Ré-identification de personnes : Application aux réseaux de caméras à champs disjoints

Boris Meden

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Présentée et soutenue par :
Boris Meden
le mardi 15 janvier 2013

Titre :

Ré-identification de personnes :
Application aux réseaux de caméras à champs disjoints

École doctorale et discipline ou spécialité :
EDSYS : Informatique 4200018

Unité de recherche :
LAAS-CNRS / CEA-LIST

Directeur(s) de Thèse :

Frédéric Lerasle
Patrick Sayd

Jury :

Serge Miguet	Rapporteur/Président
Patrick Pérez	Rapporteur
François Brémond	Examinateur
Lionnel Joussemet	Examinateur

Remerciements

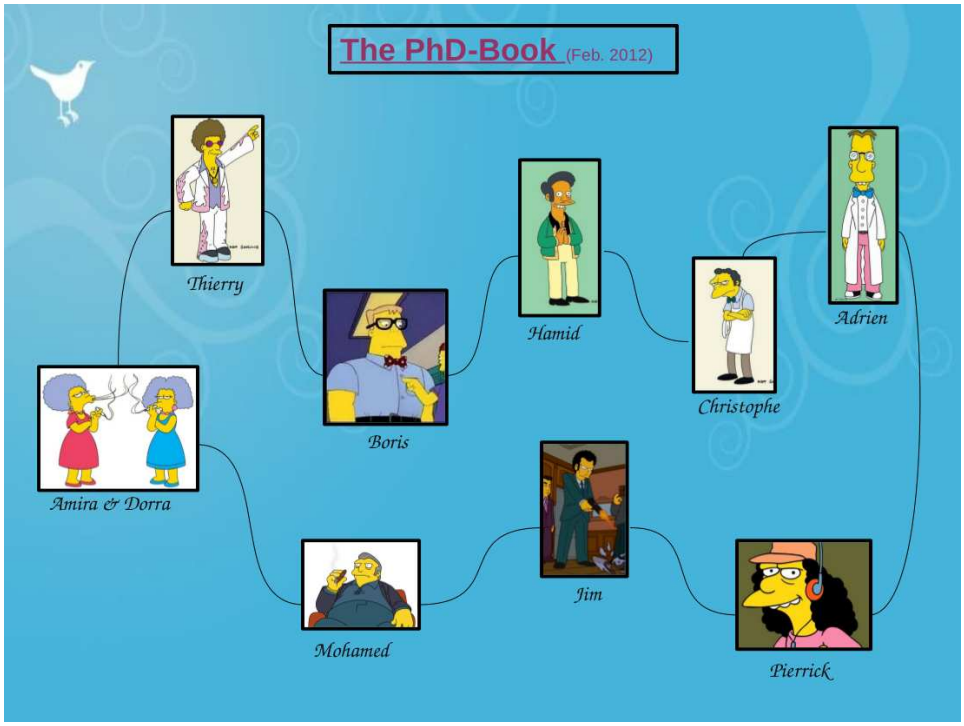
Ce manuscrit de thèse présente des travaux de recherche sur le suivi de personnes dans des réseaux de caméras à champs de vue disjoints. Ces travaux ont été réalisés au sein du Laboratoire Vision et Ingénierie des Contenus (LVIC) du CEA, LIST à Gif-sur-Yvette, et ont reçu un encadrement académique du LAAS-CNRS à Toulouse. Je remercie MM. François GASPARD et Michel DEVY, respectivement chef du LVIC et chef de groupe au LAAS, pour m'avoir donner l'opportunité de réaliser cette thèse.

Je remercie ensuite mes rapporteurs, MM. Patrick PÉREZ et Serge MIGUET, respectivement Directeur de Recherches à l'INRIA Rennes et Chercheur Senior à Technicolor, et Professeur et Directeur de Recherches au LIRIS (Laboratoire d'InfoRmatique en Image et Systèmes d'information à Lyon) pour avoir accepté de rapporter cette thèse. Je remercie M. François Brémond, Professeur à l'INRIA Sophia Antipolis, ainsi que M. Lionnel Joussemet, CEO de Diotasoft S.A.S. (mon nouveau patron :)) pour avoir accepté de prendre part à mon jury de thèse.

Passons aux choses sérieuses, il s'en est passé du temps depuis ce premier point de thèse tous ensemble en novembre 2009 dans le PMU à côté d'Antony RER. Je te remercie Patrick (Sayd) pour ton encadrement et notamment les sorties point de thèse/jogging, et surtout le Patron, Fred (Lerasle), pour avoir dirigé, bien qu'à distance, cette thèse, de main de maître. Fred, je te remercie pour tout, la motivation

que tu m'as transmise tout au long de ces trois années, les très nombreuses relectures, toujours impeccables, en flux tendu pour la fin de rédaction de thèse, toi le jour, moi la nuit, et les coups de fil remonteurs de moral les veilles de conf., et pas grave si c'est un dimanche aprem. Merci vous deux pour tout !

Je remercie ensuite ma promo de thèse : MoMo, Disco Stu et George, et j'attends impatiemment vos soutenances (et pots ;-)). Et parce qu'une thèse, ça se réalise avant tout dans un labo, je remercie mes collègues et amis du quotidien : la team des D'jeuns, Adrien (call sign Papa), Hamid (Apu) et Pierrick (Otto Bus) (on ne relâche pas la pression, c'est pas encore tout à fait terminé pour vous !), les thésards de 3D, Amira, Dorra et Jim, le grand-frère P'tit Lu (Grand Lu maintenant !), les potes : Romain, Vincent PhD et Steve, une mention particulière pour LVIC Racing Club (où l'on aime à pratiquer un foutcheubol virileumaiskorrect) ainsi que l'ensemble des équipes de vidéosurveillance et de 3D. Merci à tous !



Résumé

Cette thèse s'inscrit dans le contexte de la vidéosurveillance "intelligente", et s'intéresse à la supervision de réseaux de caméras à champs disjoints, contrainte classique lorsque l'on souhaite limiter l'instrumentation du bâtiment. Il s'agit là de l'un des cas d'application du problème de la ré-identification de personnes. À ce titre, la thèse propose une approche se démarquant de l'état de l'art qui traite classiquement le problème sous l'aspect description, via la mise en correspondance de signatures image à image. Nous l'abordons ici sous l'aspect filtrage : comment intégrer la ré-identification de personne dans un processus de suivi multi-pistes, de manière à maintenir des identités de pistes cohérentes, malgré des discontinuités dans l'observation. Nous considérons ainsi une approche suivi et mises en correspondance, au niveau caméra et utilisons ce module pour ensuite raisonner au niveau du réseau.

Nous décrivons dans un premier temps les approches classiques de ré-identification, abordées sous l'aspect description. Nous proposons ensuite un formalisme de filtrage particulaire à états continus et discret pour estimer conjointement position et identité de la cible au cours du temps, dans chacune des caméras. Un second étage de traitement permet d'intégrer la topologie du réseau et les temps d'apparition pour optimiser la ré-identification au sein du réseau.

Nous démontrons la faisabilité de l'approche en grande partie sur des données issues de réseaux de caméras déployés au sein du laboratoire, étant donné le manque de données publiques concernant ce domaine. Nous prévoyons de mettre en accès public ces banques de données.

Mots clés : Ré-identification, suivi visuel, descripteurs image, réseau de caméras, champs de vue disjoints, filtrage particulaire, vidéosurveillance.

Abstract

This thesis deals with intelligent videosurveillance, and focus on the supervision of camera networks with nonoverlapping fields of view, a classical constraint when it comes to limitate the building instrumentation. It is one of the use-case of the pedestrian re-identification problem. On that point, the thesis distinguishes itself from state of the art methods, which treat the problem from the descriptor perspective through image to image signatures comparison. Here we consider it from a bayesian filtering perspective : how to plug re-identification in a complete multi-target tracking process, in order to maintain targets identities, in spite of observation discontinuities. Thus we consider tracking and signature comparison, at the camera level, and use that module to take decisions at the network level.

We describe first the classical re-identification approaches, based on the description. Then, we propose a mixed-state particle filter framework to estimate jointly the targets positions and their identities in the cameras. A second stage of processing integrates the network topology and optimise the re-identifications in the network.

Considering the lack of public data in nonoverlapping camera network, we mainly demonstrate our approach on camera networks deployed at the lab. A publication of these data is in progress.

Key words : Re-identification, visual tracking, image descriptors, camera networks, non-overlapping fields of view, particle filtering, videosurveillance.

Table des matières

Remerciements iii

Résumé v

Abstract vii

Notations 1

1 Introduction et contexte des travaux 3
- 1.1 Contexte général et enjeux . . . 4
- 1.2 Que dit la législation ? . . . 5
  - 1.2.1 Droit à l'image (loi du 21 janvier 1995) . . . 7
- 1.3 Positionnement de nos travaux . . . 9
  - 1.3.1 Vue d'ensemble . . . 9
  - 1.3.2 Suivi multi-cibles . . . 11
  - 1.3.3 Ré-identification . . . 11
- 1.4 Contributions et organisation du manuscrit . . . 12

I Traitements au niveau des caméras 17

2 Modèles de ré-identification 19
- 2.1 Introduction . . . 20
- 2.2 Ré-identification par caractéristiques biométriques . . . 20
- 2.3 Au delà de la biométrie : positionnement du problème . . . 22
  - 2.3.1 Un problème de rang : focalisation sur la ré-identification entre deux caméras . . . 22
  - 2.3.2 Quelques premiers constats sur la ré-identification . . . 26

2.4 Ré-identification entre paires de caméras : méthodes supervisées . 27
2.4.1 Fonction de transfert colorimétrique entre caméras 27
2.4.2 Apprentissage statistique pour la ré-identification 30
2.4.2.1 Travaux inspirés de [Gray et Tao, 2008] 33
2.4.2.2 Étude comparative 33
2.4.3 Limitations des méthodes supervisées 34
2.5 Méthodes non-supervisées pour la ré-identification dans un réseau 34
2.5.1 Principe de l'accumulation de caractéristiques locales dirigée par les symétries 35
2.5.1.1 Axes de symétrie/asymétrie 36
2.5.1.2 Extraction des descripteurs 37
2.5.1.3 Comparaison de signatures 38
2.5.2 Extensions directes de l'approche 39
2.6 Choix de notre représentation 40
2.6.1 Influence des composantes de SDALF 42
2.6.2 Perspectives pour notre contexte de surveillance de réseaux de caméras 43
2.7 Conclusion 43

3 Estimation bayésienne de suivi et ré-identification dans une caméra 45
3.1 Introduction 46
3.2 État de l'art 47
3.3 Filtrage bayésien récursif 50
3.3.1 Formalisation du problème 50
3.3.2 Approximation particulaire 51
3.4 Extension au filtrage particulaire à état mixte 53
3.4.1 Modèle de prédiction à état mixte 54
3.4.2 Exploitation de la mesure dans le cas d'un état mixte . . . 54
3.5 Transition vers le suivi multi-cibles 56
3.5.1 Gestion des détections 57
3.5.2 Modèle d'observation intégrant les détections 58
3.5.3 Notion d'*identité* d'une cible 58
3.6 Suivi et ré-identification conjoints 59
3.6.1 Association traqueurs mixtes/détections 59
3.6.2 Modèle d'observation mixte intégrant les détections . . . 60

3.7 Implémentation . . . 62
3.7.1 Modélisation de l'apparence d'une cible . . . 62
3.7.2 Descriptions des identités du réseau . . . 62
3.7.3 Vecteur d'état . . . 63
3.7.4 Modèle de mouvement . . . 63
3.7.5 Modèle d'observation . . . 64
3.7.6 Caractérisation des paramètres libres de notre système . . . 64
3.8 Évaluations et analyses associées . . . 65
3.8.1 Jeux de données . . . 65
3.8.2 Critères et modalités évalués . . . 67
3.8.3 Performances de la méthode d'échantillonnage mixte . . . 68
3.8.4 Performances du suivi par ré-identification . . . 69
3.8.4.1 Performances quantitatives . . . 69
3.9 Conclusion . . . 72

II Système décisionnel haut-niveau 73

4 Supervision des identités : une approche réseau 75
4.1 Introduction . . . 76
4.2 État de l'art et positionnement des travaux . . . 76
4.2.1 Suivi de cibles multiples par logique différée à partir d'observations continues . . . 76
4.2.1.1 Principes d'association de détections . . . 77
4.2.1.2 Suivi monoculaire par logique différée . . . 77
4.2.2 Suivi à partir d'observations discontinues : réseaux à champs disjoints . . . 78
4.2.3 Notre approche . . . 80
4.3 Définitions . . . 82
4.3.1 Modélisation du réseau de caméra . . . 82
4.3.2 Données propres aux superviseurs . . . 83
4.4 Approche MAP trajectoriel . . . 83
4.4.1 Formalisation de la programmation dynamique . . . 84
4.4.2 MAP trajectoriel : mise en oeuvre . . . 85
4.4.2.1 Intégration temporelle . . . 85

4.4.2.2 Exclusivité de l'association 86
4.4.2.3 Optimisation des tracklets sur une séquence de suivi 87
4.4.3 Bilan du superviseur MAPT 88
4.5 Approche MCMC sur les trajectoires 90
4.5.1 Association de données MCMC 90
4.5.2 Formulation du problème 92
4.5.3 Modèle de vraisemblance 93
4.5.4 MCMC Data Association dirigé par apparence et topologie 94
4.6 Évaluations et discussions associées 95
4.6.1 Performances du MAPT 96
4.6.1.1 Performances quantitatives 96
4.6.1.2 Limitations du MAPT 97
4.6.2 Performances du MCMC 98
4.6.2.1 Tests sur données de synthèse 98
4.6.2.2 Tests sur données réelles 99
4.7 Conclusion 99

5 Vers un système évolutif 101
5.1 Introduction 102
5.2 Construction de la base d'identités 102
5.3 Filtrage des échantillonnages d'identité 103
5.4 Projection de la base d'identité par fonctions de transfert de luminance 104
5.5 Apprentissage statistique de modèle 105
5.6 Reconfiguration du réseau face à un capteur défaillant 105
5.7 Extensions des travaux 108
5.7.1 Reconnaissance d'activités / détection d'évènements dans un réseau de caméras 108
5.7.2 Au-delà du champ disjoint : utilisation de caméras PTZ . . 108
5.8 Conclusion 109

6 Conclusions et perspectives 111

Bibliographie 115

Notations

Probabilités

Les notations adoptées dans cette thèse suivent celles utilisées par Michael Isard dans [Isard, 1998] :

p	Loi de probabilité d'une variable continue
P	Probabilité d'une variable discrète
t	Lettre utilisée en indice pour désigner le temps discret
j	Lettre utilisée en exposant pour désigner un des J objets
i	Lettre utilisée en exposant pour désigner une des N particules
$\mathbf{x}$	Lettre représentant le vecteur d'état continu d'un système dynamique
id	Abréviation représentant le paramètre discret du vecteur d'état
$\mathbf{X}$	Lettre représentant le vecteur d'état mixte, faisant intervenir paramètres continus et paramètres discrets
$\mathbf{Z}$	Lettre représentant l'observation, généralement l'information image

Acronymes

MAPT	Maximum a posteriori trajectoriel
MSR	Mixed-state ré-identification
MCMC	Monte-Carlo Markov Chain
MOT	Multiple Object Tracking
MHT	Multiple Hypothesis Tracker
NOFOV	Non Overlapping Fields of View
REID	Reidentification
SIR	Sampling Importance Resampling

ELF	Ensemble of Localized Features [Gray et Tao, 2008]
SDALF	Symetry Driven Accumulation of Local Features [Farenzena *et al.*, 2010]

MOTP	Multiple Object Tracking Precision (métrique)
MOTA	Multiple Object Tracking Accuracy (métrique)
CMC	Cumulative Match Characteristic
nAUC	normalized Area Under the Curve
HSV	Hue Saturation Value
RGB	Red Green Blue
YCbCr	Luminance Chrominance bleue Chrominance rouge
ROI	Region Of Interest
BTF	Brightness Transfert Function
DBN	Dynamic Bayesian Model
DPM	Deformable Part Model [Felzenszwalb *et al.*, 2010]
EM	Expectation Maximisation
HMM	Hidden Markov Model
HOG	Histograms of Oriented Gradients [Dalal et Triggs, 2005]
KLT	Kanade–Lucas–Tomasi feature tracker [Shi et Tomasi, 1994]
PRDC	Probabilistic Relative Distance Comparison [Zheng *et al.*, 2011]
STEL	STructure ELement [Jojic *et al.*, 2009]
SVM	Support Vector Machine
VIPeR	Viewpoint Independant Pedestrian Recognition [Gray *et al.*, 2007]
RFID	Radio Frequency Identification
PTZ	Pan Tilt Zoom (caméra)

CHAPITRE

1 Introduction et contexte des travaux

Ce chapitre pose les bases de l'étude menée au cours de cette thèse et décrit la problématique abordée. Il introduit le domaine de la Vision par Ordinateur et sa déclinaison en Vidéosurveillance. Ici, l'accent est mis sur la supervision de réseaux de caméras, sujet de la thèse. Le chapitre présente ensuite les objectifs fonctionnels visés par notre système, à savoir le suivi d'objets multiples dans un réseau de caméras. Un panel des systèmes répondant à cette problématique est esquissé. Ils s'appuient sur des fonctionnalités perceptuelles canoniques (détection, classification, suivi) dont l'état de l'art associé se retrouvera détaillé au fil des chapitres de ce manuscrit. Ce chapitre vise à appréhender notre système complet et justifier/discuter ses fonctionnalités sous-jacentes.

1.1 Contexte général et enjeux

Bien loin de la description apocalyptique esquissée par George Orwell dans son roman d'anticipation *Nineteen Eighty-Four* [Orwell, 1949] décrivant les abus d'un état totalitaire usant notamment de vidéosurveillance pour se maintenir en place, il est tout de même intéressant de noter que c'est la ville de l'écrivain, Londres, qui compte à ce jour le plus de caméras de vidéosurveillance par habitant.

Les premières utilisations de caméras de vidéosurveillance remontent aux années 1950. Leur exploitation au sein de systèmes de télévision en circuit fermé (TVCF) sont apparues en 1970 mais c'est à partir des années 1990 que la vidéosurveillance a vraiment commencé à s'implanter.

Démocratisation de l'installation de caméras Les attentats de septembre 2001 aux États-Unis et de 2005 à Londres ont contribué à l'explosion du nombre de caméras ([Gouaillier et Fleurant, 2009], [Frost et Sullivan, 2007]). D'après les sources de Norris *et al.* [Norris *et al.*, 2004], un rapport faisait état d'approximativement quatre millions de caméras (toutes caméras confondues) en Grande-Bretagne, soit une caméra pour dix-sept habitants. Devant une telle quantité de caméras, les agents de sécurité ne peuvent pas toutes les surveiller en même temps. Ils sont contraints de ne regarder que ponctuellement chacune d'entre elles en permutant régulièrement (l'image 1.2 illustre la complexité de la tâche). Une étude de Dee et Velastin [Dee et Velastin, 2008] rapporte qu'à un instant donné et dans le meilleur des cas, une caméra sur quatre est sous l'observation d'un opérateur et dans le pire des cas ce ratio peut tomber à une caméra pour soixante-dix-huit. De plus, la répétitivité de la tâche, la faible fréquence des situations dangereuses ou anormales et la lassitude occasionnée, rendent le travail de surveillance particulièrement difficile et inefficace. Certaines estimations de Gouaillier et Fleurant [Gouaillier et Fleurant, 2009] font état d'une chance sur mille de réagir en direct à un évènement anormal. Pour ces raisons, ce travail se doit d'être assisté, voire automatisé via des systèmes de vidéosurveillance ou de vidéo-assistance intelligents. En outre, nombre de ces caméras sont utilisées pour la surveillance d'espaces publics très fréquentés (supermarchés, gares, axes routiers, rues ou places piétonnes, etc). Ce type de scène où les risques sont d'autant plus grands que le nombre de véhicules ou de personnes est important, est aussi particulièrement difficile à surveiller à cause justement de cette forte densité.

Au-delà de la surveillance d'espaces publics, ces caméras sont également déployées à domicile pour l'aide au maintien des personnes, la surveillance de sites sensibles, les installations domestiques pour particuliers... Dans toutes ces applications, le système doit être en mesure de fournir une localisation qualitative des usagers du lieu surveillé. Ensuite, l'utilisation de cette information sera dépendante de l'application. Dans une maison de retraite, le système tendrait à prévenir le per-

sonnel encadrant, alors que sur un site sensible, une alarme serait déclenchée.

Ainsi, de plus en plus de caméras sont installées. Cependant elles partagent souvent peu de champs de vue avec leurs voisines car elles surveillent des espaces de plus en plus larges.

Les solutions du marché, vision et autre Il existe de nombreuses technologies permettant d'obtenir une localisation qualitative d'individus. Parmi ceux-ci, nous pouvons lister le GPS (pour « Global Positionning System »), les puces RFID (pour « Radio Frequency IDentification »), l'utilisation d'un capteur laser... Chacune de ces technologies présente son lot d'avantages et d'inconvénients. Par exemple, le GPS fonctionne mal en intérieur, les puces RFID ont un champ d'action restreint et doivent être portées par les individus à suivre, donc nécessitent leur coopération...

Pour une surveillance d'intérieur, de lieux à grande échelle, la vidéosurveillance présente de nombreux avantages : mise en oeuvre simple, pas de perte de signal, pas de coopération nécessaire...

1.2 Que dit la législation ?

Cependant, la vidéosurveillance d'espaces publics comme privés induit des questions de droit à l'image. Et de ce point de vue, la législation n'est pas identique dans tous les pays. En France, la législation est particulièrement pointue. La surveillance de lieu public par les forces de l'ordre doit explicitement être signalée par des panneaux tels que l'exemple de la figure 1.1. Ainsi, la population est prévenue que son droit à l'image entre en jeu.

(a) (b)

FIGURE 1.1 – Exemple de panneaux d'avertissement (a) en France indiquant que le métro parisien est une zone vidéosurveillée, et (b) dans les transports en commun de Grande-Bretagne.

1.2.1 Droit à l'image (loi du 21 janvier 1995)

Loi n°95-73 du 21 janvier 1995 d'orientation et de programmation relative à la sécurité : articles 10 (Extrait) *« ... La transmission et l'enregistrement d'images prises sur la voie publique, par le moyen de la vidéo surveillance, peuvent être mis en oeuvre par les autorités publiques compétentes aux fins d'assurer la protection des bâtiments et installations publics et de leurs abords, la sauvegarde des installations utiles à la défense nationale, la régulation du trafic routier, la constatation des infractions aux règles de la circulation ou la prévention des atteintes à la sécurité des personnes et des biens dans des lieux particulièrement exposés à des risques d'agression ou de vol. La même faculté est ouverte aux autorités publiques aux fins de prévention d'actes de terrorisme ainsi que, pour la protection des abords immédiats de leurs bâtiments et installations, aux autres personnes morales, dans les lieux susceptibles d'être exposés à des actes de terrorisme. Il peut être également procédé à ces opérations dans des lieux et établissements ouverts au public aux fins d'y assurer la sécurité des personnes et des biens lorsque ces lieux et établissements sont particulièrement exposés à des risques d'agression ou de vol ou sont susceptibles d'être exposés à des actes de terrorisme. Les opérations de vidéo surveillance de la voie publique sont réalisées de telle sorte qu'elles ne visualisent pas les images de l'intérieur des immeubles d'habitation ni, de façon spécifique, celles de leurs entrées. Le public est informé de manière claire et permanente de l'existence du système de vidéo surveillance et de l'autorité ou de la personne responsable. ... »*

La vidéosurveillance présente en effet un aspect intrusif, et c'est sur ce point que la législation vient fixer des limites. L'utilisation de traitements automatiques de la vidéo vise à réduire cet aspect. Classiquement, les algorithmes cherchent à détecter des anomalies dans la scène qu'ils surveillent. On peut facilement imaginer ne stocker que les flux correspondant aux événements détectés.

Le fort développement que connaissent les systèmes de vidéosurveillance s'accompagne d'une augmentation de l'activité de la recherche. L'objectif est de développer des systèmes « intelligents » de vidéosurveillance qui puissent remplacer la vidéosurveillance traditionnelle (figure 1.2). En effet, il est humainement difficile pour un opérateur de surveiller simultanément un grand nombre de caméras et de ne pas rater un événement qui ne dure que quelques secondes. Ainsi, aujourd'hui, le but des travaux de recherche en vidéosurveillance est de pouvoir, dans la mesure du possible, accomplir automatiquement des tâches de surveillance. Si l'on s'intéresse seulement à la surveillance de personnes, celles-ci sont liées aux

FIGURE 1.2 – Illustration d'un mur vidéo classique.

thématiques suivantes [Hu *et al.*, 2004] :

1) Contrôle des accès. Dans certains lieux de haute sécurité comme les bases militaires, seules les personnes habilitées sont autorisées à entrer. Après la constitution d'une base de données biométriques des personnes habilitées, lorsqu'un visiteur se présente, le système pourra obtenir automatiquement les caractéristiques de la personne telles que sa taille, l'apparence de son visage à partir d'images prises en temps réel et décider si la personne est autorisée ou non à entrer dans le bâtiment.

2) Identification de personnes. L'identification des personnes à distance par un système de surveillance intelligent peut aider la police dans la recherche des personnes suspectes. La police peut construire une base de données biométriques des suspects et placer des systèmes de vidéo-surveillance à des endroits où les personnes recherchées ont l'habitude d'être comme, par exemple, les stations de métros, les casinos, etc. Le système doit pouvoir traiter automatiquement les personnes aperçues et juger si elles sont suspectes ou non.

3) Analyse et statistique des flux et congestion. En se basant sur la détection de personne, les systèmes de vidéo-surveillance peuvent automatiquement déterminer et analyser le flux de personnes dans des lieux publics tels que les centres commerciaux ou des sites touristiques afin de prévenir les problèmes de congestion.

4) Détection et alerte en cas d'anomalie. Dans certaines situations, il est important de pouvoir analyser et déterminer si le comportement d'une personne ou d'un groupe de personne est normal ou non (vol dans un supermarché, agression dans un parking, dégradation de biens...), si des objets sont abandonnés par des

individus,... Lorsqu'un comportement suspect est détecté, le système peut alerter les services de sécurité qui pourront intervenir le plus rapidement possible.

5) Actimétrie à domicile. Dans le contexte de l'aide au maintien des personnes à domicile, les applications mettront en jeu des techniques de reconnaissance de posture avec pour objectif la détection de chutes.

1.3 Positionnement de nos travaux

Parmi les différentes problématiques citées ci-dessus, nos travaux portent sur l'identification de personnes. Le sujet traite de la surveillance et du suivi de personnes en environnement intérieurs/extérieurs à large échelle. La surveillance à l'aide de caméras à champs de vue disjoints (abrégé NOFOV dans la suite, pour « Non Overlapping Fields of View ») présente des avantages certains : la redondance d'information permet plus de robustesse dans les tâches à accomplir, notamment face aux occultations engendrées par les éléments de la scène, et parfois même par les autres cibles suivies. Toutefois, les contraintes matérielles/économiques limitent en général le nombre de caméras et empêchent une couverture exhaustive de l'espace, souvent à large échelle. Ceci engendre des discontinuités dans le champ de vue du réseau et on parle alors de réseaux à champs de vue disjoints comme illustré en figure 1.3.

1.3.1 Vue d'ensemble

La littérature est très riche sur les réseaux de caméras. Par exemple, des travaux récents s'intéressent à l'inférence de « modèles d'activités » à partir de trajectoires reconstruites en réseaux à champs disjoints. Wang et al. [Wang *et al.*, 2010] regroupent les trajectoires de différentes caméras en "activités" selon leurs distributions et leurs directions de mouvements sans chercher à résoudre le problème d'association caméra à caméra. L'approche se base sur des trajectoires intra-caméras d'objets d'intérêts et les regroupe. Les applications visées concernent la surveillance routière et donc permettent l'utilisation d'un module bas niveau de suivi. A l'inverse, Loy et al. [Loy *et al.*, 2010] s'intéressent à des vidéos ne permettant pas l'utilisation d'un tel module, de par leur densité de personnes et leur faible résolution. Ils adoptent donc une approche de segmentation sémantique de leurs champs de vue, à partir de soustraction de fond.

Il s'agit là d'approches sur les réseaux de caméras qui ne rentrent pas dans notre contexte applicatif. En effet, nous écartons le cas de la surveillance de foules et faisons l'hypothèse de pouvoir disposer d'algorithmes de suivi d'individus. Dans ce cadre là, un système de surveillance est composé de plusieurs fonctions canon-

FIGURE 1.3 – Exemple de réseau de caméras à champs de vue disjoints déployé dans un bâtiment et en zone extérieure, présentant de forts changements de point de vue, avec la position des caméras illustrée sur un plan du bâtiment.

iques [Javed et Shah, 2008, Gong *et al.*, 2011] :

- ▷ la **détection** de personnes. L'objectif est de localiser les personnes présentes dans une image, généralement avec des boites englobantes. Il n'y a ici pas d'analyse temporelle, le traitement ne concerne qu'une seule image ;
- ▷ le **suivi** de personnes. Contrairement à la détection, le problème est ici temporel. L'objectif est de fournir au cours du temps les positions (dans le plan image ou dans le plan du sol, selon le type de suivi) des cibles mobiles, typiquement des personnes dans un environnement humain. Les résultats d'un détecteur de personnes vont pouvoir venir aider ces méthodes ;
- ▷ la **ré-identification** de personnes, qui consiste à reconnaître un individu parmi un ensemble de personnes, dans différents lieux, au sein d'une caméra ou d'un réseau de caméras sans champs de vue communs.

Les travaux présentés dans cette thèse ont pour objectif le suivi de personnes en environnements intérieurs à large échelle. Les contraintes matérielles/économiques limitent en général le nombre de caméras et empêchent une couverture totale de l'espace, ce qui engendre des discontinuités dans le champ de vue du réseau. On parle de réseaux à champs de vue disjoints comme illustré en figure 1.3.

L'enjeu pour le processus de suivi est alors de gérer ces discontinuités dans le champ de vue du réseau pour assurer la cohérence spatio-temporelle des trajectoires inférées des cibles. Le verrou auquel nous faisons face est double. Le premier élément à mettre en place est un système de suivi de cibles multiples. Dans

le reste du manuscrit, nous utiliserons l'acronyme MOT (pour « Multiple Object Tracking ») pour désigner ces techniques. Par ailleurs, le réseau proposé présente des champs de vue disjoints, *i.e.* l'information visuelle disponible sur les cibles sera discontinue. Un système de MOT seul ne peut faire face à ce genre de discontinuités. La notion sous-jacente est celle de ré-identification : être capable de reconnaître un individu déjà observé, et ce, dans la même caméra suite à une sortie du champ de vue ou dans une autre, avec un capteur présentant une réponse colorimétrique différente, un point de vue différent, et une scène différente. Dans la suite, nous utiliserons l'acronyme REID pour désigner le terme de ré-identification.

1.3.2 Suivi multi-cibles

Le MOT vise à caractériser la trajectoires des cibles par analyse du flux vidéo. Un tel système repose classiquement sur un détecteur, qui l'alimente et lui permet d'initialiser et de terminer des pistes de suivi sans intervention de l'utilisateur. À partir du moment où il y a plus d'une cible dans la scène, l'algorithme fait face à un problème d'association de données. Les techniques classiques reposent sur la définition d'un modèle d'apparence pour représenter les cibles, les discriminer de la scène et les discriminer entre elles, et sur une stratégie de recherche dans le flux vidéo. Nous distinguons à ce niveau deux stratégies possibles. La recherche peut se faire avec une logique séquentielle, *i.e.* les traitements sont effectués à la cadence image du flux vidéo (*e.g.* les méthodes markoviennes relèvent de cette logique) ou avec une logique différée, *i.e.* les traitements d'association et de suivi sont déportés à la fin d'une fenêtre temporelle, au cours de laquelle l'algorithme a agrégé des informations. D'une manière similaire, nous ferons la distinction entre les approches distribuées, où chaque traqueur est instancié sur chaque nouvelle cible et indépendant des autres, aux approches centralisées, où l'estimation de l'état des diverses cibles est inféré via un seul filtre, de manière conjointe.

1.3.3 Ré-identification

À la lumière de ces travaux, le problème de la REID dans un réseau de caméras présente des caractéristiques similaires au problème du MOT. En effet, les deux se présentent comme une mise en correspondance temporelle de détections décrites de manière appropriée. Cependant, la REID est une version relaxée du MOT, car les contraintes spatio-temporelles liant les détections à apparier en REID sont beaucoup plus faibles. Le problème de la REID se décompose donc en trois sous-problèmes.

- ▷ Problème N°1 : Quelles informations de la silhouette seront invariantes entre différentes caméras, comment les obtenir, et quelle distance utiliser pour la comparaison ? En bref, comment générer une fonction de similarité ayant

du sens entre plusieurs caméras ?

- ▷ Problème N°2 : Comment ré-identifier des cibles au sein d'une caméra ? De simples détections peuvent-elles suffire ou faut-il réaliser un suivi des pistes ? La ré-identification peut-elle bénéficier de ce suivi ?
- ▷ Problème N°3 : Comment la connaissance du réseau peut-elle permettre de contraindre l'appariement à réaliser, et ainsi simplifier la combinatoire d'association ?

Lorsque l'on s'intéresse à la surveillance d'un réseau de caméra, une base d'images de piétons déjà segmentés n'est pas un jeu de donnée suffisant. En effet, ceci occulte les problèmes N°2 et 3 de la classification ci-dessus. Or il n'existe aucunes vidéos publiques concernant les réseaux de caméras à champs disjoints, les travaux les plus proches de l'application visée [Kuo *et al.*, 2010a, Matei *et al.*, 2011] n'ayant pas rendu leurs jeux de données publics. Le tableau 1.1 résume les différents jeux de données concernant le MOT et la REID.

Nous avons donc dû réaliser nos propres séquences vidéos : NOFOV0, NOFOV1 et NOFOV2. Avec la législation française, l'acquisition de ces données vidéos a dû donner lieu à de nombreux affichages pour prévenir les personnels des acquisitions en cours, dans les locaux du CEA-LIST, ainsi que, dans une moindre mesure, ceux du LAAS-CNRS. Le chapitre 3 présente nos séquences en détails. Chacune d'elles met en scène un groupe de piétons évoluant dans un bâtiment et ses alentours (une ou deux caméras en extérieur sur un parking, selon les séquences). Une première caméra représente le hall d'entrée, vu comme point d'entrée des personnes dans le réseau. Les individus évoluent entre les différentes caméras, ce qui génère un problème conjoint de MOT et de REID dans le réseau.

1.4 Contributions et organisation du manuscrit

Pour proposer une localisation qualitative des personnes évoluant dans un réseau de caméras tel que celui de la figure 1.3, le système visé doit pouvoir suivre les cibles dans les caméras, *i.e.* faire du MOT, ainsi que les ré-identifier entre les différentes caméras pour combler les discontinuités d'observation inhérentes à ce genre de réseaux. La figure 1.4 présente un synoptique du système que nous proposons pour répondre à ce problème.

Au niveau caméra, le système se base sur le détecteur de piéton de Dalal et Triggs dans [Dalal et Triggs, 2005] : HOG. Ces détections viennent alimenter des filtres distribués à état mixte MSR (pour « Mixed-State Re-identification ») inférant position et identité dans une base pour les cibles qu'ils suivent. Les traitements au niveau de la caméra doivent relever de la logique séquentielle pour pouvoir fournir une réponse à chaque instant image. L'algorithme de MOT est dérivé de [Breitenstein *et al.*, 2009], et incrémenté par une inférence d'identité. Ces filtres

	Séquence	nombre de cibles	nombre de caméras	statut	difficulté
	Vidéos				
MOT	iLIDS [a]	?	?	payant	
MOT	CAVIAR [b]	72	2	public	
MOT	**PETS S2L1** [c]	10	1	public	
MOT+REID	[Kuo *et al.*, 2010a]	?	3	privé	-
MOT+REID	[Matei *et al.*, 2011]	2759 véhicules	8	privé	- -
MOT+REID	**NOFOV0**	16	5	en cours	-
MOT+REID	**NOFOV1**	7	3	en cours	-
MOT+REID	**NOFOV2**	12	5	en cours	+
	Images				
REID	**VIPeR**	632	2	public	+ +
REID	ETHZ REID [d]	80	1	public	- -
REID	iLIDS REID [e]	?	2	payant	?
REID	CAVIAR REID [f]	50/22	2/1	public	-

TABLE 1.1 – Jeux de données pour MOT et REID. La difficulté concerne la REID et est évaluée par rapport aux changements de points de vue des caméras. Nous mentionnons en gras les bases que nous avons utilisées et qui seront présentées en détails dans la suite.

a. http://www.homeoffice.gov.uk/science-research/hosdb/i-lids/
b. http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/
c. http://www.cvg.rdg.ac.uk/PETS2009/a.html
d. http://homepages.dcc.ufmg.br/~william/datasets.html
e. http://www.homeoffice.gov.uk/science-research/hosdb/i-lids/
f. http://www.lorisbazzani.info/code-datasets/caviar4reid/

MSR produisent des distributions d'identités, reflétant leur « avis » sur l'identité de la cible qu'ils suivent.

