DOCTORAT EN CO-ACCREDITATION
TELECOM SUDPARIS ET L'UNIVERSITE EVRY VAL D'ESSONNE

Spécialité : Informatique

Ecole doctorale : Sciences et Ingénierie

Présentée par

Mohamed IBN KHEDHER

Pour obtenir le grade de
DOCTEUR DE TELECOM SUDPARIS

Ré-identification de personnes à partir des séquences vidéo

Soutenue le : 01/07/2014 devant le jury composé de :

Rapporteurs :	Mme Catherine Achard	UPMC - France
	Mme Najoua Ben Amara	ENISo - Tunisie
Examinateurs :	Mr Bruno Defude	Télécom SudParis - France
	Mr Patrick Sayd	CEA-List - France
	Mr Bogdan Stanciulescu	MINES-ParisTech - France
Encadrant:	Mr Mounim El Yacoubi	Télécom SudParis - France
Directrice de thèse:	Mme Bernadette Dorizzi	Télécom SudParis - France

Thèse N° 2014TELE0018

Résumé

De nos jours, un grand nombre de caméras sont installées dans des lieux privés et publics pour faire face à l'augmentation de la délinquance et de la criminalité. L'exploitation automatique de l'énorme quantité de données collectées est un défi actuel. Cette thèse s'inscrit dans le contexte de l'analyse automatique de la vidéo surveillance et s'intéresse au problème de la ré-identification de personnes dans un réseau de caméras à champs disjoints. La ré-identification consiste à déterminer si une personne quitte le champ de vue d'une caméra et réapparait dans une autre. Ce problème est particulièrement difficile car l'apparence de la personne peut changer entre deux prises de vue de manière significative à cause de différents facteurs tels que la variation des conditions d'illumination, la variation des angles de vue et l'imprécision des régions d'intérêt détectées. L'objectif principal de cette thèse est de proposer un système de ré-identification assurant une certaine robustesse aux facteurs de complexité de la ré-identification.

Nous proposons pour cela d'exploiter la nature du mouvement de la personne dans la vidéo pour la ré-identifier. Il s'agit d'une nouvelle approche se démarquant de l'état de l'art qui traite le problème de la ré-identification par des approches fondées sur l'apparence. Plus précisément, la contribution principale de la thèse consiste à profiter de la nature complémentaire de l'apparence de la personne et du style de son mouvement dans la vidéo pour la décrire d'une manière robuste permettant de prendre en compte la complexité de la tâche de ré-identification. Les contributions majeures proposées concernent deux phases du Système de Ré-identification : la description de la personne et l'appariement des primitives.

Nous considérons deux scénarios de ré-identification différant par le degré de complexité des conditions d'enregistrement. Dans le scénario simple, nous étudions la faisabilité de deux approches : une approche biométrique fondée sur la démarche et une approche d'apparence fondée sur des points d'intérêt spatiaux et des primitives de couleurs. Dans le scénario complexe, nous proposons de fusionner des primitives d'apparence et de mouvement afin d'améliorer la robustesse en ré-identification. Le schéma de fusion proposé est fondé sur le calcul de la somme pondérée des vecteurs de votes des deux ensembles de primitives et ensuite l'application de la règle de vote majoritaire. Pour la description, nous décrivons le mouvement par des points d'intérêt spatio-temporels et l'apparence par des points d'intérêt spatiaux. Quant à l'appariement, nous proposons d'utiliser la représentation parcimonieuse comme méthode d'appariement local entre les points d'intérêts (spatiaux ou spatiotemporels). Nous proposons également une analyse permettant d'identifier les sources d'erreurs principales de notre système dans le but de dégager les pistes d'amélioration les plus prometteuses.

Les améliorations proposées ont été évaluées sur des bases publiques. Les résultats atteints par nos approches montrent des performances comparables à celles des systèmes existants.

Mots clés : Ré-identification, mouvement, apparence, fusion, descripteurs image, descripteurs vidéo, description locale, description globale, représentation parcimonieuse.

Abstract

Nowadays, a large number of videosurveillance cameras are installed in both private and public places to deal with the increase in delinquency and crime. The automatic exploitation of the huge amount of collected data is a current challenge. This thesis addresses video surveillance and focuses on the problem of human re-identification through a network of cameras with non overlapping fields of view. Human re-identification is defined as the task of determining if a person leaving the field of one camera reappears in another. It is particularly difficult because of persons' significant appearance can change within different cameras vision fields due to such changes in illumination conditions, variation in viewing angle and the imprecision of regions of interest detected. The main objective of this thesis is to propose a re-identification system ensuring robustness to complexity factors of the re -identification.

The title of the thesis, re -identification of persons from video sequences, mirrors the purpose of the work : we propose to exploit the complementary nature of appearance and motion features to re-identify person. This is a new approach for the re-identification problem that is usually treated by appearance methods only.

In this context, the main contribution of this thesis is to take advantage of the complementary of person's appearance and style of movement that leads to a description that is more robust with respect to various complexity factors. The major contributions proposed in this work include : person's description and features matching.

First we study the re-identification problem and classify it into two scenarios : simple and complex. In the simple scenario, we study the feasibility of two approaches : a biometric approach based on gait and an appearance approach based on spatial Interest Points and color features. In the complex scenario, we propose to exploit a fusion strategy of two complementary features provided by appearance and motion descriptions. We describe motion using spatiotemporal IPs, and use the spatial IPs for describing the appearance. For feature matching, we use sparse representation as a local matching method between interest points. The fusion strategy is based on the weighted sum of matched interest points votes and then applying the rule of majority vote.

Moreover, we have carried out an error analysis to identify the sources of errors in our proposed system to identify the most promising areas for improvement.

Key words : Re-identification, motion, appearance, fusion, image descriptor, video descriptor, local description, global description, sparse representation.

Remerciements

Cette thèse a été réalisée au sein de l'équipe Intermédia du département Électronique et Physique (EPH) de TELECOM SudParis.

J'adresse tous mes remerciements à l'ensemble des personnes qui ont contribué à la réalisation et l'amélioration de mes travaux de thèse ainsi qu'à la rédaction de mon manuscrit.

Je remercie tout particulièrement ma directrice de thèse, Madame Bernadette Dorizzi, qui m'a chaleureusement accueilli dans son laboratoire, et a accepté de diriger ma thèse. Je la remercie pour son investissement et pour son soutien sur le plan scientifique aussi bien que moral. Je la remercie également pour les séminaires enrichissants qu'elle organise au sein de notre équipe.

Je voudrais également remercier Monsieur Mounim A. El Yacoubi qui a encadré ma thèse. Je le remercie de tout mon cœur pour ses précieux conseils, son écoute et son soutien tout au long de ma thèse. Mr Mounim a toujours répondu présent à toutes mes sollicitations malgré son emploie du temps très chargé.

Je remercie tous les membres du jury pour l'intérêt qu'ils accordent à l'évalutation de mon travail de recherche. Je remercie mes rapporteurs Madame Najoua Ben Amara et Madame Catherine Achard d'avoir accepté de lire et juger mon manuscrit. Je remercie également Monsieur Bruno Defude, Monsieur Patrick Sayd et Monsiuer Bogdan Stanciulescu d'avoir accepté d'être membres de mon jury de thèse.

Enfin, merci à tous mes proches en Tunisie et en France pour le soutien qu'ils m'ont apporté durant cette thèse.

A la mémoire de mes grands-parents.

A mes chers parents Brahim et Ommezzine, ceux qui m'ont élevé avec amour et patience. Vous êtes la source de mes joies et le secret de mon courage. Vous serez toujours le modèle de ma vie. Merci pour votre confiance en moi.

A ma femme Houda, celle que je veux chérir, celle qui m'a toujours épaulé et encouragé, et celle qui a partagé avec moi les moments de joie et tristesse durant cette thèse.

A mon fils Mouadh, celui que j'aime. Je le remercie pour son calme durant la préparation de cette thèse ainsi que pour son BAAA (cette belle « phrase » je n'arrive toujours pas à comprendre).

A mes frères et mes soeurs.

A toute la famille Khedher et Jmila .

A tous mes amis à TSP et l'ENSI.

Table des matières

Table des figures

Liste des tableaux

Acronymes

2D	Bi-dimensionnel
3D	Tri-dimensionnel
ACP	Analyse en Composantes Principales
arbre-KD	arbre K-Dimensions
BBF	Best Bin First
BoF	Bag of Features
BTF	Brightness Transfert Function
CASIA	Chinese Academy of Sciences, Institute of Automation
CAVIAR4REID	Context Aware Vision using Image-based Active Recognition for Re-identification
CBTF	Cumulative Brightness Transfert Function
CMC	Cumulative Match Characteristic
CPS	Custom Pictorial Structures
DC	Descente par CoordonnÃ©es
DoG	Difference of Gaussians
ETHZ-REID	Eidgenossische Technische Hochschule Zurich for Re-identification
GLOH	Gradient Location and Orientation Histogram
GMM	Modèlle de Mixture de Gaussiennes
HOF	Histogram of Optical Flow
HOG	Histograms of Oriented Gradients
HsL	signal Hauteur sur Longueur
HSV	Hue Saturation Value
HMM	Modèle(s) de Markov cachée(s) - Hidden Markov Model(s)
i-LIDS-REID	Imagery Library for Intelligent Detection Systems for Re-identification
ISM	Implicit Shape Model
KD	K Dimensional
KNN	K-Nearest Neighbors
LARS	Large Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
LBP	Local Binary Pattern
DCD	Dominant Color Descriptor
LMNN	Large Margin Nearest Neighbor

LoG	Laplacian of Gaussian
MEI	Motion Energy Image
MHI	Motion History Images
MP	Matching Pursuit
OMP	Orthogonal Matching Pursuit
MSCR	Maximally Stable Colour Regions
RCCA	Regularized Canonical Correlation Analysis
RGB	Red Green Blue
RHSP	Recurrent High-Structured Patches
ROI	Region Of Interest
pdf	fonction de densité de probabilité - Probality Density Function
PI	Point d'Intérêt
PRDC	Probabilistic Relative Distance Comparison
PRID-2011	Person Re-ID 2011
PS	Pictorial Structures
SC	Shape-Context
SDALF	System Drive, Accumulation of Local Features
SIFT	Scale Invariant Features Transform
STIP	Space-Time Interest Points
SURF	Speeded Up Robust Features
SUSAN	Smallest Univalue Segment Assimilating Nucleus
SVM	Support Vector Machine
VIPeR	Viewpoint Independant Pedestrian Recognition
YCbCr	Luminance Chrominance bleue Chrominance rouge

Chapitre 1

Introduction

Contents

1.1 Contexte

La vidéosurveillance connaît de nos jours une forte expansion tant sur le plan technologique qu'économique. Elle est devenue l'un des maillons essentiels des politiques de sécurité des gouvernements. Cette évolution répond au besoin de tout citoyen à la sécurité face à l'augmentation de la délinquance et de la criminalité. Les attentats de septembre 2001 aux États-Unis et de 2005 à Londres ont contribué à l'explosion du nombre de caméras. Londres compte à ce jour le plus de caméras de vidéosurveillance par habitant. Prenons le cas de cambriolage des locaux d'habitation en France, il est en hausse de 7% par rapport à 2012 (INHESJ, 2013). Ainsi, le gouvernement français s'est doté des moyens nécessaires pour lutter contre l'insécurité ; la vidéosurveillance en est un. En effet, «*selon le ministère de l'intérieur, le taux d'élucidation des crimes et délits commis sur la voie publique peut en outre être multiplié par deux dans les villes vidéo-protégées*» (LeFigaro, 2009). La vidéosurveillance est désormais nécessaire pour surveiller tant les lieux publics que les lieux privés.

Dans ce contexte, des réseaux de caméras sont installés dans la rue, les centres commerciaux, les transports en commun, les bureaux, les aéroports, les immeubles d'habitation, etc. La Commission Nationale de l'Informatique et des Libertés (CNIL) affirme qu'en 2012, environ 935000 caméras ont été installées en France (CNIL, 2012). Dans la ville d'Evry qui connaît souvent des faits divers de délinquance, régulièrement rapportés dans les médias, 65 caméras scrutent la ville dans l'espérance de définitivement dissuader les criminels et renvoient leurs images sur des écrans dans un centre de contrôle (figure 1.1).

FIGURE 1.1 – Evry se donne les moyens pour lutter contre l'insécurité.

Un système de vidéosurveillance consiste essentiellement alors à surveiller en même temps tous ces écrans. Cependant, l'augmentation du nombre de caméras installées rend difficile la tâche d'exploitation manuelle des données produites par ces caméras. Pour aider les agents de sécurité à explorer ces données, il est donc nécessaire de rendre la tâche de vidéosurveillance plus intelligente en automatisant certaines de ses fonctions. Parmi ces dernières, on cite la détection des objets, la détection des personnes, la reconnaissance des événements et des actions humaines, le suivi des personnes, etc. Une autre application consiste à reconnaître les personnes qui quittent le champ de vue d'une caméra et réapparaissent dans celui d'une autre. Le système de vidéosurveillance doit, alors, être capable de ré-identifier la personne et de continuer le suivi.

1.1.1 Etapes d'un système de ré-identification

Dans cette thèse, nous nous intéressons à la ré-identification des personnes dans un réseau de deux caméras à champs de vue disjoints où le temps entre les enregistrements des deux caméras est court (environ quelques minutes). Un système générique de ré-identification est composé de quatre étapes principales (figure 1.2) : détection des personnes à ré-identifier, extraction des primitives, appariement des primitives et ré-identification des personnes. Nous décrivons dans la suite le principe de chaque étape.

a) **Détection des personnes à ré-identifier :** cette étape dépend énormément de la complexité de la scène. Dans une scène simple (essentiellement une seule personne passe), la personne peut être détectée par une méthode de faible complexité telle que la suppression du fond, la détection du mouvement, etc. Quand la scène devient complexe (essentiellement plusieurs personnes passent), cette étape devient plus difficile et une méthode de suivi des personnes peut être utilisée. Dans cette thèse, nous ne considérons pas ce cas complexe, nous utilisons plutôt des images résultant d'un algorithme de détection des personnes. Dans le scénario complexe, cet algorithme n'est pas parfait et une mauvaise detection des personnes peut se produire. En effet, les images peuvent contenir par exemple plusieurs personnes, seulement une partie du corps, etc.

b) **Extraction des primitives :** cette étape consiste à décrire la personne par un ensemble de primitives approprié au type de la scène. Dans la littérature, ces primitives sont extraites des images et décrivent pour la plupart l'apparence de la personne par une description de couleur, de texture et/ou de forme.

c) **Appariement des primitives** : une fois les primitives extraites des données enregistrées, elles sont appariées en définissant un score de correspondance. L'appariement des primitives peut être fondé sur une méthode supervisée ou non-supervisée.

d) **Ré-identification des personnes :** en utilisant le score de correspondance, le système de ré-identification doit affecter une identité référence à chaque personne test dont l'identité est au départ inconnue.

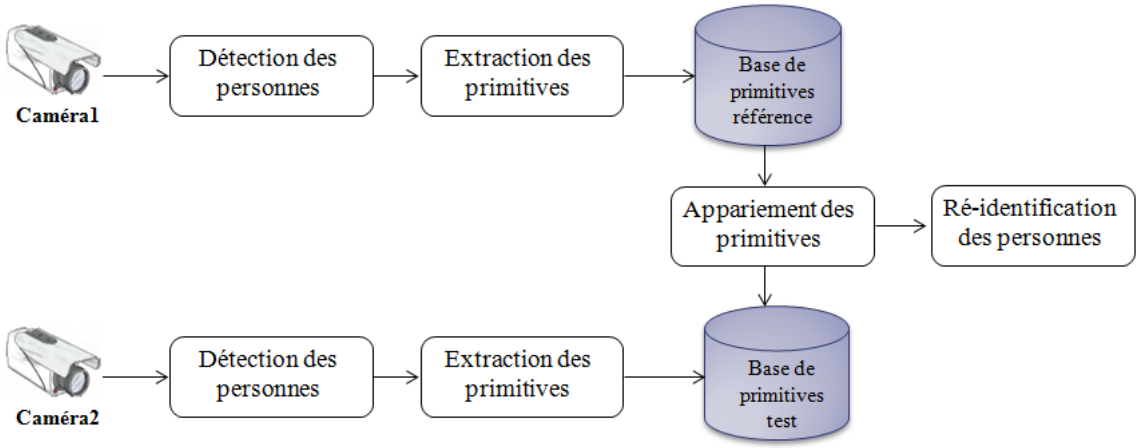


FIGURE 1.2 – Etapes d'un système générique de ré-identification dans un réseau de deux caméras.

1.1.2 Comparaison entre l'identification et la ré-identification

Comme montré précédemment, la ré-identification est similaire à une tâche biométrique car elle consiste à associer une identité référence à une identité test inconnue. La différence majeure entre l'identification et la ré-identification est qu'un système d'identification comprend une phase d'enrôlement où l'identité de la personne est enregistrée alors qu'en ré-identification, on veut reconnaître si une personne passant dans le champ de vue d'une caméra réapparaît dans une autre.

Concernant la difficulté des deux systèmes, le niveau de difficulté de la ré-identification est différent de celui de la tâche d'identification biométrique classique. La table 1.1 montre une comparaison entre l'identification et la ré-identification selon trois critères : nombre d'échantillons référence disponibles par personne, type des conditions d'acquisition et temps entre l'acquisition des données de test et de référence. Les deux premiers critères rendent la tâche d'identification moins complexe alors que le troisième rend la tâche de ré-identification moins complexe.

TABLE 1.1 – Comparaison entre l'identification et la ré-identification.

Critère de comparaison	**Identification**	**Ré-identification**
Echantillons référence par personne	Plusieurs	Un seul
Conditions d'acquisition	Conditions contrôlées (éclairage, fond, etc.)	Conditions aléatoires : peuvent être simples mais sont généralement complexes
Temps entre l'acquisition des données de test et de référence	Beaucoup de temps	Généralement dans le même jour

1.1.3 Différents scenarios de la ré-identification

Selon les conditions d'acquisition des données, la ré-identification peut correspondre à plusieurs scénarios de degrés de complexité différents. On en cite deux : simple et complexe (figure 1.3).

- Dans le scénario simple, des contraintes sont fixées sur le passage de la personne dans le champ de vue de la caméra. De plus, les conditions du milieu d'enregistrement sont contrôlées. En effet, la personne doit marcher seule dans une direction bien définie sans porter aucun objet. La base de données est composée d'un nombre réduit de personnes filmées, à deux instants différents, par une seule caméra fixe dans des conditions d'éclairage stables (la même caméra est utilisée en référence et en test).

- Dans le scénario complexe, on n'a aucune contrainte sur le passage de la personne dans le champ de vue de la caméra. De plus, les conditions du milieu d'enregistrement ne sont pas contrôlées. En effet, plusieurs personnes peuvent passer simultanément en portant ou non des objets de petites dimensions (sac, téléphone, etc.) ou de grandes dimensions (poussette, matériel acheté, etc.). En outre, la base de données est composée d'un nombre de personnes largement plus important que dans le scénario simple ; elles sont filmées par deux caméras installées dans deux endroits différents et dans des conditions d'éclairage non contrôlées.

(a) Scénario simple (b) Scénario complexe

FIGURE 1.3 – Scénarios de la ré-identification.

La table 1.2 résume une comparaison entre le scénario simple et le scénario complexe selon un ensemble de critères.

1.1.4 Problématique du système de ré-identification

En générant une signature (représentation) pour chaque personne filmée par les deux caméras A et B (notons S_A^i et S_B^i les signatures correspondant à la personne i extraites respectivement des caméras A et B), un système de ré-identification doit trouver pour chaque signature S_A^i une signature correspondante S_B^j. (Dans le scénario simple $A = B$, la caméra filme deux angles de vue à deux instants différents).

La résolution de la requête précédente est une tâche très difficile, principalement en raison des conditions d'acquisition non contrôlées qui affectent l'apparence de la personne sous les deux caméras, et à la petite quantité de données disponible par personne et par caméra. Dans la suite, nous présentons les principaux facteurs de complexité de la tâche de ré-identification.

TABLE 1.2 – Comparaisons des scenarios de ré-identification.

Critère de comparaison	**Scénario simple**	**Scénario complexe**
Nombre de caméras d'enregistrement	Une seule	Deux
Conditions d'éclairage	Stable	Variables
Fond	Stable	Variable
Occultations	Pas d'occultation	Occultations fortement possibles
Nombre de personnes	Réduit	Plusieurs dizaines
Direction de la marche	Bien définie	Aléatoire sans aucune contrainte
Angle de vue	Variable	Variable
Porter des objets	Non permis	Possible

1.2 Facteurs de complexité de la ré-identification

La correspondance des signatures des personnes dans un réseau de caméras à champs de vues disjoints est la tâche la plus difficile pour tous les systèmes d'analyse vidéo. En effet, un système de ré-identification doit faire face à la variation de l'apparence causée par différents facteurs tels que la variation des conditions d'illumination, l'occultation, la variation des angles de vue, la variabilité du fond, la pose, la qualité de détection des personnes, etc. (figure 1.4)

Conditions d'illumination :

La variation des conditions d'illumination est un défi critique d'un système de vidéo surveillance. Cette variation peut être due à la diversité des paramètres optiques des caméras ou des conditions d'éclairage du milieu d'enregistrement. Ainsi, les conditions d'illumination changent d'une façon significative l'apparence d'une personne.

Occultations :

Le passage des personnes dans le champ de vue de la caméra sans aucune contrainte peut produire des occultations. Ces occultations peuvent être causées par des objets, par d'autres personnes ou par des structures de l'environnement. En cas d'occultation, si certaines caractéristiques importantes de la personne sont invisibles, sa ré-identification devient difficile et peut échouer.

Changement de l'angle de vue :

Selon l'angle de vue à partir duquel la personne est observée, des parties du corps humain peuvent apparaître différemment. En outre, des parties observées dans une caméra peuvent être cachées dans une autre.

Changement de pose :

Le changement de pose causé par la variation des angles de vue et par l'articulation du corps humain conduit à des apparences significativement différentes de la même personne vue par différentes caméras.

Détection des personnes :

Bien que dans cette thèse, nous ne nous intéressons pas à la détection des personnes avant de les ré-identifier, notons que cette tâche reste un grand défi pour un système de ré-identification. La détection des personnes peut être réalisée par un algorithme statique (Dalal et Triggs, 2005) ou par un algorithme de suivi des personnes (Kalman, 1960). Une mauvaise détection peut être causée par le passage des personnes dans une foule, le bruit du fond, etc. En outre, elle dépend d'autres facteurs tels que les conditions d'illumination, des poses, des angles de vues, etc. La ré-identification d'une personne est difficile si certaines de ses parties sont détectées dans une caméra et non détectées dans une autre.

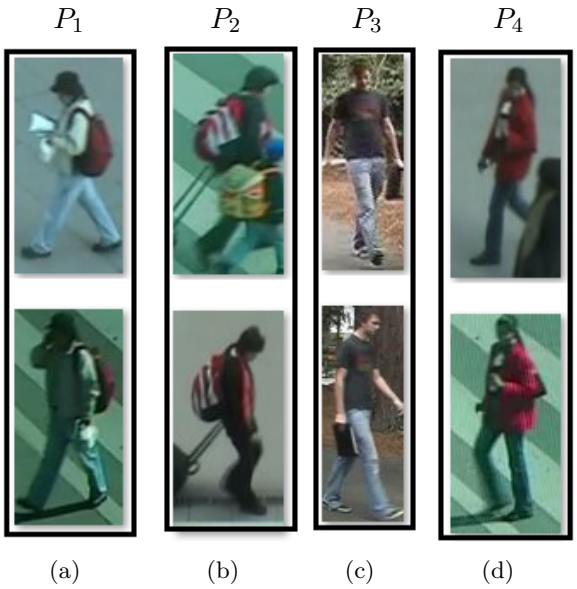


FIGURE 1.4 – Exemple de facteurs de complexité de la ré-identification : (a) conditions d'illumination, (b) occultation, (c) angle de vue et (d) détection des personnes.

1.3 Contributions

Dans le cadre de cette thèse, nous nous intéressons à la ré-identification des personnes dans différents scénarios et par conséquent dans différentes conditions d'acquisition. Chaque scénario présente ses propres facteurs de complexité.

L'objectif de la thèse est de trouver une description de la personne et une méthode d'appariement des descripteurs, qui soient les plus robustes possibles aux différents facteurs de complexité de la ré-identification. Les contributions de cette thèse sont les suivantes :

- L'état de l'art réalisé durant cette thèse a montré que les systèmes de ré-identification sont fondés essentiellement sur des descripteurs d'apparence extraits de l'image. Notre première contribution consiste à exploiter la nature complémentaire de l'apparence de la personne et le style de son mouvement dans la vidéo. De ce fait, nous proposons une description de la personne qui tient compte de l'apparence et du mouvement en même temps.

- La deuxième contribution consiste à proposer une méthode d'appariement des primitives robuste au bruit. L'idée est d'utiliser la représentation parcimonieuse comme méthode d'appariement local entre les points d'intérêt spatiaux liés à l'apparence d'une part et les points d'intérêt spatiotemporels liés au mouvement d'autre part.

- Nous avons étudié la faisabilité d'un système biométrique fondé sur la démarche, dans un scénario simple, et nous avons évalué sa performance selon certains facteurs de complexité de la ré-identification.

1.4 Plan de thèse

La suite de ce manuscrit est organisée en cinq chapitres.

- Le chapitre 2 présente un état de l'art sur la ré-identification des personnes. Tout d'abord, il montre les limites des approches d'identification biométriques standards pour la ré-identification. Puis, il décrit les approches fondées sur l'apparence en les classifiant en approches non-supervisées et approches supervisées. Ensuite, il présente les différentes bases de données utilisées pour évaluer ces approches. Finalement, nous discutons l'approche que nous avons choisie et son positionnement par rapport à l'état de l'art.

- Le chapitre 3 considère l'exploitation de la complémentarité apparence/mouvement pour décrire la personne d'une manière appropriée. Il présente les différentes primitives extraites de l'image (respectivement de la vidéo) pour décrire l'apparence (respectivement le mouvement). Pour chaque type de primitives, deux représentations ont été étudiées : une représentation locale et une représentation globale.

Dans le reste du manuscrit, deux scénarios de ré-identification ont été étudiés : simple et complexe. Pour chacun, un ensemble de primitives décrites dans le chapitre 3 ont été exploitées.

- Le chapitre 4 traite la ré-identification des personnes dan un scénario simple. La première partie étudie la faisabilité d'un système biométrique fondé sur la démarche, essentiellement en fonction de la différence angulaire entre les angles de vue de test et de référence. Ce système décrit implicitement le mouvement de la personne par le modèle stochastique HMM. La deuxième partie se focalise sur la ré-identification par l'apparence. Cette dernière est décrite par les points d'intérêt spatiaux (SURF) et les histogrammes de couleurs (RGB).

- Le chapitre 5 traite la ré-identification des personnes dans un scénario complexe. Il considère la fusion des descripteurs d'apparence et de mouvement pour décrire la personne. Le mouvement est décrit explicitement par les points d'intérêt spatiotemporels ; l'apparence est décrite par les points d'intérêt spatiaux et les histogrammes de couleurs. Quant à l'appariement des points d'intérêt, il est fondé sur la représentation parcimonieuse. En outre, ce chapitre étudie la faisabilité du filtrage des appariements non fiables et identifie un ensemble de sources d'erreurs de l'approche.

- Finalement, le chapitre 6 conclut ce manuscrit et discute les perspectives de cette étude sur le court et long terme.

Chapitre 2

Etat de l'art

Contents

2.1 Introduction

La ré-identification des personnes consiste à suivre une personne dans un réseau de caméras à champs de vue disjoints. Comme vu dans le chapitre précédent, la tâche de ré-identification est difficile et présente un défi lié à la variation de l'apparence de la personne sous les champs de vue des caméras disjoints (variation de poses, caméras à différents paramètres, variation de l'éclairage, etc.). La question qui se pose alors est : comment peut-on décrire d'une manière appropriée une personne observée par deux caméras installées dans deux endroits différents. Les caractéristiques biométriques efficaces pour l'identification de personnes sont écartées dans les applications de vidéosurveillance à cause de ces contraintes d'application. L'alternative dans la littérature consiste à utiliser des informations d'apparence globale de la personne (contrairement aux approches biométriques qui souvent exploitent l'apparence d'une seule partie du corps humain). Le problème de ré-identification dans un réseau de deux caméras (A et B) peut être traité par deux types d'approches. La première catégorie concerne les approches supervisées qui consistent à exploiter des connaissances *a priori* d'une base d'apprentissage filmée par les deux caméras (A et B) où le problème de ré-identification est résolu. La deuxième

catégorie concerne les approches non-supervisées (dite aussi « approches directes ») qui consistent à ré-identifier les personnes sans cette connaissance *a priori* c-à-d sans recours à une base d'apprentissage. Dans la suite du manuscrit, nous appelons *paires d'images positives* : deux images de la même personne prises par les deux caméras A et B ; *paires d'images négatives* : deux images de différentes personnes prises par les deux caméras A et B ; *paires d'images test* : deux images de deux personnes inconnues, n'appartenant pas à la base d'apprentissage et prises par les deux caméras A et B. Dans le reste de ce chapitre, la section 2.2 montre les limitations de la biométrie et son inadéquation à la tâche de ré-identification. Dans la section 2.3, nous présentons une vue générale sur les représentations de la personne et les différentes méthodes d'appariement de ces représentations couramment utilisées dans la ré-identification. La section 2.4 décrit les deux classes d'approches : approches non-supervisées et approches supervisées. La section 2.5 présente les différentes bases de données utilisées pour évaluer les approches de ré-identification. Finalement, nous discutons l'approche que nous avons choisie et son positionnement par rapport à l'état de l'art.

2.2 Ré-identification fondée sur des primitives biométriques

La biométrie est la science qui consiste à identifier les êtres humains par leurs caractéristiques physiques ou comportementales. Généralement, la biométrie est appliquée dans des domaines de sécurité de données, de contrôle d'accès, d'authentification, etc. Parmi ces caractéristiques, on peut citer l'iris, le visage, la démarche, etc. L'identification des personnes par ces caractéristiques biométriques est bien avancée dans la littérature. Par contre, ce type d'approches impose des contraintes pratiques et techniques. Pratiquement, les systèmes d'identification par l'iris et par le visage nécessitent la coopération de la personne à identifier. Celle ci doit regarder la caméra tout en étant suffisamment proche d'elle. Techniquement, ces systèmes nécessitent des images de haute résolution dans le cas de visage ou des capteurs spécifiques dans le cas de l'iris. Quant à un système d'identification par la démarche, bien qu'il ne nécessite aucune coopération de la part des personnes à identifier ; il impose des contraintes sur le mouvement de la personne dans le champ de vue de la caméra (par exemple des contraintes sur la direction de la marche). Donc, il est difficile de l'appliquer dans un contexte de ré-identification. En effet, dans un scénario de ré-identification, durant son passage par le champ de vue de la caméra, la personne peut s'arrêter, changer la vitesse de marche, passer avec des objets, ramasser un objet, marcher dans des directions différentes, faire un demi tour, etc. Autrement dit, dans un scénario de ré-identification, on n'a aucune contrainte sur le comportement de la personne durant son passage par le champ de vue de la caméra.

A notre connaissance, la démarche est le seul comportement biométrique qui a été utilisé explicitement pour la ré-identification (Skog, 2010). Dans ce dernier travail, plusieurs méthodes d'extraction des primitives exploitant la démarche ont été examinées. Ces méthodes ont été testées sur deux bases de données. La première est filmée par des caméras de surveillance placées à l'extérieur, dans une scène urbaine (figure 2.1). La deuxième est filmée par une seule caméra de résolution supérieure à celle de la première (figure 2.2). Le système conçu pour la ré-identification est fondé sur trois étapes : 1) suppression du fond, 2) extraction des primitives où chaque séquence de silhouettes est transformée en une représentation de marche et 3) comparaison des primitives par une méthode simple de classification. Les méthodes testées dans ce travail sont : l'image énergie de la démarche («Gait Energy Image» (Han et Bhanu, 2006)), l'image énergie active («Active Energy

Image» (Zhang *et al.*, 2010)), la transformée de Fourrier pour un volume de silhouette (Yu Ohara et Yagi, 2004) et l'image énergie de différence des images («Frame Difference Energy Image» (Chen *et al.*, 2009)). Les résultats montrent que les performances des méthodes diffèrent beaucoup entre les deux bases de données, et sont toutes plus élevées sur la deuxième base de données filmée par une seule caméra que sur la première. En effet, la deuxième base de données est filmée par une seule caméra et dans des conditions contrôlées, et donc la ré-identification est réalisée dans un scénario simple. Quant à la première base de données, les personnes sont filmées par différentes caméras plus éloignées des personnes que dans le cas précédent, et donc la ré-identification est réalisée dans un scénario plus complexe.

(a) Caméra 1 (b) Caméra 2 (c) Caméra 3

FIGURE 2.1 – Champs de vue des caméras de la base de données 1.

FIGURE 2.2 – Champs de vue de la caméra de la base de données 2.

Nous concluons que la démarche est efficace pour la ré-identification dans un scénario simple où la personne est filmée par la même caméra, lorsqu'elle est vue par des angles de vue proches et que l'on dispose d'une durée de marche suffisante. Dans le cas où on désire ré-identifier des personnes par deux caméras ayant différents paramètres photométriques, installées dans différents endroits ; de plus, si on n'a aucune contrainte sur les comportements des personnes, les caractéristiques biométrique sont écartées.

2.3 Vue générale sur les approches de ré-identification fondées sur l'apparence

Les approches fondées sur une apparence globale de la personne ont moins de contraintes pratiques et techniques. On extrait des primitives liées à l'apparence. Sachant que les personnes sont capturées par plusieurs caméras et dans différentes conditions, il est nécessaire que la description de la personne soit invariante aux paramètres de la caméra, à

l'orientation et au changement d'éclairage. Une approche de ré-identification est définie en répondant à deux questions : 1) comment représenter une personne (caractérisée par une seule image ou un ensemble d'images clés) et 2) comment apparier les représentations des personnes ?

Le problème majeur des approches fondées sur l'apparence est le changement d'éclairage. Pour remédier à ce problème, différentes méthodes ont été proposées dans la littérature.

Dans la suite, on montre comment la problématique de l'éclairage a été traitée dans la littérature. Ensuite, les représentations et les méthodes d'appariement des représentations couramment utilisées sont discutées.

2.3.1 Problématique de changement des conditions d'éclairage dans un réseau de caméra

Le changement des conditions d'éclairage ainsi que l'utilisation de différentes caméras pour une application de ré-identification représente un grand défi. En effet, les images d'une même personne prises par deux caméras différentes peuvent présenter une différence significative au niveau des couleurs. Deux principales méthodes ont été proposées dans la littérature. La première estime une fonction colorimétrique sur une base d'apprentissage (ensemble de paires d'images mises en correspondance *a priori*) en faisant une liaison entre les plages des couleurs des différentes caméras. La deuxième est fondée sur une normalisation des couleurs ou égalisation des histogrammes sans aucune information sur les plages des couleurs des caméras *a priori*.

Concernant les méthodes fondées sur une fonction colorimétrique, (Porikli, 2003) propose une méthode de calibration colorimétrique entre différentes caméras, appelée «Brightness Transfert Function» (BTF). Cette approche est fondée sur une matrice de corrélation entre les histogrammes de couleurs des images d'une même personne prises par deux caméras différentes. (Gilbert et Bowden, 2006) utilise le même principe en intégrant une phase d'apprentissage pour mettre à jour le changement d'illumination entre caméras. Cette méthode est fondée sur une initialisation de la fonction colorimétrique et nécessite un très grand nombre d'images en apprentissage (entre 5000 et 1000). (Javed *et al.*, 2005) et plus tard (Javed *et al.*, 2008) proposent une extension de (Porikli, 2003). D'abord, les auteurs estiment un BTF pour chaque paire d'images positives. Puis, une Analyse en Composantes Principales (ACP) est appliquée pour trouver la meilleure représentation des changements des couleurs entre caméras. Les méthodes précédentes supposent que les deux personnes sont observées depuis deux angles de vue légèrement différents. (Prosser *et al.*, 2008) traitent le cas où les personnes sont observées par des angles de vue différents. Il propose un BTF Cumulatif (CBTF) en appliquant le BTF de (Porikli, 2003) sur l'histogramme de couleur cumulatif de plusieurs images de la même personne prises par la même caméra. Cette technique a été appliquée dans un contexte de ré-identification (Prosser *et al.*, 2008).

La deuxième famille de méthodes est fondée sur la normalisation des couleurs ou égalisation des histogrammes. Pour la normalisation des couleurs, l'espace de couleur RGB (pour «Red–Green–Blue») est souvent utilisé dans (Bauml et Stiefelhagen, 2011, Du *et al.*, 2012, Truong Cong *et al.*, 2010a) ; la couleur d'un canal de chaque pixel est divisée par la somme des couleurs de tous les canaux du même pixel.

$$R_{i,n} = \frac{R_i}{R_i + G_i + B_i}, \quad G_{i,n} = \frac{G_i}{R_i + G_i + B_i}, \quad B_{i,n} = \frac{B_i}{R_i + G_i + B_i}$$

où pour un pixel i donné, R_i, G_i, B_i sont les valeurs des canaux RGB avant la normalisation ; $R_{i,n}$, $G_{i,n}$, $B_{i,n}$ sont les valeurs des canaux RGB après la normalisation.

Une autre technique, dite «grey world normalisation», est appliquée en ré-identification dans (Truong Cong *et al.*, 2009, 2010a). Elle consiste à diviser la couleur d'un canal d'un pixel par la somme des couleurs du même canal de tous les pixels.

$$R_{i,n} = \frac{R_i}{\text{mean}(R)}, \quad G_{i,n} = \frac{G_i}{\text{mean}(G)}, \quad B_{i,n} = \frac{B_i}{\text{mean}(B)}$$

Similaire à la méthode précédente, la normalisation affine est utilisée dans (Bauml et Stiefelhagen, 2011, Truong Cong *et al.*, 2009, 2010a) ; elle consiste à centrer et réduire la couleur d'un canal d'un pixel comme suit :

$$R_{i,n} = \frac{R_i - \text{mean}(R)}{\text{std}(R)}, \quad G_{i,n} = \frac{G_i - \text{mean}(G)}{\text{std}(G)}, \quad B_{i,n} = \frac{B_i - \text{mean}(B)}{\text{std}(B)}$$

où $\text{mean}(x)$ est une fonction qui retourne la moyenne du vecteur x, $\text{std}(x)$ est une fonction qui retourne l'écart type du vecteur x.

Quant aux techniques d'égalisation des histogrammes (Finlayson *et al.*, 2005), elles supposent qu'un changement d'illumination conserve l'ordre de classement des réponses des capteurs (les valeurs des pixels dans notre cas). Elles consistent à convertir l'histogramme H_k d'un canal k en un vecteur M_k mesurant l'ordre de classement de chaque composante de l'histogramme (Equation 2.1). En ré-identification, cette technique est utilisée dans (Bak *et al.*, 2010b, Truong Cong *et al.*, 2009, 2010a).

$$M_k(i) = \frac{\sum_{u=0}^{i} H_k(u)}{\sum_{u=0}^{N} H_k(u)} \tag{2.1}$$

2.3.2 Représentations des personnes

Quelque soit l'approche de ré-identification, elle consiste à représenter la région d'intérêt (ROI pour «Region Of Interest») où se trouve la personne à ré-identifier par des informations liées à son apparence. L'apparence d'une personne est généralement représentée par des caractéristiques couleurs, textures, formes ou combinaison de ces caractéristiques. En ré-identification, la ROI est décrite par ses couleurs et ses textures, séparées ou combinées ; alors que la forme géométrique du corps est toujours utilisée comme une information additionnelle à celle de couleur, de texture ou les deux.

Les caractéristiques couleurs sont fondées généralement sur les histogrammes de couleurs dans différents espaces de couleurs et sur les couleurs dominantes de la ROI. RGB, HSV (pour «Hue Saturation Value») et YCbCr (pour «Luminance, Chrominance bleue et Chrominance rouge») sont les espaces de couleurs les plus fréquents pour décrire une région de l'image (Farenzena *et al.*, 2010, Javed *et al.*, 2008, Berdugo *et al.*, 2010, Gray et Tao, 2008, Prosser *et al.*, 2010, Zheng *et al.*, 2011) ou pour décrire un point d'intérêt (de Oliveira et de Souza Pio, 2009, Gheissari *et al.*, 2006). La représentation par couleur

dominante cherche les couleurs principales de la ROI en regroupant ses pixels en un nombre fixe de régions (Madden *et al.*, 2007, Bak *et al.*, 2010a) ou en cherchant itérativement les régions les plus stables en couleur (Farenzena *et al.*, 2010).

La texture d'une image est définie selon (Sklansky, 1978) comme suit : "une région dans une image a une texture constante si un ensemble de statistiques locales ou d'autres propriétés de la fonction image sont constantes, varient légèrement ou sont presque périodiques". Différentes caractéristiques de textures sont utilisées pour décrire une région de l'image ou un point d'intérêt. On cite l'utilisation fréquente des filtres de Gabor et de Schmid (Prosser *et al.*, 2010, Zheng *et al.*, 2011, Gray et Tao, 2008) et de LBP (pour «Local Binary Pattern») (Hirzer *et al.*, 2012, An *et al.*, 2013) pour décrire une bande horizontale de l'image ; et les primitives de Haar (Hamdoun *et al.*, 2008, de Oliveira et de Souza Pio, 2009) pour décrire un point d'intérêt. Généralement, les caractéristiques de textures diffèrent par leurs invariances aux transformations géométriques et au changement d'éclairage.

La forme géométrique du corps humain est utile et souvent combinée avec la couleur ou la texture pour enrichir sa description. La hauteur de la personne est utilisée comme une information de forme dans (Park *et al.*, 2006). (Wang *et al.*, 2007) modélise la forme géométrique par les distances entre les pixels de la silhouette et un point référence sur la tête.

Pour assurer une description discriminante de la ROI, différentes caractéristiques sont combinées. Dans (Prosser *et al.*, 2010, Zheng *et al.*, 2011, Gray et Tao, 2008), des primitives de couleur (RGB, HS, YCbCr) et de textures (filtre de Gabor et filtre de Schmid) sont combinées pour représenter une bande de la ROI. (Farenzena *et al.*, 2010) cherche pour chaque partie de la ROI les régions les plus stables en couleur et les régions les plus structurées en texture.

Dans la section suivante, nous montrons comment ces représentations sont appariées pour ré-identifier les personnes.

2.3.3 Appariement des représentations de personnes

Une méthode d'appariement des représentations consiste à associer une personne vue par une caméra A à une personne vue par une caméra B. L'appariement des personnes peut être effectué d'une façon globale ou locale. Un appariement global consiste à calculer des similarités entre des images alors que l'appariement local est fondé sur le calcul des similarités entre des caractéristiques locales extraites de l'image. Cette similarité peut être calculée directement à travers les distances usuelles comme elle peut être apprise sur une base d'apprentissage.

Comme mesure de similarité à travers les distances usuelles, (Bak *et al.*, 2010a) mesure la similarité entre les couleurs dominantes de deux images par la distance euclidienne. (Farenzena *et al.*, 2010) décrit la ROI par trois primitives, ensuite il introduit la similarité entre deux ROIs comme la somme pondérée de trois distances (une distance par primitive) : la distance Bhattacharyya pour calculer la similarité entre les primitives fondées sur les histogrammes et la distance euclidienne pour les autres. (Truong Cong *et al.*, 2010b) exploite la représentation parcimonieuse pour apparier des ROIs. Le principe est de représenter la ROI test par une combinaison des ROIs références à l'aide d'un vecteur de

coefficient ; la ROI test est identifiée comme l'identité référence que les ROIs références associées permettent mieux de reconstruire la ROI test. (Hamdoun *et al.*, 2008, de Oliveira et de Souza Pio, 2009) proposent un appariement local fondé sur les points d'intérêt et le calcul des distances. Dans ces deux travaux, deux personnes ne sont appariées que si elles ont suffisamment de points d'intérêt similaires. Pour calculer cette similarité, (Hamdoun *et al.*, 2008) calcule la somme des différences des valeurs absolues des deux descripteurs de points d'intérêt, alors que (de Oliveira et de Souza Pio, 2009) calcule la distance euclidienne entre ces descripteurs.

D'autres auteurs ont visé d'estimer la similarité entre deux personnes sur une base d'apprentissage afin d'apparier une nouvelle personne (qui n'appartient pas à la base d'apprentissage) d'une manière robuste. Dans ce type de méthodes, la base d'apprentissage contient des paires d'images positives et des paires d'images négatives. Dans (Gray et Tao, 2008), la ROI est décrite par plusieurs primitives, ensuite une similarité est calculée grâce à l'algorithme Adaboost appris sur une base d'apprentissage. Cet algorithme d'apprentissage apprend le poids de chaque primitive. Dans (Prosser *et al.*, 2010) inspiré de (Gray et Tao, 2008), les poids des primitives sont appris par la méthode «RankSVM». Dans (Dikmen *et al.*, 2010), l'algorithme du k-plus proche voisin (KNN pour «K-Nearest Neighbors») est utilisé pour l'appariement, il prend comme mesure de similarité la distance de Mahalanobis appris sur une base d'apprentissage. Dans (Zheng *et al.*, 2011), la similarité entre les personnes est définie à partir d'une fonction probabiliste appris sur une base d'apprentissage qui maximise les distances entre les paires d'images négatives et minimise les distances entre les paires d'images positives.

