméras 18-17.



(e) Caméras 40-17.



(f) Caméras 40-18.

FIGURE 4.1 – Courbes CMC par paires de caméras pour les caméras 17, 18 et 40 de la base HDA.

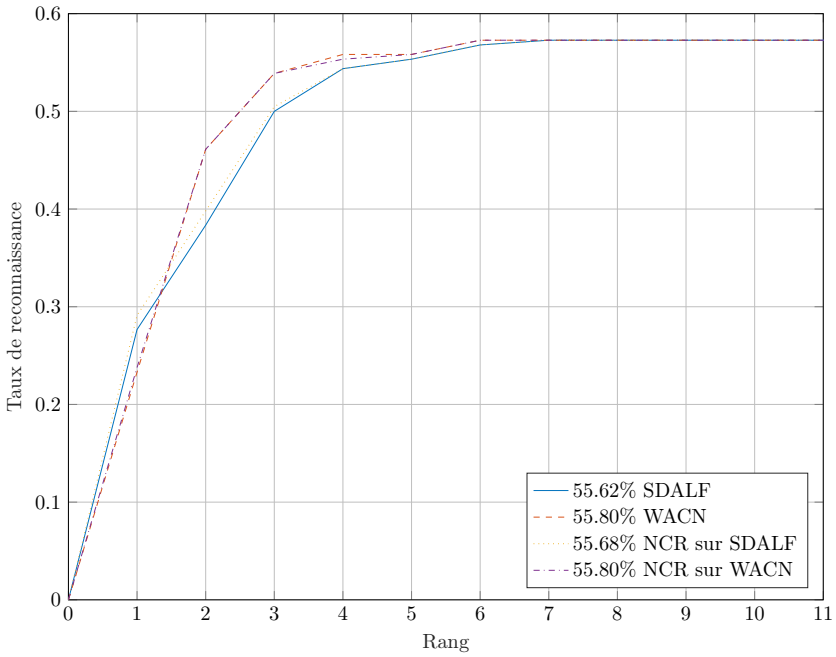


FIGURE 4.2 – Courbes CMC globale pour les caméras 17, 18 et 40 de la base HDA.

(a) Caméras 17-18. (b) Caméras 17-40. (c) Caméras 18-40.

FIGURE 4.3 – Exemple d'appariement par SDALF sur la base HDA. L'identité 10 de la caméra 17 est appariée avec l'identité 2 de la caméra 18. L'identité 10 de la caméra 17 est appariée avec l'identité 1 de la caméra 40. L'identité 2 de la caméra 18 est appariée avec l'identité 46 de la caméra 40.

4.2 Définition et formalisation de notre approche D-NCR

Comme évoqué précédemment, les temps de transit et l'horodatage permettent de filtrer les appariements impossibles en raison des contraintes de temps. Ainsi, considérant deux détections de personnes dans deux caméras distinctes du réseau, un appariement ne pourra être plausible que si la différence d'horodatage des détections est supérieure ou égale au temps de transit minimum.

(a) Caméras 17-18. (b) Caméras 17-40. (c) Caméras 18-40.

FIGURE 4.4 – Exemple d'appariement par NCR sur SDALF sur la base HDA. L'identité 10 de la caméra 17 est appariée avec l'identité 2 de la caméra 18. L'identité 10 de la caméra 17 est appariée avec l'identité 1 de la caméra 40. L'identité 2 de la caméra 18 est appariée avec l'identité 1 de la caméra 40.

De même, la connaissance a priori de la topologie réseau permet d'apparier les détections dans un ordre rationnel. Deux caméras sont dites topologiquement connectées lorsqu'une personne peut passer de l'une à l'autre sans passer par le champ de vue d'aucune autre caméra. La prise en compte de la topologie est pertinente dans les environnements humains structurés (e.g. bâtiments, villes). Ainsi, si les caméras (A, B) et (B, C) sont connectées (et non (A, C)), un appariement entre deux détections d'une personne cible sur les caméras A et C ne sera réalisé que si un appariement a été préalablement réalisé pour cette personne entre les caméras A et B, puis entre les caméras B et C. Ces appariements successifs caractérisent donc le chemin de la personne dans le graphe topologique. Ainsi, chaque détection dans une caméra sera au plus appariée deux fois : une première fois avec une détection sur une caméra indiquant la provenance de la personne, une deuxième fois avec une détection sur une caméra indiquant la destination de la personne.

La section suivante présente une formalisation originale, sous forme de programme linéaire en nombres entiers (PLNE), du problème de ré-ID avec prises en compte explicite de la topologie et du temps de transit inter-caméras.

4.2.1 Formalisation

Notons $\mathcal{G}(V, E)$ le graphe représentant la structure topologique du réseau de caméras. V est l'ensemble des caméras $\{1, \ldots, m\}$ du réseau, auxquelles nous ajoutons deux caméras fictives 0 et $m + 1$. Il est supposé que, chaque fois qu'une personne entre (quitte, respectivement) le réseau de caméras, sa première (dernière, respectivement) détection se produit sur 0 ($m + 1$, respectivement). E est l'ensemble des arêtes tel qu'une arête entre deux sommets p et q de V indique qu'une personne peut être détectée par la caméra p, puis par la caméra q, sans aucune détection par une troisième caméra $r \neq p, q$. Chaque arête (p, q) de $\mathcal{G}$ est pondérée par $t^{p,q}$ qui correspond au temps de transit minimal pour passer de la caméra p à la caméra q

(et vice versa). Par hypothèse, nous supposons que les arcs $(0, q)$ et $(q, m+1)$ sont dans V $\forall q \in V$. De plus, $t^{0,q} = t^{q,m+1} = 0$ $\forall q \in V$.

À chaque sommet q de $\mathcal{G}$ est associé un ensemble de détections de personnes. La détection de la $i^{\text{ème}}$ personne observée par la caméra q est notée comme précédemment $\mathcal{P}_i^q$. Une fenêtre temporelle $[\underline{t}_i^q, \overline{t}_i^q]$ lui est associée, où $\underline{t}_i^q$ est le temps de début de la détection, et $\overline{t}_i^q$, son temps de fin. De plus, pour toute détection $\mathcal{P}_i^q$, deux détections fictives $\mathcal{P}_i^0$ et $\mathcal{P}_i^{m+1}$ sont également créées, avec $\left[\underline{t}_i^0, \overline{t}_i^0\right] = [0, 0]$ et $\left[\underline{t}_i^{m+1}, \overline{t}_i^{m+1}\right] = [t^{\max}, t^{\max}]$ ($t^{\max}$ étant le plus grand temps de fin de détection). Comme expliqué ci-dessous, $\mathcal{P}_i^0$ et $\mathcal{P}_i^{m+1}$ sont utilisés respectivement pour décider si $\mathcal{P}_i^q$ est la première détection d'une personne dans le réseau, ou, inversement, sa dernière détection. L'ensemble des détections de personnes est noté P.

Nous introduisons à présent le deuxième graphe $G(P, X)$, correspondant au graphe de détection de personnes, où chaque détection $\mathcal{P}_i^q \in P$ est un sommet de G. Il existe un arc entre la détection $\mathcal{P}_i^p$ et la détection $\mathcal{P}_j^q$ dans X si et seulement si $\underline{t}_j^q > \overline{t}_i^p$ et $\underline{t}_j^q - \overline{t}_i^p \geq t^{p,q}$. De plus, X peut être partitionné en deux sous-ensembles Y et Z (c'est-a-dire, $X = Y \cup Z$) tels que : l'arc $\left(\mathcal{P}_i^p, \mathcal{P}_j^q\right)$ appartient à Y si l'arête (p, q) est dans E, sinon il est dans Z. À chaque arc de X est associé un bénéfice $c_{i,j}^{p,q}$ qui correspond au score de similarité entre les détections de personne $\mathcal{P}_i^p$ et $\mathcal{P}_j^q$. Notez que $\forall \mathcal{P}_i^q \in P$, $c_{i,i}^{0,p}$ et $c_{i,i}^{p,m+1}$ égal à 0 par hypothèse.

La figure 4.5 illustre la formalisation précédente sur un exemple simple de 4 caméras et 7 détections : 1 détection, 2 détections, 3 détections et 1 détection respectivement pour les caméras 1, 2, 3 et 4. Le graphe $\mathcal{G}(V, E)$ (resp. $G(P, X)$) est représenté en vert (resp. en bleu) tandis que les arcs appartenant à Z sont en pointillés. Dans cet exemple, l'ensemble E est composé des arcs $(1, 2)$, $(1, 3)$, $(1, 4)$, $(2, 3)$ et $(2, 4)$. L'ensemble X est composé de 17 arcs, 13 appartenant à Y et 4 à Z. Les caméras fictives 0 et $m + 1$ ne sont pas représentées pour favoriser la lisibilité de la figure.

Sous ce formalisme, le problème de ré-ID s'apparente à un problème de partitionnement en cliques à profit maximum dans un graphe orienté avec contraintes de flots. Une solution au problème de ré-ID est un ensemble de chemins disjoints dans G qui utilisent seulement les arcs de Y et partent d'une détection du sommet source pour finir à une détection du sommet puits. Chaque chemin disjoint peut être considéré comme l'itinéraire emprunté par une personne à l'intérieur du réseau de caméras. Une solution est optimale si elle maximise le profit total, qui est la somme du profit $c_{i,j}^{p,q}$ des arcs de Y utilisés par les chemins, plus la somme des profits des arcs de Z impliqués dans la fermeture transitive de chaque chemin.

4.2.2 Complexité du problème

Le problème précédent est voisin du problème de partition en cliques [Gramm 2009] qui est NP-difficile. La formalisation NCR [Das 2014, Chakraborty 2016] est équivalente à un problème de partition en cliques de profit maximum, également NP-difficile. Dans notre cas, le graphe est orienté et composé

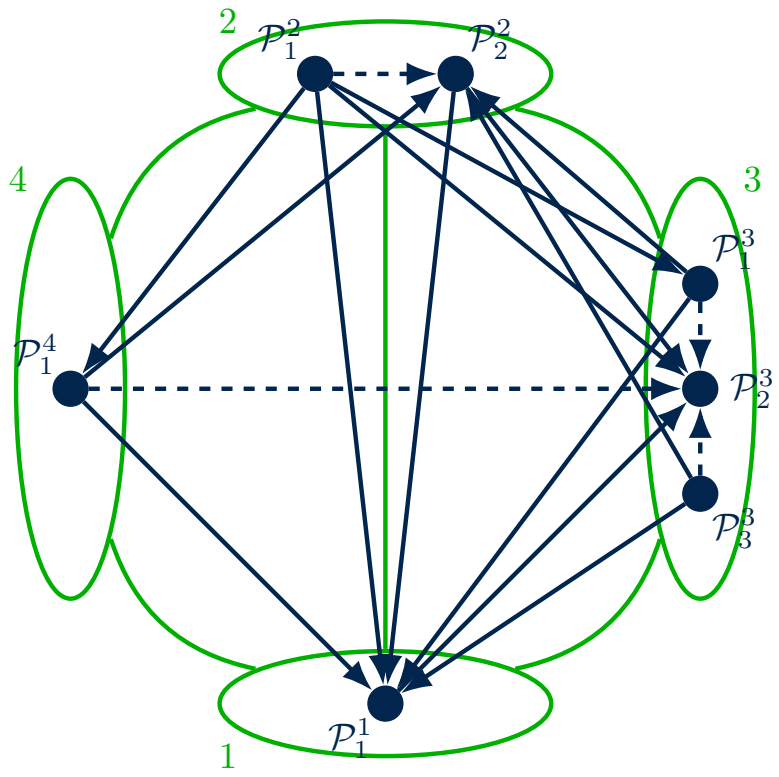


FIGURE 4.5 – Exemple de modélisation en graphes de la ré-ID de personnes pour un réseau de 4 caméras et 7 détections.

de deux catégories d'arcs : ceux sur lesquels circule un flot et les autres. Nous n'avons pas trouvé d'étude de complexité sur ce problème particulier de partitionnement en cliques de profit maximum dans des graphes pondérés et orientés. Si uniquement les arcs sur lesquels circule un flot sont considérés, le problème peut être résolu en temps polynomial par une recherche de chemins disjoints à profit maximum [Robertson 1995] en utilisant un des algorithmes de recherche de flot maximum à coût minimum (par exemple [Goldberg 1990]). Cette caractéristique de notre problème nous amène à penser que notre problème est plus simple que le problème de partition en cliques traité dans NCR. Nous conjecturons qu'il est également NP-Difficile même si cette question de complexité demeure ouverte.

4.2.3 Formulation PLNE

Dans cette partie, une formulation PLNE du problème d'optimisation précédent est proposée. Pour chaque paire de détection $\mathcal{P}_i^p \in P$ et $\mathcal{P}_j^q \in P$, la variable de décision binaire $x_{i,j}^{p,q}$ est introduite : elle est égale à 1 si l'arc $\left(\mathcal{P}_i^p, \mathcal{P}_j^q\right) \in X$ est utilisé dans un chemin disjoint de Y ou sa fermeture transitive dans Z, 0 sinon.