Au niveau du réseau, ces distributions d'identité sont filtrées par un superviseur ayant la connaissance de tous les filtres MSR actifs, ainsi que de la topologie du réseau. Les traitements ne nécessitant pas de réponse immédiate à ce niveau là, le superviseur relève lui de la logique différée, et prend le temps d'agréger suffisamment d'information avant de produire une décision de ré-identification.

Il est enfin envisagé d'exploiter cette décision renforcée de REID, pour permettre des traitements supplémentaires, comme une exploitation de cette information dans les filtres de suivi. Nous nous sommes placés ici dans une démarche ascendante, des capteurs vers le superviseur. Nous explorons avec ces briques les possibilités d'une démarche double-flux, *i.e.* faire redescendre les décisions vers les éléments de traitements locaux aux capteurs.

Dans ce cadre, le plan de la thèse se décline comme suit :

Le chapitre 2 traite de la première question soulevée par la ré-identification : comment décrire de manière à la fois précise et robuste aux changements de caméras l'apparence d'une personne, supposée ici *a priori* segmentée. La littérature est con-

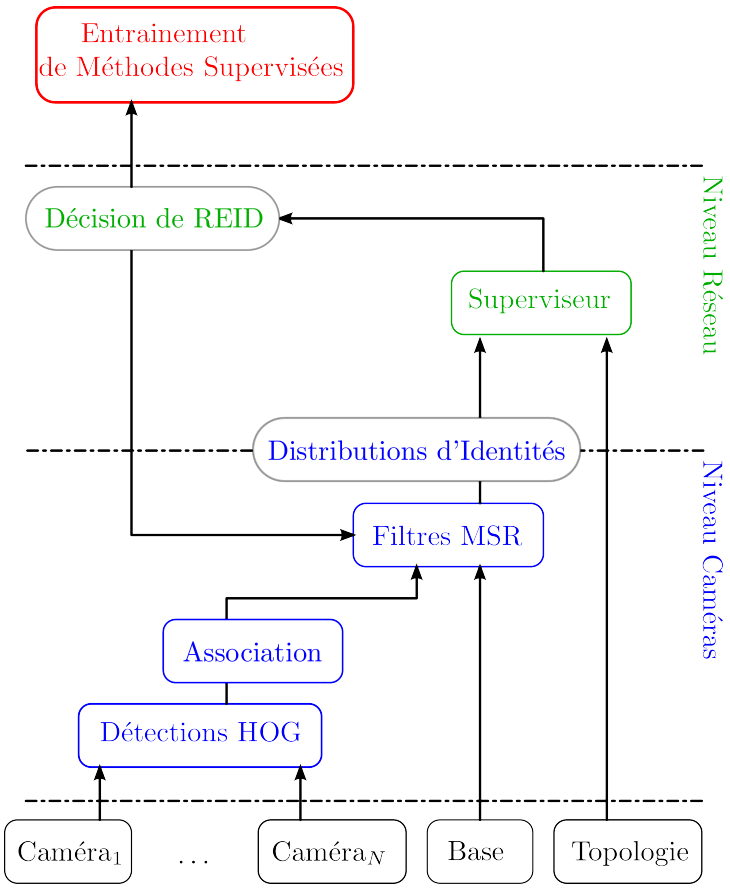


FIGURE 1.4 – Diagramme synoptique de notre approche complète. N caméras génèrent des observations (a). Chacune est surveillée par des traqueurs locaux à la caméra, lançant des cibles de suivi (b), tout en ayant connaissance de la base de personnes à ré-identifier (c). Ces traqueurs génèrent un ensemble de tracklets (ou cibles de suivi dans une caméra) (d) associées à leurs zones de début et de fin ainsi que leurs temps. Le superviseur (e) a la connaissance de l'ensemble du réseau et centralise toutes ces informations. Il a aussi la connaissance de la topologie du réseau (f), *i.e.* de la relation de voisinage entre les caméras à champs de vue disjoints. Ceci lui permet de fournir une réponse de ré-identification optimisée (g).

séquente sur ce sujet et nous consacrons ce chapitre à la présentation des approches classiques, ainsi que leurs contraintes d'application. Il en résulte une classification de ces méthodes, et le choix justifié d'un descripteur pour nos travaux.

Le chapitre 3 présente la philosophie de l'approche que nous proposons dans cette thèse : considérer la ré-identification comme un problème de suivi multi-cibles généralisé. Nous introduisons dans ce chapitre un processus d'estimation conjointe entre position et identité de la cible, travaillant au niveau de chacune des

caméras. En effet, avant de considérer le réseau, il est essentiel d'être capable de ré-identifier au niveau des caméras.

Le chapitre 4 concerne l'extension au réseau via un superviseur. Ce chapitre propose deux stratégies d'optimisation des ré-identifications proposées, en s'appuyant sur la topologie du réseau supposée connue *a priori*. Nous montrerons que l'intégration d'un processus de suivi au niveau des caméras et la localisation qualitative des cibles induite permet d'exploiter une topologie relativement fine du réseau, en zone d'entrée/sortie des caméras.

Finalement, le chapitre 5 énumère plusieurs extensions de l'approche développée, en terme de démarche descendante, du superviseur vers les capteurs.

Liste de publications Nous donnons ci-dessous la liste des publications réalisées au cours de cette thèse :

- [Meden *et al.*, 2012a] Meden, B., Lerasle, F. et Sayd, P. (2012a). MCMC Supervision for People Re-Identification in Nonoverlapping Cameras. *In Proceedings of the British Machine Vision Conference*, Guildford, Angleterre
- [Meden *et al.*, 2012c] Meden, B., Sayd, P. et Lerasle, F. (2012c). Suivi par ré-identification dans un réseau de caméras à champs disjoints. *In Actes du congrès francophone sur la Reconnaissance des Formes et l'Intelligence Artificielle (RFIA)*, Lyon
- [Meden *et al.*, 2013] Meden, B., Lerasle, F. et Sayd, P. (2013). Suivi par ré-identification dans un réseau de caméras à champs disjoints. *Traitement du Signal (à paraître)*
- [Meden *et al.*, 2012b] Meden, B., Lerasle, F. et Sayd, P. (2012b). Tracking-by-Reidentification in a Non-Overlapping Field of View Camera Network. *In VISAPP*, Rome, Italie
- [Achard *et al.*, 2012] Achard, C., Ambellouis, S., Meden, B., Lefebvre, S. et Truong Cong, D. N. (2012). *Suivi d'objets d'intérêt dans un réseau de caméras*. Hermès Science Publications, http://www.editions-hermes.fr/
- [Meden *et al.*, 2011b] Meden, B., Sayd, P. et Lerasle, F. (2011b). Mixed-State Particle Filtering for Simultaneous Tracking and Re-Identification in Non-Overlapping Camera Networks. *In Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, Ystad, Suède
- [Meden *et al.*, 2011a] Meden, B., Sayd, P. et Lerasle, F. (2011a). Mixed-state condensation pour suivi et ré-identification simultanés dans des réseaux de caméras à champs de vue disjoints. *In Actes du congrès francophone des jeunes chercheurs en vision par ordinateur (ORASIS)*, Praz-sur-Arly
- Rédaction en cours d'un article pour le journal Computer Vision and Image Understanding (CVIU)

Première partie

Traitements au niveau des caméras

CHAPITRE

2

Modèles de ré-identification

Lorsque l'on parle de ré-identification entre différentes caméras, l'invariance de descriptions est le premier élément qui vient à l'esprit. La littérature est très riche sur le sujet et ce chapitre propose une présentation relativement exhaustive de ces travaux. Nous distinguons principalement les approches se focalisant sur une paire de caméras et proposant un apprentissage dédié de la signature et les approches directes. Nous conservons à l'esprit notre contexte applicatif : la ré-identification de réseaux de caméras à champs disjoints.

2.1 Introduction

Lorsque l'on souhaite décrire un objet en vue de sa ré-identification lorsqu'on le reverra, dans des conditions éventuellement différentes, il est nécessaire de se créer une représentation de cet objet, et de ses caractéristiques spécifiques, qui le distingueront des autres objets présents. Au-delà de l'introduction « formelle » de la thèse, donnée au chapitre précédent, nous nous concentrons ici sur un des aspects clés de la ré-identification : le choix de la représentation. Comment décrire de manière appropriée un individu que l'on sera amené à observer sous différents points de vue par la suite, éventuellement depuis des caméras différentes, pour être en mesure de le reconnaître ?

Dans notre contexte applicatif de la surveillance de réseaux à champs disjoints, il existe peu de stratégies de ré-identification clairement identifiées. La signature choisie doit notamment être peu gourmande en temps de calcul. Nous réalisons donc ici une étude bibliographique détaillée sur les descripteurs de ré-identification proposés par la littérature, en vue d'en choisir un pour notre application.

Dans la suite du manuscrit nous appellerons méthode supervisée toute méthode de ré-identification se basant sur un apprentissage à partir d'images vérité terrain appariées entre deux caméras. Par opposition, nous qualifierons de non-supervisée toute méthode ne reposant pas sur cet *a priori*.

Dans ce contexte applicatif, la section 2.2 évoque le cas de la biométrie. Ces méthodes semblant répondre à ce problème seront écartées car inapplicables dans le cas multi-caméra. Ensuite, la section 2.3 pose le problème de la ré-identification de personnes tel qu'il est envisagé dans le contexte de la vidéosurveillance. Les sections 2.4 et 2.5 déclinent en détails les deux principales classes d'approches de ré-identification que nous distinguons pour répondre à ce problème : les méthodes s'appuyant sur des appariements inter-images (méthodes supervisées) et les méthodes sans connaissance *a priori*, *i.e.* directes. La section 2.6 discute, aux vues des évaluations réalisées, de la technique la plus appropriée pour notre contexte applicatif de surveillance dans les réseaux de caméras et finalement la section 2.7 résume et conclut ce chapitre.

2.2 Ré-identification par caractéristiques biométriques

Biométrie : Le mot biométrie signifie littéralement « mesure du vivant », et désigne dans un sens très large l'étude quantitative des êtres vivants. Parmi les principaux domaines d'application de la biométrie, on peut citer l'agronomie, l'anthropologie, l'écologie et la médecine.

L'usage de ce terme se rapporte de plus en plus à l'usage de ces techniques à des fins de reconnaissance, d'authentification et d'identification. Le Petit Robert

la définit comme une « Science qui étudie à l'aide de mathématiques (statistiques, probabilités) les variations biologiques à l'intérieur d'un groupe déterminé ».

En Vision par Ordinateur, et plus particulièrement en vidéosurveillance, les techniques de biométrie classiquement utilisées reposent sur une reconnaissance faciale. Toutefois, ces méthodes requièrent des conditions d'acquisition particulièrement contraintes comme une résolution importante du visage, une vue de face pour les méthodes 2D, *etc.* Des états de l'art récents sur ces méthodes sont présentés dans [Jafri et Arabnia, 2009, Germa *et al.*, 2009]. La figure 2.1 illustre ce contexte applicatif à travers le portique destiné aux aéroports élaboré par la firme Morpho. Herold *et al.* présentent dans [Herold *et al.*, 2011] une technique de suivi de visage en 3D au sein du portique, à des fins d'identification faciale. L'un des pré-requis est que la personne regarde la caméra. Le passage dans un portique étroit rend cette tâche plus aisée.

FIGURE 2.1 – Portique d'aéroport dans lequel est effectué la solution de suivi de visage à des fins de reconnaissance faciale par l'entreprise Morpho [Herold *et al.*, 2011].

L'analyse de démarche [Kim *et al.*, 2008] fait aussi partie des caractéristiques biométriques. Toutefois, il s'agit là d'investigations assez récentes, et difficilement applicables aux réseaux de caméras, ne supposant aucune collaboration de la part des usagers observés (contrairement au passage dans un portique tel que celui présenté en figure 2.1).

Dans notre contexte applicatif, la ré-identification intervient dans un réseau, *i.e.* entre plusieurs caméras aux réponses photométriques différentes et déployéees sur des lieux différents, donc avec une pose et un fond différents. De plus, l'environnement est beaucoup moins contraint. Il est en général très peu probable d'obtenir des images de visages fronto-parallèles, ou d'être en mesure de reconstruire des visages en 3D, nos caméras d'ambiance ayant un large champ de vue. Pour toutes ces raisons, les techniques de biométrie telles que la reconnaissance faciale ou l' analyse de démarche sont donc écartées de notre étude.

2.3 Au delà de la biométrie : positionnement du problème

Ayant écarté les approches du type « biométrie à la volée » pour leur inadéquation à notre problème, les verrous pour notre application demeurent toutefois :

- ▷ comment décrire un individu pour le distinguer de ses semblables ? On retrouve l'un des problèmes auxquels fait face le MOT. Savoir identifier les personnes est un cas général de l'association de données du MOT.
- ▷ comment intégrer dans cette description une robustesse aux changements de point de vue et à la réponse photométrique (capteurs différents) ?

Les attributs de description classiquement utilisés pour caractériser un individu en REID [Gray *et al.*, 2007] sont proches de ceux utilisés en MOT, avec les distributions couleurs, introduites pour le suivi par [Pérez *et al.*, 2002, Nummiaro *et al.*, 2003] et la texture, *e.g.* au travers des points d'intérêt du traqueur KLT [Shi et Tomasi, 1994].

Au-delà de la description, le choix de la région ainsi décrite est aussi très important. En effet, une distribution couleur perd la notion de spatialité des pixels. Décrire plusieurs régions par des distributions couleurs indépendantes augmente la précision de la description [Pérez *et al.*, 2002]. Une autre solution consiste à décrire l'information par des distributions calculées dans un espace à cinq dimensions associant les trois canaux couleurs, *e.g.* RGB, aux positions x et y des pixels considérés. Ainsi la distribution modélise de manière conjointe la distribution de couleurs et sa répartition spatiale. Birchfield et Rangarajan proposent une version discrète de type histogramme étendu (appelé spatiogrammes par les auteurs) dans [Birchfield et Rangarajan, 2005] et Dickinson *et al.* modélisent ceci par des mixutre de gaussiennes dans [Dickinson *et al.*, 2009]. Nous verrons dans la suite que les travaux sur la ré-identification favorisent les modèles externalisant la localisation des signatures calculées. Ceci s'explique en partie par le fait que les approches basées sur de l'apprentissage statistique nécessitent des caractéristiques simples et nombreuses (*e.g.* les classifieurs faibles pour Adaboost [Freund et Schapire, 1995]).

2.3.1 Un problème de rang : focalisation sur la ré-identification entre deux caméras

Lorsque l'on considère la ré-identification au sein d'une paire de caméras, le problème se définit classiquement comme : retrouver la bijection existant entre les deux ensembles d'observations issus des deux caméras. L'hypothèse sous-jacente est celle du « *closed world* », selon laquelle les caméras observent les mêmes personnes. Cette hypothèse est généralement réaliste, mais impacte les conditions d'acquisitions des données, ou d'installation des caméras, *e.g.* avec une caméra surveillant l'unique entrée d'un bâtiment, et permettant de constituer une base des personnes présentes.

C'est en 2006, avec les travaux de Gheissari *et al.* dans [Gheissari *et al.*, 2006] que le communauté Vision commence véritablement à définir le problème comme tel. Ces derniers proposent deux signatures dédiées ré-identification, s'appuyant sur une segmentation en triangle de la silhouette. Le focus est ici mis sur une bonne localisation des signaux à comparer. Par ailleurs et principalement, il s'agit des premiers travaux introduisant les caractéristiques d'appariements cumulées (CMC pour « Cumulative Match Characteristic ») pour évaluer les performances. Ces outils sont issus du domaine de la biométrie, où l'on cherche à quantifier la justesse d'appariements R-R entre deux ensembles de cibles segmentées *a priori*, appelés galerie et requêtes.

Une matrice de confusion est calculée entre la galerie et les requêtes, comme présenté en figure 2.2.

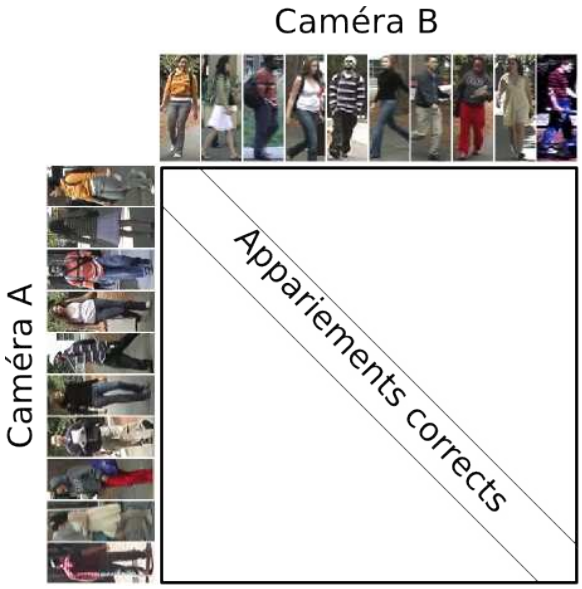


FIGURE 2.2 – Calcul de la matrice de confusion pour des exemples de la base VIPeR. La galerie et les requêtes présentent chacunes une seule occurence de chaque identité. Les appariements corrects sont sur la diagonale.

Courbes CMC Une courbe CMC représente la probabilité de trouver un appariement correct parmi les r meilleurs appariements, pour $r = 1 \ldots R$. La variable r est appelée le rang de ré-identification. Une courbe CMC atteignant les 100% au rang 1 représente les performances maximales atteignables par le système. L'algorithme 1 explicite le calcul d'une CMC, et la figure 2.3 illustre respectivement les cas idéal et usuel. D'autre part, Bazzani utilise dans [Bazzani, 2012] le score nAUC (pour « normalized Area Under the Curve »), l'aire sous la courbe CMC, exprimée en %. Tous les auteurs ne l'utilisent pas, mais nous reportons ce score à

chaque fois que cela est possible. Ceci permet d'exprimer les performances d'une méthode par un simple scalaire.

Algorithme 1: Algorithme de construction d'une courbre CMC.

Données : R exemples de galerie
R exemples de requêtes
S matrice de confusion entre la galerie et les requêtes. Mémorisation de l'indice de la requête dans des paires $[(similarité(i,j), indice_{requête})]_{1 \leq i,j \leq R}$
Résultat : courbe CMC

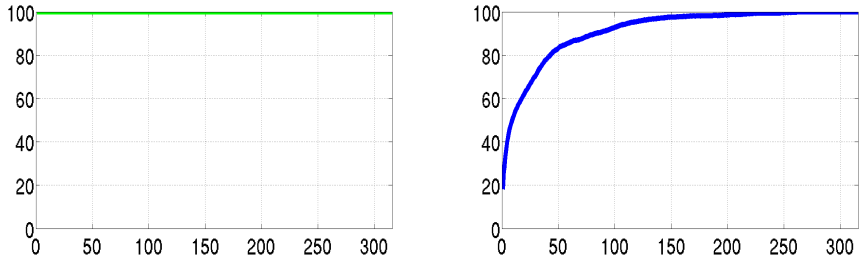
Trier les lignes de S selon la similarité en mémorisant les indices des requêtes ;
`// Parcours de la galerie`
pour *i = 1... R* **faire**
  `// Parcours des rangs de ré-identification`
  pour *j = 1... R* **faire**
    `// Si l'appariement galerie/requête est correct`
    si $S(i,j).indice == i$ **alors**
      $CMC(j) \leftarrow CMC(j) + 1$;
      break ;
    fin
  fin
fin
`// Accumulation des résultats`
pour *j = 1... R* **faire**
  $CMC(j) \leftarrow \sum_{i=1}^{j} CMC(i)$;
fin

Bolle *et al.* justifient dans [Bolle *et al.*, 2005] le lien entre la CMC et la Receiver Operating Curve (ROC) classiquement utilisée pour évaluer des problèmes de classification binaire tels que la détection de piétons.

Avec ce mode d'évaluation insistant sur les performances propres du modèle de description, sont apparues des bases d'images dédiées à la ré-identification entre deux caméras.

Base VIPeR Gray *et al.* proposent dans [Gray *et al.*, 2007] la base d'images publique VIPeR [1] (pour « **V**iewpoint **I**nvariant **Pe**destrian **R**ecognition »), le premier jeu de données dédié à la ré-identification de personnes. Cette base est con-

1. http://vision.soe.ucsc.edu/node/178



(a) cas idéal (b) performances usuelles

FIGURE 2.3 – Différentes courbes CMC où (a) 100% des paires sont correctement associées au premier essai et (b) 16% au rang 1 et il faut attendre le rang 270 (sur 316) pour atteindre 100%.

stituée de 632 personnes, vues une fois chacune dans deux caméras différentes, nommées galerie et requêtes. La figure 2.4 illustre quelques échantillons de cette base bien connue. Le challenge soulevé par ce jeu de données est de parvenir à apparier un maximum de silhouettes correctes de la première vers la deuxième caméra.

FIGURE 2.4 – Quelques exemples issus des deux caméras de la base VIPeR. La première ligne représente la galerie et la seconde les requêtes correspondantes.

Base ETHZ_REID Par la suite, deux autres bases d'images ont été proposées : ETHZ_REID [2] publiée originellement dans [Ess *et al.*, 2007] pour du suivi et dont les silhouettes ont été extraites par [Schwartz et Davis, 2009] et iLIDS_REID, extraite des séquences iLIDS par [Zheng *et al.*, 2009].

ETHZ_REID est une base d'images publique, divisée en trois sous-ensembles, composés de 83, 35 et 28 identités respectivement. Chacune de ces sous-bases est acquise depuis une caméra mobile dans une rue. Les images galeries et requêtes

2. http://homepages.dcc.ufmg.br/~william/datasets.html

des personnes proviennent donc du même capteur, mais avec des points de vues différents. La figure 2.5 présente le type d'images de cette base. Les difficultés présentées sont moindres relativement à VIPeR, car (i) le capteur est le même, et (ii) les changements de poses de caméra autorisés par la caméra mobile sont bien moins importants que ceux présentés par VIPeR.

iLIDS_REID est une base d'images privée, qui requiert la licence iLIDS MCTS, dont nous ne disposons pas. Nous nous contentons donc seulement de la mentionner.

FIGURE 2.5 – Quelques exemples issues de la base ETHZ_REID.

2.3.2 Quelques premiers constats sur la ré-identification

Le problème de la ré-identification sous-tend bien des questions de description et de comparaison. Le test d'un algorithme de ré-identification par courbes CMC sur l'une des bases d'images citées ci-dessus évalue sa capacité à retrouver la surjection existant entre deux ensembles, la galerie et les requêtes. En se plaçant dans le cas particulier d'une paire de caméras, une partie de la littérature s'est intéressée à l'intégration de connaissances *a priori* dans l'algorithme de mise en correspondance. Cet *a priori* prend la forme de paires vérité terrain de silhouettes, qui seront utilisées pour apprendre des invariants entre signatures inter-caméras.

Nous distinguons donc dans la suite deux classes d'approches :

- ▷ les méthodes exploitant la connaissance *a priori* d'un ensemble d'apprentissage pour se dédier à une paire de caméra. Nous discuterons par la suite la mise à l'échelle pour un réseau de caméras. Nous les qualifierons de méthodes supervisées.
- ▷ Les méthodes de descriptions directes, *i.e.* sans appariement préalable. Nous

les qualifierons de méthodes non-supervisées.

2.4 Ré-identification entre paires de caméras : méthodes supervisées

Lorsque l'on se focalise sur une paire de caméras, il devient envisageable de disposer d'un ensemble de paires d'images mises en correspondance *a priori*, *i.e.* des paires vérité terrain pour lesquelles le problème de la ré-identification est résolu. Disposer d'une telle base labellisée permet de recourir à un apprentissage. Deux stratégies d'apprentissage supervisé sont à distinguer dans la littérature : d'abord, celui d'une fonction de transfert colorimétrique pour la paire de caméras concernées, puis celui d'un modèle de description construit par apprentissage statistique pour être spécifiquement invariant entre ces deux caméras (que nous pouvons voir comme la généralisation du premier).

2.4.1 Fonction de transfert colorimétrique entre caméras

Dans [Porikli, 2003], Porikli propose une méthode initiale de calibration colorimétrique entre différentes caméras qu'il nomme « Brightness Transfer Function » (BTF). Il suggère que le changement d'illumination entre les vues peut être modélisé par une matrice de corrélation entre les histogrammes couleurs des deux images à mettre en correspondance. Il n'y a pas de changement de pose entre les deux vues, seulement d'illumination. La matrice répertorie les distances entre les valeurs des cellules des histogrammes couleur. Le calcul du chemin de coût minimal d'un coin de la matrice à l'autre fournit la fonction de transfert. Pour des images RGB, la méthode calcule une fonction par canal couleur.

Gilbert *et al.* utilisent aussi des matrices de corrélation et étendent dans [Gilbert et Bowden, 2006] ce concept en incorporant une méthode d'apprentissage en ligne pour mettre à jour les changements d'illumination entre les caméras. Toutefois, leur méthode se base sur une bonne initialisation de la fonction de transfert, et requiert entre 5000 et 10000 trajectoires d'entrainement acquises *a priori*.

Javed *et al.* proposent aussi une extension des travaux de Porikli dans [Javed *et al.*, 2005, Javed *et al.*, 2008], avec une application au problème de la ré-identification entre caméras à champs disjoints. Les travaux de Javed *et al.* [Javed *et al.*, 2005] basés sur les réponses radiométriques de caméras ont prouvé théoriquement que les BTF reliant deux caméras font partie d'un sous-espace de faible dimension. Supposant une approximation polynômiale de la réponse radiométrique, cette dimension est bornée par le degré de ce polynôme plus un [Javed *et al.*, 2005]. Ainsi, supposant disposer de paires d'observations des mêmes personnes dans les deux capteurs, il est possible d'estimer sur ces données cette fonction de transfert. Ils

estiment donc plusieurs fonctions de transfert, une par appariement dont ils disposent. Puis ils réalisent une analyse en composantes principales pour obtenir la fonction de transfert représentant au mieux le changement de caméras.

Lors du calcul d'une BTF entre différentes caméras, les objets d'intérêts ne sont pas non plus vus sous le même point de vue, *i.e.* les proportions des différentes couleurs ne sont plus forcément identiques. Pour dépasser cette difficulté, Prosser *et al.* proposent dans [Prosser *et al.*, 2008] une BTF cumulative. Pour ce faire, ils accumulent dans un même histogramme plusieurs images de la même personne dans un même capteur avant d'appliquer la méthode de Porikli entre ces histogrammes cumulés. Contrairement à [Javed *et al.*, 2005], plutôt que de calculer une moyenne dans l'espace des fonctions de transfert (par PCA probabiliste), la moyenne est calculée avant le calcul de la BTF, ce qui permet une meilleure prise en compte des cas rares [Prosser *et al.*, 2008].

Ci-dessous, nous détaillons rapidement la méthode de [Prosser *et al.*, 2008], que nous avons implémentée pour présenter des résultats sur VIPeR. Prosser *et al.* calculent un histogramme cumulé $\hat{H}_i$ pour les 256 niveaux de couleurs $B_1, ..., B_m, ..., B_{256}$ de N exemples d'entrainement dans la vue i. Ensuite ils accumulent les valeurs de ces cellules de manière à produire la distribution cumulée. L'équation (2.1) résume ces deux étapes : les histogrammes I_l des l^{e} exemples sont accumulés sur les N exemples et jusqu'au niveau de couleur B_m :

$$\hat{H}_i(B_m) = \sum_{k=1}^{m} \sum_{l=1}^{N} I_l(B_k) \tag{2.1}$$

Un histogramme cumulé similaire $\hat{H}_j$ est réalisé à partir des exemples de la vue j, ils sont normalisés par rapport au nombre de pixels intervenant, puis mis en correspondance à partir de la relation de base de la fonction de transfert. Considérons les observations d'un même objet vu dans les deux caméras. Comme il s'agit du même objet, la mise en correspondance des niveaux de couleur dans les deux vues donne la relation $\hat{H}_i(B_i) = \hat{H}_j(B_j) = \hat{H}_j(f_{ij}(B_i))$, où f_{ij} représente la fonction de transfert des niveaux de couleurs de l'image i vers l'image j. Ainsi, nous déduisons :

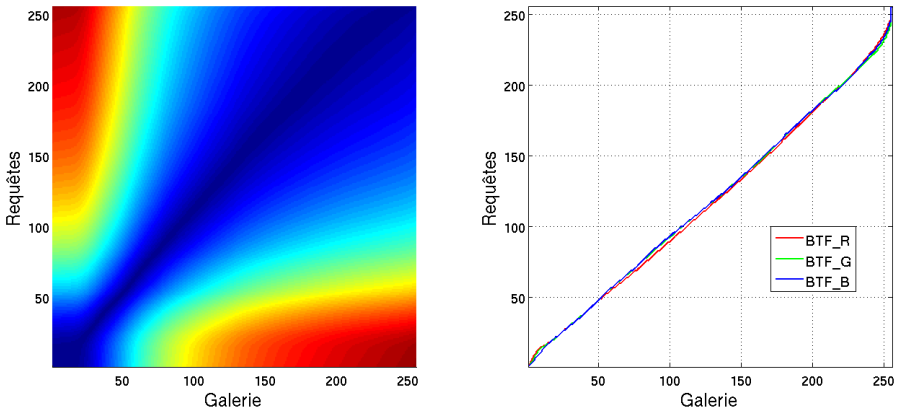
$$f_{ij}(B_i) = \hat{H}_j^{-1}(\hat{H}_i(B_i)) \tag{2.2}$$

Devant le peu de détails donné par la littérature pour le calcul de $\hat{H}_j^{-1}$, nous utilisons la méthode de Porikli [Porikli, 2003] pour le calcul de la mise en correspondance. Nous calculons une matrice de corrélation $C_{256\times256} = [c_{km}]_{1\leq k,m\leq 256}$ entre les histogrammes cumulés :

$$c_{km} = |\hat{H}_i(B_k) - \hat{H}_j(B_m)| \tag{2.3}$$

Le calcul du plus court chemin d'un coin à l'autre de la matrice de corrélation fournit la BTF, représentant la mise en correspondance des niveaux de couleur d'une

caméra avec ceux de l'autre. Ce calcul est effectué de manière indépendante pour chacun des canaux couleur. La figure 2.6(a) représente la matrice de corrélation du canal rouge, calculé sur 316 paires de la base VIPeR avec la méthode de [Prosser *et al.*, 2008]. La figure 2.6(b) représente les BTF des 3 canaux, rouge, vert et bleu, estimées avec la méthode du plus court chemin de [Porikli, 2003].



(a) (b)

FIGURE 2.6 – Matrice de corrélation entre histogrammes cumulés du canal rouge (a) calculé sur VIPeR, et (b) BTF déduite sur les 3 canaux.

Finalement, nous avons calculé un appariement sur la partie test de la base VIPeR, à partir de distributions couleur RGB, localisées spatialement dans la silhouettes (cinq bandes horizontales non recouvrantes). Nous avons ensuite appliqué ces BTF à la galerie et recalculé l'appariement. La figure 2.7 donne les courbes CMC avec et sans application de la BTF. Le descripteur choisi est volontairement simple, ce qui explique les performances médiocre, mais on s'intéresse aux performances relatives. Bien que la base VIPeR ne présente pas de changements d'illumination conséquents (la BTF est proche de la première bissectrice, d'équation $y = x$, représentant la BTF unité), l'application de la BTF induit toutefois un gain dans les performances : le nAUC monte de 22.01% à 23.05%.

Il est à noter que dans la littérature [Porikli, 2003, Javed *et al.*, 2005, Gilbert et Bowden, 2006, Prosser *et al.*, 2008], la technique des BTF est toujours évaluée sur des données privées.

Pour chacune des méthodes listées ci-dessus, une fois la BTF calculée, le principe est de ramener les signatures colorimétriques issues de l'une des caméras dans la base de l'autre pour augmenter la pertinence de leur comparaison. Il s'agit là d'une méthode de correction des distributions couleurs calculées dans un capteur, par rapport à un autre. Ainsi, la couleur étant l'une des caractéristiques dominantes pour la REID en réseau de caméras [Gheissari *et al.*, 2006], la méthode

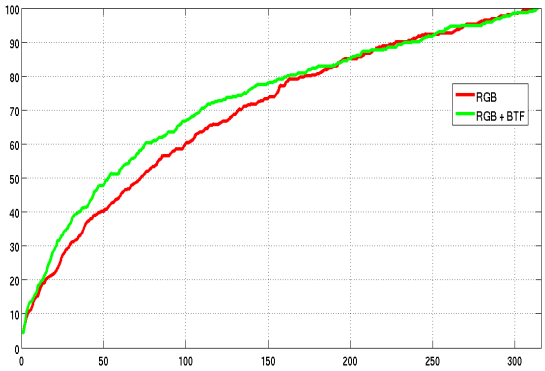


FIGURE 2.7 – Courbe CMC sur VIPeR avec un appariement d'histogrammes couleurs (courbe rouge, nAUC = 22.01%) et les mêmes histogrammes calculés sur les images » corrigées en couleur»(courbe verte, nAUC = 23.05%).

des BTF présente de nombreux avantages en terme d'augmentation des performances de reconnaissance. Nous avons montré ici que même sur des bases complexes comme VIPeR, la méthode des BTF induit un gain en faisant augmenter le nAUC de 22.01% à 23.05%.

Nous avons utilisé ici des paires d'images vérité terrain comme *a priori* pour réaliser l'apprentissage d'une fonction de transfert globale. Nous allons décrire une autre stratégie pour tirer parti de cette connaissance *a priori*.

2.4.2 Apprentissage statistique pour la ré-identification

Le second type d'apprentissage exploré pour résoudre le problème de la ré-identification est l'apprentissage statistique pour construire une signature invariante au changement de caméra, construite sur la base des invariances observées dans l'ensemble d'apprentissage. Un état de l'art complet sur la théorie de l'apprentissage statistique dépasse largement le cadre de ce manuscrit. Nous nous contentons de citer brièvement deux des grandes classes d'approches que sont Adaboost, proposé dans [Freund et Schapire, 1995] et les SVM (pour « Support Vector Machine ») proposés dans [Hearst *et al.*, 1998]. Dans le domaine de la Vision par Ordinateur, les SVMs ont été appliqués avec succès en combinaison aux histogrammes d'orientations des gradients HOG, dans le détecteur de piétons de Dalal et Triggs, dans [Dalal et Triggs, 2005], alors que Adaboost est connu en particulier pour son couplage aux ondelettes de Haar dans la détection de visages par Viola et Jones dans [Viola et Jones, 2001].

Nous citons ici des travaux majeurs qui exploitent Adaboost. Nous rappelons donc brièvement son principe général. Il existe beaucoup de variations (discrète,

réelle, *etc.*). Nous ne rentrons pas dans ces détails.

Adaboost : Adaboost est un algorithme de classification binaire entre deux classes, les positifs et les négatifs (*e.g.* les visages et les non-visages pour [Viola et Jones, 2001]). Il repose sur la sélection itérative de classifieurs faibles h_i^{weak} en fonction d'un ensemble d'apprentissage labellisé $\{(x_j, y_j)\}$ où x_j représente le j^{e} exemple de l'ensemble d'apprentissage et y_j son label (*i.e.* positif ou négatif). Un classifieur faible prédit le label d'un exemple et doit seulement avoir des performances légèrement meilleures que le hasard. Adaboost sélectionne parmi ces N classifieurs faibles ceux qui discriminent le mieux les exemples positifs des négatifs sur l'ensemble d'apprentissage et les pondère α_i en fonction. Ainsi il produit un classifieur fort :

$$h^{strong}(x) = \sum_{i=1}^{N} \alpha_i h_i^{weak}(x)$$

Une fois l'algorithme entrainé, il peut classifier des exemples x dont le label n'est pas connu.