Dans la section suivante, un état de l'art non exhaustif des approches de ré-identification est présenté.

2.4 Classification des approches de ré-identification

Les approches de ré-identification peuvent être regroupées selon plusieurs critères : 1) le nombre d'images par personne, 2) le type de représentation de l'image utilisé pour la ré-identification et 3) l'existence ou l'absence d'un ensemble de paires d'images mises en correspondance *a priori* menant à des approches supervisées et non supervisées.

2.4.1 Approches mono-échantillon vs approches multi-échantillons

Selon le nombre d'images disponibles par personne, on peut classifier les approches de ré-identification en deux familles. La première concerne les approches mono-échantillon où la signature d'une personne (sa représentation) est extraite à partir d'une seule image. Parmi ces approches, on cite (Park *et al.*, 2006, Cai *et al.*, 2008, Wang *et al.*, 2007, Gray et Tao, 2008, Prosser *et al.*, 2010, Zheng *et al.*, 2011, Dikmen *et al.*, 2010, Schwartz et Davis, 2009, An *et al.*, 2013). La deuxième famille concerne les approches multi-échantillons où plusieurs images servent à calculer la signature d'une personne. Parmi ces approches, on cite (Gheissari *et al.*, 2006, Hamdoun *et al.*, 2008, de Oliveira et de Souza Pio, 2009, Bauml et Stiefelhagen, 2011, Jungling et Arens, 2011, Truong Cong *et al.*, 2010b, Huang *et al.*, 2008, Cheng *et al.*, 2011, Farenzena *et al.*, 2010, Souded, 2013, Truong Cong *et al.*, 2010a, 2009, Hirzer *et al.*, 2011).

Cette classification «mono-échantillon vs multi-échantillons», adoptée par quelques auteurs, reste très vague. En fait, souvent, une approche multi-échantillons peut être appliquée dans un contexte où une seule image est disponible. Et souvent, une approche mono-échantillon peut être appliquée dans un contexte multi-échantillons en étendant la mesure de similarité entre deux images en une similarité entre deux ensembles d'images.

2.4.2 Approches globales vs approches locales

Dans les approches globales, on exploite toute l'information de l'image pour calculer la signature de la personne à l'aide d'une seule représentation. Elle consiste d'abord à diviser l'image en des régions et ensuite à concaténer les représentations de toutes les régions pour former la signature de l'image. Les régions peuvent être des bandes horizontales (Truong Cong *et al.*, 2009, 2010a, Gray et Tao, 2008, Prosser *et al.*, 2010, Huang *et al.*, 2008, Park *et al.*, 2006, Gray et Tao, 2008, Prosser *et al.*, 2010, Ijiri *et al.*, 2012, Zheng *et al.*, 2011), une grille des régions avec chevauchement (Dikmen *et al.*, 2010, Hirzer *et al.*, 2012) ou des régions correspondant aux parties du corps humain (Cheng *et al.*, 2011, Farenzena *et al.*, 2010, Souded, 2013, An *et al.*, 2013). La représentation globale peut être aussi fondée sur une extraction des caractéristiques locales. Dans ce cas toutes les caractéristiques locales extraites de l'image sont combinées dans une représentation globale sous forme d'un seul vecteur (Jungling et Arens, 2011, Bauml et Stiefelhagen, 2011). Les approches locales représentent l'image ou la vidéo par plusieurs vecteurs caractéristiques. Chaque vecteur décrit une région ou un point d'intérêt localement détecté (Gheissari *et al.*, 2006, Hamdoun *et al.*, 2008, de Oliveira et de Souza Pio, 2009).

2.4.3 Approches supervisées vs approches non supervisées

De même, les approches de ré-identification peuvent être classifiées selon l'existence ou l'absence d'un ensemble de paires d'images mises en correspondance *a priori*. Deux familles d'approches sont alors distinguées : approches supervisées et approches non-supervisées. La première catégorie, dite approches supervisées, consiste à diviser la base de données en deux parties : une partie pour l'apprentissage et une partie pour le test. La base d'apprentissage est alors constituée des paires d'images mises en correspondances *a priori*. Ces informations *a priori* sont exploitées pour améliorer la ré-identification des personnes sur la partie de test. Contrairement à la première catégorie, les approches non supervisées consistent à apparier les personnes de la partie de test sans aucune information *a priori*.

Ce critère d'absence ou de présence d'une phase d'apprentissage est un critère pertinent pour caractériser une approche de ré-identification. Pour cette raison, on adopte dans cet état de l'art la classification en approches non-supervisées et approches supervisées.

Dans la suite, quelques approches non-supervisées et supervisées sont présentées en décrivant pour chacune les différentes primitives utilisées ainsi que la méthode d'appariement. Dans cette section, toutes les méthodes s'appliquent sur la ROI ; les mots “image” et “ROI” signifient la région d'intérêt où se trouve la personne à ré-identifier.

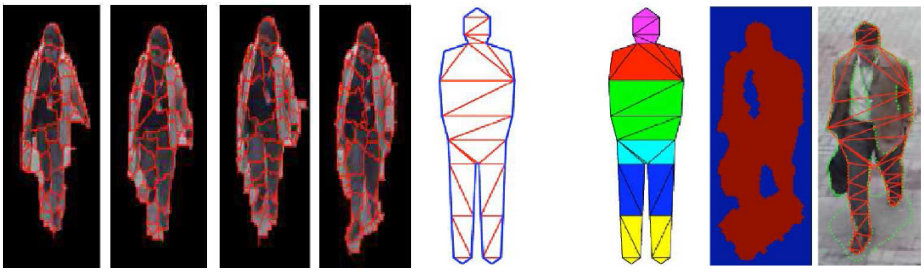
2.4.3.1 Approches de ré-identification non-supervisées

Les choix de représentations des personnes ainsi que la façon dont ils sont combinés pour obtenir une représentation robuste sont un défi. Cette représentation peut être par

points d'intérêt ou par division en régions.

La représentation par Points d'Intérêt (PI) consiste à décrire la ROI par des points saillants. Pour décrire ces points, des caractéristiques sont extraites d'une petite région centrée sur chaque point. Dans cette catégorie, on cite (Gheissari *et al.*, 2006, de Oliveira et de Souza Pio, 2009, Hamdoun *et al.*, 2008, Jungling et Arens, 2011, Bauml et Stiefelhagen, 2011).

- L'approche de (Gheissari *et al.*, 2006) est une méthode multi-échantillons qui consiste à modéliser la personne par un graphe où chaque nœud représente une région décrivant un point d'intérêt. D'abord, le détecteur de PIs «Hessienne affine invariante» (Mikolajczyk et Schmid, 2004) est appliqué. La description d'un PI combine deux types de primitives : la couleur et des informations structurelles (texture). La description couleur est fondée sur les deux canaux teinte et saturation de l'espace de couleur HSV. Quant à la description de l'information structurelle, elle est fondée sur les contours résultant du détecteur de contour de Canny (Canny, 1986). Ensuite, une segmentation triangulaire est appliquée pour modéliser la ROI par un graphe où chaque nœud décrit une région décrivant un PI (figure 2.3). Sachant que plusieurs images par personne sont disponibles, un algorithme dynamique est utilisé pour ajuster le modèle de la personne. Le graphe permet de filtrer les points d'intérêt qui n'appartiennent pas au modèle. Quant à la l'appariement des modèles, il s'agit de définir une similarité entre deux ensembles de points d'intérêt. Cette similarité est égale au cardinal de l'ensemble des correspondances des PIs retenues. Une correspondance est retenue si la distance euclidienne entre un PI test et le PI référence le plus proche est inférieure à un seuil.

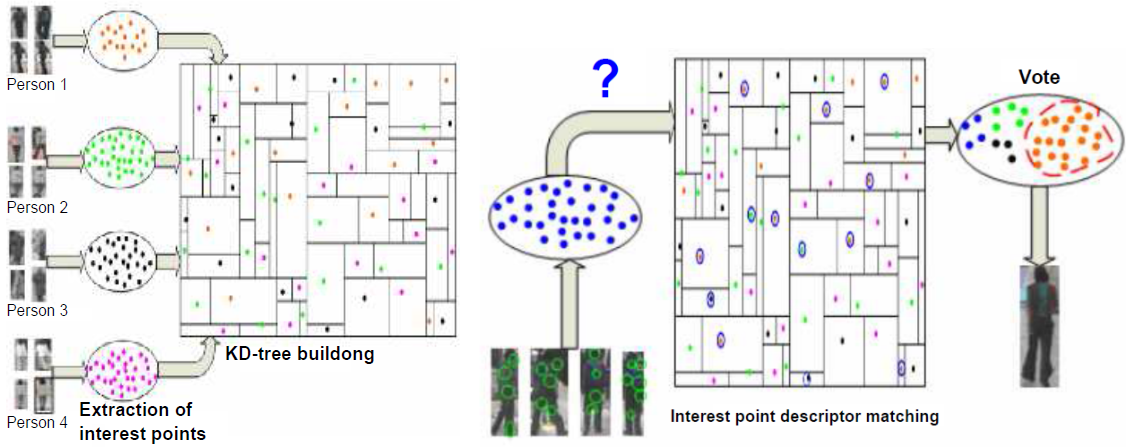


(a) Segmentation (b) Modèle triangulaire (c) Ajustement du modèle

FIGURE 2.3 – (a) Segmentation de quelques images. (b) Gauche : modèle d'une personne décomposé en des triangles. Les contours bleus correspondent aux bordures de la personne, et les droites rouges correspondent aux bords intérieurs. À droite : partitionnement de la ROI en des parties. (c) Gauche : masque de l'arrière plan. Droite : ajustement du modèle (Gheissari *et al.*, 2006).

- Les points d'intérêt, à savoir SURF (pour «Speeded Up Robust Features»)(Bay *et al.*, 2006), sont exploités dans (Hamdoun *et al.*, 2008). La méthode consiste à apparier des SURFs collectés sur des courtes séquences vidéo. D'abord, un modèle référence est construit pour toutes les personnes (figure 2.4-a). Ce modèle est équivalent à une base de référence contenant tous les SURFs de toutes les personnes références filmées par la caméra A. Les auteurs n'ont pas pris toutes les images de la vidéo mais uniquement des images espacées d'une demi-seconde. Ensuite, une

requête est construite. Elle contient tous les SURFs d'une personne inconnue filmée par la caméra B. Quant à l'appariement des descripteurs, la similarité entre deux SURFs décrits par deux vecteurs de même dimension est égale à la somme des valeurs absolues des différences des éléments deux à deux. Afin de ré-identifier la requête (2.4-b), chaque SURF de cette dernière est apparié au SURF référence le plus proche à l'aide de la technique arbre-KD (pour «arbre K-Dimensions»)qui assure un appariement robuste et rapide. Ensuite, un modèle de filtrage est appliqué aux paires mise en correspondance. Si la paire est retenue alors un vote est ajouté à l'identité du SURF référence correspondante. Le principe de filtrage est similaire au filtrage proposé par Lowe (Lowe, 2001) dans une application de reconnaissance des objets dans des images fixes. Il consiste à comparer les distances au SURF référence le plus proche d_1 et seconde SURF référence le plus proche d_2. En effet, la paire est retenue si $\frac{d_1}{d_2} \leq seuil$ et est écartée si $\frac{d_1}{d_2} > seuil$ (le seuil choisi dans ce travail est égale à 0.8). Finalement, la requête est ré-identifiée comme la personne ayant le plus de votes. Les auteurs ne valident une ré-identification que si le nombre de votes de l'identité référence reconnue est supérieur à un seuil.



(a) (b)

FIGURE 2.4 – (a) Vue schématique de la construction du modèle : chaque personne référence est représentée par 4 images, tous les SURFs sont collectés et unis pour construire un arbre-KD. (b) Ré-identification d'une requête : chaque SURF est apparié au point référence le plus proche ; et s'il est retenu alors un vote est ajouté à la personne référence correspondante. Dans cet exemple, la séquence requête est reconnue comme la première personne référence (Hamdoun *et al.*, 2008).

- (de Oliveira et de Souza Pio, 2009) modélisent la ROI par des SURFs en ajoutant une description couleur. La description couleur est fondée sur la composante teinte (H) de l'espace de couleurs HSV, invariant à la luminosité et proposé dans (Swain et Ballard, 1990).

$$H = \arccos \frac{\log(R) - log(G)}{\log(R) + \log(G) - 2 * log(B)}$$

où R, G et B sont les canaux de l'espace de couleur RGB. Quant à l'appariement des PIs, la méthode n'utilise pas un arbre-KD pour accélérer le calcul comme dans (Hamdoun *et al.*, 2008). L'appariement entre une requête et une référence où chacune est représentée par un ensemble de PIs (m respectivement p) est validé si au

moins 2 PIs sont similaires. Deux PIs sont dits similaires si la distance euclidienne entre eux est inférieure à un seuil empirique.

- Dans (Bauml et Stiefelhagen, 2011), plusieurs détecteurs et descripteurs de PIs sont testés et comparés (figure 2.5). D'abord, un détecteur de PI est appliqué. Ensuite, chaque ROI est représentée par un histogramme où chaque composante décrit le nombre d'occurrence d'un centre. Les centres sont calculés sur l'ensemble de référence à l'aide de l'algorithme k-means. Cette représentation est connue dans la littérature par « Bag of Features (BoF)». Le descripteur final d'une personne représentée par une séquence vidéo est égal à la moyenne des descripteurs normalisés de toutes les ROIs de cette personne. L'appariement des descripteurs est fondé sur l'algorithme KNN. La figure 2.5 montre les résultats de quelques détecteurs de PIs.

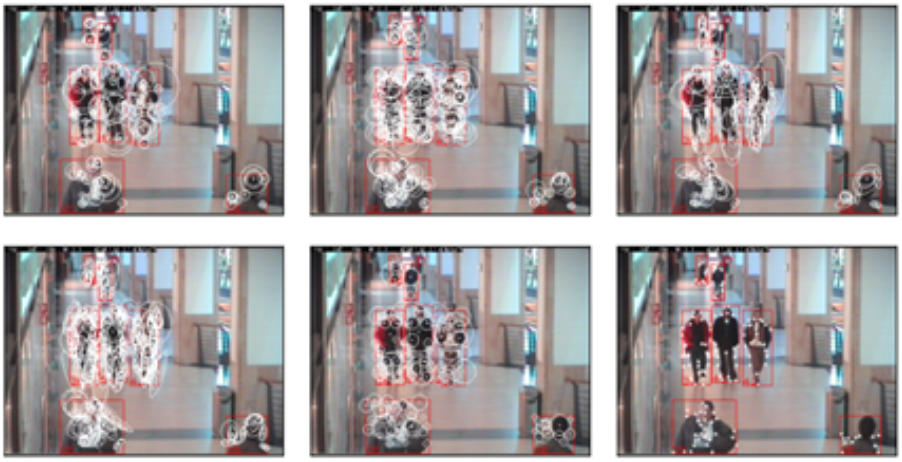


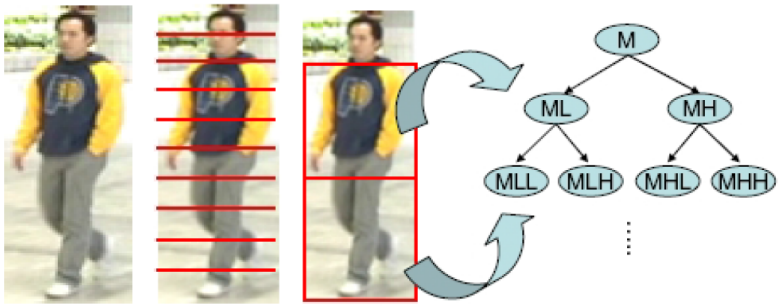
FIGURE 2.5 – Les réponses des différents détecteurs de points d'intérêt. De haut en bas et de gauche à droite : Harris, Harris-Laplace, Hessian-Laplace, Harris-Affine, Hessian-Affine et Fast-Hessian (Bauml et Stiefelhagen, 2011).

- Dans (Jungling et Arens, 2011), les auteurs proposent un système qui comprend trois étapes : 1) détection de la ROI, 2) suivi de la personne et 3) sa ré-identification. Les auteurs utilisent un modèle implicite de forme (ISM pour «Implicit Shape Model») pour la détection et le suivi des personnes alors que la ré-identification est fondée sur les points d'intérêt SIFT (pour «Scale Invariant Features Transform»). Pour décrire un PI, les auteurs ajoutent au descripteur SIFT une description spatiale (la position de PI dans l'image). En appariement, chaque personne est décrite par un BoF. Sachant que le descripteur SIFT fondé sur les gradients n'est pas robuste au changement de la direction de la marche, les auteurs proposent une transformation miroir du gradient en fonction de la direction de marche de la personne en référence et en test. Dans ce travail, la robustesse de la méthode aux changements des angles de vue est évaluée. La méthode a été aussi testée dans un contexte d'acquisition par caméra infrarouge (Jungling et Arens, 2010).

Contrairement à une représentation par PIs, la représentation de la ROI par division en régions consiste à exploiter toute l'information disponible. D'abord, la ROI est découpée en régions. Ensuite, la concaténation des représentations de toutes les régions définit la représentation de la ROI.

- (Truong Cong *et al.*, 2010b) décrivent une méthode à base de régions. D'abord, la ROI est découpée en «P» bandes horizontalement équidistantes. Ensuite, chaque bande est décrite par les couleurs moyennes des trois canaux RGB. Une ROI est alors représentée par un vecteur de dimension $3 * P$, concaténation des descriptions de ses P-bandes. En appariement, une ROI requête est représentée comme une combinaison parcimonieuse de toutes les ROIs références. A l'aide de cette représentation de dimension égale au nombre de ROIs références, une erreur de reconstruction est calculée pour chaque identité référence en utilisant uniquement les coefficients correspondant à cette identité. La requête est reconnue comme l'identité référence minimisant l'erreur de reconstruction. Les auteurs ont utilisé le même algorithme d'appariement par la représentation parcimonieuse présenté dans (Wright *et al.*, 2009) et utilisé pour la reconnaissance des visages.

- (Huang *et al.*, 2008) segmentent la ROI en trois parties de haut en bas (à 1/5ème et 3/5ème de la hauteur). Les trois parties correspondent idéalement à la tête, le tronc et les jambes. Dans ce travail, les informations d'apparence sont extraites uniquement des deux parties inférieures en ignorant la tête. En représentation, une stratégie d'échantillonnage de couleur est appliquée pour obtenir une structure d'arbre contenant les médianes des couleurs. Dans cette structure, les histogrammes fils sont calculés par la séparation des régions en partie supérieure et partie inférieure. La représentation d'une ROI est finalement obtenue par la fusion des vecteurs médians acquis auprès des canaux RGB. Quant à l'appariement, la distance Kullback-Leibler est calculée pour apparier les ROIs. La figure 2.6 montre la transformation d'une ROI en une représentation arbre.

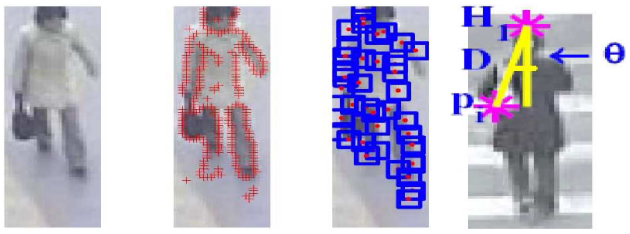


(a) (b) (c)

FIGURE 2.6 – (a) Image originale. (b) Segmentation en bandes horizontales (Bird *et al.*, 2005). (c) Construction du structure arbre extrait à partir de la partie supérieure et inférieure du corps (Huang *et al.*, 2008).

- Dans (Cai *et al.*, 2008), l'image est segmentée en des régions extraites de l'image contour. Les contours sont détectés par le détecteur de Canny (Canny, 1986). Ensuite, des régions rectangulaires sont sélectionnées autour des contours. Quant à la représentation, l'apparence de chaque région est décrite par ses couleurs dominantes et leurs fréquences d'apparition dans la région. Une contrainte spatiale est ajoutée pour décrire une région. Cette contrainte spatiale est codée par la distance entre un point référence et un point de contour, normalisée par la hauteur de la silhouette. Dans ce travail, le point au sommet de la tête est pris comme un point de référence. Les régions de la ROI requête et référence sont appariées en définissant un coût. Ce coût d'appariement tient compte de l'apparence des deux régions et de leurs

positions. La similarité entre deux images est égale à la moyenne des coûts des K meilleurs appariements.



(a) (b)

FIGURE 2.7 – (a) De gauche à droite : image originale, résultat du détecteur de contour de Canny, régions décrivant les contours. (b) Contrainte spatiale (Cai *et al.*, 2008).

- (Bak *et al.*, 2010b) décrivent une approche qui s'appuie sur les covariances spatiales des parties du corps humain. Pour détecter ces parties, un détecteur fondé sur l'histogramme de gradients orientés est appliqué. Il détecte 6 parties correspondant idéalement aux parties du corps : tête, tronc, bras droit, bras gauche, jambe droite et jambe gauche. En représentation, chaque région est décrite par une matrice de covariance. Une méthode d'appariement pyramidale (figure 2.8) à noyau est ensuite appliquée (Grauman et Darrell, 2005).

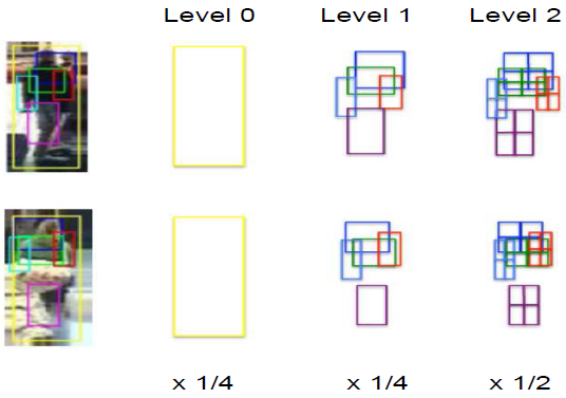


FIGURE 2.8 – Exemple de construction d'une pyramide à trois niveaux. La colonne de gauche représente deux images avec les parties du corps détectés. « Level 0 » correspond au corps entier. «Level 1» et «Level 2» correspondent respectivement au reste des parties du corps et l'intérieur de ces parties (Bak *et al.*, 2010b).

- (Cheng *et al.*, 2011) formulent le problème de ré-identification comme un appariement des parties du corps humain, partie par partie. Cette formulation est justifiée par le fait qu'une partie du corps peut être observée dans une caméra et cachée (partiellement ou totalement) dans une autre. Cette méthode consiste d'abord à décomposer le corps en 6 parties correspondant idéalement à : tête, poitrine, cuisses et jambes. Pour la détection, la méthode utilisée est fondée sur des structures picturales (PS pour «Pictorial Structures») (Felzenszwalb et Huttenlocher, 2005). Une PS repose essentiellement sur deux étapes : capture de l'apparence locale des différentes parties du corps et une deuxième étape qui gère l'articulation entre ces différentes parties. Dans un contexte multi-échantillons, chaque partie du corps se retrouve plusieurs fois (une fois par image). Les auteurs proposent un algorithme

itératif appelé CPS (pour «Custom Pictorial Structures») qui exploite toutes les occurrences pour mettre à jour le modèle du corps (figure 2.9). Quant à la représentation, chaque partie est décrite par deux descriptions couleur : histogramme de couleur HSV et les couleurs des régions les plus stables (MSCR pour «Maximally Stable Colour Regions») (Forssén, 2007). Quant à l'appariement, la similarité entre deux représentations est définie comme la somme pondérée de deux distances (une distance entre HSV et une distance entre MSCR). Dans un cas de multi-échantillons, la similarité entre deux séquences d'images est égale à la distance minimale entre les images requêtes et références prises deux à deux.

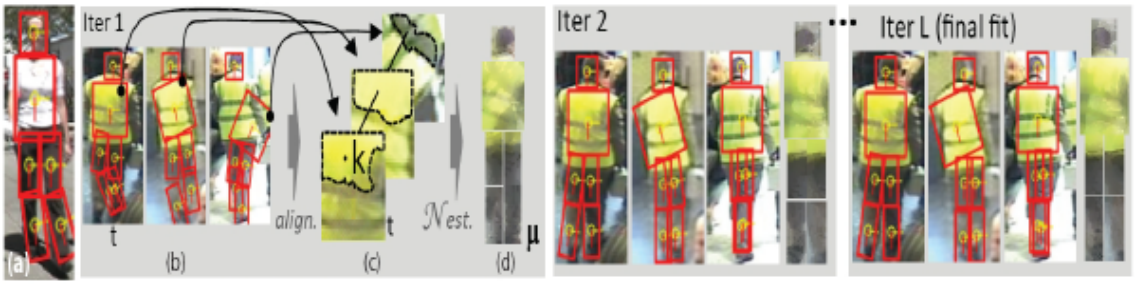


FIGURE 2.9 – CPS : méthode itérative de localiser les parties du corps humain (Cheng *et al.*, 2011).

- (Park *et al.*, 2006) proposent un système dit «ViSE» (Visual Search Engine). D'abord, la ROI est détectée et ensuite décrite par trois caractéristiques : la hauteur du corps, le rapport hauteur/largeur et une caractéristique couleur. Les deux premières primitives sont faciles à extraire dès qu'on trouve la ROI. Ensuite, la ROI est divisée en trois parties de haut en bas (à 1/5ème et 3/5ème de la hauteur). Idéalement, ces trois parties correspondent à la tête, la chemise et le pantalon. Dans ce travail, uniquement les couleurs de la chemise et du pantalon sont prises en compte. Chaque pantalon ou chemise est caractérisé par un histogramme de 10 composantes dans l'espace de couleurs HSV (en utilisant uniquement le canal « teinte » (H)). Les histogrammes construits sont de dimension dix correspondant aux couleurs suivantes : rouge, brun, jaune, vert, bleu, violet, rose, blanc, noir et gris. La couleur ayant le plus de pixels est retenue comme couleur finale d'une telle partie. Le schéma de la méthode ViSE est présenté dans la figure 2.10.

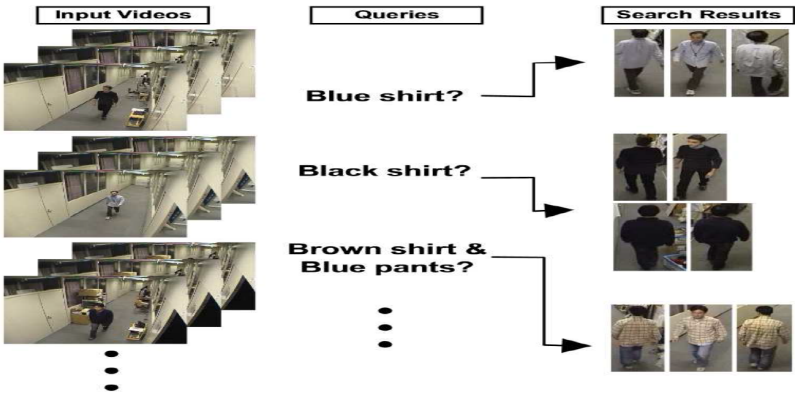


FIGURE 2.10 – Schéma de la méthode ViSE (Park *et al.*, 2006).

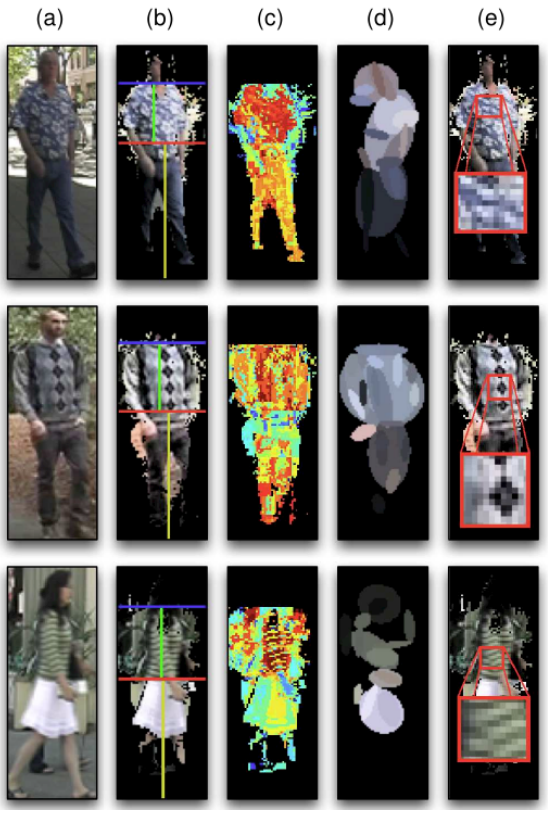


FIGURE 2.11 – Illustration du descripteur SDALF (Farenzena *et al.*, 2010). (a) Étant donnée une image, (b) SDALF localise les parties du corps ainsi que les axes de symétrie. Ensuite, les descriptions de l'apparence sont extraits : (c) histogrammes HSV pondérés par la distance à l'axe de symétrie, (d) MSCR et (e) RHSP.

- (Farenzena *et al.*, 2010) proposent d'exploiter la symétrie de la ROI pour la diviser en 5 régions. En effet, deux axes de symétrie horizontaux sont recherchés. Ils séparent respectivement la tête du tronc et le tronc des jambes. Ensuite, deux autres axes de symétrie verticaux sont cherchés, le premier divise le tronc et le deuxième divise les jambes. Ensuite, trois caractéristiques complémentaires sont extraites pour chaque région. Dans la littérature cette méthode est connue sous le nom «SDALF» (pour « System Drive, Accumulation of Local Features »). Quant à la représentation, elle combine deux descriptions couleurs et une description texture. Une première description couleur consiste à chercher pour chaque région les zones les plus stables en couleur (MSCR) (Forssén, 2007). MSCR consiste à trouver itérativement un ensemble de zones de couleur semblables. Ensuite, chaque zone est décrite par sa surface, la position du centre de gravité, la couleur moyenne et la matrice des moments d'ordre 2. La deuxième description couleur est un histogramme de couleurs pondéré. Il s'agit d'extraire un histogramme dans l'espace de couleur HSV pour chaque région où les valeurs des pixels sont pondérées par la distance à l'axe de symétrie verticale. Quant à la description texture dite «patches de structure récurrente» (RHSP pour «Recurrent High-Structured Patches»), elle est proposée par (Farenzena *et al.*, 2010). Elle consiste d'abord à segmenter chaque région en des zones de même dimension. Ensuite, un module de filtrage est appliqué pour garder les zones les plus structurées. La figure 2.11 montre la génération des trois descriptions. Quant à l'appariement, la

similarité entre deux images est définie comme la somme pondérée des trois distances (une distance entre HSV, une distance entre MSCR et une distance entre RHSP).

- (Souded, 2013) étend la méthode SDALF en proposant des prétraitements ou adaptation des paramètres. Parmi les contributions de ce travail, on en cite deux : 1) la première consiste à ajouter une normalisation des couleurs afin de minimiser l'influence des changements d'éclairage entre les données de test et de référence, 2) la deuxième contribution consiste à adapter les pondérations de chacune des trois descriptions. En effet, dans (Farenzena *et al.*, 2010), une pondération constante est considérée sur toute la base de données. Par contre dans ce travail, une pondération est considérée pour chaque personne sans aucune phase d'apprentissage *a priori*.

- Dans (Hirzer *et al.*, 2011), chaque personne est représentée par deux modèles : descriptif et discriminant. Le modèle discriminant est appris en s'appuyant sur le modèle descriptif. En effet, pour chaque image, un modèle descriptif d'apparence est appliqué. Il consiste à découper l'image en 7 bandes horizontales et pour chaque bande calculer un descripteur de covariance (Tuzel *et al.*, 2006). En utilisant cette description, un premier classement des images de référence est généré. La méthode utilise les 50 premières images références proches de l'image test pour décider si la personne test existe dans la base de référence. Sinon, un deuxième modèle discriminant est appliqué. Ce modèle discriminant est fondé sur l'algorithme Adaboost qui prend en entrée les descripteurs de covariance des images (Tuzel *et al.*, 2006) et d'ondelettes de Haar. Comme déjà mentionné précédemment, ce type d'algorithme nécessite des images positives et négatives. Dans ce cas, les images négatives correspondent aux dernières images du classement donné par le modèle descriptif et les images positives correspondent à des images extraites de la séquence test sur les quelles on a effectué quelques modifications (des transformations géométriques et des déplacements) pour assurer une certaine variation. Cette approche utilise une méthode d'apprentissage pour sélectionner les primitives les plus discriminantes mais en n'utilisant que les données de test (l'ensemble référence de la base de test). C'est pourquoi, elle est classée dans les méthodes non-supervisées. La figure 2.12 visualise le principe de la méthode et montre le rôle du modèle descriptif pour apprendre le modèle discriminant.

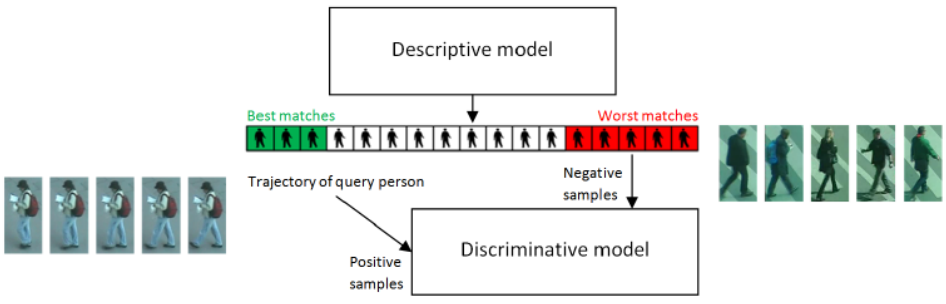


FIGURE 2.12 – Après l'application d'un modèle descriptif pour obtenir un classement initial, un modèle discriminant peut être utilisé pour affiner le résultat (Hirzer *et al.*, 2011).

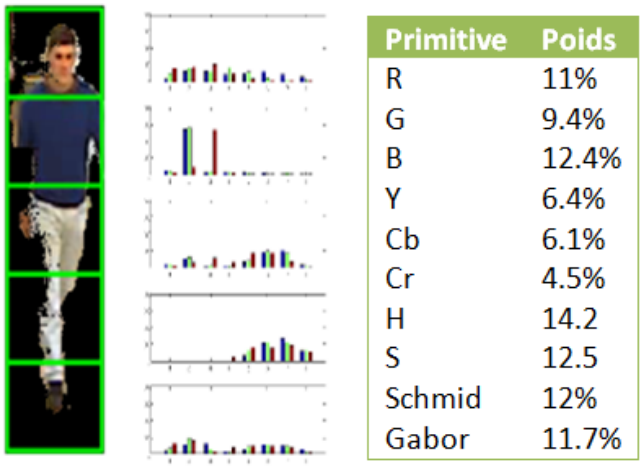
Dans la section suivante, nous décrivons quelques approches supervisées utilisées pour la ré-identification.

2.4.3.2 Approches de ré-identification supervisées

Les méthodes supervisées exigent l'existence d'une base d'apprentissage composée de paires d'images mises en correspondance *a priori*. Ces informations de correspondance *a priori* permettent de trouver le meilleur moyen d'apparier les personnes de test. On distingue deux méthodes d'apprentissage selon l'étape où l'apprentissage est effectué. La première catégorie (méthodes d'apprentissage de métrique) consiste à apprendre la distance entre deux images. La deuxième catégorie (méthodes discriminantes) consiste à trouver les représentations discriminantes de la ROI. Dans ces deux catégories, le choix de la métrique ou de la représentation discriminante se fait de façon à maximiser la similarité entre les paires d'images négatives et minimiser la similarité entre les paires d'images positives. Dans le reste de cette section, on appelle base d'apprentissage un ensemble des paires d'images prises par les deux caméras A et B et mises en correspondance *a priori*.

Apprentissage de métrique

- (Gray et Tao, 2008) introduit l'«Ensemble of Localized Features» (ELF). L'idée principale de ce travail consiste à combiner un ensemble de primitives pour représenter la ROI et estimer un poids pour chaque primitive. L'estimation de ces poids est faite à l'aide de l'algorithme Adaboost appris sur une base d'apprentissage. Il délivre en sortie les primitives les plus discriminantes en affectant un poids pour chaque primitive. La mesure de similarité entre une paire d'images test est fondée sur les poids déjà appris. Quant à la représentation des ROIs, la méthode consiste à découper l'image en 5 bandes horizontales sans chevauchement. Ensuite, pour représenter une bande, des histogrammes RGB, YCbCr et HS ainsi que des filtres de Gabor et Schmid sont calculés. L'algorithme d'Adaboost est alors appris par les primitives déjà calculées en générant un poids pour chacun. Une fois appris, l'algorithme peut classifier une paire d'image test inconnue. La figure 2.13 montre le principe de l'étape d'extraction de primitives ainsi que les poids des primitives donnés par l'algorithme d'Adaboost.



(a)

Primitive	Poids
R	11%
G	9.4%
B	12.4%
Y	6.4%
Cb	6.1%
Cr	4.5%
H	14.2
S	12.5
Schmid	12%
Gabor	11.7%

(b)

FIGURE 2.13 – (a) Extraction des primitives (figure issue de (Meden, 2013)). (b) Poids des primitives (Gray et Tao, 2008).

- La méthode de (Prosser *et al.*, 2010) est inspirée de (Gray et Tao, 2008) et appelé «Ensemble RankSVM». L'idée présentée est très similaire à ELF (Gray et Tao,

2008). En effet, un algorithme d'apprentissage permet de sélectionner les primitives les plus discriminantes. La méthode consiste d'abord à segmenter la ROI en des bandes horizontales pour capter plus ou moins la tête, le tronc supérieur, le tronc inférieur et les jambes. Pour représenter une bande, le même ensemble de primitives que celui présenté dans (Gray et Tao, 2008) est extrait. La différence majeure entre ce travail et (Gray et Tao, 2008) est qu'un «Ensemble RankSVM» vient remplacer l'Adaboost pour l'apprentissage. «Ensemble RankSVM» est une extension de la méthode «RankSVM» (Joachims 2002) où le problème de ré-identification est formulé comme un problème de classement.

- (Dikmen *et al.*, 2010) proposent une méthode pour apprendre la distance de Mahalanobis sur une base d'apprentissage et ensuite apparier une paire d'image test par l'algorithme du k-plus-proche-voisin. Cette méthode, appelée LMNN-R (pour «LMNN with Rejection»), est l'extension de la méthode LMNN (pour «Large Margin Nearest Neighbor») (Weinberger 2009), en introduisant la notion de rejet (aucun voisin n'est retenu si tous les voisins sont au-delà d'une distance). Le principe de cette méthode est de découper densément la ROI en des régions rectangulaires de dimension 8x24 pixels. Ensuite, de chaque région, on extrait deux histogrammes de couleurs (RGB et HSV) où chaque canal est représenté par un histogramme de 8 composantes. Finalement, une ACP est appliquée pour réduire la dimensionnalité.

- (Hirzer *et al.*, 2012) formulent le problème de ré-identification comme un apprentissage de transition entre les deux caméras A et B. La représentation de la ROI combine des primitives de couleur et de texture. Les primitives couleurs sont fondées sur deux histogrammes de couleurs HSV et Lab. Quant à la primitive texture, elle est fondée sur une analyse LBP. Plus précisément, chaque ROI est divisée en des régions rectangulaires de dimension 8x16 avec un chevauchement de 50%. Pour chaque région, les moyennes des canaux couleurs sont quantifiées à des valeurs de 0 à 40. En outre, l'histogramme LBP est généré à partir de l'image intensité de cette région. Les représentations de toutes les régions sont concaténées pour représenter la ROI par un vecteur de grande dimension (figure 2.14). De ce fait, une ACP est appliquée pour réduire sa dimensionnalité. En classification, le cœur de la méthode est similaire à LMNN. Il consiste à apprendre la distance de Mahalanobis tout en proposant une fonction objective qui tient compte de la dissimilarité des imposteurs.

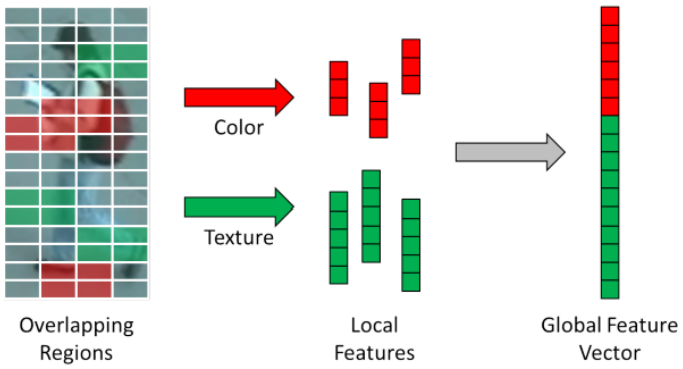


FIGURE 2.14 – Extraction des primitives sur des régions chevauchées avec concaténation dans un seul vecteur (Hirzer *et al.*, 2012).

- (Ijiri *et al.*, 2012) proposent une méthode d'apprentissage de métrique, dite analyse de composante à large marge (LMCA pour «Large Margin Component Analysis»). LMCA vise à minimiser une fonction métrique tout en séparant les ROIs de différentes classes par une grande marge dans un nouvel espace de grande dimension. Afin de représenter la ROI, celle-ci est d'abord découpée en des bandes horizontales (Bird *et al.*, 2005) ; ensuite un histogramme de couleur dans l'espace HSV est extrait de chaque bande. Les histogrammes de toutes les bandes sont concaténés pour former la représentation de la ROI. En classification, plusieurs distances sont investiguées (Chi-Square à noyau, Bhattacharyya à noyau, etc.) pour calculer des similarités entres les projections des représentations des ROIs dans le nouvel espace. Cette approche est illustrée par la figure 2.15.

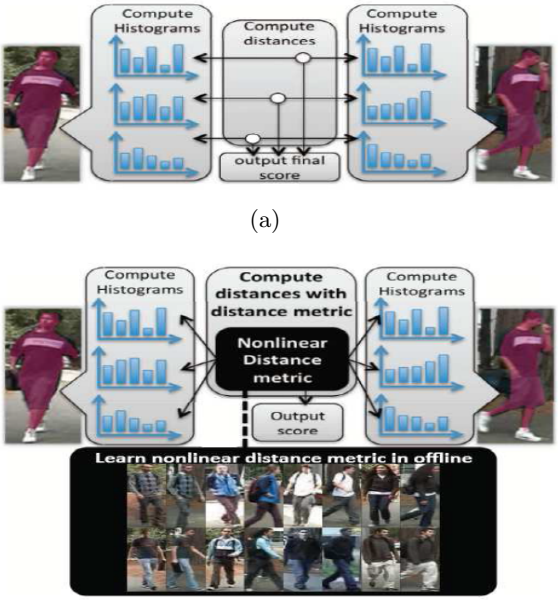


FIGURE 2.15 – Comparaison schématique entre (a) sans utilisation d'un apprentissage métrique, et (b) en utilisant un apprentissage métrique non linéaire (Ijiri *et al.*, 2012).

- (Zheng *et al.*, 2011) formulent le problème de ré-identification comme un problème de comparaison probabiliste relative à des distances (PRDC pour «Probabilistic Relative Distance Comparison»). Ce problème probabiliste est appris sur une base d'apprentissage en maximisant les distances entre les paires d'images négatives et minimisant les distances entre les paires d'images positives. Quant à la représentation, chaque image est décrite par un mélange de primitives de couleur et de texture. Plus précisément, chaque image est découpée en six bandes horizontales. De chaque bande, trois histogrammes de couleur (RGB, YCbCr et HSV) sont extraits ainsi que deux primitives de texture (filtre de Gabor et filtre de Schmid) représentées par des histogrammes. Au total, 29 canaux sont extraits et chacun est représenté sur un histogramme de 16 composantes pour obtenir finalement un vecteur de dimension 2784 décrivant chaque image.

- (An *et al.*, 2013) propose une méthode de calcul de similarité entre une paire d'images test dans un espace de projection appris sur une base d'apprentissage. L'espace de projection est construit par la méthode «Regularized Canonical Correlation Analysis

(RCCA)» (Leurgans 1993). Cette méthode construit un espace de projection où la corrélation entre les images positives d'apprentissage prises par la caméra A et celles prises par la caméra B est maximale. Quant à la représentation initiale d'une image, avant l'application de RCCA, elle consiste à découper l'image en 155 régions de 8x16 avec chevauchement de 50%. De chaque région, les moyennes des canaux couleurs HSV et Lab ainsi qu'un histogramme LBP sur 8 composantes sont extraits. Toutes ces primitives sont concaténées pour représenter une image. Une fois que l'espace RCCA est créé, chaque image test est projetée sur cet espace en gardant uniquement les projections sur les 50 premiers vecteurs propres. Sa représentation finale est formée par le calcul des similarités entre cette image et les images d'apprentissage prises par la même caméra en utilisant l'inverse de la distance euclidienne. Quant à la classification, la similarité entre une paire d'images test est calculée par la distance cosinus entre leurs projections dans l'espace RCCA. La figure 2.16 montre le principe de représentation d'une paire d'images test.