Maximiser

$$\sum_{\left(\mathcal{P}_i^p, \mathcal{P}_j^q\right) \in X} c_{i,j}^{p,q} x_{i,j}^{p,q} \tag{4.5}$$

Sous contraintes que

$$x_{0,j}^q + \sum_{\left(P_i^p, P_j^q\right) \in Y} x_{i,j}^{p,q} = 1 \quad \forall \mathcal{P}_j^q \in P \tag{4.6}$$

$$\sum_{\left(P_i^p, P_j^q\right) \in Y} x_{i,j}^{p,q} + x_{i,m+1}^{p} = 1 \quad \forall \mathcal{P}_i^p \in P \tag{4.7}$$

$$x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \quad \begin{array}{l} \forall \left(\mathcal{P}_i^p, \mathcal{P}_j^q\right) \in X, \\ \forall \left(P_i^p, P_k^r\right) \in Y, \\ \forall \left(\mathcal{P}_k^r, \mathcal{P}_j^q\right) \in Y \end{array} \tag{4.8}$$

$$x_{i,j}^{p,q} \leq x_{i,k}^{p,r} \quad \begin{array}{l} \forall \left(\mathcal{P}_i^p, \mathcal{P}_j^q\right) \in X, \\ \forall \left(\mathcal{P}_i^p, \mathcal{P}_k^r\right) \in Y \end{array} \tag{4.9}$$

$$x_{i,j}^{p,q} \leq x_{k,j}^{r,q} \quad \begin{array}{l} \forall \left(\mathcal{P}_i^p, \mathcal{P}_j^q\right) \in X, \\ \forall \left(\mathcal{P}_k^r, \mathcal{P}_j^q\right) \in Y \end{array} \tag{4.10}$$

$$\forall x_{i,j}^{p,q} \in \{0, 1\}$$

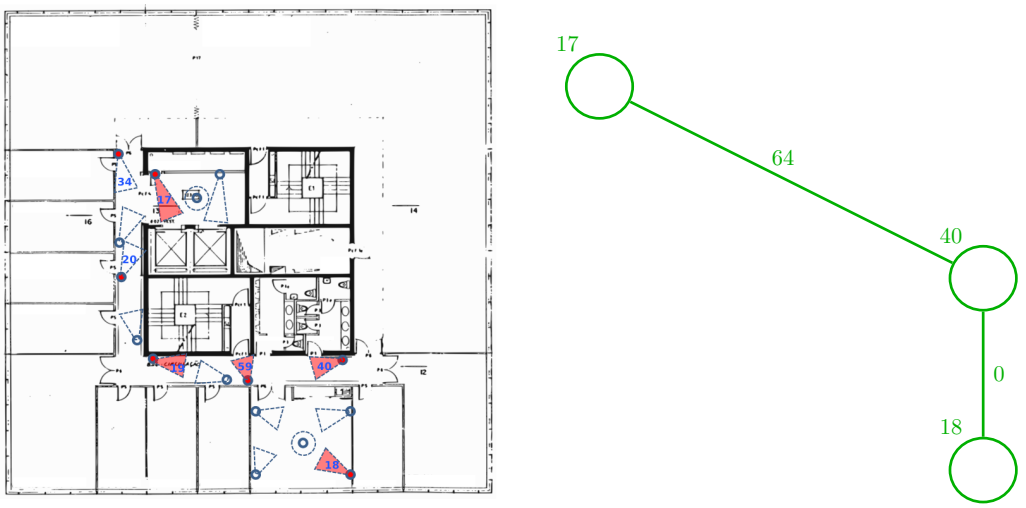
La fonction objectif (4.5) établit le bénéfice total associé à une solution. Les contraintes (4.6)-(4.7) sont des contraintes de flot : elles garantissent qu'une solution forme toujours un ensemble de chemins disjoints dans Y. Enfin, les contraintes (4.8)-(4.10) garantissent que seuls les arcs de Z correspondant à la fermeture transitive d'un chemin sont sélectionnés dans une solution.

La formulation D-NCR diffère essentiellement de celle proposée par la méthode NCR par le fait que le graphe est orienté (*directed* en anglais). Néanmoins, les inégalités (4.8)-(4.10) assurant la fermeture transitive sont semblables (hormis les inégalités (4.9) et (4.10) qui ont été ajoutées car elles nous semblaient manquantes dans la formulation de Das *et al.* [Das 2014])

4.3 Expérimentations préliminaires et évaluations associées

A l'instar de la méthode NCR (voir section 4.1), nous évaluons D-NCR sur la base de données de personnes HDA [Nambiar 2014, Figueira 2015] qui inclut également une description de la topologie du réseau de caméras. Comme précédemment, nous limitons notre étude à un sous-ensemble constitué de trois caméras. Plus précisément, nous nous concentrons sur les vidéos des caméras 17, 18 et 40 au 8$^{\text{ème}}$ étage. Le graphe topologique $\mathcal{G}(V, E)$ décrivant la structure du réseau de caméras

est présenté sur la figure 4.6b. Pour cet exemple, les arcs $(17, 40)$ et $(18, 40)$ sont dans E, mais pas l'arc $(17, 18)$. Ces séquences vidéo comportent chacune 11, 38 et 54 *tracklets* de personnes annotées, respectivement pour les caméras 17, 18 et 40. Nous évaluons toutes les combinaisons possibles de paires de caméras (17-18, 17-40, 18-17, 18-40, 40-17 et 40-18). Les horodatages de chaque image ont été construits à partir des horodatages de la première image et les fréquences d'acquisition de chaque caméra. En l'absence de données précises, les temps minimum de transit entre les paires de caméras 17-40 et 18-40 ont été estimés à partir des horodatages et de la vérité terrain. Ces valeurs estimées sont de 64 et 0 unité(s) de temps, respectivement. La dernière valeur nulle s'explique par le fait que la salle surveillée par la caméra 18 et le couloir surveillé par la caméra 40 sont directement connectés par une porte, la transition entre les deux étant donc instantanée.



(a) Emplacement des caméras au 8e étage. (b) Graphe de topologie du réseau $\mathcal{G}(V, E)$.

FIGURE 4.6 – Topologie pour nos expérimentations pour les caméras 17, 18 et 40 de la base HDA.

Nous concentrons nos expériences sur des vidéos de trois durées différentes : 10, 15 et 30 minutes. Pour chaque échantillon, la détection temporelle médiane du *tracklet* est utilisée pour représenter l'échantillon. Pour chaque paire de caméras, nous comparons chaque échantillon de la première caméra avec chaque échantillon de la seconde. Comme pour la méthode NCR, deux descripteurs sont considérés : SDALF [Farenzena 2010] et WACN [Martinel 2012], implémentés grâce aux boîtes à outils MATLAB fournies par les auteurs. Pour chacune des deux familles de descripteurs, les performances obtenues par NCR [Das 2014, Chakraborty 2016] et notre approche D-NCR sont comparées et discutées.

Les modèles PLNE de NCR et D-NCR ont été implémentés en C++ en utilisant le solveur PLNE commercial Gurobi 7.5.1. Tous les tests expérimentaux sont effectués sur un processeur Intel® Core™ i5-4670 CPU de 3,4 GHz avec 16 Go de mémoire RAM DDR3 1600 MHz. Aucun GPU n'est utilisé.

Instance	Échantillons par caméra	NCR sur SDALF	D-NCR sur SDALF	NCR sur WACN	D-NCR sur WACN
0-10 minutes	11, 38 and 54	0.20s	4.23s	0.18s	4.55s
10-20 minutes	17, 27 and 43	1.49s	0.95s	3.27s	0.98s
20-30 minutes	25, 26 and 31	5.53s	0.77s	5.36s	0.79s
0-15 minutes	23, 53 and 73	15.82s	13.25s	22.54s	13.95s
15-30 minutes	28, 40 and 56	51.31s	5.26s	107.84s	5.17s
0-30 minutes	51, 90 and 126	61031s	188.05s	>45087s	201.39s

TABLE 4.2 – Instances relatives aux intervalles de temps, le nombre d'échantillons par caméra (caméras 17, 18 et 40 resp.) et le temps d'optimisation pour chaque modèle.

Une discussion sur la taille des instances PLNE et leur temps d'optimisation est fournie dans la sous-section suivante. Dans un deuxième temps, l'instance vidéo de 10 minutes est utilisée pour analyser l'impact des contraintes de topologie et de temps de transit sur la précision de la nouvelle identification. Enfin, une évaluation des performances est fournie pour les instances vidéo de 15 et 30 minutes.

4.3.1 Étude de la taille du problème et du temps d'optimisation

HDA proposant des vidéos de 30 minutes, 3 instances de vidéo de 10 minutes peuvent être considérées (0-10 , 10-20 et 20-30 minutes), 2 vidéos de 15 minutes (0-15 minutes et 15-30 minutes) et 1 vidéo de 30 minutes. La table 4.2 présente le nombre d'échantillons pour chacune des instances et le temps de traitement requis pour l'optimisation, pour NCR et D-NCR. Nous observons en premier lieu que D-NCR est beaucoup plus rapide que NCR afin de résoudre le problème de ré-identification. Cela tend à confirmer qu'une modélisation par un graphe acyclique orienté du problème de ré-identification est plus pertinent en termes de performance de calcul. Pour l'instance complète de 30 minutes, nous observons également que D-NCR s'adapte avec succès à cette énorme instance puisqu'une solution optimale est fournie par le solveur au bout de quelques minutes (tandis que le solveur ne parvient pas à fournir une solution optimale pour NCR après plus d'un jour de calcul). Cette expérience démontre que D-NCR est clairement plus approprié que NCR pour des applications avec contraintes de temps car celui-ci permet de traiter des vidéos de longue durée, ce qui n'est pas le cas pour NCR.

Pour approfondir, nous avons analysé la taille des instances. Le nombre maximum de variables pour une instance de notre problème peut être calculé comme la somme des produits des *tracklets* par paire de caméras, auquel il faut rajouter les variables de décision ajoutées pour les caméras fictives. Pour les contraintes (4.6)-(4.7), nous avons au plus autant de contraintes que de *tracklets* dans toutes les caméras. Pour les contraintes (4.8)-(4.10) nous avons autant de contraintes que de variables, exception faite des variables associées aux caméras fictives. Pour l'instance de 10 minutes nous avons donc au maximum 5407 contraintes. Pour l'instance de 30 minutes, le nombre de variables passe à 36178, pour 35911 contraintes. Cette taille reste tout à fait abordable pour le solveur mais il reste nécessaire de considérer des instances avec davantage de caméras pour consolider cette analyse (la base HDA permet de construire des réseaux de 5 à 8 caméras).

4.3.2 Contraintes de topologie et temps de transit

Pour mettre en évidence l'effet des contraintes de topologie et temps de transit, nous focalisons sur l'instance des 10 premières minutes (0-10 minutes). Nous avaons testé notre modèle D-NCR avec et sans prise en compte des temps de transit minimum afin de tester leur effet sur les performances de ré-ID. Les expériences sont menées sur HDA avec une mesure de la précision de premier rang. Les taux de reconnaissance sont synthétisés dans la table 4.3. Ces taux montrent que la topologie du réseau et les contraintes de temps améliorent bien les performances de ré-ID. En considérant la colonne de performance globale, nous observons un gain de $3,4\%$ (7/206) (resp. $4,34\%$ 9/206) par rapport à l'approche classique de ré-ID avec SDALF (resp. avec WACN) et $1,94\%$ (4/206) (resp. $3,88\%$ 8/206) par rapport à la méthode NCR sur SDALF (resp. sur WACN). Clairement, notre modèle D-NCR semble plus précis, même si les gains de performance peuvent être hétérogènes en fonction des paires de caméras considérées. Plus précisément, les contraintes topologiques ont clairement une incidence sur les performances intrinsèques des caméras 17 et 18. En outre, l'ajout de contraintes topologiques induit des gains pour la paire de caméras 17-18 : $9,09\%$ (1/11) (resp. $5,27\%$ 2/38) par rapport à NCR sur SDALF pour la paire de caméras 17-18 (resp. 18-17) et $9,09\%$ (1/11) (resp. $7,89\%$ 3/38) par rapport à NCR sur WACN pour la paire de caméras 17-18 (18-17). Les contraintes de temps de transit minimum ont également un impact sur les performances de la paire de caméras 40-17, car un gain de $1,85\%$ (1/54) (resp. $3,70\%$ 2/54) par rapport à NCR sur SDALF (resp. sur WACN) est observé.

4.3.3 Évaluation des performances

Pour une analyse plus approfondie des performances de D-NCR, nous nous intéressons maintenant aux instances de 15 minutes et à l'instance complète de 30 minutes. Comme précédemment, nous mesurons les taux de reconnaissance de premier rang. Pour les instances de 15 minutes, ils sont présentés dans la table 4.4. En considérant la performance globale, nous observons un gain de $2,4\%$ (13/546) (resp. $1,5\%$ 8/546) par rapport à l'approche classique de ré-ID avec SDALF (resp. avec WACN) et 2% (11/546) (resp. $1,6\%$ 9/546) par rapport à la méthode NCR sur SDALF (resp. sur WACN). Un point intéressant concerne la différence entre le ré-ID traditionnelle par optimisation locale et l'optimisation globale que nous proposons. Ceci est visible pour la paire de caméras 17-18, où D-NCR a 1 (resp. 2) vrai(s) positif(s) en moins par rapport à la ré-ID classique, mais globalement, 13 (resp. 8) vrais positifs en plus pour SDALF (resp. WACN).