En terme d'apprentissage statistique pour la ré-identification, Gray et Tao ont introduit l' « Ensemble of Localized Features » dans [Gray et Tao, 2008]. L'idée se résume à ne pas prédéfinir une signature d'apparence donnée, mais calculer un ensemble de caractéristiques et laisser un algorithme Adaboost, entrainé sur un ensemble d'apprentissage sélectionner les caractéristiques les plus invariantes. Dans ce cas particulier, les exemples positifs sont des paires de silhouettes correctement associées entre les deux caméras, et les négatifs, des paires mal associées. Nous détaillons ci-dessous les caractéristiques de description et le mode d'entrainement.

Ensemble de caractéristiques de description : Plus formellement, chaque silhouette de personne est découpée en cinq bandes horizontales non-recouvrantes. Sur chacune sont calculées des distributions couleurs, estimées par des histogrammes dans les espaces RGB, YCbCr et HS, ainsi que des informations de texture avec des ondelettes de Schmid et de Gabor. La figure 2.8 donne un exemple de cette localisation en « bandes » de l'information, avec le calcul des histogrammes RGB correspondants.

Un classifieur faible calcule sa caractéristique dédiée sur les deux silhouettes, et regarde si la réponse est similaire ou non, en prenant en compte le label de la paire d'images.

Entrainement du modèle : La méthode requiert de disposer d'un ensemble de silhouettes de personnes, vues dans les deux caméras, et labellisées, *i.e.* pour lesquelles le problème de la ré-identification a déjà été résolu. Étant donné l'optique système et ré-identification appliquée aux réseaux de caméra NOFOV, il

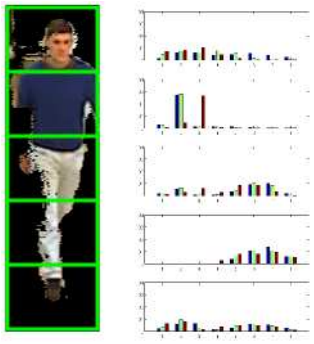


FIGURE 2.8 – Découpage de la silhouette en bandes non-recouvrantes et calcul de distributions couleurs sur ces bandes.

s'agit ici d'un point important et qui donnera lieu à des discussions dans la suite du manuscrit : pour être applicable, une telle méthode requiert un ensemble labellisé de personnes, vues dans les différentes caméras.

Ensuite, les caractéristiques locales sont utilisées dans un algorithme d'apprentissage de type Adaboost. Le principe est de rechercher parmi toutes les caractéristiques proposées, lesquelles sont partagées dans les paires de silhouettes labellisées positives tout en discriminant bien les paires labellisées négatives.

La table 2.1, représente l'importance accordée par le boosting suite à son entrainement, aux différentes caractéristiques calculées. Nous remarquons ici que les canaux H (teinte) et S (saturation) sont particulièrement privilégiés par Adaboost.

« Feature Channel »	« Percent of classifier weight »
R	11.0%
G	9.4%
B	12.4%
Y	6.4%
Cb	6.1%
Cr	4.5%
H	14.2%
S	12.5%
Schmid	12.0%
Gabor	11.7%

TABLE 2.1 – Répartition des caractéristiques choisies par le boosting sur la partie d'entrainement de la base VIPeR, selon [Gray et Tao, 2008].

Étant donné le succès rencontré par cette méthode sur la base VIPeR, ces investigations ont initié de nombreux travaux dans la littérature.

2.4.2.1 Travaux inspirés de [Gray et Tao, 2008]

Ce chapitre se voulant faire un état de l'art relativement exhaustif sur les méthodes de ré-identification, nous dressons dans cette sous-section un rapide panorama des évolutions proposées en terme de ré-identification entre une paire de caméra par apprentissage statistique.

RankSVM Dans [Prosser *et al.*, 2010], Prosser *et al.* ont appliqué l'algorithme du rankSVM [Joachims, 2002] au problème de la ré-identification. La formulation du problème est similaire à [Gray et Tao, 2008], avec le calcul des mêmes caractéristiques d'apparence, pour la même localisation dans des bandes sur la silhouette. Ici, les SVM viennent remplacer Adaboost pour l'algorithme d'apprentissage.

Comparaison de distances relatives probabilistes Par la suite, Zheng *et al.* ont formulé le problème dans [Zheng *et al.*, 2011] comme un apprentissage de distance plutôt que d'une pondération des caractéristiques. À ce titre, ils proposent la PRDC (pour « Probabilistic Relative Distance Comparison ») qui cherche à minimiser les distances entre les silhouettes de paires vérité terrain.

2.4.2.2 Étude comparative

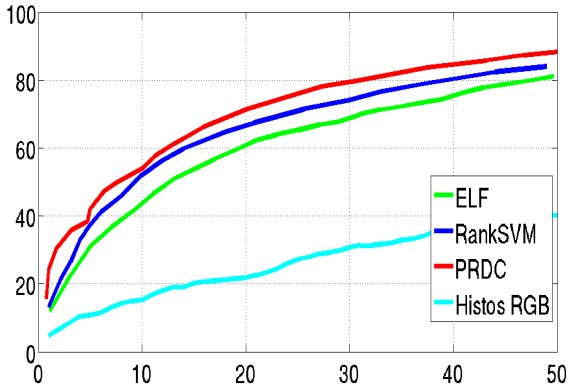


FIGURE 2.9 – Courbes CMC sur la base VIPeR avec 316 exemples d'entrainement et 316 de test, pour les approches ELF [Gray et Tao, 2008], RankSVM [Prosser *et al.*, 2010] et PRDC [Zheng *et al.*, 2011], comparées à une méthode d'appariement simple de distributions RGB.

La figure 2.9 présente les courbes CMC des principales méthodes d'apprentissage pré-citées, sur les base VIPeR, avec 316 exemples d'apprentissage. Au fur et à mesure, les techniques d'apprentissage statistiques dédiées à la ré-identificaiton se

sont affinées et c'est la PRDC qui présente les meilleures performances sur la base VIPeR. Par ailleurs, les performances sont vraiment supérieures à un appariement d'histogrammes RGB non supervisé.

2.4.3 Limitations des méthodes supervisées

Ces méthodes supervisées présentent cependant une limitation forte. En effet, qu'il s'agisse des BTF ou des apprentissages statistiques, ces méthodes reposent sur la mise en correspondance de silhouettes observées par différents capteurs. Dans le cadre d'un réseau à champs disjoints, cela suppose de disposer d'appariements de silhouettes de personnes, *i.e.* avoir déjà résolu le problème de la ré-identification sur un ensemble d'entrainement, qui plus est, relativement conséquent [Gilbert et Bowden, 2006] pour être pertinent.

Le second problème soulevé par ces méthodes est celui de la stabilité temporelle de la fonction calculée. En effet, le modèle appris est valable pour les exemples d'apprentissage considérés. En environnement non contrôlé, les conditions d'illumination sont amenées à changer de manière indépendante dans chaque caméra, la validité du modèle n'est pas assurée sur une longue période. Comme suggéré dans [Gilbert et Bowden, 2006], une mise à jour de cette BTF devient donc nécessaire.

Devant la difficulté de la construction de l'ensemble d'apprentissage pour un réseau tel que celui présenté en figure 2.10, nous rejetons pour le moment les méthodes d'apprentissage supervisé pour notre objectif de surveillance de réseaux de caméras, car non applicables directement. Nous montrerons dans la suite que le système que nous mettons en place pourra bénéficier des avantages de telles méthodes, une fois qu'il aura construit automatiquement ces ensembles d'apprentissage.

2.5 Méthodes non-supervisées pour la ré-identification dans un réseau

Par opposition aux approches reposant sur l'apprentissage, intrinsèquement dédiées à la paire de caméra considérée, certains travaux que nous listons ici présentent des approches directes. Les travaux [Madden *et al.*, 2007, Ilyas *et al.*, 2010] ont utilisé les distributions couleurs pour décrire directement l'apparence de personnes entre plusieurs caméras, alors que [Hamdoun *et al.*, 2008] ont utilisé des points d'intérêt. Bak *et al.* ont proposé dans [Bak *et al.*, 2010] de décrire la couleur et la texture de l'apparence de manière conjointe au sein de matrices de covariance. Achard *et al.* proposent une catégorisation exhaustive de toutes les variantes de descriptions proposées dans [Achard *et al.*, 2012].

Nous nous focalisons ici sur les travaux de Farenzena *et al.* dans [Farenzena

FIGURE 2.10 – Exemples de ré-identification telle qu'elle est envisagée dans notre réseau (le code couleur fournit les appariements corrects).

et al., 2010] qui proposent une signature « directe », au sens où elle ne requiert pas de phase d'entrainement, obtenant des performances similaires aux approches basées apprentissage décrites en section 2.4. Nous détaillons dans cette partie le principe de cette signature, appelée SDALF (pour « Symmetry Driven Accumulation of Local Features »), ainsi que la manière dont ces travaux ont influencé notre recherche.

2.5.1 Principe de l'accumulation de caractéristiques locales dirigée par les symétries

Cette signature se compose de trois signaux complémentaires décrivant l'apparence d'une personne, et calculés relativement à des symétries de la silhouette.

Ces axes permettent aux auteurs de définir une localisation des signaux calculés plus fine qu'un simple découpage en bandes non recouvrantes de la silhouette. Nous commençons par décrire l'obtention de ces symétries, en accord avec [Farenzena *et al.*, 2010], puis nous présentons les signaux et enfin la manière dont deux signatures sont comparées.

2.5.1.1 Axes de symétrie/asymétrie

Le calcul des symétries de la silhouette suppose d'avoir obtenu une segmentation fond/forme. L'article travaillant uniquement sur des bases d'images, la technique appliquée est le STEL modèle (pour « STructure ELement ») [Jojic *et al.*, 2009], une technique de segmentation non supervisée qui se base sur la recherche d' « éléments de structure ». Dans un cas de suivi temporel, le STEL modèle sera remplacé par l'approche de modélisation du fond par mélange de gaussiennes de Stauffer et Grimson [Stauffer et Grimson, 1999].

Farenzena *et al.* définissent deux opérateurs. L'opérateur chromatique bilatéral :

$$C(i,\delta) = \sum_{B[i-\delta,i+\delta]} d^2(p_i, \hat{p}_i)$$

où $d(.,.)$ est la distance euclidienne évaluée entre les valeurs HSV des pixels p_i et $\hat{p}_i$, situés symétriquement par rapport à l'axe horizontal, à la hauteur i. $B_{[i-\delta,i+\delta]}$ est la fenêtre glissante dans laquelle sont calculées les distances chromatiques entre les pixels. Le paramètre de largeur de fenêtre glissante δ est proportionnel à la largeur de la boite et fixé à $\delta = J/4$. Le deuxième est l'opérateur de couverture spatiale :

$$S(i,\delta) = \frac{1}{J\delta}|A(B_{[i-\delta,i]}) - A(B_{[i,i+\delta]})|$$

où $A(B_{[i-\delta,i]})$ représente le ratio de zone de premier plan (issue de la segmentation fond/forme) présent dans la boite de largeur J et de hauteur $[i-\delta, i]$.

La figure 2.11 détaille les différents axes de symétrie, ainsi que les zones dans lesquelles les distances entre les pixels sont calculées.

Les hauteurs des axes d'asymétries (séparations tronc/jambes *i.e.* « Torso/Legs » et tête/tronc *i.e.* « Head/Torso ») sont respectivement calculées à partir des opérateurs ci-dessus.

$$i_{TL} = \arg\min_i(1 - C(i,\delta) + S(i,\delta))$$

$$i_{HT} = \arg\min_i(-S(i,\delta))$$

Les axes horizontaux étant calculés, deux axes verticaux sont estimés dans les parties torse et jambes respectivement. Pour $k = 1, 2$, nous avons :

$$j_{LRk} = \arg\min_j(C(j,\delta) + S(j,\delta))$$

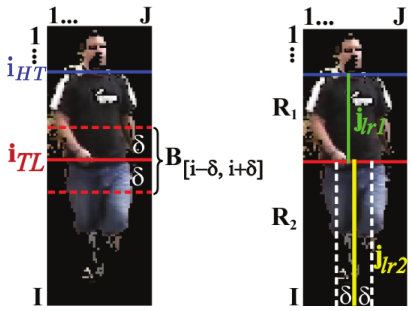


FIGURE 2.11 – Axes de symétries [Farenzena *et al.*, 2010].

En résumé, cette méthode se base sur le résultat de la segmentation fond/forme, ainsi que sur des comparaisons de pixels.

2.5.1.2 Extraction des descripteurs

Les signaux calculés dans les zones définies par ces axes sont au nombre de trois. Nous les détaillons ci-dessous.

Régions maximales de couleur stable MSCR SDALF extrait tout d'abord une information couleur moyenne, en recherchant les régions maximales de couleur stable. Il s'agit là des travaux de Forssén dans [Forssén, 2007] sur le MSCR (pour « Maximally Stable Color Region »). Ici, trois MSCR sont calculées sur le masque de forme, une pour la tête, une pour le torse et une pour les jambes. La méthode extrait des ellipses de couleur moyenne sur la silhouette. Ces ellipses sont définies par les triplets :

$$\{localisation, taille, couleur\ moyenne\}$$

FIGURE 2.12 – Exemple de blobs de couleurs moyennes détectés par le MSCR [Forssén, 2007] sur une silhouette.

Le MSCR détecte un ensemble de régions elliptiques dans l'image, comme illustré par la figure 2.12, en réalisant un «clustering » agglomératif sur les pixels

de l'image. Chaque étape assemble des pixels voisins possédant une couleur similaire, en seuillant selon une distance chromatique maximale entre les couleurs. Ces caractéristiques présentent de bonnes propriétés pour l'appariement : l'invariance à l'échelle et aux transformations affines dans les images couleurs.

Histogrammes couleurs pondérés SDALF extrait une information couleur plus détaillée avec le calcul d'histogrammes couleurs HSV (espace décorrélant chromaticité et luminance, notamment établi comme prépondérant par Adaboost 2.1, pour la méthode de Gray et Tao [Gray et Tao, 2008]), pondérés par la distance des pixels à l'axe de symétrie vertical et localisés dans les parties tronc et jambes, du masque de forme. Ceci rappelle les distributions localisées utilisée dans [Pérez *et al.*, 2002, Nummiaro *et al.*, 2003] pour du MOT. Ici la précision de la localisation est accrue par le calcul des symétries de la silhouette.

Patches de structure récurrente Finalement, SDALF extrait une information de texture, au travers de patches récurrents [Farenzena *et al.*, 2010]. Étant donnée la faible contribution de cette composante (voir plus loin la figure 2.16), nous ne détaillons pas son calcul ici.

2.5.1.3 Comparaison de signatures

La distance entre deux signatures SDALF I_A et I_B est calculée par l'équation (2.4). Il s'agit d'une somme pondérée entre les comparaisons des différentes composantes.

$$\begin{aligned} d_{SDALF}(I_A, I_B) = & \beta_{WH} \cdot d_{WH}(WH(I_A), WH(I_B)) + \\ & \beta_{MSCR} \cdot d_{MSCR}(MSCR(I_A), MSCR(I_B)) + \\ & \beta_{RHSP} \cdot d_{RHSP}(RHSP(I_A), RHSP(I_B)) \end{aligned} \tag{2.4}$$

où d_{WH} est la distance de Bhattacharyya [Bhattacharyya, 1943] entre les histogrammes des parties, ainsi d_{RHSP} est la distance calculée entre les patchs de texture.

La comparaison entre les ellipses des MSCR est réalisée comme suit :

$$d_{MSCR} = \sum_{b \in I_B} \min_{a \in I_A} \left(\gamma \cdot d_y^{ab} + (1 - \gamma) \cdot d_c^{ab} \right) \tag{2.5}$$

où d_y^{ab} est la distance euclidienne entre les ordonnées des centres des ellipses a et b, et d_c^{ab} la distance euclidienne entre les couleurs des ces deux mêmes ellipses. Le paramètre γ règle l'importance de la localisation verticale par rapport à la couleur, dans la comparaison de MSCR. Les poids sont fixés empiriquement à $\beta_{WH} = 0.4$, $\beta_{MSCR} = 0.4$, $\beta_{RHSP} = 0.2$ et $\gamma = 0.4$, selon [Farenzena *et al.*, 2010].

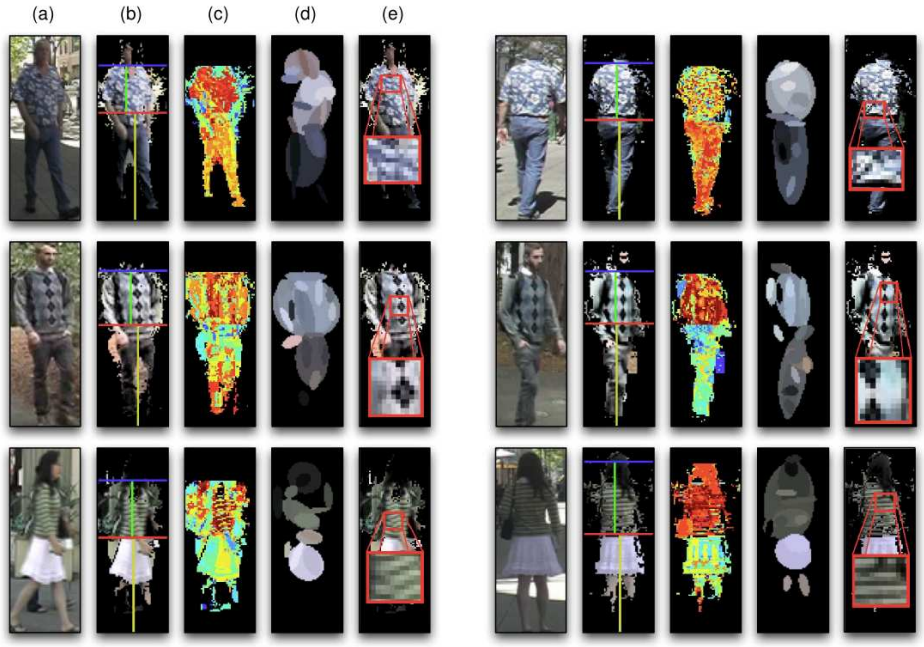


FIGURE 2.13 – Illustration du descripteur SDALF. (a) Étant donnée une image, (b) SDALF localise les parties intéressantes du corps de la personne : tête, tronc, jambes, ainsi que les axes de symétrie de ces différentes parties. Ensuite les aspects complémentaires de l'apparence sont extraits : (c) histogrammes HSV pondérés par leur distance à l'axe de symétrie, (d) Maximally Stable Color Region (descripteur proposé par [Forssén, 2007]) et (e) des patchs de texture récurrente. Le but est de correctement apparier les descripteurs de la même persone (première et sixième colonne). Illustration issue de [Bazzani, 2012].

2.5.2 Extensions directes de l'approche

Souhaitant être exhaustif dans cet état de l'art, nous citons dans la suite des travaux récents sur le problème de la ré-identification dans des bases d'images. Il s'agit dans les deux cas d'une extension de l'approche SDALF.

SDALF par parties Les récents progrès des détecteurs de personnes avec le passage aux détecteurs par parties, tels que le « Deformable Part Model » de Felszenswalb *et al.* [Felzenszwalb *et al.*, 2010] ou le « Pictorial Structure » [Felzenszwalb et Huttenlocher, 2005] permettent une localisation relativement précise des membres d'une personne. L'intérêt pour le problème de ré-identification abordé au niveau description est immédiat : ceci permet de localiser et donc comparer des sous-parties mobiles (pour les membres) de la silhouette. Les performances de la comparaison sont accrues. À ce sujet, Cheng *et al.* on proposé dans [Cheng *et al.*,

2011] une variation du « Pictorial Structure », pour prendre en compte le calcul d'un SDALF dédié détecteur par partie. Ceci résulte en des performances accrues, comme le témoigne la figure 2.14.

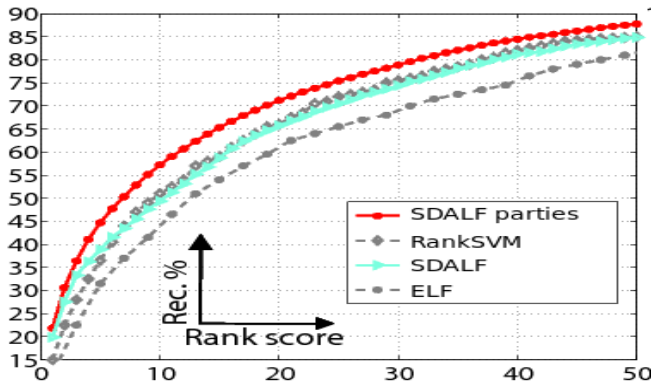


FIGURE 2.14 – Performance du SDALF par parties comparé à SDALF original (nAUC = 92.24%), rankSVM (nAUC = 92.36%) et ELF (nAUC = 90.85%), sur la base VIPeR, issu de [Cheng *et al.*, 2011].

Il est clair qu'une méthode de ré-identification doit s'appuyer sur une localisation aussi précise que possible de l'information et ces travaux font partie des plus récents en terme de ré-identification dans des bases d'images. Toutefois, étant donné notre contexte applicatif de surveillance de réseau et les temps de calculs actuels d'une telle méthode de détection, nous écartons cette approche pour notre problème.

Unification de méthodes supervisées et de méthodes non-supervisées Une seconde extension directe réside dans les travaux de Layne *et al.* . Dans [Layne *et al.*, 2012], ils proposent de compléter la signature SDALF par des notions d'apparence plus haut niveau. Il entraine un classifieur pour reconnaître des éléments de l'apparence tels qu'un sac à dos, une cravate. Cet article propose d'incrémenter SDALF par un apprentissage supervisé. Ici encore, ceci apporte un gain sur les performances.

2.6 Choix de notre représentation

Nous terminons ce chapitre état de l'art par une discussion sur la meilleure représentation à adopter pour notre contexte de surveillance de réseaux de caméras. Le tableau 2.2 vient synthétiser les différents approches présentées dans ce chapitre.

Il apparaît clairement qu'une méthode supervisée ne sera pas applicable directement pour nous, en raison des contraintes d'entrainement.

	Travaux	Temps de calcul	Description	Mise à l'échelle	Performance
non supervisé	SDALF [Farenzena *et al.*, 2010]	+	symétries couleur/texture	++	++
	SDALF par partie [Cheng *et al.*, 2011]	- -	parties couleur/texture	++	+++
supervisé	ELF [Gray et Tao, 2008]	- -	bandes couleur/texture	- -	+
	RankSVM [Prosser *et al.*, 2010]	- -	bandes couleur/texture	- -	++
	PRDC [Zheng *et al.*, 2011]	- -	bandes couleur/texture	- -	+++
	SDALF supervisé [Layne *et al.*, 2012]	- -	bandes couleur/texture	-	+++

TABLE 2.2 – Tableau récapitulatif des différents modèles pour la ré-identification.

Nous comparons maintenant les performances de SDALF à celles des méthodes supervisées. La figure 2.15 présente les résultats de SDALF sur la base VIPeR, testée sur 316 exemples, par rappport à la méthode supervisée initiale ELF [Gray et Tao, 2008], entrainée sur 316 exemples et testée sur les 316 autres. Nous constatons ici que malgré l'absence d'*a priori* pour SDALF, ses performances sont supérieures à celle d'une méthode supervisée. Ceci valide le choix de SDALF comme modèle de représentation : il est moins contraint qu'une méthode supervisée, tout en atteignant de bonnes performances de ré-identification.

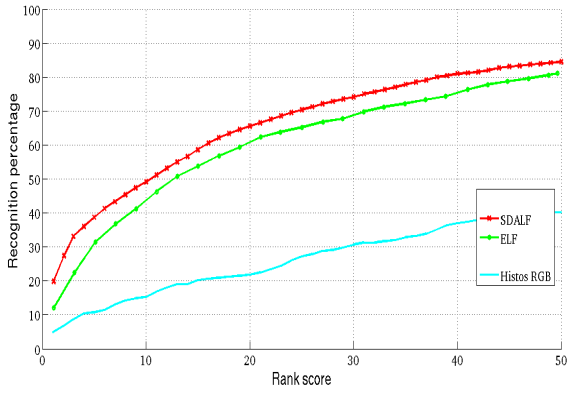


FIGURE 2.15 – Courbes CMC sur la base VIPeR comparant les performances de SDALF (non-supervisée) avec ELF (supervisée), et avec des bandes d'histogrammes RGB.

2.6.1 Influence des composantes de SDALF

Nous testons maintenant l'importance des différentes composantes de SDALF. Pour cela, nous lançons les tests sur la base VIPeR, en proposant différentes pondérations dans l'importance des composantes. La figure 2.16 présente les courbes CMC de SDALF [3] sur la base VIPeR avec différentes pondérations. Nous montrons ici que les histogrammes couleurs pondérés et le MSCR obtiennent déjà des résultats satisfaisants. Par ailleurs, nous montrons ici l'influence moindre de la composante de texture (sur la base particulière VIPeR, ne pas utiliser la composante texture résulte même en un score nAUC légèrement supérieur).

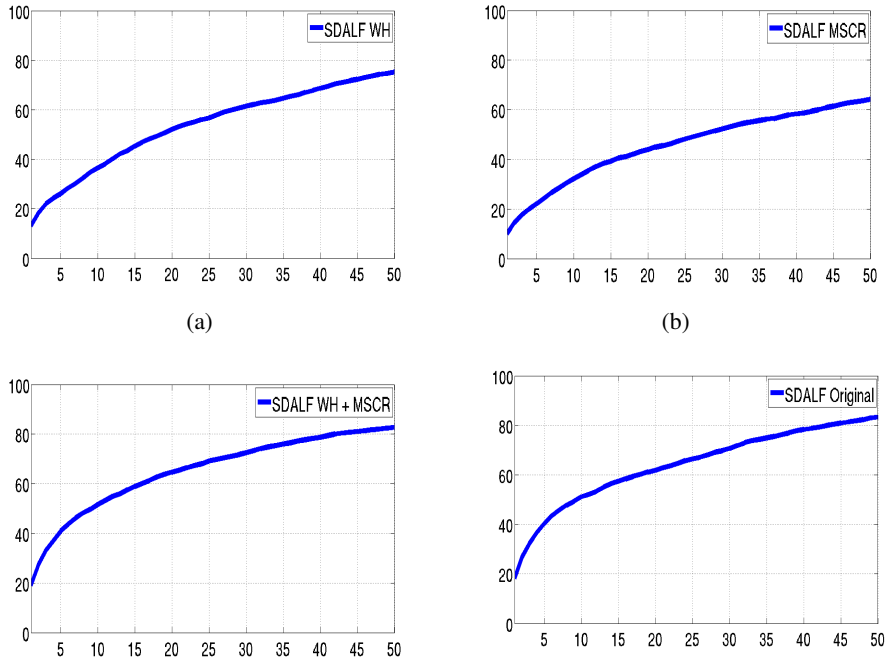


FIGURE 2.16 – Différentes pondérations des composantes de SDALF : (a) uniquement WHSV (nAUC = 89.11%), (b) uniquement MSCR (nAUC = 83.56%), (c) WHSV et MSCR avec importance égale (nAUC = 91.81%) et (d) comme dans [Farenzena *et al.*, 2010], la pondération originale de SDALF $\beta_{WH} = 0.4$ $\beta_{MSCR} = 0.4$ $\beta_{RHSP} = 0.2$ (nAUC = 91.57%).

3. codes disponibles sur http://www.lorisbazzani.info

2.6.2 Perspectives pour notre contexte de surveillance de réseaux de caméras

Dans le chapitre 3, nous allons considérer le problème du MOT. Souhaitant fusionner MOT et REID, nous avons besoin d'une signature efficace pour la ré-identification. Le tableau 2.2 reprend les constats de ce chapitre sur les différentes approches de l'état de l'art. La signature que nous choisissons ne peut être supervisée. Au regard des résultats de la figure 2.16, et du coût en temps de calcul du MSCR, nous choisissons une signature de type SDALF dégradée avec :

$$\beta_{WH} = 1,\ \beta_{MSCR} = 0.0,\ \beta_{RHSP} = 0.0.$$

2.7 Conclusion

Ce chapitre a réalisé un état de l'art sur les descripteurs pour la ré-identification. Rapidement, une partie de la communauté s'est focalisée sur le problème particulier de la ré-identification au sein d'une paire de caméras. Il est en effet alors possible de dédier la méthode aux deux caméras considérées à l'aide d'un ensemble de paires d'images d'entrainement, ce qui se traduit par des performances accrues. Toutefois au niveau d'un réseau de caméras, la combinatoire d'association des caméras rend nécessaire l'automatisation d'un tel entrainement. Quelques approches récentes [Kuo *et al.*, 2010a, Matei *et al.*, 2011] proposent des systèmes spécifiques adressant en partie l'automatisation de la collection d'exemples d'entrainement. Nous discutons ces solutions au chapitre 5.

Nous avons mené ici une étude bibliographique assez exhaustive, analysé les performances relatives des descripteurs existants, complété par nos propres évaluations pour identifier le descripteur le plus opportun dans notre contexte applicatif. La stratégie retenue, SDALF, est non supervisée, se base sur le calcul d'axes de symétrie et localise ainsi des informations de texture et de couleur. Par ailleurs, le chapitre 5 se basera sur l'étude menée ici, et notamment celle des descripteurs supervisés, pour proposer des extensions au système que nous développons ici.

Alors que les approches existantes négligent les temps de calculs, nous visons a intégrer ce descripteur dans un traqueur donc les contraintes temporelles sont importantes. Nous avons sélectionné la partie principale du descripteur, réalisant les meilleures performances tout en restant acceptable en termes de temps de calcul. Comme évalué en section 2.6, nous utilisons uniquement les composantes réellement significatives de la signature.

Se basant sur cette signature, le chapitre suivant propose une stratégie de suivi MOT et de ré-identification conjointe, relativement à une base de personnes connues comme étant présentes dans le réseau (la galerie). L'objectif ici est d'intégrer des contraintes spatio-temporelles dans le processus de ré-identification.

CHAPITRE

3

Estimation bayésienne de suivi et ré-identification dans une caméra

Dans ce chapitre 3, nous nous intéressons à une stratégie applicable pour la ré-identification de personnes, par une caméra d'ambiance fixe, dans un contexte de vidéosurveillance de bâtiment. Pour cela nous posons la ré-identification comme un problème d'estimation d'un état discret. Nous présentons le filtrage particulaire à état mixte et nous montrons comment le dériver pour répondre à ce problème. Nous faisons ici l'hypothèse de disposer pour toutes les personnes présentes dans l'espace surveillé par le réseau, d'images clés construites dans une caméra (typiquement surveillant un hall d'entrée de bâtiment). Nous focalisons sur comment les ré-identifier après une perte d'observabilité via une autre caméra du réseau. De là, nous proposons une méthode de suivi et ré-identification (par rapport à une base) conjoints. Dans la suite, nous généralisons cette méthode au cas multi-cibles. Il en résulte une méthode automatique de suivi et ré-identification conjoints de cibles multiples par vision monoculaire.

3.1 Introduction

Dans la vidéosurveillance intelligente, le suivi de personnes est une fonctionnalité essentielle. En effet, tout mécanisme haut niveau comme l'analyse d'activité ou la compréhension de scènes dynamiques, nécessite une localisation précise de ces personnes au cours du temps. Ce chapitre se focalise sur l'étape de suivi mono-caméra, mono- et multi-cibles. L'originalité de notre approche, présentée dans ce chapitre, est l'insertion d'un processus de ré-identification au sein de l'algorithme de suivi. L'objectif est d'associer une identité (provenant du réseau de caméras) aux cibles de suivi. Ainsi, une personne quittant la scène puis y revenant engendre classiquement deux cibles distinctes de suivi ; dans notre cas, elle se verra affecter la même identité.

Le suivi mono-caméra d'objets d'intérêt en mouvement est une problématique essentielle de la vidéosurveillance intelligente. L'objectif se décline comme la capacité à retourner à chaque instant-image la position du/des objets établis comme étant à suivre. Nous supposons ici l'initialisation du processus fournie par un algorithme de détection. Contrairement au problème de détection, le suivi est un processus spatio-temporel. Il faut parvenir à construire la trajectoire de la cible soit *a posteriori* à partir de la séquence d'images soit au fur et à mesure que les images arrivent, dans un contexte temps réel.

Dans un premier temps, il faut être en mesure de décrire l'apparence de l'objet d'intérêt que l'on souhaite suivre dans le reste de la séquence vidéo. Cette description doit être discriminante par rapport à la scène, ainsi que par rapport aux éventuels autres objets mobiles présents, tout en tolérant les variations que subira l'apparence de l'objet au cours du temps (i.e. dans le cas du suivi de piéton, l'objet d'intérêt est déformable et la variation d'apparence est forte). Le chapitre 2 a énuméré les descripteurs usités dans le contexte applicatif de la ré-identification de personnes. À la lumière de cet état de l'art, notre choix s'est porté sur une version restreinte de la signature SDALF, présentée en section 2.6.

Disposant d'un modèle d'apparence, il faut être en mesure de rechercher dans les images suivantes la/les zones « ressemblant » le plus à l'objet initial. À ce niveau, une calibration du système peut permettre d'effectuer la recherche dans le plan du sol et de contraindre de ce fait sa taille dans l'image. Ainsi, l'espace de recherche peut être le plan image tout comme le monde 3D (l'observation provenant toutefois toujours de l'image). L'espace étant défini, se pose maintenant la question du choix de la stratégie de suivi. La section 3.2 passe en revue différentes méthodes de l'état de l'art déterministe et non déterministe, se basant sur un calcul de similarité défini par le modèle d'apparence adopté. Cette estimée peut par ailleurs donner lieu à une mise à jour en ligne du modèle d'apparence. À ce niveau là, nous disposons d'un suivi mono-cible fonctionnel. Son extension au suivi multi-objets induit des problèmes d'échanges de cibles et nécessite une gestion de

cette combinatoire. Elle passe généralement par l'ajout d'un détecteur alimentant l'algorithme de suivi pour les décisions de créations et suppressions de cibles.

Les défis d'un algorithme de suivi MOT s'énoncent comme :
- ▷ initialisation automatique de cibles multiples ;
- ▷ gestion correcte des croisements / problème d'association de données ;
- ▷ mise à jour du modèle de représentation ;
- ▷ terminaison automatique des cibles.

Dans le chapitre précédent, nous avons abordé le problème de la représentation d'une identité à des fins de ré-identification. Les approches classiques de la ré-identification tendent à proposer un modèle de représentation de piéton qui soit le plus invariant possible, qu'il soit préalablement appris ou non (voir chapitre 2). Quel que soit le mode de représentation, la ré-identification donne lieu à des comparaisons de descripteurs.

Nous présentons ici une approche de la ré-identification de personne basée filtrage. Nous posons le problème comme une généralisation du suivi visuel. En effet, le suivi visuel markovien peut se définir comme la recherche récursive dans le flux vidéo d'une région d'intérêt, en se basant sur une mesure de similarité et des hypothèses de continuité spatio-temporelle. Les problèmes de gestion de cibles multiples et de mise à jour du modèle de représentation mis de côté, nous avons un problème similaire à celui de la ré-identification. La ré-identification telle que présentée au chapitre 2 se base uniquement sur des comparaisons de descripteurs. Nous proposons ici de réaliser les comparaisons de descripteurs de la REID au sein des filtres prenant en charge le suivi MOT. En ce sens, nous envisageons la REID dans un réseau de caméras comme une généralisation du MOT.

La section 3.2 présente un rapide état de l'art sur le suivi d'objets en contexte mono-caméra afin de positionner notre approche. De celui-ci, nous dérivons le principe de notre approche de suivi et ré-identification. Nous présentons ensuite en section 3.3 l'adaptation du filtrage particulaire au suivi visuel, et sa variante estimant un vecteur d'état mixte en section 3.4. Nous proposons alors en section 3.5 une extension au suivi multi-cibles. Nous généralisons cette approche avec un formalisme de filtrage particulaire à état mixte pour la ré-identification en section 3.6. Finalement, nous présentons les détails d'implémentation de notre approche en section 3.7, et l'évaluation de la méthode proposée en section 3.8. La section 3.9 résume les contributions et conclut ce chapitre.