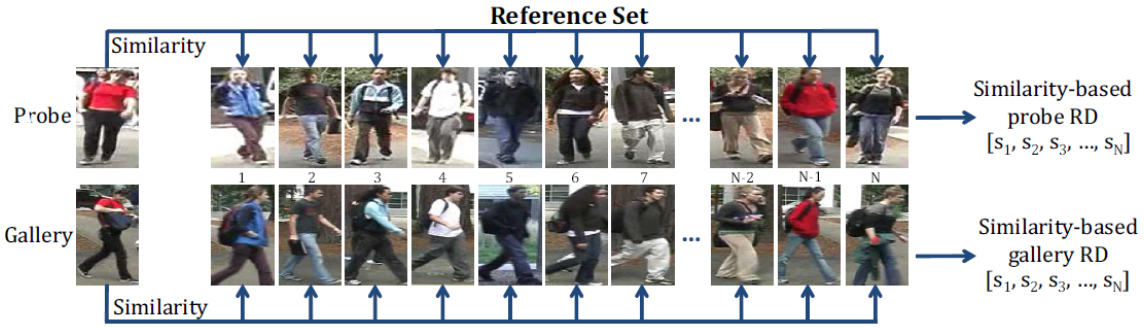


FIGURE 2.16 – La représentation d'une image de la base de test est générée par le calcul des similarités entre cette image et les images d'apprentissages prises par la même caméra (An *et al.*, 2013).

D'autres méthodes d'apprentissage de métrique, proposées dans la littérature pour d'autres problématiques, ont été testées dans le contexte de la ré-identification. Parmi ces méthodes, on cite «Large Margin Nearest Neighbor (LMNN)» (Weinberger 2009), «Information Theoretic Metric Learning (ITML)» (Davis 2007), et «Logistic Discriminant Metric Learning (LDML)» (Guillaumin 2009). LMNN consiste à apprendre la distance de Mahalanobis en maximisant la distance entre les paires négatives et en minimisant la distance entre les paires positives et ensuite à l'appliquer dans l'algorithme KNN. ITML consiste à apprendre la distance Mahalanobis, en formulant le problème d'apprentissage comme un problème de maximisation d'entropie. LDML est fondée sur une formulation probabiliste modélisée par une fonction sigmoïde et la distance de Mahalanobis.

Méthodes discriminantes

Ce type de méthodes propose d'améliorer la représentation de la ROI sur une base d'apprentissage. Parmi ces approches on trouve surtout les méthodes discriminantes de réduction de dimensionnalité.

- (Truong Cong *et al.*, 2009) introduisent une méthode discriminante non linéaire de réduction de dimensionnalité fondée sur les graphes et apprise sur une base d'apprentissage. La ROI est décrite d'abord par une combinaison de trois primitives de

couleur : 1) l'histogramme de couleur RGB, 2) un spatiogramme qui est une généralisation de l'histogramme RGB en ajoutant des moments d'ordre supérieur et 3) *color/path-length* : où chaque pixel est présenté par deux paramètres (v, l) où v représente la couleur du pixel et l représente la distance entre ce pixel et un point référence de la ROI. Ensuite, une réduction de dimensionnalité est appliquée. Pour assurer l'invariance des représentations aux conditions de luminosité et aux paramètres des caméras, trois techniques de normalisation de l'histogramme RGB sont comparées. Dans un contexte multi-échantillons, chaque personne est représentée par un ensemble de points (les projections dans le nouvel espace réduit) dont le centre sert à comparer les signatures. Dans ce cas, la similarité entre deux personnes est égale à la distance entre les deux centres correspondants.

- (Truong Cong *et al.*, 2010a), extension de (Truong Cong *et al.*, 2009), propose un nouveau descripteur de couleur. La ROI est découpée horizontalement en «P» bandes et de chaque bande les moyennes des canaux RGB sont extraites. En effet, chaque image est représentée par un vecteur de dimension $3 * P$. Ensuite, une réduction de dimensionnalité similaire à celle de (Truong Cong *et al.*, 2009) est appliquée. Quant à l'appariement des personnes, il est fondé sur la méthode «machine à vecteurs support» (SVM pour «Support Vector Machine»). La figure 2.17 montre la division de la ROI en régions ainsi que la location des couleurs dans chaque région.

(a) (b) (c)

FIGURE 2.17 – (a) image originale, (b) localisation de la silhouette, (c) répartition des couleurs dans la silhouette (Truong Cong *et al.*, 2010a).

Une autre catégorie peu abordée dans la littérature consiste à ajouter des informations liées au contexte aux apparences. Les informations peuvent être liées au voisinage de la personne à ré-identifier. De ce fait, on ne tient pas compte uniquement des apparences de la personne à ré-identifier mais aussi de ses voisinages. Ce type de méthodes s'applique sur des groupes de personnes (Zheng *et al.*, 2009). L'information additionnelle peut être liée aussi à la liaison espace-temps en se basant sur l'hypothèse qu'une personne ne peut être dans deux positions différentes à un instant donné (Haruyuki *et al.*, 2012). Le contexte d'application de ce type d'approches est très loin de notre scénario de ré-identification où on désire ré-identifier les personnes après quelques minutes de leurs enregistrement. Donc, l'information des voisinages de la personne à ré-identifier ou sa position dans la scène ne sont plus utilisables.

La table 2.1 résume les différentes approches de ré-identification citées dans ce chapitre en montrant leurs catégories et le type de représentations extraites.

TABLE 2.1 – Tableau récapitulatifs des approches de ré-identification.

Référence	**Classification**		**Représentation**		
	Non Supervisée	Supervisée	Couleur	Texture	Forme
(Gheissari *et al.*, 2006)	✓		✓	✓	
(Hamdoun *et al.*, 2008)	✓			✓	
(de Oliveira et de Souza Pio, 2009)	✓		✓	✓	
(Bauml et Stiefelhagen, 2011)	✓			✓	
(Jungling et Arens, 2010)	✓			✓	✓
(Truong Cong *et al.*, 2010b)	✓		✓		
(Huang *et al.*, 2008)	✓		✓		
(Cai *et al.*, 2008)	✓		✓		✓
(Bak *et al.*, 2010b)	✓		✓		
(Cheng *et al.*, 2011)	✓		✓		
(Park *et al.*, 2006)	✓		✓		✓
(Wang *et al.*, 2007)	✓		✓		✓
(Farenzena *et al.*, 2010)	✓		✓	✓	
(Souded, 2013)	✓		✓	✓	
(Hirzer *et al.*, 2011)	✓		✓	✓	
(Gray et Tao, 2008)		✓	✓	✓	
(Prosser *et al.*, 2010)		✓	✓	✓	
(Dikmen *et al.*, 2010)		✓	✓		
(Hirzer *et al.*, 2012)		✓	✓	✓	
(Ijiri *et al.*, 2012)		✓	✓		
(Zheng *et al.*, 2011)		✓	✓	✓	
(An *et al.*, 2013)		✓	✓	✓	
(Truong Cong *et al.*, 2009)		✓	✓		✓
(Truong Cong *et al.*, 2010a)		✓	✓		
(Schwartz et Davis, 2009)		✓	✓	✓	

2.5 Base de données en ré-identification

Dans cette section, nous présentons les bases de données les plus utilisées dans l'état de l'art pour évaluer les approches de ré-identification. Les bases de données peuvent être regroupées en deux catégories : bases de données mono-échantillon et bases de données multi-échantillons. Dans le cas multi-échantillons, une personne est représentée par un ensemble d'images clés (par exemple, CAVIAR4REID et ETHZ) ou par une séquence vidéo (par exemple PRID-2011 (version multi-échantillons) ou CASIA-A). Comme base de données mono-échantillon, où chaque personne est représentée par une seule image, on trouve VIPeR et PRID-2011 (version mono-échantillon). Dans la suite, nous décrivons les différentes bases de données.

2.5.1 Base de données multi-échantillons

CASIA-A (Chinese Academy of Sciences, Institute of Automation) (CASIA, 2001)
C'est une base de données publique créée en 2001. Elle est composée de 20 personnes.

Pour chaque personne, 12 séquences vidéo ont été enregistrées. En effet, la personne parcourt deux fois, six trajectoires ayant différentes directions. Les six directions sont les suivantes : 0˚, 90˚, 135˚, 180˚, 270˚et 315˚. Puisque deux séquences sont disponibles pour chaque direction de marche, le protocole d'évaluation consiste à utiliser une séquence en test et une séquence en référence. Cette base de données a été utilisée originellement dans un scénario d'identification. Par contre, elle est adaptée à l'évaluation des approches de ré-identification en changeant la direction de marche en test et en référence. Le nombre de caméras n'est pas mentionné dans la description de cette base de données mais en regardant les vidéos on voit qu'il s'agit d'une seule caméra utilisée pour filmer toutes les directions de marche.

PRID-2011 (Person Re-ID 2011), version multi-échantillons : (Hirzer *et al.*, 2011) C'est une base de données publique créée en 2011. Ses séquences vidéo sont obtenues par deux caméras de surveillance (A et B) placées dans la rue. 385 personnes sont filmées par la caméra A et 749 personnes sont filmées par la caméra B (200 personnes sont filmées par les deux caméras). L'évaluation typique d'un système de ré-identification sur cette base de données consiste à chercher les 200 personnes passant dans un champ de vue d'une caméra parmi toutes les personnes passant dans le champ de vue de l'autre caméra.

Cela signifie qu'il y a deux procédures d'évaluation possibles :

- Procédure d'évaluation de A à B :
  - Ensemble de test : les 200 premières personnes de A.
  - Ensemble de référence : les 749 personnes de B.
- Procédure d'évaluation de B à A :
  - Ensemble de test : les 200 premières personnes de B.
  - Ensemble de référence : les 385 personnes de A

CAVIAR4REID (Context Aware Vision using Image-based Active Recognition for Re-identification) (Cheng *et al.*, 2011)
Cette base de données publique a été créée en 2011, et est extraite de la base de données CAVIAR (CAVIAR, 2003). CAVIAR est composée de 26 séquences filmées dans le hall d'entrée d'un centre commercial à Lisbonne à partir de deux angles de vue différents et à des instants différents. Elle contient des images de gens marchant seuls, rencontrant d'autres personnes, entrant et sortant des magasins. À partir de CAVAIR, CAVIAR4REID est construite. 72 personnes sont sélectionnées : 50 d'entre elles sont filmées sous deux angles de vue différents (10 images sont fournies pour chaque angle de vue par personne), et 22 personnes sont filmées avec un seul angle de vue (de même 10 images sont fournies pour cette vue par personne). Le protocole d'évaluation consiste à chercher les 50 personnes passant dans le champ de vue d'une caméra parmi les mêmes 50 personnes passant dans le champ de vue de l'autre caméra.

ETHZ-REID (en Allemand : Eidgenössische Technische Hochschule Zürich) (Schwartz et Davis, 2009)
C'est une base de données publique créée en 2009, publiée originellement dans (Ess *et al.*, 2007) et utilisée dans une application de détection des personnes dans des rues bondées. Les images de cette base de données ont été extraites par (Schwartz et Davis, 2009) et adaptées à un scénario de ré-identification. ETHZ_REID est divisée en trois sous ensembles ETHZ1 (83 personnes, 4857 images), ETHZ2 (35 personnes, 1936 images) et ETHZ3 (28

personnes, 1762 images). Chacun de ces sous-ensembles est filmé par une caméra mobile. Ainsi, les images référence et test proviennent de la même caméra mais avec des angles de vue différents. Malgré le fait que cette base de données soit bien utilisée pour évaluer les approches de ré-identification, elle reste non adaptée à un scénario de ré-identification réel car les images référence et test sont filmées par la même caméra.

i-LIDS-REID (Imagery Library for Intelligent Detection Systems) (Zheng *et al.*, 2009)
C'est une base de données privée filmée dans un hall d'arrivée à un aéroport. Elle est créée pour évaluer deux scénarios : 1) détection des évènements dans les zones critiques de l'aéroport et 2) suivi des personnes. (Zheng *et al.*, 2009) ont extrait un sous-ensemble de séquences vidéo d'i-LIDS et pour l'utiliser dans un scénario de ré-identification. Il est composé de 479 images correspondant aux 119 personnes et enregistrées par différentes caméras.

2.5.2 Base de données mono-échantillon

VIPeR : (Viewpoint Invariant Pedestrian Recognition) (Gray et Tao, 2008)

C'est une base de données publique. Elle comprend 632 personnes filmées par deux caméras disjointes sous différentes vues (l'angle de vue change de 45° jusqu'à 180°), différentes poses et différentes conditions d'illumination. Comme protocole d'évaluation, les résultats sur VIPeR sont typiquement produits par la moyenne de 10 tests, chacun prenant une partition de 316 personnes choisies aléatoirement. Si la méthode est supervisée, les 316 personnes restantes servent pour la phase d'apprentissage.

PRID-2011 (Person Re-ID 2011), version mono-échantillon : (Hirzer *et al.*, 2011)
Cette base de données a la même description et le même protocole d'évaluation que la base de données Prid-2011 version multi-échantillons.

La table 2.2 résume les différentes bases de données mono-échantillon et multi-échantillons citées dans cette section.

2.6 Choix de notre méthode par rapport à l'état de l'art

L'objectif de cette thèse est d'exploiter l'aspect complémentaire de l'apparence de la personne et de son type de mouvement pour la ré-identifier. L'état de l'art a montré que la plupart des méthodes de ré-identification sont fondées sur l'apparence globale de la personne ; et les plus sophistiquées combinent des caractéristiques de couleurs et de textures pour décrire la personne. Quant au mouvement de la personne dans le cas où les personnes sont représentées par des séquences vidéo, il n'a jamais été utilisé en ré-identification. Nous proposons dans cette thèse des approches multi-échantillons où on décrit la personne par des caractéristiques d'apparence extraites des images et des caractéristiques de mouvement extraites de la vidéo.

TABLE 2.2 – Récapitulatifs des bases de données de ré-identification.

Nombre d'échantillons par personne	**Base de données**	**Référence**	**Nombre de personnes**
multi-échantillons	CASIA-A	(CASIA, 2001)	20
	Prid-2011_Multi	(Hirzer *et al.*, 2011)	749 en référence et 200 en test
	CAVIAR4REID	(Cheng *et al.*, 2011)	72 dont 50 passent par les 2 caméras
	ETHZ_REID	(Schwartz et Davis, 2009)	Trois sous bases de données filmées par des caméras mobiles : 1. 83 personnes 2. 35 personnes 3. 28 personnes
	i-Lids_REID	(Zheng *et al.*, 2009)	119
mono-échantillon	VIPeR	(Gray et Tao, 2008)	632
	Prid-2011_Mono	(Hirzer *et al.*, 2011)	749 en référence et 200 en test

D'abord, quel type d'approche faut-il choisir, supervisée ou non supervisée ?

Comme vu précédemment, une approche supervisée nécessite la division de la base de données en deux parties : une partie d'apprentissage et une partie de test. Pour comparer notre approche avec l'état de l'art, il faut utiliser les mêmes partitions. Sachant que toutes les méthodes évaluées sur les bases de données multi-échantillons auquel nous avons accès (CASIA-A, CAVIAR4REID, Prid-20011) sont non-supervisées (toute la base de données est utilisée comme une base de test), on a choisi de proposer une approche non-supervisée pour pouvoir la comparer avec l'état de l'art.

Ensuite, quel type de représentation de la ROI doit-on choisir ; points d'intérêt ou division en régions ?

Une approche globale consiste à diviser la ROI en des régions. La signature de la ROI est la concaténation des représentations de toutes les régions qui la composent. Ceci signifie qu'une approche globale traite toutes les informations de la ROI de la même façon. Par contre, une représentation locale fondée sur les PIs ne considère que les points saillants, les points où il y'a l'information.

Prenons un exemple de notre contexte applicatif, une personne pousse une poussette. Quelque soit la méthode de division en régions, elle va affecter la poussette à une ou plusieurs régions et la traiter comme un partie du corps humain. Par contre une méthode locale ne considère que les points saillants de ROI (inclus la poussette) et pourra les traiter indépendamment. Pour cette raison, on a choisi de représenter la ROI par des PIs.

Enfin, quel type d'appariement des points d'intérêt doit-on choisir ; local ou global ?

Etant donnée une personne représentée par un ensemble de PIs, un appariement global consiste à représenter la personne par un seul vecteur sous forme d'histogramme. Les composantes du vecteur décrivent les occurrences d'un dictionnaire des classes des PIs appris sur la base de référence. Un appariement local consiste à apparier les points d'intérêt indépendamment.

Dans notre contexte applicatif, les personnes ont des poses différentes, sont filmées par différentes caméras. En outre, les conditions de luminosité sont très variées et les personnes peuvent marcher dans différentes directions. Par conséquent, une même personne peut avoir deux représentations globales très variées sous deux conditions différentes. Donc, on a choisi d'apparier localement les PIs en cherchant les PIs similaires entre la séquence vidéo test et la séquence vidéo référence.

Dans le chapitre suivant, nous détaillons l'étape d'extraction des primitives en décrivant les différentes primitives choisies pour la ré-identification.

Chapitre 3

Extraction des primitives

Contents

3.1 Introduction

L'extraction des primitives permet de transformer les données initiales disponibles (image ou vidéo) en une représentation discriminante, robuste au changement de l'éclairage, au changement de l'angle de vue, à la pose de la personne, etc. En littérature, cette représentation consiste à extraire de l'image de la personne des caractéristiques d'apparences. Ces caractéristiques peuvent être des primitives de couleurs (des vêtements, de la tête ou de la peau), de texture ou de forme géométrique du corps humain. Outre l'exploitation de ces caractéristiques d'apparence, l'objectif de cette thèse est d'intégrer dans le système de ré-identification des caractéristiques liées au mouvement de la personne afin d'exploiter la nature complémentaire de l'apparence statique de la personne et de la nature de son mouvement dans la vidéo.

Dans un scénario simple, le mouvement est décrit implicitement par un modèle stochastique qui simule l'aspect temporel de la marche de la personne, alors que dans un scénario complexe, le mouvement est décrit explicitement par des primitives à savoir les points d'intérêt spatio-temporels. L'utilisation de ces primitives ne nécessite aucun modèle à définir d'avance et aucun prétraitement postérieur tel que la soustraction du fond. Quant à l'apparence, elle est décrite, dans le scénario simple et complexe, par deux types de primitives à savoir les histogrammes de couleurs des différentes régions de l'image et les points d'intérêt spatiaux. Les PIs spatiaux sont d'une part l'équivalent des PIs spatiotemporels en 2D et d'autre part ils ont déjà montré des bonnes performances en ré-identification.

La mise en correspondance des PIs (spatiaux ou spatiotemporels) peut être locale ou globale selon la représentation de l'image/vidéo (figure 3.1). La mise en correspondance locale consiste à représenter l'image/vidéo par un ensemble de PIs et les apparier les uns indépendamment des autres. Quant à la mise en correspondance globale, elle consiste à représenter l'image/vidéo par un histogramme de PIs et ensuite à apparier les histogrammes.

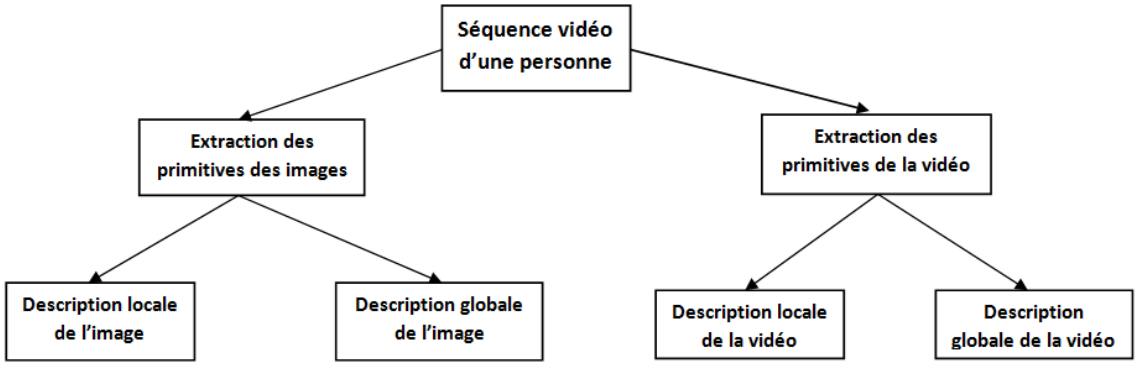


FIGURE 3.1 – Différentes types de représentations de l'image/vidéo.

Dans ce chapitre, nous montrons dans la section 3.2 les critères d'extraction des caractéristiques par points d'intérêt vis-à-vis la division en régions, ensuite nous présentons dans les sections 3.3 et 3.4 les primitives extraites respectivement pour décrire l'apparence et le mouvement.

3.2 Critères d'extraction des PIs vis-à-vis la division de l'image en des régions

La représentation de l'image par régions consiste à diviser l'image en des régions et à concaténer toutes les descriptions des régions dans un seul vecteur. Par contre, les points d'intérêt devraient représenter les zones les plus informatives de l'image/vidéo. Pour y arriver, deux étapes sont importantes pour générer les PIs : 1) détection des PIs et 2) description des PIs en exploitant leurs voisinages.

Un détecteur de PI doit générer les PIs qui représentent bien l'image et assurent un bon appariement plus tard. Un détecteur optimal de PI doit posséder les caractéristiques suivantes :

- **Une bonne répétabilité des PIs :** étant données deux images de la même personne prises par deux caméras différentes, le détecteur doit assurer un pourcentage important des PIs qui se répètent dans les deux images.
- **Quantité des PIs suffisante et optimale : Suffisante :** la quantité des PIs doit refléter la quantité d'information dans les images. **Optimale :** le détecteur ne doit détecter que les points saillants de l'image/vidéo (les points où il y a de l'information).
- **Efficacité :** la détection des PIs doit être réalisée en temps réel.

Quant à la description des PIs, elle doit être robuste aux variations relatives à la réidentification. En conclusion, la qualité de la représentation par PIs dépend conjointement des performances du détecteur et du descripteur. Ces deux derniers sont complémentaires pour réussir un bon appariement des PIs similaires dans deux images différentes. La figure

3.2 montre un exemple d'un bon appariement des PIs.

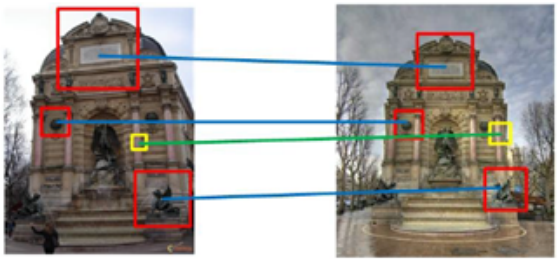


FIGURE 3.2 – Exemple d'un bon appariement de PIs sous deux conditions différentes.

Dans le reste du chapitre, nous présentons les différentes primitives extraites de l'image et de la vidéo pour décrire l'apparence (section 3.3) et le mouvement (section 3.4) de la personne .

3.3 Description de l'apparence à partir de l'image

Pour décrire l'apparence d'une personne, deux représentations sont utilisées : locale et globale. Ces deux représentations sont étudiées et comparées. Nous introduisons un bref état de l'art sur chacune d'elles tout en détaillant les primitives extraites.

3.3.1 Description locale par points d'intérêt 2D

La description locale d'une image exploite quelques PIs localement détectés ou des patchs (petite régions) extraits de l'image. Ces PIs ou patchs sont les régions saillantes de l'image. Nous décrivons ci-dessous les PIs 2D les plus utilisés dans la littérature, puis nous détaillons les PIs considérés dans cette thèse.

3.3.1.1 Etat de l'art des points d'intérêt 2D

De nombreuses méthodes de détection et description de points d'intérêt 2D sont utilisées dans la littérature. Ces méthodes sont généralement appliquées sur l'image intensité (image en niveau de gris). Dans cette section, nous présentons les détecteurs et les descripteurs les plus populaires.

Détecteurs

Deux types de détecteurs de PIs ont été présentés dans la littérature : détecteurs de coins et détecteurs de blobs. La première catégorie considère des coins comme PIs. Les coins d'une image correspondent aux pixels ayant des changements d'intensité dans toutes les directions. Cependant, les coins de l'image sont parfois épais et donc difficilement détectables par ce type de détecteurs. La deuxième catégorie surmonte ce problème en détectant des petites zones d'intérêt (blobs) plutôt que des pixels. Dans les deux catégories, les travaux initiaux furent fondés sur des détecteurs à échelle fixe. Ils supposent qu'il n'y a aucun changement d'échelle prévu sur la caméra. Plus tard, (Lindeberg, 1998) a introduit la notion d'analyse multi-échelle de l'image, qui permet d'apparier deux images de différentes échelles. Rapidement les détecteurs à échelle fixe ont évolué vers une version multi-échelle.

Détecteur de coins Parmi ces détecteurs, on cite le détecteur de Harris (Harris et Stephens, 1988) et le détecteur SUSAN (pour « Smallest Univalue Segment Assimilating Nucleus») (Smith et Brady, 1997). Le détecteur de Harris est fondé sur le calcul des gradients locaux pour chercher les coins. Un pixel est considéré comme un PI si ses valeurs de gradients dans toutes les directions sont grandes. Quant au détecteur de SUSAN, il est fondé sur des statistiques de la similarité des intensités des pixels au voisinage d'un pixel noyau. Selon le pourcentage des pixels ayant une intensité similaire au pixel noyau, ce dernier est considéré soit comme un pixel de contour, soit comme un pixel appartenant à une région homogène soit comme un PI. Ces deux détecteurs ne tiennent pas compte du changement d'échelle entre les images. Plus tard, après la proposition de l'analyse multi-échelle des images dans (Lindeberg, 1998), le détecteur de Harris a été adapté dans (Mikolajczyk et Schmid, 2004) pour être invariant à l'échelle en définissant le détecteur Harris-Laplace. Ce dernier détecte des PIs invariants aux transformations euclidiennes (rotation, translation) et à l'échelle. Son idée principale consiste à appliquer le détecteur de Harris à plusieurs échelles (construction d'une pyramide d'images). Les PIs sélectionnés correspondent aux maximums locaux d'une fonction qui fait intervenir le déterminant et la trace d'une matrice à base de gradients.

Détecteur de blobs Ce type de détecteurs permet de rechercher des petites zones homogènes en intensité plutôt que de se limiter à des coins. Ces détecteurs comblent les lacunes des détecteurs de coins au niveau des zones lisses sans transition apparente. Le premier détecteur de blobs utilisé dans la littérature est fondé sur la matrice Hessienne. C'est un détecteur à échelle fixe. Son critère de sélection des PIs est fondé sur le déterminant de la Hessienne. En effet, les PIs retenus correspondent aux maximums locaux du déterminant de la Hessienne calculé pour chaque pixel. Par analogie au Harris-Laplace, Hessienne-Laplace intègre la notion d'échelle en appliquant la Hessienne sur plusieurs échelles et garde les maximums locaux d'une fonction espace-échelle.

D'autre part, des travaux ont été proposés pour approximer le Laplacien de Gaussienne (LoG pour «Laplacian of Gaussian») qui intervient dans le calcul du déterminant de la Hessienne. Dans (Lowe, 2001), le LoG est approximé par la différence des Gaussiennes (DoG pour «Difference of Gaussians»). Ce détecteur multi-échelle fondé sur le DoG est connu dans la littérature par SIFT. L'utilisation de DoG pour le calcul du déterminant de la Hessienne rend la détection des PIs plus rapide car elle évite le calcul des dérivées secondes de l'image. SIFT est connu par son invariance aux transformations euclidiennes (rotation, translation) et à l'échelle. (Bay *et al.*, 2006) proposent le détecteur SURF. Ce détecteur propose une méthode rapide pour approximer le LoG par le calcul des produits de convolution en utilisant un ensemble de filtres de type «box» (figure 3.7-a). L'utilisation de l'image intégrale proposée dans (Bay *et al.*, 2006) rend le calcul des produits de convolution très rapide.

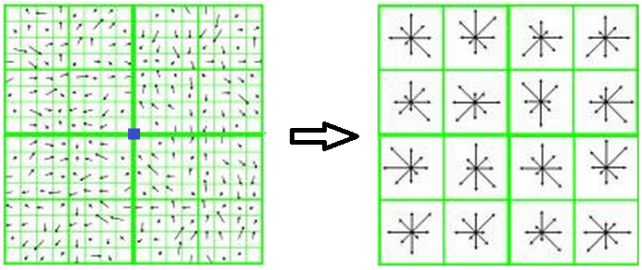
Descripteurs

Différents descripteurs ont été utilisés dans la littérature pour décrire les PIs. Certains descripteurs exploitent directement les intensités des pixels au voisinage du PI en les concaténant dans un seul vecteur ou en construisant un histogramme de distribution de ces intensités.

Schmid et al. (Schmid et Mohr, 1997) ont proposé une description pour le détecteur de Harris. À partir d'une région au voisinage du PI, un ensemble d'opérateurs différentiels

est appliqué pour extraire une description invariante à la rotation (Florack *et al.*, 1996).

D'autres descriptions, comme le descripteur SIFT (figure 3.3), sont fondés sur le calcul de gradients des pixels de la région entourant le PI. La taille de la région considérée est un multiple de l'échelle du PI et son orientation est celle de l'orientation dominante des gradients des pixels de cette région. Ensuite, cette région est divisée en 4 ou 16 zones. Pour chaque zone, les orientations des gradients sont quantifiées sur 8 valeurs et ensuite un histogramme de 8 composantes est calculé où chaque composante représente la somme des amplitudes des gradients des pixels associés à une orientation donnée.



(a) (b)

FIGURE 3.3 – Construction du descripteur SIFT : (a) l'image gradient des pixels autour du PI en bleu, (b) descripteur SIFT du PI : 16 histogrammes de 8 orientations.

(Mikolajczyk et Schmid, 2005) ont proposé le descripteur GLOH (pour «Gradient Location and Orientation Histogram») proche du descripteur SIFT sauf que la région considérée au voisinage du PI est divisée en une grille log-polaire au lieu d'une grille de régions rectangulaires. Un autre descripteur dit Shape-Context (SC) (Belongie *et al.*, 2002) est fondé sur le contour de Canny au lieu des gradients. Sa construction est similaire au SIFT.

(Bay *et al.*, 2006) ont proposé le descripteur SURF, fondé sur le calcul des réponses d'ondelettes de Haar des pixels d'une région entourant le PI. Comme dans le SIFT, cette région est divisée en une grille de zones. En chaque zone, des caractéristiques sont extraites. Le descripteur final consiste à concaténer les caractéristiques extraites de toutes les zones.

3.3.1.2 Points d'intérêt 2D choisis

L'état de l'art a mentionné que SURF et SIFT sont les PIs les plus populaires. Ils ont montré de bonnes performances dans la reconnaissance des objets. Ils sont tous les deux robustes aux transformations géométriques et à l'échelle et ont une bonne répétabilité. Néanmoins, SURF est le plus robuste au changement des angles de vues et le plus rapide (Tuytelaars et Mikolajczyk, 2008). Il a aussi été utilisé pour la ré-identification et a montré des bonnes performances (Hamdoun *et al.*, 2008, de Oliveira et de Souza Pio, 2009). Pour ces raisons, nous sélectionnons SURF pour décrire l'apparence locale de la personne. Dans le cas où il n'y a pas de variation d'échelle dans les images, nous sélectionnons les PIs générés par le détecteur de Harris. Ce sont des PIs non invariants à l'échelle mais ils ont une très bonne répétabilité. Ces PIs sont proposés dans la littérature sans description associée, nous profitons dans ce travail des avantages des SURF pour les décrire. Nous détaillons dans la suite le principe de détection et description de SURF ainsi que le détecteur de Harris.

Extraction des SURFs

L'extraction des SURFs est composée de deux étapes principales : une étape de détection et une étape de description. D'abord, le détecteur analyse l'image et renvoie un ensemble de PIs. Ensuite, au voisinage de chaque PI, un vecteur descripteur est calculé. L'étape de détection est fondée sur la matrice Hessienne, alors que la description exploite les réponses d'ondelettes de Haar. Dans la suite, le détecteur et le descripteur SURF sont détaillés.

Détection Le détecteur SURF repose sur l'approximation du déterminant de la matrice Hessienne et l'utilisation de l'image intégrale. Il présente un bon compromis entre la robustesse aux transformations géométriques et le temps de calcul. Ce détecteur intègre l'échelle dans la matrice Hessienne.

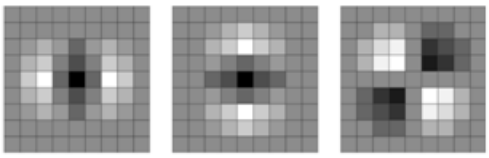
Soit p un pixel d'une image intensité I de coordonnées (x, y), la matrice Hessienne $H_\sigma(p)$ en p et à l'échelle σ est donnée par l'expression suivante :

$$H_\sigma(p) = \begin{bmatrix} L_{xx}(p,\sigma) & L_{xy}(p,\sigma) \\ L_{yx}(p,\sigma) & L_{yy}(p,\sigma) \end{bmatrix}$$

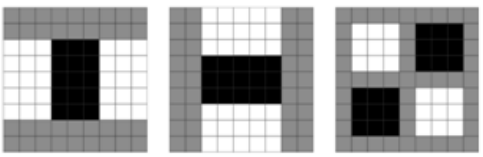
où $L_{xx}(p, \sigma)$ dénote le produit de convolution de l'image intensité I par la dérivée de deuxième ordre d'une Gaussienne $\frac{\partial^2 g(\sigma)}{\partial x^2}$, similairement à L_{yy} et L_{yx}. Ces dérivées sont connues dans la littérature par *LoG*. Comme critère de sélection de PIs, SURF utilise le déterminant de la Hessienne. Par contre au lieu d'utiliser le LoG pour approximer le déterminant de la Hessienne, il propose d'utiliser un ensemble de filtres de type «box» (figure 3.4). Le déterminant approximé est donné par l'expression suivante :

$$\det(H_{approx}) = D_{xx}D_{yy} - (wD_{xy})^2$$

où D_{xx}, D_{yy} et D_{xy} dénotent les résultats du produit de convolution de l'image par un ensemble de filtres de type «box» dans les directions x, y et xy et w est une constante choisie empiriquement.



(a)



(b)

FIGURE 3.4 – Approximation de LoG. De gauche à droite : (a) L_{xx}, L_{yy}, L_{xy}, (b) : approximation de L_{xx}, L_{yy}, L_{xy}. Elles sont appelées D_{xx}, D_{yy}, D_{xy} (Bay *et al.*, 2006).

La spécificité de la méthode SURF est l'utilisation de l'image intégrale pour accélérer le calcul du produit de convolution. La valeur de l'image intégrale $I_\Sigma(x)$ en une position

$p = (x, y)$ représente la somme des intensités des pixels du rectangle formé par l'origine O et p.

$$I_\Sigma\left(p = (x, y)\right) = \sum_{i=0}^{x}\sum_{j=0}^{y} I(i, j)$$

En utilisant l'image intégrale, uniquement 4 opérations d'additions sont nécessaires pour calculer la somme Σ des intensités d'une région rectangulaire indépendamment de ses dimensions (figure 3.5).

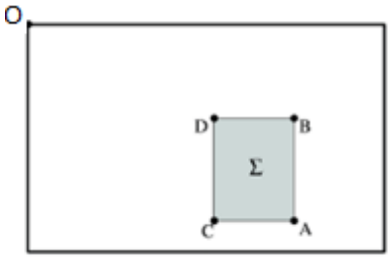


FIGURE 3.5 – Exemple d'utilisation de l'image intégrale (Σ = $I_\Sigma(A)$ - $I_\Sigma(B)$ - $I_\Sigma(C)$ + $I_\Sigma(D)$).

Une fois estimé un déterminant en chaque pixel de l'image I, les maximums sont recherchés en espace et en échelle (recherche dans des petits voisinages, typiquement des volumes de 3x3x3 pixels). Ces maximums correspondent aux PIs SURF. Ensuite, chaque PI est décrit comme suit.

Description Le descripteur SURF capture des informations caractérisant la région autour d'un PI détecté avec l'échelle σ. La description se déroule en deux phases : attribution d'une orientation et extraction du descripteur.

Tout d'abord, afin de réaliser une invariance de l'image à la rotation, une orientation caractéristique du descripteur est estimée. En effet, pour chaque pixel d'une région circulaire de rayon $6^*\sigma$ autour du PI, les réponses d'ondelettes de Haar sont calculées et ensuite pondérées par une gaussienne centrée au PI. Chaque réponse pondérée est interprétée comme un vecteur de deux éléments : d_x et d_y. Ces derniers sont les réponses d'ondelette de Haar au long de l'axe des abscisses et de l'axe des ordonnées respectivement. Ensuite, en utilisant une approche à base de fenêtre glissante, chacune des six fenêtres d'ouverture angulaire égale à $\frac{\pi}{3}$ est représentée par un vecteur de deux composantes : la composante horizontale (respectivement verticale) est égale à la somme de tous les d_x (respectivement tous les d_y) de tous les pixels de la fenêtre. Le vecteur ayant une amplitude maximale sur toutes les fenêtres détermine l'orientation caractéristique du descripteur (figure 3.6).

Dans la deuxième étape, l'extraction du descripteur, nous considérons une région carrée autour du PI, orientée vers l'angle déjà calculé dans l'étape précédente (angle θ dans la figure 3.7-b). La taille de cette région est choisie en fonction de σ. Ensuite, elle est divisée en une grille de 4x4 pour former 16 sous-régions. De chaque sous-région, les réponses d_y et d_x des ondelettes de Haar sont calculées pour 5x5 pixels échantillonnés uniformément de la sous-région. Pour ces 25 points, nous calculons les 4 éléments $v_1,\ v_2,\ v_3$ et v_4 comme suit (figure 3.7) :

$$v_1 = \sum d_x,\ v_2 = \sum d_y,\ v_3 = \sum |d_x|\,,\ v_4 = \sum |d_y|$$

Finalement, tous les éléments collectés à partir de chacune des 16 sous-régions forment le descripteur SURF de 64 dimensions.

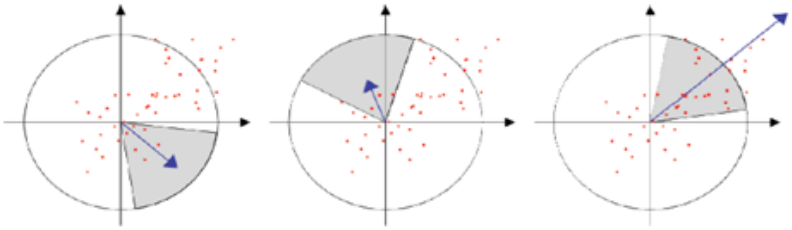
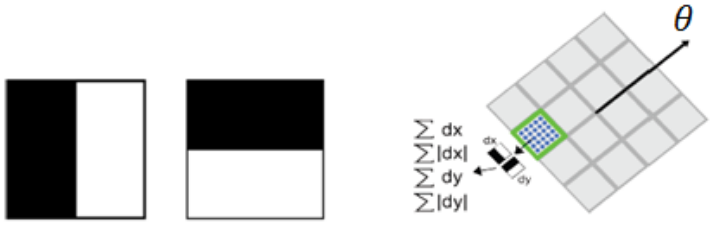


FIGURE 3.6 – Attribution d'une orientation. Dans cet exemple, l'orientation du vecteur bleu à droite est choisie comme l'orientation caractéristique (figure extraite du (Tuytelaars et Mikolajczyk, 2008)).



(a) (b)

FIGURE 3.7 – (a) Filtres de type «box» pour calculer d_x (à gauche) et d_y (à droite). Les pixels noirs ont un poids -1 et les pixels blancs ont un poids +1. (b) Extraction des descripteurs : (à droite) 4x4 sous-régions autour du PI. La sous-région verte est une sous-région orientée vers θ avec les 5x5 réponses des ondelettes de Haar, (à gauche) éléments extraits de chaque sous-région.

Extraction des Harris

Le détecteur de Harris est un détecteur de coins fondé sur la matrice de second moment, dite aussi la matrice d'auto-corrélation dans la littérature. Étant donnée une image intensité I et p un pixel de I de coordonnées (x, y), la matrice de second moment en p est définie ainsi :

$$M(p) = \begin{bmatrix} I_x^2(p) & I_x I_y(p) \\ I_x I_y(p) & I_y^2(p) \end{bmatrix}$$

où I_x et I_y sont les dérivées partielles de I dans respectivement la direction horizontale et la direction verticale.

Algorithmiquement, le détecteur de Harris est composé des étapes suivantes. D'abord, l'image I est filtrée par une gaussienne, ensuite la matrice de second moment est calculée pour chaque pixel de l'image filtrée. La nouvelle matrice est appelée la matrice de Harris, son expression en un pixel p est la suivante :

$$H(p) = G_\sigma * M(p) = G_\sigma * \begin{bmatrix} I_x^2(p) & I_x I_y(p) \\ I_x I_y(p) & I_y^2(p) \end{bmatrix}$$

où G_σ est un filtre gaussien défini par l'expression suivante :

$$G_\sigma = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

Physiquement, la valeur de la matrice de Harris en un pixel mesure la distribution des gradients au voisinage du pixel (Tuytelaars et Mikolajczyk, 2008).

Ensuite, (Harris et Stephens, 1988) ramènent le problème de détection des PIs à l'étude des valeurs propres de la matrice de Harris. Les valeurs propres de cette matrice représentent les variations de l'intensité de l'image autour du pixel dans les deux directions associées aux deux valeurs propres. Soient λ_1 et λ_2 les deux valeurs propres de la matrice de Harris. Selon les valeurs de λ_1 et λ_2, trois cas peuvent se présenter (figure 3.8) :

a) Si les deux valeurs propres sont faibles → faible changement d'intensité quelle que soit la direction → le pixel appartient à une région homogène.

b) Si une des valeurs propres est très grande par rapport à la deuxième → un changement d'intensité dans une seule direction → le pixel est un point de contour.

c) Si les deux valeurs propres sont très grandes → il n'y a pas une direction privilégiée → changement d'intensité significatif dans les deux directions → le pixel est considéré comme un coin (PI).

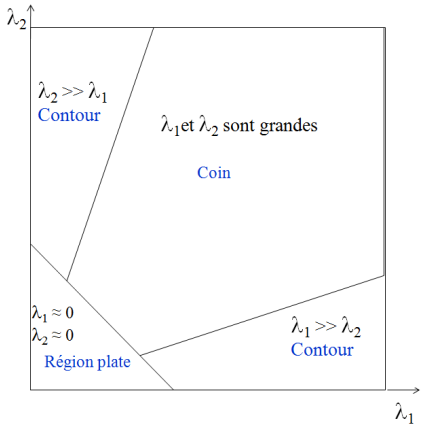


FIGURE 3.8 – Distribution de types des pixels en fonction des valeurs propres de la matrice de Harris (Harris et Stephens, 1988).

Pour tenir compte de ces trois cas, un critère de sélection de PIs est introduit dans (Harris et Stephens, 1988). Il s'agit de mesurer une fonction de détection de coins «c» qui combine les deux valeurs propres de la matrice de Harris en ce pixel. La fonction de détection de coins (en anglais «Cornerness Function») d'un pixel p est donnée par l'expression suivante :

$$c = \det(H(p)) - k \ \text{trace}(H(p))$$

où $H(p)$ est la matrice de Harris en p, $\det(H(p))$ est le déterminant de $H(p)$ et $\text{trace}(H(p))$ est la trace de $H(p)$, k est un paramètre permettant de gérer la sensibilité de détection des PIs. Selon plusieurs auteurs, la valeur empirique de k est égale à 0.04.

Finalement, après avoir calculé la valeur de «c» pour chaque pixel, les PIs correspondant aux maximums locaux de la fonction «c» (théoriquement les pixels dont les matrices de Harris ont deux grandes valeurs propres). Pour trouver ces maximums, une étape de suppression de non-maximum est appliquée et elle est généralement précédée par une étape de seuillage pour filtrer les pixels candidats à être des PIs. Ce détecteur génère un grand nombre de coins et il est connu par sa répétabilité suffisante.

Quant à la description des PIs, le détecteur de Harris est proposé dans la littérature sans descripteur associé. Pour les comparer aux SURFs, nous avons décrit ces PIs par le descripteur SURF. Comme le détecteur de Harris génère des PIs sans échelle, alors que le descripteur de SURF nécessite une échelle de détection pour calculer le descripteur, nous avons utilisé une échelle égale à 1 pour tous les PIs. Ce choix est justifié par le fait que le détecteur de Harris est appliqué dans un contexte où il y'a pas de variation d'échelles entre les images.

3.3.2 Description globale

La description globale de l'image consiste à combiner toutes les caractéristiques extraites de l'image dans un seul vecteur de dimension fixe. Ces caractéristiques peuvent être des statistiques globales sur l'intégrité de l'image, une combinaison des caractéristiques de l'image après découpage en des régions ou des caractéristiques extraites localement (PIs dans notre cas). Dans la suite, nous présentons un bref état de l'art des méthodes globales de description de l'image. Ensuite, nous détaillons les méthodes sélectionnées dans ce travail.

3.3.2.1 Etat de l'art

La représentation globale par des caractéristiques statistiques consiste à décrire l'image à partir des mesures statistiques de ses pixels. Parmi ces méthodes, on cite les histogrammes dans différents espaces de couleurs (RGB, HSV, YCbCr, etc.). Ces histogrammes sont normalisés pour assurer une invariance aux variations de l'éclairage. On cite aussi les moments de l'image qui sont des caractéristiques statistiques donnant une information globale sur les variations des intensités de l'image. Par exemple, la variance ou la moyenne des intensités sont des moments très utilisés. Ces statistiques peuvent être déterminées en analysant l'image directement ou après un filtrage préalable. La table 3.1 montre quelques caractéristiques statistiques de l'image qui se déduisent de l'intensité n et de la probabilité $p(n)$.

TABLE 3.1 – Quelques moments statistiques.