Les taux de reconnaissance de premier rang pour l'instance complète de 30 minutes sont présentés dans la table 4.5. En considérant la performance global, nous observons un gain de $2,1\%$ (11/534) (resp. $1,1\%$ 6/534) par rapport à l'approche classique de ré-ID avec SDALF (resp. avec WACN) et $2,4\%$ (13/534) (respectivement $1,3\%$ (7/534)) par rapport à la méthode NCR sur SDALF (resp. sur WACN). Nous pouvons aussi observer dans les résultats la différence entre l'optimisation lo-

	17-18	18-17	17-40	40-17	18-40	40-18	Global
SDALF	18.18% (2/11)	21.05% (8/38)	27.27% (3/11)	22.22% (12/54)	34.21% (13/38)	35.19% (19/54)	27.67% (57/206)
NCR sur SDALF	18.18% (2/11)	21.05% (8/38)	27.27% (3/11)	24.07% (13/54)	36.84% (14/38)	37.04% (20/54)	29.13% (60/206)
D-NCR sur SDALF*	27.27% (3/11)	26.32% (10/38)	27.27% (3/11)	24.07% (13/54)	36.84% (14/38)	35.19% (19/54)	30.10% (62/206)
D-NCR sur SDALF	27.27% (3/11)	26.32% (10/38)	27.27% (3/11)	25.93% (14/54)	36.84% (14/38)	37.04% (20/54)	31.07% (64/206)
WACN	9.09% (1/11)	15.79% (6/38)	36.36% (4/11)	11.11% (6/54)	39.47% (15/38)	29.63% (16/54)	23.30% (48/206)
NCR sur WACN	9.09% (1/11)	15.79% (6/38)	36.36% (4/11)	11.11% (6/54)	39.47% (15/38)	31.48% (17/54)	23.79% (49/206)
D-NCR sur WACN*	18.18% (2/11)	23.68% (9/38)	36.36% (4/11)	12.96% (7/54)	44.74% (17/38)	31.48% (17/54)	27.19% (56/206)
D-NCR sur WACN	18.18% (2/11)	23.68% (9/38)	36.36% (4/11)	14.82% (8/54)	44.74% (17/38)	31.48% (17/54)	27.67% (57/206)

* Sans contraintes de temps de transit minimum.

TABLE 4.3 – Taux de reconnaissance de premier rang par paires de caméras sur l'instance de 0-10 minutes.

	17-18	18-17	17-40	40-17	18-40	40-18	Global
SDALF	37.26% (19/51)	29.03% (27/93)	33.33% (17/51)	27.13% (35/129)	36.56% (34/93)	45.74% (59/129)	34.98% (191/546)
NCR sur SDALF	37.26% (19/51)	29.03% (27/93)	33.33% (17/51)	28.68% (37/129)	36.56% (34/93)	45.74% (59/129)	35.35% (193/546)
D-NCR sur SDALF	33.33% (17/51)	34.41% (32/93)	37.25% (19/51)	29.46% (38/129)	39.78% (37/93)	47.29% (61/129)	37.36% (204/546)
WACN	35.29% (18/51)	25.81% (24/93)	27.45% (14/51)	23.26% (30/129)	38.71% (36/93)	38.76% (50/129)	31.50% (172/546)
NCR sur WACN	35.29% (18/51)	25.81% (24/93)	27.45% (14/51)	23.26% (30/129)	37.63% (35/93)	38.76% (50/129)	31.32% (171/546)
D-NCR sur WACN	33.33% (17/51)	32.26% (30/93)	27.45% (14/51)	24.81% (32/129)	39.78% (37/93)	38.76% (50/129)	32.97% (180/546)

TABLE 4.4 – Taux de reconnaissance de premier rang par paires de caméras sur les instances de 15 minutes.

cale classique et l'optimisation globale. Dans ce cas, cela peut être vu pour la paire de caméras 17-40, où D-NCR sur WACN a 1 vrai positif en moins par rapport à la re-ID classique avec WACN mais, globalement, 6 vrais positifs en plus.

4.3.4 Exemples de scénario

Nous montrons ci-après deux exemples d'appariements erronés inférés par SDALF mais évités par notre stratégie D-NCR grâce aux contraintes de type réseau.

(a) Caméra 17 identité 4 - caméra 18 identité 10.

(b) Caméra 17 identité 7 - caméra 18 identité 36.

(c) Caméra 17 identité 9 - caméra 18 identité 20.

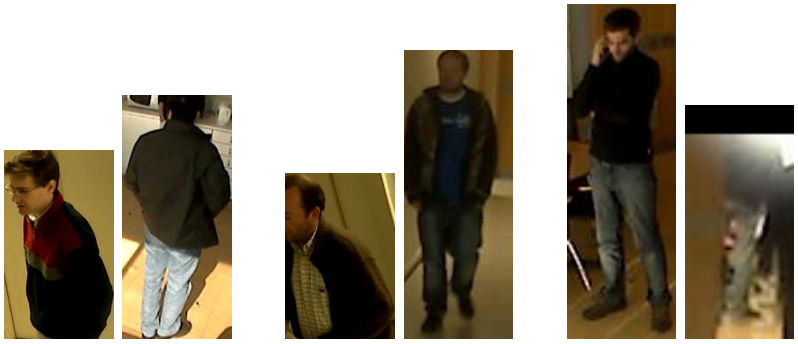
FIGURE 4.7 – Exemples d'appariements par SDALF sur la base HDA qui sont corrigées par des contraintes de topologie.

La figure 4.7 illustre des exemples d'appariements faits par la ré-ID classique avec SDALF entre les caméras 17 et 18. Ces appariements sont évités grâce aux contraintes de topologie de réseau intégrées dans notre formulation. Dans la ré-ID classique, la paire de caméras 17-18 est traitée comme toutes les paires de caméras, c'est-à-dire qu'il est permis qu'une personne passe directement de la caméra 17 à la caméra 18. Or, sur le graphe figure 4.6b, ceci n'est plus possible dans notre approche D-NCR car il est nécessaire de transiter au préalable par la caméra 40. En d'autres termes, tout appariement entre les caméras 17-18 doit être la conséquence des appariements inférés entre les caméras 17-40 et 18-40.

La figure 4.8 illustre un autre exemple d'appariements inférés par ré-ID classique avec SDALF. Ces appariements sont filtrés par les contraintes temporelles de notre stratégie D-NCR. En effet, dans la ré-ID classique, l'horodatage des identités n'est pas pris en compte ce qui peut conduire à apparier des identités incohérentes temporellement sur deux caméras. Les appariements figure 4.8 illustrent ce cas. Nos contraintes temporelles permettent explicitement d'éviter ces erreurs.

	17-18	18-17	17-40	40-17	18-40	40-18	Global
SDALF	33.33% (17/51)	34.44% (31/90)	19.61% (10/51)	32.54% (41/126)	35.56% (32/90)	46.03% (58/126)	35.39% (189/534)
NCR sur SDALF	33.33% (17/51)	34.44% (31/90)	19.61% (10/51)	32.54% (41/126)	34.44% (31/90)	45.24% (57/126)	35.02% (187/534)
D-NCR sur SDALF	37.25% (19/51)	36.67% (33/90)	21.57% (11/51)	34.92% (43/126)	37.78% (34/90)	46.83% (59/126)	37.45% (200/534)
WACN	31.37% (16/51)	27.78% (25/90)	23.53% (12/51)	32.54% (41/126)	33.33% (30/90)	35.71% (45/126)	31.65% (169/534)
NCR sur WACN	33.33% (17/51)	27.78% (25/90)	19.61% (10/51)	32.54% (41/126)	33.33% (30/90)	35.71% (45/126)	31.46% (168/534)
D-NCR sur WACN	33.33% (17/51)	31.11% (28/90)	21.57% (11/51)	34.13% (43/126)	34.44% (31/90)	35.71% (45/126)	32.77% (175/534)

TABLE 4.5 – Taux de reconnaissance de premier rang par paires de caméras sur l'instance de 30 minutes.



(a) Caméra 17 identité 2 - caméra 18 identité 29.

(b) Caméra 17 identité 6 - caméra 40 identité 40.

(c) Caméra 18 identité 1 - caméra 40 identité 10.

FIGURE 4.8 – Exemples d'appariements inférés par SDALF sur la base HDA mais évités par nos contraintes temporelles.

4.4 Perspectives d'évolution et poursuite des investigations

Les expérimentations préliminaires détaillées dans la partie précédente valident la cohérence de notre approche pour la ré-ID et confirment sa pertinence. Cependant, de nouvelles investigations sont à faire. Tout d'abord, une campagne d'expérimentation plus poussée et une analyse plus rigoureuse seraient à mener pour évaluer l'efficacité de notre méthode et mieux cerner ses limites, notamment relativement à la taille des instances de problèmes pouvant être résolues en temps acceptable. D'autre part, un meilleur paramétrage de nos modèles d'optimisation serait également une source potentielle d'amélioration à étudier. Enfin, d'autres méthodes d'optimisation pourraient être utilisées pour résoudre plus efficacement notre problème.

4.4.1 Besoin de plus d'expérimentations

Comme indiqué auparavant, un des principaux avantages de notre méthode est qu'elle peut être utilisée quels que soient les descripteurs et métriques utilisés. Toutefois, il serait pertinent d'évaluer notre approche sur un ensemble plus vaste de descripteurs et métriques pour déterminer quelles sont les associations les plus judicieuses en termes de performances absolues de ré-ID mais aussi en termes de temps de calcul. Il serait ainsi intéressant de mesurer l'impact de notre modélisation en fonction du pouvoir discriminant du descripteur.

De plus, la base de données HDA pourrait être mise à profit pour extraire plusieurs catégories d'instances de complexité croissante en jouant tout à la fois sur la taille du réseau de caméras et la longueur des séquences vidéos. Enfin, il serait pertinent d'évaluer sur une autre base de données afin de corroborer les résultats obtenus.

4.4.2 Paramétrisation du modèle d'optimisation

Dans nos expérimentations préliminaires, les temps minimum de transit entre caméras ont tous été fixés à zéro car ils n'étaient pas connus dans la base de données HDA. De ce fait, le renforcement de cohérence temporelle de la ré-ID est faible dans nos expérimentations préliminaires. Afin d'obtenir une ré-ID plus performante, il serait intéressant de caractériser, par une analyse statistique de la base de données, des temps de transit minimum et/ou moyen, voire d'associer une distribution de probabilité à ces grandeurs. Utiliser un temps de transit minimum supérieur à zéro permet d'améliorer la cohérence temporelle de la ré-ID. Utiliser un temps moyen de transit peut être préférable mais risque de conduire à un sur-renforcement induisant le rejet de bonnes solutions. Introduire une fonction de probabilité permettrait de contrôler ce risque et de le chiffrer.

Un autre paramètre pouvant influencer grandement l'efficacité de notre méthode concerne les scores de similarité associés aux appariements mettant en jeu les caméras fictives 0 et $m + 1$. Rappelons que la caméra 0 associe à chaque détection une détection soeur virtuelle telle que, si la détection est appariée avec sa soeur, on déduirait que la personne entre pour la première fois dans le réseau. De même, la caméra $m + 1$ associe à chaque détection une autre détection virtuelle qui, en cas d'appariement, indique que la personne sort définitivement du réseau. Or, ce choix de modélisation n'est pas nécessairement judicieux car il existe des réseaux de caméra pour lesquels on sait pertinemment qu'il est impossible qu'une personne entre ou sorte au niveau d'une caméra. Cette information pourrait donc être facilement intégrée dans notre modèle en évitant de générer les détections virtuelles soeurs lorsque ce n'est pas nécessaire. De plus, dans notre modèle, les scores de similarité associés aux détections virtuelles sont fixés à zéro, ce qui est la pire valeur de similarité possible. Une conséquence malheureuse de ce choix est que deux chemins disjoints et temporellement consistants, correspondant aux parcours de deux personnes distinctes dans le réseau de caméras, seront mis bout-à-bout par notre modèle afin d'augmenter la valeur de la fonction objectif (le score de similarité entre la dernière détection de la première personne et la première détection de la deuxième personne étant strictement supérieur à zéro). Ici aussi, il est intéressant d'améliorer le paramétrage des scores de similarité sur les détections virtuelles. On pourrait par exemple fixer comme valeur le minimum de tous les scores de similarité, ce qui améliorerait la cohérence de la solution. De façon générale, une étude des scores de similarité doit être mené afin d'associer aux appariements virtuels des récompenses qui soient les plus grandes possibles, sans toutefois biaiser la performance.

4.4.3 Autres méthodes d'optimisation

Nous avons utilisé dans notre approche un solveur de programmation mathématique pour résoudre notre problème d'optimisation. Même si ces solveurs sont de plus en plus performants, ils n'utilisent peu ou pas les propriétés du modèle. Or, dans le cadre de notre étude, il est tentant d'utiliser d'autres outils.