3.2 État de l'art

Suivi mono-cible Un état de l'art sur le suivi d'objet en général dépasse largement les objectifs de ce chapitre. [Yilmaz *et al.*, 2006] et [Maggio *et al.*, 2012]

proposent un état de l'art détaillé dans ce domaine. Pour notre part, nous nous contentons de présenter ici les quelques papiers dont découlent, plus ou moins directement, l'approche de suivi et ré-identification que nous proposons.

L'algorithme de filtrage particulaire SIR (pour « Sampling Importance Resampling ») est issu des travaux de Gordon *et al.* dans [Gordon *et al.*, 1993]. Son introduction dans la communauté Vision est due à Isard et Blake dans [Isard et Blake, 1996].

Les travaux initiaux relèvent de la logique séquentielle avec des modèles markoviens d'ordre 1. L'intérêt de ces algorithmes, *e.g.* de filtrage particulaire (CONDENSATION), a été établi par les travaux fondateurs de Isard et Blake dans [Isard et Blake, 1998b], qui ont adapté le SIR au suivi visuel, avec une association de contours en tant que fonction de vraisemblance. La section 3.3 détaille le processus de filtrage de la CONDENSATION.

À modèle d'apparence égal, Pérez *et al.* ont montré dans [Pérez *et al.*, 2002] l'intérêt des méthodes stochastiques de type filtrage particulaire face aux méthodes déterministes telles que le *mean-shift tracking* [Comaniciu *et al.*, 2000] pour le suivi visuel.

Suivi multi-cibles Rapidement, la communauté s'est intéressée au cas des cibles multiples, notamment dans [Isard et Blake, 2001]. À ce niveau, les algorithmes se rapprochent du système automatique que nous visons : l'objectif est toujours de suivre, mais en ajoutant de nouvelles capacités telles que le lancement automatique des traqueurs, la gestion des situations de croisements et d'occultations...

Ensuite Okuma *et al.* [Okuma *et al.*, 2004] combinent l'algorithme de [Vermaak *et al.*, 2003] et un détecteur de piéton basé boosting pour donner la première approche de *suivi-par-détection*.

Contrairement à une stratégie de filtres distribués, [Khan *et al.*, 2005] et [Smith *et al.*, 2005] proposent un échantillonnage MCMC plus efficace pour les vecteurs d'états agrégeant plusieurs cibles, éventuellement à taille variable pour [Smith *et al.*, 2005]. Par opposition aux approches distribuées, nous parlerons de méthodes centralisées : un seul vecteur d'état est estimé. L'algorithme se base sur les résultats d'un détecteur pour estimer à chaque instant le nombre de personnes présentes dans la scène, et les suivre.

Parmi les méthodes classiques de l'état de l'art en terme de suivi multi-cibles mono-caméra, Breitenstein *et al.* proposent dans [Breitenstein *et al.*, 2010] une méthode markovienne basée sur des détections HOG [Dalal et Triggs, 2005]. L'implémentation GPU du détecteur [Prisacariu et Reid, 2009] rend la méthode temps-réel.

Détecteur	Distribué	Centralisé
SDF		[Khan *et al.*, 2005] [Smith *et al.*, 2005]
HOG	[Breitenstein *et al.*, 2010]	[Okuma *et al.*, 2004]

TABLE 3.1 – Tableau récapitulatif des différents systèmes de MOT markoviens présentés.

Suivi et identification Le sujet du suivi et de l'identification conjoints dans des réseaux de caméras à champs de vue joints *e.g.* [Qu *et al.*, 2007] est très similaire au MOT mono-caméra : la mise en relation des différents flux vidéos des capteurs vient classiquement avec l'utilisation de la calibration du système, permettant de travailler dans un repère 3D commun. Ainsi l'identification des cibles d'une caméra aux autres se base uniquement sur un critère géométrique.

Dès lors que le suivi présente des discontinuités d'observation, *e.g.* lorsque la personne quitte le champ de vue de la caméra puis y réapparait, une technique de ré-identification est nécessaire. Germa *et al. e.g.* proposent dans [Germa *et al.*, 2010] une fusion de données entre plusieurs capteurs pour un suivi avec identification des cibles à partir d'un robot guide. Le robot peut perdre les cibles de son champ de vue, mais des puces RFID, ainsi que de la reconnaissance faciale lui permet de ré-identifier ses « interlocuteurs ».

Par essence, un suivi multi-cibles induit une notion d'identité. Un traqueur est dédié à une cible. On dénote ici par le terme cible, une personne physique lors de son temps d'observation par la caméra concernée. Pour un algorithme de suivi classique, la réapparition de cette même personne physique suite à une sortie du champ de vue de la caméra, donnera lieu à la création d'une nouvelle cible. Nous sommes ici confronté au problème de la ré-identification : être capable de dire qu'il s'agit de la même personne que précédemment, malgré une période de non-observabilité (suffisamment longue pour induire la terminaison du suivi temporel).

Le tableau 4.1 récapitule les systèmes MOT que nous avons présentés. La distinction est faite entre les approches centralisées et les approches distribuées. Étant donné que nous visons une extension au réseau dans le chapitre 4, notre méthode de suivi doit être modulaire. Nous privilégions donc une approche distribuée avec des filtres indépendants. Nous verrons au chapitre suivant que ce choix nous permet d'introduire des interactions au niveau du réseau entre ces filtres. Nous avons discuté au chapitre 2 du choix de la signature pour le suivi. Nous utilisons le SDALF simplifié que nous avons présenté. Ainsi, notre approche, inspirée de [Breitenstein *et al.*, 2009], repose sur des filtres particulaires distribués mais ajoute la composante ré-identification via une variable discrète relative à l'identifiant de la cible dans une base d'identités. On parle alors de filtrage particulaire à état mixte.

3.3 Filtrage bayésien récursif

3.3.1 Formalisation du problème

Considérons l'étude d'un système dynamique évoluant temporellement. À l'instant t, l'état de ce système est noté $\mathbf{x}_t$. Ce système est complètement défini par sa densité de probabilité initiale $p(\mathbf{x}_0|\mathbf{Z}_0) = p(\mathbf{x}_0)$, aussi appelé *a priori*, et par deux équations : un modèle d'évolution (3.1) et un modèle d'observation (3.2).

$$\begin{cases} \mathbf{x}_t = \mathbf{f}_t(\mathbf{x}_{t-1}, \mathbf{v}_{t-1}) & (3.1) \\ \mathbf{Z}_t = \mathbf{h}_t(\mathbf{x}_t, \mathbf{n}_t) & (3.2) \end{cases}$$

où $\mathbf{v}_t$ et $\mathbf{n}_t$ sont des bruits respectivement de prédiction et de mesure, et $\mathbf{f}_t$ et $\mathbf{h}_t$ sont les fonctions respectivement d'évolution et de mesure, tous ces termes étant exprimés à l'instant t. Le problème se pose comme l'estimation au cours du temps de la suite d'états du système $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$.

En particulier, nous recherchons une suite d'estimées de $\mathbf{x}_t$, à partir de l'ensemble de mesures disponibles $\mathbf{Z}_{1:t} = \{\mathbf{Z}_i, i = 1, \ldots, t\}$ jusqu'au temps t.

D'un point de vue bayésien, le problème du suivi se traduit par le calcul récursif de la densité de probabilité de l'état $\mathbf{x}_k$ au temps t, à partir de $\mathbf{Z}_{1:t}$. Ainsi, nous souhaitons obtenir la densité de probabilité $p(\mathbf{x}_t|\mathbf{Z}_{1:t})$. Le processus est supposé initialisé par l'a priori $p(\mathbf{x}_0|\mathbf{Z}_0) = p(\mathbf{x}_0)$. À partir de là, la densité $p(\mathbf{x}_t|\mathbf{Z}_{1:t})$ peut être obtenue récursivement, en deux étapes : prédiction (3.1), puis mise à jour selon la mesure (3.2).

Comme (3.2) décrit un processus de Markov, on déduit la simplification $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{Z}_{1:t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})$. Par ailleurs, l'évolution probabilistique de l'état $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ est donnée par (3.2) et la connaissance du bruit $\mathbf{v}_{t-1}$. À partir de là, et supposant disposer de la densité $p(\mathbf{x}_{t-1}|\mathbf{Z}_{1:t-1})$ au temps $t-1$, l'équation de Chapman-Kolmogorov permet d'obtenir la densité *a priori*, prédite pour le temps t :

$$p(\mathbf{x}_t|\mathbf{Z}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{1:t-1}) p(\mathbf{x}_{t-1}|\mathbf{Z}_{1:t-1}) d\mathbf{x}_{t-1}. \quad (3.3)$$

Au temps t, après la phase de prédiction, la mesure $\mathbf{Z}_t$ devient disponible et peut être utilisée pour mettre à jour la prédiction grace au théorème de Bayes. Les observations sont supposées indépendantes conditionnellement au processus d'état, et leur distribution ne dépend que de l'état au même instant, ce qui se traduit par :

$$p(\mathbf{x}_{0:t}|\mathbf{Z}_{1:t}) = \frac{p(\mathbf{Z}_t|\mathbf{x}_{0:t}, \mathbf{Z}_{1:t-1}) p(\mathbf{x}_{0:t}|\mathbf{Z}_{1:t-1})}{p(\mathbf{Z}_t|\mathbf{Z}_{1:t-1})} \quad (3.4)$$

, où la constante de normalisation est :

$$p(\mathbf{Z}_t|\mathbf{Z}_{1:t-1}) = \int p(\mathbf{Z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{Z}_{1:t-1})d\mathbf{x}_t. \tag{3.5}$$

Cette constante de normalisation dépend de la fonction de vraisemblance $p(\mathbf{Z}_t|\mathbf{x}_t)$ définie par le modèle de mesure (3.2) et la connaissance de la statistique de $\mathbf{n}_t$. Dans l'étape de mise à jour (3.4), la mesure $\mathbf{Z}_t$ est utilisée pour modifier la densité prédite et obtenir la densité *a posteriori* pour l'état courant.

Les relations de récurrence (3.3) et (3.4) forment la base de la solution bayésienne optimale, au sens où elles propagent récursivement une densité *a posteriori* de l'état de manière exacte. Cela demeure cependant une formulation théorique car en pratique, il est rare de pouvoir déterminer analytiquement une telle densité. Le filtre de Kalman, et ses hypothèses très contraignantes de linéarité et de gaussiannité des modèles d'évolution et de mesure est un exemple de résolution exacte (dans le cas où le système accepte un tel modèle).

3.3.2 Approximation particulaire

Le filtrage particulaire a été utilisé dans de nombreux domaines, allant de la mécanique des fluides à l'analyse statistique de données, en passant par l'économétrie. Ici, nous nous contentons simplement de présenter une trame grossière de son cheminement vers ses applications en Vision par ordinateur et notamment en suivi visuel, en accord avec la présentation faite par MacCormick [MacCormick, 2000].

Les techniques de filtrage bayésien séquentielles basées sur un échantillonnage aléatoire, sont en partie apparues avec les travaux en automatique de Handschin et Mayne [Handschin et Mayne, 1969, Handschin, 1970]. Toutefois, leur formulation ne faisait pas intervenir d'étape de ré-échantillonnage, donnant ainsi une dégénérescence de l'algorithme. Le filtre à particules Sampling Importance Resampling (abrégé SIR), dû à Rubin [Rubin *et al.*, 1988] introduit une étape de ré-échantillonnage des particules.

Pour modéliser des distributions non gaussiennes, Gordon *et al.* introduisent dans [Gordon *et al.*, 1993] le formalisme de filtrage particulaire. Les techniques de filtrage particulaire sont des méthodes de simulation séquentielles de type Monte Carlo permettant l'estimation du vecteur d'état d'un système Markovien non linéaire soumis à des excitations aléatoires possiblement non Gaussiennes [Doucet *et al.*, 2001, Arulampalam *et al.*, 2002]. En tant qu'estimateurs bayésiens, leur but est d'estimer récursivement la densité de probabilité *a posteriori* $p(\mathbf{x}_t|\mathbf{Z}_{1:t})$ du vecteur d'état $\mathbf{x}_t$ à l'instant t conditionné sur l'ensemble des mesures $\mathbf{Z}_{1:t} = \mathbf{Z}_1, ..., \mathbf{Z}_t$, une connaissance *a priori* de la distribution du vecteur d'état initial $\mathbf{x}_0$ pouvant

être également prise en compte. À chaque instant image t, la densité $p(\mathbf{x}_t|\mathbf{Z}_{1:t})$ est approximée au moyen de la distribution ponctuelle :

$$p(\mathbf{x}_t|\mathbf{Z}_{1:t}) \approx \sum_{i=1}^{N} \pi_t^{(i)} \delta(\mathbf{x}_t - \mathbf{x}_t^{(i)}), \text{ avec } \sum_{i=1}^{N} \pi_t^{(i)} = 1, \tag{3.6}$$

exprimant la sélection d'une valeur -ou particule- $\mathbf{x}_t$ avec la probabilité -ou poids- π_t où i = 1, . . . , N est l'index de la particule. Les moments conditionnels de $\mathbf{x}_t$, tels que l'estimateur du minimum d'erreur quadratique moyenne (ou MMSE, pour « Minimum Mean Square Error ») $E[\mathbf{x}_t|\mathbf{Z}_{1:t}] = \sum_{i=1}^{N} \pi_t^{(i)} \mathbf{x}_t^{(i)}$, peuvent alors être approchés par ceux de la variable aléatoire ponctuelle de densité de probabilité (3.6). Ainsi, nos différents filtres sont basés sur cet estimateur MMSE. Les particules $\mathbf{x}_t^{(i)}$ évoluent stochastiquement dans le temps. Elles sont échantillonnées selon une fonction d'importance visant a explorer adaptativement les zones « pertinentes » de l'espace d'état.

L'algorithme 2 donne les équations du filtrage SIR générique. Il est à noter que cet algorithme fait intervenir une fonction de proposition q lors de la propagation des particules (équation (3.7)). Dans le cadre de la CONDENSATION, la fonction de proposition utilisée est la dynamique du système :

$$q(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)}) = p(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})$$

Ceci induit donc une simplification de l'équation (3.8) de mise à jour des poids des particules qui devient :

$$\pi_t^{(i)} \propto p(\mathbf{Z}_t|\mathbf{x}_t^{(i)})$$

La figure 3.1 illustre son application au suivi visuel : la CONDENSATION, initié par Isard et Blake dans [Isard et Blake, 1996].

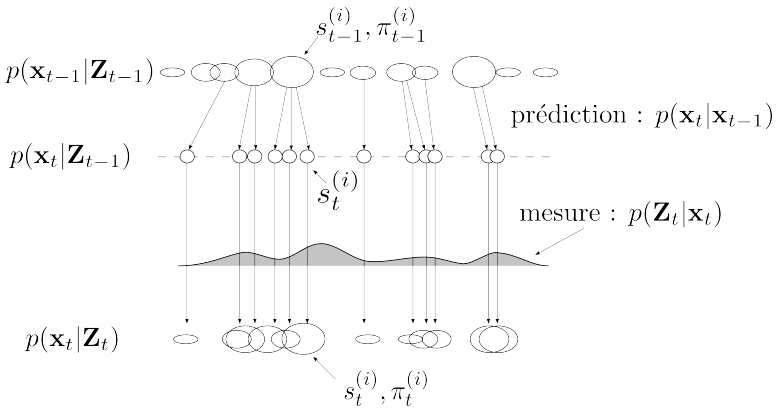


FIGURE 3.1 – Principe de la CONDENSATION, telle que présentée dans [Isard et Blake, 1998b].

Algorithme 2: Algorithme générique de filtrage particulaire (SIR).

Données : Mesure à t : $\mathbf{Z}_t$,
ensemble de particules pondérées à $t-1$: $\{\mathbf{x}_{t-1}^{(i)}, \pi_{t-1}^{(i)}\}_{i=1\ldots N}$.

Résultat : Ensemble de particules pondérées à t : $\{\mathbf{x}_t^{(i)}, \pi_t^{(i)}\}_{i=1\ldots N}$.

si *t=0 (INITIALISATION)* **alors**

Échantillonner $\mathbf{x}_0^{(1)}, \ldots, \mathbf{x}_0^{(i)}, \ldots, \mathbf{x}_0^{(N)}$ i.i.d. selon $p(\mathbf{x}_0)$, et poser $\pi_0^{(i)} = \frac{1}{N}$

fin

si $t \geq 1$ **alors**

pour $i = 1, \ldots, N$ **faire**

propager la particule $\mathbf{x}_{t-1}^{(i)}$ en simulant de manière indépendante

$$\mathbf{x}_t^{(i)} \sim q(\mathbf{x}_t | \mathbf{x}_{t-1}^{(i)}, \mathbf{Z}_t) \tag{3.7}$$

Mettre à jour les poids $\pi_t^{(i)}$ selon l'équation

$$\pi_t^{(i)} \propto \frac{p(\mathbf{Z}_t|\mathbf{x}_t^{(i)}) p(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})}{q(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)}, \mathbf{Z}_t)} \tag{3.8}$$

préalablement à une étape de normalisation assurant que $\sum_{i=1}^{N} \pi_t^{(i)} = 1$

fin

Le nuage $\{\mathbf{x}_t^{(i)}, \pi_t^{(i)}\}_{i=1\ldots N}$ permet d'approcher la loi de filtrage par

$$p(\mathbf{x}_t|\mathbf{Z}_t) \simeq \sum_{i=1}^{N} \pi_t^{(i)} \delta(\mathbf{x}_t - \mathbf{x}_t^{(i)}) \tag{3.9}$$

Rééchantillonner $\{\mathbf{x}_t^{(i)}, \pi_t^{(i)}\}$ selon $P(\tilde{\mathbf{x}}_t^{(i)} = \mathbf{x}_t^{(j)}) = \pi_t^{(j)}$, de façon à obtenir un ensemble de particules pondérées $\{\tilde{\mathbf{x}}_t^{(i)}, \frac{1}{N}\}$ tel que $\sum_{i=1}^{N} \pi_t^{(i)} \delta(\mathbf{x}_t - \mathbf{x}_t^{(i)})$ et $\frac{1}{N} \sum_{i=1}^{N} \delta(\mathbf{x}_t - \tilde{\mathbf{x}}_t^{(i)})$ approximent $p(\mathbf{x}_t|\mathbf{Z}_t)$.
Affecter $\mathbf{x}_t^{(i)}$ et $\pi_t^{(i)}$ avec $\tilde{\mathbf{x}}_t^{(i)}$ et $\frac{1}{N}$.

fin

3.4 Extension au filtrage particulaire à état mixte

Notre problème de suivi et ré-identification conjoint de cible met intrinsèquement en jeu des variables continues (la position des cibles) et des variables discrètes (l'identité de ces cibles). Dans cette section, nous présentons une technique d'échantillonnage dans un espace d'état mixte, en partie continu (position con-

tinue) et en partie discret (ce paramètre pourra représenter l'identité dans une base), de la cible traitée. De cette manière, le filtre proposé infère des distributions de probabilité mixtes.

Ce filtrage à état mixte a été introduit par Isard et Blake comme extension de la CONDENSATION dans [Isard et Blake, 1998a] dans un contexte de banque de modèles de dynamique. Le principe est de traiter différents types de particules, différentes par le modèle de mouvement qui leur est associé. L'algorithme 3 détaille la gestion du vecteur d'état mixte et son échantillonnage en deux phases. L'article montre qu'autoriser différents modèles de mouvements permet de mieux répartir les particules et d'estimer aussi le modèle de mouvement représentant le mieux celui du système à l'instant donné, *i.e.* celui pour lequel le plus particules ont « voté ». Ce principe a été repris par Brèthes *et al.* dans [Brèthes *et al.*, 2004] pour estimer le template décrivant au mieux la main d'un opérateur lors d'une interaction homme-machine, parmi un ensemble de configurations prédéfinies. Là encore, le template choisi est celui pour lequel le maximum de particules a « voté ».

3.4.1 Modèle de prédiction à état mixte

Nous cherchons à estimer un vecteur d'état mixte, ajoutant un terme discret aux paramètres continus appartenant à un espace à k dimensions, soit :

$$\mathbf{X} = (\mathbf{x}, id)^{\mathsf{T}},\ \mathbf{x} \in \mathbb{R}^k,\ id \in \{1, \dots, N_{id}\} \tag{3.10}$$

Étant donné ce vecteur d'état étendu, la densité du processus d'échantillonnage à l'image t peut être décomposée comme dans [Isard et Blake, 1998a] :

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}) = p(\mathbf{x}_t|id_t, \mathbf{X}_{t-1}) \cdot P(id_t|\mathbf{X}_{t-1}) \tag{3.11}$$

avec :

$$P(id_t|\mathbf{X}_{t-1}) : P(id_t = j|\mathbf{x}_{t-1}, id_{t-1} = i) = T_{ij}(\mathbf{x}_{t-1}) \tag{3.12}$$

$$p(\mathbf{x}_t|id_t, \mathbf{X}_{t-1}) : p(\mathbf{x}_t|\mathbf{x}_{t-1}, id_{t-1} = i, id_t = j) = p_{ij}(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{3.13}$$

où $T_{ij}(\mathbf{x}_{t-1})$ est la probabilité de transition de l'identité i vers j, appliquée au paramètre discret d'identité. De son côté, $p_{ij}(\mathbf{x}_t|\mathbf{x}_{t-1})$ est l'échantillonnage de la loi appliquée à la partie continue de l'état.

3.4.2 Exploitation de la mesure dans le cas d'un état mixte

Suite à la propagation mixte des particules et l'application du modèle d'observation, l'estimation de l'état mixte est ensuite un processus en deux étapes. Le

Algorithme 3: Algorithme de la CONDENSATION à état mixte.

Données : Mesure à t : $\mathbf{Z}_t$,
ensemble de particules pondérées à $t-1$: $\{\mathbf{X}_{t-1}^{(i)}, \pi_{t-1}^{(i)}\}_{i=1\ldots N}$

Résultat : Ensemble de particules pondérées à t : $\{\mathbf{X}_t^{(i)}, \pi_t^{(i)}\}_{i=1\ldots N}$

si *t=0 (INITIALISATION)* **alors**

`// Distribuer les états discret de manière équiprobables entre les particules`

Échantillonner $\mathbf{x}_0^{(1)}, \ldots, \mathbf{x}_0^{(i)}, \ldots, \mathbf{x}_0^{(N)}$ i.i.d. selon $p(\mathbf{x}_0)$, et poser $\pi_0^{(i)} = \frac{1}{N}$

fin

si $t \geq 1$ **alors**

pour $i = 1, \ldots, N$ **faire**

1. Échantillonner selon $p(\mathbf{X}_t|\mathbf{X}_{t-1} = s_t^{(i)})$ (équation (3.11));
   (a) Échantillonner selon les probabilités de transition des identités (équation (3.12)) pour trouver $id_t^{(i)}$;
   (b) Échantillonner selon l'équation (3.13) la partie continue de l'état : $\mathbf{x}_{t-1}^{(i)}$;
2. Mettre à jour les poids $\pi_t^{(i)}$ selon l'équation :

$$\pi_t^{(i)} \propto p(\mathbf{Z}_t|\mathbf{X}_t^{(i)}) \tag{3.14}$$

préalablement à une étape de normalisation assurant que $\sum_{i=1}^{N} \pi_t^{(i)} = 1$

fin

Rééchantillonner $\{\mathbf{X}_t^{(i)}, \pi_t^{(i)}\}$ selon $P(\tilde{\mathbf{X}}_t^{(i)} = \mathbf{X}_t^{(j)}) = \pi_t^{(j)}$, de façons à obtenir un ensemble de particules pondérées $\{\tilde{\mathbf{X}}_t^{(i)}, \frac{1}{N}\}$ tel que $\sum_{i=1}^{N} \pi_t^{(i)} \delta(\mathbf{X}_t - \mathbf{X}_t^{(i)})$ et $\frac{1}{N}\sum_{i=1}^{N} \delta(\mathbf{X}_t - \tilde{\mathbf{X}}_t^{(i)})$ approximent $p(\mathbf{X}_t|\mathbf{Z}_t)$. Affecter $\mathbf{X}_t^{(i)}$ et $\pi_t^{(i)}$ avec $\tilde{\mathbf{X}}_t^{(i)}$ et $\frac{1}{N}$.

L'estimation de l'état se fait en deux étapes :
(a) estimation de la partie discrète (3.15) ;
(b) estimation de la partie continue sur les sous-ensemble correspondant (3.16).

fin

processus commence par calculer le Maximum A Posteriori sur le paramètre dis-

cret :

$$\begin{aligned} \hat{id}_t &= \arg\max_j P(id_t = j|\mathbf{Z}_t) \\ &= \arg\max_j \sum_{i\in\boldsymbol{\Upsilon}_j} \pi_t^{(i)}, \text{ où } \boldsymbol{\Upsilon}_j = \left\{n|\mathbf{X}_t^{(i)} = (\mathbf{x}_t^{(i)}, j)^\mathsf{T}\right\}_{1\leq n\leq N} \end{aligned} \tag{3.15}$$

Ensuite, les composantes continues sont estimées sur le sous-ensemble de particules qui possèdent le paramètre discret le plus vraisemblable.

$$\hat{\mathbf{x}}_t = \sum_{i\in\hat{\boldsymbol{\Upsilon}}} \pi_t^{(i)} \cdot \mathbf{x}_t^{(i)} / \sum_{i\in\hat{\boldsymbol{\Upsilon}}} \pi_t^{(i)}, \text{ où } \hat{\boldsymbol{\Upsilon}} = \left\{i|\mathbf{X}_t^{(i)} = (\mathbf{x}_t^{(i)}, \hat{id}_t)^\mathsf{T}\right\} \tag{3.16}$$

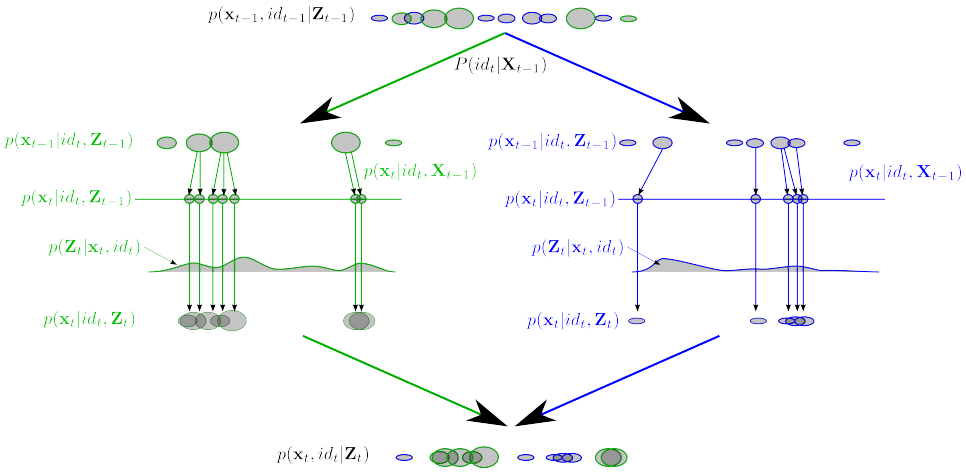


FIGURE 3.2 – Schématisation du mixed-state CONDENSATION. Nous illustrons ici le cas d'un paramètre discret pouvant prendre deux valeurs différentes (dont les évolutions sont représentées en vert et bleu respectivement).

3.5 Transition vers le suivi multi-cibles

Toutefois, le contexte applicatif que nous visons fait intervenir des cibles multiples. En ce sens, la CONDENSATION ne répond pas complètement à notre problème. Nous présentons donc ici en détail l'algorithme de suivi multi-cibles proposé par Breitenstein *et al.* dans [Breitenstein *et al.*, 2010]. Il s'agit d'une approche par filtres particulaires distribués, qui répond au cahier des charges que nous nous sommes fixé. Par ailleurs, la performance de cette approche a été clairement établie sur un vaste ensemble de séquences [Breitenstein *et al.*, 2010], ainsi que des études par rapport à d'autres formalismes, en logique séquentielle avec [Okuma *et al.*, 2004] et en logique différée avec [Leibe *et al.*, 2007, Berclaz *et al.*, 2006, Huang *et al.*, 2008, Wu et Nevatia, 2007].

3.5.1 Gestion des détections

Breitenstein *et al.* privilégient une stratégie « tracking-by-detection » via le détecteur classique proposé dans [Dalal et Triggs, 2005].

Détection de piéton par histogrammes d'orientation de gradients Le détecteur par histogrammes d'orientation de gradients (HOG) publié par Dalal et Triggs dans [Dalal et Triggs, 2005] est une approche par fenêtre glissante. Durant l'entrainement, le détecteur divise la boite englobante à traiter en cellules de tailles constantes et calcule un histogramme d'orientation de gradients pour chacune. Suite à une normalisation, toutes ces caractéristiques sont accumulées dans un vecteur, utilisé pour entrainer un SVM linéaire.

La phase de détection est similaire. L'image à traiter est scannée par fenêtre glissante à différentes échelles. Les caractéristiques HOG sont calculées sur les boites englobantes à tester, puis classifiées comme piéton ou non-piéton par le SVM. Les détections finales sont produites suite à une phase de suppression des non-maximum.

Association aux détections/Gestion du caractère multi-cibles En accord avec [Breitenstein *et al.*, 2010], ces détections HOG sont intégrées dans le processus de suivi par une étape préalable d'association aux traqueurs. À la fin de cette étape, chaque traqueur est potentiellement (il est possible qu'un traqueur ne reçoive aucune détection) associé à une détection qui va servir à la mise à jour de ses particules. Pour ce faire, nous construisons une matrice d'association entre les détections (lignes) et les traqueurs (colonnes). Le score de chaque paire détection d, traqueur tr donné par l'équation (3.17), fait intervenir :

- ▷ la distance des particules du traqueur à la détection évaluée sous une loi normale $p_{\mathcal{N}}(.) \sim \mathcal{N}(., \sigma^2)$,
- ▷ l'aire de la boîte du traqueur $\mathcal{A}(tr)$ relativement à celle de la détection aussi évaluée sous une loi normale,
- ▷ l'évaluation du modèle d'apparence du traqueur en la détection ($w_{App}(.)$).

$$S(d,tr) = \underbrace{\sum_{p_i \in tr}^{N} p_{\mathcal{N}}(d - p_i)}_{\text{distance euclidienne}} \times \underbrace{p_{\mathcal{N}}\left(\frac{|\mathcal{A}(tr) - \mathcal{A}(d)|}{\mathcal{A}(tr)}\right)}_{\text{taille relative}} \times \underbrace{w_{App}(d,tr)}_{\text{modèle d'apparence}} \tag{3.17}$$

Ainsi, le traqueur et la détection doivent présenter simultanément une cohérence en terme de position, de taille et de contenu colorimétrique. Une fois cette matrice de similarité construite, il faut réaliser l'appariement. En pratique, à la manière de [Breitenstein *et al.*, 2010], une heuristique gloutonne de complexité en $\mathcal{O}(n)$ est généralement suffisante ([Wu et Nevatia, 2007] tirent des conclusions similaires) par rapport à la solution optimale fournie par l'algorithme Hongrois [Kuhn, 1955],

de complexité $\mathcal{O}(n^3)$ (avec n la plus grande dimension de la matrice d'association). L'heuristique consiste en une extraction itérative des maxima, avec suppression de leurs lignes et colonnes. Elle est itérée tant que les maxima sont supérieurs au seuil d'appariement.

Initialisations / terminaisons automatiques de traqueurs Toute détection récurrente donne lieu à l'instanciation d'un nouveau traqueur. Par ailleurs, tout traqueur n'ayant pas de détection associée sur un intervalle de temps supérieur au seuil de suppression se voit arrêté définitivement.

3.5.2 Modèle d'observation intégrant les détections

En accord avec [Breitenstein *et al.*, 2010], le poids $\pi_{tr}^{(i)}$ attribué à la i^{e} particule p_i du traqueur tr est calculé en intégrant : (i) la distance de la particule à la détection d^* qui lui a été associée, (ii) la similarité colorimétrique au modèle d'apparence du traqueur $w_{App}(.)$.

$$\pi_{tr}^{(i)} = \underbrace{\alpha \cdot \mathcal{I}(tr) \cdot p_{\mathcal{N}}(d^* - p_i)}_{\text{distance à la détection}} + \underbrace{\beta \cdot w_{App}(d, tr)}_{\text{modèle d'apparence}} \tag{3.18}$$

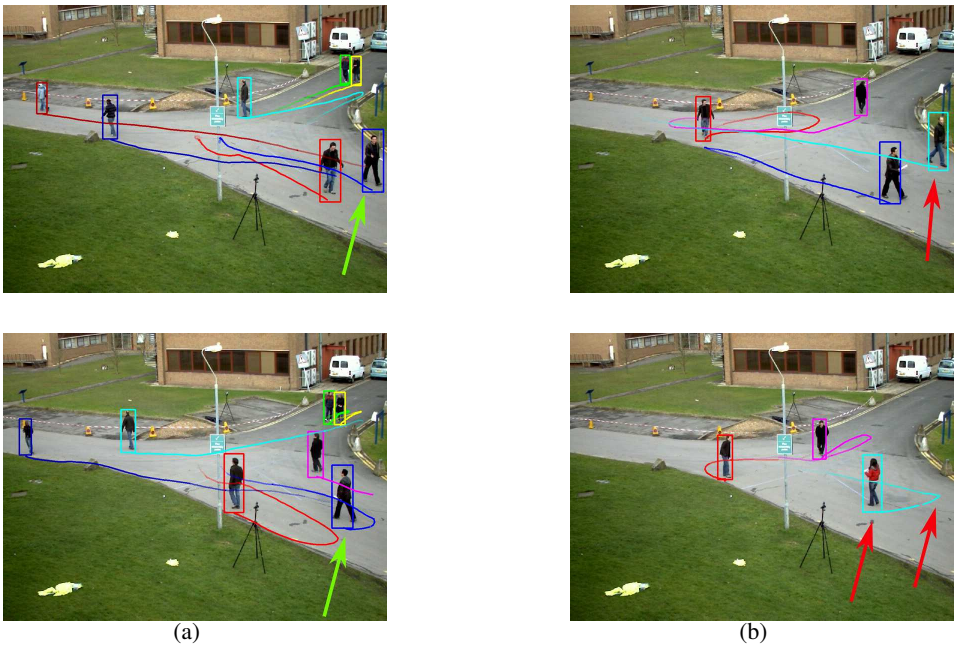
où α et β sont des coefficients dont la somme est égale à 1, et $\mathcal{I}(tr)$ un booléen signifiant l'existence ou non d'une détection associée au traqueur.

3.5.3 Notion d'*identité* d'une cible

Par essence, un algorithme de suivi multi-cibles fait intervenir une notion d'identité, pour distinguer un traqueur d'un autre. Toutefois, cette identité ne sera valable que pour le temps d'apparition de la cible et prendra fin avec l'arrêt de son traqueur dédié. Si cette même cible est amenée à revenir dans la scène après un certain temps d'absence, elle se verra affecter un nouveau traqueur, et ainsi une nouvelle identité. Nous parlons alors d'*identité locale*, par opposition aux *identités* au sens du réseau, que nous visons pour notre suivi et ré-identification.

Les figures 3.3 et 3.4, présentent la limitation d'une simple gestion d'*identités locales* et l'apport de notre modalité de ré-identification. Sur la figure 3.3, lorsqu'une personne sort et une personne différente entre, les trajectoires sont raboutées. Un simple critère spatial est utilisé dans [Breitenstein *et al.*, 2010]. La figure 3.3 (b) met en exergue cette limitation lorsque la personne sortie n'est pas la même que celle qui entre. Le traqueur serait également pris en défaut si la personne suivie réapparait dans une autre région de l'image, typiquement dans un réseau de couloir.

Pour notre approche (figure 3.4), à chaque instant, chaque traqueur propose une distribution de probabilité d'identité observée. Ceci permet d'accepter des périodes de non observabilité comme une sortie de caméra puis de ré-initialiser le traqueur



(a) (b)

FIGURE 3.3 – Résultats issus de [Breitenstein *et al.*, 2010] (a) Entre les images 204 et 241 la personne fléchée sort puis entre à nouveau. Un traqueur est maintenu et apparie la bonne personne, sur un critère spatial. (b) La même situation se produit entre les images 390 et 445. Mais cette fois, une nouvelle personne entre, un traqueur déjà existant lui est affecté, la trajectoire est reprise. Il s'agit d'une erreur de ré-identification.

avec le bon identifiant. Lorsqu'une personne entre, le traqueur qui la suit va converger vers des identités de la base.