Caractéristiques	**Formule**
Moyenne	μ_1
Variance	$\sigma^2 = \mu_2$
Contraste	$\frac{\max(n)-\min(n)}{\max(n)+\min(n)}$
Energie	$w = \sum^n p^2(n)$
Entropie	$E = -\sum^n p(n) \log p(n)$
Moments d'ordre k	$\mu_k = \sum^n n^k p(n)$

La représentation globale d'une image peut aussi être la concaténation des descriptions de ses différentes régions. Elle consiste à décomposer l'image en plusieurs régions de taille fixe ou variable.

Quand l'image est divisée en des régions de même taille (régions rectangulaires ou des bandes horizontales), des caractéristiques sont extraites de chaque région et ensuite concaténées pour former la représentation globale de l'image. Parmi ces caractéristiques, on cite l'utilisation de l'histogramme de gradients orientés (HOG pour «Histogram of Oriented Gradients») (Dalal et Triggs, 2005) dans une application de détection des personnes ; les pseudo-Haar (Oren *et al.*, 1997) sont aussi utilisées pour détecter des objets ou des visages (Viola et Jones, 2001) dans une image. En ré-identification, on cite comme caractéristiques extraites des régions, les histogrammes de couleur et les filtres de texture (Gray et Tao, 2008, Hirzer *et al.*, 2012).

Quant à la division de l'image en régions de différentes tailles, l'idée consiste à supposer que le corps humain est composé de régions non homogènes. Les régions sont généralement liées aux différentes parties du corps humain. La figure 3.9 montre deux exemples de division de l'image en des régions, proposés dans (Alonso *et al.*, 2007) et (Shashua *et al.*, 2004). Ces deux approches sont utilisées dans une application de description des piétons. En ré-identification, quelques méthodes de l'état de l'art consistent à diviser l'image en des régions qui correspondent idéalement aux différentes parties du corps humain (Cheng *et al.*, 2011, Huang *et al.*, 2008).

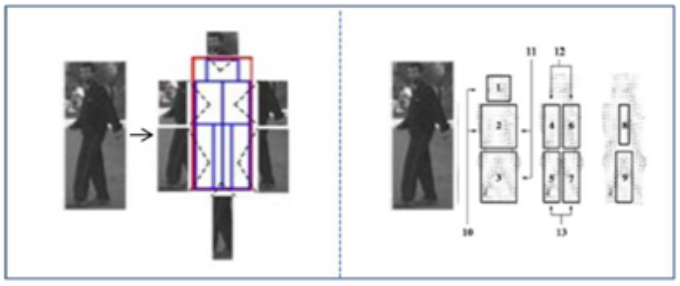


FIGURE 3.9 – Deux exemples de division de l'image en régions (proposés respectivement dans (Alonso *et al.*, 2007) et (Shashua *et al.*, 2004)).

La représentation globale peut être aussi fondée sur une extraction des caractéristiques locales telles que les PIs. En effet, les PIs détectés dans une image ne sont pas ordonnés et leur nombre est variable. Dans ce cas, la similarité entre deux images ne peut être calculée que si toutes les caractéristiques locales extraites de l'image sont transformées en une seule représentation de dimension fixe pour toutes les images. La méthode la plus répandue est proposée dans (Sivic et Zisserman, 2003), et consiste à décrire l'image par un histogramme où chaque composante décrit la fréquence d'apparition d'un mot visuel. Ces mots visuels sont les centres des classes (clusters) apprises préalablement par un algorithme de classification non supervisée.

3.3.2.2 Descriptions globales choisies

Deux représentations globales ont été considérées dans notre travail à savoir l'histogramme des mots visuels par PIs et l'histogramme de couleurs. Dans la suite, nous détaillons l'extraction des deux descriptions.

Histogramme des mots visuels par PIs : Cette description globale prend en entrée les points d'intérêt SURF ou Harris. Ensuite, elle représente l'image par un seul vecteur

de dimension fixe, choisie empiriquement. Cette représentation est inspirée de la représentation «sac de mots» introduite dans (Joachims, 1998) pour représenter un document texte. La représentation par un «sac de mots» consiste à représenter un document texte par un histogramme décrivant la fréquence d'apparition de chaque mot d'un vocabulaire. Par analogie au domaine de traitement d'image, cette approche a été utilisée pour représenter une image par un histogramme décrivant le nombre d'apparition de chaque PI d'un vocabulaire (Sivic et Zisserman, 2003). La méthodologie reste la même mais son nom est devenu «sac de mots visuels» (BoF pour «Bag of Features»). La génération des BoFs repose sur 3 étapes principales (figure 3.10) :

a) Le PIs sont extraits sur l'ensemble des images de référence.

b) Un partitionnement en k classes (clusters) de tous les PIs références est créé par l'algorithme k-means (**?**). Les centres de ces classes sont les mots visuels et représenteront les k dimensions de BoF qui servent à décrire l'image.
k-means est un algorithme de partitionnement itératif qui minimise la somme des distances entre chaque PI et son centre associé.
Les mots visuels sont appris sur la base de référence et servent à représenter toutes les images de référence et de test.

c) Pour caractériser une image référence ou test, chaque PI de l'image est affecté au mot visuel le plus proche par la distance euclidienne en incrémentant le nombre d'occurrences de ce mot visuel. L'image est caractérisée par l'histogramme décrivant le nombre d'occurrence de chaque mot visuel.

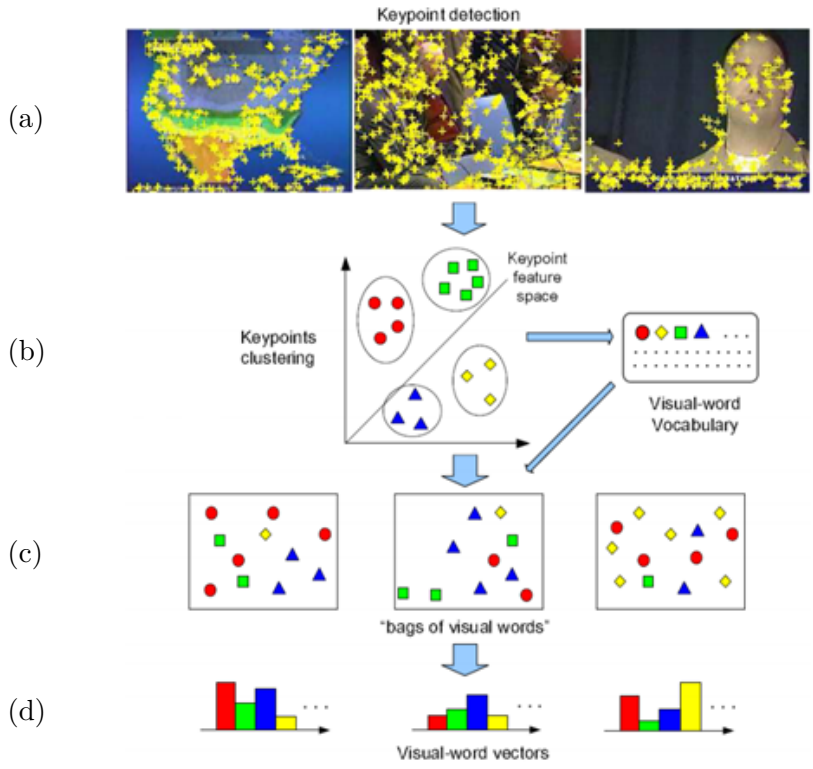


FIGURE 3.10 – Principe de génération des BoFs (figure issue de (Yang *et al.*, 2007). (a) Extraction des PIs des images références. (b) Partitionnement en k classes. (c) Affectation des PIs aux classes. (d) Construction des BoFs.

Le seul paramètre à fixer empiriquement dans cette représentation est le nombre de classes k à créer par k-means. Il ne doit pas être très grand pour éviter que les classes vides (ou de faible cardinal) soient nombreuses ; et il ne doit pas être très faible pour assurer une représentation assez variée.

Histogramme de couleurs dans l'espace RGB : L'histogramme de couleurs mesure la distribution des couleurs de l'image ; par contre il ne tient pas compte de l'information spatiale des pixels dans sa construction. Pour intégrer une information spatiale, nous proposons de calculer un histogramme de couleurs pour chaque région de l'image et ensuite nous concaténons les histogrammes de toutes les régions dans un seul vecteur. Etant donnée une image couleur I où le pixel à la position (i, j) est représenté par un triplet $(I_R(i,j), I_G(i,j), I_B(i,j)) \in [0..255]^3$ correspondant aux couleurs des canaux RGB, la construction de son histogramme de couleur se déroule en trois étapes :

1. Les couleurs des pixels dans l'intervalle [0..255] sont quantifiées dans l'intervalle [0..n], où n est une valeur empirique.
2. Pour chaque région de l'image de hauteur H et de largeur W, on construit les sous-histogrammes h_R, h_G et h_B comme suit :

$$h_R(c) = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} \delta(I_R(i,j), c) \ , \ c \in \ [0..n]$$

$$h_G(c) = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} \delta(I_G(i,j), c) \ , \ c \in \ [0..n]$$

$$h_B(c) = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} \delta(I_B(i,j), c) \ , \ c \in \ [0..n]$$

   où $\delta(x, y) = 1$ si $x = y$, et 0 sinon.
3. La concaténation des sous-histogrammes h_R, h_G et h_B définit l'histogramme de couleur de la région.

3.4 Description de mouvement à partir de la vidéo

Le mouvement d'une personne peut être décrit de deux façons. 1) Description implicite par un modèle stochastique qui représente l'aspect temporel de la vidéo à partir des séquences d'observations. 2) Description explicite à partir des primitives, celles-ci capturent l'information spatiotemporelle pertinente pour décrire le type du mouvement.

3.4.1 Description implicite à partir d'un modèle

Le type de mouvement peut être décrit implicitement par un modèle à savoir les modèles à base d'automates d'états. Ces derniers permettent de modéliser la séquentialité du mouvement à partir des séquences d'observations de vecteurs descripteurs. Plusieurs modèles statistiques ont été proposés dans la littérature et se divisent en deux familles : méthodes discriminantes et méthodes génératives (Bishop, 2006). Nous proposons d'utiliser la méthode des «modèles de Markov à états cachés» (Hidden Markov Models (HMM)) qui est de loin la méthode la plus utilisée en reconnaissance de mouvements (Sorel, 2012). Tout d'abord, chaque HMM apprend le type du mouvement d'une personne en lui donnant

des séquences d'observations *type* de son mouvement. Ensuite, ces HMMs appris sont utilisés pour reconnaître une personne à partir d'une séquence d'observations d'identité inconnue. Les observations entrée du HMM sont des caractéristiques de même dimension, extraites des images. Les représentations les plus adéquates de l'état de l'art sont les représentations globales. Dans notre cas, nous représentons l'image par le descripteur HOG (Dalal et Triggs, 2005) qui a montré de bons résultats dans une application de détection des personnes.

L'histogramme des gradients orientés (HOG)

C'est une caractéristique décrivant la texture globale de l'image. Elle a été utilisée initialement dans (Dalal et Triggs, 2005) pour la détection des personnes. Cette représentation consiste à diviser l'image en une grille de régions et ensuite concaténer les HOGs de ces régions dans un seul vecteur. Etant donnée une image intensité I, le calcul de son HOG se déroule en cinq étapes principales : 1) calcul de l'image gradient 2) division de l'image en une grille de régions 3) calcul d'un HOG par région, 4) normalisation du HOG de chaque région et 5) concaténation des HOGs de toutes les régions.

Calcul de l'image gradient

En mathématiques, le gradient est un vecteur représentant la variation d'une fonction par rapport à la variation de ses différents paramètres (Gradient, edia). Selon cette définition, le gradient d'une image est un vecteur représentant la variation de l'intensité par rapport au déplacement dans la direction horizontale et verticale. En un pixel p en position (x, y) de I, le gradient est un vecteur à deux composantes $(G_V(x, y), G_H(x, y))$. $G_V(x, y)$ est la composante verticale du gradient décrivant la variation verticale des intensités de l'image autour de p. $G_H(x, y)$ est la composante horizontale du gradient décrivant la variation horizontale des intensités de l'image autour de p. Pour trouver ces deux composantes en p, deux filtres centrés en p sont appliqués :

- Horizontalement, on applique le filtre $(-1\ 0\ 1)$
- Verticalement, on applique le filtre $(-1\ 0\ 1)^T$

En appliquant ces deux filtres, $G_H(x, y)$ et $G_V(x, y)$ sont donnés par les deux expressions suivantes.

$$G_H(x, y) = I(x + 1, y) - I(x - 1, y)$$
$$G_V(x, y) = I(x, y + 1) - I(x, y - 1)$$

Le calcul d'un gradient pour chaque pixel de l'image définit l'image gradient. Les composantes G_H (respectivement G_V) de tous les pixels d'une image intensité (figure 3.11-b) sont affichées dans la figure 3.11-c (respectivement la figure 3.11-d). Les zones noires des deux figures (3.11-c et 3.11-d) correspondent aux zones homogènes de la figure 3.11-b .

Comme tout vecteur, le gradient en p est défini par sa norme $N_G(x, y)$ et son orientation $\theta_G(x, y)$ données par les deux expressions suivantes en fonction de $G_H(x, y)$ et $G_V(x, y)$:

$$\theta_G(x, y) = \text{atan}\,\frac{G_H(x, y)}{G_V(x, y)}$$

$$N_G(x, y) = \sqrt{G_H(x, y)^2 + G_V(x, y)^2}$$

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

FIGURE 3.11 – (a) Image originale, (b) image intensité, (c) gradients horizontaux, (d) gradients verticaux.

Division de l'image en une grille de régions

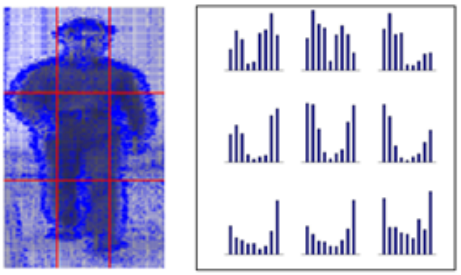
Après le calcul de l'image gradient, cette dernière est découpée en une grille de régions couvrant toute l'image. Deux divisions sont possibles : 1) fixer les dimensions des régions : dans ce cas l'image est normalisée afin d'avoir le même nombre de régions quelles que soient les dimensions de l'image, 2) fixer les dimensions de la grille : on obtient le même nombre de régions quelles que soient ses dimensions. Dans nos expériences, nous avons fixé les dimensions de la grille en 3x3 au lieu de fixer les dimensions des régions (figure 3.12-a).

Calcul d'un HOG par région

Pour chaque région, un HOG est calculé en utilisant les gradients de tous ses pixels. Chaque pixel en position (x, y) participe au calcul de son HOG de n composantes de la manière suivante :

$$HOG(a) = HOG(a) + N_G(x, y)$$

où $a \in [1, n]$ et vérifiant la condition suivante : $\theta_G(x, y) \in [\Delta_a, \Delta_{a+1}[$ et $\Delta_a = \frac{(a-1)\pi}{n}$. Le nombre de composantes du HOG est paramétrable. Il permet de régler la précision de l'orientation des gradients. La figure 3.12-b montre la configuration du HOG utilisé dans ce travail. L'image gradient est divisée en une grille de 3x3 régions, et de chaque région on extrait un HOG de 9 composantes.



(a) (b)

FIGURE 3.12 – (a) : Image gradient divisée en 3X3 régions. (b) Un HOG de 9 composantes est extrait de chaque région.

Normalisation du HOG de chaque région

Cette étape consiste à normaliser le HOG de chaque région indépendamment des autres. (Dalal et Triggs, 2005) mentionne plusieurs normalisations en se basant sur les normes L_1 et L_2 :

- Norme L_1 : $\overline{HOG} = \frac{HOG}{\|HOG\|_1 + \varepsilon}$
- Norme L_2 : $\overline{HOG} = \frac{HOG}{\sqrt{\|HOG\|_2^2 + \varepsilon}}$

où $\overline{HOG}$ dénote le HOG normalisé, et ε est un terme de régularisation. Dans nos expériences, nous avons normalisé les HOGs en utilisant la norme L_2.

Concaténation des HOGs

Après avoir calculé les $\{\overline{HOG}_i,\ 1 \leq i \leq M\}$ normalisés des M régions de l'image, le HOG de l'image HOG_{image} consiste à concaténer les $\overline{HOG}_i$ des M régions :

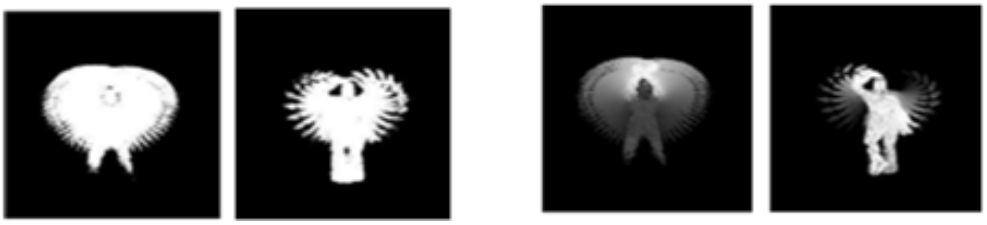
$$HOG_{image} = [\overline{HOG}_1, \overline{HOG}_2, \ldots, \overline{HOG}_M]$$

3.4.2 Description à partir des primitives

Le mouvement peut être décrit par des primitives extraites de la vidéo. Dans ce cadre, deux catégories d'approches existent dans l'état de l'art.

La première catégorie concerne les approches fondées sur un modèle cinématique ou de forme. Un modèle cinématique est un modèle 3D qui établit des joints (angles) entre les différents segments du corps humain (Atine, 2004, Wang *et al.*, 2003a). Quant au modèle de forme, il consiste à représenter les segments du corps humain par des formes géométriques 2D telles que des rectangles (modèles rectangulaires (İkizler et Duygulu, 2007)), ou des formes géométriques 3D telles que des cylindres (Pehlivan et Duygulu, 2009). L'avantage de ces approches est le fait qu'elles décrivent explicitement le mouvement de la personne ; mais malheureusement, elles sont souvent coûteuses en raison du grand nombre de paramètres qui doivent être calculés, de plus elles sont non robustes aux occultations, nécessitent le suivi de la personne et ont des difficultés pour déterminer les positions des joints dans les bras et les jambes (Yang *et al.*, 2008). La deuxième catégorie d'approches ne nécessite aucun modèle à définir d'avance. Ces approches extraient directement des primitives de la vidéo sans passer par un modèle et sans la nécessité de la soustraction du fond. Dans ce cas, les primitives peuvent décrire le mouvement par une représentation globale ou locale.

La représentation globale consiste à transformer toute la séquence vidéo en une seule image temporelle (appelée aussi "carte"). L'état de l'art mentionne deux représentations populaires fondées sur les images temporelles (Bobick et Davis, 1996, 2001) : l'image énergie (MEI pour «Motion Energy Image») et l'image historique du mouvement (MHI pour «Motion History Images»). Les MEIs (figure 3.13-a) sont des images binaires qui encodent les emplacements du mouvement, alors que, les MHIs (figure 3.13-b) sont des images en niveau de gris qui encodent l'historique du mouvement dans une séquence vidéo. Malgré le fait que cette représentation soit riche en information, elle nécessite la détection de la personne dans la scène et la suppression du fond.



(a) (b)

FIGURE 3.13 – Représentations globales des vidéos : (a) MEIs, (b) MHIs (figure issue de (Weinland *et al.*, 2011)).

Quant à la représentation locale, elle consiste à exploiter uniquement les points en mouvement de la vidéo. Ces points d'intérêt spatio-temporels sont extraits facilement de n'importe quelle résolution de la vidéo sans suppression du fond préalablement. Cette représentation est plus robuste à l'occultation et au changement des angles de vues. Elle est parcimonieuse, mais il n'est pas certain que ces points soient suffisants pour représenter la vidéo. Pour ces raisons, nous nous intéressons à la représentation de la vidéo par les PIs spatiotemporels (3D) que nous comptons enrichir par d'autres représentations. Nous présentons dans la suite les PIs 3D les plus utilisés pour décrire la vidéo. Ensuite, nous détaillons les PIs 3D sélectionnés dans ce travail.

3.4.2.1 Etat de l'art des points d'intérêt 3D

Les PIs 3D sont vus théoriquement comme extension des PIs 2D. Différents PIs 3D ont été proposés dans la littérature pour décrire une vidéo ou une image 3D. Dans notre contexte applicatif, les PIs 3D servent à décrire une vidéo. De manière similaire aux PIs 2D, ils nécessitent une phase de détection et une phase de description.

Détecteurs

Les détecteurs de PIs 3D partagent les mêmes principes que les détecteurs de PIs 2D à savoir, la définition d'un critère de sélection de PIs et l'application d'un algorithme de suppression des non-maximums.

(Laptev, 2005) a proposé le détecteur Harris 3D connu dans la littérature par STIP (pour «Space-Time Interest Points»). C'est l'extension spatio-temporelle du Harris 2D. De même, son critère de sélection est une combinaison de la trace et du déterminant de la matrice Harris 3D (Equation 3.1).

Pour caractériser les mouvements périodiques, (Dollár *et al.*, 2005) ont proposé un détecteur de PIs fondé sur le filtre de Gabor et le filtre Gaussien. Ce détecteur, connu par Cuboïdes, a été appliqué initialement pour caractériser d'une part les mouvements d'un animal et d'autre part, les actions et les expressions faciales d'une personne.

(Willems *et al.*, 2008) ont proposé un détecteur fondé sur la Hessienne 3D. C'est l'extension spatiotemporelle du détecteur 2D fondé sur la Hessienne et décrit dans la section 3.3.1.1. Comme critère de sélection, ce détecteur mesure le déterminant de la Hessienne 3D pour chaque voxel (pixel 3D). Ensuite, un algorithme de suppression de non-maximums est appliqué pour sélectionner les maximums du critère en espace, temps et échelle.

(Scovanner *et al.*, 2007) ont suggéré une extension spatiotemporelle du détecteur SIFT. Ce détecteur partage les mêmes propriétés que SIFT 2D. Il est fondé sur l'approximation du déterminant de la Hessienne 3D par DoG.

Descripteurs

Différents descripteurs ont été proposés dans la littérature. Les plus populaires sont fondés sur le HOG et le HOF (pour «Histogram of Optical Flow») pour décrire respectivement la variation spatiotemporelle de la texture et du mouvement de la personne.

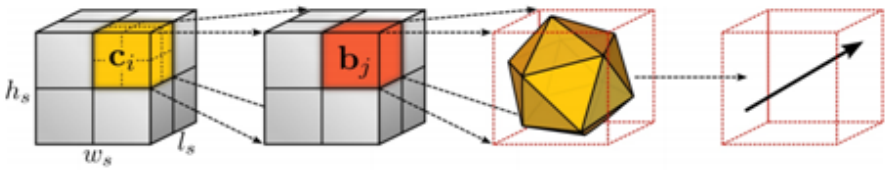
Le descripteur SIFT 3D (Scovanner *et al.*, 2007), extension du descripteur SIFT 2D, est fondé sur un calcul des histogrammes des gradients spatiotemporels orientés. Le gradient spatiotemporel (L_x, L_y, L_t) calculé en un voxel en position (x, y, t) est défini par trois paramètres : sa magnitude ($m_{3D}(x, y, t)$) et son orientation définie par les deux angles ($\theta(x, y, t)$ et $\varphi(x, y, t)$) :

$$m_{3D}(x,y,t) = \sqrt{L_x^2 + L_y^2 + L_t^2},\ \theta(x,y,t) = \tan^{-1}(\frac{L_y}{L_x}),\ \varphi(x,y,t) = \tan^{-1}(\frac{L_t}{\sqrt{L_x^2 + L_y^2}})$$

Pour générer le descripteur SIFT 3D, un cube est défini autour du PI et divisé en une grille de 4x4x4 cellules. Les valeurs de θ et φ sont quantifiées respectivement à 4 et 8 valeurs. Ensuite, un histogramme 2D est calculé pour chaque cellule similairement au SIFT 2D où chaque composante (bin) correspond à une combinaison des angles (θ, φ). La concaténation des histogrammes de toutes les cellules définit le descripteur SIFT 3D de dimension (4x8) x (4x4x4) = 2048.

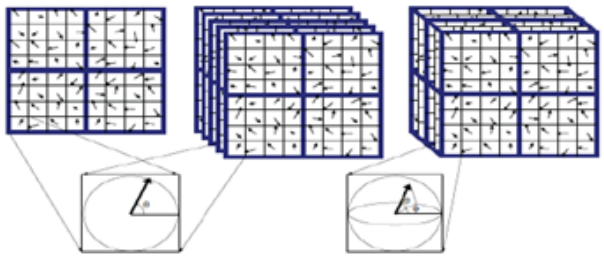
(Kläser *et al.*, 2008) propose le descripteur HOG 3D inspiré du descripteur SIFT 3D. Il consiste à diviser le cube entourant le PI en une grille de 4x4x3 cellules (figure 3.14-a). Ensuite un histogramme de 8 composantes est extrait pour chaque cellule. Finalement, chaque histogramme d'une cellule est normalisé par sa norme L_2 et concaténé dans le vecteur descripteur HOG 3D. Pour calculer l'histogramme d'une cellule, cette dernière est divisée en une grille de sous-blocs et ensuite un histogramme est calculé pour chaque sous-bloc (figure 3.14-b). L'histogramme d'une cellule est alors égal à la somme des histogrammes de tous ces sous-blocs. Le calcul de l'histogramme d'un sous-bloc est similaire au SIFT 3D. L'unique différence correspond aux différentes combinaisons des angles θ et φ associées aux composantes de l'histogramme. Dans ce cas, des polyèdres (figure 3.14-c) sont utilisés pour refléter les combinaisons des angles θ et φ. Pour calculer la somme des amplitudes des gradients des voxels dans un polyèdre, les vidéos intégrales, similaires aux images intégrales exploitées par SURF, sont utilisées. (figure 3.14-d).

(Laptev *et al.*, 2008) ont introduit le descripteur HOG/HOF. Cette description combine deux types de caractéristiques : HOG pour décrire la texture et HOF pour décrire le mouvement. Le calcul du HOG est similaire à celui de HOG pour les images (décrit dans la Section 3.4.1) ; l'unique différence est que les points participant au calcul de HOG appartiennent à un cube au lieu d'un rectangle. HOF est calculé de la même façon que HOG en calculant un vecteur de mouvement spatial au lieu d'un gradient. Un PI est finalement décrit par la concaténation des deux descriptions HOG et HOF. La figure 3.15 montre la différence entre SIFT 2D, HOG et SIFT 3D.



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

FIGURE 3.14 – Principe de calcul du descripteur HOG 3D (Kläser *et al.*, 2008).



(a) (b) (c)

FIGURE 3.15 – Comparaisons des descripteurs : pour (a) SIFT 2D et (b) HOG, le gradient est spatial alors que pour (c) SIFT 3D, le gradient est spatiotemporel (Scovanner *et al.*, 2007).

(Willems *et al.*, 2008) ont proposé le descripteur E-SURF (Extended SURF) extension du descripteur SURF 2D sur les vidéos. Comme pour tous les descripteurs 3D, un cube autour du PI est divisé en une grille tridimensionnelle de cellules. De chaque cellule, on collecte la somme des réponses des ondelettes de Haar échantillonnées uniformément au long des trois axes : $\sum d_x$, $\sum d_y$ et $\sum d_t$.

3.4.2.2 Points d'intérêt 3D choisis

Deux types de PIs 3D sont sélectionnés de l'état de l'art pour décrire le mouvement : STIP proposé par (Laptev, 2005) et Cuboïdes proposé par (Dollár *et al.*, 2005). Ces deux PIs ont été proposés originellement pour la reconnaissance des actions humaines où ils ont montré de bons résultats. L'utilisation des Cuboïdes est aussi motivée par le fait qu'en ré-identification, le mouvement des personnes est considéré comme périodique (le mouvement de marche).

Dans un contexte de reconnaissance des actions humaines, la base de données est généralement acquise par une seule caméra avec le même angle de vue et les personnes effectuent des mouvements très variés. En fait, les descripteurs standards associés aux STIPs et Cuboïdes exploitent essentiellement les vecteurs de mouvements. Ce descripteur de mouvement est adéquat au problème de reconnaissance des actions humaines où on désire reconnaître l'action faite par la personne plutôt que son identité. Dans un contexte de ré-identification, les changements des angles de vue, de l'éclairage et de la pose rendent ce descripteur peu robuste. Pour ces raisons, nous avons adapté le choix des descripteurs des PIs à notre contexte. Nous avons étudié la description des STIPs et des Cuboïdes par

d'autres descripteurs à savoir SURF et HOG Laptev2008.

Pour comparer cette description locale du mouvement à une représentation globale, nous choisissons de représenter la vidéo globalement par des PIs. Pour cela, la technique BoF a été utilisée avec le même principe de construction détaillé précédemment (section 3.3.2.2) dans le cas où on représente une image globalement. L'unique différence est que dans ce cas nous extrayons des PIs 3D au lieu des PIs 2D dans le cas des images.

Nous décrivons dans la suite les deux PIs sélectionnés ainsi que leurs descripteurs standards.

Points d'intérêt spatiotemporels (STIP)

Les STIPs sont des PIs spatiotemporels qui correspondent aux points en mouvements dans une vidéo.

Détecteur Le détecteur des STIPs est fondé sur la matrice Harris 3D donnée par l'équation 3.1. Ce détecteur est l'extension spatiotemporelle du détecteur de coins Harris 2D. La phase de détection se déroule en deux étapes. D'abord, on calcule la matrice Harris 3D, notée μ, en tout point (x, y, t) de la vidéo comme suit :

$$\mu = G\left(., s\sigma^2, s\tau^2\right) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_z^2 \end{pmatrix} \tag{3.1}$$

où σ et τ sont les échelles spatiale et temporelle ; G est une Gaussienne utilisée pour lisser l'image ; s est un paramètre qui lie l'échelle de G aux σ et τ. Les dérivées premières L_x, L_y et L_t d'une séquence vidéo v sont définies comme suit :

$$L_x(.;;\sigma^2,\tau^2) = \partial_x(G * v), \quad L_y(.;;\sigma^2,\tau^2) = \partial_y(G * v), \quad L_t(.;;\sigma^2,\tau^2) = \partial_t(G * v)$$

Plusieurs combinaisons d'échelles σ et τ dans les deux ensembles suivants ont été utilisées :

$$(\sigma^2,\tau^2) \in (SxT), \quad S = \{4, 8, 16, 32, 64\}, \quad T = \{2, 4\}$$

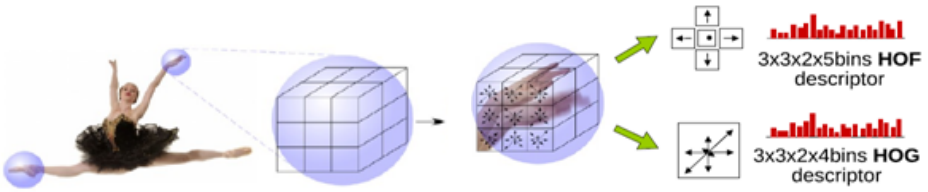
Ensuite, un critère R de sélection de PIs est défini, il est fondé sur la trace et le déterminant de la matrice Harris 3D. Ces PIs correspondent aux maximums en espace, temps et échelle de R :

$$R = \det(\mu) \ - k * \text{trace}^3(\mu), \quad R > 0$$

où k est une constante empirique. La valeur de k recommandée est $k \approx 0.005$.

Descripteur Originellement, un STIP est décrit par le descripteur HOG/HOF. Pour le calculer, un cube de dimension $(\Delta_x, \Delta_y, \Delta_t)$ est défini autour du STIP. Ces dimensions sont proportionnelles aux échelles de détection (σ et τ) : $\Delta_x = \Delta_y = 2l\sigma$ et $\Delta_t = 2l\tau$. Ensuite, ce cube est divisé en une grille de (n_x x n_y x n_t) cellules. De chaque cellule, un HOG de 4 composantes et un HOF de 5 composantes sont calculés. La concaténation des HOGs et HOFs normalisés de toutes les cellules définit la description HOG/HOF d'un STIP. La configuration standard du descripteur utilise les paramètres suivants : $l = 9$,

$n_x = n_y = 3$ et $n_t = 2$. La figure 3.16 montre le principe de construction du descripteur HOG/HOF.



(a) (b)

FIGURE 3.16 – Principe de construction du descripteur HOG/HOF : (a) un cube autour du STIP est devisé en une grille de cellules, (b) un HOG et HOF sont calculés pour chaque cellule. (Laptev *et al.*, 2008).

Les Cuboïdes

Les Cuboïdes sont des PIs spatiotemporels suggérés pour caractériser les mouvements périodiques dans les vidéos.

Détecteur La détection des cuboïdes est fondée sur des filtres Gaussiens et des filtres de Gabor. Le critère R de sélection des Cuboïdes est le suivant :

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$$

$$h_{ev}(t, \tau, \omega) = -cos(2\pi t\omega)e^{-\frac{t^2}{\tau^2}}$$

$$h_{od}(t, \tau, \omega) = -\sin(2\pi t\omega)e^{-\frac{t^2}{\tau^2}}$$

où $w = \frac{4}{\tau}$, $G(x, y, \sigma)$ est une Gaussienne de lissage, h_{ev} et h_{od} définissent les filtres de Gabor, τ et σ sont les échelles spatiale et temporelle.

Après avoir calculé la réponse R pour chaque point de la vidéo, les cuboïdes correspondent aux maximums de R en espace, temps et échelle.

Descripteur (Dollár *et al.*, 2005) ont proposé différents descripteurs pour décrire les Cuboïdes. Ces descripteurs correspondent à la concaténation des gradients ou des intensités des pixels du Cuboïde sans passer par des histogrammes comme dans le cas de HOG. La version originale de cette description utilise des Cuboïdes de dimensions 13x13x13. Pour réduire la dimensionnalité des descripteurs, les auteurs ont appliqué une ACP.

3.5 Conclusion

Dans ce chapitre, nous avons considéré la nature complémentaire de l'apparence de la personne et le type de son mouvement dans la vidéo pour la décrire. Selon la complexité du scénario de ré-identification, le mouvement est décrit soit implicitement par un modèle stochastique dans un scénario simple, soit par des PIs spatiotemporels robustes dans les conditions non contrôlées. Vu les avantages des descriptions fondées sur des points d'intérêt, nous avons enrichi la description de mouvement par une description d'apparence

fondée sur les PIs spatiaux. L'apparence est aussi décrite par les couleurs ; cette description peut être enrichissante dans des conditions d'éclairage stables.

Dans chacun des deux cas (description du mouvement ou de l'apparence), deux représentations ont été étudiées : une représentation locale et une représentation globale. Pour chaque représentation, nous avons, tout d'abord, produit un état de l'art non exhaustif et présenté les méthodes sélectionnées. La table 3.2 présente les différentes primitives sélectionnées de l'état de l'art.

TABLE 3.2 – caractéristiques des primitives extraites.

Description d'apparence		Description du mouvement	
Locale	Globale	Locale	Globale
SURF PIs de Harris	HOG RGB BoF de SURF/Harris	STIP Cuboïdes	BoF de STIP/Cuboïdes

Nous présentons dans les chapitres suivants des scénarios de ré-identification de complexités différentes. Pour chaque scénario, un ensemble de primitives est sélectionné pour décrire la personne à ré-identifier.

Chapitre 4

Ré-identification des personnes dans un scénario simple

Contents

4.1 Introduction

Nous traitons dans ce chapitre la ré-identification des personnes dans un scénario simple. Dans ce type de scénarios, d'une part, des contraintes sont fixées sur le passage de la personne dans le champ de vue de la caméra ; et d'autre part, les conditions du milieu de capture sont contrôlées. La table 4.1 illustre les différentes conditions liées à la personne et au milieu de capture.

Ces conditions contrôlées encouragent, d'après l'état de l'art, l'utilisation des caractéristiques biométriques pour reconnaître la personne. Notre objectif est d'étudier dans ce scénario simple la faisabilité d'un système de ré-identification fondé sur la biométrie. Nous avons choisi la démarche, la caractéristique biométrique la plus adéquate à un scénario de ré-identification puisque elle ne nécessite aucune coopération de la personne pour son ré-identification. Cette non-coopérativité est obligatoire pour mettre en œuvre un système de ré-identification.

TABLE 4.1 – Conditions d'un scénario simple.

Contraintes du milieu de capture	**Contraintes sur le passage de la personne**
– Eclairage stable. – Les données de référence et de test sont filmées par une seule caméra. – Base de données d'un nombre réduit de personnes. – Fond stable.	– Marcher dans une direction bien définie. – Ne pas porter des objets. – Une seule personne passe à la fois.

D'abord, nous avons évalué la robustesse du système à la variation des angles de vue. Ensuite, ce système biométrique a été comparé à un système fondé sur l'apparence. Dans ce scénario, l'apparence est décrite par les PIs spatiaux et les histogrammes de couleurs. L'utilisation des PIs spatiaux est encouragée par leurs robustesses aux changements des angles de vue. Quant à l'exploitation des histogrammes de couleurs sensibles aux conditions d'enregistrement, elle est motivée par les conditions contrôlées du scénario simple. En effet, plusieurs conditions du scénario simple promeuvent l'utilisation de la couleur : 1) stabilité du contraste car la base de données est filmée par une seule caméra, 2) absence d'occultation, 3) le nombre réduit de personnes augmente la probabilité que ces personnes puissent être distinguées par leurs couleurs (des vêtements, de la tête et de la peau).

Deux types de correspondances de PIs ont été comparés : locale et globale. La correspondance locale consiste à apparier les PIs, SURFs, les uns indépendamment des autres. Ensuite, en filtrant les paires appariées les moins fiables. Finalement, le principe de ré-identification est fondé sur la règle de décision par le vote majoritaire. Quant à la correspondance globale, elle consiste à transformer tous les PIs d'une personne en un seul vecteur (histogramme de PIs) et ensuite apparier les histogrammes de PIs par la distance Chi-Square (χ_2) au lieu d'apparier les PIs.

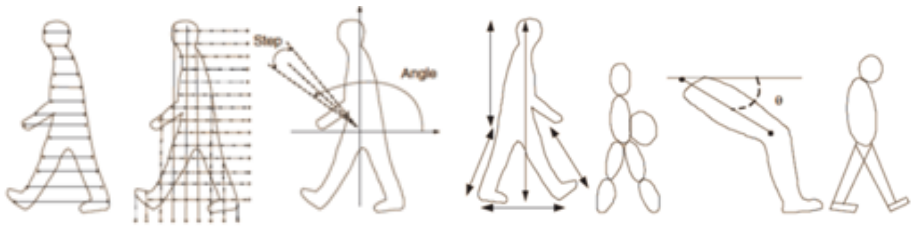
Tout au long de ce chapitre, nous montrons dans quelles conditions, la biométrie, les PIs et les couleurs peuvent être utiles pour la ré-identification. Dans la suite, la section 4.2 décrit le système biométrique de ré-identification par la démarche, la section 4.3.2 décrit le système de ré-identification par l'apparence et finalement dans la section 4.4 nous résumons les conditions de faisabilité de chaque système.

4.2 Ré-identification par la démarche

Nous avons mentionné dans le chapitre de l'état de l'art que les caractéristiques biométriques nécessitent des contraintes techniques et pratiques. En particulier, le système d'identification fondé sur la démarche induit moins de contraintes, surtout qu'il ne nécessite aucune coopération de la part des personnes à identifier. Dans cette section, nous voulons étudier la faisabilité d'un système de ré-identification fondé sur la démarche dans un scénario simple.

La motivation d'identification de la personne par sa démarche est initialisée par une étude présentée dans (Johansson, 1973). En effet, en utilisant une technique appelée "points lumineux en mouvement (Moving Light Displays)", l'auteur a démontré que les êtres humains peuvent être reconnus par leur façon de marcher. De nos jours, la démarche a montré une efficacité pour identifier les personnes, et plusieurs systèmes d'indentification fondés sur la démarche ont été développés (Wang *et al.*, 2010). Ces systèmes sont fondés sur deux étapes principales : 1) représentation de la ROI et 2) classification d'une séquence de marche (séquence de ROIs).

La représentation de la ROI a été traitée par deux familles d'approches. La première famille d'approches concerne les approches fondées sur un modèle qui consiste à représenter la ROI par un modèle cinématique ou de forme (section 3.4.2). La deuxième famille d'approches consiste à extraire directement des primitives de la ROI sans passer par un modèle à l'avance (dites aussi «approches holistiques»). La figure 4.1 illustre des exemples des deux familles d'approches.



(a) (b)

FIGURE 4.1 – (a) Approches holistiques. (b) Approches fondées sur un modèle (Boulgouris *et al.*, 2005).

Quant à la classification d'une séquence de marche, elle a été traitée par deux catégories d'approches : statique et temporelle. Dans la première catégorie, la séquence de marche est représentée par un seul vecteur caractéristique de taille fixe ou par un ensemble fixe d'images clés. Ces caractéristiques sont appariées à l'aide des classifieurs statiques tels que l'algorithme KNN (Collins *et al.*, 2002, Cunado *et al.*, 2003) et le classificateur discriminant SVM (Dadashi *et al.*, 2009, Xue *et al.*, 2010). Dans la deuxième catégorie, la séquence de marche est représentée par un nombre variable de vecteurs caractéristiques. Dans cette catégorie, on cite comme classifieur la distance élastique (DTW pour «Dynamic Time Warping») qui mesure une similarité entre deux séquences temporelles de longueurs différentes (Vega et Sarkar, 2003, Wang *et al.*, 2003b) et les modèles markoviens cachés qui tiennent compte de l'aspect temporel de la séquence de démarche (He et Debrunner, 2000, Sundaresan *et al.*, 2003).

Notre apport essentiel est d'exploiter le mouvement de la personne dans un système de ré-identification fondé sur la démarche. Pour cela, nous avons utilisé un modèle statistique permettant de modéliser l'aspect temporel de la vidéo. Le modèle sélectionné de l'état de l'art est le modèle de Markov caché qui est très répandu en reconnaissance de la voix et de gestes. Ce modèle consiste à apprendre la démarche d'une personne à partir des séquences d'observation *type*. Sachant qu'en ré-identification, une seule séquence vidéo est disponible par personne, nous avons exploité l'aspect périodique de la démarche pour augmenter le nombre de séquences par personne. L'exploitation de ce caractère périodique consiste à diviser la séquence vidéo en des périodes de marche. Les observations du HMM

sont des caractéristiques de même dimension extraites des images. Dans cette partie, nous avons sélectionné le descripteur global HOG (Dalal et Triggs, 2005) qui a montré de bons résultats dans la détection des personnes.

Le système proposé est composé des étapes principales suivantes : détection de la région d'intérêt, division en périodes, extraction des primitives et ré-identification par HMM. La figure 4.2 présente l'organigramme du système proposé.

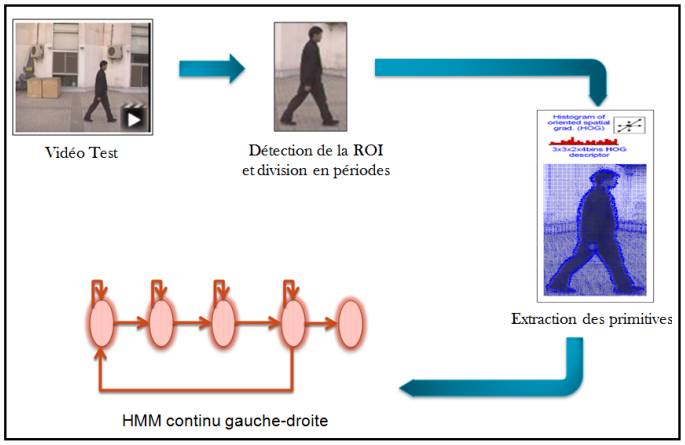


FIGURE 4.2 – Organigramme du système biométrique.

4.2.1 Détection de la ROI

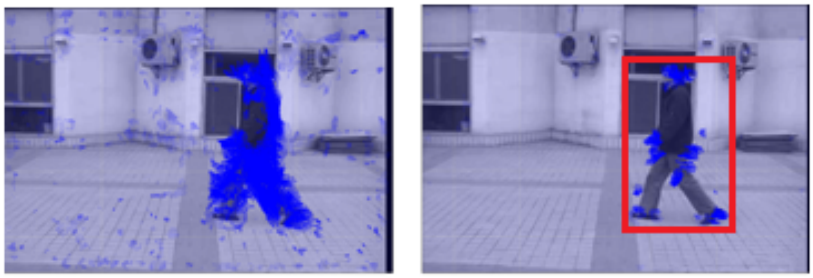
La première étape du système est la détection de la ROI. Dans un scénario simple où le fond est stable, le mouvement principal dans l'image est concentré sur la personne. La figure 4.3 montre un exemple de fond dans un scénario simple.

FIGURE 4.3 – Exemple d'images où le fond est simple.

Pour cela, nous avons proposé une méthode fondée sur la détection du mouvement pour localiser la personne. Cette méthode consiste essentiellement en trois étapes :

1. Calcul du flot optique par la méthode de (Lucas et Kanade, 1981) : pour une image capturée d'une vidéo à l'instant t, notée I_t, le calcul de l'image mouvement correspondante consiste à chercher un vecteur de mouvement de deux composantes $(F_V(x,y), F_H(x,y))$ pour chaque pixel en position (x,y) où $F_V(x,y)$ est la composante verticale du vecteur de mouvement décrivant le déplacement vertical du pixel dans l'image suivante temporellement I_{t+1}, $F_H(x,y)$ est la composante horizontale du vecteur de mouvement décrivant le déplacement horizontal du pixel dans l'image I_{t+1} (figure 4.4-a). Un pixel en position (x,y) dans I_t est alors retrouvé en position $(x + F_V(x,y), y + F_H(x,y))$ dans l'image I_{t+1}.