Une première technique que nous pourrions utiliser est basée sur la propriété que la relaxation des contraintes (4.8)-(4.10) réduit notre problème à une recherche classique de chemins disjoints [Robertson 1995] à profit maximum, ainsi que mentionné précédemment. Or, ce problème peut être résolu très efficacement grâce à des algorithmes de recherche de flot maximum à coût minimum (par exemple, [Goldberg 1990]). Une idée serait donc d'exploiter cette propriété au sein d'une recherche arborescente de type *branch-and-bound*. Nous considérons une arborescence binaire où chaque nœud est séparé en deux nœud fils considérant un arc $\left(\mathcal{P}_i^p, \mathcal{P}_j^q\right) \in Y$ de sorte que $x_{i,j}^{p,q} = 0$ ou 1 dans chacune des branches. Une solution du problème relâché est trouvée à l'aide d'un algorithme de flot. Cette solution fournit une borne supérieure de notre problème (toutes les variables $x_{i,j}^{p,q}$ associées à des arcs de Z $\left(\mathcal{P}_i^p, \mathcal{P}_j^q\right) \in X$ non fixés à 0 lors du branchement étant naturellement forcées à 1 du fait du sens de l'optimisation). A partir de la solution relâchée, il est facile de calculer une nouvelle solution en empêchant les variables $x_{i,j}^{p,q}$ associées à des arcs de Z d'être mises à 1 si l'arc $\left(\mathcal{P}_i^p, \mathcal{P}_j^q\right) \in X$ n'est pas impliqué dans une fermeture transitive d'un chemin. on obtient alors une borne inférieure. Nous pensons que cette recherche arborescente pourrait s'avérer plus efficace en termes de temps de calcul et de taille de problème traitable qu'un solveur de programmation mathématique classique.

La méthode ad-hoc évoquée au paragraphe précédent est une méthode exacte qui garantit, comme le solveur de programmation mathématique, l'optimalité de la solution trouvée. Elle peut donc nécessiter un temps de calcul conséquent. Si on accepte à présent de ne plus garantir l'optimalité de la solution trouvée et que l'on se contente d'une "bonne" solution, des méthodes beaucoup plus performantes, capables de traiter en très peu de temps des instances de très grande taille peuvent être utilisées. Il s'agit de méthodes heuristiques ou métaheuristiques qui permettent de converger rapidement vers une solution. Nous renvoyons le lecteur vers le livre de Bozorg-Haddad *et al.* [BozorgHaddad 2017] pour une description plus approfondie de ces méthodes.

4.5 Conclusions

Ce chapitre présente une modélisation originale de ré-ID cohérente au réseau avec des considérations explicites de temps de transition et de topologie. Notre formalisation D-NCR est prometteuse ; elle améliore les performances par rapport aux stratégies classiques de ré-ID, dont NCR. Notons également que notre *framework* peut être étendu à tout type de descripteur et métrique de similarité. Il est aussi indépendant du contexte applicatif et peut donc être utilisé dans d'autres applications reliées à l'association d'identités dans des réseaux.

Les travaux présentés ici ont donné lieu à une publication : congrès annuel de la société Française de Recherche Opérationelle et d'Aide à la Décision (ROADEF) [BarbosaAnda 2018b].

Ces travaux ouvrent de belles perspectives. Les nombreuses améliorations devraient induire des gains substantiels de performances en relatif (comme pointées ici) mais aussi en absolu. Celles-ci sont listées en section 4.4.

Conclusions générales et perspectives

Nos travaux se focalisent sur le développement d'un système de vidéo-surveillance à partir d'un réseau de caméras perspectives et à champs disjoints. Deux modalités de détection de personnes et de ré-ID au sein du réseau sont prototypées dans ce contexte applicatif. Ces modalités sont originales eu égard à leur cahier des charges respectif et leur modélisation qui s'inspire de techniques de Recherche Opérationnelle.

Concernant la modalité détection de personnes, nous avons étudié un détecteur type *soft-cascade* inspiré du détecteur proposé par Dollár *et al.* [Dollár 2014]. Le chapitre 2 focalise sur l'impact des seuils de détection de la *soft-cascade* sur : (1) les performances de détection et (2) le temps de réponse de la *soft-cascade*. Une approche originale basée sur la PLNE est proposée afin de minimiser le temps de réponse de la *soft-cascade*, la performance de détection désirée étant imposée. Nous avons montré que ce problème est NP-difficile et mis en évidence certaines de ses propriétés. Celles-ci nous ont permis de reformuler notre problème et de proposer des heuristiques de résolution approchée. Ces heuristiques sont basées sur une recherche locale dans un graphe. Plusieurs variantes ont été proposées. La première consiste à définir un voisinage dans l'espace de recherche des seuils de détection. La deuxième consiste à réduire la longueur de la *soft-cascade* et, par conséquent, l'espace de recherche des seuils de détection. Une troisième variante a été proposée qui combine les deux idées. Cette stratégie hybride a été implémentée pour optimiser les seuils de détection d'une *soft-cascade* puis évaluée sur les bases publiques d'images INRIA [Dalal 2005] et Caltech [Dollár 2009a, Dollár 2012]. Les évaluations associées ont montré que notre méthode permet effectivement de réduire le temps de réponse de la *soft-cascade* sans affecter les performances de détection. Ces travaux ont été publiés dans ICORES 2016 [BarbosaAnda 2016a], ROADEF 2016 [BarbosaAnda 2016b] et WACV 2018 [BarbosaAnda 2018a].

Concernant la modalité de ré-ID de personnes, nous avons répertorié les stratégies de décision avec un focus sur les méthodes globales de ré-ID qui assurent une meilleure cohérence au niveau réseau. En s'inspirant de la stratégie NCR [Das 2014, Chakraborty 2016], nous avons proposé au chapitre 4 une modélisation originale du problème de Ré-ID avec prise en compte des contraintes de topologie et de temps de transit. Les contraintes de topologie réseau utilisées intègrent les diverses trajectoires possibles entre chaque paire de caméras. Les contraintes temporelles donnent une orientation aux appariements et garantissent le respect des temps minimum de transit entre caméras. Des expérimentations ont été réalisées sur la base de données publique HDA [Nambiar 2014, Figueira 2015]. Des évaluations préliminaires ont montré que les contraintes de topologie réseau et temporelles améliorent sensiblement les performances globales de ré-ID. Plusieurs pistes ont été énumérées afin de

mieux paramétrer notre modèle. Celles-ci devraient à terme améliorer substantiellement les performances de ré-ID. Enfin, nous avons également présenté une méthode de résolution basée sur une recherche arborescente qui permettra de résoudre des instances de grande taille. Ces travaux ont été publiés dans ROADEF 2018 [BarbosaAnda 2018b].

Perspectives

Si notre détecteur de personnes a montré des bonnes performances et temps de réponse, des améliorations sont toujours envisageables. Nous avons identifié trois axes d'investigation : (1) l'amélioration des méthodes de recherche de solution, (2) l'incorporation de classificateurs faibles hétérogènes, et (3) la généralisation à un ordre quelconque des classificateurs faibles.

Sur le point (1), des investigations plus approfondies pourraient être menées pour déterminer la forme et la taille de l'enveloppe de l'espace de recherche la plus performante pour notre méthode : il s'agit de déterminer un sous-espace de recherche offrant le meilleur compromis entre la taille et la qualité des solutions qu'il contient. D'autres investigations peuvent être envisagées pour trouver des nouvelles heuristiques de résolution approchée du problème.

Concernant le point (2), notre modélisation peut intuitivement s'adapter à des classificateurs faibles hétérogènes. Pour se faire il faut choisir une méthode de construction de la *soft-cascade* qui permet de choisir les classificateurs faibles appropriés pour le détecteur, à l'instar de Mekonnen *et al.* [Mekonnen 2014]. Pour le point (3), le modèle PLNE peut être généralisé pour choisir le classificateur faible à utiliser a chaque niveau de la *soft-cascade* d'une base de classificateurs faibles (possiblement hétérogènes) et ses seuil de classification respectif. Ce problème est plus complexe puisqu'il intègre en plus une dimension d'affectation. Sa résolution nécessite la conception de nouvelles méthodes approchées de recherche de solutions.

Comme évoqué au chapitre 4, de nombreuses investigations sont envisagées concernant notre méthodologie de ré-ID de personnes. Nous avons identifié trois axes de travail : le réglage des paramètres libres de notre modélisation, l'intégration des contraintes de topologie réseau et temporelles restantes, enfin l'implémentation et l'évaluation de la méthode de solution proposée. Nous pourrons étudier l'impact de la pondération des arcs vers la caméra fictive de sortie dans le modèle sur les performances. D'autres considérations topologiques et temporelles restent à exploiter. Dans nos futures expérimentations, le temps minimal de transit entre caméras reste à intégrer avec des valeurs pertinentes. Une contrainte topologique supplémentaire concerne les entrées et sorties de personnes du réseau de caméras, car celles-ci ne sont possibles que sur des caméras spécifiques. Cette nouvelle contrainte permettra de mieux caractériser les débuts et fins des trajectoires des personnes au sein du réseau de caméras. Enfin, nous proposons une variante basée sur une recherche arborescente. Celle-ci reste à implémenter et doit permettre de gérer des instances plus grandes en termes de nombre de caméras et durée d'observation.

Ces travaux focalisent sur les modalités détection de personnes et ré-ID de personnes du système complet de vision ; celles-ci sont traitées indépendamment. L'objectif ultime est de traiter simultanément la chaîne complète c'est-à-dire depuis l'acquisition du flux jusqu'à l'inférence de ré-ID. Des travaux antérieurs [Mekonnen 2018] ont traité conjointement les modalités détection et suivi de personnes. Nous pouvons faire de même en ajoutant la modalité ré-ID de personnes. Après avoir étudié le système complet, les diverses modalités associées pourront être intégrées sur notre plateforme multi-caméras qui instrumente le bâtiment ADREAM du LAAS-CNRS. Cette plateforme, aujourd'hui opérationnelle, est composée de 5 caméras synchrones couleurs type blackfly, déployées sur deux étages du bâtiment. Cette intégration *online* constituerait l'aboutissement ultime de ce travail.

Bibliographie

[Alexe 2010] B. Alexe, T. Deselaers et V. Ferrari. « What is an object ? » Dans : *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Juin 2010, p. 73–80 (Cité en page 17.).

[Arampatzis 2005] T. Arampatzis, J. Lygeros et S. Manesis. « A Survey of Applications of Wireless Sensors and Wireless Sensor Networks ». Dans : *Proceedings of the 2005 IEEE International Symposium on, Mediterrean Conference on Control and Automation Intelligent Control, 2005.* Juin 2005, p. 719–724 (Cité en page 1.).

[Avraham 2012] Tamar Avraham, Ilya Gurvich, Michael Lindenbaum et Shaul Markovitch. « Learning Implicit Transfer for Person Re-identification ». Dans : *Computer Vision – ECCV 2012. Workshops and Demonstrations*. Sous la dir. d'Andrea Fusiello, Vittorio Murino et Rita Cucchiara. Berlin, Heidelberg : Springer Berlin Heidelberg, 2012, p. 381–390 (Cité en page 73.).

[Baltieri 2011] Davide Baltieri, Roberto Vezzani et Rita Cucchiara. « 3DPeS : 3D People Dataset for Surveillance and Forensics ». Dans : *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*. J-HGBU '11. Scottsdale, Arizona, USA : ACM, 2011, p. 59–64 (Cité en pages 78 et 79.).

[BarbosaAnda 2016a] F. Barbosa-Anda, C. Briand, F. Lerasle et A. Mekonnen. « Mean Response-Time Minimization of a Soft-Cascade Detector ». Dans : *Int. Conf. on Operations Research and Enterprise Systems (ICORES'16)*. Rome, Italy, fév. 2016, p. 252–260 (Cité en pages 5, 41, 65 et 109.).

[BarbosaAnda 2016b] Francisco Rodolfo Barbosa-Anda, Cyril Briand, Frédéric Lerasle et Alhayat Ali Mekonnen. « Minimisation du Temps de Réponse moyen d'une Cascade de Détection ». Dans : *Congrès annuel de la société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF)*. Compiègne, France, fév. 2016, 2p. (Cité en pages 5, 65 et 109.).

[BarbosaAnda 2018a] F. R. Barbosa-Anda, F. Lerasle, C. Briand et A. A. Mekonnen. « Soft-Cascade Learning with Explicit Computation Time Considerations ». Dans : *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Mar. 2018, p. 1234–1243 (Cité en pages 5, 50, 65 et 109.).

[BarbosaAnda 2018b] Francisco Rodolfo Barbosa-Anda, Cyril Briand et Frédéric Lerasle. « Partitionnement en cliques à profit maximum de graphes orientés avec contraintes de flot ». Dans : *Congrès annuel de la société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF)*. Lorient, France, fév. 2018, 2p. (Cité en pages 5, 107 et 110.).

[Bazzani 2010] L. Bazzani, M. Cristani, A. Perina, M. Farenzena et V. Murino. « Multiple-Shot Person Re-identification by HPE Signature ». Dans : *2010 20th International Conference on Pattern Recognition*. Août 2010, p. 1413–1416 (Cité en pages 70, 72, 75 et 76.).

[Bazzani 2013] Loris Bazzani, Marco Cristani et Vittorio Murino. « Symmetry-driven accumulation of local features for human characterization and re-identification ». Dans : *Computer Vision and Image Understanding* 117.2 (2013), p. 130–144 (Cité en pages 73 et 82.).

[Beauchemin 1995] S. S. Beauchemin et J. L. Barron. « The Computation of Optical Flow ». Dans : *ACM Comput. Surv.* 27.3 (sept. 1995), p. 433–466 (Cité en page 14.).

[Benenson 2015] Rodrigo Benenson, Mohamed Omran, Jan Hosang et Bernt Schiele. « Ten Years of Pedestrian Detection, What Have We Learned ? » Dans : *Computer Vision - ECCV 2014 Workshops*. Sous la dir. de Lourdes Agapito, Michael M. Bronstein et Carsten Rother. Cham : Springer International Publishing, 2015, p. 613–627 (Cité en page 80.).