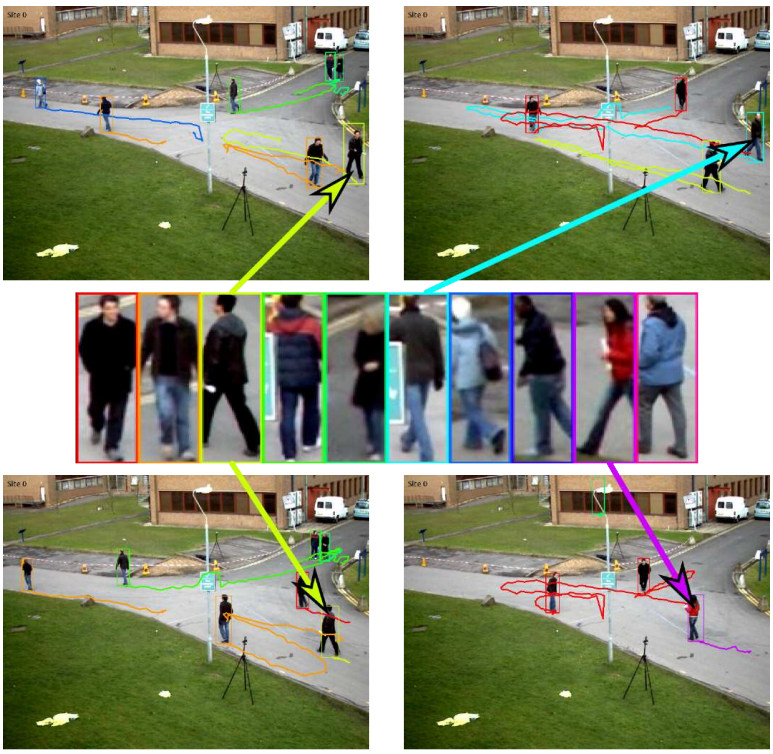
3.6 Suivi et ré-identification conjoints

Pour gérer les *identités* représentées par une base, nous étendons l'approche de [Breitenstein *et al.*, 2010] avec le formalisme du filtrage particulaire à état mixte. Ceci engendre les modifications suivantes dans l'algorithme.

3.6.1 Association traqueurs mixtes/détections

Le coût de l'association aux détections fait maintenant intervenir un paramètre lié à l'identité de la cible (noté en bleu).

- ▷ la distance des particules du traqueur à la détection évaluée sous une loi normale $p_{\mathcal{N}}(.) \sim \mathcal{N}(., \sigma^2)$,



(a) (b)

FIGURE 3.4 – Intérêt de la ré-identification intégrée au suivi multi-cibles : le système ré-identifie la piste suivie par rapport à la base d'identités et permet de détecter que les personnes entrantes et sortantes sont les mêmes en (a) et différentes en (b).

- ▷ l'aire de la boîte du traqueur $\mathcal{A}(tr)$ relativement à celle de la détection aussi évaluée sous une loi normale,
- ▷ l'évaluation du modèle d'apparence du traqueur en la détection $(w_{App}(.))$.
- ▷ les scores des modèles d'identités évalués sur la détection $(w_{Id}(.))$ pondérés par les cardinaux des sous-ensembles de particules Υ_j comme définis par l'équation (3.15).

$$S(d,tr) = \underbrace{\sum_{p\in tr}^{N} p_{\mathcal{N}}(d-p)}_{\text{distance euclidienne}} \times \underbrace{p_{\mathcal{N}}\left(\frac{|\mathcal{A}(tr)-\mathcal{A}(d)|}{\mathcal{A}(tr)}\right)}_{\text{taille relative}} \times \underbrace{w_{App}(d,tr)}_{\text{modèle d'apparence}} \times \underbrace{\sum_{j=1}^{N_{id}} \Upsilon_j \cdot w_{Id}(d,j)}_{\text{distributions d'identités}} \quad (3.19)$$

3.6.2 Modèle d'observation mixte intégrant les détections

Après l'étape d'échantillonnage, les nouvelles positions de particules sont évaluées. La vraisemblance temporelle de la CONDENSATION (illustration de l'algo-

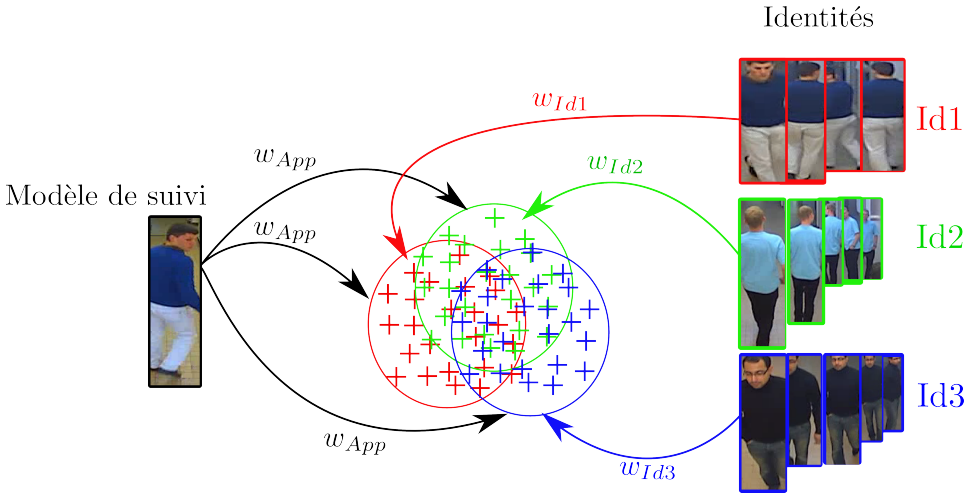


FIGURE 3.5 – Nuage de particules représentant un vecteur d'état mixte (les croix, dont l'identité est indicée par la couleur), et pondération pour la ré-identification associée : la référence de suivi classique (gauche) génère la vraisemblance w_{App}, alors que l'identité référée dans la base permet de calculer la vraisemblance d'identité w_{Id}.

rithme originel en figure 3.1) $p(\mathbf{Z}_t|\mathbf{x}_t^{(i)})$ est approximée par $w_{App}^{(i)}(t)$.

Le formalisme du Mixed-State CONDENSATION adapté à la ré-identification fournit une vraisemblance additionnelle, pondérant les particules relativement à leur identité de référence, $p(\mathbf{Z}_t|\mathbf{x}_t^{(i)}, y_t^{(i)}) : w_{Id}^{(i)}(t)$.

w_{App} est évaluée sur des descripteurs issus de la même caméra, à des instants proches, alors que pour w_{Id}, ils proviennent de deux caméras différentes. Les ordres de grandeur de ces vraisemblances ne sont pas les mêmes. Pour pouvoir toutefois combiner ces vraisemblances, nous les normalisons chacune sur l'ensemble des particules avant l'étape de ré-échantillonnage du filtre. D'une manière proche de [Javed *et al.*, 2005] et [Gilbert et Bowden, 2006] qui proposent l'estimation d'une fonction de transfert colorimétrique inter-caméras, nous utilisons l'hypothèse de « *closed world* », selon laquelle l'identité que nous cherchons fait partie de la base. La normalisation que nous introduisons ici nous permet de relever la réponse de la caméra d'apprentissage et de ré-identifier les cibles actives de suivi. Nous évitons l'estimation coûteuse et propre à une paire de caméras de la BTF (décrite en section 2.4) et nous garantissons ainsi le changement de caméras, connu comme ré-identification. Nous notons w_{Id}^* et w_{App}^* les vraisemblances normalisées.

Si w_{App}^* est supérieur à un seuil (*i.e.* si la particule est intéressante, sinon nous conservons cette faible vraisemblance temporelle comme vraisemblance combinée), nous combinons ces deux similarités normalisées pour obtenir l'expression

de la vraisemblance de particule qui sera injectée dans l'étape de pondération du filtre à particules (incrément noté en bleu) :

$$\pi_t^{(i)} = \alpha \cdot w_{App}^{*(i)}(t) + (1-\alpha) \cdot w_{Id}^{*(i)}(t), \ \forall\, i = 1, \ldots, N. \tag{3.20}$$

Ce faisant, nous donnons de l'importance aux particules correctement positionnées, qui possèdent la bonne identité.

Le poids $\pi_{tr}^{(i)}$ attribué à la i^{e} particule p_i du traqueur tr est calculé en intégrant la distance de la particule à la détection d^* qui lui a été associée, la similarité colorimétrique au modèle d'apparence du traqueur $w_{App}(.)$ et la similarité colorimétrique à l'identité de la particule $w_{Id}(.)$. $Id(i)$ représente l'identité choisie par i. Il s'agit du terme mixte. Ainsi, l'équation (3.20) du cas mono-cible se reformule (incrément noté en bleu) :

$$\pi_{tr}^{(i)} = \underbrace{\alpha \cdot \mathcal{I}(tr) \cdot p_{\mathcal{N}}(d^* - p_i)}_{\text{distance à la détection}} + \underbrace{\beta \cdot w_{App}^*(d, tr)}_{\text{modèle d'apparence}} + \underbrace{\gamma \cdot w_{Id}^*(d, id(i))}_{\text{identité}} \tag{3.21}$$

où α, β et γ sont des coefficients dont la somme est égale à 1, et $\mathcal{I}(tr)$ un booléen signifiant l'existence ou non d'une détection associée au traqueur (l'équation rend ainsi compte de la possibilité de l'absence de détection). L'introduction d'une similarité relative à l'identité dans la pondération de la particule permet de diriger le nuage de particule vers les identités les plus probables au vu des observations reçues. En ce sens, chaque traqueur maintient une multi-modalité sur les identités les plus vraisemblables pour la personne qu'il suit.

Dans la suite du manuscrit, nous emploierons l'acronyme MSR pour « Mixed-State Re-identification »pour désigner la méthode de CONDENSATION à état mixte adaptée à la ré-identification.

3.7 Implémentation

3.7.1 Modélisation de l'apparence d'une cible

Nous utilisons le modèle d'apparence décrit en section 2.6 : SDALF restreint. Il se compose de distributions couleurs dans l'espace HSV, ainsi que d'ellipses de couleurs moyennes MSCR. Ceci nous permet de calculer les similarités par rapport au modèle d'apparence d'un traqueur ainsi que par rapport à une identité de la base, notées respectivement $w_{App}(.)$ et $w_{Id}(.)$.

3.7.2 Descriptions des identités du réseau

Tout algorithme de ré-identification nécessite d'avoir vu une personne au préalable pour être capable de la ré-identifier. Nous supposons ici la phase de constitution d'une telle base acquise. Pour cela, nous extrayons une collection d'images

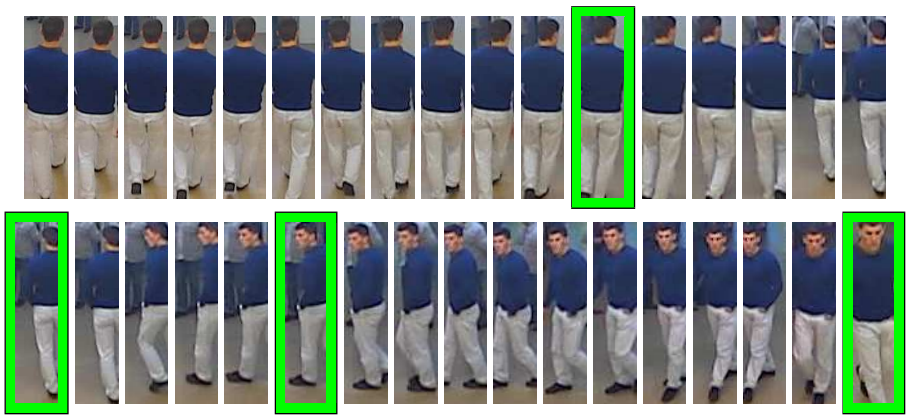


FIGURE 3.6 – Quelques ROI issues d'une séquence de suivi (ramenées à la même échelle pour la représentation). Les cadres verts mettent en valeur les quatre images clés retenues par la méthode pour décrire la séquence. Elles capturent les plus grandes variations d'apparence de la cible au cours de la séquence.

clés d'une des caméras du réseau (*e.g.* positionnée dans le hall d'entrée du bâtiment à surveiller), et utilisons celles-ci comme descriptions de nos identités. Le choix des images clés est réalisé par K-means sur des séquences de suivi dans la caméra choisie comme caméra d'entrée. Ainsi, ces images clés encodent la variabilité d'apparence obtenue pour cette identité au cours de son suivi initial, comme le présente la figure 3.6.

3.7.3 Vecteur d'état

La partie continue de l'état $\mathbf{x} = [x, y, v_x, v_y]^\mathsf{T}$ se compose de la position dans le plan image (ou sur le plan du sol si l'on dispose des calibrations intrinsèque et extrinsèque) $(x, y)^\mathsf{T}$ et du vecteur vitesse $(v_x, v_y)^\mathsf{T}$. La partie discrète id renvoie à l'une des N_{id} identités de la base. Le suivi se passe dans le plan image, et la dimension des boîtes de suivi est fixée et mise à jour sur les détections associées à ces traqueurs. Le modèle d'apparence est lui aussi mis à jour sur la détection associée.

3.7.4 Modèle de mouvement

Le paramètre discret d'identité id est échantillonné selon une matrice de transition $T = [T_{ij}]_{1 \leq i,j \leq N}$. Elle est construite sur l'ensemble des images clés. L'élément T_{ij} est la similarité $w_{id}(.)$ entre les identités i et j de la base, calculée entre les images clés les plus dissemblables.

Les particules sont propagées selon un modèle de mouvement d'ordre 1 : $p_{ij}(\mathbf{x}_t|\mathbf{x}_{t-1})$:

$$\begin{cases} (x,y)_t = (x,y)_{t-1} + (v_x,v_y)_{t-1} \cdot \Delta t + \epsilon_{(x,y)} \\ (v_x,v_y)_t = (v_x,v_y)_{t-1} + \epsilon_{(v_x,v_y)} \end{cases} \tag{3.22}$$

où les bruits $\epsilon_{(x,y)}$ et $\epsilon_{(v_x,v_y)}$ suivent des lois normales et où Δt est l'intervalle de temps séparant deux images.

3.7.5 Modèle d'observation

Les vraisemblances temporelles (équations (3.20) et (3.21)) sont calculées comme :

$$w_{App}^{(i)}(t) = \exp\{-K_1 \cdot \sum_{j=1}^{N_c} D_{SDALF}^2 \left(s_t^{(i)}(j), s_{model}(j) \right)\}, \forall\, i = 1, \ldots, N \tag{3.23}$$

où N_c est le nombre de distributions de couleurs par cible, $s_{model}(.)$ l'ensemble des distributions de couleurs de la référence temporelle, $s_t^{(i)}(.)$ l'ensemble des distributions de couleur de la particule courante, N le nombre de particules, et $K_1 = 1/(2 \cdot \sigma_{App}^2)$ est une constante de normalisation. Empriquement nous avons fixé $\sigma_{App} = 0.2$.

Les vraisemblances d'identité (équations (3.20) et (3.21)) sont calculées comme :

$$w_{Id}^{(n)}(t) = \exp\{-K_2 \cdot \min_{i \in N_y} \sum_{j=1}^{N_c} D_{SDALF}^2 \left(s_t^{(i)}(j), s_{identity}(j, y_t^{(i)}, k) \right)\}, \forall\, i = 1, \ldots, N \tag{3.24}$$

où N_y est le cardinal de la classe d'images clés de l'identité $y_t^{(i)}$ ($y_t^{(i)}$ étant l'identité assignée à la i-ième particule au temps t), N_c le nombre de distributions de couleurs par cible, $s_{identity}(., y_t^{(i)}, k)$ est l'ensemble de distributions de couleurs de la k-ième image clé de l'identité $y_t^{(i)}$ dans la base, $s_t^{(i)}(.)$ est l'ensemble de distributions de couleurs de la particule courante, N le nombre de particules, et $K_2 = 1/(2 \cdot \sigma_{Id}^2)$ est une constante de normalisation. Empriquement nous avons aussi fixé $\sigma_{Id} = 0.2$.

La figure 3.5 résume le principe de ces deux vraisemblances par particule. Chacune est évaluée relativement à la référence temporelle de suivi (w_{App}), mais aussi (w_{Id}) relativement à son identité (décrite par une collection d'images clés). Dans les deux cas, D_{SDALF} est la distance entre des signatures SDALF [Farenzena *et al.*, 2010], comme définie au chapitre 2.

3.7.6 Caractérisation des paramètres libres de notre système

Notre réseau IP fournit une fréquence moyenne de 16 images par seconde pour le flux vidéo à traiter. Nous fixons donc $\Delta t = 1/16s$ dans le modèle d'évolu-

tion des filtres particulaires. Dans le modèle d'observation des particules, équation (3.21), nous fixons de manière empirique :

$$\begin{cases} \alpha = 0.90,\ \beta = 0.05 \text{ et } \gamma = 0.05 & \text{si } \mathcal{I}(tr) = 1 \text{ [Breitenstein } \textit{et al.}\text{, 2010]} \\ \alpha = 0.0,\ \beta = 0.8 \text{ et } \gamma = 0.2 & \text{sinon, filtre MSR simple.} \end{cases}$$

Le nombre de particules est fixé à 400 dans l'ensemble des expériences réalisées.

Le tableau 3.2 synthétise l'ensemble des paramètres libres du système que nous proposons, avec leurs valeurs associées que nous avons fixées empiriquement. Les notations sont les mêmes que dans la présentation de la méthode.

	Paramètre	Notation	Valeur
$\mathcal{I}(tr) = 1$	Coefficient distance à la détection	α	0.8
	Coefficient w_{App}	β	0.05
	Coefficient w_{Id}	γ	0.05
$\mathcal{I}(tr) = 0$	Coefficient distance à la détection	α	0.0
	Coefficient w_{App}	β	0.8
	Coefficient w_{Id}	γ	0.2
	Nombre de particules	N	400
	Écart-type observation d'apparence w_{App}	σ_{App}	0.2
	Écart-type observation d'identité w_{Id}	σ_{Id}	0.2
	Dimension espace continu	k	4

TABLE 3.2 – Tableau récapitulatif des différents paramètres libres.

3.8 Évaluations et analyses associées

3.8.1 Jeux de données

Nous évaluons les différentes composantes de notre approche sur quatre jeux de données, reflétant différents contextes et différents niveaux de difficultés.

PETS'09- Tout d'abord nous testons le traqueur dédié à chaque noeud/caméra, sans, puis avec le module de ré-identification actif, sur la séquence PETS'09 S2L1 [1]. Cette séquence publique, longue de 795 images, présente un espace ouvert, dans lequel évoluent 10 individus, avec croisements et entrées/sorties. Ayant labellisé ce jeu de données, nous sommes en mesure de quantifier les résultats de notre algorithme de suivi.

1. http ://www.cvg.rdg.ac.uk/PETS2009/a.html

FIGURE 3.7 – Base des 10 identités de la séquence PETS.

La figure 3.7 présente la base d'identités utilisée pour traiter la séquence PETS. Il s'agit ici d'un contexte monocaméra. Les images de la base sont donc issues de la caméra où est réalisé le suivi. Il s'agit d'un « toy example », dont le seul but est de valider notre implémentation de [Breitenstein *et al.*, 2010], notamment par rapport à l'intégration du MSR. Nous présentons ici quelques images de la séquence, pour chaque identité.

NOFOV0- Ce jeu de données représente des acquisitions préliminaires qui ont permis l'élaboration de la méthode. Comme nous l'avons mentionné au chapitre 1, il n'existe pas de jeu de données publics présentant de réseaux à champs disjoints. Nous avons donc dû les réaliser nous-mêmes. Le jeu de données présente 5 caméras à champs disjoint, dont une en extérieur et la figure 3.8 présente la base d'identité de ce réseau. Toutefois, le scenario était relativement simple et en contexte monocible. Nous utilisons ce jeu de données uniquement pour la validation de la stratégie MSR, par rapport à une stratégie naïve de suivi puis de ré-identification.

FIGURE 3.8 – Base des 16 identités de la séquence NOFOV0.

NOFOV1- Au niveau du réseau à champs disjoints et étant donnée l'absence de jeux de données publics associés, nous évaluerons la composante de supervision sur une séquence privée notée NOFOV1 (pour « Non Overlapping Field Of View »). La séquence présente un ensemble de 7 personnes transitant entre 3 caméras. Il n'y a pas de champs de vue commun entre les caméras, deux sont placées en intérieur de bâtiment, alors que la $3^{ème}$ surveille un espace ouvert extérieur avec une configuration similaire à la séquence PETS'09. La séquence représente 837 images.

La figure 3.9 présente un exemple de base d'identités.

FIGURE 3.9 – Base des 7 identités de la séquence NOFOV1 (caméra 1).

NOFOV2- Cette acquisition présente un ensemble de 12 personnes évoluant dans un réseau de cinq caméras. Deux sont placées dans un couloir, une dans une salle de réunion, et les deux dernières sont montées sur des mâts extérieurs à 5m de haut, surveillant un parking dans une configuration proche du PETS'09 Dataset. Il n'y a aucun champ recouvrant entre les caméras. La séquence représente 4000 images, et la vérité terrain représente 11454 boites englobantes de piétons étiquetées. Notre objectif est de rendre ces données publiques à terme. Cette séquence sera utilisée au chapitre 4.

FIGURE 3.10 – Base des 12 identités de la séquence NOFOV2 (caméra 0).

3.8.2 Critères et modalités évalués

Nous utilisons les métriques CLEAR MOT [Bernardin et Stiefelhagen, 2008] pour la quantification des résultats de suivi. Nous obtenons un score de précision : MOTP (pour « Multi-Object Tracking Precision »), calculé comme le rapport de l'intersection sur l'union des boîtes de suivi avec celles de la vérité terrain, et un score d'« accuracy » : MOTA (pour « Multi-Object Tracking Accuracy ») prenant en compte les faux positifs, les faux négatifs et les changements de cibles des traqueurs.

MOTP- La « Multiple Object Tracking Precision » se définit comme :

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}$$

Il s'agit de l'erreur totale dans les positions estimées pour les paires cibles-positions vérité terrain sur toutes les images de la séquence, moyennée par le nombre d'appariements faits. Cela met en avant la propension de l'algorithme à estimer

précisément la position des objets, indépendamment de sa capacité à reconnaître une configuration de cibles, conserver des trajectoires consistantes...

MOTA- La « Multiple Object Tracking Accuracy » se définit comme :

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t},$$

où m_t, fp_t, et mme_t sont respectivement les nombres de cibles manquées, de faux positifs et de mauvaises associations, au temps t. Ce critère peut être vu comme dérivé des trois ratios d'erreur suivants. Le ratio de cibles manquées dans la séquence, calculé par rapport au nombre total de cibles présentes dans l'ensemble de la séquence est donné par :

$$\bar{m} = \frac{\sum_t m_t}{\sum_t g_t},$$

le ratio de faux positifs est donné par :

$$\bar{fp} = \frac{\sum_t fp_t}{\sum_t g_t},$$

et le ratio de mauvaises associations par :

$$\bar{mme} = \frac{\sum_t mme_t}{\sum_t g_t}.$$

TRR- De plus nous évaluons les capacités de ré-identification par un taux de ré-identification correcte TRR (pour True Re-identification Rate), calculé comme le rapport du nombre de ré-identifications correctes sur le nombre de ré-identifications totales. Étant donné que le superviseur opèrera sur une fenêtre temporelle, les critères TRR sont estimés en fin de fenêtre temporelle.

Dans la suite nous évaluons notre algorithme de suivi et ré-identification sur les séquences que nous venons de présenter, avec les métriques que nous venons de détailler.

3.8.3 Performances de la méthode d'échantillonnage mixte

Comme expliqué en section 3.6, notre approche Mixed-State amène une nouvelle réponse au problème de ré-identification. Pour évaluer cette approche, nous commençons par une comparaison avec celle de l'état de l'art, dans le cas de suivi d'une seule cible. Nous calculons les taux de ré-identification pour toutes les identités de la base, dans chacune des caméras avec une stratégie image par image à chaque instant et avec notre stratégie. Dans les deux cas, nous utilisons le même

descripteur car nous évaluons les stratégies d'appariement. Pour la comparaison image à image, nous utilisons un suivi sans ré-identification, et à chaque nouvelle estimation de position, nous la comparons à chacune des entrées de la base. À ceci nous confrontons notre stratégie MSR. La vérité terrain sur les identités permet d'obtenir une réponse binaire pour chaque stratégie, pour chaque cible et à chaque instant. Pour chaque caméra, nous sommons les résultats obtenus sur les cibles, ce qui donne des taux de ré-identification par image. Nous moyennons ensuite ces taux sur la séquence. De plus les taux sont moyennés sur cinq répétitions de chaque suivi pour prendre en compte la nature stochastique du filtrage particulaire. Le tableau 3.3 résume les résultats obtenus.

TABLE 3.3 – Taux de ré-identification entre caméras pour la stratégie triviale et la stratégie Mixed-State.

	Suivi puis REID	MSR
Site #0 vers #0	0.96	**0.98**
Site #0 vers #1	0.40	**0.46**
Site #0 vers #2	0.66	**0.81**
Site #0 vers #3	0.65	**0.71**
Site #0 vers #4	0.30	**0.34**

Nous observons différents taux de ré-identification dépendants de la caméra considérée (un exemple de chacune de ces caméras est donné en figure 3.11). Le site #0 est celui où les identités ont été apprises, donc les descripteurs sont vraiment similaires. Ceci explique le taux approchant les 100%. En revanche, les sites #1 et #4 sont relativement différents en terme de pose de caméra et de contexte (le site #4 étant de plus en extérieur et présentant de forts changements d'illuminations). Cependant, nous constatons pour toutes les caméras, que en relatif la stratégie conjointe MSR est toujours supérieure.

3.8.4 Performances du suivi par ré-identification

La séquence NOFOV0 ne présentant que des passages de personnes seules, nous ne testons pas notre suivi MOT sur cette séquence.

3.8.4.1 Performances quantitatives

Le tableau 3.4 présente nos résultats quantitatifs sur les séquences PETS'09 et NOFOV1. PETS'09 nous a permis dans un premier temps de valider notre implémentation de [Breitenstein *et al.*, 2010], dont certains aspects n'ont pas été implémentés (utilisation de la confiance du détecteur dans le modèle d'observation, mais cet aspect pourra facilement être ajouté, et modèle d'apparence de type « boosting

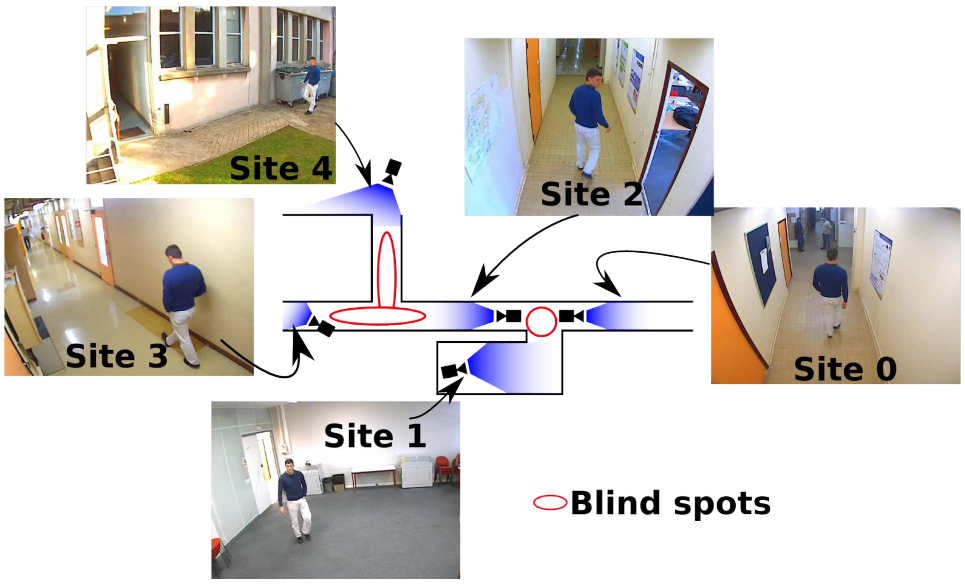


FIGURE 3.11 – Réseau NOFOV0.

online », car il est remplacé par notre SDALF simplifié de la section 2.6).

Cependant, notre approche dispose d'une modalité supplémentaire avec la notion d'*identité*. Nous montrons d'abord que l'introduction du filtrage particulaire à état mixte ne dégrade pas les performances de suivi, en comparant MOTP et MOTA pour notre implémentation sans et avec le module de ré-identification actif. Puis cette modalité supplémentaire permet d'exprimer le taux de ré-identification pour la séquence.

L'aspect stochastique du filtrage particulaire est pris en compte : le tableau 3.4 présente les résultats moyens de chaque score, sur un ensemble de 10 répétitions.

Séquence PETS'09	MOTP	MOTA	TRR
Suivi-par-détection [Breitenstein *et al.*, 2010]	**56.3%**	**79.7%**	-
Suivi-par-détection implémenté	42.7%	77.9%	-
Suivi-par-Réidentification	42.5%	77.7%	**59.7%**

TABLE 3.4 – Résultats de suivi selon les métriques CLEAR MOT [Bernardin et Stiefelhagen, 2008] et taux de ré-identification sur la séquence monocaméra PETS'09 S2L1. Nous donnons ici les Multi-Object Tracking Precision (MOTP), Multi-Object Tracking Accuracy (MOTA), et True Re-identification Rate (TRR) définis en section 3.8.2.

Le tableau 3.5 présente les taux de ré-identification par caméra, puis au niveau du réseau global sur la séquence NOFOV1. La base étant construite à partir de la caméra 1, ceci explique les taux de ré-identification supérieurs dans cette caméra.

Le superviseur présenté au chapitre 4 va logiquement tirer parti de ces performances : chaque identité correctement ré-identifiée va contraindre le système dans la suite de la topologie.

Séquence NOFOV	cam0	cam1	cam2	réseau
Suivi-par-Réidentification	43.7%	67.3%	55.5%	54.6%

TABLE 3.5 – Taux de ré-identifications correctes TRR pour chacune des caméras du réseau NOFOV1.

3.9 Conclusion

Dans ce chapitre, nous avons présenté une stratégie de suivi de personnes multiples, pour un contexte de vidéosurveillance sur une caméra fixe, permettant de ré-identifier les cibles par rapport à une base d'individus donnée *a priori*.

Comme présenté dans la thèse de MacCormick [MacCormick, 2000], les probabilités et statistiques fondent la base des algorithmes de suivi visuel, et peuvent aussi être utilisées pour interpréter les résultats de ces mêmes algorithmes. Ces algorithmes produisent leurs résultats sous la forme de distribution de probabilités, que l'on peut interpréter comme la croyance du système sur l'état du monde qu'il observe. La CONDENSATION à état mixte pour la ré-identification ne fait pas exception à la règle, au sens où elle produit une densité de probabilité de présence discrète (position de la cible), ainsi qu'une distribution de probabilité d'identité, elle aussi discrète, indicée sur la base de personnes connues. Le chapitre qui suit se base sur ces distributions d'identité pour optimiser la ré-identification.

La contribution principale exposée ici est une stratégie de filtrage bayésien à état mixte pour répondre de manière conjointe aux problèmes de suivi et de ré-identification de personnes, en supposant disposer au préalable d'une base d'identités potentielles sur laquelle échantillonner. Les approches classiques de ré-identification [Gray et Tao, 2008, Farenzena *et al.*, 2010] posent le problème en terme de description de l'identité. Nous montrons dans cette partie qu'une telle stratégie est complémentaire d'un descripteur performant. Nous comparons cette stratégie à la méthode naïve, et montrons son intérêt.

La seconde contribution de ce chapitre concerne l'extension de ce formalisme de ré-identification au problème du suivi multi-cibles. Il ressort de ce chapitre une méthode de suivi multi-cibles automatique, capable de proposer des distributions d'identités pour les cibles suivies, relativement à une base d'identités fournie en entrée. Les travaux de ce chapitre ont été publiés dans [Meden *et al.*, 2011b] pour la partie filtrage à état mixte et [Meden *et al.*, 2012b] pour le suivi et identification de cibles multiples.

À ce niveau, les questions soulevées sont nombreuses :

- ▷ l'application de cette approche à un réseau de caméra, et les contraintes supplémentaires que l'on peut y ajouter ;
- ▷ la construction de la base d'identités ;
- ▷ comment utiliser les distributions de probabilité d'identité générées.

Le chapitre 4 se positionne non plus au niveau de la caméra, mais au niveau du réseau. La topologie est donnée *a priori*, et nous allons voir comment répondre aux problèmes précédemment soulevés.

Deuxième partie

Système décisionnel haut-niveau

CHAPITRE

4

Supervision des identités : une approche réseau

Dans ce chapitre 4, nous nous plaçons désormais à l'échelle du réseau et non plus de chaque caméra. Nous disposons donc d'une vue d'ensemble des différents filtres distribués à état mixte, inférant suivi et ré-identification des personnes au sein de chaque caméra. De plus, nous supposons la relation topologique existant entre les différentes caméras connue *a priori*. À partir de là, nous proposons deux stratégies de supervision des filtres qui vérifient leurs cohérences, à la fois entre eux et relativement à la topologie du réseau. Nous introduisons ainsi des interactions entre les filtres distribués, au niveau de leur paramètre discret relatif aux identités.

4.1 Introduction

Ce chapitre aborde un second étage de filtrage, au niveau du réseau de caméras. Dans le chapitre 2, nous avons énuméré les signatures visuelles exploitées pour la ré-identification. Le chapitre 3 a présenté une stratégie distribuée de filtrage dans un espace d'état mixte, en vue de réaliser de manière conjointe, au niveau caméra, suivi et ré-identification par rapport à une base d'identités. Dans ce chapitre, nous décrivons notre stratégie de ré-identification au niveau du réseau en corrélant les sorties des traqueurs et la topologie des lieux/caméras. La topologie entre les zones d'entrée/sorties du réseau de caméras est supposée connue *a priori*. Nous abordons ici le Problème N°3 énoncé au chapitre 1 : Comment la connaissance du réseau et de la localisation des identités à un instant donné peut-elle permettre de contraindre l'appariement à réaliser, et ainsi simplifier la combinatoire d'association traqueurs/identités ? Nous formulons ici ces contraintes.

Contrairement aux traitements intra-caméra du chapitre 3, le superviseur ne fournit pas une réponse en ligne. L'idée ici est de raisonner sur une fenêtre temporelle pour agréger un maximum d'information avant de produire une décision de ré-identification la plus sûre possible. À ce titre, la section 4.2 présente un état de l'art dédié aux approches de suivi d'objets multiples par logique différée, et leur application aux réseaux de caméras à champs de vue disjoints. La section 4.3 présente la modélisation du réseau que nous adoptons et les sections 4.4 et 4.5, respectivement les deux stratégies de supervision que nous proposons. Finalement, les sections 4.6 et 4.7 respectivement présentent nos résultats et concluent ce chapitre.

4.2 État de l'art et positionnement des travaux

4.2.1 Suivi de cibles multiples par logique différée à partir d'observations continues

Le suivi de cibles multiples induit un problème d'association temporelle de détections comme l'illustre la figure 4.1. Pour un problème d'association à un instant donné, une solution répandue est la méthode hongroise [Burgeois et Lasalle, 1971], que nous approximons par son heuristique gloutonne au chapitre 3, il s'agit là d'une association à un instant donné. Il existe des approches qui raisonnent au niveau d'une fenêtre temporelle et relaxent l'hypothèse markovienne d'ordre 1. Si l'on considère la figure 4.1, cela revient à raisonner sur un intervalle de temps $[t, t+T]$, avec $T > 1$, et considérer l'ensemble de détections obtenues sur cette période pour remonter aux trajectoires des objets ayant généré ces détections.