2. Filtrage des pixels ayant des vecteurs de mouvement de faible amplitude : ces pixels correspondent principalement à l'arrière-plan (figure 4.4-b).
3. Opération de cadrage : cadrer la zone où il y a du mouvement en parcourant les dimensions de l'image mouvement pour trouver les premiers pixels dont les vecteurs de mouvement ont une amplitude excédant un seuil empirique. La zone cadrée en rouge de la figure 4.4-b correspond à la ROI cherchée.



(a) (b)

FIGURE 4.4 – (a) : Image mouvement sans filtrage, (b) : image mouvement avec filtrage et ROI en rouge.

4.2.2 Division en périodes

Bien que le style de la marche est différent d'une personne à une autre, le processus de la marche est le même pour tous les êtres humains. Pour simplifier l'étude de ce processus, (Boulgouris *et al.*, 2005) ont considéré quatre postures principales de la marche : 1) les deux pieds touchent le sol, la jambe droite à l'avant (*right double support*), 2) les jambes sont au même niveau par rapport à la caméra, le pied droit touche le sol (*right midstance)*, 3) les deux pieds touchent le sol, la jambe gauche à l'avant (*left double support*) et 4) les jambes sont au même niveau par rapport à la caméra, le pied gauche touche le sol (*left midstance*). Un cycle de marche, dit aussi période, est définie par l'ensemble de silhouettes entre deux *midstances* consécutives de même type (figure 4.5).

La division des séquences en des périodes a deux avantages : 1) c'est la solution proposée au problème suivant : d'une part, un HMM nécessite un nombre relativement grand de séquences d'observations et d'autre part, une seule séquence d'apprentissage est disponible par personne. En découpant la séquence d'apprentissage en des périodes, chacune sera considérée comme une séquence d'observations pour HMM, 2) elle assure la synchronisation entre les périodes d'apprentissage et les périodes de test. En fait, chaque période commence et termine par les mêmes postures des silhouettes.

FIGURE 4.5 – Postures principale d'une période (Boulgouris *et al.*, 2005).

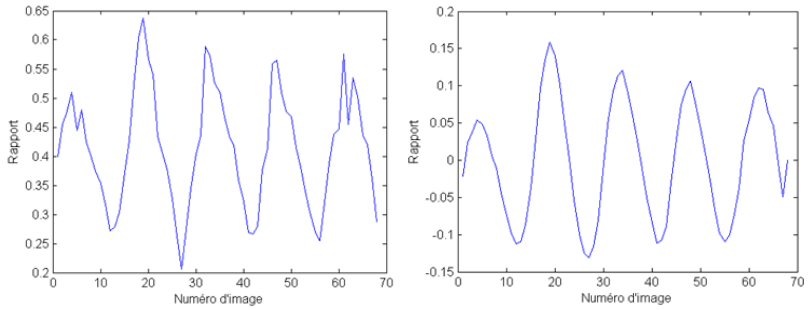
Plusieurs méthodes de division en des périodes ont été proposées dans la littérature. Dans ce travail, nous avons considéré le signal du rapport «*Hauteur sur Longueur*» (HsL) des ROIs. En effet, ce rapport atteint un minimum local en posture de «*Double Support*» (4.6-a) et atteint un maximum local en posture de «*Midstance*» (4.6-b). Une demi période est alors définie par les postures entre deux minimums locaux successifs ou deux maximums locaux successifs, une période étant associée à deux demi périodes successives.

Pour trouver ces maximums, la courbe du signal HsL subit deux étapes de prétraitement :

1. **Une étape de filtrage :** un filtre moyenneur est appliqué pour réduire les bruits. Ces bruits peuvent être dus à une mauvaise détection de la ROI.
2. **Une étape de normalisation :** une normalisation de la courbe est appliquée en rendant sa moyenne nulle. Cette normalisation permet de tenir compte du fait que les maximums de la courbe peuvent avoir des valeurs très variées. Un exemple de signal HsL, avant et après prétraitement, est présenté dans la figure 4.7.

(a) (b)

FIGURE 4.6 – (a) Signal des rapports HsL est maximal. (b) Signal des rapports HsL est minimal.



(a) (b)

FIGURE 4.7 – (a) Signal HsL avant prétraitement. (b) Signal après prétraitement.

4.2.3 Extraction des primitives

De chaque ROI, le descripteur de texture HOG est extrait (voir section 3.4.1). Chaque ROI est représentée par un vecteur de 81 composantes. Pour réduire la dimension des primitives, une ACP est appliquée. La figure 4.8 montre deux exemples des images gradient sous deux angles de vue différents.

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

FIGURE 4.8 – (a) et (c) : images originales. (b) et (d) : images gradient en 3x3 régions.

4.2.4 Introduction au HMM

Les modèles de Markov cachés (HMM) sont des modèles graphiques génératifs. Ils sont très répandus dans la classification dynamique telle que l'identification de la parole (Rabiner, 1989). Un HMM modélise un processus markovien qui est un système à temps discret qui, à chaque instant, il se trouve dans un état parmi N états distincts. Les transitions entre ces états sont modélisées par une loi de probabilité. Ainsi, la probabilité d'être à un état i ne dépend que de l'état précédent $i-1$. En outre, un HMM modélise un ensemble de séquences d'observation où l'état de chaque observation est caché mais lié à une fonction de densité de probabilité.

Soit $O = \{o_1; o_2; \ldots; o_T\}$ une séquence de T observations, elle peut être modélisée par un HMM qui se caractérise par les paramètres suivants :

- N, le nombre d'états du modèle. Généralement, ces états peuvent ou non avoir une signification physique. Ils sont notés par l'ensemble $S = S_1; S_2; \ldots; S_N$ et on note s_t l'état à un instant t.
- $A = [a_{ij}], 1 \leq i; j \leq N$, est la matrice des probabilités des transitions d'un état à un autre, avec :

$$a_{ij} = P(s_{t+1} = S_j | s_t = S_i), 1 \leq i; j \leq N$$

- $B = [b_j(o_t)], 1 \leq j \leq N; 1 \leq t \leq T$, est la matrice des probabilités d'apparition d'une observation sachant que le modèle est dans un état donné, avec :

$$b_j(o_t) = P(o_t | s_t = S_j), 1 \leq j \leq N; 1 \leq t \leq T$$

- $\Pi = [\pi_i], 1 \leq i \leq N$ est le vecteur des probabilités initiales avec :

$$\pi_i = P(s_1 = s_i)$$

  π_i est la probabilité d'être à l'état i à l'instant $t = 1$.

En résumé, un HMM est définit par le nombre d'état N et le triplet (A, B et Π).

Selon le type des observations (primitives) extraites, HMM peut être discret ou continu. Dans le cas discret, les primitives sont quantifiées et converties en des symboles d'observations. Quant au cas continu, les primitives sont utilisées directement comme des observa-

tions sans aucune transformation. Dans notre cas, les primitives extraites de l'image ou de la vidéo sont continues. Nous choisissons de modéliser une personne par un HMM continu car il évite la perte de données résultant de l'étape de quantification.

En effet, nous modélisons $P(o_t|s_t = S_j)$ par une mixture de M gaussiennes (GMM pour «Gaussian Mixture Model») caractérisée par trois paramètres : 1) les pondérations α_{kj}, 2) les moyennes μ_{kj} et 3) les variances σ_{kj} où $j = 1, \ldots, N$ et $k = 1, \ldots, M$. Le nombre de gaussiennes par mélange est constant pour tous les états et choisi empiriquement. Ainsi, on peut écrire $b_j(o_t)$ comme suit :

$$b_j(o_t) = P(o_t|s_t = S_j) = \sum_{k=1}^{M} \alpha_{kj} N(o_t, \mu_{kj}, \sigma_{kj})$$

Dans ce cas, estimer la matrice des probabilités d'observations est équivalente à estimer les paramètres du GMM pour chaque état j : α_{kj}, μ_{kj} et σ_{kj}, $k = 1, \ldots, M$ où α_{kj} sont des scalaires vérifiant : $(\sum_{k=1}^{M} \alpha_{kj}) = 1$, μ_{kj} est un vecteur de dimension D, σ_{kj} est une matrice de dimension DxD et D est la dimension d'une observation. Dans nos expériences, pour réduire le nombre de paramètres à estimer, on suppose que σ_{kj} est une matrice diagonale.

4.2.5 Principe de ré-identification par HMM

En ré-identification, on veut ré-identifier une ou plusieurs personnes test à partir d'un ensemble de Q personnes référence. La ré-identification avec HMM s'opère en deux étapes fondamentales : apprentissage et classification.

- **Apprentissage :** pour chaque personne q ($1 \leq q \leq Q$), on construit le HMM λ_q qui maximise la vraisemblance de ses données d'apprentissage.

- **Classification :** pour une séquence d'observations O d'une personne inconnue, la vraisemblance de cette séquence est calculée pour toutes les Q personnes : $P(O|\lambda_q)$ où $1 \leq q \leq Q$. La personne ayant la séquence O est ré-identifiée comme la personne q^* qui maximise sa vraisemblance.

$$q^* = \text{argmax}_{1 \leq q \leq Q}[P(O|\lambda_q)]$$

En ré-identification, chaque personne est représentée par une seule séquence vidéo, en référence et en test, et après division en périodes, nous aurons un ensemble de séquences d'observations. Dans ce cas, $O = \{O_1; \ldots; O_L\}$ où L est le nombre de périodes de la séquence vidéo. Une observation est le HOG de dimension D d'une image donnée. Quant au calcul de $P(O|\lambda_q)$, on suppose que les séquences d'observations sont indépendantes et par conséquent elle s'écrit comme suit :

$$P(O|\lambda_q) = \sum_{l=1}^{L} P(O_l|\lambda_q)$$

4.2.5.1 Apprentissage des paramètres des HMMs

L'apprentissage du HMM λ_q associé à la personne q consiste à déterminer le triplet (A_q, B_q, Π_q) qui maximisent la probabilité $P(O|\lambda_q)$. L'apprentissage de ces paramètres s'opère en deux phases : 1) initialisation des paramètres et 2) ré-estimation des paramètres d'une manière itérative.

Initialisation des paramètres Le principe d'initialisation d'un HMM de N états est le suivant :

- Initialisation des probabilités initiales $\Pi_p = (1, 0, \ldots, 0)$: on oblige le HMM d'être à l'état 1 à $t = 1$. Avec cette initialisation, les valeurs de Π_p sont inchangées durant la ré-estimation des paramètres. Ce choix est justifié par le fait que les périodes commencent par les mêmes postures qu'on affecte à l'état 1.
- Initialisation de la matrice des probabilités des transitions d'une manière uniforme :

$$\forall i, j \in [1..N], \quad A(i,j) = \begin{cases} 0.5 & \text{si } j = i \text{ ou } j = (i \bmod N) + 1 \\ 0 & \text{sinon.} \end{cases}$$

- Initialisation de la matrice des probabilités des observations : les paramètres des gaussiennes sont initialisés uniformément. En effet, pour un HMM donné, les séquences d'apprentissage sont uniformément segmentées en N sous-séquences. Toutes les sous-séquences correspondant à un état donné sont utilisées pour estimer les paramètres des gaussiennes correspondants à cet état.

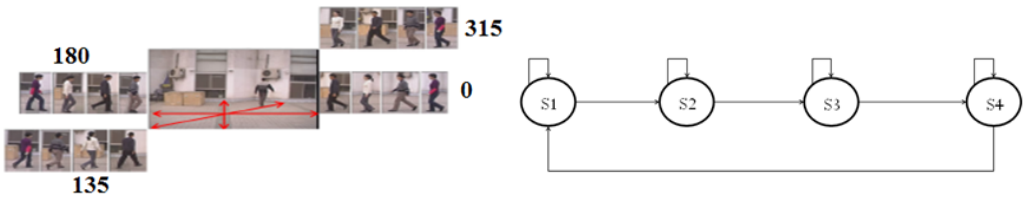
Ré-estimation des paramètres : L'algorithme Baum-Welch (Rabiner, 1989) est utilisé pour raffiner itérativement les matrices A_q et B_q. À chaque itération, l'algorithme Baum-Welch maximise $P(O|\lambda_q)$, et comme paramètre de convergence de l'algorithme, nous avons choisi un nombre d'itérations maximal égal à 20.

4.2.5.2 Classification avec HMM

Etant donnée une séquence vidéo V d'identité inconnue et après avoir appris les HMMs λ_q pour chaque personne référence, nous cherchons le HMM λ_j qui maximise la probabilité$P(O|\lambda_q)$ où O est l'ensemble des séquences d'observations extraites de V. $P(O|\lambda_q)$ est calculée par l'algorithme Forward (Rabiner, 1989).

4.2.6 Expériences et résultats

En évaluation, nous avons étudié la faisabilité du système biométrique en fonction de la différence angulaire entre les angles de vue d'apprentissage et de test. Nous avons utilisé la base de données CASIA-A (section 2.5.1). Comme mentionne l'état de l'art, elle est composée de 20 personnes marchant dans 6 directions différentes ; dans cette partie, nous tenons compte uniquement des 4 directions (2 directions latérales et 2 directions diagonales). La figure 4.9 montre le modèle HMM utilisé ainsi que les directions de CASIA-A explorées dans l'évaluation. Comme implémentation de HMM, nous nous sommes servis de celle de Murphy (Murphy, 1998).



(a) (b)

FIGURE 4.9 – (a) Exemple d'images de CASIA-A. (b) Le modèle HMM utilisé (le nombre d'états est à titre indicatif).

Pour étudier la robustesse de ce système au changement des angles de vue, deux scénarios simples ont été définis : 1) un premier scénario simple où l'angle de vue en test est égal à l'angle de vue en apprentissage et 2) un deuxième scénario simple où l'angle de vue en test est légèrement différent à l'angle de vue en apprentissage.

Ce système dépend de trois paramètres : 1) N : nombre d'états des HMMs, 2) M : nombre de gaussiennes par état et 3) R : taille des descripteurs après la réduction par ACP.

Nous avons testé deux configurations :
- Configuration 1 : N=3, M=2 et R=50
- Configuration 2 : N=1, M=1 et R=50

Pour la première configuration, on a environ 6 fois plus de paramètres à estimer par rapport la deuxième configuration. Pour les deux configurations, nous avons retenu 50 éléments du descripteur HOG suite à la réduction par ACP, c'est ce qui correspond à plus de 90% de la variance des vecteurs HOGs par rapport à leur centre de gravité. Les tables 4.2 et 4.3 résument les résultats des deux configurations évaluées sur les deux scénarios.

TABLE 4.2 – Taux de ré-identification du scénario simple 1.

	Apprentissage : 0°, Test : 0°	Apprentissage : 180°, Test : 180°
Configuration 1	95%	95%
Configuration 2	100%	100%

TABLE 4.3 – Taux de ré-identification du scénario simple 2.

	Apprentissage : 0°, Test : 315°	Apprentissage : 180°, Test : 135°
Configuration 1	35%	25%
Configuration 2	50%	20%

D'une part, les résultats ont montré, que dans le cas où l'angle de vue en apprentissage est égal à celui du test, la ré-identification est parfaite pour les deux configurations. Par contre, pour une différence angulaire égale à 45°, le taux de ré-identification chute jusqu'à 20%.

D'autre part, un HMM modélise théoriquement mieux que GMM l'aspect temporel de la vidéo, par contre en évaluation, ils ont des performances proches.

4.2.7 Faisabilité d'un système biométrique

En synthèse, nous constatons que le système biométrique performe mieux 1) quand les données d'apprentissage et de test sont prises par le même angle de vue et 2) quand assez de données sont disponibles pour apprendre les HMMs.

En effet, ce système biométrique réalise une ré-identification presque parfaite quand on apprend les HMMs sur un angle de vue et on les teste sur le même angle de vue. Par contre la performance du système se dégrade rapidement dès qu'on change légèrement l'angle de vue de test. Ceci est justifié par le non robustesse des gradients au changement des angles de vue.

D'autre part, le système biométrique nécessite une quantité de données suffisante pour apprendre la marche des personnes. Sur la base de données d'apprentissage étudiée, les séquences vidéo sont de longueur moyenne égale à 4.8 secondes, équivalent à 2 périodes en moyenne par séquence vidéo (table 4.4). Pour bien apprendre la marche d'une personne, elle doit marcher largement suffisamment dans les champs de vue de la caméra ; une dizaine de périodes peut être nécessaire. Ce manque de données justifie le fait que HMM et GMM donnent des performances proches malgré le potentiel des HMMs à modéliser la marche par rapport au GMM.

TABLE 4.4 – Nombre de périodes par HMM.

Nombre de périodes par séquence vidéo	1	2	3
Nombre de séquences vidéo (HMMs) associées	12	27	1

4.3 Ré-identification par l'apparence

L'alternative des méthodes biométriques est l'utilisation des caractéristiques d'apparence. Comme primitives d'apparence robustes aux changements des angles de vue, nous avons sélectionné les histogrammes de couleurs et les PIs. Deux représentations sont comparées : 1) représentation locale fondée sur la correspondance des PIs SURFs et 2) représentation globale fondée sur la correspondance des BoFs de PIs ou des histogrammes de couleur.

4.3.1 Description locale

Nous traitons dans cette partie la robustesse de l'apparence aux changements des angles de vue. Les points d'intérêt SURF, robustes aux changements des angles de vue, sont utilisés. Nous avons procédé à la ré-identification à partir des séquences vidéo entières (approche multi-échantillons) afin d'augmenter leur reproductibilité et réduire leur instabilité. Dans ce contexte, nous avons proposé un système de ré-identification fondé sur la correspondance des SURFs. Ensuite, nous avons étudié la performance des SURFs et analysé les résultats en fonction de la différence angulaire entre les angles de vue de référence et de test. Pour une mise en correspondance (matching) fiable, nous avons proposé une méthode de sélection automatique des correspondances SURF (filtrage), fondée sur une modélisation probabiliste. Cette modélisation est fondée sur le rapport de deux GMMs appris sur la base de référence. Les 2 GMMs modélisent respectivement la distribution des distances résultant des correspondances de deux séquences de la même personne et de différentes personnes. Dans la suite, nous détaillons les différentes étapes du système réalisé, ensuite nous présentons les résultats obtenus.

4.3.1.1 Description du système d'apparence

Etant donnée une séquence vidéo test d'identité inconnue et une base de référence, l'objectif d'un système de ré-identification est d'évaluer si une séquence de la même personne apparaissant dans la séquence test existe dans la base de référence. Ce système se déroule essentiellement en cinq étapes : 1) détection de la région d'intérêt, 2) extraction des SURFs, 3) correspondance des paires de SURFs, 4) filtrage des correspondances à l'aide de GMM et 5) ré-identification par la règle du vote majoritaire. La figure 4.10 montre l'organigramme du système.

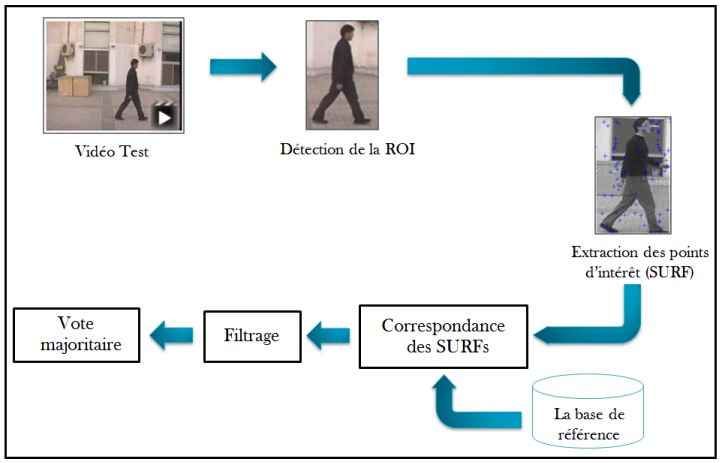


FIGURE 4.10 – Organigramme du système d'apparence.

4.3.1.1.1 Détection de la ROI Trois directions différentes de passage dans le champ de vue de la caméra sont considérées : latérale, diagonale et frontale. Pour les directions latérales et diagonales, le mouvement peut être exploité pour localiser la personne car visuellement tous les membres de la personne sont en mouvement. Par contre, quand la direction est frontale (figure 4.11), la personne semble stable et il y'a peu de mouvement dans la partie inférieure du corps et le mouvement est presque nul sur la partie supérieure du corps. Dans ce cas, la personne ne peut pas être localisée facilement par son mouvement. Par conséquence, nous avons utilisé une méthode de localisation fondée sur les silhouettes binaires pour toutes les directions. Cette méthode dépend effectivement de la qualité de binarisation. Dans ce scénario simple, le fond est stable et les silhouettes binaires disponibles avec la base de données ont une bonne qualité (figure 4.12-b). Pour localiser la silhouette, nous parcourons les dimensions de l'image binaire pour trouver les premiers pixels blancs et déduire par la suite le cadrage (figure 4.12-c).

FIGURE 4.11 – Passage frontale dans le champ de vue de la caméra.

(a) (b) (c)

FIGURE 4.12 – Détection de la ROI : (a) image originale, (b) silhouette binaire, (c) ROI.

4.3.1.1.2 Extraction des SURFs La ROI est décrite par un ensemble de PIs extraits par la méthode SURF (figure 4.13). Comme mentionné dans la section 3.3.1.2, SURF est connu par sa robustesse aux changements des angles de vue, son invariance aux transformations géométriques et sa rapidité. Chaque SURF est décrit par un descripteur de dimension 64 fondé sur les réponses aux ondelettes de Haar. La table 4.5 montre une statistique sur le nombre moyen des SURFs détectés en référence et en test.

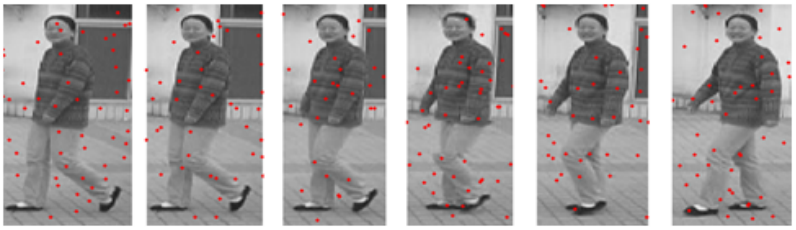


FIGURE 4.13 – Exemples des SURFs détectés.

TABLE 4.5 – Statistiques des SURFs détectés sur CASIA-A.

	Test	Référence
Nombre moyen des SURFs par image	81	79
Nombre moyen des SURFs par séquence	6600	6416

4.3.1.1.3 Correspondance des paires de SURFs Le principe d'appariement des SURFs est fondé sur le schéma suivant : chaque SURF de la séquence test est apparié au SURF le plus proche de la base de référence en utilisant la distance euclidienne (figure 4.14 où d_{min} est la distance euclidienne minimale entre un SURF test et la base de référence).

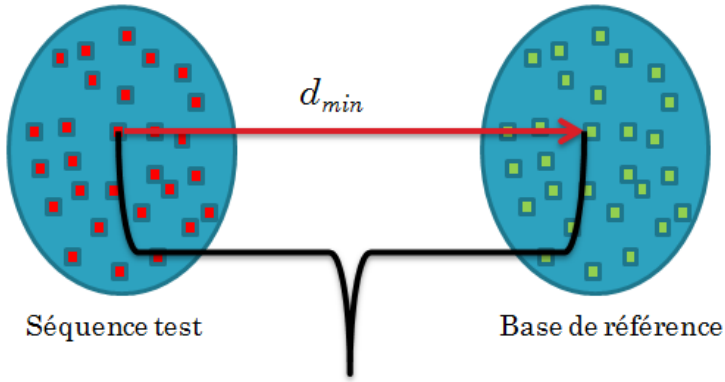


FIGURE 4.14 – Principe de correspondance des paires de SURFs.

4.3.1.1.4 Filtrage des correspondances à l'aide de GMM Cette étape consiste à filtrer les paires appariées les moins fiables qui sont probablement relatives par exemple à 1) l'appariement des PIs de fond avec des PIs appartenant à la silhouette, 2) l'appariement des PIs appartenant à deux parties différentes de la silhouette et 3) l'appariement des PIs

appartenant à des personnes différentes.

En littérature, (Hamdoun *et al.*, 2008, de Oliveira et de Souza Pio, 2009) ont proposé deux approches de ré-identification fondées sur l'appariement des PIs. Pour valider une ré-identification, deux filtrages successifs sont appliqués en utilisant des seuils empiriquement choisis :

1. Le premier filtrage concerne le nombre d'appariements. En effet, si le nombre de paires appariées entre la requête et la référence est inferieur à un seuil, la requête est considérée comme une séquence inconnue, inexistante dans la référence.
2. Le deuxième filtrage concerne les PIs. En effet, une correspondance entre deux PIs p_0 et p_1 ($p_0 \in$ référence et $p_1 \in$ test) est retenue, si $d(p_0, p_1) < c * d(p_0, p_i)\ \forall p_i \in$ référence, où c est un coefficient empirique $c < 1$ et $d(.,.)$ est la distance euclidienne.

La condition du deuxième filtrage est proposée dans (Lowe, 2001) pour classifier des objets dans des images 2D ; par contre son application dans une vidéo souffre du problème suivant : soit une bonne correspondance entre p_0 et p_1 et soient p_1 et p_2 les PIs références les plus proches de p_0. En raison de la répétabilité des PIs dans une vidéo, p_1 et p_2 peuvent être proches et vérifient $d(p_0, p_1) \approx d(p_0, p_2)$. Comme la condition de Lowe n'est pas vérifiée, cette correspondance est filtrée même si il s'agit d'une bonne correspondance. En outre, la condition de Lowe souffre du choix empirique du seuil c.

Dans ce travail, nous proposons de développer une méthode de filtrage des correspondances de PIs qui ne souffre pas de ces inconvénients. Ainsi, nous considérons deux modèles de mélanges gaussiens (GMM) : GMM_1 modélisant la distribution des distances entre les PIs associés à la même personne, et GMM_2 modélisant la distribution des distances entre les PIs associés à des personnes différentes. La décision d'accepter ou de rejeter une paire avec la distance d est fondée sur le rapport de vraisemblance LR entre les deux GMMs.

$$LR = \frac{P(d|GMM_1)}{P(d|GMM_2)}$$

La paire est conservée si $LR > 1$ et est éliminée si $LR \leq 1$. Comme montré précédemment, ce mécanisme de décision est automatique et aucun réglage de seuil empirique n'est nécessaire.

Apprentissage des GMMs

GMM est un modèle probabiliste qui peut approximer une distribution de probabilité, étant donné un nombre suffisant de composantes. Dans le cas où la matrice de covariance est diagonale, la fonction de densité de probabilité associée est définie comme suit :

$$p(x) = \sum_{g=1}^{G} c_g \prod_{f=1}^{F} \frac{1}{\sqrt{2\pi {\sigma_{gf}}^2}} . \exp(-\frac{1}{2}(\frac{x_f - \mu_{gf}}{\sigma_{gf}})^2)$$

où G est le nombre de composantes du mélange, c_g est le poids associé à la composante g, F est la dimension du vecteur descripteur x, μ_{gf} et σ_{gf} sont la moyenne et l'écart type de x correspondant à la composante g. Dans notre cas, GMM_1 et GMM_2 sont univariées et modélisent la distribution des distances entre les SURFs qui se correspondent. La probabilité d'une distance d par rapport à chaque GMM est :

$$p(d/GMM_i) = \sum_{g=1}^{G} c_{ig} \frac{1}{\sqrt{2\pi\, \sigma_{ig}^2}} . \exp(-\tfrac{1}{2}\left(\tfrac{d-\mu_{\mathrm{ig}}}{\sigma_{\mathrm{ig}}}\right)^2) \quad i = 1; 2$$

où μ_{ig} et σ_{ig} sont la moyenne et l'écart type de la composante g de GMM_i.

Pour former les deux GMMs, les séquences vidéo de référence peuvent être facilement exploitées comme expliqué ci-dessous. L'apprentissage des deux GMMs est effectué sur une base de référence composée de 20 personnes. Chaque personne fournit une séquence pour chacun des six angles de vue (figure 4.16). Pour chaque combinaison d'angles de vue dans la base de référence ($angle_1$, $angle_2$) (36 combinaisons) et pour chaque personne "P_1" de la base de référence, nous appliquons les deux étapes suivantes :

1. **Correspondance 1 :** nous considérons les deux séquences de la personne "P_1", correspondant aux deux angles de vue ($angle_1$ et $angle_2$). Puis, les SURFs de ces deux séquences sont appariés et les distances des correspondances sont ajoutées à l'ensemble S_{same} constitué par les distances de correspondance des séquences de la même personne.
2. **Correspondance 2 :** nous choisissons, au hasard, une autre personne "P_2" de la base de référence et nous sélectionnons la séquence de "P_2" avec l'angle de vue "$angle_2$". Les distances résultant de la correspondance des SURFs entre la séquence de la personne "P_1" avec l'angle de vue "$angle_1$" et la séquence de la personne "P_2" avec l'angle de vue "$angle_2$" contribuent à la construction de l'ensemble S_{diff} composé par les distances de correspondance des séquences de différentes personnes.

Cette stratégie soulève une question : quand $angle_1$ et $angle_2$ sont égaux, une seule séquence de référence est disponible et "Correspondance 1" n'est plus réalisable. Pour surmonter ce problème, nous divisons, dans ce cas, la séquence de référence en deux sous-séquences, l'une contenant les images impaires et l'autre contenant les images paires (afin de conserver l'aspect temporel de la vidéo) et nous effectuons la correspondance entre elles afin de générer des distances contribuant à S_{same}. Une fois les deux ensembles S_{same} et S_{diff} générés, les paramètres de GMM_1 et GMM_2 sont estimés en tenant compte d'un certain nombre de composantes proches du nombre d'angles de vue disponibles dans la base de référence.

Filtrage des correspondances

Les deux GMMs sont appris par le principe précédent. Ensuite, les correspondances sont filtrées en utilisant le critère de vraisemblance LR. Dans notre cas, 37.08% des correspondances sont filtrées.

4.3.1.1.5 Ré-identification par vote majoritaire En ré-identification, les paires de SURFs retenues sont soumises à la règle de décision par le vote majoritaire. En effet, pour chaque paire retenue, un vote est ajouté à la personne associée au SURF référence. La personne qui obtient la majorité des votes est considérée comme la personne ré-identifiée (figure 4.15).

4.3.1.2 Expériences et résultats

4.3.1.2.1 Base de données Pour évaluer ce système, nous avons utilisé la base de donnée multi-échantillons CASIA-A (section 2.5.1). Elle est adaptée à l'évaluation de la ré-identification en fonction du changement de l'angle de vue entre la référence et le test car elle contient des vidéos de 20 personnes se déplaçant en référence et en test dans 6 directions différentes : 0°, 90°, 135°, 180°, 270° et 315° (figure 4.16).

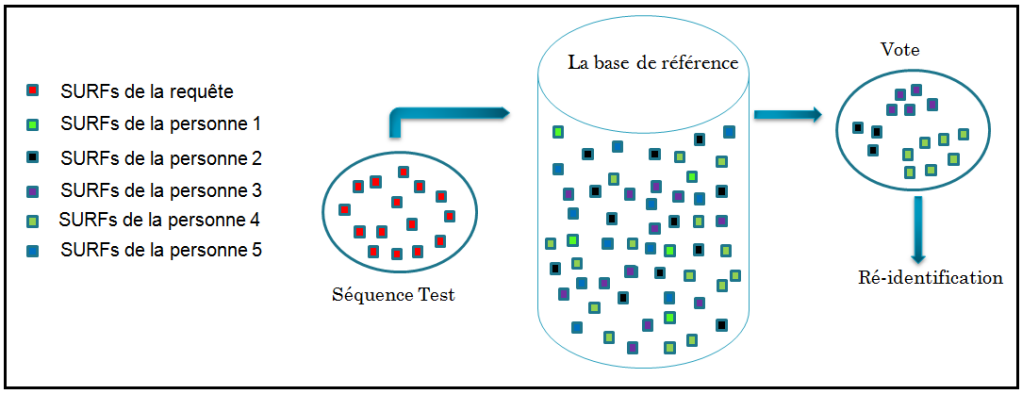


FIGURE 4.15 – Exemple de vote majoritaire (Dans cet exemple, la personne requête est reconnue comme la personne 4).

FIGURE 4.16 – La base de données CASIA-A.

4.3.1.2.2 Résultats L'évaluation est réalisée pour chaque combinaison possible (angle test, angle référence). CASIA-A contient 6 angles de vue test et 6 angles de vue référence, 36 expériences ont été réalisées. Les résultats sont présentés dans la figure 4.17 et la table 4.6. La figure 4.17 montre la performance de la ré-identification en fonction de la différence angulaire entre les vues de test et de référence. La table 4.6 montre les résultats de différentes combinaisons des vues de référence (colonnes) et vues de test (lignes). La performance du système est présentée par le taux de classification correcte (CCR pour «Correct Classification Rate») défini par le rapport entre le nombre de personnes correctement ré-identifiées et le nombre total de personnes testées.

TABLE 4.6 – CCR pour les différentes combinaisons des angles de vue de CASIA-A.

Angle	**0°**	**90°**	**135°**	**180°**	**270°**	**315°**
0°	100	40	90	100	35	90
90°	75	100	100	75	90	95
135°	90	70	100	100	70	100
180°	100	50	95	100	30	85
270°	70	95	95	75	100	95
315°	100	60	100	85	70	100

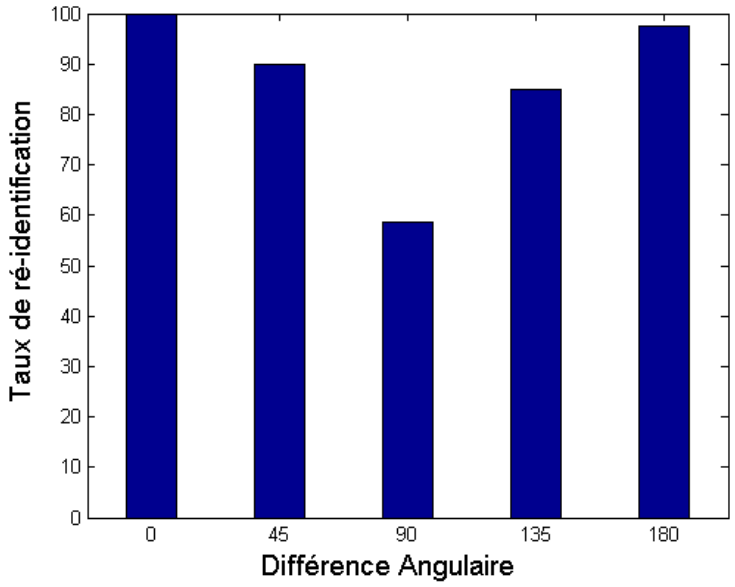


FIGURE 4.17 – Evolution du CCR en fonction de la différence angulaire sur CASIA-A.

La figure 4.17 montre que la ré-identification est presque parfaite lorsque les vues de test et de référence sont identiques ou symétriques (différence angulaire = 180°). Cela prouve la robustesse des SURFs lorsque les séquences de référence et de test sont similaires. Par contre, la performance de la ré-identification diminue avec l'augmentation de la différence angulaire.

En effet, elle est meilleure lorsque les vues de test et de référence partagent certaines parties visibles (une petite différence angulaire). Par exemple, prenons l'angle "0°" comme angle de référence, la table 4.6 montre que les meilleurs résultats sont obtenus pour les angles de test "0°" et "180°" (100% et 100%), puis les résultats diminuent légèrement avec les angles de test "135°" et "315°" (90% et 100%), puis de manière plus significative avec les angles de test "90°" et "270°" (75% et 70%). Ce résultat est compatible avec les exemples de la base de données (figure 4.16), où nous pouvons voir que les postures de l'angle de vue "0 °" sont similaires à celles de "180°" et partagent une partie commune avec les postures des angles de vue "135°" et "315°" ; cette partie commune diminue ensuite encore pour les angles de vue "90°" et "270°".

En outre, les résultats ont montré que dans certains cas, on arrive à ré-identifier des personnes même si on ne voit pas leurs visages et même pour une grande différence angulaire. Par exemple, prenons l'angle "0°" comme angle de référence, la table 4.6 montre que pour l'angle de test "90°" où le visage est totalement caché, on arrive à ré-identifier 40% des personnes. Cela prouve que le système de ré-identification est différent d'un système biométrique dans le sens qu'il exploite toute information utile à reconnaître la personne et non uniquement les caractéristiques physiques ou comportementales de la personne.

4.3.1.2.3 Comparaison avec l'état de l'art Nos résultats ont été comparés avec l'approche présentée dans (Jungling et Arens, 2011), fondée sur un modèle implicite de forme à partir des SIFTs (à notre connaissance, c'est le seul papier contenant des résultats en ré-identification obtenus sur la même base de données et en utilisant le même protocole d'évaluation). La comparaison des résultats est présentée dans la figure 4.18.

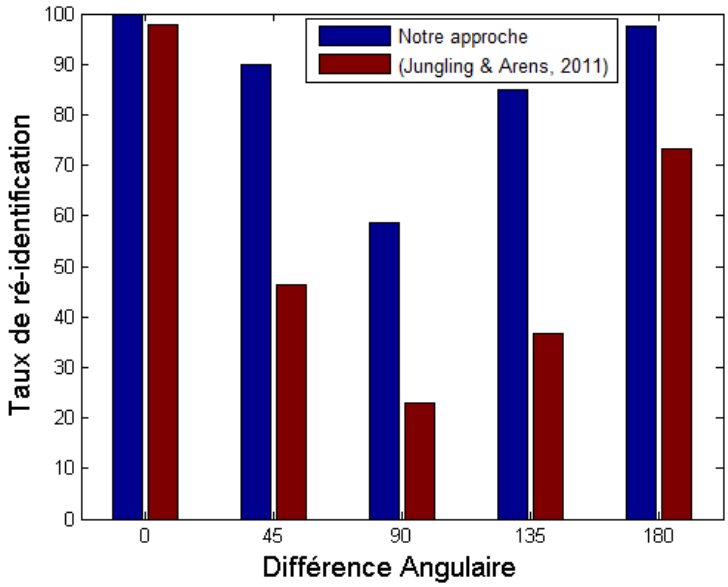


FIGURE 4.18 – Comparaison des résultats avec l'état de l'art.

La comparaison montre que notre méthode surmonte de manière significative (Jungling et Arens, 2011), en particulier lorsque les vues de référence et de test sont différentes. Cela montre la robustesse de la correspondance des SURFs sous différents angles de vue lorsque tout le contenu des séquences de référence et de test est exploité pour l'appariement et lorsqu'un mécanisme pour rejeter/accepter une correspondance (filtrage) est sélectionné.

Un autre résultat sur CASIA-A a été présenté dans (Eisenbach *et al.*, 2012). Nous ne nous pouvons pas comparer avec ce dernier travail car il n'utilise pas le même protocole d'évaluation.

4.3.2 Description globale

Dans cette partie, nous comparons la représentation locale de l'apparence à une représentation globale. Deux représentations globales sont étudiées : 1) une représentation fondée sur les PIs en utilisant la technique BoF (section 3.3.2.2) et 2) une représentation fondée sur les histogrammes de couleurs (section 3.3.2.2).

Le principe de ré-identification avec ces représentations globales est identique à celui du système précédent fondé sur les SURFs dans le sens qu'il consiste à apparier directement les descripteurs. Par contre, il en diffère en deux points : 1) aucun filtrage n'est appliqué et 2) la distance Chi-Square est utilisée au lieu de la distance euclidienne pour calculer la similarité entre les descripteurs.

En effet, le filtrage est appliqué dans le cas des SURFs pour filtrer les paires de PIs les moins fiables. Dans le cas global, cette supposition n'est plus vraie car tous les SURFs participent à construction de l'histogramme de PIs.

Quant au calcul de similarité, plusieurs distances entre les histogrammes ont été exploitées dans la littérature telles que la distance Chi-Square, la distance d'intersection des

histogrammes et la distance de Bhattacharyya. Nous avons utilisé la distance Chi-Square qui a déjà été employée pour mesurer des similarités entres les BoFs dans le domaine de reconnaissance des actions (Laptev *et al.*, 2008, Kläser *et al.*, 2008). Elle permet une normalisation implicite des descripteurs. La distance Chi-Square (χ_2) entre les deux histogrammes H^i et H^j de même dimension d est définie comme suit :

$$\chi_2(H^i, H^j) = \frac{1}{2}\sum_{k=1}^{d}\frac{\left(H_k^i - H_k^j\right)^2}{\left(H_k^i + H_k^j\right)}$$

4.3.2.1 Description par l'histogramme BoF de PIs

4.3.2.1.1 Extraction des primitives Nous avons vu dans la section 3.3.2.2 que la construction de BoF ne dépend que d'un seul paramètre «k» qui reflète la dimension du descripteur («k» est le nombre de classes (clusters) référence générées par l'algorithme k-means). Dans ce travail, nous avons testé plusieurs valeurs de k allant de 5 à 1000. Quelles que soit la valeur de k, le nombre de BoFs par séquence est fixe (BoF par image). La table 4.7 montre le nombre moyen de BoFs par séquence référence et test.

TABLE 4.7 – Nombre moyen de BoFs par séquence.

Test	Référence
81 BoFs/séquence	78 BoFs/séquence

4.3.2.1.2 Résultats Nous avons suivi le même protocole d'évaluation que le système fondé sur les SURFs (section 4.3.1.2.2) et la même base de données CASIA-A. En outre, nous avons étudié la performance du système en fonction du nombre de classes k : des petites et grandes valeurs de k ont été testées (10, 50, 100, 500 et 1000). Les résultats sont présentés de deux façons : 1) CCR en fonction de la différence angulaire entre les angles de vue en test et en référence (figure 4.19) et 2) taux de ré-identification global : c'est la moyenne des CCRs des 36 expériences réalisées (figure 4.20). Dans les deux figures (4.19 et 4.20), les résultats correspondant à chaque valeur de k sont représentés par la même couleur.

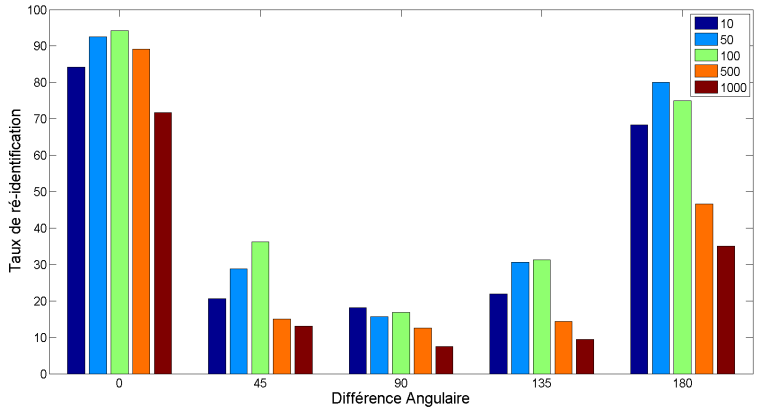


FIGURE 4.19 – CCR en fonction de la différence angulaire et la dimension des BoFs (k).

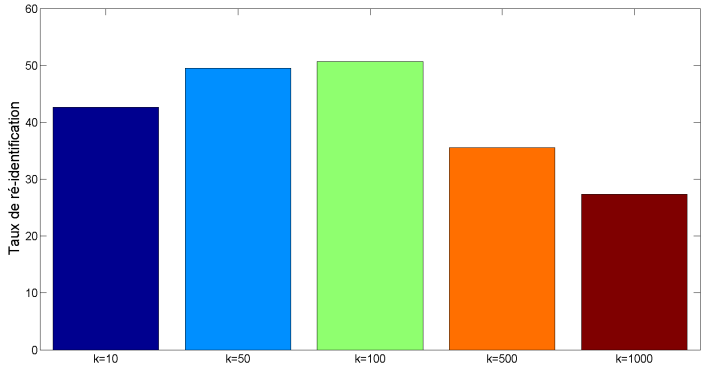


FIGURE 4.20 – Taux globaux de ré-identification en fonction de k.

La figure 4.19 montre que généralement les résultats s'améliorent en augmentant le nombre de classes k mais à partir de $k = 100$, les résultats commencent à se dégrader. En effet, pour k petit, les classes apprises ne sont pas représentatives, alors que pour k grand, la représentation est devenue très variée. Cela explique que le meilleur taux global de ré-identification est obtenu pour un nombre moyen de classes ($k = 100$) (figure 4.20). La table 4.8 montre les CCRs des 36 expériences réalisées avec $k = 100$.

TABLE 4.8 – CCR pour les différentes combinaisons des angles de vue de CASIA-A ($k = 100$).

Angle	**0˚**	**90˚**	**135˚**	**180˚**	**270˚**	**315˚**
0˚	90	20	45	75	20	50
90˚	10	100	50	10	75	40
135˚	20	30	95	20	20	75
180˚	60	35	50	80	15	45
270˚	10	90	40	15	100	40
315˚	30	35	75	5	20	100

Pour $k = 100$, la ré-identification est presque parfaite lorsque les vues de test et de référence sont identiques ou symétriques (différence angulaire = 180°) (figure 4.19). Dans ces deux cas, le taux de ré-identification excède 70%. Par contre, la performance du système se dégrade rapidement avec l'augmentation de la différence angulaire. Elle atteint 10% pour une différence angulaire = 90° et 5% pour une différence angulaire = 135° (table 4.8).