[Bennewitz 2005] M. Bennewitz, F. Faber, D. Joho, M. Schreiber et S. Behnke. « Towards a humanoid museum guide robot that interacts with multiple persons ». Dans : *5th IEEE-RAS International Conference on Humanoid Robots, 2005.* Déc. 2005, p. 418–423 (Cité en page 1.).

[Bialkowski 2012] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes et P. Lucey. « A Database for Person Re-Identification in Multi-Camera Surveillance Networks ». Dans : *2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*. Déc. 2012, p. 1–8 (Cité en pages 79 et 80.).

[Bouma 2013] Henri Bouma, Jan Baan, Sander Landsmeer anda Chris Kruszynski, Gert van Antwerpen et Judith Dijk. « Real-time tracking and fast retrieval of persons in multiple surveillance cameras of a shopping mall ». Dans : *Proc.SPIE*. T. 8756. 2013, (Cité en page 1.).

[Bourdev 2005] L. Bourdev et J. Brandt. « Robust object detection via soft cascade ». Dans : *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. T. 2. Juin 2005, 236–243 vol. 2 (Cité en pages 11, 22, 38 et 40.).

[BozorgHaddad 2017] Omid Bozorg-Haddad, Mohammad Solgi et Hugo A. Loáiciga. *Meta-Heuristic and Evolutionary Algorithms for Engineering Optimization*. Wiley-Blackwell, 2017. eprint : `https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119387053` (Cité en page 107.).

[Breitenstein 2011] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier et L. Van Gool. « Online Multiperson Tracking-by-Detection from a Single, Uncalibrated Camera ». Dans : *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33.9 (2011), p. 1820–1833 (Cité en page 9.).

[Broggi 2000] A. Broggi, M. Bertozzi, A. Fascioli et M. Sechi. « Shape-based pedestrian detection ». Dans : *Proceedings of the IEEE Intelligent Vehicles Symposium 2000 (Cat. No.00TH8511)*. 2000, p. 215–220 (Cité en page 15.).

[Broggi 2006] A. Broggi, R. I. Fedriga, A. Tagliati, T. Graf et M. Meinecke. « Pedestrian Detection on a Moving Vehicle : an Investigation about Near Infra-Red Images ». Dans : *2006 IEEE Intelligent Vehicles Symposium*. 2006, p. 431–436 (Cité en page 15.).

[Cai 2015] Z. Cai, M. Saberian et N. Vasconcelos. « Learning Complexity-Aware Cascades for Deep Pedestrian Detection ». Dans : *2015 IEEE International Conference on Computer Vision (ICCV)*. Déc. 2015, p. 3361–3369 (Cité en page 34.).

[Cao 2016a] J. Cao, Y. Pang et X. Li. « Pedestrian Detection Inspired by Appearance Constancy and Shape Symmetry ». Dans : *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2016, p. 1316–1324 (Cité en pages 30, 31, 34, 58, 60, 61, 64 et 65.).

[Cao 2016b] J Cao, Y. Pang et X. Li. « Pedestrian Detection Inspired by Appearance Constancy and Shape Symmetry ». Dans : *IEEE Trans. on Image Processing* 25.12 (déc. 2016), p. 5538–5551 (Cité en page 65.).

[Chakraborty 2016] A. Chakraborty, A. Das et A. K. Roy-Chowdhury. « Network Consistent Data Association ». Dans : *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI'16)* 38.9 (sept. 2016), p. 1859–1871 (Cité en pages 70, 77, 81, 89, 90, 95, 98 et 109.).

[Chen 2015] Ying-Cong Chen, Wei-Shi Zheng et Jianhuang Lai. *Mirror Representation for Modeling View-Specific Transform in Person Re-Identification*. 2015 (Cité en page 74.).

[Cheng 2011] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani et Vittorio Murino. « Custom Pictorial Structures for Re-identification ». Dans : *Proceedings of the British Machine Vision Conference*. BMVA Press, 2011, p. 68.1–68.11 (Cité en pages 78 et 79.).

[Dalal 2005] N. Dalal et B. Triggs. « Histograms of oriented gradients for human detection ». Dans : *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. T. 1. Juin 2005, 886–893 vol. 1 (Cité en pages 11, 15, 18, 19, 21, 22, 23, 24, 29, 30, 31, 56 et 109.).

[Dalal 2006a] Navneet Dalal. « Finding People in Images and Videos ». Theses. Institut National Polytechnique de Grenoble - INPG, juil. 2006 (Cité en page 11.).

[Dalal 2006b] Navneet Dalal, Bill Triggs et Cordelia Schmid. « Human Detection Using Oriented Histograms of Flow and Appearance ». Dans : *Computer Vision – ECCV 2006 : 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part II*. Sous la dir. d'Aleš Leonardis, Horst Bischof et Axel Pinz. Berlin, Heidelberg : Springer Berlin Heidelberg, 2006, p. 428–441 (Cité en pages 19 et 21.).

[Das 2014] Abir Das, Anirban Chakraborty et Amit K. Roy-Chowdhury. « Consistent Re-identification in a Camera Network ». Dans : *Europ. Conf. on Computer Vision (ECCV'14)*. Sous la dir. de David Fleet, Tomas Pajdla, Bernt Schiele et Tinne Tuytelaars. Cham : Springer International Publishing, 2014, p. 330–345 (Cité en pages 70, 72, 77, 79, 80, 81, 82, 88, 90, 95, 97, 98 et 109.).

[Davis 2007] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra et Inderjit S. Dhillon. « Information-theoretic Metric Learning ». Dans : *Proceedings of the 24th International Conference on Machine Learning*. ICML '07. Corvalis, Oregon, USA : ACM, 2007, p. 209–216 (Cité en pages 73 et 74.).

[Dollár 2008] Piotr Dollár, Boris Babenko, Serge Belongie, Pietro Perona et Zhuowen Tu. « Multiple Component Learning for Object Detection ». Dans : *Computer Vision – ECCV 2008 : 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II*. Sous la dir. de David Forsyth, Philip Torr et Andrew Zisserman. Berlin, Heidelberg : Springer Berlin Heidelberg, 2008, p. 211–224 (Cité en page 16.).

[Dollár 2009a] P. Dollár, C. Wojek, B. Schiele et P. Perona. « Pedestrian detection : A benchmark ». Dans : *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Juin 2009, p. 304–311 (Cité en pages 11, 15, 22, 56 et 109.).

[Dollár 2009b] Piotr Dollár, Zhuowen Tu, Pietro Perona et Serge Belongie. « Integral Channel Features ». Dans : *Proceedings of the British Machine Vision Conference*. doi :10.5244/C.23.91. BMVA Press, 2009, p. 91.1–91.11 (Cité en pages 20 et 21.).

[Dollár 2012] P. Dollár, C. Wojek, B. Schiele et P. Perona. « Pedestrian Detection : An Evaluation of the State of the Art ». Dans : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.4 (avr. 2012), p. 743–761 (Cité en pages 9, 10, 11, 15, 16, 18, 19, 22, 25, 27, 29, 30, 31, 32, 33, 34, 40, 56, 57, 61 et 109.).

[Dollár 2014] P. Dollár, R. Appel, S. Belongie et P. Perona. « Fast Feature Pyramids for Object Detection ». Dans : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.8 (août 2014), p. 1532–1545 (Cité en pages 20, 21, 22, 23, 29, 30, 31, 40, 56, 57 et 109.).

[Dollár 2016] Piotr Dollár. *Piotr's Computer Vision Matlab Toolbox (PMT)*. `https://github.com/pdollar/toolbox`. 2016 (Cité en pages 29 et 56.).

[DOrazio 2015] Tiziana D'Orazio et Cataldo Guaragnella. « A Survey of Automatic Event Detection in Multi-Camera Third Generation Surveillance Systems ». Dans : *International Journal of Pattern Recognition and Artificial Intelligence* 29.01 (2015), p. 1555001. eprint : `https://doi.org/10.1142/S0218001415550010` (Cité en page 1.).

[Enzweiler 2009] M. Enzweiler et D. M. Gavrila. « Monocular Pedestrian Detection : Survey and Experiments ». Dans : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.12 (déc. 2009), p. 2179–2195 (Cité en page 15.).

[Ess 2010] A. Ess, K. Schindler, B. Leibe et L. Van Gool. « Object Detection and Tracking for Autonomous Navigation in Dynamic Environments ». Dans : *The International Journal of Robotics Research* 29.14 (2010), p. 1707–1725 (Cité en page 9.).

[Everett 2003] H. R. Everett. « Robotic security systems ». Dans : *IEEE Instrumentation Measurement Magazine* 6.4 (déc. 2003), p. 30–34 (Cité en page 1.).

[Everingham 2010] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn et A. Zisserman. « The Pascal Visual Object Classes (VOC) Challenge ». Dans : *International Journal of Computer Vision* 88.2 (2010), p. 303–338 (Cité en page 10.).

[Everingham 2015] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn et Andrew Zisserman. « The Pascal Visual Object Classes Challenge : A Retrospective ». Dans : *International Journal of Computer Vision* 111.1 (jan. 2015), p. 98–136 (Cité en page 10.).

[Farenzena 2010] M. Farenzena, L. Bazzani, A. Perina, V. Murino et M. Cristani. « Person re-identification by symmetry-driven accumulation of local features ». Dans : *Int. Conf. Computer Vision and Pattern Recognition (CVPR'10)*. Juin 2010, p. 2360–2367 (Cité en pages 70, 72, 73, 75, 76, 82, 83, 89 et 98.).

[Felzenszwalb 2005] Pedro F. Felzenszwalb et Daniel P. Huttenlocher. « Pictorial Structures for Object Recognition ». Dans : *International Journal of Computer Vision* 61.1 (jan. 2005), p. 55–79 (Cité en page 15.).

[Felzenszwalb 2010] P. F. Felzenszwalb, R. B. Girshick, D. McAllester et D. Ramanan. « Object Detection with Discriminatively Trained Part-Based Models ». Dans : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.9 (sept. 2010), p. 1627–1645 (Cité en pages 11, 16, 19, 22, 23, 30 et 31.).

[Figueira 2015] Dario Figueira, Matteo Taiana, Athira Nambiar, Jacinto Nascimento et Alexandre Bernardino. « The HDA+ Data Set for Research on Fully Automated Re-identification Systems ». Dans : *Europ. Conf. on Computer Vision (ECCV'15) - Workshops*. Sous la dir. de Lourdes Agapito, Michael M. Bronstein et Carsten Rother. Cham : Springer International Publishing, 2015, p. 241–255 (Cité en pages 79, 80, 82, 89, 97 et 109.).

[Fischler 1973] M. A. Fischler et R. A. Elschlager. « The Representation and Matching of Pictorial Structures ». Dans : *IEEE Transactions on Computers* C-22.1 (jan. 1973), p. 67–92 (Cité en page 15.).

[Foucher 2011] Samuel Foucher, Marc Lalonde et Langis Gagnon. « A system for airport surveillance : detection of people running, abandoned objects, and pointing gestures ». Dans : *Proc.SPIE*. T. 8056. 2011, (Cité en page 1.).

[Garey 1979] Michael R. Garey et David S. Johnson. *Computers and Intractability : A Guide to the Theory of NP-Completeness*. New York, NY, USA : W. H. Freeman & Co., 1979 (Cité en page 43.).

[Gavrila 1999] D. M. Gavrila et V. Philomin. « Real-time object detection for ldquo ;smart rdquo ; vehicles ». Dans : *Proceedings of the Seventh IEEE International Conference on Computer Vision*. T. 1. 1999, 87–93 vol.1 (Cité en page 15.).

[Gavrila 2000] D. M. Gavrila. « Pedestrian Detection from a Moving Vehicle ». Dans : *Computer Vision — ECCV 2000 : 6th European Conference on Computer Vision Dublin, Ireland, June 26–July 1, 2000 Proceedings, Part II*. Sous la dir. de David Vernon. Berlin, Heidelberg : Springer Berlin Heidelberg, 2000, p. 37–49 (Cité en page 15.).

[Gerónimo 2007] David Gerónimo, Antonio López, Daniel Ponsa et Angel D. Sappa. « Haar Wavelets and Edge Orientation Histograms for On–Board Pedestrian Detection ». Dans : *Pattern Recognition and Image Analysis : Third Iberian Conference, IbPRIA 2007, Girona, Spain, June 6-8, 2007, Proceedings, Part I*. Sous la dir. de Joan Martí, José Miguel Benedí, Ana Maria Mendonça et Joan Serrat. Berlin, Heidelberg : Springer Berlin Heidelberg, 2007, p. 418–425 (Cité en pages 18 et 19.).

[Gerónimo 2010a] D. Gerónimo, A. M. López, A. D. Sappa et T. Graf. « Survey of Pedestrian Detection for Advanced Driver Assistance Systems ». Dans : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.7 (juil. 2010), p. 1239–1258 (Cité en pages 9, 11, 15, 18 et 21.).

[Gerónimo 2010b] David Gerónimo, Angel D. Sappa, Daniel Ponsa et Antonio M. López. « 2D-3D-based On-board Pedestrian Detection System ». Dans : *Comput. Vis. Image Underst.* 114.5 (mai 2010), p. 583–595 (Cité en page 17.).