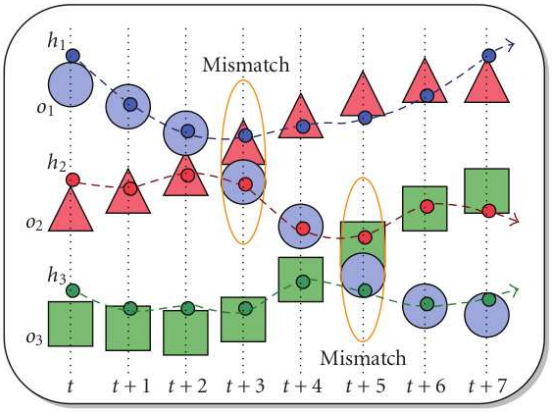


FIGURE 4.1 – Associations temporelle de trois objets $\{o_i\}_{1\leq i\leq 3}$ et trois hypothèses de suivi $\{h_i\}_{1\leq i\leq 3}$. Deux échecs d'associations (« mismatches ») sont mis en avant. (illustration issue de [Bernardin et Stiefelhagen, 2008])

4.2.1.1 Principes d'association de détections

Avant l'avènement de la Vision par ordinateur, le suivi d'objets multiples (MOT pour Multiple Object Tracking) a été largement traité par les communautés Radar et Traitement du Signal, avec les travaux originels de Reid [Reid, 1979] sur le Multi-Hypothesis Tracker (MHT). L'objectif est de filtrer sur une fenêtre temporelle un ensemble de détections en environnement encombré pour produire un suivi spatio-temporel des cibles ayant généré ces détections. Le MHT crée un arbre des possibilités, au fur et à mesure que le temps évolue. Les travaux suivants ont cherché à limiter l'explosion combinatoire inhérente à cet algorithme. En effet, ils explorent l'ensemble des assignations les plus probables avant de produire leur réponse. A ce titre, Cox et Hingorani [Cox et Hingorani, 1996] ont proposé une implémentation plus efficace du MHT, plus tard à nouveau reformulé par Danchick et Newnam [Danchick et Newnam, 2006]. Les améliorations reposent principalement sur des simplifications de l'arbre des possibilités.

Suivant les idées de Pasula *et al.* [Pasula *et al.*, 1999] sur les réseaux de capteurs, Oh *et al.* [Oh *et al.*, 2004] proposent d'explorer la combinatoire d'association à l'aide de « Monte Carlo Markov Chain Data Association » (abrégé MCMCDA) et ont proposé un algorithme MCMCDA et prouvé que la distribution converge vers l'optimum.

4.2.1.2 Suivi monoculaire par logique différée

Ces techniques d'associations de données sont appliquées au domaine de la Vision avec des associations de *tracklets*, définies pour la première fois dans les

travaux de Stauffer [Stauffer, 2003]. Il définit une *tracklet* en réalisant des correspondances image à image. Ensuite, ces *tracklets* sont associées sur une fenêtre temporelle par une heuristique, dans une matrice d'association étendue prenant en compte les initialisations et terminaisons de pistes. L'appariement est résolu par la méthode hongroise. Kaucic *et al.* dans [Kaucic *et al.*, 2005] et Huang *et al.* dans [Huang *et al.*, 2008] emploient un formalisme de suivi très similaire. Sur un formalisme proche, Xing *et al.* obtiennent dans [Xing *et al.*, 2009] leurs *tracklets* en faisant évoluer des filtres à particules locaux. Dans la même veine que les travaux précédents, l'appariement de *tracklets* est réalisé par méthode hongroise.

De leur côté, Berclaz *et al.* réalisent dans [Berclaz *et al.*, 2006] une optimisation de l'appariement avec l'algorithme de Viterbi, alors que Leibe *et al.* proposent dans [Leibe *et al.*, 2007] un couplage de type programmation quadratique booléenne entre détections et suivis, et une résolution avec un algorithme dérivant de « Expectation-Maximisation ».

Mais c'est finalement l'algorithme MCMCDA ([Pasula *et al.*, 1999, Oh *et al.*, 2004]) qui retient le plus l'attention, notamment avec les travaux de Yu *et al.* [Yu *et al.*, 2007] sur l'association de données à partir de blobs de soustraction de fond, pour générer des cibles de suivi. Cette approche est encore étendue par Ge *et al.* dans [Ge et Collins, 2008]. Là où Yu *et al.* recourent à un programme linéaire, Ge *et al.* intègrent le calcul des paramètres du modèle dans le formalisme bayésien de l'approche. Finalement, dans les travaux récents, Benfold et Reid proposent dans [Benfold et Reid, 2011] un suivi multi-cibles temps réel, modulo la durée de la fenêtre temporelle, utilisant un MCMCDA sur des détections HOG [Prisacariu et Reid, 2009].

Par rapport à la démarche que nous suivons dans ce manuscrit, il est à noter que ces méthodes adressent par essence le suivi de cibles multiples et, en accord avec la classification du chapitre 3, elles sont toutes centralisées. Nous retenons de ce panel de travaux sur le MOT monoculaire que la littérature privilégie la méthode hongroise associée à une formalisation de *tracklets* et MCMCDA par rapport aux autres approches.

Cependant nos investigations portent ici sur le réseau. Nous souhaitons réaliser des associations de données entre plusieurs capteurs, sur la base des filtres MSR présentés au chapitre 3. Nous focalisons donc maintenant sur notre contexte applicatif : le suivi dans les réseaux de caméras à champs de vue disjoints.

4.2.2 Suivi à partir d'observations discontinues : réseaux à champs disjoints

Dans un réseau de caméras, lorsque les trajectoires de cibles présentent des discontinuités dues à leur non-observabilité, *e.g.* entre différentes caméras non recouvrantes, l'application de la ré-identification devient une nécessité. Les travaux

de Huang et Russel dans [Huang et Russell, 1997] figurent parmi les premières formalisations de ce problème. Ils ont formalisé une ré-identification de véhicules en réalisant des associations dans un espace probabilisé construit sur des cibles de tailles et couleurs moyennes similaires. Ensuite, Pasula *et al.* [Pasula *et al.*, 1999] ont prouvé la convergence de l'échantillonnage MCMC pour explorer la combinatoire d'association, tandis que Kettnaker *et al.* [Kettnaker et Zabih, 1999] ont décomposé le problème multi-caméra en un programme linéaire. Zajdel *et al.* [Zajdel et Kröse, 2005] ont adopté un formalisme proche de celui de Pasula *et al.* mais l'ont résolu à l'aide de réseaux de Bayes dynamiques (DBN pour « Dynamic Bayes Network ») pour évaluer la vraisemblance des hypothèses et un algorithme EM pour apprendre les paramètres du modèle.

En termes de réseaux NOFOV, à notre connaissance, Matei *et al.* dans [Matei *et al.*, 2011] et Kuo *et al.* dans [Kuo *et al.*, 2010b] sont les seuls à essayer d'unifier MOT et modalités de ré-identification telles que présentées au chapitre 2 dans un système complet. Matei *et al.* se positionnent sur le suivi de véhicules et traitent des réseaux tels que celui présenté en figure 4.2. Le cas de véhicules en contexte urbain induit un modèle de mouvement linéaire et la supposition d'une vitesse constante. Ces contraintes leur permettent de formaliser un MHT utilisant cinématique et apparence.

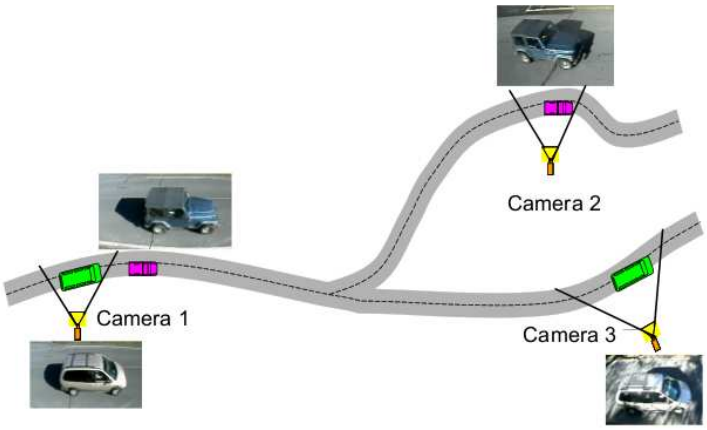


FIGURE 4.2 – Illustration du réseau de 3 caméras privé de Matei *et al.* dans [Matei *et al.*, 2011]. .

Au contraire, Kuo *et al.* traitent le cas de piétons. La figure 4.3 donne un exemple de leur contexte applicatif. Ils adoptent une ré-identification similaire à celle de [Gray et Tao, 2008] en entrainant un boosting sur des paires de silhouettes correctement appariées entre deux caméras. Le système proposé travaille en deux phases : tout d'abord, des contraintes spatio-temporelles d'apparition dans les caméras permettent de générer des appariements de silhouettes. Ensuite, un algo-

FIGURE 4.3 – Illustration du réseau de 3 caméras privé de Kuo *et al.* dans [Kuo *et al.*, 2010b]. L'apparence des personnes est relativement similaire dans chacune des caméras, et la topologie du réseau est linéaire, avec des caméras seulement reliées à leur voisine directe.

rithme de type MIL-boost (pour « Multiple Instance Learning-boost ») apprend les caractéristiques invariantes dans le changement de caméra tout en tolérant du bruit dans la labellisation, *i.e.* dans les appariements de silhouettes donnés comme exemples positifs. Le suivi est réalisé par intégration temporelle de *tracklets*. A la fin de la séquence, les classifieurs de ré-identification sont utilisés avec l'algorithme Hongrois [Burgeois et Lasalle, 1971] pour apparier entre les caméras. La méthode est automatique, bien que la phase de labellisation repose sur des contraintes très faibles en contexte difficile. Par ailleurs, la ré-identification est vue comme un problème de Maximum A Posteriori (MAP), et résolu en énumérant toutes les possibilités d'associations. L'explosion combinatoire d'association est un frein à l'extensibilité d'une telle méthode.

4.2.3 Notre approche

Ré-identification dans un réseau À ce niveau, il convient de rappeler le parallèle que nous avons établi entre REID et MOT dans le chapitre d'introduction. En effet, les deux se présentent comme une mise en correspondance temporelle de silhouettes décrites de manière appropriée. Cependant, la REID est une version relaxée du MOT, car les contraintes spatio-temporelles liant les silhouettes à apparier en REID sont beaucoup plus faibles. Le problème de la REID se décompose donc en deux sous-problèmes.

- ▷ Quelles informations de la silhouette seront invariantes entre différentes caméras, comment les obtenir, et quelle distance utiliser pour la comparaison ? En bref, comment générer une fonction de similarité ayant du sens entre plusieurs caméras ? Le chapitre 2 disserte sur ce problème.
- ▷ Comment la connaissance du réseau peut-elle contraindre l'appariement à réaliser, et ainsi simplifier la combinatoire d'association ?

Lorsque les identités évoluent dans le réseau et que l'on se pose la question de « qui est qui ? », on est rapidement amené à comparer des observations provenant de caméras différentes. Il y a alors deux approches possibles, que nous avons présentées au chapitre 2. L'apprentissage de caractéristiques invariantes en-

tre deux caméras est la méthode fournissant les meilleures résultats sur les bases d'images classiques. Toutefois, nous avons écarté cette classe d'approches car il faut disposer d'exemples d'entrainement labellisés, *i.e.* pour qui le problème de la ré-identification a été résolu. La seconde approche est la ré-identification directe, avec la comparaison d'une signature fixe. Là encore, nous avons un problème de cohérence lorsque l'on compare des signatures provenant de caméras différentes.

Ré-identification récursive versus globale Lorsque l'on passe à un réseau de caméras, le problème de la ré-identification se complexifie. En effet, il n'y a plus un seul changement de caméra à considérer. Toutefois, la littérature [Kuo *et al.*, 2010b, Matei *et al.*, 2011, Chen *et al.*, 2008] considère toujours les ré-identifications caméra à caméra. Ce faisant, ils réalisent des ré-identifications locales et n'ont à aucun moment de notion d'identité dans le réseau. Nous parlons alors de ré-identification récursive, car la trajectoire d'une identité peut être reconstruite en mettant bout à bout les trajectoires intra caméra, pour finalement remonter au moment où elle est entrée dans le réseau. Comme toute approche récursive, la moindre erreur fausse l'ensemble.

Par opposition à ceci, nous qualifions notre approche de globale. Nous avons une notion d'identité dans le réseau, définie par la base d'identité. Les filtres MSR se comparent toujours à la caméra d'entrée. Ainsi, les estimations d'identités des filtres MSR sont indépendantes. L'obtention de la trajectoire d'une identité dans le réseau n'est en aucun cas récursive ici, et nous avons à chaque instant une estimation de sa position dans le réseau.

Or dans cette partie supervision, nous avons à comparer des observations disparates. Pour conserver toutefois l'homogénéïté, nous travaillons non pas sur des comparaisons de descripteurs, mais sur les résultats de filtres de suivi et identification. En ce sens, nous nous démarquons des approches classiques e.g. [Kuo *et al.*, 2010b, Matei *et al.*, 2011, Chen *et al.*, 2008]. Nos filtres MSR établissent des comparaisons à une base issue de la même caméra, et produisent une distribution de probabilité sur l'ensemble des identités testées, sans prendre en considération ni l'historique d'apparition de cette identité, ni les réponses des autres filtres.

De par cette étude bibliographique représentative des méthodes existantes en logique différée (synthétisée par le tableau 4.1), nous tirons les enseignements suivants : (i) l'unification MOT/REID au sein d'un réseau de caméras est un sujet encore peu exploré et (ii) parmi les méthodes classiques d'associations de données, la méthode hongroise associée à une formalisation par *tracklets* et le MCMCDA sont celles qui rencontrent le plus de succès.

Forts de ces constats, nous proposons deux formalismes pour notre superviseur, l'un utilisant une résolution par méthode hongroise, et l'autre utilisant le MCMCDA. Tous deux se basent sur les résultats des filtres MSR, et sur notre définition

	MOT	REID	MOT+REID
MHT	[Cox et Hingorani, 1996][Danchick et Newnam, 2006]	[Matei *et al.*, 2011]	
Agrégation + Hungarian	[Kaucic *et al.*, 2005] [Huang *et al.*, 2008] [Xing *et al.*, 2009]		[Kuo *et al.*, 2010b]
MCMCDA	[Yu *et al.*, 2007] [Ge et Collins, 2008] [Benfold et Reid, 2011]	[Pasula *et al.*, 1999]	

TABLE 4.1 – Tableau récapitulatif des différents systèmes de MOT par logique différée présentés.

de la topologie du réseau.

Nous proposons dans ce chapitre de superviser ces probabilités d'identités, entre elles, et au regard de leur existence dans la topologie du réseau de caméras. Nous présentons ici deux types de superviseurs. Tout d'abord le MAPT (pour Maximum A Posteriori Trajectoriel) est inspiré de la programmation dynamique. Nous présentons ensuite le superviseur basé MCMC, plus complexe à mettre en oeuvre, mais travaillant sur un horizon temporel plus important.

4.3 Définitions

Avant de formaliser nos principes de supervision, nous définissons les éléments qui vont les caractériser : la modélisation du réseau de caméra, commune aux deux formalisations, et les données manipulées par les superviseurs.

4.3.1 Modélisation du réseau de caméra

Dans cette partie, nous supposons disposer de la topologie du réseau sur lequel nous travaillons. Cette topologie est représentée par un graphe non orienté $G = (V, E)$ où les noeuds V représentent les zones d'entrées/sorties des caméras, et les arêtes E donnent les transitions possibles entre ces zones, comme illustré par la figure 4.4. Dans le cas de notre réseau et compte tenu de la précision de nos traqueurs, les zones d'entrées/sorties peuvent être définies grossièrement par des portions de plans image. Des configurations plus complexes pourraient mériter une modélisation plus fine.

Par ailleurs, cette connaissance *a priori* pourrait être apprise en ligne à l'aide de méthodes telles que [Gilbert et Bowden, 2006] ou [Chen *et al.*, 2008].

Nous localisons dans cette topologie les traqueurs, en regardant dans quelle zone tombe le bas de la boite englobant la personne cible. À ce niveau, la méthode

gagnerait en précision avec un échantillonnage des particules sur le plan du sol. Toutefois, comme notre définition des zones d'entrée/sortie est large, un traitement 2D est suffisant.

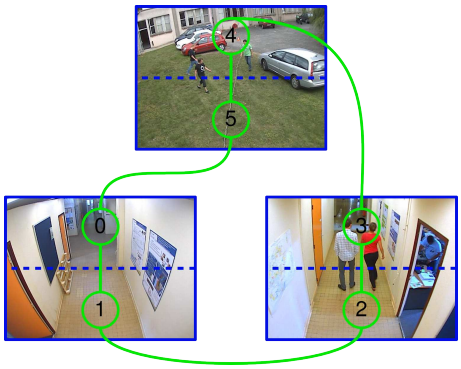


FIGURE 4.4 – Modélisation de la topologie du réseau de caméras. Un graphe non orienté relie les zones d'entrées/sortie des caméras.

4.3.2 Données propres aux superviseurs

Notion de *tracklet* d'identité : En nous basant sur l'étude bibliographique réalisée en section 4.2, nous définissons ici la notion de *tracklet* d'identité, agrégée sur une fenêtre temporelle de durée T. Contrairement à la version métrique, nous désignons ici le paramètre discret représentant l'identité de la piste. Le MAPT optimise ces *tracklets* par méthode hongroise.

Notion de piste : Le superviseur MCMC agit dans une temporalité différente. Il attend que les cibles des traqueurs disparaissent. Ceci génère ce que nous appelons dans la suite une piste de suivi, définie par ses temps de début et de fin, zones de début et de fin, et distribution d'identité.

Notion de chemin : Dans la suite, nous appelons chemin les regroupements de pistes produits par le MCMC. Chaque chemin est associé à une identité. Affecter une piste à un chemin équivaut à lui attribuer l'identité de ce chemin.

4.4 Approche MAP trajectoriel

Cette première méthode est inspirée des travaux de Lev-Tov et Moses dans [Lev-Tov et Moses, 2010]. Ils proposent un filtrage particulaire dont l'état est une position dans le graphe topologique du réseau. Les noeuds du graphe peuvent représenter des caméras, tout comme des zones d'ombre du réseau. Ils formalisent

une vraisemblance pour ces deux types de localisation, et font évoluer les particules dans le graphe pour suivre leurs différentes cibles. Ensuite, ils reconstruisent les chemins empruntés par les cibles avec l'algorithme de Viterbi sur les k meilleures hypothèses de positions. Toutefois dans l'article, la solution est uniquement testée en simulation, avec une simulation des mesures images donnant les vraisemblances dans les noeuds caméras.

Nous retenons de cet article l'idée de lisser temporellement le paramètre d'identité relativement à la topologie du réseau, aux autres cibles présentes et à l'historique de ré-identification. Nous proposons de lisser les ré-identifications grâce à un estimateur type MAP Trajectoriel. Nous commençons par présenter brièvement les principes de la programmation dynamique desquels cet estimateur est inspiré.

4.4.1 Formalisation de la programmation dynamique

La programmation dynamique est une technique algorithmique pour optimiser des sommes de fonctions monotones croissantes sous contrainte. Elle a été désignée par ce terme pour la première fois dans les années 1940 par Richard Bellman. Elle s'applique à des problèmes d'optimisation dont la fonction objectif se décrit comme « la somme de fonctions monotones croissantes des ressources ».

Elle s'appuie sur un principe simple : toute solution optimale s'appuie elle-même sur des sous-problèmes résolus localement de façon optimale. Concrètement, cela signifie que l'on va pouvoir déduire la solution optimale d'un problème en combinant des solutions optimales d'une série de sous problèmes.

Algorithme de Viterbi L'algorithme de Viterbi [Fornay, 1973] permet de corriger, dans une certaine mesure, les erreurs survenues lors d'une transmission à travers un canal bruité. Son utilisation s'appuie sur la connaissance du canal bruité, par le biais d'un modèle de Markov caché (HMM pour « Hidden Markov Model ») qui le définit. Le HMM (un exemple est donné en figure 4.5) est un automate résumant les états non observables possibles, les probabilités de transitions entre états, et les probabilités d'émissions d'observations de ces différents états.

Pour ce faire, l'algorithme de Viterbi énumère les états possibles au cours du temps dans un treillis tels que celui de la figure 4.6. À chaque instant, il choisit l'état le plus vraisemblable en regard des probabilités d'observation et de transition. Le chemin optimal pour la programmation dynamique est construit de manière récursive.

Dans la suite, nous allons voir comment les principes de la programmation dynamique nous permettent de formaliser la notion de *tracklet* d'identité.

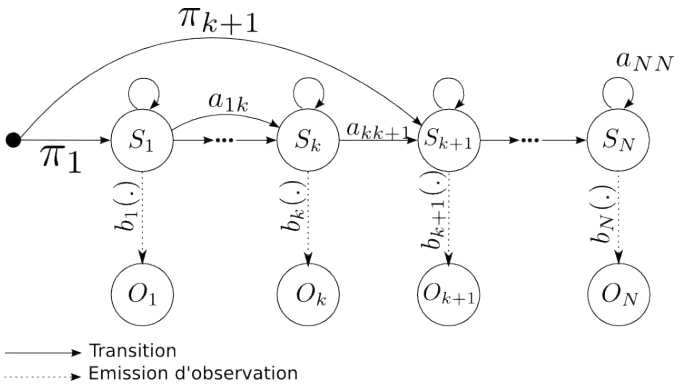


FIGURE 4.5 – Schéma déployé d'un HMM.

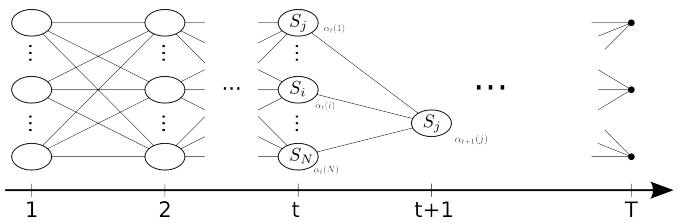


FIGURE 4.6 – Treillis de parcours pour l'algorithme de Viterbi, indicé sur le temps.

4.4.2 MAP trajectoriel : mise en oeuvre

Le chapitre 3 a présenté une stratégie de ré-identification intégrée au suivi et nous avons montré que cette stratégie intégrée est plus performante qu'une stratégie de suivi puis ré-identification. Sa limitation réside toutefois dans le caractère distribué des filtres à état mixte. En effet, les densités de probabilités sur l'identité de la personne suivie sont indépendantes d'un filtre à l'autre. Deux filtres peuvent produire simultanément la même identité. Nous souhaitons ici contraindre ceci de manière à produire un appariement filtre/identité exclusif via leur interaction au niveau du réseau.

4.4.2.1 Intégration temporelle

Chaque traqueur produit à chaque instant une distribution discrète de probabilités sur l'ensemble des identités, calculée comme le ratio de particules dédiées à une identité. Ces probabilités sont agrégées sur une fenêtre temporelle notée T dans un formalisme de type Programmation Dynamique. À l'instar de [Stauffer, 2003] nous parlons alors de *tracklet*. Cependant ici, ces *tracklets* concernent les identités.

L'utilisation de la topologie du réseau intervient à ce niveau. Elle permet de

supprimer les associations traqueur/identité impossibles. Nous partons d'une localisation initiale des identités dans le réseau. À chaque terminaison de traqueur, nous mettons à jour cette localisation avec sa ré-identification. Nous utilisons cette localisation pour mettre à zéro les associations incohérentes avec celle-ci. Une association est dite incohérente si la zone d'entrée/sortie dans laquelle se trouve le traqueur n'est pas connexe avec la dernière localisation enregistrée de l'identité testée.

Dans notre cas, les états S_k sont les identités, les probabilités de l'état sont données par les distributions MSR :

$$P(id_t = S_i|Y) = \text{Card}\left(\Upsilon_{tr,S_i}\right),$$

où, en accord avec les notations adoptées au chapitre 3, Υ_{tr,S_i} représente l'ensemble de particules du traqueur tr ayant l'identité S_i. Nous sommes toutefois dans un cas particulier de Viterbi, car nous considérons des identités de cibles. Ainsi, il n'est pas réaliste pour une cible de changer d'identité. Les probabilités de transition se voient donc imposer de demeurer dans l'état où l'automate est :

$$P(id_t = S_i, id_{t+1} = S_j|Y) = \begin{cases} 1 & \text{si } S_i = S_j \\ 0 & \text{sinon.} \end{cases}$$

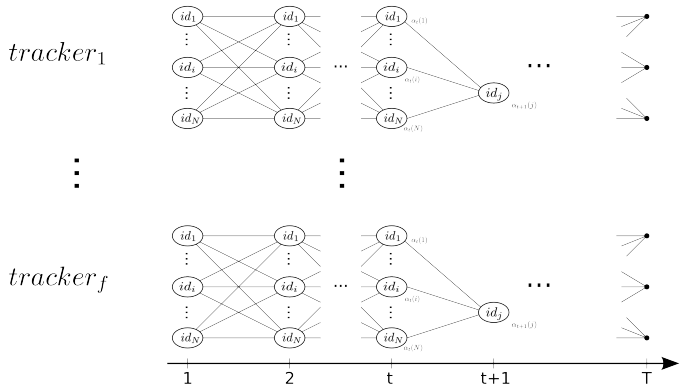


FIGURE 4.7 – Chaque filtre MSR construit un treillis sur l'ensemble des identités sur lesquelles il échantillonne.

4.4.2.2 Exclusivité de l'association

Une affectation exclusive par méthode hongroise [Burgeois et Lasalle, 1971], travaillant à partir de la fonction de similarité (4.1) permet d'obtenir une association

exclusive traqueurs/identités en fin de fenêtre temporelle. Ce score d'association est calculé pour toute paire possible traqueur/identité, sur une fenêtre temporelle de taille T. La topologie et les ré-identifications précédentes interviennent pour supprimer des possibilités. Finalement, l'association traqueurs/détections impose une exclusivité entre les paires résultantes.

$$S(tr_{t_0+T}, id_{t_0+T}) = p\left(id_{t_0+T} | zone(tr_{t_0})\right) \cdot \sqrt[T]{\prod_{t=t_0+1}^{t_0+T} \text{Card}\left(\Upsilon_{tr,id_t}\right)} \qquad (4.1)$$

$$\text{où } \Upsilon_{tr,id_t} = \left\{ i \mid \mathbf{X}_t^{(i)} = (\mathbf{x}_t^{(i)}, id_t)^\intercal \right\}$$

et où

$$p\left(id|zone(tr)\right) = \begin{cases} 1 & \text{si localization[id]=zone(tr) ;} \\ p_{\mathcal{N}}(d_{topo}(localization[id], zone(tr))) & \text{sinon.} \end{cases}$$

Ici, $zone(.)$ représente la zone courante du traqueur considéré dans le graphe topologique du réseau, et $localization[id]$ représente la dernière zone où l'identité id a été déclarée présente. Par ailleurs, d_{topo} représente la distance topologique sur le graphe du réseau.

La gestion des identités au sein du suivi permet d'éviter les problèmes de combinatoire inhérents à la gestion de cibles multiples et de maintenir constamment à jour la répartition dans le réseau des *identités globales* évoluant dedans.

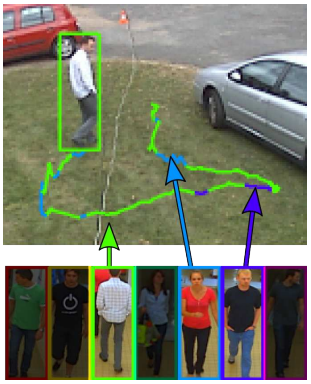


FIGURE 4.8 – Tracklets d'identités au cours d'une séquence de suivi. (meilleur rendu en version couleur)

4.4.2.3 Optimisation des tracklets sur une séquence de suivi

Ces affectations supervisées interviennent à la fin de chaque fenêtre temporelle. Chaque fenêtre temporelle fournit une identification pour toute sa durée. Nous

obtenons ici des *tracklet* d'identités. La figure 4.8 présente les différentes *tracklet* d'identités inférées par le superviseur pour une séquence de suivi. Les différentes couleurs réfèrent à différentes identités de la base.

Pour ne pas biaiser le processus de ré-identification sur le début de la séquence de suivi, nous venons remettre la distribution des identités du filtre à état mixte à l'équiprobabilité à chaque fin de fenêtre temporelle. Ainsi, le processus de recherche d'identité de la cible propose à nouveau toutes les identités de la base, et converge à nouveau vers une en particulier, selon les observations qu'il reçoit.

Pour chacun des traqueurs actifs nous mémorisons ces ré-identifications dans un accumulateur indexé sur les identités de la base. Suivant les principes de la programmation dynamique, à chaque instant l'identité affectée à ce traqueur est le mode prédominant dans cet accumulateur (vote majoritaire). De la même manière, lorsqu'un traqueur s'arrête, il se voit affecter l'identité ayant eu le plus de votes dans son accumulateur, et la localisation dans le graphe topologique de cette identité est mise à jour.

Algorithme 4: Algorithme du MAPT

Données : Topologie, $\{\text{Card}\left(\Upsilon_{tr,id_t}\right), \forall tr, \forall id\}$.
Résultat : Associations traqueurs/identités exclusive.

pour $t = 1$ *à* T **faire**
  Pour chaque traqueur, accumuler les résultats d'identités selon l'équation (4.1) ;
  Ceci revient à pour chaque traqueur, calculer les N meilleurs chemins selon Viterbi ;
fin
En fin de fenêtre temporelle, réaliser une association traqueurs/identités selon la méthode hongroise ;

4.4.3 Bilan du superviseur MAPT

Nous avons présenté ici une première méthode de supervision des réponses des filtres MSR. La décision produite intègre (i) les probabilités de ré-identification des filtres MSR, (ii) un historique de ré-identification dans le graphe topologique du réseau de caméras et (iii) une décision centralisée bloquant toute ubiquité d'identité.

Les limitations d'une telle approche résident dans le caractère incrémental de la méthode. Chaque décision prise est optimale, mais est ensuite entérinée et ne peut être remise en cause par la suite. C'est pour lever cette limitation que nous avons envisagé un superviseur basé sur une approche MCMC.

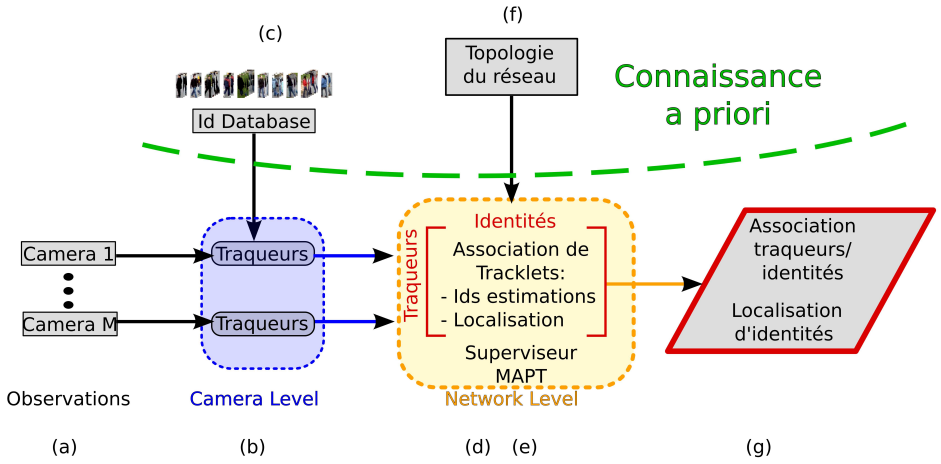


FIGURE 4.9 – Schéma de principe du superviseur MAPT : les observations des caméras (a) sont traitées localement par des traqueurs MSR (b) utilisant la base d'identités (c). Les distributions d'identités estimées par ces traqueurs de manière indépendantes sont agrégées dans une matrice d'association traqueurs/identités, pondérées par les relations topologiques (f) aux dernières localisations des identités dans le réseau (d). Cette matrice est construite sur une fenêtre temporelle et ses composantes représentent nos *tracklets* d'identités. L'association est résolue par méthode hongroise. Ceci constitue le MAPT (e). Il introduit une interaction entre les filtres MSR.

	Paramètre	Notation	Valeur
	Fenêtre temporelle MAPT	T	7 images
	Vraisemblance toplogie	σ_{Topo}	$\sqrt{5}$

TABLE 4.2 – Tableau récapitulatif des différents paramètres libres du MAPT.

4.5 Approche MCMC sur les trajectoires

La première méthode de supervision que nous avons proposée se base sur les principes de la programmation dynamique, *i.e.* fournit le chemin de vraisemblance maximale connaissant tout l'historique de parcours. Ce processus est « simple » à mettre en place, cependant, il ne permet pas de remettre en cause d'éventuelles erreurs. Par ailleurs, le MAPT est calculé au cours des suivis. L'objectif est toujours l'introduction d'interactions au niveau des paramètres d'identité relativement à la topologie, mais nous souhaitons une approche plus globale. Nous proposons maintenant de travailler au niveau des pistes de suivi terminées et de se donner la possibilité de remettre en cause des décisions de ré-identification sur une fenêtre temporelle d'optimisation.

Contrairement à [Matei *et al.*, 2011] qui utilisent un classifieur de ré-identification (*cf.* section 2.4) avec donc une ré-identification récursive et un algorithme MHT pour énumérer les possibilités, nous construisons sur les filtres MSR pour une ré-identification globale et optons pour une association de données MCMC, prouvée comme étant plus efficace par [Oh *et al.*, 2004].

La figure 4.10 illustre le principe de notre superviseur MCMC. Sur la gauche, nous avons les caméras du réseau, dans lesquelles des filtres MSR sont actifs. Sur la droite, nous représentons les pistes de suivi terminées, superposées à la topologie du réseau. Il s'agit des données dont va disposer le superviseur MCMC. La couleur des pistes indique leur identité. En regard de la topologie et des autres pistes, le MCMC va explorer la combinatoire d'association de pistes terminées.

Les méthodes MCMC sont des méthodes générales pour générer des échantillons à partir d'une distribution Π sur un espace Ω en construisant une chaîne de Markov $\mathcal{M}$ ayant ses états $\omega \in \Omega$ et une distribution stationnaire $\Pi(\omega)$.

4.5.1 Association de données MCMC

L'association de données MCMC est issue des travaux de [Pasula *et al.*, 1999] et [Oh *et al.*, 2004].

Nous décrivons maintenant un algorithme MCMC connu sous le nom d'algorithme de Metropolis-Hastings. Si nous sommes dans l'état $\omega \in \Omega$, nous proposons $\omega' \in \Omega$ suivant la distribution de proposition $q(\omega, \omega')$. Le mouvement est accepté

avec une probabilité d'acceptation $A(\omega, \omega')$ où

$$A(\omega, \omega') = \min(1, \frac{\Pi(\omega')q(\omega', \omega)}{\Pi(\omega)q(\omega, \omega')}), \tag{4.2}$$

sinon, l'échantillonneur demeure dans l'état ω. Avec cette construction, la condition d'équilibre est statisfaite i.e., pour tout $\omega, \omega' \in \Omega$ avec $\omega = \omega$,

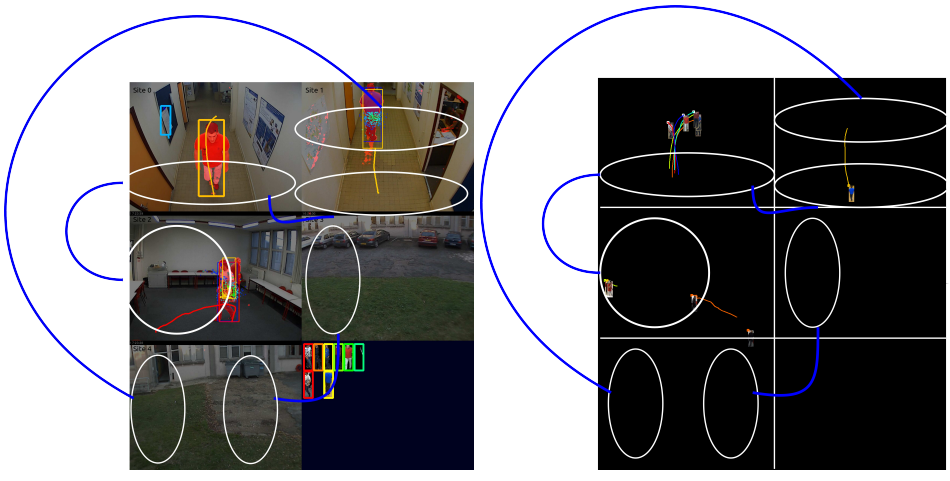
$$Q(\omega, \omega') = \Pi(\omega)P(\omega, \omega') = \Pi(\omega')P(\omega, \omega'), \tag{4.3}$$

où $P(\omega, \omega') = q(\omega, \omega')A(\omega, \omega')$ est la probabilité de transition de ω vers ω' .