La comparaison des résultats des deux méthodes (locale et globale) fondées sur les SURFs est présentée dans la figure 4.21. Pour toute différence angulaire, la performance de la méthode locale est meilleure que la méthode globale. Même pour des petites différences angulaires, les CCRs des deux méthodes sont proportionnelles avec facteur de 4. Ainsi, la personne change sa représentation globale quand elle change sa direction de marche. Dans ce cas, les caractéristiques locales peuvent mieux capturer les régions similaires entre deux images de la même personne prise sous deux angles de vue différents.

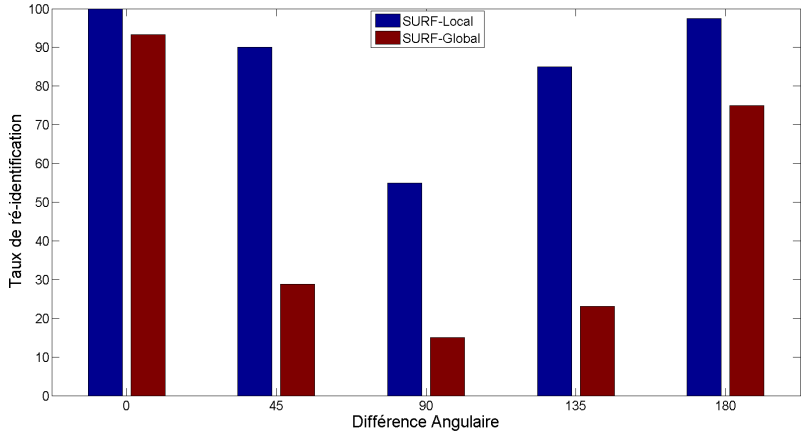


FIGURE 4.21 – Comparaison des résultats des SURFs : local vs global.

4.3.2.2 Description par histogramme de couleurs

4.3.2.2.1 Extraction des primitives La deuxième représentation globale testée est fondée sur les histogrammes de couleurs RGB. L'extraction de ce descripteur est décrite dans la section 3.3.2.2. Brièvement, il consiste à concaténer les histogrammes de couleurs de différentes régions de l'image ; le calcul de l'histogramme de chaque région est fondé sur la concaténation de 3 sous-histogrammes, de même dimension, correspondant aux canaux R, G et B (figure 4.22). Nous avons étudié les performances de cette description en fonction de trois paramètres :

1. Le chevauchement des régions de l'image.
2. Le nombre de régions par image.
3. La résolution de l'histogramme *i.e.* le nombre de composantes (bin) par sous-histogramme.

En évaluation, une configuration est définie par un triplet (n_x, n_y, b) où n_x (respectivement n_y) représente la dimension horizontale (respectivement verticale) de la grille divisant l'image, et b est le nombre de composantes par sous-histogramme. Pour une configuration donnée, la dimension du descripteur est égale à $D = n_x * n_y * b * 3$.

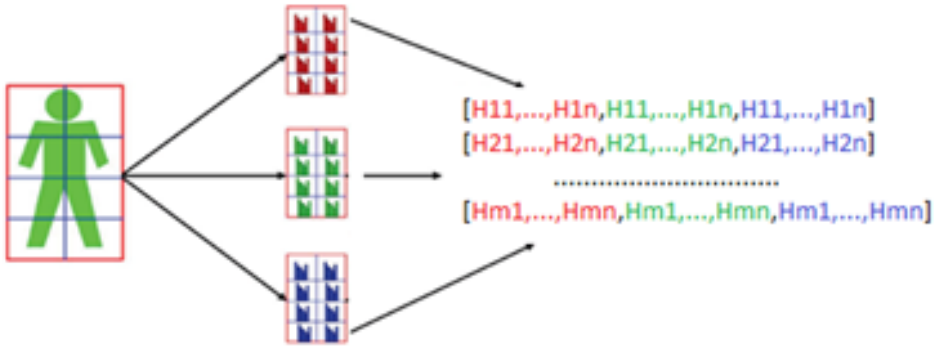


FIGURE 4.22 – Principe de construction de l'histogramme de couleurs.

4.3.2.2.2 Résultats Dans une première expérience, nous avons évalué l'influence de la division de l'image en une grille de régions *avec* ou *sans* chevauchement. Les régions d'une colonne ou ligne de la grille peuvent être non chevauchées (figure 4.23-a) ou chevauchées (figure 4.23-b). Pour les deux types de division, nous avons utilisé la même configuration (3, 3, 16). Les CCRs des deux types de division sont présentés dans la figure 4.24.

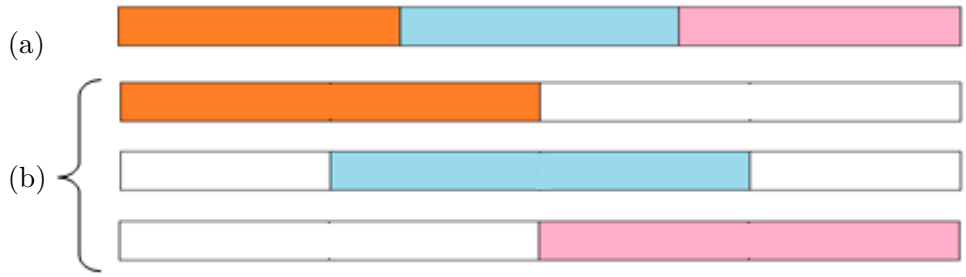


FIGURE 4.23 – Division en 3 régions : (a) sans chevauchement, (b) avec chevauchement.

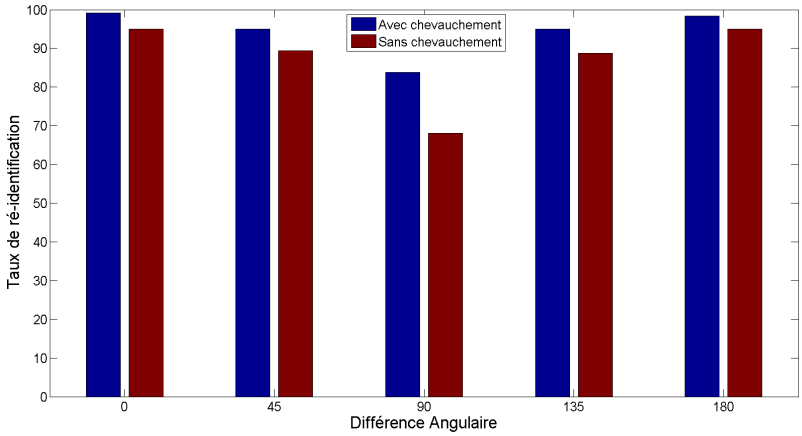


FIGURE 4.24 – CCR en fonction du type de division de l'image en des régions.

Les résultats prouvent que le descripteur généré par une division avec chevauchement est plus discriminant que sans chevauchement. En effet, le chevauchement des régions permet de tenir compte de la corrélation entre les régions.

Dans une deuxième expérience, et en gardant la division avec chevauchement, nous avons étudié également l'influence de l'augmentation du nombre de régions verticales et horizontales (figure 4.25). Les taux globaux de ré-identification des différents partitionnements sont représentés dans la figure 4.26. La table 4.9 présente la dimension du descripteur en fonction des partitionnements.

TABLE 4.9 – Dimension du descripteur en fonction des partitionnements.

Partitionnement $(n_x - n_y)$	$1-1$	$1-3$	$1-6$	$1-8$	$2-1$	$2-3$	$2-6$	$2-8$
Dimension (D)	48	144	288	384	96	288	576	768

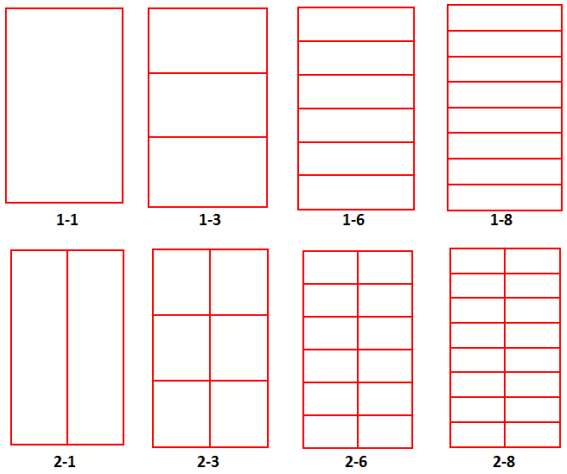


FIGURE 4.25 – Différents partitionnement de l'image.

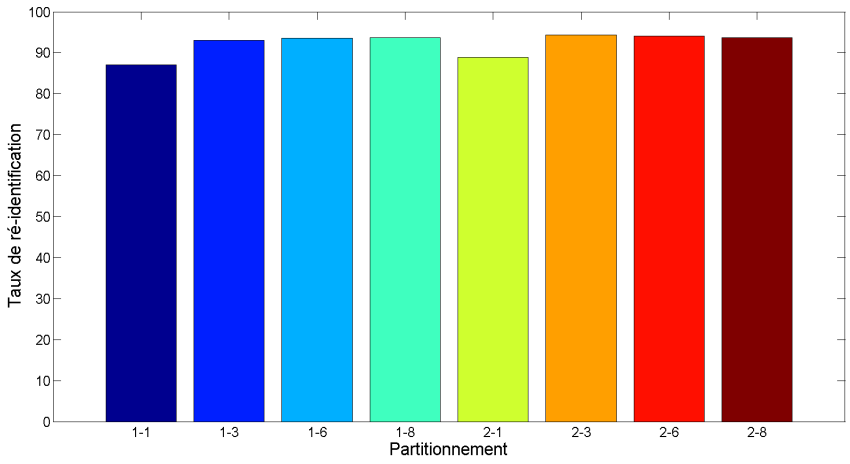


FIGURE 4.26 – Taux globaux de ré-identification en fonction des partitionnements.

La figure 4.26 montre que la performance du descripteur s'améliore de 6% dés une première division de l'image en trois régions horizontales. Ensuite, la performance du système reste presque stable en augmentant le nombre de régions horizontales ou verticales. Pour des partitionnements ayant la même performance, nous avons retenu le partitionnement ayant une dimension minimale du descripteur. Il s'agit du partitionnement $(1-6)$ qui donne un taux de ré-identification global égal à 93.5%.

Dans une troisième expérience en gardant le partitionnement $(1-6)$ avec chevauchement de régions, nous avons étudié également l'influence du nombre de composantes (b) par sous-histogramme. Différentes valeurs de b (table 4.10) ont été testées dont les résultats sont présentés dans la figure 4.27.

TABLE 4.10 – Dimension du descripteur en fonction de b.

Nombre de bins (b)	4	8	16	24	32
Dimension (D)	72	144	288	432	576

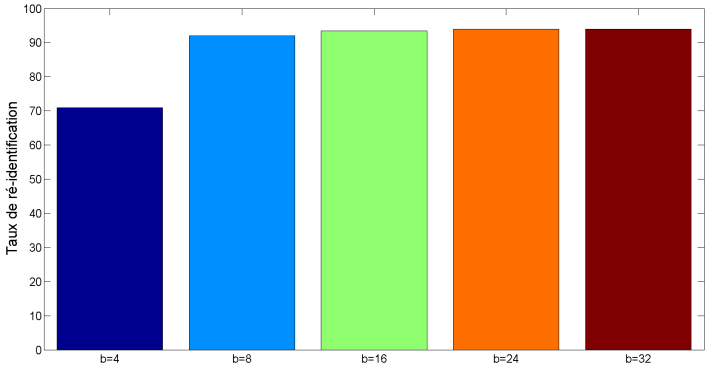


FIGURE 4.27 – Taux globaux de ré-identification en fonction de b.

En augmentant, la résolution de l'histogramme, la performance du système garde une stabilité à partir de $b = 8$. La table 4.11 montre les CCRs des 36 expériences réalisées pour la meilleure configuration retenue après les trois études faites. Il s'agit de la configuration $(1, 6, 8)$ avec chevauchement de régions.

TABLE 4.11 – Matrice de confusion de la meilleure configuration $(1, 6, 8)$.

Angle	**0˚**	**90˚**	**135˚**	**180˚**	**270˚**	**315˚**
0˚	100	85	95	100	80	95
90˚	80	95	100	85	90	95
135˚	95	90	100	95	85	100
180˚	100	85	95	100	80	95
270˚	90	100	100	90	100	95
315˚	95	90	100	95	90	100

4.4 Conclusion

Dans ce chapitre, nous avons évalué la faisabilité de deux approches de ré-identification dans un scénario simple : approche fondée sur un modèle biométrique et approche fondée sur l'apparence.

Quant au système biométrique, il consiste à modéliser l'aspect dynamique de la marche à l'aide d'un modèle stochastique (HMM) qui prend en entrée des séquences d'observations fondées sur le descripteur image HOG. Ce système est capable de ré-identifier les personnes quand les angles de vue de référence et de test sont similaires et quand la personne marche suffisamment dans le champ de vue de la caméra. Notre conclusion est que la modélisation de la marche par un HMM aurait un potentiel significatif en ré-identification si les angles de vue de référence et de test sont similaires, et si les personnes marchent suffisamment dans le champ de vue de la caméra, la deuxième condition n'étant pas satisfaite dans la base de données que nous avons considérée.

Pour une différence plus élevée entre les angles de vue de référence et de test, les approches fondées sur l'apparence seront l'alternative du système biométrique. Deux types de primitives ont été testés : les PIs SURF et les histogrammes de couleurs RGB. La

comparaison des performances de ces primitives d'apparence est présentée dans la figure 4.28.

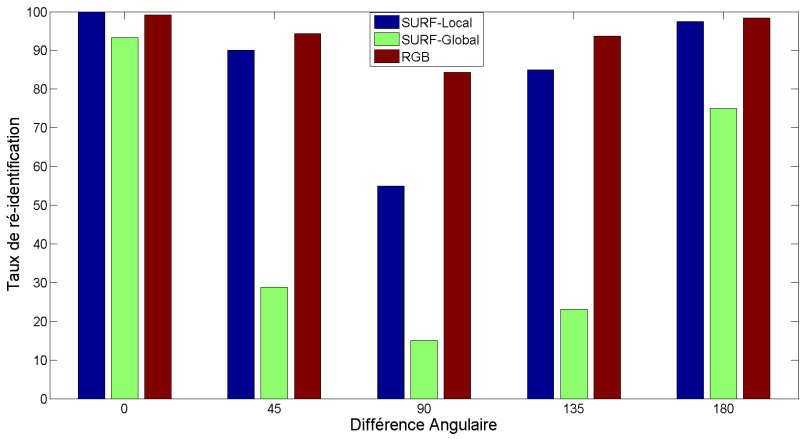


FIGURE 4.28 – Comparaison des méthodes fondées sur l'apparence.

Les PIs spatiaux ont montré une bonne robustesse même aux grandes différences angulaires. L'appariement local de ces PIs est plus robuste qu'un appariement global. Ceci est justifié par le fait que localement chaque PI de la requête participe par un vote de manière indépendante alors que globalement les BoFs des PIs sont très instables en raison de la provenance variée des PIs (personne, fond, objet, etc.).

Quand aux couleurs, le fait d'utiliser une seule caméra fixe avec un grand contraste rend l'effet des couleurs stable. Ceci prouve les performances obtenues par les primitives de couleurs. Ces conditions de stabilité de l'éclairage ne sont plus valides dans le scénario complexe.

Chapitre 5

Ré-identification des personnes dans un scénario complexe

Contents

5.1 Introduction

Après avoir étudié la faisabilité de certaines approches dans un scénario simple, nous traitons dans ce chapitre la ré-identification dans un scénario complexe. Dans un milieu complexe, le nombre de personnes est largement plus important que dans un scénario simple, les caméras d'enregistrement sont placées à l'extérieur dans des emplacements différents, les conditions de luminosité sont non contrôlées, plusieurs personnes peuvent passer simultanément dans les champs de vue des caméras et la direction de marche des personnes peut varier de manière significative.

L'objectif de ce chapitre est de mettre en œuvre un système de ré-identification convenable au scénario complexe. Le système biométrique, performant dans certaines conditions du scénario simple, est non adéquat à un scénario complexe. En effet, ce système nécessite que la personne ait marché suffisamment longtemps dans le champ de vue de la caméra alors que dans les bases de données actuelles la personne ne marche en moyenne que 3 secondes. De plus, ce système nécessite aussi que la personne ne change pas de direction de marche alors que dans le scénario complexe on n'a aucune contrainte sur la direction de marche de la personne. Nous proposons alors d'améliorer les systèmes de l'état de l'art fondés sur l'apparence en optimisant ses étapes les plus critiques à savoir la description de la personne et l'appariement des primitives.

Pour optimiser l'étape de la description, nous proposons d'exploiter la complémentarité de l'apparence de la personne et le style de son mouvement dans la vidéo. Pour décrire le mouvement, nous avons choisi les points d'intérêt spatio-temporels ; l'apparence est décrite par les PIs spatiaux. En outre, nous avons étudié la faisabilité d'autres descriptions de l'apparence à savoir les histogrammes de couleurs et d'autres types de PIs spatiaux.

Quant à l'étape d'appariement, nous nous intéressons à l'amélioration des méthodes d'appariement des PIs fondées essentiellement sur le calcul du plus proche voisin par des distances usuelles. Notre idée consiste à exploiter l'information de plusieurs PIs, ceux les plus corrélés de la base de référence, plutôt qu'un seul voisin pour apparier un PI test (spatial ou spatiotemporel). Dans ce travail, nous proposons comme méthode d'appariement la représentation parcimonieuse (RP) qui permet d'approximer un PI test par la combinaison, la plus proche, des PIs référence. Une utilisation des PIs référence sélectionnés consiste à chercher l'identité dominante (l'identité qui minimise le résiduel de reconstruction d'un PI) qui serait l'identité du PI test. En s'appuyant sur la RP, nous avons étudié la faisabilité de certains mécanismes de filtrage des appariements, fondés essentiellement sur le résiduel de reconstruction du PI test.

Une fois les deux systèmes d'apparence et de mouvement réalisés, nous proposons de les fusionner afin d'améliorer la robustesse en ré-identification. Le schéma de la fusion proposée est fondé sur le calcul de la somme pondérée des vecteurs de votes des deux systèmes et ensuite l'application de la règle de vote majoritaire pour reconnaître l'identité de la personne test.

Nous proposons également une analyse d'erreurs permettant d'identifier les sources d'erreurs principales de notre système pour dégager les voies d'amélioration les plus prometteuses.

Dans la suite, nous présentons notre approche de ré-identification fondée sur la RP par des primitives d'apparence (respectivement de mouvement) dans la section 5.2 (respectivement 5.3). Nous décrivons dans la section 5.4 le schéma proposé pour la fusion des deux approches. Pour chacune de ces primitives ainsi que pour la fusion, nous présentons les résultats obtenus essentiellement sur la base de données PRID-2011, et nous les comparons à l'état de l'art. Ensuite, une analyse des erreurs du système est présentée dans la section 5.5 et finalement la section 5.6 conclut le chapitre.

5.2 Ré-identification par l'apparence

Dans l'état de l'art, l'apparence d'une personne est décrite par des primtives de couleurs, de textures, de formes ou combinasion de ces primitives. Il n'y a aucun type de primitives systématiquement plus performant que les autres : chacun a des avantages et des limitations. Nous choisissons de décrire l'apparence par des PIs spatiaux, SURF, qui sont d'une part robustes à certains facteurs de complexité de la ré-identification à savoir la luminosité et le changement de l'angle de vue, et d'autre part ils ont déjà montré de bonnes performances en ré-identification.

Pour adapter les systèmes de l'état de l'art, fondés sur les PIs, aux conditions réelles d'un scénario complexe, nous nous sommes intéressés, dans un premier lieu, à l'amélioration de l'étape de mise en correspondances des PIs. Dans (Hamdoun *et al.*, 2008, de Oliveira et de Souza Pio, 2009), un SURF test est apparié directement au SURF référence le plus proche indépendamment de la nature des séquences vidéo. Par contre, dans un scénario complexe, les SURFs sont beaucoup plus bruités et ambigus et l'information du SURF référence le plus proche est insuffisante pour identifier le SURF test. Nous proposons d'exploiter la RP pour mieux approximer un PI test et ensuite le classifier. Nous décrivons dans la suite les principes de la représentation parcimonieuse ainsi que son application dans le contexte de PIs.

5.2.1 Représentation parcimonieuse

D'une manière générale, un vecteur est considéré parcimonieux si la majorité de ses coefficients sont nuls. Certains auteurs parlent de *p-parcimonie* : «un ensemble de signaux de dimension D de $\mathbb{R}^D$ est p-parcimonieux dans une base orthogonale de dimension $D >> p$, si on peut représenter, avec une bonne approximation, un signal quelconque de cet ensemble, à l'aide d'environ p composantes de cette base» (Martin, 2010).

L'objectif de la représentation parcimonieuse est de trouver une approximation d'un signal faisant intervenir le moins d'éléments d'un dictionnaire choisi préalablement. Parmi plusieurs solutions parcimonieuses, la convergence vers la solution optimale peut s'avérer complexe quand les signaux sont bruités. Cette problématique a fait l'objet de plusieurs études approfondies (Fuchs, 2005, Donoho *et al.*, 2006, Tropp, 2004).

5.2.1.1 Principe de la représentation parcimonieuse

Étant donné un signal $y \in \mathbb{R}^D$ et un dictionnaire $\Phi \in \mathbb{R}^{D \times K}$, ($K$ est le nombre de colonnes du dictionnaire où chaque colonne est un signal de dimension D), alors il existe une infinité de solutions du vecteur α, tel que :

$$y = \Phi\alpha \tag{5.1}$$

Le but est de trouver parmi l'ensemble des solutions possibles, celle qui est la plus parcimonieuse i.e. celle pour laquelle le vecteur α a un faible nombre d'éléments non-nuls. Théoriquement, la parcimonie est mesurée à l'aide de la norme l_0 correspondant au nombre d'éléments non-nuls d'un vecteur :

$$||x||_0 = \text{card}\{k, \text{ tels que } x_k \neq 0\}$$

La solution la plus parcimonieuse α_s de l'équation 5.1 peut être recherchée en minimisant la norme l_0 de α. Par conséquent, la formulation initiale du problème de la représentation parcimonieuse peut être écrite comme suit :

$$\alpha_s = \min_{\alpha} \|\alpha\|_0 \quad \text{sous} \quad y = \Phi\alpha \tag{5.2}$$

5.2.1.2 Algorithmes de représentation parcimonieuse

Pour résoudre l'équation 5.2, de nombreux algorithmes ont été proposés dans l'état de l'art. Ils peuvent être classés en deux groupes : algorithmes approximatifs et algorithmes exacts.

Les algorithmes approximatifs sont des algorithmes gloutons, itératifs et cherchent à trouver une approximation de l'équation 5.2. Ils raffinent itérativement l'estimation de α_s en sélectionnant un ou plusieurs atomes (colonnes) du dictionnaire séquentiellement pour approximer le signal y. Le principe de ces algorithmes est alors de sélectionner pas à pas les atomes les plus corrélés avec le signal y. On cite comme exemple d'algorithmes, le «Matching Pursuit» (MP) (Mallat et Zhang, 1993) et le «Orthogonal Matching Pursuit» (OMP) (Mallat et Zhang, 1993). La différence entre les deux algorithmes ainsi cités, réside dans la mise à jour des coefficients de la représentation parcimonieuse.

Quant aux algorithmes exacts, ils suggèrent de résoudre de manière exacte l'équation 5.2. Ils proposent une convexification du problème en remplaçant la norme l_0 par la norme l_1 (équation 5.3). En effet, le problème d'optimisation de la norme l_0 dans l'équation 5.2 est NP-complexe et des études récentes ont montré que la recherche de la solution du problème d'optimisation de la norme l_0 est équivalente au problème d'optimisation de la norme l_1.

$$\alpha_s = \min_{\alpha} \|\alpha\|_1 \quad \text{sous} \quad y = \Phi\alpha \tag{5.3}$$

En réalité, la contrainte d'égalité du problème $y = \Phi\alpha$ est trop forte si on veut obtenir une représentation parcimonieuse. En supposant que les données sont bruitées, y s'écrit comme suit :

$$y = b + e$$

où e est un bruit blanc gaussien, b est le signal source et y est le signal bruité observé. Plusieurs critères ont été proposés pour trouver une représentation parcimonieuse de y. On cite trois critères trouvés dans la littérature :

$$\min_{\alpha} \frac{1}{2}\|\Phi\alpha - y\|_2^2 \quad \text{sous} \quad \|\alpha\|_1 < \delta \tag{5.4}$$

$$\min_{\alpha} \|\alpha\|_1 \quad \text{sous} \quad \|\Phi\alpha - y\|_2^2 < \varepsilon \tag{5.5}$$

$$\min_{\alpha} \left(\|\Phi\alpha - y\|_2^2 + \lambda\|\alpha\|_1\right) \tag{5.6}$$

où δ, ε et λ sont des paramètres à fixer empiriquement. Pour des valeurs bien choisies de ces paramètres, les trois critères sont équivalents.

1. Le premier critère (équation 5.4) est introduit par Tibshirani en 1996 (Tibshirani,

1996). Il porte sur le nombre maximal de coefficients non-nuls participants à la représentation parcimonieuse. L'approche proposée dans (Tibshirani, 1996) est dénommée LASSO (pour «Least Absolute Shrinkage and Selection Operator»).

2. Le deuxième critère (équation 5.5) est connu sous le nom de «Basis Pursuit Denoising» (Fuchs, 1997, 1998). A l'inverse du LASSO, on fixe l'erreur de reconstruction au lieu de fixer le nombre maximal de coefficients non-nuls.
3. L'équation 5.6 décrit le troisième critère qui permet d'établir un compromis entre la valeur de l'erreur de reconstruction (premier terme) et le nombre de coefficients non-nuls (deuxième terme). Le paramètre λ sert à ajuster le poids que l'on souhaite donner à chacun des deux éléments de la somme. Plus la valeur de λ est grande, plus la parcimonie est privilégiée aux dépens de la qualité de la reconstruction. Inversement, si la valeur de λ est très petite, l'approximation obtenue sera de bonne qualité mais peu parcimonieuse.

Dans ce travail, nous avons considéré le troisième critère, i.e. le plus général, donc la solution parcimonieuse α_s de l'équation 5.3 est :

$$\alpha_s = \min_{\alpha}(\|\Phi\alpha - y\|_2^2 + \lambda\|\alpha\|_1) \tag{5.7}$$

La fonction $f : \alpha \mapsto \|\Phi\alpha - y\|_2^2 + \lambda\|\alpha\|_1$ est convexe et l'équation 5.7 devient un problème de minimisation de f, et peut être écrite comme suit :

$$\alpha_s = \arg\min_{\alpha}(f(\alpha)) \quad \text{avec} \quad \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K) \tag{5.8}$$

Plusieurs algorithmes ont été proposés dans l'état de l'art pour résoudre l'équation 5.8 tels que l'algorithme Descente par Coordonnées (DC) proposé dans (Friedman *et al.*, 2007) et l'algorithme LARS (pour «Large Angle Regression») (Bradley Efron et Tibshirani, 2004).

5.2.1.3 Représentation parcimonieuse dans un contexte de PIs

Dans un contexte de PIs, nous comptons exploiter la RP pour les classifier (apparier). Étant donné une personne requête et un ensemble de PIs référence associés à M identités (personnes), nous comptons chercher une RP pour chaque PI de la requête. Tout d'abord, l'ensemble de PIs référence $S_{i,j}$ est rangé dans une matrice Φ appelée dictionnaire où chaque colonne du dictionnaire est une description d'un PI : $S_{i,j} \in \mathbb{R}^D, i = 1 \ldots M, j = 1 \ldots k_i$, où k_i désigne le nombre de PIs référence associés à la $i^{ème}$ personne référence, et $K = k_1 + k_2 + \cdots + k_M$ désigne le nombre de PIs de la base de référence. Les descriptions des k_i PIs de la $i^{ème}$ personne référence constituent les colonnes de la matrice Φ_i :

$$\Phi_i = [S_{i,1}; S_{i,2}; \ldots; S_{i,k_i}] \tag{5.9}$$

Tous les K PIs de la base de référence sont combinés pour former la matrice Φ :

$$\Phi = [\Phi_1; \Phi_2; \ldots; \Phi_M] = [S_{1,1}; S_{1,2}; \ldots; S_{M,k_M}] \tag{5.10}$$

Par conséquent, nous désirons représenter un PI y comme une combinaison linéaire des PIs référence :

$$y = \Phi\alpha_s = [\Phi_1, \Phi_2, \ldots \Phi_M]\,\alpha_s \tag{5.11}$$

À ce stade, l'équation 5.8 est appliquée pour trouver α_s qui est un vecteur de coefficients parcimonieux, et il peut être écrit idéalement ainsi :

$$\alpha_s = [0; \ldots; 0; \alpha_{i,1}; \alpha_{i,2}; \ldots; \alpha_{i,k_i}; 0; \ldots; 0]^T \tag{5.12}$$

Dans ce cas idéal, α_s a des coefficients non-nuls uniquement associés à l'$i^{ème}$ identité correspondant à l'identité réelle de y. Cependant dans le cas où les PIs sont bruités, les coefficients associés à d'autres identités peuvent être non-nuls. Les coefficients non-nuls de α_s peuvent être utilisés pour déterminer l'identité de y.

La RP est le cœur de la méthode de mise en correspondance des SURFs. C'est une étape cruciale du système proposé pour la ré-identification. Dans la suite, nous décrivons les étapes de ce système.

5.2.2 Description du système proposé

Notre approche se déroule essentiellement en trois étapes : 1) extraction des primitives (SURFs), 2) classification des SURFs par la RP et 3) ré-identification de la personne par la règle de vote majoritaire. La figure 5.1 présente l'organigramme de l'approche proposée.

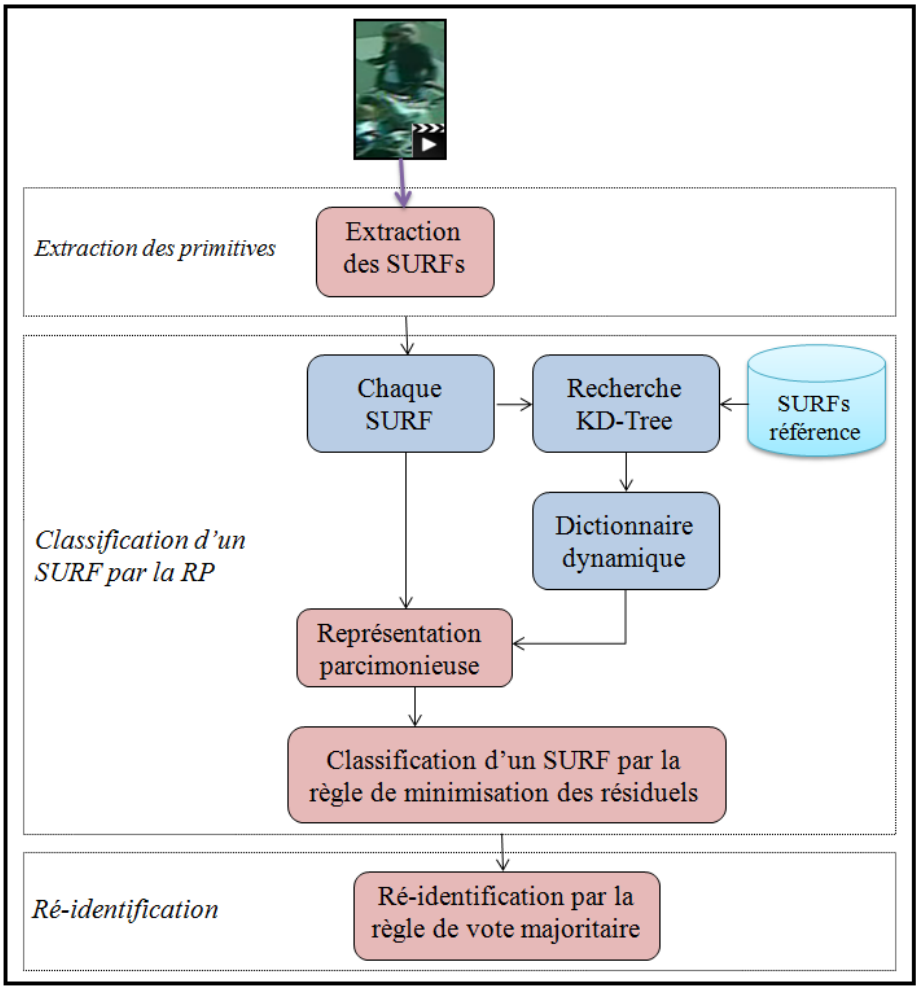


FIGURE 5.1 – Etapes du système de ré-identification fondé sur l'apparence.

5.2.2.1 Extraction des primitives

Nous avons extrait les mêmes primitives SURF exploitées dans le scénario simple. Ils sont encore utilisés dans ce scénario complexe vu leurs robustesses prouvées précédemment. Les principes de détection et description des SURFs sont décrits dans la section 3.3.1.2. Ensuite, chaque SURF est classifié via la RP comme montré ci-dessous.

5.2.2.2 Classification d'un SURF par la RP

La RP des signaux a été étudiée depuis deux décennies, mais elle n'est devenue populaire que récemment après son utilisation dans une application de reconnaissance de visages (Wright *et al.*, 2009). Ensuite, elle a été également utilisée dans d'autres domaines de vision par ordinateur telles que la reconnaissance de la démarche (Gong *et al.*, 2011), la reconnaissance de la parole (Gemmeke *et al.*, 2011) et la ré-identification des personnes (Truong Cong *et al.*, 2010b).

L'exploitation de la RP dans ce travail est différente de (Wright *et al.*, 2009) en 2 points :

1. La RP dans (Wright *et al.*, 2009) est globale (RP pour tout le visage), alors que la nôtre est locale (RP pour chaque SURF de la requête). De même, notre approche est différente de (Truong Cong *et al.*, 2010b) où le système de ré-identification est fondé sur une RP globale calculée pour chaque silhouette.
2. Dans (Wright *et al.*, 2009), pour chaque RP, le dictionnaire se compose de toutes les données de référence, tandis que dans notre approche, nous utilisons un dictionnaire réduit et dynamique, composé de quelques SURFs référence choisis comme expliqué plus tard. Ce choix du dictionnaire rend la méthode adaptée à des grandes bases de données.

Plus précisément, notre contribution réside dans la manière d'utiliser la RP pour classifier les SURFs. En effet, notre idée consiste à calculer pour chaque SURF test une RP locale, indépendamment des autres SURFs. Pour calculer une RP, un dictionnaire dynamique de petite taille est généré. Ce dictionnaire est dynamique dans le sens où il change pour chaque SURF. Donc, pour classifier un SURF, trois étapes sont appliquées : 1) construction du dictionnaire, 2) calcul de la RP et 3) affectation d'une identité.

5.2.2.2.1 Construction du dictionnaire D'une manière générale, le dictionnaire est composé de tous les éléments de la base de référence. Dans notre cas, la base de référence est de grande taille et composée d'environ un million de SURFs ; le temps de calcul pour trouver une RP d'un SURF est par conséquence énorme. Dans ce travail, un dictionnaire dynamique A est choisi pour chaque SURF, composé de N plus proches SURFs de la base de référence. Le dictionnaire A est de dimension NxD où chaque colonne représente une description SURF de dimension $D = 64$ et N est fixé empiriquement à 200. Le principe de construction du dictionnaire est illustré par l'algorithme 1.

Pour accélérer la recherche des plus proches voisins, un arbre KD (pour «K-Dimensions») est utilisé (Friedman *et al.*, 1977). Dans la suite, nous décrivons le principe de construction d'un arbre-KD ainsi que le principe de recherche des plus proches voisins par l'arbre-KD.

Algorithme 1 : CONSTRUCTION DU DICTIONNAIRE.

Données : ϕ **:** Matrice contenant tous les SURFs référence
N : Nombre de SURFs du dictionnaire
q : SURF requête
Résultat : A : dictionnaire
`// Construction de l'arbre KD`
$tree$ = Construction-arbre-KD(ϕ);
`// Tri des éléments de` ϕ `selon leurs distances à q par la recherche successive des plus proches voisins`
$\phi_{trié}$ = Tri-arbre-KD($tree, q$);
`// Construction du dictionnaire`
$A = [A_i]_{1 \leq i \leq N}$ = les N premiers SURFs de $\phi_{trié}$

Principe de construction de l'arbre-KD :

L'arbre-KD est une structure de données qui permet de stocker des vecteurs de dimension D (Bentley, 1975). L'arbre-KD est construit d'une manière récursive : à chaque nœud, on partitionne les données en deux sous-ensembles selon la valeur de l'une de ses composantes jusqu'à atteindre un critère d'arrêt (un exemple est présenté dans la figure 5.2). Algorithmiquement, le principe de construction d'un arbre-KD est résumé par les points suivants :

- Recherche de la dimension qui servira à séparer les données : il s'agit de la dimension où les données sont les plus éparpillées.
- Recherche du pivot : il s'agit du vecteur dont la valeur associée à la dimension sélectionnée permet de séparer les données en deux. Ce pivot sélectionné serait un nœud de l'arbre.
- Construction récursive de l'arbre :
  - ➢ Un sous-arbre est créé à gauche du nœud avec les données inférieures ou égales au pivot.
  - ➢ Un sous-arbre est créé à droite du nœud avec les données supérieures au pivot.
  - → Dans cette construction, la comparaison des vecteurs avec le pivot consiste à comparer leurs valeurs associées à la dimension de séparation.
- Critère d'arrêt : dès qu'un sous-arbre contient un nombre d'éléments inférieur à un seuil, une feuille contenant ces éléments est créée.

Principe de recherche des plus proches voisins par l'arbre-KD :

L'arbre-KD a été utilisé pour accélérer la recherche des plus proches voisins. Il a été employé pour cet objectif pour la première fois dans (Friedman *et al.*, 1977). Pour un vecteur donné, la recherche de son plus proche voisin (ou de ses plus proches voisins) à partir des données stockées dans un arbre-KD est exacte, mais reste un peu lente si on veut obtenir des résultats en temps réel quand on a un arbre-KD avec un grand nombre de niveaux. Pour remédier à ce problème, un critère a été ajouté, dans l'état de l'art, pour accélérer la recherche, c'est le nombre maximal des nœuds à visiter pour trouver le plus proche voisin. Le résultat approximatif de la recherche du KNN avec l'ajout de ce critère est obtenu rapidement et il est proche de celui de la recherche du KNN sans approximation. Dans notre travail, le nombre maximal à visiter pour trouver le plus proche voisin est fixé empiriquement à 1000.

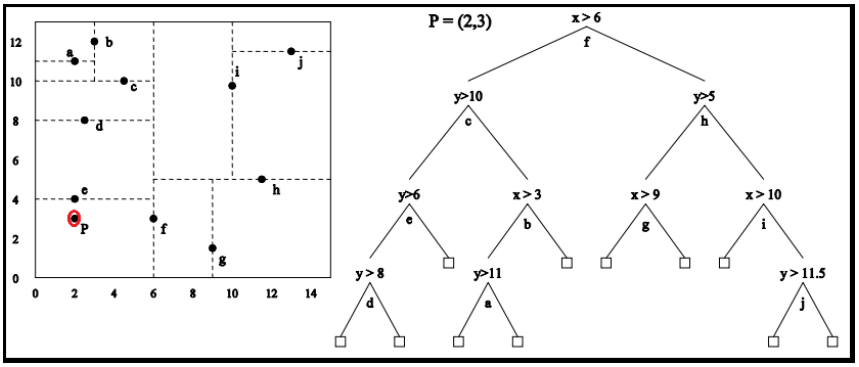


FIGURE 5.2 – Exemple de construction d'un arbre-KD ($K = 2$). P est un vecteur requête. (figure issue de (Thomas-Dietterich, 2005)).

Plusieurs heuristiques ont été utilisées pour chercher les plus proches voisins d'un vecteur requête via l'arbre-KD. Dans ce travail, nous utilisons l'algorithme BBF (pour «Best Bin First») (Beis et Lowe, 1997). Pour une requête donnée, le principe de BBF se déroule en deux étapes :

- Parcourir l'arbre du nœud principal jusqu'aux feuilles en faisant un test de comparaison à chaque nœud : si la valeur du pivot à la dimension de séparation est supérieure à la valeur de la requête à la même dimension, alors on parcourt le sous-arbre gauche, sinon on parcourt le sous arbre droite. Les nœuds visités sont enregistrés dans une file d'attente de priorité (FAP).
- Soit d la distance entre la requête et le dernier élément enregistré dans la FAP. Au départ, cet élément est initialisé au plus proche voisin. Ensuite, dans une étape de *backtracking*, chaque nœuds de la FAP est comparée à la requête :
  - ➢ si la distance entre eux est inférieur à d alors 1) le sous-arbre n'est pas encore parcouru est ajouté au FAP et 2) ce nœud devient le plus proche voisin actuel.
  - ➢ Sinon, on passe au nœud suivant de la FAP.
  - ➢ Le *backtracking* est arrêté si la FAP est vide ou s'il on atteint le nombre maximal des nœuds à visiter.

La table 5.1 montre le déroulement de l'algorithme de recherche du plus proche voisins du vecteur $P = (2, 3)$ dans l'arbre-2D présenté dans la figure 5.2. Au départ, le FAP contient les nœuds {d,e,c,f} et f est le plus proche voisin actuel. En *backtracking*, on a $\text{dist}(d, P) > \text{dist}(e, P)$, $\text{dist}(e, P) < \text{dist}(c, P)$ et $\text{dist}(e, P) < \text{dist}(f, P)$ alors aucun sous-arbre n'est ajouté à la FAP. Le *backtracking* est arrêté quand le FAP est vide, le plus proche voisin actuel de P est e

TABLE 5.1 – Exemple de recherche du plus proche voisin dans un arbre-2D.

Contenu de la FAP	**Comparaisons**	**Plus proche voisin actuel**
{d,e,c,f}	Etat initial	d
{d,e,c,f}	$\text{dist}(d, P) > \text{dist}(e, P)$	e
{e,c,f}	$\text{dist}(d, P) > \text{dist}(e, P)$	e
{c,f}	$\text{dist}(e, P) < \text{dist}(c, P)$	e
{f}	$\text{dist}(e, P) < \text{dist}(f, P)$	e
{}	Etat d'arrêt	e

5.2.2.2.2 Représentation parcimonieuse Comme les conditions de ré-identification sont complexes, les éléments (atomes) du dictionnaire sont bruités et ne sont pas fortement corrélés. En effet, pour trouver la RP (algorithme 2), nous avons utilisé l'algorithme DC, qui d'après la littérature, est performant dans le cas où la corrélation entre les éléments du dictionnaire est petite (Mairal, 2009).

Algorithme 2 : Représentation parcimonieuse (RP).

Données : q : un SURF requête
A : le dictionnaire correspond à q
λ : paramètre empirique
Résultat : α_s : vecteur de coefficients parcimonieux
`// Résoudre l'équation suivante par l'algorithme DC`
$\alpha_s = \min_{\alpha}(\frac{1}{2}||y - A\alpha||_2^2 + \lambda||\alpha||_1)$

L'algorithme DC prend en entrée un SURF requête q et le dictionnaire A qui lui correspond et fournit en sortie une RP sous forme d'un vecteur parcimonieux α_s dont la plupart des coefficients sont égaux à zéro (figure 5.3).

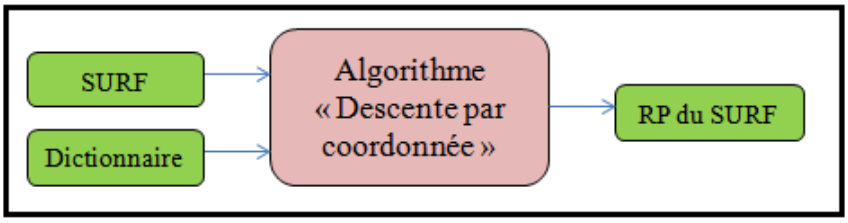


Figure 5.3 – Entrées/Sorties de l'algorithme DC.

C'est un algorithme cyclique dont le principe consiste à initialiser d'abord le vecteur de coefficients α (α^0) et ensuite choisir à chaque itération j une seule coordonnée de α^j (α_i^j) de façon que la fonction $f_i : \alpha \mapsto \|\Phi\alpha_i - y\|_2^2 + \lambda\|\alpha_i\|_1$ soit minimale. Finalement, la valeur de α_i^j est mise à jour à la valeur $(\alpha_i^j)^*$ par la méthode de descente de coordonnées (Friedman *et al.*, 2007). Ce principe de fonctionnement de DC est illustré par l'algorithme 3.

Algorithme 3 : Descente par coordonnées (DC).

Données : $J \in \mathbb{N}$, $\alpha^0 \in \mathbb{R}^N$
Résultat : α^J
pour $j \leftarrow 0$ **jusqu'à** $J - 1$ **faire**
  $i \leftarrow choix\_coordonnée(\alpha^j)$;
  `// mise à jour de` (α_i^j)
  $\alpha^{j+1} \leftarrow [\alpha_1^j, \ldots, (\alpha_i^j)^*, \ldots, \alpha_N^j]$
fin

5.2.2.2.3 Affectation d'un SURF à une identité Les coefficients non-nuls de la RP sont utilisés pour attribuer une identité au SURF test. Pour connaître son identité, un résiduel est calculé pour chaque identité i ayant au moins un coefficient non-nul dans la RP de la manière suivante :

Soit q un SURF test, A le dictionnaire lui correspondant, α_s la RP de q par rapport à A et L le nombre d'identités ayant au moins un coefficient non-nul dans α_s. D'abord, nous calculons $x_i, 1 \leq i \leq L$ comme suit :

$$x_i = [0, \ldots 0, \alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,k_i}, 0, \ldots, 0] \tag{5.13}$$

où x_i est un vecteur de coefficients obtenus à partir de α_s dont tous les éléments sont nuls à l'exception de ceux qui sont associés à l'identité i. L'identité dominante j satisfait l'égalité suivante :

$$j = \arg\min_i \|Ax_i - q\|_2^2 \tag{5.14}$$

j correspond à l'identité minimisant l'erreur de reconstruction de q. Enfin, q est identifiée comme étant un SURF de la personne j satisfaisant l'équation 5.14. L'algorithme 4 résume la procédure d'affectation d'un SURF à une identité.