[Gheissari 2006] N. Gheissari, T. B. Sebastian et R. Hartley. « Person Reidentification Using Spatiotemporal Appearance ». Dans : *2006 IEEE Computer Society Conference on Computer Vision and Pattern Re-*

cognition (CVPR'06). T. 2. 2006, p. 1528–1535 (Cité en pages 70 et 71.).

[Girshick 2014] R. Girshick, J. Donahue, T. Darrell et J. Malik. « Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation ». Dans : *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Juin 2014, p. 580–587 (Cité en pages 20, 21, 22, 23 et 24.).

[Goldberg 1990] Andrew V. Goldberg et Robert E. Tarjan. « Finding Minimum-Cost Circulations by Successive Approximation ». Dans : *Mathematics of Operations Research* 15.3 (1990), p. 430–466 (Cité en pages 96 et 107.).

[Gou 2017] M. Gou, S. Karanam, W. Liu, O. Camps et R. J. Radke. « DukeMTMC4ReID : A Large-Scale Multi-camera Person Re-identification Dataset ». Dans : *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Juil. 2017, p. 1425–1434 (Cité en pages 79 et 80.).

[Gramm 2009] Jens Gramm, Jiong Guo, Falk Hüffner et Rolf Niedermeier. « Data Reduction and Exact Algorithms for Clique Cover ». Dans : *J. Exp. Algorithmics* 13 (fév. 2009), 2 :2.2–2 :2.15 (Cité en page 95.).

[Gray 2008] Douglas Gray et Hai Tao. « Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features ». Dans : *Computer Vision – ECCV 2008*. Sous la dir. de David Forsyth, Philip Torr et Andrew Zisserman. Berlin, Heidelberg : Springer Berlin Heidelberg, 2008, p. 262–275 (Cité en pages 71, 73, 78 et 79.).

[Han 2006] Ju Han et Bir Bhanu. « Individual recognition using gait energy image ». Dans : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.2 (fév. 2006), p. 316–322 (Cité en page 76.).

[Han 2017] Bing Han et Xiaoyu Wang. « Detection for Power line Inspection ». Dans : *MATEC Web Conf.* 100 (2017), p. 03010 (Cité en page 22.).

[Hirzer 2011] Martin Hirzer, Csaba Beleznai, Peter M. Roth et Horst Bischof. « Person Re-identification by Descriptive and Discriminative Classification ». Dans :

Proceedings of the 17th Scandinavian Conference on Image Analysis. SCIA'11. Ystad, Sweden : Springer-Verlag, 2011, p. 91–102 (Cité en pages 78 et 79.).

[Hoiem 2012] Derek Hoiem, Yodsawalai Chodpathumwan et Qieyun Dai. « Diagnosing Error in Object Detectors ». Dans : *Computer Vision – ECCV 2012 : 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III*. Sous la dir. d'Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato et Cordelia Schmid. Berlin, Heidelberg : Springer Berlin Heidelberg, 2012, p. 340–353 (Cité en pages 27 et 28.).

[Hosang 2015] J. Hosang, M. Omran, R. Benenson et B. Schiele. « Taking a deeper look at pedestrians ». Dans : *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2015, p. 4073–4082 (Cité en page 11.).

[Huang 1997] Timothy Huang et Stuart Russell. « Object Identification in a Bayesian Context ». Dans : *Proceedings of the Fifteenth International Joint Conference on Artifical Intelligence - Volume 2*. IJCAI'97. Nagoya, Japan : Morgan Kaufmann Publishers Inc., 1997, p. 1276–1282 (Cité en page 70.).

[Hussain 2010] Sibt-ul Hussain et Bill Triggs. « Feature Sets and Dimensionality Reduction for Visual Object Detection ». Dans : *Proceedings of the British Machine Vision Conference*. doi :10.5244/C.24.112. BMVA Press, 2010, p. 112.1–112.10 (Cité en page 20.).

[Jayawardena 2010] C. Jayawardena, I. H. Kuo, U. Unger, A. Igic, R. Wong, C. I. Watson, R. Q. Stafford, E. Broadbent, P. Tiwari, J. Warren, J. Sohn et B. A. MacDonald. « Deployment of a service robot to help older people ». Dans : *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Oct. 2010, p. 5990–5995 (Cité en page 1.).

[Jourdheuil 2012] Loïc Jourdheuil, Nicolas Allezard, Thierry Chateau et Thierry Chesnais. « HETEROGENEOUS ADABOOST WITH REAL-TIME CONSTRAINTS - Application to the Detection of Pedestrians by Stereovision ». Dans : *Proceedings*

of the International Conference on Computer Vision Theory and Applications - Volume 1 : VISAPP, (VISIGRAPP 2012). INSTICC. SciTePress, 2012, p. 539–546 (Cité en page 20.).

[Karanam 2015] S. Karanam, Y. Li et R. J. Radke. « Person Re-Identification with Discriminatively Trained Viewpoint Invariant Dictionaries ». Dans : *Int. Conf. on Computer Vision (ICCV'15)*. Déc. 2015, p. 4516–4524 (Cité en pages 71 et 76.).

[Karanam 2018] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps et R. J. Radke. « A Systematic Evaluation and Benchmark for Person Re-Identification : Features, Metrics, and Datasets ». Dans : *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018), p. 1–1 (Cité en pages 69, 79 et 80.).

[Kawanishi 2014] Yasutomo Kawanishi, Yang Wu, Masayuki Mukunoki et Michihiko Minoh. « Shinpuhkan2014 : A Multi-Camera Pedestrian Dataset for Tracking People across Multiple Cameras ». Dans : *the 20th Korea-Japan Joint Workshop on Frontiers of Computer Vision*. Fév. 2014 (Cité en page 79.).

[Klaeser 2008] A. Klaeser, M. Marszalek et C. Schmid. « A Spatio-Temporal Descriptor Based on 3D-Gradients ». Dans : *Proceedings of the British Machine Vision Conference*. doi :10.5244/C.22.99. BMVA Press, 2008, p. 99.1–99.10 (Cité en page 76.).

[Köstinger 2012] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth et H. Bischof. « Large scale metric learning from equivalence constraints ». Dans : *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Juin 2012, p. 2288–2295 (Cité en page 74.).

[Layne 2012] Ryan Layne, Tim Hospedales et Shaogang Gong. « Person Re-identification by Attributes ». Dans : *Proceedings of the British Machine Vision Conference*. BMVA Press, 2012, p. 24.1–24.11 (Cité en page 72.).

[Layne 2015] Ryan Layne, Timothy M. Hospedales et Shaogang Gong. « Investigating Open-World Person Re-identification Using a Drone ». Dans : *Computer Vision - ECCV 2014 Workshops*. Sous la dir. de Lourdes Agapito, Michael M. Bronstein et Carsten

Rother. Cham : Springer International Publishing, 2015, p. 225–240 (Cité en page 79.).

[LealTaixé 2015] L. Leal-Taixé, A. Milan, I. Reid, S. Roth et K. Schindler. « MOTChallenge 2015 : Towards a Benchmark for Multi-Target Tracking ». Dans : *arXiv :1504.01942 [cs]* (avr. 2015). arXiv : 1504.01942 (Cité en pages 10 et 56.).

[Leibe 2008] Bastian Leibe, Aleš Leonardis et Bernt Schiele. « Robust Object Detection with Interleaved Categorization and Segmentation ». Dans : *International Journal of Computer Vision* 77.1 (mai 2008), p. 259–289 (Cité en page 16.).

[Levi 2004] K. Levi et Y. Weiss. « Learning object detection from a small number of examples : the importance of good features ». Dans : *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* T. 2. Juin 2004, (Cité en pages 18 et 21.).

[Li 2013a] W. Li et X. Wang. « Locally Aligned Feature Transforms across Views ». Dans : *2013 IEEE Conference on Computer Vision and Pattern Recognition.* Juin 2013, p. 3594–3601 (Cité en pages 79 et 80.).

[Li 2013b] Wei Li, Rui Zhao et Xiaogang Wang. « Human Reidentification with Transferred Metric Learning ». Dans : *Computer Vision – ACCV 2012.* Sous la dir. de Kyoung Mu Lee, Yasuyuki Matsushita, James M. Rehg et Zhanyi Hu. Berlin, Heidelberg : Springer Berlin Heidelberg, 2013, p. 31–44 (Cité en pages 79 et 80.).

[Li 2013c] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao et J. R. Smith. « Learning Locally-Adaptive Decision Functions for Person Verification ». Dans : *2013 IEEE Conference on Computer Vision and Pattern Recognition.* Juin 2013, p. 3610–3617 (Cité en pages 72 et 76.).

[Li 2014] W. Li, R. Zhao, T. Xiao et X. Wang. « DeepReID : Deep Filter Pairing Neural Network for Person Reidentification ». Dans : *2014 IEEE Conference on Computer Vision and Pattern Recognition.* Juin 2014, p. 152–159 (Cité en pages 70, 75, 79 et 80.).

[Liao 2015] S. Liao, Y. Hu, Xiangyu Zhu et S. Z. Li. « Person re-identification by Local Maximal Occurrence representation and metric learning ». Dans : *Int. Conf. Computer Vision and Pattern Recognition (CVPR'15)*. Juin 2015, p. 2197–2206 (Cité en pages 72, 73 et 74.).

[Lienhart 2002] R. Lienhart et J. Maydt. « An extended set of Haar-like features for rapid object detection ». Dans : *Proceedings. International Conference on Image Processing*. T. 1. 2002, (Cité en pages 18 et 21.).

[Lin 2017] J. Lin, L. Ren, J. Lu, F. Feng et J. Zhou. « Consistent-Aware Deep Learning for Person Re-identification in a Camera Network ». Dans : *Int. Conf. Computer Vision and Pattern Recognition (CVPR'17)*. Juil. 2017, p. 3396–3405 (Cité en pages 70, 78 et 87.).

[Lisanti 2015] G. Lisanti, I. Masi, A. D. Bagdanov et A. D. Bimbo. « Person Re-Identification by Iterative Re-Weighted Sparse Ranking ». Dans : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.8 (août 2015), p. 1629–1642 (Cité en page 73.).

[Liu 2012a] Chunxiao Liu, Shaogang Gong, Chen Change Loy et Xinggang Lin. « Person Re-identification : What Features Are Important ? » Dans : *Computer Vision – ECCV 2012. Workshops and Demonstrations*. Sous la dir. d'Andrea Fusiello, Vittorio Murino et Rita Cucchiara. Berlin, Heidelberg : Springer Berlin Heidelberg, 2012, p. 391–401 (Cité en page 72.).

[Liu 2012b] Xiao Liu, Mingli Song, Qi Zhao, Dacheng Tao, Chun Chen et Jiajun Bu. « Attribute-restricted latent topic model for person re-identification ». Dans : *Pattern Recognition* 45.12 (2012), p. 4204–4213 (Cité en page 72.).

[Liu 2014] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen et J. Bu. « Semi-supervised Coupled Dictionary Learning for Person Re-identification ». Dans : *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Juin 2014, p. 3550–3557 (Cité en page 72.).

[Liu 2016] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu et Alexander C. Berg. « SSD : Single Shot MultiBox Detector ». Dans : *Computer Vision – ECCV 2016*. Sous la dir. de Bastian Leibe, Jiri Matas, Nicu Sebe et Max Welling. Cham : Springer International Publishing, 2016, p. 21–37 (Cité en page 18.).

[Lowe 1999] D. G. Lowe. « Object recognition from local scale-invariant features ». Dans : *Proceedings of the Seventh IEEE International Conference on Computer Vision*. T. 2. Sept. 1999, 1150–1157 vol.2 (Cité en page 72.).

[Lowe 2004] David G. Lowe. « Distinctive Image Features from Scale-Invariant Keypoints ». Dans : *International Journal of Computer Vision* 60.2 (nov. 2004), p. 91–110 (Cité en page 72.).

[Loy 2010] Chen Change Loy, Tao Xiang et Shaogang Gong. « Time-Delayed Correlation Analysis for Multi-Camera Activity Understanding ». Dans : *International Journal of Computer Vision* 90.1 (oct. 2010), p. 106–129 (Cité en pages 78 et 79.).

[Loy 2013] C. C. Loy, C. Liu et S. Gong. « Person re-identification by manifold ranking ». Dans : *2013 IEEE International Conference on Image Processing*. Sept. 2013, p. 3567–3571 (Cité en pages 78 et 79.).

[Ma 2012] Bingpeng Ma, Yu Su et Frédéric Jurie. « Local Descriptors Encoded by Fisher Vectors for Person Re-identification ». Dans : *Computer Vision – ECCV 2012. Workshops and Demonstrations*. Sous la dir. d'Andrea Fusiello, Vittorio Murino et Rita Cucchiara. Berlin, Heidelberg : Springer Berlin Heidelberg, 2012, p. 413–422 (Cité en pages 72 et 73.).

[Ma 2014] Bingpeng Ma, Yu Su et Frédéric Jurie. « Covariance descriptor based on bio-inspired features for person re-identification and face verification ». Dans : *Image and Vision Computing* 32.6 (2014), p. 379–390 (Cité en pages 72 et 73.).

[Manning 2008] Christopher D. Manning, Prabhakar Raghavan et Hinrich Schütze. *Introduction to Information Retrieval*. New York, NY, USA : Cambridge University Press, 2008 (Cité en pages 27 et 28.).