L'algorithme 5 résume ceci.

Si $\mathcal{M}$ est irréductible et apériodique, alors, selon le théorème ergodique [Roberts, 1996], $\mathcal{M}$ converge vers sa distribution stationnaire.

L'un des intérêts de l'algorithme est que l'équation (4.2) requiert uniquement



(a) (b)

FIGURE 4.10 – Vues des 5 caméras du réseau NOFOV2 (a) avec des traqueurs actifs et la superposition du graphe topologique du réseau et (b) niveau d'abstraction du superviseur MCMC : la topologie sur laquelle sont représentées les pistes de suivi terminées. Ce sont les identités associées à ces pistes que le MCMC optimise.

la capacité à calculer le ratio $\frac{\Pi(\omega')}{\Pi(\omega)}$, tout en évitant le besoin de normaliser Π.

Algorithme 5: Algorithme de Metropolis-Hastings

Données : Y, n_{mc}, $w^* = w_0$.
Résultat : w^*.

pour $n = 1$ *à* n_{mc} **faire**
  Tirer un mouvement à appliquer à w ;
  Proposer w' selon $q(w'|w)$;
  Tirer U selon Unif[0,1];
  si $U < A(w, w')$ **alors**
    $w_n = w'$
  sinon
    $w_n = w$
  fin
  si $p(w_n|Y) > p(w^*|Y)$ **alors**
    $w^* = w_n$
  fin
fin

4.5.2 Formulation du problème

Soit $Y = \{y_k = (ids_k, t_k^{in}, t_k^{out}, a_k^{in}, a_k^{out}), k = 1, \ldots, K\}$ un ensemble de K pistes générées par nos filtres à état mixte dans les caméras (voir chapitre 3), où ids_k est la distribution d'identités, t_k^{in} et t_k^{out} sont les dates d'apparition et de disparition et a_k^{in} et a_k^{out} sont les zones topologiques d'apparition et de disparition. Contrairement à [Matei *et al.*, 2011], nous suivons des piétons, *i.e.* sans mouvement prédéfini (contrairement à des voitures évoluant sur des voies de circulation définies), induisant une durée de visibilité dans la caméra totalement dépendante de l'individu considéré. Notre approche se base sur ce constat et, plutôt que d'avoir une fenêtre temporelle de durée fixe, attend que le superviseur ait réuni un ensemble de K pistes avant d'entreprendre son optimisation d'association de données.

Nous définissons le problème comme l'inférence de N cibles au niveau du réseau, à partir de l'ensemble des pistes Y, où N est le nombre d'identités évoluant dans le réseau, connu et fixe. L'équation (4.4) résume ceci, où τ_0 est l'ensemble des fausses alarmes, τ_n constitue le n^{e} chemin (associé à une identité) dans le réseau, définie comme une combinaison de pistes dans les caméras.

$$\omega = \{\tau_0, \tau_1, \ldots, \tau_N\} \tag{4.4}$$

Chaque τ_n dans ω est défini comme une combinaison de pistes de caméra. Dans notre cadre de travail, nous définissons le problème du suivi dans le réseau comme un problème de maximisation a posteriori (MAP) de l'assignation des pistes aux

identités, étant donné l'ensemble d'observation Y :

$$\omega^* = \arg\max_{\omega}(p(\omega|Y)) \text{ , où } \omega \sim p(\omega|Y) \propto p(Y|\omega)p(\omega) \tag{4.5}$$

4.5.3 Modèle de vraisemblance

Le modèle de vraisemblance que nous proposons $p(Y|\omega)$ se compose de deux termes : une partie topologique et une partie issue des distributions d'identités des filtres MSR : $p(Y|\omega) = \mathcal{P}_{Topo}(Y|\omega) \cdot \mathcal{P}_{MSR}(Y|\omega)$

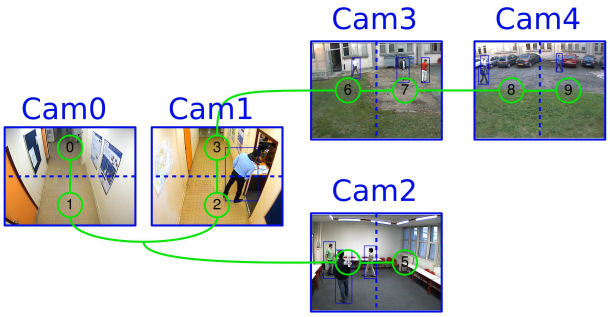


FIGURE 4.11 – Graphe topologique du réseau de test : NOFOV2.

Vraisemblance topologique : Les chemins les plus courts sur le graphe sont obtenus avec l'algorithme de Dijsktra [Dijkstra, 1971]. Comme les caméras sont statiques, la topologie est fixée, et ces distances peuvent être pré-calculées hors ligne et stockées dans une base de données. La figure (4.11) donne un exemple de topologie sur notre réseau privé de test.

$$\mathcal{P}_{Topo}(Y|\omega) = \prod_{i=1}^{|\tau_n|-1} p_{\mathcal{N}}(d_{topo}(a_{i-1}^{out}, a_i^{in})), \tag{4.6}$$

où $d_{topo}(.)$ est la distance entre deux noeuds du graphe topologique, a_i^{in} (*resp.* a_i^{out}) sont les zones de début (*resp.* fin) de la i^{eme} piste, $p_{\mathcal{N}}(.)$ est un noyau gaussien d'écart-type σ_{Topo} transformant la distance en une similarité normalisée entre 0 et 1 et $|\tau_n|$ est le cardinal de l'ensemble de pistes τ_n.

Distributions d'identités : En plus des contraintes topologiques entre pistes, nous ajoutons des caractéristiques d'apparence. Cependant, comparer directement les descripteurs génère un problème d'homogénéïté. En effet des pistes issues de la même caméra peuvent se ressembler plus que des pistes issues de différentes, sans pour autant garantir que les premières représentent la même identité. Il s'agit là du problème de descriptions inter-caméra. À ce stade, Matei *et al.* [Matei *et al.*,

2011] proposent d'utiliser des scores de classifieurs inter-caméras. Supposant que ces scores soient comparables entre eux, nous rejetons toutefois pour le moment une telle méthode, car elle nécessite un entrainement supervisé (voir chapitre 2). À la place, nous utilisons la croyance des traqueurs à état mixte concernant la piste considérée.

$$\mathcal{P}_{MSR}(Y|\omega) = \prod_{i=1}^{|\tau_n|-1} ids_i(id), \tag{4.7}$$

où ids_i est la distribution de probabilité discrète sur l'identité dans la base, pour la i^{e} piste. Ainsi, $ids_i(id)$ représente la probabilité que la piste i ait l'identité id.

4.5.4 MCMC Data Association dirigé par apparence et topologie

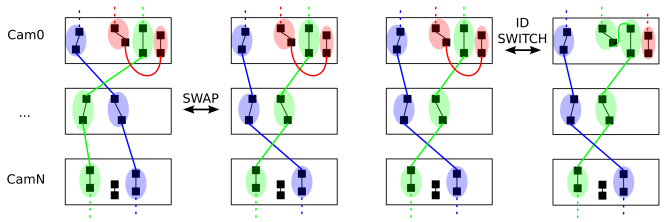


FIGURE 4.12 – Illustration des mouvements de Swap et de Id Switch utilisés dans le MCMCDA entre des caméras à champs disjoints (la temporalité n'est pas représentée dans le schéma).

Nous avons recours à l'algorithme de Metropolis Hastings pour échantillonner à partir de l'équation (4.5). Par définition, une série de mouvements de proposition réversible définit une chaîne de Markov irréductible, apériodique, et qui converge vers une distribution stationnaire selon le théorème ergodique. Dans notre cas, la distribution stationnaire $\pi(\omega)$ est définie par l'équation (4.5.3), et le ratio d'acceptation pour la j^{e} itération est calculé comme

$$p(\omega_j \leftarrow \omega^*) = \min\left(\frac{\pi(\omega^*)q(\omega_{j-1}|\omega^*)}{\pi(\omega_{j-1})q(\omega^*|\omega_{j-1})}, 1\right) \tag{4.8}$$

Les distributions de propositions $q(\omega, \omega')$ consistent en deux paires de mouvements réversibles comme illustré en figure (4.12).

Mouvement Id Switch : Dans un mouvement Id Switch, une piste y_{switch} et un chemin τ_{new} (différent du chemin contenant y_{switch}) sont choisis de manière aléatoire. Le mouvement proposé est un changement d'identité pour la piste y_{switch}. Ce faisant la piste passe d'un chemin à un autre, modifiant les longueurs de ceux-ci.

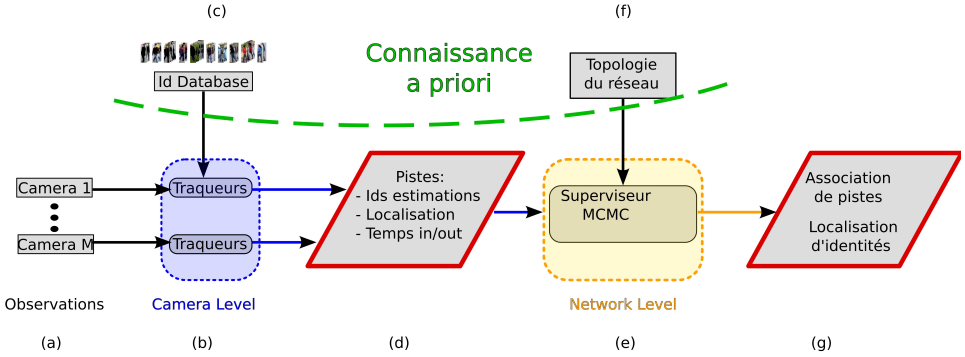


FIGURE 4.13 – Schéma de principe du superviseur MCMC : les observations des caméras (a) sont traitées localement par des traqueurs MSR (b) utilisant la base d'identités (c). Le MCMC attend que les traqueurs terminent de suivre. Ceci génère des pistes (d), représentées par leurs distributions d'identités, leurs localisation en début et fin de piste dans la topologie (f), et leur temps de début et de fin. Lorsque le MCMC a agrégé suffisamment de pistes, il lance son algorithme d'optimisation.

Mouvement Swap : Dans un mouvement de swap, deux pistes y_i et y_j tirées aléatoirement dans deux chemins différents sont échangées.

Le tableau 4.3 synthétise l'ensemble des paramètres libres du système que nous proposons, avec leurs valeurs associées que nous avons fixées empiriquement. Les notations sont les mêmes que dans la présentation de la méthode.

	Paramètre	Notation	Valeur
	Vraisemblance toplogie	σ_{Topo}	$\sqrt{5}$
	Nombre d'itérations	n_{mc}	100000
	Proportion de mouvements switch		0.5
	Proportion de mouvements swap		0.5

TABLE 4.3 – Tableau récapitulatif des différents paramètres libres des superviseurs.

4.6 Évaluations et discussions associées

La figure 4.14 représente la sortie de notre méthode sur la séquence NOFOV1, pour les deux types de supervision, avec les suivis dans les images et la localisation des identités dans la topologie du réseau. Les identités non-observées, sont les identités de la base que les filtres considèrent ne pas être en train de suivre. Ces identités sont localisées dans les zones « blind spots » adjacentes à leur dernière zone d'observation pour affichage (représentées en niveaux de gris dans la carte

du bâtiment). Cette localisation est utilisée pour contraindre les ré-identifications futures (*cf.* section 4.5).

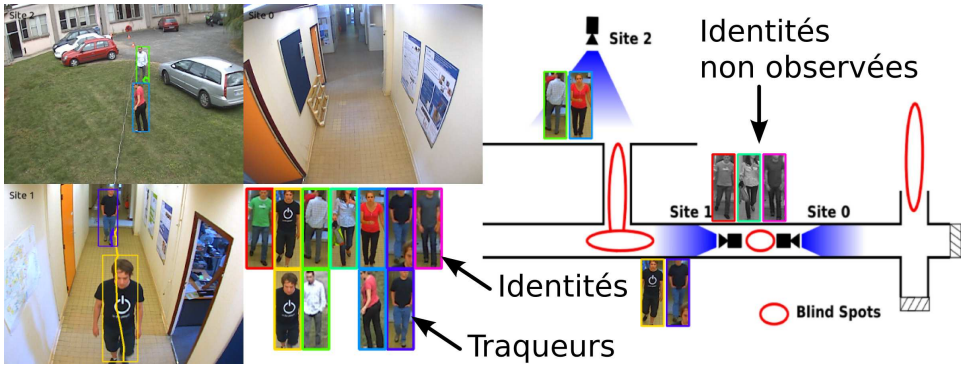


FIGURE 4.14 – Exemple de suivi dans le réseau avec maintien d'*identités globales* sur les cibles, permettant de les localiser dans la topologie du réseau.

4.6.1 Performances du MAPT

4.6.1.1 Performances quantitatives

Le tableau 4.4 présente nos résultats quantitatifs sur PETS'09 pour la stratégie MAPT. Nous comparons les résultats de ré-identification par filtre MSR obtenus au chapitre 3 à l'approche filtres supervisés par le MAPT, dans laquelle l'exclusivité entre les ré-identifications est imposée (section 4.4.2). Cette contrainte d'exclusivité induit de meilleurs scores de ré-identification.

L'aspect stochastique du filtrage particulaire est pris en compte : le tableau 4.4 présente les résultats moyens de chaque score, sur un ensemble de 10 répétitions. L'écart type observé est faible et montre la répétabilité du processus.

Séquence PETS'09	MOTP	MOTA	TRR
Suivi-par-Réidentification	42.5%	77.7%	59.7%
Suivi-par-Réidentification + MAPT	42.4%	75.9%	**64%**

TABLE 4.4 – Résultats de suivi selon les métriques CLEAR MOT [Bernardin et Stiefelhagen, 2008] et taux de ré-identification sur la séquence monocaméra PETS'09 S2L1. Nous donnons ici les Multi-Object Tracking Precision (MOTP), Multi-Object Tracking Accuracy (MOTA), et True Re-identification Rate (TRR) définis au chapitre 3.

Le tableau 4.5 présente les résultats du MAPT sur le réseau NOFOV1 (Suivi-par-Réidentification + MAPT), comparés à l'approche filtres MSR seuls (Suivi-

Séquence NOFOV	cam0	cam1	cam2	réseau
Suivi-par-Réidentification	43.7%	67.3%	55.5%	54.6%
Suivi-par-Réidentification + MAPT	**67.7%**	**76.9%**	**63.8%**	**68.2%**

TABLE 4.5 – Taux de ré-identifications correctes TRR pour chacune des caméras du réseau NOFOV1 : comparaison des approches sans, et avec superviseur MAPT sur le réseau.

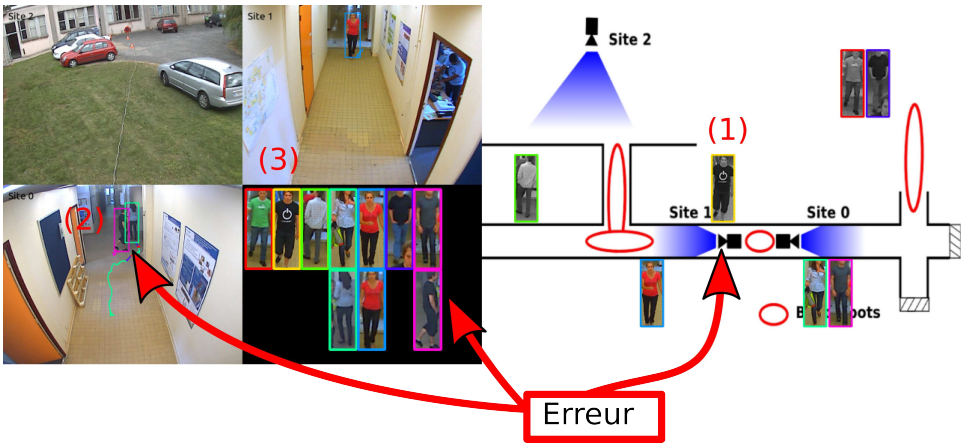


FIGURE 4.15 – Mise en défaut du superviseur MAPT. Nous sommes confrontés ici à une erreur de ré-identification (1), l'identité (2) de la base a été mal localisée dans la topologie (1). Dans le suivi courant (3), ceci décroit la probabilité de l'identité (2) et ainsi force le MAPT à se tromper.

par-Réidentification). L'analyse de ce tableau montre des gains d'au moins 10% selon les caméras, ainsi qu'en moyenne dans le réseau.

4.6.1.2 Limitations du MAPT

Le MAPT est une première solution pour introduire de l'interaction entre nos filtres MSR distribués.

Cependant, le MAPT s'appuie sur la programmation dynamique, *i.e.* il ne remet pas en cause une décision prise. En ce sens, il améliore les résultats du MSR en levant les ambiguïtés entre traqueur, mais introduit aussi une faiblesse par rapport aux erreurs de ré-identification comme l'illustre la figure 4.15. Ce cas présente une erreur de ré-identification au cours d'un suivi car l'identité correcte est supposée être présente dans un noeud différent du graphe topologique.

4.6.2 Performances du MCMC

Au-delà du MAPT, nous avons proposé une formulation MCMC du superviseur, ne travaillant non plus sur des *tracklets*, mais sur des pistes, *i.e.* des séquences de suivi intra-caméras. Le MCMC cherche à répartir les pistes en des chemins les plus cohérents possibles, du point de vue de vraisemblance d'identité et de vraisemblance topologique. Ce faisant, il s'autorise à permuter des pistes, et remettre en cause des ré-identifications déjà effectuées.

4.6.2.1 Tests sur données de synthèse

L'objectif du MCMC est d'imposer une robustesse au bruit sur les distributions d'identités issues des filtres MSR. La mise en oeuvre étant ici plus complexe, nous commençons par valider la formalisation en synthèse. Le but ici est d'évaluer le niveau de bruit tolérable dans la réponse des filtres MSR, permettant à l'algorithme de fournir toutefois une bonne solution. Dans cette partie, nous simulons le graphe d'un réseau et nous l'explorons de manière aléatoire. À chaque intersection, la cible choisit sa destination avec équiprobabilité entre toutes les possibilités. Ainsi nous générons des distributions d'identité aléatoires.

$$ids(i) = \begin{cases} \max(1 - abs(\epsilon_i), 0) & \text{if } i = id_{GT}, \\ \min(abs(\epsilon_i), 1) & \text{else.} \end{cases}$$

Dans les deux cas $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. De cette manière, nous construisons un ensemble de pistes
$Y_{synth} = \{y_k = (ids_k, t_k^{in}, t_k^{out}, a_k^{in}, a_k^{out}), k = 1, \ldots, K\}$ et nous optimisons cet ensemble avec notre algorithme MCMC.

La figure 4.16 présente les résultats pour un réseau de synthèse à 30 caméras, entre 10 et 40 identités évoluant dedans et 100000 itérations de l'algorithme de Metropolis Hastings. Chaque identité apparait environ 20 fois dans le réseau, générant ainsi chacune une vingtaine de pistes.

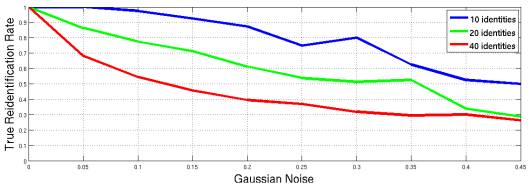


FIGURE 4.16 – Évolution du TRR issu du filtrage MCMC sur données de synthèse, en fonction du bruit gaussien appliqué sur les vecteurs d'identité.

Cette partie valide la formalisation du superviseur MCMC. En effet, même avec des mesures incorrectes dans les distributions d'identités (la valeur maximale

n'étant pas sur l'identité vérité terrain), la connaissance de la topologie permet au superviseur de rattrapper ces erreurs.

4.6.2.2 Tests sur données réelles

Nous présentons les taux de ré-identifications pour la séquence NOFOV2. Nous comparons ici la méthode se basant uniquement sur les informations de couleur introduite dans le filtrage particulaire (résultats du chapitre 3), avec le système supervisé que nous proposons en section 4.5 et son ajout de contraintes topologiques.

Le tableau 4.6 présente les taux de ré-identification par caméra, puis au niveau du réseau global. La base étant construite à partir de la caméra 0, ceci explique les taux de ré-identification supérieurs dans cette caméra. Ces résultats illustrent l'apport du superviseur : chaque identité correctement ré-identifiée contraint le système dans la suite de la topologie.

NOFOV Sequence	cam0	cam1	cam2	cam3	cam4
Suivi-par-Réidentification	88.7%	65.3%	58.5%	54.6%	54.0%
Suivi-par-Réidentification + MCMC	**90.6%**	**76.2%**	**68.2%**	**63.8%**	**62.0%**

TABLE 4.6 – Taux de ré-identifications correctes TRR pour chacune des caméras du réseau NOFOV2 : comparaison des approches sans, et avec superviseur MCMC sur le réseau.

4.7 Conclusion

Dans ce chapitre, deux stratégies de supervision sont proposées. Dans un contexte de vidéosurveillance dans un réseau de caméras fixes, dans lesquelles des filtres MSR distribués sont actifs et suivent et ré-identifient leurs cibles, l'objectif est d'introduire des interactions entre les filtres, au niveau du paramètre d'identité. Cette interaction dans le réseau est rendue possible par l'utilisation de la topologie entre les zones d'entrée/sortie des caméras. Cette topologie vient contraindre *a posteriori* les estimations d'identités par les filtres MSR.

Les stratégies présentées reposent toutes les deux sur la logique différée, avec des temporalités propres. L'interaction d'identités de type MAPT est relativement directe. Elle consiste en l'agrégation sur une fenêtre temporelle des scores de ré-identification des filtres en des *tracklets* d'identités. À la fin de la fenêtre, une décision exclusive d'identification est prise par association filtre/identité à l'aide de l'heuristique gloutonne approximant l'algorithme hongrois.

L'interaction MCMC travaille elle sur des pistes intra caméra. Lorsque le superviseur a réuni suffisamment de pistes, décrites par leurs positions dans le graphe

topologique et leurs distributions d'identités MSR, il lance une optimisation des appariements de pistes. Ce faisant, il remet en cause des décisions de ré-identification face aux croyances des autres filtres, et la continuité topologique des pistes précédentes. Nous avons discuté les avantages de chacune des deux méthodes. Des évaluations quantitatives entre les deux formalisations sont en cours.

Les travaux de ce chapitre ont été publiés dans [Meden *et al.*, 2012b] pour le MAPT et dans [Meden *et al.*, 2012a] pour le superviseur MCMC.

Les questions demeurant non adressées à ce niveau sont :

- ▷ comment construire automatiquement la base d'identités ?
- ▷ comment redescendre de l'information vers les filtres MSR ?
- ▷ et quelles sont les possibilités dégagées par un tel superviseur ?

Le chapitre 5 vient proposer quelques premiers éléments de réponse.

CHAPITRE

5

Vers un système évolutif

Ce chapitre propose des travaux prospectifs visant à étendre la stratégie de ré-identification basée filtrage, présentée dans cette thèse. L'objectif principal exploré ici concerne la redescente de la connaissance haut niveau acquise par le superviseur vers les unités de traitement locales, *i.e.* vers les filtres MSR.

5.1 Introduction

Ce chapitre se veut prospectif et propose des extensions de l'approche que nous avons décrite dans ce manuscrit. En premier lieu, la section 5.2 vient présenter une heuristique de construction de base d'identités, en accord avec la stratégie globale que nous avons adoptée. Les sections suivantes considèrent différentes manières d'utiliser les informations de ré-identification du superviseur. La section 5.3 propose un mode d'échantillonnage des identités par les filtres MSR, lié aux décisions de ré-identification superviseur et à la topologie du réseau. La section 5.4 montre comment les paires d'images appariées (ré-identifiées) par le superviseur permettraient de calculer des BTF entre les caméras du réseau, et comment intégrer cette connaissance supplémentaire dans le système. D'une manière similaire, la section 5.5 considère le cas d'un apprentissage statistique pour la ré-identification, avec des modèles comme ceux proposés dans [Gray et Tao, 2008]. Les sections 5.6 et 5.7, considèrent respectivement les problèmes de reconfigurabilité du réseau, et les ouvertures possibles de la problématique de surveillance par réseaux à champs disjoints. Enfin, la section 5.8 vient conclure ce chapitre de perspectives.

5.2 Construction de la base d'identités

Le système que nous avons présenté jusqu'alors repose sur l'hypothèse de disposer d'images clés des personnes présentes dans le réseau, issues de l'une des caméras. Nous rappelons ici que ceci est cohérent avec l'hypothèse de « closed world », classiquement admise dans les systèmes de ré-identification actuels, et avec le déploiement d'un réseau de caméras, en supposant un point de passage obligatoire pour les personnes amenées à transiter dans le réseau. Ce point de passage sera le lieu de pose de la caméra de construction de la base.

Dans le chapitre 4, nous avons introduit une topologie de réseau reliant zones d'entrée/sorties des caméras. Nous proposons ici une heuristique simple et toutefois efficace pour la construction de la base. En accord avec le paragraphe précédent, nous supposons avoir une caméra d'entrée par laquelle toutes les identités vont entrer dans le réseau. Nous localisons les traqueurs instanciés dans la zone d'entrée correspondante, et utilisons les images de suivi de ces traqueurs pour constituer la base d'identités.

Nous avons défini MSR comme réalisant une ré-identification globale, *i.e.* ayant des images clés des personnes du réseau, issues d'une caméra, et comparant les hypothèses des particules toujours relativement à cette base. Lors de ses comparaisons de w_{Id} dans les équations (3.20) et (3.21), le MSR fait l'hypothèse que les images clés de la base sont toutes issues de la même caméra. Nous pouvons cependant imaginer différentes bases issues de différentes caméras et mises en compétition

au niveau du superviseur.

5.3 Filtrage des échantillonnages d'identité

Dans cette section, nous supposons disposer d'une réponse de ré-identification du superviseur prenant la forme de paires identité/position dans la topologie. Nous posons la question de la redescente d'information, du superviseur vers les filtres MSR. Lorsqu'un filtre MSR est lancé, la distribution initiale du paramètre d'identité est à l'équiprobabilité entre toutes les entrées de la base. Connaissant une estimée des positions topologiques des identités, nous pouvons introduire cet *a priori* dans l'échantillonnage des filtres. Nous proposons de remplacer l'équiprobabilité entre les identifiants par un tirage fonction de la distance du traqueur aux identités dans la topologie.

L'algorithme 6 se base sur les distances topologiques précalculées dans le réseau au chapitre 4. Pour tout nouveau traqueur instancié dans un noeud du réseau, il lui fournit la distribution initiale d'identité correspondant à la position estimée des identités dans le réseau via le superviseur.

Algorithme 6: Algorithme d'échantillonnage initial d'identité dirigé par la topologie et le superviseur.

Données : Localisation topologique des identités : $\{a_{id}\}_{1 \leq id \leq N}$, localisation du traqueur à instancier : a_{tr}.
Résultat : Répartition initiale des identités : $sampling[1-N]$.

pour *tout traqueur* tr *à instancier* **faire**
  Calculer la somme de toutes les distances topologiques, évaluées sous un noyau gaussien :
  $S = \sum_{id=1}^{N} p_{\mathcal{N}}(d_{topo}(a_{tr}, a_{id}))$;
  pour $id = 1$ *à* N **faire**
    $sampling[id] \leftarrow p_{\mathcal{N}}(d_{topo}(a_{tr}, a_{id}))/S$;
  fin
fin

Une telle méthode oriente la distribution d'identité inférée par un filtre MSR vers les identités les plus proches de ce filtre dans la topologie du réseau.

5.4 Projection de la base d'identité par fonctions de transfert de luminance

Nous souhaitons toutefois aller plus loin dans l'utilisation de la connaissance acquise par le superviseur. Au chapitre 2, nous avons dans un premier temps écarté les méthodes supervisées, pour relaxer l'hypothèse contraignante de appariements entre plusieurs caméras pour servir d'exemples d'apprentissage. Nous disposons ici de la réponse du superviseur et nous proposons dans un premier temps d'entrainement une BTF entre les caméras disposant de suffisamment de paires d'entrainement.

Deux approches sont possibles ici :

- ▷ continuer dans notre approche globale et calculer toutes les BTF relativement à la caméra d'entrée. Ce faisant, les filtres MSR bénéficierons d'images corrigées en couleur ;
- ▷ calculer la fonction de passage caméra à caméra pour adopter une stratégie récursive.

Concernant les BTF, étant donné qu'il s'agit d'une correction des conditions d'observation et non pas d'une description en elle même, nous suggérons la première approche. Nous les voyons comme une amélioration de la signature utilisée par les filtres MSR.

Nous calculons les BTF entre la caméra 0 (lieu de construction de la base du MSR) et chacune des autres caméras. La figure 5.1 rappelle les champs de vue de ce réseau et la figure 5.2 présente les BTF calculées pour les trois canaux R, G et B. Nous utilisons la méthode que nous avons présentée dans la section 2.4, qui consiste en le calcul d'histogrammes cumulés d'exemples de personnes vues dans les deux caméras. Ensuite, ces histogrammes cumulés sont mis en correspondance via une matrice de corrélation pour calculer la fonction de passage des niveaux de couleur de l'un à l'autre.

À titre d'exemple, nous présentons en figure 5.3 des silhouettes sur lesquelles ont été appliquées les BTF correspondantes. On note une transformation des niveaux de couleurs.

Dans la section 2.4, nous avons vu que des paires appariées de silhouettes peuvent aussi donner lieu à l'apprentissage d'un modèle de ré-identification.

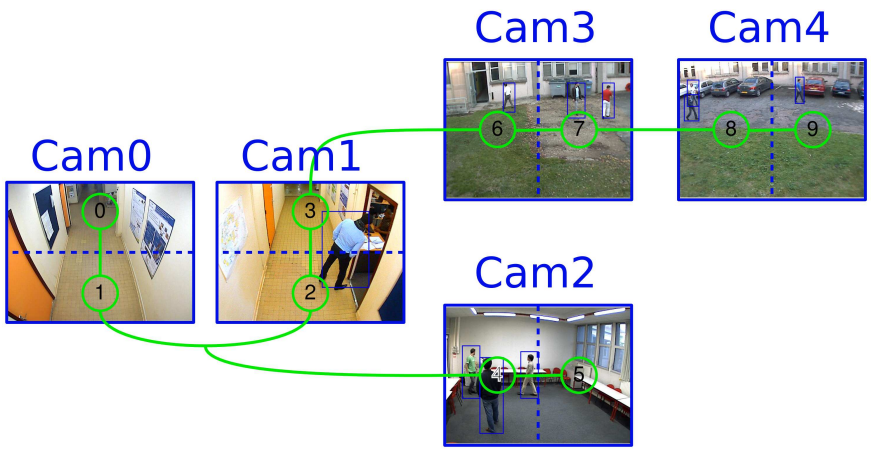


FIGURE 5.1 – Différentes caméras du réseau NOFOV2.

5.5 Apprentissage statistique de modèle

Nous nous basons dans cette section sur les travaux de Kuo et Nevatia dans [Kuo *et al.*, 2010a]. En effet, ils présentent un système de suivi et ré-identification pour réseau de caméras, utilisant un apprentissage statistique caméra à caméra comme descripteur de REID. Ils expliquent que leur système commence par un suivi simple dans chacune des caméras, que les cibles sont mises en correspondance selon une heuristique sur les temps d'apparitions et que ceci leur génère des appariements « faibles ». Pour remédier à l'incertitude des appariements, ils ont recours à algorithme MIL-boost (pour « Multiple Instance Learning boost »), qui ne suppose pas disposer de données parfaitement labellisées [Dietterich *et al.*, 1997, Viola *et al.*, 2006, Babenko *et al.*, 2011].

Étant donné que dans une telle approche, un classifieur doit être entrainé pour chaque paire de caméras adjacentes, une telle heuristique d'appariement initial présentera des difficultés face à la mise à l'échelle. Notre réponse supervisée propose des appariement plus forts.

5.6 Reconfiguration du réseau face à un capteur défaillant

Le problème d'apprentissage non supervisé de relations spatio-temporelles entre les caméras, ou inférence de topologie [Tieu *et al.*, 2005], implique l'estimation des zones d'entrées/sorties des caméras, des probabilités de transitions reliant ces zones, ainsi que les temps nécessaires. Les travaux de Makris *et al.* [Makris *et al.*,

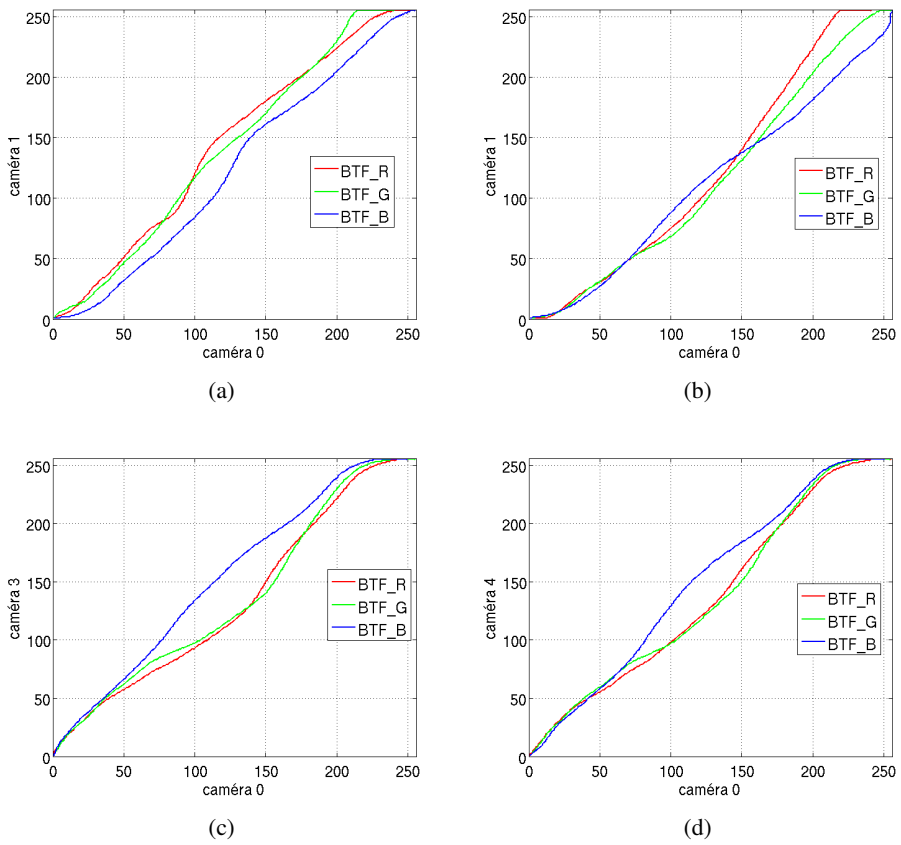
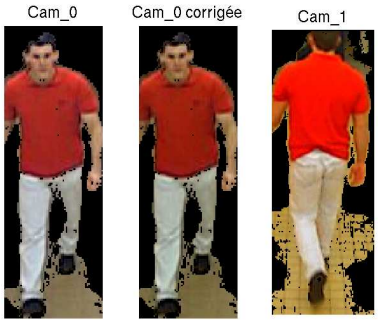
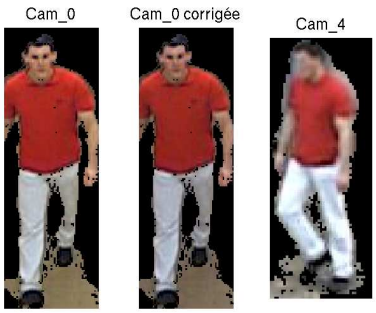


FIGURE 5.2 – BTF calculées sur les caméras du réseau NOFOV2 entre : (a) caméra 0 et 1, (b) caméra 0 et 2, (c) caméra 0 et 3, (d) caméra 0 et 4.