Algorithme 4 : AFFECTATION D'UNE IDENTITÉ.

Données : α_s **:** la RP du SURF requête q correspond au dictionnaire A
q : SURF requête
Résultat : j : identité de q
pour *chaque identité i ayant k_i coefficients non-nuls* **faire**
  `// Calculer` x_i
  $x_i = [0, \ldots, 0, \alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,k_i}, 0, \ldots 0]$;
  `// Calculer l'erreur de reconstruction` r_i
  $r_i = ||Ax_i - q||_2^2$;
fin
$j = \arg\min_i(r_i)$;

5.2.2.3 Vote Majoritaire

Finalement, après avoir identifié tous les SURF d'une requête, nous utilisons la règle de vote majoritaire pour connaître l'identité de la séquence requête. Il s'agit de la même méthode utilisée dans le système d'apparence proposé dans le scénario simple. Le principe est que pour chaque SURF test q, un vote est ajouté à la personne associée à l'identité référence j minimisant son erreur de reconstruction. La personne qui obtient la majorité des votes est considérée comme la personne ré-identifiée.

5.2.3 Expériences et résultats

Ce système de ré-identification a été évalué sur des bases de données filmées dans des conditions complexes. La performance du système dépend du niveau de parcimonie attribué à la RP. Pour un test donné, plusieurs valeurs empiriques de λ ont été testées. Ces valeurs de λ correspondent à des cas où la RP est trop parcimonieuse et des cas où elle est peu parcimonieuse. Comme implémentation de la RP, nous nous sommes servis de celle de la boîte à outils statistiques de MATLAB-2012. Nous présentons ci-dessous la mesure de performance adoptée et les bases de données sélectionnées de l'état de l'art. Par la suite, nous montrons les résultats obtenus sur les bases de données choisies.

Mesure de performance

La performance d'une approche est mesurée par la courbe CMC (pour «Cumulative Match Characteristic»). Comme on compte ré-identifier une personne requête parmi N personnes référence, la courbe CMC représente la probabilité de trouver la bonne personne référence parmi les r meilleurs appariements. r est appelé le rang de ré-identification. Au rang 1, la valeur de la courbe CMC est appelée CCR (pour «Correct Classification Rate»).

Autrement dit, la courbe CMC donne le pourcentage des personnes ré-identifiées en fonction du rang de ré-identification. La courbe CMC est croissante et converge vers la valeur 100% de taux de ré-identification (figure 5.4).

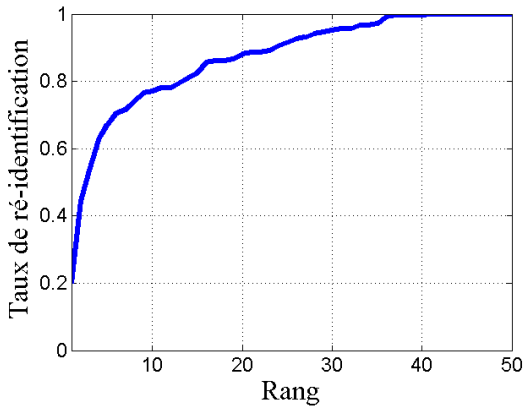


FIGURE 5.4 – Exemple d'une courbe CMC.

Bases de données

Quatre bases de données multi-échantillons sont adéquates à un scénario complexe : PRID-2011, CAVIAR4REID, ETHZ-REID et i-LIDS-REID. La base de données i-LIDS-REID est privée et requiert la licence "iLIDS MCTS" dont nous ne disposons pas. Quant à la base de données ETHZ-REID, elle est filmée par une seule caméra et elle ne correspond donc pas à un scénario réel de ré-identification. Pour cela, nous avons évalué notre approche sur les deux bases de données : CAVIAR4REID et PRID-2011.

5.2.3.1 Résultats obtenus sur CAVIAR4REID

CAVIAR4REID est une base de données multi-échantillons composée de 72 personnes dont 50 personnes sont filmées par deux caméras disjointes (chaque caméra fournit 10 images par personne). D'après sa description (section 2.5.1), les données de CAVIAR4REID sont adéquates à un scénario de ré-identification complexe car elles sont filmées par deux caméras différentes et installées dans deux endroits différents. En outre, les images de CAVIAR4REID sont de différentes échelles et ont un éclairage variable (figure 5.5). Sur CAVIAR4REID, nous avons détecté en moyenne 247 SURFs par personne test et 270 SURFs par personne référence.

En évaluation, nous avons choisi empiriquement $\lambda = 10^{-2}$. Dans ce cas, le nombre moyen des coefficients non-nuls d'une RP est égale à 12.

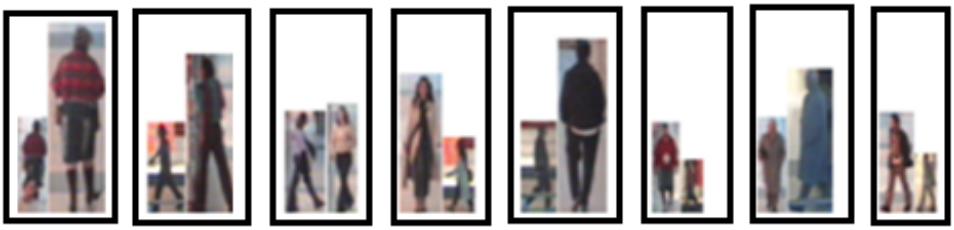


FIGURE 5.5 – Exemple d'images de CAVIAR4REID. Chaque paire décrit le même individu.

La figure 5.6 montre la courbe CMC de notre approche fondée sur la RP comparée à celle obtenue par l'approche proposée dans le scénario simple (méthode fondée sur le plus proche voisin (1-NN)). La table 5.2 montre les CCRs des différentes méthodes de l'état de l'art évaluées sur CAVIAR4REID.

TABLE 5.2 – Comparaison des résultats sur CAVIAR4REID.

Approches	**CCR(%)**
Auteurs de (Bak *et al.*, 2010b) apparaît dans Bazzani *et al.* (2013)	10
(Farenzena *et al.*, 2010)	10
(Kviatkovsky *et al.*, 2013)	10
Approche du chapite 4 (1-NN)	16
(Chen *et al.*, 2009)	17
(**?**)	19
Notre approche : RP	20

Les approches fondées sur des caractéristiques d'apparence telles que (SDALF) (Farenzena *et al.*, 2010) et MRGC (Bak *et al.*, 2010b) atteignent 10% de CCR. L'approche fondée sur la correspondance des SURFs par la méthode 1-NN (chapitre 4) réalise un CCR égale à 16%. Quant à l'approche fondée sur la RP, elle atteint un CCR de 20% qui est légèrement meilleur comparé à l'approche présentée dans Chen *et al.* (2009) et fondée essentiellement sur des caractéristiques de couleurs. Les sources des courbes CMC des autres approches sur CAVIAR4REID (table 5.2)ne sont pas disponibles pour les reproduire avec les nôtres sur la même figure. Ces résultats illustrent l'apport de l'appariement fondé sur la RP par rapport à l'appariement par le 1-NN.

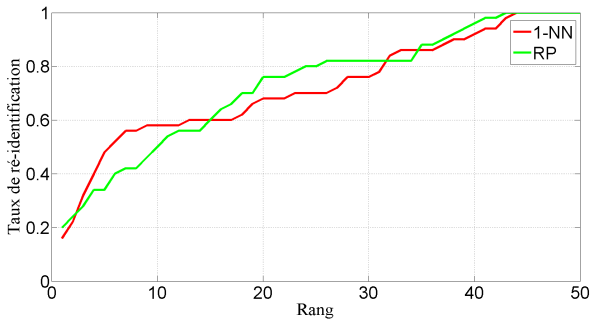


FIGURE 5.6 – Courbes CMC obtenues sur CAVIAR4REID.

Dans l'état de l'art, CAVIAR4REID a été évaluée en utilisant un autre protocole. Ce

dernier consiste à diviser la base de données en un ensemble d'apprentissage (14 personnes) et un ensemble de test (36 personnes). Ce protocole concerne généralement les approches supervisées. Parmi ces approches, on cite (Pedagadi *et al.*, 2013) qui atteint un CCR = 36.19% et (Liu *et al.*, 2014) qui atteint un CCR = 49.1%.

5.2.3.2 Résultats obtenus sur PRID-2011.

PRID-2011 est une base de données multi-échantillons, composée de 749 personnes en référence et 385 en test filmées par deux caméras disjointes. 200 personnes sont communes à la base de référence et de test. D'après sa description (section 2.5.1), les données de PRID-2011 sont adéquates à un scénario complexe. En effet, elle est composée de plusieurs centaines de personnes filmées dans des conditions réelles et non contrôlées ; les deux caméras de surveillance sont installées dans la rue et dans deux endroits différents. En outre, les images de PRID-2011 possèdent une grande variation d'éclairage et les personnes passent parfois en foule et portent des objets (figure 5.7).

(a) Personne 1 (b) Personne 2

FIGURE 5.7 – Exemples d'images de PRID-2011. Rangées supérieures et inférieures correspondent aux différentes caméras.

La table 5.3 présente l'évolution des performances de notre approche en fonction des valeurs de λ choisies empiriquement. En outre, elle illustre le niveau de parcimonie pour chaque valeur de λ. La figure 5.8 montre que le meilleur résultat obtenu correspond à une valeur de $\lambda = 10^{-3}$. En augmentant λ, la RP devient très parcimonieuse. D'un autre côté, en diminuant λ la RP n'est plus parcimonieuse. Pour $\lambda = 10^{-3}$, la RP contient en moyenne 22 coefficients non-nuls.

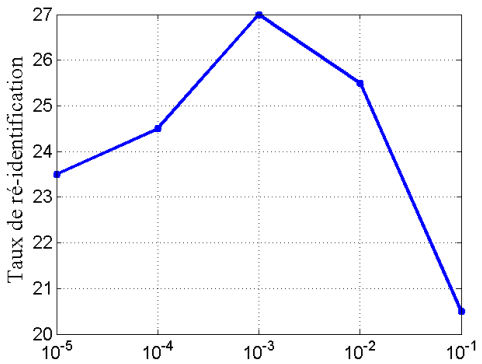


FIGURE 5.8 – Taux de ré-identification en fonction de λ.

TABLE 5.3 – Résultats en fonction de la parcimonie.

λ	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
CCR (%)	23.5	24.5	27	25.5	20.5
Nombre moyen de coefficients non-nuls	186	73	22	14	1

La figure 5.9 présente les courbes CMC (de rang 1 au rang 10) des approches de l'état de l'art évaluées sur PRID-2011. La table 5.4 présente une comparaison des CCRs des approches présentées dans la figure 5.9. La table 5.4 montre que notre approche fondée sur la correspondance des SURFs par la RP apporte au rang 1 une amélioration de 4.5% par rapport à l'approche fondée sur la correspondance des SURFs par le 1-NN. Le CCR de notre approche dépasse favorablement l'approche proposée dans (Hirzer *et al.*, 2011) qui combine deux modèles d'apparences.

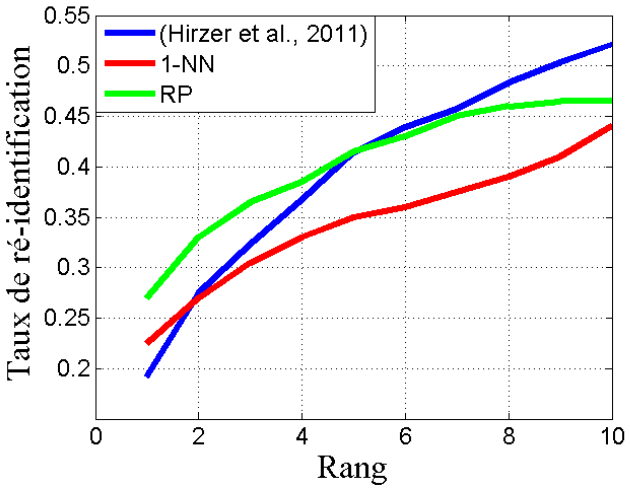


FIGURE 5.9 – Courbes CMC sur PRID-2011.

TABLE 5.4 – Comparaison des résultats sur PRID-2011.

Approches	**CCR (%)**
(Hirzer *et al.*, 2011)	19.18
Approche du chapite 4 (1-NN)	22.50
Notre approche : RP	27.00

5.2.3.3 Analyse des résultats : apport de la RP

L'approche proposée est une adaptation de la méthode d'appariement des SURFs à un scénario complexe : il s'agit de la correspondance par la RP. Sur la base de données PRID-2011, la performance de notre approche surpasse celle de l'approche fondée sur le 1-NN pour tous les rangs. Au rang 1, notre approche réalise une augmentation de 4.5% en taux de ré-identification par rapport au 1-NN. Cette augmentation est très significative vu la grande taille de la base de données et prouve que la RP est plus riche que le 1-NN pour classifier un SURF. Comparant à (Hirzer *et al.*, 2011), notre approche devient moins bonne à partir du rang 6. Par contre, elle est meilleure du rang 1 au rang 5 et réalise

une amélioration de 7.82% en taux de ré-identification au rang 1. Cette amélioration est significative et montre le pouvoir de la RP comparée à (Hirzer *et al.*, 2011) qui fusionne deux classifieurs. La RP est encore meilleure sur CAVIAR4REID au rang 1. L'amélioration est faible par rapport à celle obtenue sur PRID-2011, c'est probablement parce que la base de données est petite et uniquement quelques images sont disponibles par personne, contrairement à PRID-2011.

L'apport de la RP par rapport au 1-NN est mis en valeur dans les cas où des SURFs référence sont proches d'un SURF test de la même personne mais pas dans les premiers rangs. En tenant compte de plusieurs proches voisins pour calculer la RP, l'appariement du SURF test peut être rectifié. Cette hypothèse est illustrée par l'exemple présenté dans la table 5.5. Dans ce cas, le SURF test de la personne numéro 68 (point bleu de la figure 5.10-a) est apparié à un SURF de la personne 170 (point rouge de la figure 5.10-b) par la méthode 1-NN (table 5.5). Quant à la RP, elle cherche la meilleure approximation du SURF test par la combinaison de plusieurs PIs référence. Par conséquent, la RP permet d'affecter un poids fort pour l'identité 68 et donc cette identité obtient le minimum des résiduels (table 5.5). Le SURF test est alors classifié comme un SURF de la même identité référence.

(a) (b)

FIGURE 5.10 – (a) En bleu : SURF test, (b) en rouge : le SURF référence le plus proche.

TABLE 5.5 – Statistiques sur les 9 plus proches SURFs référence d'un SURF test de la personne numéro 68.

Identités des SURFs référence	Distance par rapport le SURF test	Coefficients de la RP	Résiduels
170	**0.17**	**0.04**	**0.95**
678	0.19	0	1.00
675	0.19	0.04	0.95
404	0.20	0.31	0.73
196	0.20	0.06	0.93
68	**0.21**	**0.57**	**0.54**
681	0.21	0	—
685	0.22	0	—
439	0.22	0	—

5.2.3.4 Influence des paramètres de la RP

Pour améliorer d'avantage ce système, nous avons étudié, sur PRID-2011, l'évolution de sa performance en fonction des deux paramètres empiriquement choisis : le nombre de nœud à visiter ($nMax$) utilisé par l'arbre-KD pour sélectionner les éléments du dictionnaire et la taille du dictionnaire (N).

Dans une première expérience, différentes valeurs de $nMax$ ont été testées. La performance du système n'a pratiquement aucune dépendance avec $nMax$. Théoriquement, en augmentant $nMax$, on augmente la chance de trouver la bonne approximation des plus proches voisins. Par contre, la table 5.6 montre qu'en augmentant $nMax$ dix fois, (pour $nMax = 500$ et $nMax = 5000$), la performance du système reste stable. Ceci prouve qu'on a à peu près le même dictionnaire. Cela est justifié par le fait que dans les vidéos, les SURFs sont répétables et un SURF test peut être à la même distance de plusieurs SURFs référence. Dans ce cas, l'un de ces SURF référence peut être retrouvé sans parcourir un grand nombre de nœuds de l'arbre.

TABLE 5.6 – CCRs en fonction de $nMax$.

$nMax$	200	500	1000	5000
CCR (%)	26	26.5	27	26.5

Dans une deuxième expérience, nous avons testé des tailles différentes du dictionnaire (table 5.7). Quand le dictionnaire contient un seul élément, l'approche fondée sur la RP est équivalente à 1-NN. La performance du système s'améliore avec l'augmentation de la taille du dictionnaire. À partir d'un dictionnaire de 100 SURFs, la performance du système devient stable.

TABLE 5.7 – CCRs et temps moyen de test en fonction de D.

N	1	10	50	100	200	500
CCR (%)	22.5	24.5	25.5	27	27	27.5

Selon ces expériences, les valeurs de $nMax$ ($nMax = 1000$) et N ($N = 200$), utilisées dans nos tests, sont optimales.

5.2.4 Etudes des mécanismes de filtrage des correspondances

L'étape de filtrage consiste à écarter les correspondances les moins fiables. Dans le scénario simple, une correspondance est définie par deux PIs (référence et test). Dans le cas du scénario complexe, une correspondance est définie par un PI test et l'identité référence qui minimise l'erreur de son reconstruction par la RP. Cette erreur de reconstruction est considérée comme le score d'une correspondance.

Notre méthode de filtrage est inspirée du critère de Lowe (Lowe, 2001) : soit P un SURF test, id_1 et id_2 sont les deux meilleures identités référence qui reconstruit P avec deux résiduels respectivement r_1 et r_2 ($r_1 < r_2$). Une correspondance est dite fiable si $R = \frac{r_1}{r_2} < seuil$. ($seuil$ est une valeur empirique). La figure 5.11 visualise la distribution

de r_1 et r_2 des SURFs test de PRID-2011 alors que la figure 5.12 montre la distribution de R.

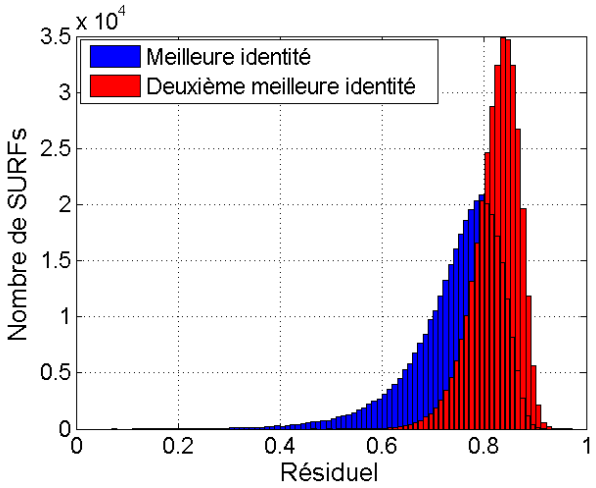


FIGURE 5.11 – Distributions des résiduels.

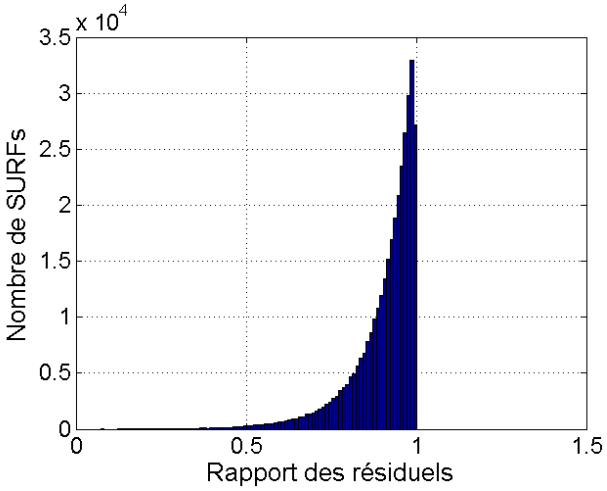


FIGURE 5.12 – Distribution de rapports des résiduels.

Intuitivement, on s'attend à que r_1 et r_2 aient deux valeurs non proches, mais la figure 5.12 montre l'inverse. Cela est dû à la grande ambigüité des SURFs. En évaluation, on arrive à obtenir des performances de 26.5% en filtrant 80% des SURFs, mais le meilleur résultat n'est obtenu qu'en utilisant presque tous les SURFs (table 5.8).

TABLE 5.8 – CCR en fonction de taux de filtrage.

seuil	0.6	0.7	0.8	0.85	0.90	0.95
CCR (%)	19.5	21	23.5	26.5	26	27
% des SURFs utilisés	2	4.9	12.4	20.4	34.2	58.1

Cette méthode de filtrage ne tient compte que de peu d'information sur les correspondances pour les filtrer malgré le fait que les SURFs sont bruités et par conséquent les

correspondances sont très bruités. La supervision du mécanisme de filtrage et la prise en compte d'autres caractéristiques de la correspondance (par exemple, le résiduel global de la reconstruction du SURF, etc..) sont d'autres voies à considérer pour obtenir un filtrage plus efficace.

5.2.5 Etudes d'autres descriptions d'apparence

Nous nous intéressons dans cette section à l'étape de description de l'apparence dans les conditions complexes. Nous avons étudié d'une part d'autre PIs à savoir les PIs de Harris (section 3.3.1.2)et d'autre part les histogrammes de couleurs évaluées dans le scénario simple. Quant aux PIs de Harris, ils sont caractérisés par une bonne répétabilité mais ils ne tiennent pas compte de la variation d'échelle entre les images. Une comparaison des SURFs et PIs de Harris est illustrée par la table 5.9.

TABLE 5.9 – Caractéristiques des SURFs et PIs de Harris.

Primitives	SURF	PIs de Harris
Détection	Détecteur de blob fondé sur la Hessienne	Détecteur de coins Harris 2D
Description	Réponses aux ondelettes de Haar	
Caractéristiques	Invariant aux transformations géométriques	Avoir une bonne répétabilité
Nombre moyen de PIs par séquence test	1733	5521
Nombre moyen de PI par séquence référence	1123	2993

La figure 5.13 représente les courbes CMC obtenues par les deux types de PIs sur PRID-2011. Elle montre que les PIs de Harris sont plus performants que les SURFs pour certains rangs, alors que les SURFs sont les plus performants pour d'autres rangs. Au rang 1, SURF réalise un CCR de 22.5% comparé à 21.5% réalisé par les PIs de Harris. Bien que le nombre des PIs de Harris soit environ égal au double des SURFs (ce qui augmente énormément la durée de la phase de ré-identification), la performance de PIs de Harris dépasse légèrement celle des SURFs uniquement pour certains rangs.

Quant aux primitives de couleurs, dans un scénario simple, elles ont montré une grande efficacité surtout quand l'éclairage est stable, une seule caméra filme les données de référence et de test, il n'y a pas d'occultations et la personne ne change pas ses vêtements (la personne n'enlève pas son manteau par exemple dans un champ de vue d'une caméra et le porte dans un autre). Dans ce scénario complexe, ces conditions ne sont plus vérifiées. En effet, l'éclairage n'est pas stable car les personnes sont filmées par deux caméras différentes et installées dans deux endroits différents ; l'occultation peut se produire car les gens passent en foule. En outre, les personnes passent parfois avec des objets ce qui influe sur les couleurs dominantes de l'image. Dans cette partie, nous nous intéressons à la faisabilité des primitives de couleurs dans un scénario complexe.

Dans ce scénario, les images d'une même personne prises par deux caméras différentes peuvent présenter une variation significative de couleurs. Dans la littérature, pour remédier

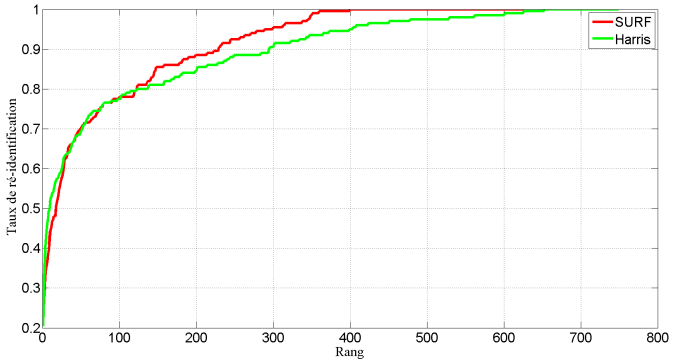


FIGURE 5.13 – Courbes CMC des SURFs et PIs de Harris sur PRID-2011.

à ce problème, des prétraitements ont été appliqués. Dans ce travail, d'une part, nous avons utilisé une méthode d'égalisation des couleurs qui consiste à centrer les couleurs des pixels comme suit :

$$R_{i,n} = R_i - mean(R),\ G_{i,n} = G_i - mean(G),\ B_{i,n} = B_i - mean(B)$$

où pour un pixel donné i, R_i, G_i, et B_i sont les valeurs des canaux RGB avant la normalisation et $R_{i,n}$, $G_{i,n}$, et $B_{i,n}$ sont les valeurs des canaux RGB après la normalisation et mean(x) est une fonction qui retourne la moyenne du vecteur x.

D'autre part, nous avons évalué la méthode d'égalisation des histogrammes (Finlayson *et al.*, 2005) décrite dans l'état de l'art (section 2.3.1). Sur PRID-2011, les courbes CMC des deux méthodes de prétraitements sont présentées dans la figure 5.14.

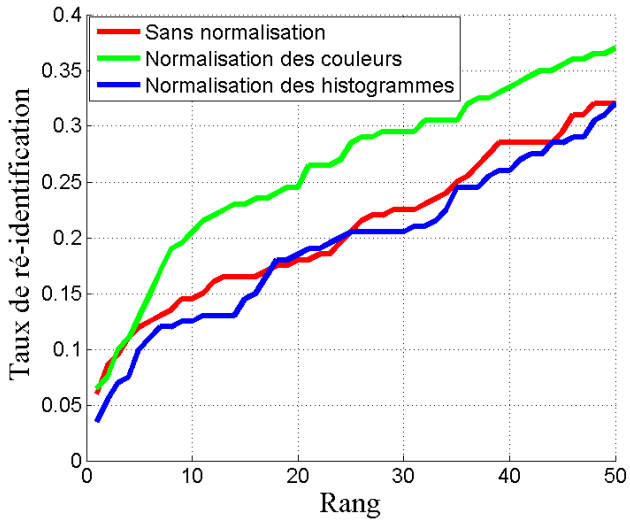


FIGURE 5.14 – Courbes CMC obtenus suite à l'application des prétraitements des couleurs.

La normalisation des couleurs permet d'améliorer le taux de ré-identification de 2%

au rang 1. Elle surpasse la performance des primitives sans prétraitement ou avec la normalisation des histogrammes. Par contre, elle reste très loin des performances des SURFs. Ceci dit, les méthodes simples de normalisation des couleurs restent incapables de remédier aux problèmes d'éclairages relatifs au scénario complexe. L'alternative dans ce cas est d'exploiter des méthodes plus complexes telles que les méthodes supervisées (section 2.3.1) qui apprennent les transformations dues à l'éclairage entre les deux caméras pour ré-identifier une personne.

5.3 Ré-identification par le mouvement

Nous nous intéressons dans cette partie à la description du mouvement de la personne par les PIs spatiotemporels. Deux types de PIs sont sélectionnés de l'état de l'art : STIP et Cuboïdes. Les étapes d'un système de ré-identification par le mouvement sont similaires à celles du système de ré-identification par l'apparence. L'unique différence correspond aux PIs en entrée du système. Ce système a été évalué uniquement sur la base de données PRID-2011 car c'est la seule base de données publique, composée par des séquences vidéo et filmée dans des conditions complexes.

5.3.1 Extraction des primitives

Le principe de détection et description des STIPs et Cuboïdes est décrit dans la section 3.4.2.2. Les caractéristiques de ces deux types de PIs sont illustrées par la table 5.10. Quant à la description des PIs, nous n'utilisons pas leurs descripteurs standards mais plutôt nous exploitons d'autres descripteurs à savoir HOG (Laptev *et al.*, 2008) et SURF.

TABLE 5.10 – Caractéristiques des STIPs et Cuboïdes.

Primitives	STIP	Cuboïdes
Détection	Harris 3D	Filtre de Gabor
Description	HOG ou SURF	
Nombre moyen de PIs par séquence test	321	1030
Nombre moyen de PIs par séquence référence	246	615

5.3.2 Expériences et résultats

L'utilisation des PIs spatiotemporels n'est possible que sur des séquences vidéo. La base de données PRID-2011 est composée de séquences vidéo de 69 images en moyenne par séquence en référence et 117 en test.

Comme la détection de ces PIs utilise une échelle temporelle, les expériences ont montré qu'une séquence vidéo doit avoir assez d'images pour pouvoir détecter des PIs spatiotemporels. Ce minimum d'images dépend effectivement de la complexité du contenu vidéo et il est empiriquement autour de 20 images.

Quant à PRID-2011, pour les 200 personnes qu'on essaye de ré-identifier, 17 séquences ont un nombre insuffisant d'images, en référence ou en test, pour détecter des PIs. Dans une première expérience, pour montrer l'apport de la description par le mouvement, nous

nous n'intéressons pas à ces 17 séquences. Ensuite, dans une deuxième expérience, pour se comparer avec l'état de l'art, nous tenons compte de toute la base de données.

5.3.2.1 Résultats des STIPs

La table 5.11 montre la performance des deux descripteurs HOG et SURFs. Elle montre que HOG est plus performant que SURF.

TABLE 5.11 – Comparaison des performances des descripteurs des STIPs.

Descripteur	**CCR (%)**
STIP : HOG	19.67
STIP : SURF	15.30

Dans la suite, les STIPs sont décrits par HOG. La figure 5.15 présente les courbes CMC obtenues par les deux méthodes d'appariement des STIPs : RP ($\lambda = 0.005$) et 1-NN. La table 5.12 présente les CCRs des deux méthodes d'appariements. Les résultats obtenus prouvent l'apport de la RP par rapport au 1-NN. En effet, la RP est légèrement plus performante que le 1-NN au rang 1 (table 5.12). Au-delà du rang 1, la RP surpasse clairement le 1-NN pour tous les rangs (figure 5.15).

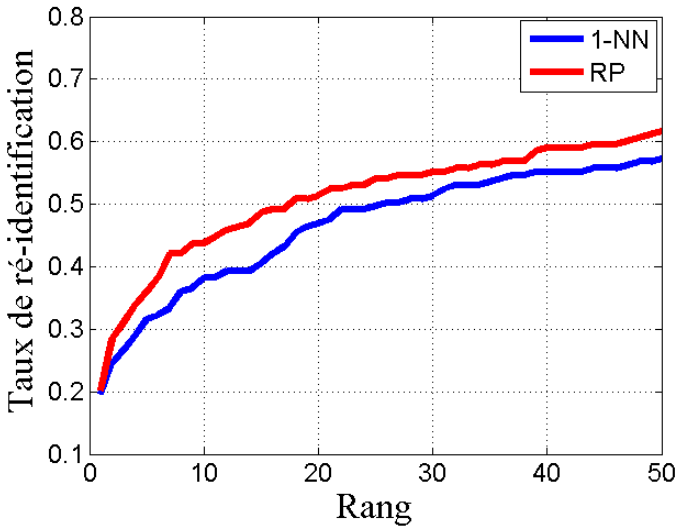


FIGURE 5.15 – Courbe CMC obtenues par les STIPs (sans les 17 séquences).

TABLE 5.12 – CCR obtenus par les STIPs (sans les 17 séquences).

Approches	**CCR (%)**
STIP : 1-NN	19.67
STIP : RP	20.22

5.3.2.2 Résultats des Cuboïdes

De même, nous avons évalué les Cuboïdes avec les deux descripteurs HOG et SURF, et les résultats obtenus ont montré que les SURFs sont plus performants. Dans la suite,

les Cuboïdes sont décrits par des SURFs. Ensuite, nous avons évalué les Cuboïdes comme entrée du système au lieu des STIPs. La figure 5.16 présente les courbes CMC des deux méthodes d'appariement alors que la table 5.13 visualise leurs CCRs. Pour une valeur optimale de λ ($\lambda = 0.01$), la figure 5.16 prouve l'efficacité de la RP par rapport au 1-NN pour tous les rangs avec une amélioration de 2.18% au rang 1 (table 5.13).

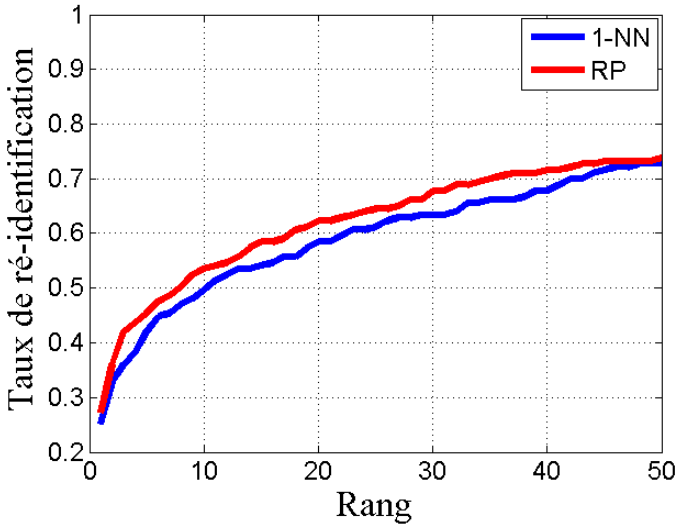


FIGURE 5.16 – Courbe CMC obtenues par les Cuboïdes (sans les 17 séquences).

TABLE 5.13 – CCR obtenus par les Cuboïdes (sans les 17 séquences).

Approches	**CCR (%)**
Cuboïdes : 1-NN	25.14
Cuboïdes : RP	27.32

5.3.2.3 Analyse des résultats

Pour les deux types de PIs, STIPs et Cuboïdes, l'appariement fondé sur la RP est plus performant que celui fondé sur le 1-NN. Cela prouve que le comportement de la RP par rapport au 1-NN est identique à celui dans le cas des SURFs. Par contre, l'amélioration réalisée par la RP est moins significative par rapport à celle réalisée dans le cas des SURFs. Elle peut être justifiée par la quantité moyenne des SURFs, par séquence référence, comparée aux STIPs et Cuboïdes (table 5.14). En effet, quand le nombre de PIs par séquence est grand, la chance de trouver une bonne approximation d'un PI test est élevée.

TABLE 5.14 – Nombre moyen de PIs par séquence.

PIs	SURF	Cuboïdes	STIP
Nombre moyen par séquence référence	1130	615	256

5.3.2.4 Comparaison des performances des PIs

Nous comparons maintenant les performances obtenues par les PIs décrivant l'apparence et PIs décrivant le mouvement. Dans une première expérience, nous nous n'intéressons pas aux 17 séquences non encodables par le mouvement, et dans une deuxième expérience, nous comparons nos approches à celles de l'état de l'art sur toute la base de données.

Concernant la première expérience, la figure 5.17 présente les courbes CMC des SURFs, STIPs et Cuboïdes. La table 5.15 présente les CCRs obtenus par ces trois types de PIs.

TABLE 5.15 – CCR obtenus par les PIs (sans les 17 séquences).

Approches	**CCR (%)**
SURF	29.51
Cuboïdes	27.32
STIP	20.22

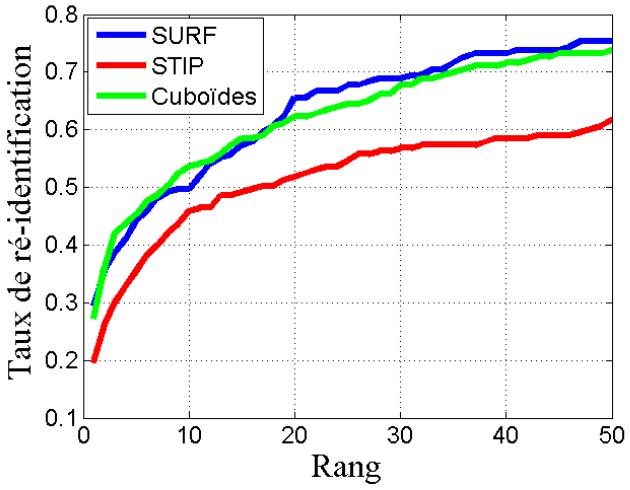


FIGURE 5.17 – Courbes CMC obtenues par les PIs sur PRID-2011 (sans les 17 séquences).

Dans une deuxième expérience, nous présentons dans la figure 5.18 et la table 5.16 les résultats obtenus par les différents PIs testés avec la RP ainsi que pour l'unique travail de l'état de l'art évalué sur PRID-2011 (Hirzer *et al.*, 2011).

TABLE 5.16 – Comparaison des CCR obtenus par les PIs avec l'état de l'art.

Approches	CCR (%)
SURF	27.00
STIP	18.00
Cuboïdes	25.00
(Hirzer *et al.*, 2011)	19.18

La figure 5.18 montre que les SURFs et les Cuboïdes ont des performances proches alors que chacun décrit un aspect spécifique de la vidéo à savoir l'apparence et le mouvement. Nous nous intéressons dans la section suivante à profiter de l'aspect complémentaire entre le mouvement et l'apparence et fusionner les performances de ces deux primitives.

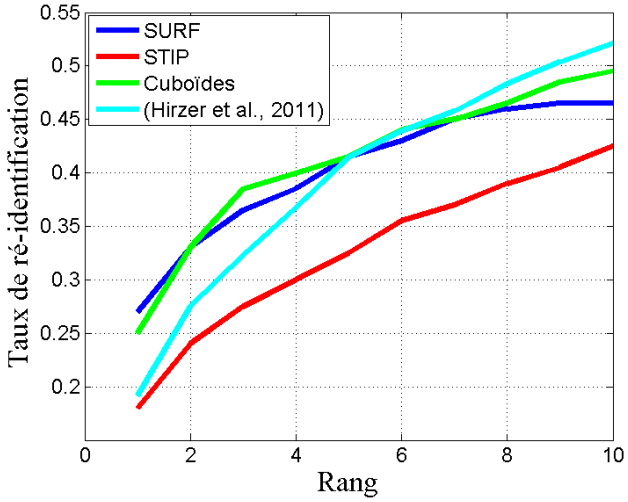


FIGURE 5.18 – Comparaison des courbes CMC obtenus par les PIs avec l'état de l'art.

5.4 Fusion d'apparence-mouvement

La fusion de deux approches peut se produire au niveau de l'une des trois étapes importantes d'un système biométrique générique : 1) extraction des primitives, 2) correspondance des primitives et 3) décision. En effet, trois niveaux de fusion sont populaires dans l'état de l'art : 1) fusion au niveau de l'extraction des primitives : les différentes primitives des approches à fusionner sont concaténées dans le même vecteur, 2) fusion au niveau de la correspondance des primitives : elle est aussi appelée fusion de scores. Dans ce cas, les scores de correspondance de différentes approches sont transformés en un seul score et 3) fusion au niveau de la décision : chaque approche à fusionner génère une décision. La décision finale consiste à combiner les différentes décisions pour générer une seule décision telle que la règle des votes majoritaires.

5.4.1 Schéma de la fusion

Notre schéma de fusion est de type fusion de décisions. En effet, chaque PI test, SURF et Cuboïdes, est apparié par la RP en utilisant les PIs référence de même type. Ensuite, deux vecteurs de votes majoritaires sont générés comme expliqué dans la section 4.3.1.1.5. Chaque vecteur de vote concerne un type de PIs. L'idée principale de la fusion consiste à calculer la somme pondérée des deux vecteurs de votes.

Soit une séquence vidéo test d'identité inconnue, décrite par n SURFs et m Cuboïdes, nous voulons reconnaître son identité comme une parmi N identités référence. Soient v et w les deux vecteurs de votes majoritaires respectivement des SURFs et des Cuboïdes. La vecteur de votes z correspondant à la fusion s'écrit comme suit :

$$z = \alpha v + (1 - \alpha)w$$

où α est un paramètre de pondération, choisi empiriquement. Finalement, la séquence test est ré-identifiée comme la séquence référence ayant le maximum de votes. La figure 5.19 montre l'organigramme de l'approche proposée.

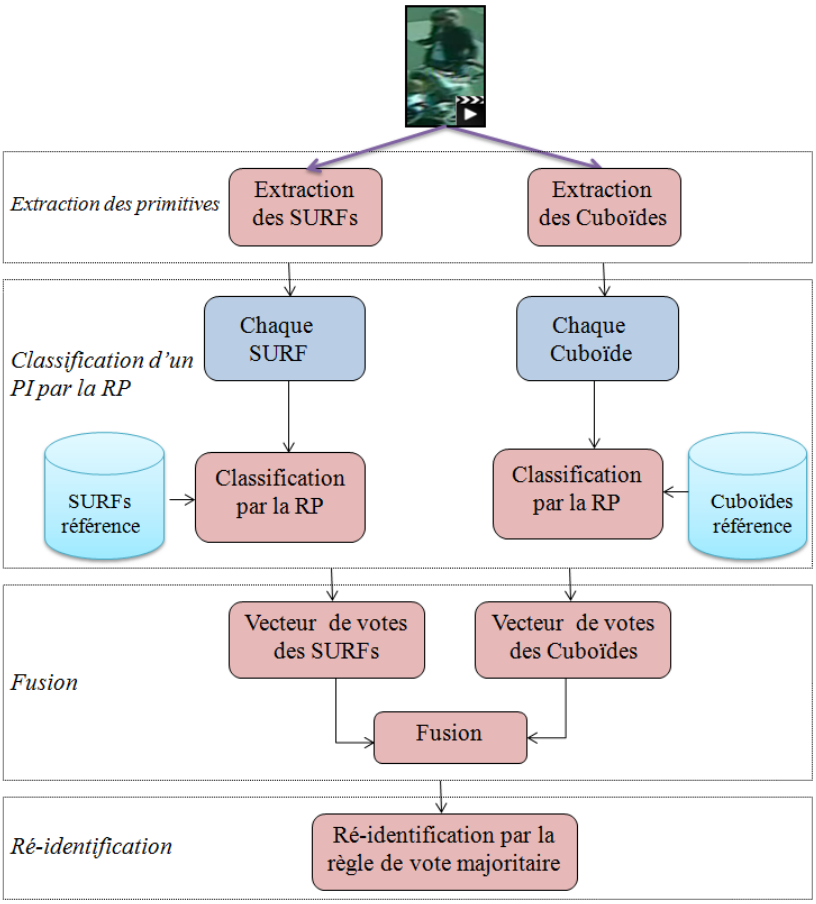


FIGURE 5.19 – Organigrammes de l'approche de la fusion.

Nous avons vu précédemment (table 5.14), que le nombre moyen des SURFs par séquence est significativement supérieur à celui des Cuboïdes (1130 SURFS vs 615 Cuboïdes). Ceci dit, chaque séquence est décrite en moyenne par 65% de SURFs et 35% de Cuboïdes. Pour en tenir compte, α a été empiriquement fixé à une valeur proche de 0.35 et comparé à $\alpha = 0.5$.

5.4.2 Résultats de la fusion

Dans une première expérience, les 17 séquences où on ne détecte pas de Cuboïdes ne sont pas considérées pendant les tests. Nous avons évalué deux valeurs de α : 0.3 et 0.5. Au rang 1, le système de fusion atteint un CCR = 33.88 % pour $\alpha = 0.3$ et un CCR = 31.14 % pour $\alpha = 0.5$. Ceci montre l'importance des poids quand les deux approches à fusionner génèrent des nombres de PIs différents. Dans le reste des expériences, la valeur de α est fixée à 0.3. La figure 5.20 présente les courbes CMC obtenues par SURF et Cuboïdes ainsi que le résultat de la fusion ; la table 5.17 montre leurs CCRs.

Ces résultats ont montré que la fusion des deux descripteurs a permis d'améliorer la performance du système de 4.37% par rapport aux SURFs seuls et 6.56% par rapport aux Cuboïdes seuls. Cette amélioration est significative vu la grande taille de la base de données. En outre, elle montre expérimentalement la complémentarité entre l'aspect statique de l'apparence d'une personne d'une part et l'aspect dynamique de son type de mouvement d'autre part.

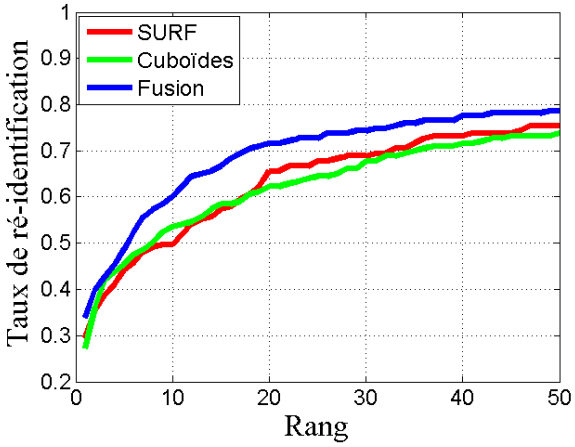


FIGURE 5.20 – Courbes CMC obtenues par SURF, Cuboïdes et la fusion (sans les 17 séquences).

TABLE 5.17 – CCR obtenus par les SURFs, Cuboïdes et la fusion (sans les 17 séquences).