[Martin 1997] Alvin F. Martin, George R. Doddington, Terri Kamm, Mark Ordowski et Mark A. Przybocki. « The DET curve in assessment of detection task performance. » Dans : *EUROSPEECH*. Sous la dir. de George Kokkinakis, Nikos Fakotakis et Evangelos Dermatas. ISCA, 1997 (Cité en pages 27 et 28.).

[Martinel 2012] N. Martinel et C. Micheloni. « Re-identify people in wide area camera network ». Dans : *Int. Conf. Computer Vision and Pattern Recognition (CVPR'12) - Workshops*. Juin 2012, p. 31–36 (Cité en pages 73, 79, 80, 82, 84, 89 et 98.).

[Matsukawa 2016] T. Matsukawa, T. Okabe, E. Suzuki et Y. Sato. « Hierarchical Gaussian Descriptor for Person Re-identification ». Dans : *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2016, p. 1363–1372 (Cité en pages 72 et 73.).

[McLaughlin 2016] N. McLaughlin, J. M. d. Rincon et P. Miller. « Recurrent Convolutional Network for Video-Based Person Re-identification ». Dans : *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2016, p. 1325–1334 (Cité en page 76.).

[Mekonnen 2014] A. A. Mekonnen, F. Lerasle, A. Herbulot et C. Briand. « People Detection with Heterogeneous Features and Explicit Optimization on Computation Time ». Dans : *2014 22nd International Conference on Pattern Recognition*. Août 2014, p. 4322–4327 (Cité en page 110.).

[Mekonnen 2018] A. A. Mekonnen et F. Lerasle. « Comparative Evaluations of Selected Tracking-by-Detection Approaches ». Dans : *IEEE Transactions on Circuits and Systems for Video Technology* (2018), p. 1–1 (Cité en page 111.).

[Micheloni 2003] C. Micheloni, G. L. Foresti et L. Snidaro. « A cooperative multicamera system for video-surveillance of parking lots ». Dans : *IEE Symposium on Intelligence Distributed Surveillance Systems (Ref. No. 2003/10062)*. Fév. 2003, p. 5/1–5/5 (Cité en page 1.).

[Mignon 2012] A. Mignon et F. Jurie. « PCCA : A new approach for distance learning from sparse pairwise constraints ». Dans : *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Juin 2012, p. 2666–2672 (Cité en pages 72 et 74.).

[Mikolajczyk 2004] Krystian Mikolajczyk, Cordelia Schmid et Andrew Zisserman. « Human Detection Based on a Probabilistic Assembly of Robust Part Detectors ». Dans : *Computer Vision - ECCV 2004 : 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I*. Sous la dir. de Tomás Pajdla et Jiří Matas. Berlin, Heidelberg : Springer Berlin Heidelberg, 2004, p. 69–82 (Cité en page 16.).

[Møgelmose 2012] A. Møgelmose, A. Prioletti, M. M. Trivedi, A. Broggi et T. B. Moeslund. « Two-stage part-based pedestrian detection ». Dans : *2012 15th International IEEE Conference on Intelligent Transportation Systems*. Sept. 2012, p. 73–77 (Cité en page 20.).

[Mohan 2001] A. Mohan, C. Papageorgiou et T. Poggio. « Example-based object detection in images by components ». Dans : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.4 (avr. 2001), p. 349–361 (Cité en page 16.).

[Mu 2008] Yadong Mu, Shuicheng Yan, Yi Liu, T. Huang et Bingfeng Zhou. « Discriminative local binary patterns for human detection in personal album ». Dans : *2008 IEEE Conference on Computer Vision and Pattern Recognition*. Juin 2008, p. 1–8 (Cité en page 19.).

[Nam 2014] Woonhyun Nam, Piotr Dollár et Joon Hee Han. « Local Decorrelation for Improved Pedestrian Detection ». Dans : *Proceedings of the 27th International Conference on Neural Information Processing Systems*. NIPS'14. Montreal, Canada : MIT Press, 2014, p. 424–432 (Cité en pages 20, 21, 22, 23, 24, 29, 30, 31 et 34.).

[Nambiar 2014] Athira Nambiar, Matteo Taiana, Dario Figueira, Jacinto C. Nascimento et Alexandre Bernardino. « A multi-camera video dataset for research on high-definition surveillance ». Dans : *Int. Journal*

of Machine Intelligence and Sensory Signal Processing 3 (déc. 2014), p. 267–286 (Cité en pages 79, 80, 82, 89, 97 et 109.).

[Nguyen 2016] Duc Thanh Nguyen, Wanqing Li et Philip O. Ogunbona. « Human detection from images and videos : A survey ». Dans : *Pattern Recognition* 51.Supplement C (2016), p. 148–175 (Cité en page 10.).

[Ojala 1996] Timo Ojala, Matti Pietikäinen et David Harwood. « A comparative study of texture measures with classification based on featured distributions ». Dans : *Pattern Recognition* 29.1 (1996), p. 51–59 (Cité en pages 19 et 21.).

[Paisitkriangkrai 2008] S. Paisitkriangkrai, C. Shen et J. Zhang. « Fast Pedestrian Detection Using a Cascade of Boosted Covariance Features ». Dans : *IEEE Transactions on Circuits and Systems for Video Technology* 18.8 (août 2008), p. 1140–1151 (Cité en page 21.).

[Paleček 2012] Karel Paleček, David Gerónimo et Frédéric Lerasle. « Pre-attention Cues for Person Detection ». Dans : *Cognitive Behavioural Systems : COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers*. Sous la dir. d'Anna Esposito, Antonietta M. Esposito, Alessandro Vinciarelli, Rüdiger Hoffmann et Vincent C. Müller. Berlin, Heidelberg : Springer Berlin Heidelberg, 2012, p. 225–235 (Cité en page 17.).

[Pallikonda Rajasekaran 2010] M. Pallikonda Rajasekaran, S. Radhakrishnan et P. Subbaraj. « Sensor grid applications in patient monitoring ». Dans : *Future Generation Computer Systems* 26.4 (2010), p. 569–575 (Cité en page 1.).

[Pan 2013] Hong Pan, Yaping Zhu et Liangzheng Xia. « Efficient and accurate face detection using heterogeneous feature descriptors and feature selection ». Dans : *Computer Vision and Image Understanding* 117.1 (2013), p. 12–28 (Cité en page 20.).

[Papageorgiou 2000] Constantine Papageorgiou et Tomaso Poggio. « A Trainable System for Object Detection ». Dans : *International Journal of Computer Vision* 38.1 (juin 2000), p. 15–33 (Cité en pages 11, 18 et 21.).

[Park 2010] Dennis Park, Deva Ramanan et Charless Fowlkes. « Multiresolution Models for Object Detection ». Dans : *Computer Vision – ECCV 2010 : 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*. Sous la dir. de Kostas Daniilidis, Petros Maragos et Nikos Paragios. Berlin, Heidelberg : Springer Berlin Heidelberg, 2010, p. 241–254 (Cité en page 16.).

[Pedagadi 2013] S. Pedagadi, J. Orwell, S. Velastin et B. Boghossian. « Local Fisher Discriminant Analysis for Pedestrian Re-identification ». Dans : *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'13)*. Juin 2013, p. 3318–3325 (Cité en pages 72 et 74.).

[Piccardi 2004] M. Piccardi. « Background subtraction techniques : a review ». Dans : *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*. T. 4. Oct. 2004, 3099–3104 vol.4 (Cité en page 14.).

[Pozzobon 1999] Alberto Pozzobon, Giuseppe Sciutto et Valerio Recagno. « Security in Ports : the User Requirements for Surveillance System ». Dans : *Advanced Video-Based Surveillance Systems*. Sous la dir. de Carlo S. Regazzoni, Gianni Fabri et Gianni Vernazza. Boston, MA : Springer US, 1999, p. 18–26 (Cité en page 1.).

[Räty 2010] T. D. Räty. « Survey on Contemporary Remote Surveillance Systems for Public Safety ». Dans : *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40.5 (sept. 2010), p. 493–515 (Cité en page 1.).

[Ren 2015] Shaoqing Ren, Kaiming He, Ross Girshick et Jian Sun. « Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks ». Dans : *Advances in Neural Information Processing Systems 28*. Sous la dir. de C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama et R. Garnett. Curran Associates, Inc., 2015, p. 91–99 (Cité en pages 24 et 30.).

[Ren 2017] S. Ren, K. He, R. Girshick et J. Sun. « Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks ». Dans : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (juin 2017), p. 1137–1149 (Cité en page 11.).

[Robertson 1995] N. Robertson et P.D. Seymour. « Graph Minors .XIII. The Disjoint Paths Problem ». Dans : *Journal of Combinatorial Theory, Series B* 63.1 (1995), p. 65–110 (Cité en pages 96 et 107.).

[Ronetti 2000] Nino Ronetti et Carlo Dambra. « Railway Station Surveillance : The Italian Case ». Dans : *Multimedia Video-Based Surveillance Systems : Requirements, Issues and Solutions*. Sous la dir. de Gian Luca Foresti, Petri Mähönen et Carlo S. Regazzoni. Boston, MA : Springer US, 2000, p. 13–20 (Cité en page 1.).

[Russakovsky 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg et Li Fei-Fei. « ImageNet Large Scale Visual Recognition Challenge ». Dans : *International Journal of Computer Vision (IJCV)* (avr. 2015), p. 1–42 (Cité en page 10.).

[Satpathy 2013] A. Satpathy, X. Jiang et H. L. Eng. « Human detection using Discriminative and Robust Local Binary Pattern ». Dans : *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Mai 2013, p. 2376–2380 (Cité en page 19.).

[Schapire 2003] Robert E. Schapire. « The Boosting Approach to Machine Learning : An Overview ». Dans : *Nonlinear Estimation and Classification*. Sous la dir. de David D. Denison, Mark H. Hansen, Christopher C. Holmes, Bani Mallick et Bin Yu. New York, NY : Springer New York, 2003, p. 149–171 (Cité en pages 18 et 20.).

[Schwartz 2009a] W. R. Schwartz et L. S. Davis. « Learning Discriminative Appearance-Based Models Using Partial Least Squares ». Dans : *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*. Oct. 2009, p. 322–329 (Cité en page 79.).

[Schwartz 2009b] W. R. Schwartz, A. Kembhavi, D. Harwood et L. S. Davis. « Human detection using partial least squares analysis ». Dans : *2009 IEEE 12th International Conference on Computer Vision*. Sept. 2009, p. 24–31 (Cité en page 20.).

[Shi 2015] Z. Shi, T. M. Hospedales et T. Xiang. « Transferring a semantic representation for person re-identification and search ». Dans : *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2015, p. 4184–4193 (Cité en page 73.).

[Sochman 2005] J. Sochman et J. Matas. « Waldboost - learning for time constrained sequential detection ». Dans : *Int Conf. on Computer Vision and Pattern Recognition (CVPR'05)*. Boston, USA, juin 2005 (Cité en page 38.).

[Spackman 1989] Kent A. Spackman. « Signal Detection Theory : Valuable Tools for Evaluating Inductive Learning ». Dans : *Proceedings of the Sixth International Workshop on Machine Learning*. Ithaca, New York, USA : Morgan Kaufmann Publishers Inc., 1989, p. 160–163 (Cité en page 27.).

[Stauffer 1999] C. Stauffer et W. E. L. Grimson. « Adaptive background mixture models for real-time tracking ». Dans : *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. T. 2. 1999, 252 Vol. 2 (Cité en page 14.).

[Su 2015] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis et W. Gao. « Multi-Task Learning with Low Rank Attribute Embedding for Person Re-Identification ». Dans : *2015 IEEE International Conference on Computer Vision (ICCV)*. Déc. 2015, p. 3739–3747 (Cité en page 72.).

[Szarvas 2005] M. Szarvas, A. Yoshizawa, M. Yamamoto et J. Ogata. « Pedestrian detection with convolutional neural networks ». Dans : *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.* Juin 2005, p. 224–229 (Cité en page 21.).

[Tang 2012] Danhang Tang, Yang Liu et Tae-kyun Kim. « Fast Pedestrian Detection by Cascaded Random Forest with Dominant Orientation Templates ». Dans :

Proceedings of the British Machine Vision Conference. BMVA Press, 2012, p. 58.1–58.11 (Cité en page 21.).

[Teutsch 2015] M. Teutsch et W. Krüger. « Robust and fast detection of moving vehicles in aerial videos using sliding windows ». Dans : *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Juin 2015, p. 26–34 (Cité en page 22.).

[Tian 2015] Y. Tian, P. Luo, X. Wang et X. Tang. « Pedestrian detection aided by deep learning semantic tasks ». Dans : *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2015, p. 5079–5087 (Cité en page 11.).

[Tomè 2016] D. Tomè, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi et S. Tubaro. « Deep Convolutional Neural Networks for pedestrian detection ». Dans : *Signal Processing : Image Communication* 47 (2016), p. 482–489 (Cité en pages 20, 21, 22, 23 et 24.).

[Tuzel 2008] O. Tuzel, F. Porikli et P. Meer. « Pedestrian Detection via Classification on Riemannian Manifolds ». Dans : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.10 (oct. 2008), p. 1713–1727 (Cité en page 21.).

[Varga 2016] Robert Varga et Sergiu Nedevschi. « Robust Pallet Detection for Automated Logistics Operations ». Dans : *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4 : VISAPP, (VISIGRAPP 2016)*. INSTICC. SciTePress, 2016, p. 470–477 (Cité en page 22.).