2004] proposent une méthode non supervisée, i.e. ne nécessitant pas de correspondances fournies par l'utilisateur, pour inférer cette topologie. À partir de suivis intra-caméra, les zones d'entrées/sorties sont inférées par mixtures de gaussiennes sur les débuts et fins de pistes, et un calcul de corrélation met les trajectoires en correspondance. Tieu *et al.* [Tieu *et al.*, 2005] poursuivent ces travaux et fournissent une formalisation probabiliste plus poussée des transitions entre zones. Ces méthodes sont non supervisées, i.e. elles permettent d'envisager un déploiement important de capteurs avec une intervention humaine minimaliste en terme de mise en fonctionnement. Toutefois, suite à cette période d'entrainement du système, aucune d'elles ne présentent d'adaptabilité au cours du temps. En réponse à ce problème, Gilbert *et al.* [Gilbert et Bowden, 2006] proposent une stratégie d'apprentissage incrémentale de la topologie du réseau, ainsi que de la fonction de transfert de luminance. Les zones d'entrées/sorties sont estimées par découpage en blocs de plus en plus fins. Ce faisant, l'approche aborde le problème de l'adaptabilité



(a)



(b)

FIGURE 5.3 – Application des BTF aux images de la caméra 0, avec (a) projection vers la caméra 1 et (b) projection vers la caméra 4.

aux changements de conditions d'observations inhérentes aux systèmes actifs sur plusieurs heures/jours. [Chen *et al.*, 2008] poursuivent ces travaux, en modélisant leurs zones d'intérêts par mixture de gaussienne (figure 5.4) à l'instar de [Makris *et al.*, 2004], en ajoutant une composante de mise à jour. Par ailleurs, leur adaptivité aux changements d'illumination requiert moins de données que [Gilbert et Bowden, 2006] et se prête donc mieux à ces changements brusques.

L'objectif de cette partie était d'aborder l'apprentissage d'a priori sur les réseaux à champs disjoints. Nous avons vu que pour les applications visées, de telles méthodes se doivent d'être non supervisées (déployables à large échelle), et adaptatives (permettant des traitements sur de longues durées). Ces méthodes se basent sur des suivis intra-caméra et fournissent des a priori spatio-temporels et colorimétriques permettant de contraindre et réduire la complexité de l'association de pistes entre caméras disjointes.

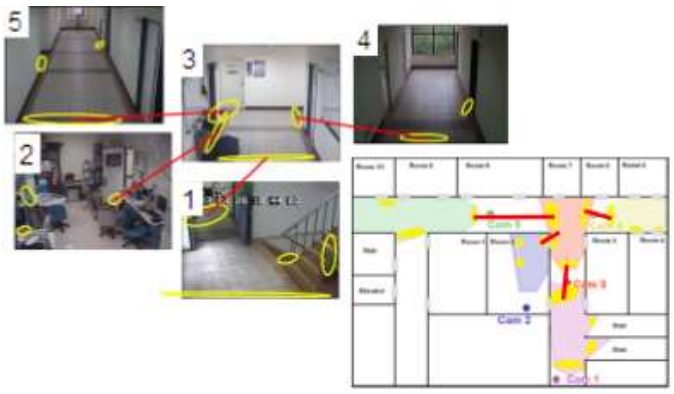


FIGURE 5.4 – Exemple du réseau de test de [Chen *et al.*, 2008] et résultats de l'inférence de zones d'entrées/sorties.

5.7 Extensions des travaux

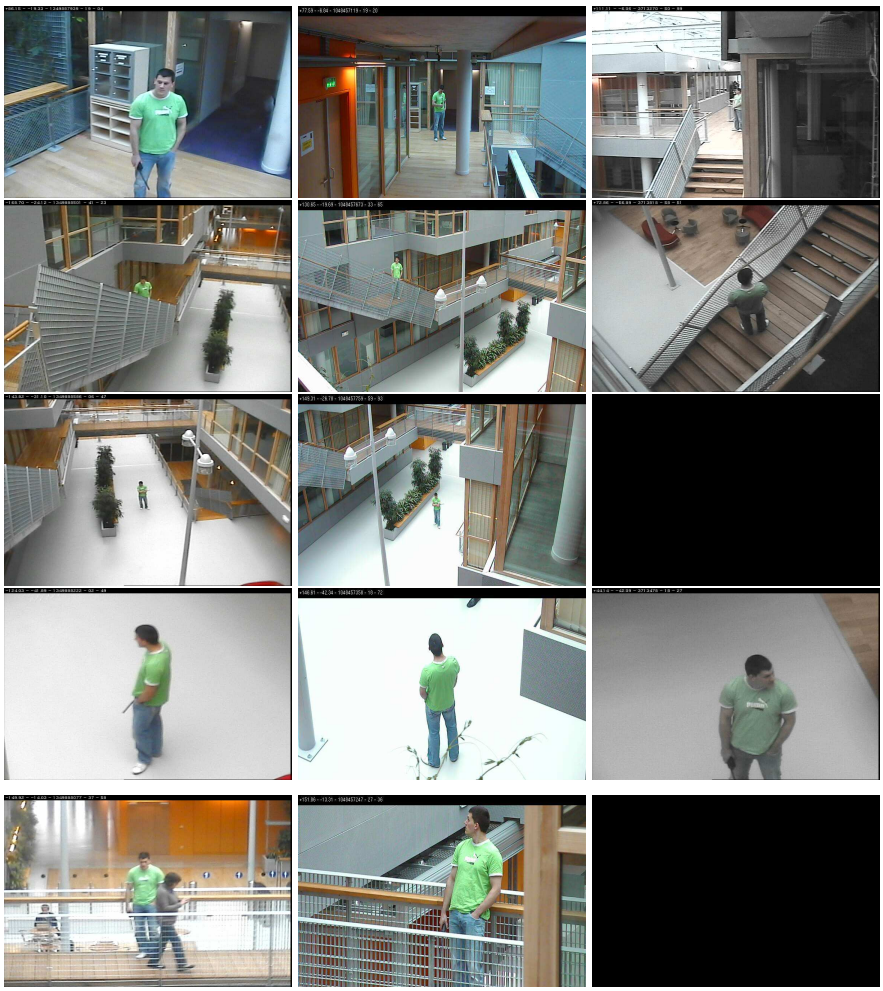
5.7.1 Reconnaissance d'activités / détection d'évènements dans un réseau de caméras

Des travaux récents, [Wang *et al.*, 2010, Loy *et al.*, 2010, Zhu *et al.*, 2012], s'intéressent à l'inférence de « modèles d'activités » à partir de trajectoires reconstruites en réseaux à champs disjoints. [Wang *et al.*, 2010] clusterise les trajectoires de différentes caméras en « activités » selon leurs distributions et leurs directions de mouvements sans chercher à résoudre le problème d'association caméra à caméra. L'approche se base sur des trajectoires intra-caméras d'objets d'intérêts et les regroupe. Les applications visées concernent la surveillance routière et donc permettent l'utilisation d'un module bas niveau de suivi.

A l'inverse, [Loy *et al.*, 2010] s'intéressent à des vidéos ne permettant pas l'utilisation d'un tel module, de par leur densité de personnes et leur faible résolution. Ils adoptent donc une approche de segmentation sémantique de leurs champs de vue, à partir de soustraction de fond.

5.7.2 Au-delà du champ disjoint : utilisation de caméras PTZ

Au-delà des réseaux à champs disjoints, le niveau supérieur en terme de limitation de l'instrumentation pour la surveillance de lieux à large échelle est le recours à des capteurs actifs, tels les caméras PTZ (pour « Pan Tilt Zoom »). En effet, ces caméras sont commandables selon deux axes de rotation et en zoom. La figure 5.5 présente un exemple des possibilités de ces capteurs. Nous montrons ici des vues des trois caméras déployées au laboratoire, pour différentes positions et niveaux de zoom.



(a) (b) (c)

FIGURE 5.5 – Différentes vues d'une personne au sein du réseau de trois caméras PTZ (a) (b) (c), déployé dans le hall du laboratoire, pour différentes positions des caméras et différents niveaux de zoom.

La thèse en cours de P.Paillet vise à explorer des stratégies de coopération entre ces caméras PTZs et nos caméras fixes.

5.8 Conclusion

Ce chapitre propose des extensions de l'approche que nous avons décrite dans ce manuscrit. Ces extensions concernent l'amélioration des performances du système, par l'introduction d'une redescente d'informations vers les traitements bas-niveau, *i.e.* les filtres MSR travaillant dans les caméras.

Ce chapitre mentionne aussi des travaux permettant la maintenance d'un tel réseau, avec la reconfiguration en cas de pannes de capteurs.

Finalement, ce chapitre ouvre le sujet de la surveillance par réseaux de caméras à champs disjoints. Ceci passe *e.g.* par l'inférence de modèles d'activité entre les caméras, ou même par le passage à des capteurs actifs, type caméras PTZ.

CHAPITRE

6

Conclusions et perspectives

Cette thèse s'inscrit dans le contexte de la « vidéosurveillance intelligente », et s'intéresse à la supervision de réseaux de caméras à champs disjoints. Il s'agit là de l'un des cas d'application du problème de la ré-identification de personnes. À ce titre, la thèse propose une approche se démarquant de l'état de l'art qui traite classiquement le problème sous l'aspect description. Nous l'abordons ici sous l'aspect filtrage : comment intégrer la ré-identification de personne dans un processus de suivi multi-cibles, aux niveaux caméras et réseau, de manière à maintenir des identités de cibles cohérentes, malgré des discontinuités dans l'observation de leurs trajectoires, inhérentes à la surveillance par réseaux à champs disjoints.

Plus concrètement, nos travaux visent à fusionner ré-identification et suivi multi-cibles pour surveiller des réseaux de caméras présentant des champs de vue disjoints.

Ce document débute par une introduction présentant le contexte et les objectifs de nos travaux. La problématique de la surveillance par réseaux de caméras est introduite au travers d'un état de l'art général sur le sujet. Ensuite, le contexte spécifique de nos travaux est décliné et nous décrivons notre approche et ses spécificités, justifions les choix faits et présentons les articulations du manuscrit.

Le deuxième chapitre présente un état de l'art relativement exhaustif des modèles de ré-identification basés sur la description de l'apparence. En effet, peu de travaux sont en lien direct avec notre problématique. Cet état de l'art sur les descripteurs était donc nécessaire pour positionner notre approche. Dans ce chapitre, nous distinguons deux grandes classes d'approches : les méthodes ayant recours à un apprentissage supervisé, et les méthodes directes. Les premières seront dans un premier temps écartées en raison des contraintes d'obtention de l'ensemble d'entrainement et notre choix se portera sur une signature non-supervisée comme modèle d'apparence pour notre algorithme de suivi et ré-identification conjoints.

Le troisième chapitre introduit le filtre à état mixte que nous proposons comme solution pour estimer conjointement position et identité des cibles présentes dans les caméras. Il s'agit là d'une méthode markovienne inférant un vecteur d'état

mixte, composé de variables continues -la position de la cible- et d'une variable discrète -l'identité de cette cible-. Plus loin dans le chapitre, ce processus est généralisé pour traiter le cas du suivi multi-cibles. Jusqu'ici, l'approche est mono-caméra, et se base sur une galerie des identités présentes dans la caméra. La donnée de cette galerie représente un *a priori* léger relativement aux méthodes de ré-identification supervisée, et n'est en aucun cas un frein à la mise à l'échelle de la méthode.

Ensuite, le chapitre 4 étend l'horizon de travail en se plaçant au niveau du réseau de caméras. À ce stade, nous bénéficions de filtres distribués à état mixte, inférant position et identité de leurs cibles. Le superviseur vient proposer une mise en interaction des filtres, par le biais de leur paramètre discret. En effet deux cibles ne peuvent avoir la même identité. Le chapitre propose deux modalités de supervision, faisant toutes deux intervenir la topologie du réseau, une donnée *a priori* facilement obtenable, dans l'interaction générée. La première se formule comme un problème de MAP sur une fenêtre temporelle au cours du suivi. Souhaitant se donner la possibilité de remettre en cause d'éventuelles erreurs des ré-identification, nous élargissons ce concept et proposons un second superviseur travaillant lui sur les pistes de suivi terminées et explorant par MCMC la combinatoire d'association inhérente à la ré-identification à ce niveau là.

Le chapitre 5 vient ouvrir le sujet, en proposant différentes perspectives de poursuite de ces travaux. Le fil conducteur de cette thèse a été de fusionner suivi de cibles multiples et ré-identification, pour proposer une application concrète de la ré-identification au domaine de la vidéosurveillance et superviser des réseaux de caméras à champs disjoints. À ce propos, nous avons écarté les approches de ré-identification supervisée comme modalité initiale de description pour notre système. Toutefois, nous montrons dans ce chapitre de perspectives, que le formalisme de superviseur que nous proposons peut servir d'oracle pour la création de l'ensemble de paires d'images associées entre les caméras, nécessaire à l'entrainement de méthodes supervisées. Par ailleurs, nous montrons que notre système s'enrichira de l'application d'une telle méthode supervisée.

Concernant directement nos travaux, plusieurs voies restent encore à explorer. Un certain nombre d'axes de recherche ont été énumérés au chapitre 5. Ces axes concernent l'exploitation de la connaissance haut-niveau du superviseur pour :

- ▷ améliorer l'efficacité de l'échantillonnage d'identité des filtres MSR ;
- ▷ incrémenter l'approche à l'aide de descripteurs supervisés (présentés en section 2.4). Ceci est rendu possible par les appariements réalisés de manière automatique par le superviseur ;
- ▷ adresser l'adaptabilité du système face à des pannes de capteurs ;
- ▷ s'intéresser à des capteurs actifs, en tant que généralisation des réseaux à champs disjoints.

À plus long terme, un tel système doit clairement passer par une phase d'ingé-

nierie pour être déployé sur un réseau étendu dans un démonstrateur, avec des unités de traitement dédiées aux caméras et un superviseur centralisé. Le découpage et la modularité du système, distribué pour le maximum de tâches possibles, ont été réfléchis ainsi, pour ne pas être bloqué par la mise à l'échelle.

Bibliographie

[Achard *et al.*, 2012] ACHARD, C., AMBELLOUIS, S., MEDEN, B., LEFEBVRE, S. et TRUONG CONG, D. N. (2012). *Suivi d'objets d'intérêt dans un réseau de caméras*. Hermès Science Publications, http ://www.editions-hermes.fr/.

[Arulampalam *et al.*, 2002] ARULAMPALAM, M., MASKELL, S., GORDON, N. et CLAPP, T. (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *Signal Processing, IEEE Transactions on*, 50(2):174–188.

[Babenko *et al.*, 2011] BABENKO, B., YANG, M. et BELONGIE, S. (2011). Robust object tracking with online multiple instance learning. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1619–1632.

[Bak *et al.*, 2010] BAK, S., CORVEE, E., BRÉMOND, F. et THONNAT, M. (2010). Person re-identification using spatial covariance regions of human body parts. *In Proceedings of the International Conference on Advanced Video and Signal Based Surveillance*.

[Bazzani, 2012] BAZZANI, L. (2012). *Beyond Multi-target Tracking*. Thèse de doctorat, PhD thesis, Università degli Studi di Verona.

[Benfold et Reid, 2011] BENFOLD, B. et REID, I. (2011). Stable multi-target tracking in real-time surveillance video. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

[Berclaz *et al.*, 2006] BERCLAZ, J., FLEURET, F. et FUA, P. (2006). Robust people tracking with global trajectory optimization. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 744–750. IEEE.

[Bernardin et Stiefelhagen, 2008] BERNARDIN, K. et STIEFELHAGEN, R. (2008). Evaluating multiple object tracking performance : the clear mot metrics. *Journal on Image and Video Processing*.

[Bhattacharyya, 1943] BHATTACHARYYA, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc*, 35(99-109):4.

[Birchfield et Rangarajan, 2005] BIRCHFIELD, S. et RANGARAJAN, S. (2005). Spatiograms versus histograms for region-based tracking. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

[Bolle *et al.*, 2005] BOLLE, R., CONNELL, J., PANKANTI, S., RATHA, N. et SENIOR, A. (2005). The relation between the roc curve and the cmc. *In Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*, pages 15–20. IEEE.

[Breitenstein *et al.*, 2009] BREITENSTEIN, M., REICHLIN, F., LEIBE, B., KOLLER-MEIER, E. et VAN GOOL, L. (2009). Robust tracking-by-detection using a detector confidence particle filter. *In Proceedings of the International Conference on Computer Vision*.

[Breitenstein *et al.*, 2010] BREITENSTEIN, M., REICHLIN, F., LEIBE, B., KOLLER-MEIER, E. et VAN GOOL, L. (2010). Online multi-person tracking-by-detection from a single, uncalibrated camera. IEEE *Transactions on Pattern Analysis and Machine Intelligence*.

[Brèthes *et al.*, 2004] BRÈTHES, L., MENEZES, P., LERASLE, F. et HAYET, J. (2004). Face tracking and hand gesture recognition for human-robot interaction. *In Proceedings of the International Conference on Robotics and Automation*, volume 2, pages 1901–1906. IEEE.

[Burgeois et Lasalle, 1971] BURGEOIS, F. et LASALLE, J.-C. (1971). An extension of the munkres algorithm for the assignment problem to rectangular matrices. *Communications of the ACM*.

[Chen *et al.*, 2008] CHEN, K., LAI, C., HUNG, Y. et CHEN, C. (2008). An adaptive learning method for target tracking across multiple cameras. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

[Cheng *et al.*, 2011] CHENG, D. S., CRISTANI, M., STOPPA, M., BAZZANI, L. et MURINO, V. (2011). Custom pictorial structures for re-identification. *In Proceedings of the British Machine Vision Conference*.

[Comaniciu *et al.*, 2000] COMANICIU, D., RAMESH, V. et MEER, P. (2000). Real-time tracking of non-rigid objects using mean shift. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 142–149. IEEE.

[Cox et Hingorani, 1996] COX, I. et HINGORANI, S. (1996). An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(2):138–150.

[Dalal et Triggs, 2005] DALAL, N. et TRIGGS, B. (2005). Histograms of oriented gradients for human detection. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

[Danchick et Newnam, 2006] DANCHICK, R. et NEWNAM, G. (2006). Reformulating reid's mht method with generalised murty k-best ranked linear assignment algorithm. *In Radar, Sonar and Navigation, IEE Proceedings-*, volume 153, pages 13–22. IET.

[Dee et Velastin, 2008] DEE, H. et VELASTIN, S. (2008). How close are we to solving the problem of automated visual surveillance ? : A review of real-world surveillance, scientific progress and evaluative mechanisms. *Machine Vision Applications*, 19(5-6):329–343.

[Dickinson *et al.*, 2009] DICKINSON, P., HUNTER, A. et APPIAH, K. (2009). A spatially distributed model for foreground segmentation. *Image and vision computing*, 27(9):1326–1335.

[Dietterich *et al.*, 1997] DIETTERICH, T., LATHROP, R. et LOZANO-PÉREZ, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71.

[Dijkstra, 1971] DIJKSTRA, E. (1971). A short introduction to the art of programming.

[Doucet *et al.*, 2001] DOUCET, A., DE FREITAS, N. et GORDON, N. (2001). *Sequential Monte Carlo methods in practice*. Springer Verlag.

[Ess *et al.*, 2007] ESS, A., LEIBE, B. et VAN GOOL, L. (2007). Depth and appearance for mobile scene analysis. *In Proceedings of the International Conference on Computer Vision*, pages 1–8. IEEE.

[Farenzena *et al.*, 2010] FARENZENA, M., BAZZANI, L., PERINA, A., MURINO, V. et CRISTANI, M. (2010). Person re-identification by symmetry-driven accumulation of local features. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

[Felzenszwalb et Huttenlocher, 2005] FELZENSZWALB, P. et HUTTENLOCHER, D. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79.

[Felzenszwalb *et al.*, 2010] FELZENSZWALB, P. F., GIRSHICK, R. B., MCALLESTER, D. et RAMANAN, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

[Fornay, 1973] FORNAY, G. (1973). The viterbi algorithm. *Proceedings of the* IEEE, 61(3):268–278.

[Forssén, 2007] FORSSÉN, P. (2007). Maximally stable colour regions for recognition and matching. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

[Freund et Schapire, 1995] FREUND, Y. et SCHAPIRE, R. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. *In Computational learning theory*, pages 23–37. Springer.

[Frost et Sullivan, 2007] FROST et SULLIVAN (2007). Eyes on the network - understanding the shift toward network-based video surveillance in asia.

[Ge et Collins, 2008] GE, W. et COLLINS, R. (2008). Multi-target data association by tracklets with unsupervised parameter estimation. *In Proceedings of the British Machine Vision Conference*, volume 96.

[Germa *et al.*, 2010] GERMA, T., LERASLE, F., OUADAH, N. et CADENAT, V. (2010). Vision and rfid data fusion for tracking people in crowds by a mobile robot. *Computer Vision and Image Understanding*, 114(6):641–651.

[Germa *et al.*, 2009] GERMA, T., LERASLE, F. et SIMON, T. (2009). Video-based face recognition and tracking from a robot companion. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(03):591–616.

[Gheissari *et al.*, 2006] GHEISSARI, N., SEBASTIAN, T. et HARTLEY, R. (2006). Person reidentification using spatiotemporal appearance. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

[Gilbert et Bowden, 2006] GILBERT, A. et BOWDEN, R. (2006). Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. *In Proceedings of the European Conference on Computer Vision*, pages 125–136.

[Gong *et al.*, 2011] GONG, S., LOY, C. et XIANG, T. (2011). Security and surveillance. *Visual Analysis of Humans*, pages 455–472.

[Gordon *et al.*, 1993] GORDON, N., SALMOND, D. et SMITH, A. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *In Radar and Signal Processing, IEE Proceedings F*, volume 140, pages 107–113. IET.

[Gouaillier et Fleurant, 2009] GOUAILLIER, V. et FLEURANT, A. (2009). Intelligent video surveillance : Promises and challenges. *Technological and Commercial Intelligence Report*.

[Gray *et al.*, 2007] GRAY, D., BRENNAN, S. et TAO, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. *In Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*.

[Gray et Tao, 2008] GRAY, D. et TAO, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. *In Proceedings of the European Conference on Computer Vision*.

[Hamdoun *et al.*, 2008] HAMDOUN, O., MOUTARDE, F., STANCIULESCU, B. et STEUX, B. (2008). Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. *In Proceedings of the International Conference on Distributed Smart Cameras*, pages 1–6. IEEE.

[Handschin, 1970] HANDSCHIN, J. (1970). Monte carlo techniques for prediction and filtering of non-linear stochastic processes. *Automatica*, 6(4):555–563.

[Handschin et Mayne, 1969] HANDSCHIN, J. et MAYNE, D. (1969). Monte carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering†. *International Journal of Control*, 9(5):547–559.

[Hearst *et al.*, 1998] HEARST, M., DUMAIS, S., OSMAN, E., PLATT, J. et SCHOLKOPF, B. (1998). Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28.

[Herold *et al.*, 2011] HEROLD, C., GENTRIC, S., MOËNNE-LOCCOZ, N. *et al.* (2011). Suivi de la pose 3d du visage en environnement multi-caméras avec un modèle tridimensionnel individualisé. *In Actes du congrès francophone des jeunes chercheurs en vision par ordinateur (ORASIS)*.

[Hu *et al.*, 2004] HU, W., TAN, T., WANG, L. et MAYBANK, S. (2004). A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C : Applications and Reviews*, 34(3):334–352.

[Huang *et al.*, 2008] HUANG, C., WU, B. et NEVATIA, R. (2008). Robust object tracking by hierarchical association of detection responses. *In Proceedings of the European Conference on Computer Vision*, pages 788–801. Springer.

[Huang et Russell, 1997] HUANG, T. et RUSSELL, S. (1997). Object identification in a bayesian context. *In Proceedings of the International Joint Conference on Artificial Intelligence*.

[Ilyas *et al.*, 2010] ILYAS, A., SCUTURICI, M. et MIGUET, S. (2010). A combined motion and appearance model for human tracking in multiple cameras environment. *In Emerging Technologies (ICET), 2010 6th International Conference on*, pages 198–203. IEEE.

[Isard, 1998] ISARD, M. (1998). *Visual motion analysis by probabilistic propagation of conditional density*. Thèse de doctorat, Department of Engineering Science, University of Oxford.

[Isard et Blake, 1996] ISARD, M. et BLAKE, A. (1996). Contour tracking by stochastic propagation of conditional density. pages 343–356. Springer.

[Isard et Blake, 1998a] ISARD, M. et BLAKE, A. (1998a). A mixed-state CONDENSATION tracker with automatic model-switching. *In Proceedings of the International Conference on Computer Vision*.

[Isard et Blake, 1998b] ISARD, M. et BLAKE, A. (1998b). Condensation-conditional density propagation for visual tracking. *International journal of computer vision*, 29(1):5–28.

[Isard et Blake, 2001] ISARD, M. et BLAKE, A. (2001). BraMBLe : a Bayesian multiple blob tracker. *In Proceedings of the International Conference on Computer Vision*.

[Jafri et Arabnia, 2009] JAFRI, R. et ARABNIA, H. (2009). A survey of face recognition techniques. *journal of Information Processing Systems*, 5.

[Javed *et al.*, 2008] JAVED, O., SHAFIQUE, K., RASHEED, Z. et SHAH, M. (2008). Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162.

[Javed *et al.*, 2005] JAVED, O., SHAFIQUE, K. et SHAH, M. (2005). Appearance modeling for tracking in multiple non-overlapping cameras. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

[Javed et Shah, 2008] JAVED, O. et SHAH, M. (2008). *Automated multi-camera surveillance : algorithms and practice*, volume 10. Springer-Verlag New York Inc.

[Joachims, 2002] JOACHIMS, T. (2002). Optimizing search engines using clickthrough data. *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM.

[Jojic *et al.*, 2009] JOJIC, N., PERINA, A., CRISTANI, M., MURINO, V. et FREY, B. (2009). Stel component analysis : Modeling spatial correlations in image class structure. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 2044–2051. IEEE.

[Kaucic *et al.*, 2005] KAUCIC, R., PERERA, A., BROOKSBY, G., KAUFHOLD, J. et HOOGS, A. (2005). A unified framework for tracking through occlusions and accross sensor gaps. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

[Kettnaker et Zabih, 1999] KETTNAKER, V. et ZABIH, R. (1999). Bayesian multi-camera surveillance. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

[Khan *et al.*, 2005] KHAN, Z., BALCH, T. et DELLAERT, F. (2005). Mcmc-based particle filtering for tracking a variable number of interacting targets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(11):1805–1819.

[Kim *et al.*, 2008] KIM, D., KIM, D. et PAIK, J. (2008). Model-based gait recognition using multiple feature detection. *In Advanced Concepts for Intelligent Vision Systems*, Lecture Notes in Computer Science.

[Kuhn, 1955] KUHN, H. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*.

[Kuo *et al.*, 2010a] KUO, C., HUANG, C. et NEVATIA, R. (2010a). Inter-camera Association of Multi-target Tracks by On-Line Learned Appearance Affinity Models. *In Proceedings of the European Conference on Computer Vision*.

[Kuo *et al.*, 2010b] KUO, C., HUANG, C. et NEVATIA, R. (2010b). Inter-camera association of multi-target tracks by on-line learned appearance affinity models. *In Proceedings of the European Conference on Computer Vision*.

[Layne *et al.*, 2012] LAYNE, R., HOSPEDALES, T., GONG, S. et MARY, Q. (2012). Person re-identification by attributes. *In Proceedings of the British Machine Vision Conference*.

[Leibe *et al.*, 2007] LEIBE, B., SCHINDLER, K. et VAN GOOL, L. (2007). Coupled detection and trajectory estimation for multi-object tracking. *In Proceedings of the International Conference on Computer Vision*, pages 1–8. IEEE.

[Lev-Tov et Moses, 2010] LEV-TOV, A. et MOSES, Y. (2010). Path recovery of a disappearing target in a large network of cameras. *In Proceedings of the International Conference on Distributed Smart Cameras*.

[Loy *et al.*, 2010] LOY, C., XIANG, T. et GONG, S. (2010). Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106–129.

[MacCormick, 2000] MACCORMICK, J. (2000). *Probabilistic models and stochastic algorithms for visual tracking*. Thèse de doctorat, PhD thesis, University of Oxford.

[Madden *et al.*, 2007] MADDEN, C., CHENG, E. et PICCARDI, M. (2007). Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Machine Vision and Applications*, 18(3):233–247.

[Maggio *et al.*, 2012] MAGGIO, S., HAUGEARD, J.-E., MEDEN, B., LUVISON, B., AUDIGIER, R., BURGER, B. et PHAM, Q.-C. (2012). *Suivi d'objets d'intérêt dans une séquence d'images*. Hermès Science Publications, http ://www.editions-hermes.fr/.

[Makris *et al.*, 2004] MAKRIS, D., ELLIS, T. et BLACK, J. (2004). Bridging the gaps between cameras. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

[Matei *et al.*, 2011] MATEI, B., SAWHNEY, H. et SAMARASEKERA, S. (2011). Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

[Meden *et al.*, 2012a] MEDEN, B., LERASLE, F. et SAYD, P. (2012a). MCMC Supervision for People Re-Identification in Nonoverlapping Cameras. *In Proceedings of the British Machine Vision Conference*, Guildford, Angleterre.

[Meden *et al.*, 2012b] MEDEN, B., LERASLE, F. et SAYD, P. (2012b). Tracking-by-Reidentification in a Non-Overlapping Field of View Camera Network. *In VISAPP*, Rome, Italie.

[Meden *et al.*, 2013] MEDEN, B., LERASLE, F. et SAYD, P. (2013). Suivi par réidentification dans un réseau de caméras à champs disjoints. *Traitement du Signal (à paraître)*.

[Meden *et al.*, 2011a] MEDEN, B., SAYD, P. et LERASLE, F. (2011a). Mixed-state condensation pour suivi et ré-identification simultanés dans des réseaux de caméras à champs de vue disjoints. *In Actes du congrès francophone des jeunes chercheurs en vision par ordinateur (ORASIS)*, Praz-sur-Arly.

[Meden *et al.*, 2011b] MEDEN, B., SAYD, P. et LERASLE, F. (2011b). Mixed-State Particle Filtering for Simultaneous Tracking and Re-Identification in Non-Overlapping Camera Networks. *In Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, Ystad, Suède.

[Meden *et al.*, 2012c] MEDEN, B., SAYD, P. et LERASLE, F. (2012c). Suivi par ré-identification dans un réseau de caméras à champs disjoints. *In Actes du congrès francophone sur la Reconnaissance des Formes et l'Intelligence Artificielle (RFIA)*, Lyon.

[Norris *et al.*, 2004] NORRIS, C., M., M. et WOOD, D. (2004). Editorial : The growth of cctv : A global perspective on the international diffusion of video surveillance in publicly accessible space. *Surveillance Society*, 2(2/3):110–135.

[Nummiaro *et al.*, 2003] NUMMIARO, K., KOLLER-MEIER, E. et VAN GOOL, L. (2003). An adaptive color-based particle filter. *Image and Vision Computing*, 21(1):99–110.

[Oh *et al.*, 2004] OH, S., RUSSELL, S. et SASTRY, S. (2004). Markov chain monte carlo data association for general multiple-target tracking problems. *In CDC*.

[Okuma *et al.*, 2004] OKUMA, K., TALEGHANI, A., DE FREITAS, N., LITTLE, J. et LOWE, D. (2004). A boosted particle filter : multitarget detection and tracking. *In Proceedings of the European Conference on Computer Vision*.

[Orwell, 1949] ORWELL, G. (1949). *Nineteen Eighty-Four*. Secker and Warburg.

[Pasula *et al.*, 1999] PASULA, H., RUSSELL, S., OSTLAND, M. et RITOV, Y. (1999). Tracking many objects with many sensors. *In Proceedings of the International Joint Conference on Artificial Intelligence*.

[Pérez *et al.*, 2002] PÉREZ, P., HUE, C., VERMAAK, J. et GANGNET, M. (2002). Color-based probabilistic tracking. *In Proceedings of the European Conference on Computer Vision*, pages 661–675, Copenhaguen, Denmark.

[Porikli, 2003] PORIKLI, F. (2003). Inter-camera color calibration by correlation model function. *In Proceedings of the International Conference on Image Processing*, volume 2, pages II–133. IEEE.

[Prisacariu et Reid, 2009] PRISACARIU, V. et REID, I. (2009). fasthog - a real-time gpu implementation of hog. Rapport technique 2310/09, Department of Engineering Science, Oxford University.

[Prosser *et al.*, 2008] PROSSER, B., GONG, S., XIANG, T. et MARY, Q. (2008). Multi-camera matching using bi-directional cumulative brightness transfer functions. *In Proceedings of the British Machine Vision Conference*.

[Prosser *et al.*, 2010] PROSSER, B., ZHENG, W., GONG, S., XIANG, T. et MARY, Q. (2010). Person Re-Identification by Support Vector Ranking. *In Proceedings of the British Machine Vision Conference*.

[Qu *et al.*, 2007] QU, W., SCHONFELD, D. et MOHAMED, M. (2007). Distributed bayesian multiple-target tracking in crowded environments using multiple collaborative cameras. *Int. Journal EURASIP*.

[Reid, 1979] REID, D. (1979). An algorithm for tracking multiple targets. *Automatic Control, IEEE Transactions on*, 24(6):843–854.

[Roberts, 1996] ROBERTS, G. (1996). Markov chain concepts related to sampling algorithms. *Markov chain Monte Carlo in practice*, 57.

[Rubin *et al.*, 1988] RUBIN, D. *et al.* (1988). Using the sir algorithm to simulate posterior distributions. *Bayesian statistics*, 3:395–402.

[Schwartz et Davis, 2009] SCHWARTZ, W. et DAVIS, L. (2009). Learning discriminative appearance-based models using partial least squares. *In Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*, pages 322–329. IEEE.

[Shi et Tomasi, 1994] SHI, J. et TOMASI, C. (1994). Good features to track. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 593–600. IEEE.

[Smith *et al.*, 2005] SMITH, K., GATICA-PEREZ, D. et ODOBEZ, J. (2005). Using particles to track varying numbers of interacting people. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

[Stauffer, 2003] STAUFFER, C. (2003). Estimating tracking sources and sinks. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition, Workshop*, volume 4, pages 35–35. IEEE.

[Stauffer et Grimson, 1999] STAUFFER, C. et GRIMSON, W. (1999). Adaptive background mixture models for real-time tracking. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*, volume 2. IEEE.

[Tieu *et al.*, 2005] TIEU, K., DALLEY, G. et GRIMSON, W. (2005). Inference of non-overlapping camera network topology by measuring statistical dependence. *In Proceedings of the International Conference on Computer Vision*. IEEE Computer Society.

[Vermaak *et al.*, 2003] VERMAAK, J., DOUCET, A. et PÉREZ, P. (2003). Maintaining multimodality through mixture tracking. *In Proceedings of the International Conference on Computer Vision*, pages 1110–1116. Ieee.

[Viola et Jones, 2001] VIOLA, P. et JONES, M. (2001). Rapid object detection using a boosted cascade of simple features. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–511. IEEE.

[Viola *et al.*, 2006] VIOLA, P., PLATT, J. et ZHANG, C. (2006). Multiple instance boosting for object detection. *Advances in neural information processing systems*, 18:1417.

[Wang *et al.*, 2010] WANG, X., TIEU, K. et GRIMSON, E. (2010). Correspondence-free activity analysis and scene modeling in multiple camera views. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(1):56–71.

[Wu et Nevatia, 2007] WU, B. et NEVATIA, R. (2007). Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. 75(2):247–266.

[Xing *et al.*, 2009] XING, J., AI, H. et LAO, S. (2009). Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 1200–1207. IEEE.

[Yilmaz *et al.*, 2006] YILMAZ, A., JAVED, O. et SHAH, M. (2006). Object tracking : A survey. *Acm Computing Surveys (CSUR)*, 38(4):13.

[Yu *et al.*, 2007] YU, Q., MEDIONI, G. et COHEN, I. (2007). Multiple target tracking using spatio-temporal markov chain monte carlo data association. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

[Zajdel et Kröse, 2005] ZAJDEL, W. et KRÖSE, B. (2005). A sequential bayesian algorithm for surveillance with nonoverlapping cameras. *International Journal of Pattern Recognition and Artificial Intelligence*.

[Zheng *et al.*, 2009] ZHENG, W., GONG, S. et XIANG, T. (2009). Associating groups of people. *In Proceedings of the British Machine Vision Conference*, volume 7.

[Zheng *et al.*, 2011] ZHENG, W., GONG, S. et XIANG, T. (2011). Person re-identification by probabilistic relative distance comparison. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

[Zhu *et al.*, 2012] ZHU, X., GONG, S. et LOY, C. (2012). Comparing visual feature coding for learning disjoint camera dependencies. *In Proceedings of the British Machine Vision Conference*.