Approches	**CCR (%)**
SURF-RP	29.51
Cuboïdes-RP	27.32
Fusion	33.88

Comparaison avec l'état de l'art

Dans une deuxième expérience, nous tenons compte de toute la base de données PRID-2011. Nous avons vu précédemment (figure 5.9) que l'approche (Hirzer *et al.*, 2011) devient plus performante que les SURFs à partir du rang 6. Après la fusion des SURFs et Cuboïdes, notre approche devient plus performante que (Hirzer *et al.*, 2011) pour tous les rangs (de rang 1 au rang 10) (figure 5.21). La table 5.18 montre les CCRs de ces approches.

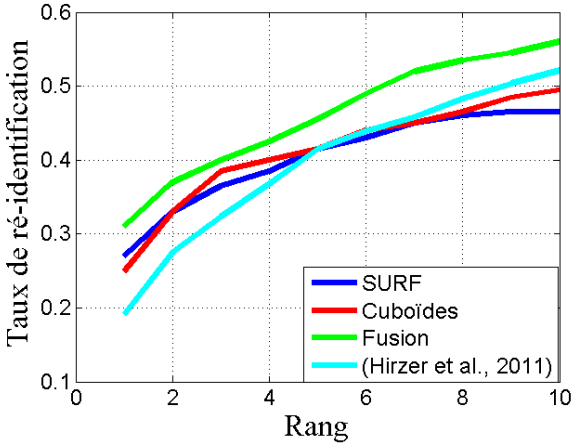


FIGURE 5.21 – Courbes CMC de la fusion comparées à l'état de l'art.

TABLE 5.18 – CCR de la fusion comparés à l'état de l'art.

Approches	**CCR (%)**
SURF-RP	27.00
Cuboïdes-RP	25.00
Fusion	31.00
(Hirzer *et al.*, 2011)	19.18

5.5 Analyse des erreurs

La tâche de ré-identification paraît difficile dans les conditions complexes car le meilleur système proposé atteint un taux de ré-identification égale à 31% au rang 1.

Parmi les sources d'erreur, on cite la dimension de la région descriptive des PIs, la luminosité, la direction de la marche, la qualité de l'étape de détection des personnes et les apparences semblables. À cause de l'absence d'une base d'évaluation, nous avons choisi 25 personnes non ré-identifiées pour identifier les sources d'erreurs.

5.5.1 Région descriptive

Dans la version originale du SURF, la région utilisée pour décrire un SURF est de taille $20^*\sigma$ où σ est l'échelle de détection du SURF. En effet, SURF est utilisé initialement pour la reconnaissance d'objets dans des images de haute résolution où $20^*\sigma$ est une valeur très petite par rapport à ses dimensions (figure 5.22). Dans notre cas, en traçant la distribution des échelles des SURFs référence de PRID-2011, environ 50% des SURFs ont une échelle de détection proche de 2 (figure 5.23). Comme les images de PRID-2011 sont de faible résolution (128x64 pixels), la dimension de la région descriptive devient très grande par rapport à celle de l'image (figure 5.24). Par conséquent, les dimensions de cette région descriptive devraient être adaptées à celles de l'image.

FIGURE 5.22 – Image de dimension : 514 x 627. Régions descriptives bleues d'arrête égale à $20^*\sigma$.

Visuellement, on ne détecte pas de variation d'échelle entre les images référence et test de PRID-2011. Pour cela, nous proposons d'utiliser une échelle égale à 1 pour tous les

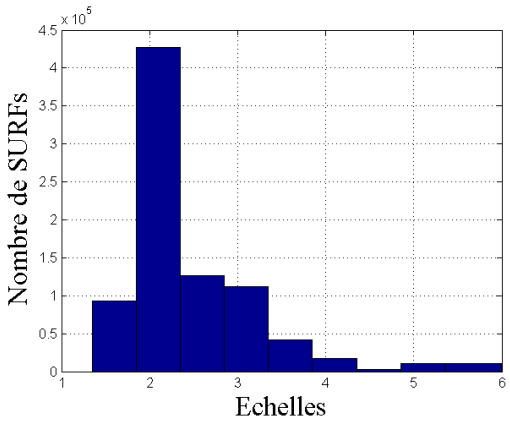
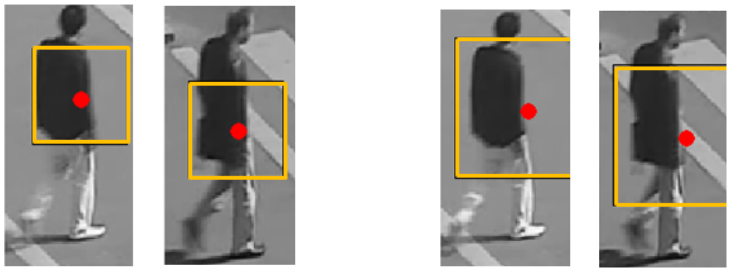


FIGURE 5.23 – Distribution des σ des SURFs référence de PRID-2011.



(a) (b)

FIGURE 5.24 – Exemple de régions descriptives en jaunes : (a) $\sigma = 2$, (b) $\sigma = 3$ (les points rouges sont des SURFs).

SURFs. Par conséquent, nous réduisons les dimensions de la région descriptive à 20x20 pixels ; dimensions adéquates à celles des images. Cette adaptation d'échelle permet d'améliorer la performance des SURFs de 3% en taux de ré-identification au rang 1 (figure 5.25). Pour les Cuboïdes, l'adaptation d'échelle n'améliore pas la performance du système (figure 5.26).

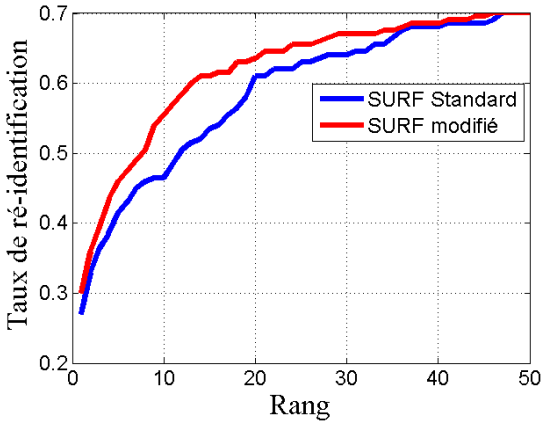


FIGURE 5.25 – Adaptation des dimensions de la région descriptive : résultats des SURFs.

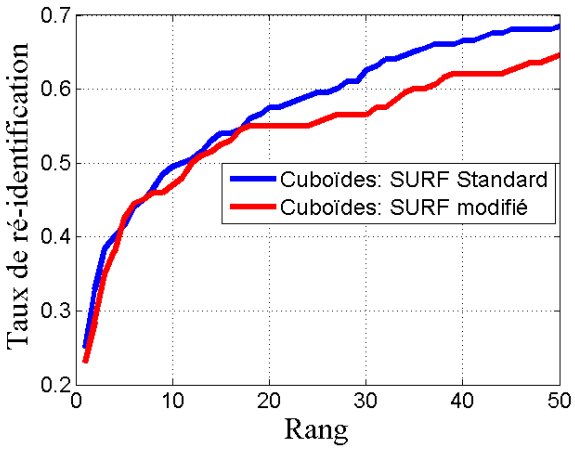
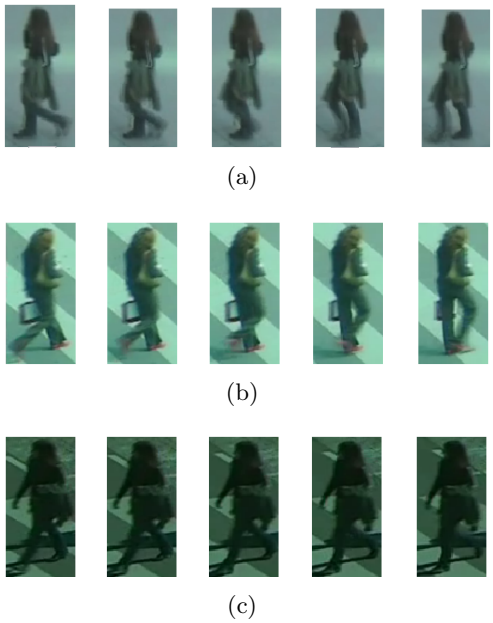


FIGURE 5.26 – Adaptation des dimensions de la région descriptive : résultats des Cuboïdes.

5.5.2 Luminosité

Comme vu précédemment, les conditions de luminosité des deux caméras sont très différentes. Ainsi, deux images de la même personne peuvent avoir deux contrastes très différents. Cela peut influer négativement sur la ré-identification des personnes. Dans la suite, nous montrons un exemple où la variation de l'éclairage entre la séquence de référence (figure 5.27-a) et la séquence de test (figure 5.27-c) est énorme. À cause de cette variation d'éclairage, la séquence test est reconnue comme correspondant à une autre personne (5.27-b).



(a)

(b)

(c)

FIGURE 5.27 – Exemple de variation de l'éclairage : (a) séquence test, (b) séquence reconnue, (c) séquence référence.

Pour remédier à cette source d'erreurs, nous proposons d'appliquer un prétraitement d'égalisation locale des histogrammes (Zuiderveld, 1994) (figure 5.28). Cette égalisation des histogrammes permet d'améliorer la performance des SURFs de 2% au rang 1 (figure

5.29) alors qu'elle n'apporte pas d'améliorations sur les Cuboïdes (figure 5.30). Vue la complexité des conditions de luminosité, des approches supervisées de normalisation de l'éclairage peuvent être abordées.

(a) (b)

FIGURE 5.28 – (a) Sans égalisation des histogrammes. (b) Avec égalisation des histogrammes.

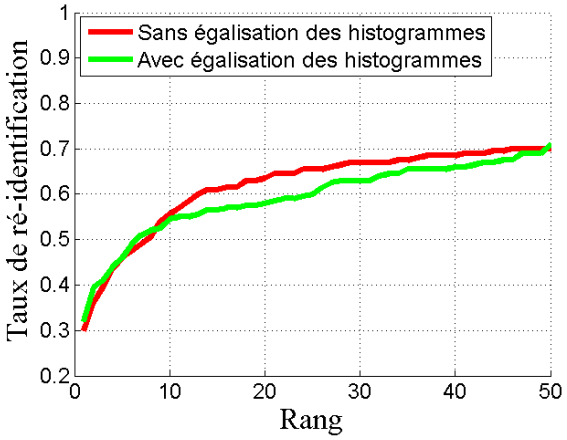


FIGURE 5.29 – Egalisation des histogrammes : résultats des SURFs.

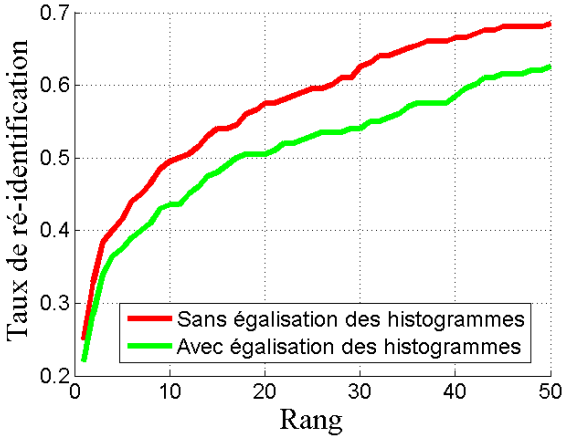


FIGURE 5.30 – Egalisation des histogrammes : résultats des Cuboïdes.

5.5.3 Direction de la marche

Certaines personnes peuvent être non ré-identifiées à cause du changement de la direction de marche entre la référence et le test. À cause de ce changement, deux séquences de

la même personne peuvent avoir des textures très variées Nous présentons ci-dessous un exemple d'une personne test (figure 5.31-a) qui est mal reconnue (figure 5.31-b) car elle marche en référence dans une direction différente à celle en test (figure 5.31-c).

(a)

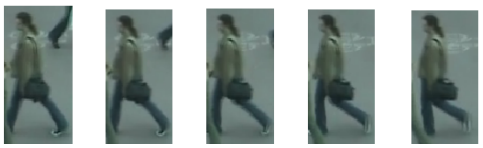
(b)

(c)

FIGURE 5.31 – Exemple de changement de direction de marche : (a) séquence test, (b) séquence reconnue, (c) séquence référence.

5.5.4 Détection de la personne

Dans un scénario complexe, les personnes peuvent passer en foule. Par conséquent, des mauvaises détection de la personne peuvent se produire et influer sur la performance du



(a)

(b)

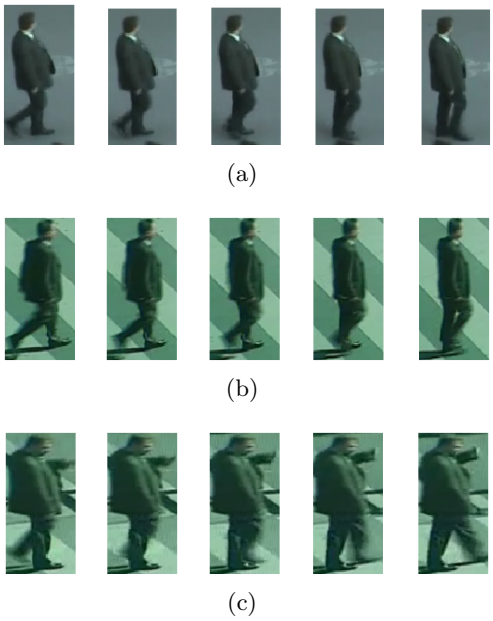
(c)

FIGURE 5.32 – Exemple de mauvaise détection de la personne : (a) séquence test, (b) séquence reconnue, (c) séquence référence.

système de ré-identification. La figure 5.32 montre un exemple d'une personne test (figure 5.32-a) qui est mal détectée en référence (Elle est détectée avec d'autres personnes). (figure 5.32-c). À cause de cette mauvaise détection, la personne est mal reconnue (figure 5.32-b).

5.5.5 Apparences semblables

Etant donné que la base de données est composée de plusieurs centaines de personnes, il est probable de trouver des personnes ayant des apparences semblables. La figure 5.33 montre un exemple où la personne test (figure 5.33-a) possède de textures des vêtements proches de celles de la personne reconnue (figure 5.33-b).



(a) (b) (c)

FIGURE 5.33 – Exemple d'apparences semblables : (a) séquence test, (b) séquence reconnue, (c) séquence référence.

5.6 Conclusion

Dans ce chapitre, nous avons proposé un système de ré-identification adéquat à un scénario complexe. Il s'agit d'adapter les systèmes de l'état de l'art, fondé sur l'apparence, aux conditions réelles de la ré-identification en optimisant ses étapes les plus critiques. Deux améliorations majeures ont été réalisées : 1) amélioration de la description et 2) amélioration de l'appariement des primitives (PIs).

Pour l'appariement des PIs, nous avons proposé une nouvelle méthode de mise en correspondance des PIs via la représentation parcimonieuse qui consiste à représenter chaque PI test comme la combinaison linéaire la plus parcimonieuse des PIs référence. Pour plus d'efficacité, un dictionnaire dynamique est sélectionné et composé d'un nombre prédéfini de plus proches PIs référence obtenus par la recherche dans un arbre-KD. Cette méthode d'appariement, en utilisant les SURFs, a été évaluée sur deux bases de données publiques. Les résultats ont montré une amélioration de 4% sur CAVIAR4REID et 4.5% sur PRID-2011 par rapport au 1-NN. Ces résultats prouvent la richesse en information relative de la

RP quand les PIs sont bruités et ambigus, ce qui est inhérent aux séquences vidéos réelles.

Quant à la deuxième contribution, nous nous sommes intéressés à l'adaptation de la description de la personne au scénario complexe à partir des séquences vidéo. De ce fait, nous avons exploité la complémentarité entre la description mouvement de la personne et son apparence. Plusieurs PIs spatiotemporels ont été étudiés pour décrire le mouvement et les Cuboïdes étaient les plus performants. En évaluation, la performance de description mouvement atteint 25% de CCR ; alors que celle de l'apparence atteint un CCR de 27%. Quant à la fusion des deux descriptions, elle a permis au système d'atteindre un CCR de 31% qui se compare très favorablement avec l'état de l'art.

Dans l'étape de l'analyse des erreurs, nous avons identifié un ensemble de sources d'erreurs. Les traitements proposés pour remédier aux problèmes de la dimension de la région descriptive des PIs et la luminosité ont montré l'aptitude d'améliorer la performance du système de ré-identification.

Conclusion et Perspectives

Cette thèse s'inscrit dans le contexte de la vidéosurveillance intelligente, et s'intéresse à la ré-identification des personnes dans un réseau de deux caméras à champs de vue disjoints. Ce problème est particulièrement difficile car l'apparence de la personne peut changer entre deux prises de vue de manière significative à cause de différents facteurs liés essentiellement aux conditions du milieu de capture et aux conditions de passage de la personne dans le champ de vue de la caméra. Selon ces conditions d'acquisition des données, la ré-identification peut correspondre à plusieurs scénarios de degré de complexité différents. Nous avons distingué et étudié deux scénarios : simple et complexe.

Dans le scénario simple, des contraintes sont imposées sur le passage de la personne dans le champ de vue de la caméra et sur les conditions du milieu d'enregistrement. D'une part, la personne doit passer seule dans le champ de vue de la caméra, dans une seule direction et sans porter aucun objet. D'autre part, la base de données est composée d'un ensemble réduit de personnes filmées, en référence et en test, par une seule caméra dans des conditions d'éclairage stables.

Dans ce scénario, nous avons étudié la faisabilité de deux approches de ré-identification : une approche fondée sur un modèle biométrique et une approche fondée sur l'apparence. Ces approches ont été évaluées en fonction de la différence angulaire entre les angles de vue de référence et les angles de vue de test. Pour une grande différence angulaire, les expériences réalisées sur la base de données CASIA-A de 20 personnes, ont montré que les approches fondées sur l'apparence sont les plus prometteuses.

Deux descriptions d'apparence ont été évaluées : les PIs spatiaux SURF et les histogrammes de couleurs dans l'espace RGB. Le choix de ces descriptions a été encouragé par les contraintes imposées par le scénario simple.

L'utilisation des SURFs est motivée par le fait que d'une part, ces derniers sont robustes aux changements des angles de vue, et d'autre part la bonne répétabilité des SURFs dans la vidéo augmente la chance de trouver un bon appariement entre un SURF référence et un SURF test. L'appariement local des SURFs est plus robuste qu'un appariement global. Ceci est justifié par le fait que localement chaque PI de la requête participe par un vote de manière indépendante alors que globalement les BoFs des PIs sont très instables en raison de la provenance variée des PIs (personne, fond, objet, etc.). Notre conclusion est que dans un scénario simple, la ré-identification via un appariement local des points d'intérêt s'avère plus robuste pour toutes les différences angulaires.

Le choix des primitives de couleurs pour décrire l'apparence est motivé essentiellement par la stabilité du contraste, l'absence d'occultation et le nombre réduit de personnes de la base de données. Les expériences réalisées ont montré que les conditions de ce scénario réduisent fortement le changement de couleurs d'une prise à l'autre. Ces primitives se sont également montrées robustes pour distinguer les personnes de la base de données réduite pour toutes les différences angulaires.

Les résultats de la ré-identification fondée sur la modélisation de la marche par un HMM sont moins bons que ceux obtenus par les approches fondées sur l'apparence. De plus, les performances du système se dégradent rapidement pour une petite différence angulaire. Ceci dit, le HMM a un potentiel réel quand les angles de vue de référence et de test sont similaires et quand les personnes marchent suffisamment dans le champ de vue de la caméra de manière à générer un nombre suffisant de périodes pour apprendre correctement les paramètres du HMM. La dégradation des performances du système est justifiée par le fait que la deuxième condition n'est pas vérifiée dans la base de données que nous avons considérée.

Le deuxième scénario étudié, scénario complexe, n'impose aucune contrainte sur le passage de la personne dans le champ de vue de la caméra ou sur les conditions d'enregistrement des données. Plus précisément, les données de référence et de test sont filmées par deux caméras installées dans deux endroits différents où les personnes marchent librement. Pour prendre en compte la complexité de ce scénario, deux contributions ont été proposées concernant la description de la personne et l'appariement des descripteurs.

La première contribution, liée à la description de la personne, consiste à exploiter la complémentarité entre les descriptions de l'apparence de la personne et de son style de mouvement. De ce fait, nous avons proposé une description de la personne qui tient compte de l'apparence et du mouvement en même temps. Nous avons décrit l'apparence (respectivement le mouvement) par des points d'intérêt spatiaux (respectivement points d'intérêt spatiotemporels). Bien que la performance du descripteur de mouvement soit légèrement moins bonne que celle du descripteur de l'apparence, la fusion entre eux a permis d'améliorer la performance générale. Les expériences, réalisées sur la grande base de données PRID-2011 de 749 personnes en référence et 200 personnes en test, ont montré que la fusion des deux descriptions a permis au système de ré-identification d'atteindre un CCR de 31% qui se compare favorablement avec l'état de l'art.

Concernant la deuxième contribution, liée à l'appariement des descripteurs, nous avons proposé d'utiliser la représentation parcimonieuse comme méthode d'appariement local entre les points d'intérêt spatiaux liés à l'apparence d'une part et les points d'intérêt spatiotemporels liés au mouvement d'autre part. Notre manière d'exploiter la représentation parcimonieuse est différente à celle utilisée dans l'état de l'art. Nous avons proposé d'utiliser une représentation parcimonieuse locale (représentation pour chaque point d'intérêt). De plus, nous avons proposé d'utiliser un dictionnaire dynamique et réduit pour chaque représentation parcimonieuse. L'évaluation de notre méthode d'appariement, sur deux bases de données publiques CAVIAR4REID et PRID-2011, a montré sa robustesse d'une part par rapport à la méthode fondée sur le 1-NN, et d'autre part par rapport aux approches populaires de l'état de l'art. En utilisant le descripteur SURF, les expériences ont montré une amélioration de 4% sur CAVIAR4REID et 4.5% sur PRID-2011 par rapport au 1-NN. Ces résultats prouvent la richesse en information relative de la représentation parcimonieuse quand les points d'intérêt sont bruités et ambigus, ce qui est inhérent aux séquences vidéo réelles.

Les primitives de couleurs n'ont pas donné de bons résultats quand les conditions d'acquisition ne sont pas contrôlées. Les méthodes simples de description de la couleur restent incapables de remédier au problème de variabilité d'éclairage relatif au scénario complexe. Ces méthodes nécessitent a priori l'utilisation des informations liées aux variations colorimétriques entre les deux caméras.

A l'issue de ce travail, plusieurs voies restent à explorer et plusieurs perspectives s'ouvrent sur les diverses problématiques traitées. Les principales perspectives concernant directement nos travaux incluent :

➢ Dans les deux scénarios étudiés, la méthode proposée pour filtrer les PIs paraît insuffisante car on utilise une seule mesure pour accepter/rejeter un appariement. Il s'agit de la distance euclidienne entre un PI test et le PI référence le plus proche dans le scénario simple, et du minimum des erreurs de reconstruction associées aux différentes identités référence dans le scénario complexe. Nous proposons d'explorer des mesures multidimensionnelles plutôt qu'une mesure pour filtrer un appariement. Dans le scénario simple, nous proposons d'exploiter par exemple les distances d'un PI test par rapport à un ensemble de PIs référence plutôt qu'un seul PI référence. Ces PIs référence peuvent être par exemple les PIs les plus proches par rapport à chaque identité référence. Dans le scénario complexe, la mesure multidimensionnelle peut correspondre aux erreurs de reconstruction associées aux identités références. Les mesures multidimensionnelles sont plus riches en information et devraient donc être mieux exploitables pour le filtrage des PIs.

➢ La représentation parcimonieuse dépend des éléments du dictionnaire. Nous avons uniquement étudié le cas où le dictionnaire est composé d'un nombre fixe de plus proches PIs référence. Nous proposons d'étudier d'autres méthodes de construction du dictionnaire qui peuvent par exemple estimer la taille du dictionnaire, fixée dans nos expériences. Si on arrive à estimer des tailles du dictionnaire plus réduites, on gagne sur la durée de la phase de ré-identification.

➢ Pour affecter un PI test à une identité référence, nous proposons d'étudier d'autres exploitations de la représentation parcimonieuse. Dans ce travail, l'identité d'un PI est considérée comme celle de l'identité référence qui minimise son erreur de reconstruction. L'affectation d'un vote binaire (1 pour l'identité qui minimise l'erreur de reconstruction et 0 pour les autres) est une décision dure vue l'ambiguïté des PIs. Nous proposons d'affecter des votes continus entre 0 et 1, proportionnels aux erreurs des reconstructions, pour les différentes identités références. La ré-identification restera fondée sur la règle du vote majoritaire.

➢ Dans ce travail, nous avons proposé des approches non supervisées pour résoudre le problème de la ré-identification. Cependant, dans des conditions complexes, une approche supervisée exploitant une base d'apprentissage comprenant des paires de séquences vidéo référence et test associées, nous permettrait d'apprendre les paramètres de transformation des couleurs entre les deux caméras, d'apprendre un modèle de filtrage des appariements les moins fiables, ou encore d'estimer la taille du dictionnaire en fonction du PI.

A plus long terme :

➢ Optimiser les performances des primitives de couleurs dans le scénario complexe et étudier leur fusion avec les PIs. Dans le scénario complexe, la complexité des conditions d'éclairage empêche les primitives de couleurs de distinguer les personnes. Nous proposons d'explorer les méthodes supervisées qui établissent d'une part une liaison entre les plages des couleurs des deux caméras, et d'autre part une liaison entre les

transformations géométriques entre les deux caméras. Le fait d'utiliser un ensemble de paires d'images, prises par les deux caméras et mises en correspondance *a priori*, aide le système à ré-identifier une nouvelle personne filmée par les mêmes caméras.

➢ Travailler sur l'étape de détection des personnes dans le scénario complexe et intégrer sa performance dans la performance globale d'un système de ré-identification. Dans le scénario complexe, même un bon algorithme de détection des personnes ne peut pas éviter les mauvaises détections causées par exemple par les occultations ou la variation des poses. Nous proposons de filtrer les images où le niveau d'occultation est très élevé. Ces mauvaises détections ne peuvent que perturber les décisions faites par le système de ré-identification plutôt que les améliorer.

Liste des publications

Conférences internationales :

(2013) Mohamed Ibn Khedher, Mounim A. El-Yacoubi, Bernadette Dorizzi : **Multi-shot SURF-based person re-identification via sparse representation.** AVSS 2013 : 159-164

(2012) Mohamed Ibn Khedher, Mounim A. El-Yacoubi, Bernadette Dorizzi : **Probabilistic matching pair selection for SURF-based person re-identification.** BIOSIG 2012 : 1-6

(2012) Mohamed Ibn Khedher, Mounim A. El-Yacoubi, Bernadette Dorizzi : **Human Action Recognition using Continuous HMMs and HOG/HOF Silhouette Representation.** ICPRAM (2) 2012 : 503-508

Articles de journaux :

(2014) Mohamed Ibn Khedher, Mounim A. El-Yacoubi, Bernadette Dorizzi : **Multiple-shot Person Re-identification Based on Fusion of Appearance and Motion Local Features.** En cours de rédaction !

Bibliographie

Alonso, I. P., Llorca, D. F., Sotelo, M. Á., Bergasa, L. M., de Toro, P. R., Nuevo, J., Ocaña, M. et Garrido, M. Á. G. (2007). Combination of feature extraction methods for svm pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 8(2):292–307.

An, L., Kafai, M., Yang, S. et Bhanu, B. (2013). Reference-based person re-identification. *Dans Proceedings of the 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 244–249.

Atine, J.-C. (2004). People action recognition in image sequences using a 3d articulated object. *Dans Proceedings of the International Conference on Image Analysis and Recognition*, pages 769–777.

Bak, S., Corvee, E., Bremond, F. et Thonnat, M. (2010a). Person re-identification using haar-based and dcd-based signature. *Dans Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–8.

Bak, S., Corvee, E., Bremond, F. et Thonnat, M. (2010b). Person re-identification using spatial covariance regions of human body parts. *Dans Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 435–440.

Bauml, M. et Stiefelhagen, R. (2011). Evaluation of local features for person re-identification in image sequences. *Dans Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 291–296.

Bay, H., Tuytelaars, T. et Gool, L. V. (2006). Surf : Speeded up robust features. *Dans Proceedings of 9th European Conference on Computer Vision*, pages 404–417.

Bazzani, L., Cristani, M. et Murino, V. (2013). Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117(2):130–144.

Beis, J. S. et Lowe, D. G. (1997). Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. *Dans Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1000–1006.

Belongie, S., Malik, J. et Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.

Berdugo, G., Soceanu, O., Moshe, Y., Rudoy, D. et Dvir, I. (2010). Object reidentification in real world scenarios across multiple non-overlapping cameras. *Dans Proceedings of the 18th European Signal Processing Conference (EUSIPCO)*, pages 806–1810.

Bird, N., Masoud, O., Papanikolopoulos, N. et Isaacs, A. (2005). Detection of loitering individuals in public transportation areas. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):167–177.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.

Bobick, A. et Davis, J. (1996). An appearance-based representation of action. *Dans Proceedings of the 13th International Conference on Pattern Recognition*, volume 1, pages 307–312.

Bobick, A. F. et Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267.

Boulgouris, N., Hatzinakos, D. et Plataniotis, K. (2005). Gait recognition : a challenging signal processing technology for biometric identification. *IEEE Signal Processing Magazine*, 22(6):78–90.

Bradley Efron, Trevor Hastie, I. J. et Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–840.

Cai, Y., Huang, K. et Tan, T. (2008). Human appearance matching across multiple non-overlapping cameras. *Dans Proceedings of the 19th International Conference on Pattern Recognition*, pages 1–4.

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

Chen, C., Liang, J., Zhao, H., Hu, H. et Tian, J. (2009). Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognition Letters*, 30(11):977–984.

Cheng, D. S., Cristani, M., Stoppa, M., Bazzani, L. et Murino, V. (2011). Custom pictorial structures for re-identification. *Dans Procedings of the British Machine Vision Conference*, pages 68.1–68.11.

Collins, R. T., Gross, R. et Shi, J. (2002). Silhouette-based human identification from body shape and gait. *Dans Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 351–356.

Cunado, D., Nixon, M. S. et Carter, J. N. (2003). Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90:1–41.

Dadashi, F., Araabi, B. et Soltanian Zadeh, H. (2009). Gait recognition using wavelet packet silhouette representation and transductive support vector machines. *Dans Proceedings of the 2nd International Congress on Image and Signal Processing*, pages 1–5.

Dalal, N. et Triggs, B. (2005). Histograms of oriented gradients for human detection. *Dans Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893.

de Oliveira, I. O. et de Souza Pio, J. L. (2009). People reidentification in a camera network. *Dans Proceedings of the 9th IEEE International Conference on Dependable, Autonomic and Secure Computing*, pages 461–466.

Dikmen, M., Akbas, E., Huang, T. et Ahuja, N. (2010). Pedestrian recognition with a learned metric. *Dans Proceedings of the 10th Asian Conference on Computer Vision*, volume Part IV, pages 501–512.

Dollár, P., Rabaud, V., Cottrell, G. et Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. *Dans 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72.

Donoho, D., Elad, M. et Temlyakov, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18.

Du, Y., Ai, H. et Lao, S. (2012). Evaluation of color spaces for person re-identification. *Dans Proceedings of the 21st International Conference on Pattern Recognition*, pages 1371–1374.

Eisenbach, M., Kolarow, A., Schenk, K., Debes, K. et Gross, H. (2012). View invariant appearance-based person reidentification using fast online feature selection and score level fusion. *Dans Proceedings of the 9th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 184–190.

Ess, A., Leibe, B. et Van Gool, L. (2007). Depth and appearance for mobile scene analysis. *Dans Proceedings of the 11th IEEE International Conference on Computer Vision*, pages 1–8.

Farenzena, M., Bazzani, L., Perina, A., Murino, V. et Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. *Dans Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2360–2367.

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

Finlayson, G. D., Hordley, S. D., Schaefer, G. et Tian, G. Y. (2005). Illuminant and device invariant colour using histogram equalisation. *Pattern Recognition*, 38(2):179–190.

Florack, L., Ter Haar Romeny, B., Viergever, M. et Koenderink, J. (1996). The gaussian scale-space paradigm and the multiscale local jet. *International Journal of Computer Vision*, 18(1):61–75.

Forssén, P.-E. (2007). Maximally stable colour regions for recognition and matching. *Dans IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Friedman, J., Hastie, T., Höfling, H. et Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J. H., Bentley, J. L. et Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3(3):209–226.

Fuchs, J.-J. (1997). Une approche à l'estimation et à l'identification simultanées. *Dans Seizième colloque Groupe d'Etudes du Traitement du Signal et des Images (GRETSI)*, pages 1273 –1276.

Fuchs, J.-J. (1998). Detection and estimation of superimposed signals. *Dans Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1649–1652.

Fuchs, J.-J. (2005). Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608.

Gemmeke, J., Virtanen, T. et Hurmalainen, A. (2011). Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2067–2080.

Gheissari, N., Sebastian, T. B. et Hartley, R. (2006). Person reidentification using spatiotemporal appearance. *Dans Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1528–1535.

Gilbert, A. et Bowden, R. (2006). Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. *Dans Proceedings of the 9th European Conference on Computer Vision*, volume Part II, pages 125–136.

Gong, M., Xu, Y., Yang, X. et Zhang, W. (2011). Gait identification by sparse representation. *Dans Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, volume 3, pages 1719–1723.

Grauman, K. et Darrell, T. (2005). The pyramid match kernel : Discriminative classification with sets of image features. *Dans Proceedings of the 10th IEEE International Conference on Computer Vision*, volume 2, pages 1458–1465.

Gray, D. et Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. *Dans Proceedings of the 10th European Conference on Computer Vision*, volume Part I, pages 262–275.

Hamdoun, O., Moutarde, F., Stanciulescu, B. et Steux, B. (2008). Person reidentification in multi-camera system by signature based on interest point descriptors collected on short video sequences. *Dans Proceedings of the 2nd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, pages 1–6.

Han, J. et Bhanu, B. (2006). Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322.

Harris, C. et Stephens, M. (1988). A combined corner and edge detector. *Dans Proceedings of the 4th Alvey Vision Conference*, pages 147–151.

Haruyuki, I., Yasushi, M. et Yasushi, Y. (2012). Group context-aware person identification in video sequences. *IPSJ Transactions on Computer Vision and Applications*, 4:87–99.

He, Q. et Debrunner, C. (2000). Individual recognition from periodic activity using hidden markov models. *Dans Proceedings of the Workshop on Human Motion (HUMO)*, pages 47–52.

Hirzer, M., Beleznai, C., Roth, P. M. et Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. *Dans Proceedings of the 17th Scandinavian conference on Image analysis*, pages 91–102.

Hirzer, M., Roth, P. et Bischof, H. (2012). Person re-identification by efficient impostor-based metric learning. *Dans Proceedings of the 9th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 203–208.

Huang, C.-H., Wu, Y.-T. et Shih, M.-Y. (2008). Unsupervised pedestrian re-identification for loitering detection. *Dans Proceedings of the 3rd Pacific Rim Symposium on Advances in Image and Video Technology*, pages 771–783.

Ijiri, Y., Lao, S., Han, T. X. et Murase, H. (2012). Human re-identification through distance metric learning based on jensen-shannon kernel. *Dans Proceedings of the International Conference on Computer Vision Theory and Applications*, volume 1, pages 603–612.

İkizler, N. et Duygulu, P. (2007). Human action recognition using distribution of oriented rectangular patches. *Dans Proceedings of the 2nd conference on Human motion : understanding, modeling, capture and animation*, pages 271–284.

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, O., Shafique, K. et Shah, M. (2005). Appearance modeling for tracking in multiple non-overlapping cameras. *Dans Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 26–33.

Joachims, T. (1998). Text categorization with support vector machines : Learning with many relevant features. *Dans Proceedings of the European Conference on Machine Learning*, pages 137–142.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211.

Jungling, K. et Arens, M. (2010). Local feature based person reidentification in infrared image sequences. *Dans Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 448–455.

Jungling, K. et Arens, M. (2011). View-invariant person re-identification with an implicit shape model. *Dans Proceedings of the 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 197–202.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

Kläser, A., Marszałek, M. et Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. *Dans Proceedings of the British Machine Vision Conference*, pages 995–1004.

Kviatkovsky, I., Adam, A. et Rivlin, E. (2013). Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1622–1634.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123.

Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B., Rennes, I., Grenoble, I. I. et Ljk, L. (2008). Learning realistic human actions from movies. *Dans Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Lindeberg, T. (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30:79–116.

Liu, X., Song, M., Tao, D., Zhou, X., Chen, C. et Bu, J. (2014). Semi-supervised coupled dictionary learning for person re-identification. *Dans Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Lowe, D. G. (2001). Local feature view clustering for 3d object recognition. *Dans Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 682–688.

Lucas, B. D. et Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *Dans Proceedings of the 7th international joint conference on Artificial intelligence*, volume 2, pages 674–679.

Madden, C., Cheng, E. D. et Piccardi, M. (2007). Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Machine Vision and Applications*, 18:233–247.

Mairal, J. (2009). Spams : a sparse modeling software, v2.4.

Mallat, S. et Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415.

Martin, A. (2010). *Représentations parcimonieuses adaptées à la compression d'images*. Thèse de doctorat, Université Rennes 1, France.

Meden, B. (2013). Application aux réseaux de caméras à champs disjoints. Mémoire de D.E.A., Université de Toulouse, France.

Mikolajczyk, K. et Schmid, C. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86.

Mikolajczyk, K. et Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630.

Murphy, K. (1998). Hidden markov model (hmm) toolbox for matlab.

Oren, M., Papageorgiou, C., Sinha, P., Osuna, E. et Poggio, T. (1997). Pedestrian detection using wavelet templates. *Dans Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 193–199.

Park, U., Jain, A. K., Kitahara, I., Kogure, K. et Hagita, N. (2006). Vise : Visual search engine using multiple networked cameras. *Dans Proceedings of the 18th International Conference on Pattern Recognition*, volume 3, pages 1204–1207.

Pedagadi, S., Orwell, J., Velastin, S. et Boghossian, B. (2013). Local fisher discriminant analysis for pedestrian re-identification. *Dans Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3318–3325.

Pehlivan, S. et Duygulu, P. (2009). 3d human pose search using oriented cylinders. *Dans Proceedings of the 12th IEEE International Conference on Computer Vision Workshops*, pages 16–22.

Porikli, F. M. (2003). Inter-camera color calibration by correlation model function. *Dans Proceedings of the International Conference on Image Processing*, volume 2, pages 133–136.

Prosser, B., Gong, S. et Xiang, T. (2008). Multi-camera matching using bi-directional cumulative brightness transfer functions. *Dans Proceedings of the British Machine Vision Conference*, pages 64.1–64.10.

Prosser, B., Zheng, W.-S., Gong, S. et Xiang, T. (2010). Person re-identification by support vector ranking. *Dans Proceedings of the British Machine Vision Conference*, pages 21.1–21.11.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Dans Proceedings of the IEEE*, volume 77, pages 257–286.

Schmid, C. et Mohr, R. (1997). Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535.

Schwartz, W. R. et Davis, L. S. (2009). Learning discriminative appearance-based models using partial least squares. *Dans Proceedings of the 22nd Brazilian Symposium on Computer Graphics and Image Processing*, pages 322–329.

Scovanner, P., Ali, S. et Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. *Dans Proceedings of the 15th International Conference on Multimedia*, pages 357–360.

Shashua, A., Gdalyahu, Y. et Hayun, G. (2004). Pedestrian detection for driving assistance systems : Single-frame classification and system level performance. *Dans Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 1–6.

Sivic, J. et Zisserman, A. (2003). Video google : a text retrieval approach to object matching in videos. *Dans Proceedings of the 9th IEEE International Conference on Computer Vision*, volume 2, pages 1470–1477.

Sklansky, J. (1978). Image segmentation and feature extraction. *IEEE Transactions on Systems, Man and Cybernetics*, 8(4):237–247.

Skog, D. (2010). Gait-based reidentification of people in urban surveillance video. Mémoire de D.E.A., Department of Information Technology, Uppsala University, Suède.

Smith, S. M. et Brady, J. M. (1997). Susan&mdash ;a new approach to low level image processing. *International Journal of Computer Vision*, 23(1):45–78.

Sorel, A. (2012). *Gestion de la variabilité morphologique pour la reconnaissance de gestes naturels à partir de données 3D.* Thèse de doctorat, Université Rennes 2, France.

Souded, M. (2013). *People Detection, Tracking and Re-identification through a video camera network*. Thèse de doctorat, Institut National de Recherche en Informatique et en Automatique (INRIA), France.

Sundaresan, A., Chowdhury, A. R. et Chellappa, R. (2003). A hidden markov model based framework for recognition of humans from gait sequences. *Dans Proceedings of the International Conference on Image Processing*, volume 2, pages 93–96.

Swain, M. et Ballard, D. (1990). Indexing via color histograms. *Dans Proceedings of the 3rd International Conference on Computer Vision*, pages 390–393.

Thomas-Dietterich (2005). The nearest neighbor algorithm. Rapport technique, Oregon State University, États-Unis.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58(1):267–288.

Tropp, J. (2004). Greed is good : algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242.

Truong Cong, D. N., Achard, C., Khoudour, L. et Douadi, L. (2009). Video sequences association for people re-identification across multiple non-overlapping cameras. *Dans Proceedings of the 15th International Conference on Image Analysis and Processing*, pages 179–189.

Truong Cong, D. N., Khoudour, L., Achard, C., Meurie, C. et Lezoray, O. (2010a). People re-identification by spectral classification of silhouettes. *Signal Processing*, 90(8): 2362–2374.

Truong Cong, N., Achard, C. et Khoudour, L. (2010b). People re-identification by classification of silhouettes based on sparse representation. *Dans Proceedings of the International Confernce on Image Processing Theory, Tools and Applications*, pages 60–65.

Tuytelaars, T. et Mikolajczyk, K. (2008). Local invariant feature detectors : A survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280.

Tuzel, O., Porikli, F. et Meer, P. (2006). Region covariance : A fast descriptor for detection and classification. *Dans Proceedings of the 9th European Conference on Computer Vision*, volume Part II, pages 589–600.

Vega, I. R. et Sarkar, S. (2003). Statistical motion model based on the change of feature relationships : Human gait-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1323–1328.

Viola, P. et Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Dans Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 511–518.

Wang, J., She, M. F., Nahavandi, S. et Kouzani, A. Z. (2010). A review of vision-based gait recognition methods for human identification. *Dans International Conference on Digital Image Computing : Techniques and Applications*, pages 320–327.

Wang, L., Ning, H., Tan, T. et Hu, W. (2003a). Fusion of static and dynamic body biometrics for gait recognition. *Dans Proceedings of the 9th IEEE International Conference on Computer Vision*, volume 2, pages 1449–1454.

Wang, L., Tan, T., Ning, H. et Hu, W. (2003b). Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518.

Wang, X., Doretto, G., Sebastian, T., Rittscher, J. et Tu, P. (2007). Shape and appearance context modeling. *Dans Proceedings of the 11th IEEE International Conference on Computer Vision*, pages 1–8.

Weinland, D., Ronfard, R. et Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241.

Willems, G., Tuytelaars, T. et Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. *Dans Proceedings of the 10th European Conference on Computer Vision*, volume Part II, pages 650–663.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S. et Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227.

Xue, Z., Ming, D., Song, W., Wan, B. et Jin, S. (2010). Infrared gait recognition based on wavelet transform and support vector machine. *Pattern Recognition*, 43(8):2904–2910.

Yang, J., Jiang, Y.-G., Hauptmann, A. G. et Ngo, C.-W. (2007). Evaluating bag-of-visual-words representations in scene classification. *Dans Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 197–206.

Yang, X., Zhou, Y., Zhang, T., Shu, G. et Yang, J. (2008). Fast communication : Gait recognition based on dynamic region analysis. *Signal Processing*, 88(9):2350–2356.

Yu Ohara, Ryusuke Sagawa, T. E. et Yagi, Y. (2004). Gait volume : Spatio-temporal analysis of walking. *Dans Proceedings of the 5th Workshop on Omnidirectional Vision, Camera Networks and Non-classical cameras*, pages 79–90.

Zhang, E., Zhao, Y. et Xiong, W. (2010). Active energy image plus 2DLPP for gait recognition. *Signal Processing*, 90(7):2295–2302.

Zheng, W.-S., Gong, S. et Xiang, T. (2009). Associating groups of people. *Dans Proceedings of the British Machine Vision Conference*, pages 23.1–23.11.

Zheng, W.-S., Gong, S. et Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. *Dans Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 649–656.

Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. *Dans Graphics Gems IV*, pages 474–485. Academic Press Professional, Inc.

Nétographie

CASIA (2001). http ://www.cbsr.ia.ac.cn/english/gait20%databases.asp.

CAVIAR (2003). http ://homepages.inf.ed.ac.uk/rbf/caviar/.

CNIL (2012). http ://www.cnil.fr/linstitution/actualite/article/article/videosurveillance-videoprotection-les-bonnes-pratiques-pour-des-systemes-plus-respectueux-de/ ?tx_ttnews%5bbackpid%5d=91&chash=282959181eb8838bc97778c066d4aa67.

Gradient (wikipedia). http ://fr.wikipedia.org/wiki/gradient.

INHESJ (2013). http ://www.inhesj.fr/sites/default/files/bulletin_annuel_2013.pdf.

LeFigaro (2009). http ://www.lefigaro.fr/actualite-france/2009/03/23/01016-20090323artfig00238-la-videosurveillance-fait-chuter-la-delinquance-de-rue-.php.