[Viola 2004] Paul Viola et Michael J. Jones. « Robust Real-Time Face Detection ». Dans : *International Journal of Computer Vision* 57.2 (mai 2004), p. 137–154 (Cité en pages 10, 11, 16, 18, 21 et 22.).

[Viola 2005] Paul Viola, Michael J. Jones et Daniel Snow. « Detecting Pedestrians Using Patterns of Motion and Appearance ». Dans : *International Journal of Computer Vision* 63.2 (juil. 2005), p. 153–161 (Cité en pages 18, 19 et 21.).

[Wald 1945] A. Wald. « Sequential Tests of Statistical Hypotheses ». Dans : *Ann. Math. Statist.* 16.2 (juin 1945), p. 117–186 (Cité en page 40.).

[Walk 2010] S. Walk, N. Majer, K. Schindler et B. Schiele. « New features and insights for pedestrian detection ». Dans : *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Juin 2010, p. 1030–1037 (Cité en pages 19, 20 et 21.).

[Wang 2009] X. Wang, T. X. Han et S. Yan. « An HOG-LBP human detector with partial occlusion handling ». Dans : *2009 IEEE 12th International Conference on Computer Vision*. Sept. 2009, p. 32–39 (Cité en pages 11 et 19.).

[Wang 2011] S. Wang, M. Lewandowski, J. Annesley et J. Orwell. « Re-identification of pedestrians with variable occlusion and scale ». Dans : *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. Nov. 2011, p. 1876–1882 (Cité en pages 78 et 79.).

[Wang 2014] Taiqing Wang, Shaogang Gong, Xiatian Zhu et Shengjin Wang. « Person Re-identification by Video Ranking ». Dans : *Computer Vision – ECCV 2014*. Sous la dir. de David Fleet, Tomas Pajdla, Bernt Schiele et Tinne Tuytelaars. Cham : Springer International Publishing, 2014, p. 688–703 (Cité en pages 76, 79 et 80.).

[Wang 2016] T. Wang, S. Gong, X. Zhu et S. Wang. « Person Re-Identification by Discriminative Selection in Video Ranking ». Dans : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.12 (déc. 2016), p. 2501–2514 (Cité en pages 79 et 80.).

[Wei 2018] Longhui Wei, Shiliang Zhang, Wen Gao et Qi Tian. « Person Transfer GAN to Bridge Domain Gap for Person Re-Identification ». Dans : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2018 (Cité en page 79.).

[Weinberger 2006] Kilian Q Weinberger, John Blitzer et Lawrence K. Saul. « Distance Metric Learning for Large Margin Nearest Neighbor Classification ». Dans : *Advances in Neural Information Processing Systems 18*. Sous la dir. d'Y. Weiss, B. Schölkopf et J. C. Platt. MIT Press, 2006, p. 1473–1480 (Cité en pages 73 et 74.).

[Weinberger 2009] Kilian Q. Weinberger et Lawrence K. Saul. « Distance Metric Learning for Large Margin Nearest Neighbor Classification ». Dans : *J. Mach. Learn. Res.* 10 (juin 2009), p. 207–244 (Cité en page 74.).

[Wojek 2008] Christian Wojek et Bernt Schiele. « A Performance Evaluation of Single and Multi-feature People Detection ». Dans : *Pattern Recognition : 30th DAGM Symposium Munich, Germany, June 10-13, 2008 Proceedings*. Sous la dir. de Gerhard Rigoll. Berlin, Heidelberg : Springer Berlin Heidelberg, 2008, p. 82–91 (Cité en pages 11, 19 et 20.).

[Wu 2008] Bo Wu et R. Nevatia. « Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection ». Dans : *2008 IEEE Conference on Computer Vision and Pattern Recognition*. Juin 2008, p. 1–8 (Cité en page 20.).

[Wu 2016] Lin Wu, Chunhua Shen et Anton van den Hengel. « Deep Recurrent Convolutional Networks for Video-based Person Re-identification : An End-to-End Approach ». Dans : *CoRR* abs/1606.01609 (2016). arXiv : 1606.01609 (Cité en page 76.).

[Wu 2017] Lin Wu et Anton van den Hengel. « Deep Linear Discriminant Analysis on Fisher Networks ». Dans : *Pattern Recogn.* 65.C (mai 2017), p. 238–250 (Cité en page 75.).

[Xiao 2003] R. Xiao, L. Zhu et H. Zhang. « Boosting chain learning for object detection ». Dans : *Int. Conf. on Computer Vision (ICCV'03)*. Nice, France, oct. 2003 (Cité en pages 38 et 40.).

[Xiao 2016] T. Xiao, H. Li, W. Ouyang et X. Wang. « Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification ». Dans : *Int. Conf. Computer Vision and Pattern Recognition (CVPR'16)*. Juin 2016, p. 1249–1258 (Cité en page 75.).

[Xiong 2014] Fei Xiong, Mengran Gou, Octavia Camps et Mario Sznaier. « Person Re-Identification Using Kernel-Based Metric Learning Methods ». Dans : *Computer Vision – ECCV 2014*. Sous la dir. de David Fleet, Tomas Pajdla, Bernt Schiele et Tinne Tuy-

telaars. Cham : Springer International Publishing, 2014, p. 1–16 (Cité en pages 72, 73 et 74.).

[Xu 2014] Yuanlu Xu, Bingpeng Ma, Rui Huang et Liang Lin. « Person Search in a Scene by Jointly Modeling People Commonness and Person Uniqueness ». Dans : *Proceedings of the 22Nd ACM International Conference on Multimedia*. MM '14. Orlando, Florida, USA : ACM, 2014, p. 937–940 (Cité en page 70.).

[Yan 2007] S. Yan, D. Xu, B. Zhang, H. j. Zhang, Q. Yang et S. Lin. « Graph Embedding and Extensions : A General Framework for Dimensionality Reduction ». Dans : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.1 (jan. 2007), p. 40–51 (Cité en page 74.).

[Yan 2016] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan et Xiaokang Yang. « Person Re-identification via Recurrent Feature Aggregation ». Dans : *Computer Vision – ECCV 2016*. Sous la dir. de Bastian Leibe, Jiri Matas, Nicu Sebe et Max Welling. Cham : Springer International Publishing, 2016, p. 701–716 (Cité en page 76.).

[Yang 2014] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi et Stan Z. Li. « Salient Color Names for Person Re-identification ». Dans : *Computer Vision – ECCV 2014*. Sous la dir. de David Fleet, Tomas Pajdla, Bernt Schiele et Tinne Tuytelaars. Cham : Springer International Publishing, 2014, p. 536–551 (Cité en pages 72 et 73.).

[Yang 2015] B. Yang, J. Yan, Z. Lei et S. Z. Li. « Convolutional Channel Features ». Dans : *2015 IEEE International Conference on Computer Vision (ICCV)*. Déc. 2015, p. 82–90 (Cité en page 34.).

[Yi 2014] D. Yi, Z. Lei, S. Liao et S. Z. Li. « Deep Metric Learning for Person Re-identification ». Dans : *2014 22nd International Conference on Pattern Recognition*. Août 2014, p. 34–39 (Cité en pages 70 et 75.).

[You 2016] J. You, A. Wu, X. Li et W. S. Zheng. « Top-Push Video-Based Person Re-identification ». Dans : *2016 IEEE Conference on Computer Vision and*

Pattern Recognition (CVPR). Juin 2016, p. 1345–1353 (Cité en page 76.).

[Zajdel 2005] W. Zajdel, Z. Zivkovic et B. J. A. Krose. « Keeping Track of Humans : Have I Seen This Person Before ? » Dans : *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. Avr. 2005, p. 2081–2086 (Cité en page 70.).

[Zambanini 2009] S. Zambanini, P. Blauensteiner et M. Kampel. « Automated multi-camera surveillance for the prevention and investigation of bank robberies in Austria : A case study ». Dans : *3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009)*. Déc. 2009, p. 1–6 (Cité en page 1.).

[Zhang 2007] Cha Zhang et Paul Viola. « Multiple-instance Pruning for Learning Efficient Cascade Detectors ». Dans : *Proceedings of the 20th International Conference on Neural Information Processing Systems*. NIPS'07. Vancouver, British Columbia, Canada : Curran Associates Inc., 2007, p. 1681–1688 (Cité en pages 11, 22, 38, 40 et 56.).

[Zhang 2009] M. Zhang et R. Alhajj. « Content-Based Image Retrieval : From the Object Detection/Recognition Point of View ». Dans : *Artificial Intelligence for Maximizing Content Based Image Retrieval*. Sous la dir. de Z. Ma. PA : Information Science Reference. Hershey, 2009, p. 115–144 (Cité en page 9.).

[Zhang 2013] Xin Zhang, Yee-Hong Yang, Zhiguang Han, Hui Wang et Chao Gao. « Object Class Detection : A Survey ». Dans : *ACM Comput. Surv.* 46.1 (2013), 10 :1–10 :53 (Cité en page 9.).

[Zhang 2015] S. Zhang, R. Benenson et B. Schiele. « Filtered channel features for pedestrian detection ». Dans : *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2015, p. 1751–1760 (Cité en pages 11 et 22.).

[Zhang 2016a] Liliang Zhang, Liang Lin, Xiaodan Liang et Kaiming He. « Is Faster R-CNN Doing Well for Pedestrian Detection ? » Dans : *Computer Vision – ECCV 2016 : 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*. Sous la dir. de Bastian Leibe, Jiri

Matas, Nicu Sebe et Max Welling. Cham : Springer International Publishing, 2016, p. 443–457 (Cité en pages 30, 31, 34, 61, 64 et 65.).

[Zhang 2016b] S. Zhang, R. Benenson, M. Omran, J. Hosang et B. Schiele. « How Far are We from Solving Pedestrian Detection ? » Dans : *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2016, p. 1259–1267 (Cité en pages 11 et 22.).

[Zhao 1999] Liang Zhao et C. Thorpe. « Stereo- and neural network-based pedestrian detection ». Dans : *Proceedings 199 IEEE/IEEJ/JSAI International Conference on Intelligent Transportation Systems (Cat. No.99TH8383)*. 1999, p. 298–303 (Cité en page 21.).

[Zhao 2000] L. Zhao et C. E. Thorpe. « Stereo- and neural network-based pedestrian detection ». Dans : *IEEE Transactions on Intelligent Transportation Systems* 1.3 (sept. 2000), p. 148–154 (Cité en page 21.).

[Zhao 2013a] R. Zhao, W. Ouyang et X. Wang. « Person Re-identification by Salience Matching ». Dans : *Int. Conf. on Computer Vision (ICCV'13)*. Déc. 2013, p. 2528–2535 (Cité en pages 72 et 73.).

[Zhao 2013b] R. Zhao, W. Ouyang et X. Wang. « Unsupervised Salience Learning for Person Re-identification ». Dans : *Int. Conf. Computer Vision and Pattern Recognition (CVPR'13)*. Juin 2013, p. 3586–3593 (Cité en pages 72 et 73.).

[Zhao 2014] R. Zhao, W. Ouyang et X. Wang. « Learning Mid-level Filters for Person Re-identification ». Dans : *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Juin 2014, p. 144–151 (Cité en pages 72 et 73.).

[Zheng 2009] Wei-Shi Zheng, Shaogang Gong et Tao Xiang. « Associating Groups of People ». Dans : *Proceedings of the British Machine Vision Conference*. BMVA Press, 2009, p. 23.1–23.11 (Cité en page 79.).

[Zheng 2012] W. S. Zheng, S. Gong et T. Xiang. « Transfer re-identification : From person to set-based verification ». Dans : *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Juin 2012, p. 2650–2657 (Cité en page 76.).

[Zheng 2015] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang et Q. Tian. « Scalable Person Re-identification : A Benchmark ». Dans : *Int. Conf. on Computer Vision (ICCV'15)*. Déc. 2015, p. 1116–1124 (Cité en pages 79 et 80.).

[Zheng 2016a] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang et Qi Tian. « MARS : A Video Benchmark for Large-Scale Person Re-Identification ». Dans : *Computer Vision – ECCV 2016*. Sous la dir. de Bastian Leibe, Jiri Matas, Nicu Sebe et Max Welling. Cham : Springer International Publishing, 2016, p. 868–884 (Cité en pages 76 et 79.).

[Zheng 2016b] Liang Zheng, Yi Yang et Alexander G. Hauptmann. « Person Re-identification : Past, Present and Future ». Dans : *CoRR* abs/1610.02984 (2016). arXiv : 1610.02984 (Cité en pages 69, 71, 73 et 74.).

[Zhu 2006] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng et S. Avidan. « Fast Human Detection Using a Cascade of Histograms of Oriented Gradients ». Dans : *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. T. 2. 2006, p. 1491–1498 (Cité en page 19.).

[Zhu 2016] Xiaoke Zhu, Xiao-Yuan Jing, Fei Wu et Hui Feng. « Video-based Person Re-identification by Simultaneously Learning Intra-video and Inter-video Distance Metrics ». Dans : *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. IJCAI'16. New York, New York, USA : AAAI Press, 2016, p. 3552–3558 (Cité en page 76.).

[Zouba 2009] N. Zouba, F. Bremond et M. Thonnat. « Multisensor Fusion for Monitoring Elderly Activities at Home ». Dans : *2009 Sixth IEEE International Conference on Advanced Video and Signal Based*

Surveillance. Sept. 2009, p. 98–103 (Cité en page 1.).