



**ÉCOLE DOCTORALE MATHÉMATIQUES, INFORMATIQUE, PHYSIQUE THÉORIQUE ET
INGÉNIERIE DES SYSTÈMES**

Laboratoire PRISME / Laboratoire MIRACL

THÈSE EN COTUTELLE INTERNATIONALE présentée par :

Ahmed DERBEL

soutenue le : 08 septembre 2014

pour obtenir le grade de :
**Docteur de l'université d'Orléans
et de l'Université de Sfax**

Discipline : Sciences et Technologies Industrielles

**Identification de personnes dans un flux vidéo
par l'apparence et la démarche**

THÈSE dirigée par :

M. Abdelmajid BEN HAMADOU Professeur, Université de Sfax
Mme Sylvie TREUILLET Maître de Conférences, Université d'Orléans

RAPPORTEURS :

M. Patrick GROS Directeur de Recherche, Inria
M. Frédéric LERASLE Professeur, Université Paul Sabatier de Toulouse

JURY :

M. Karim ABDED MERAIM	Professeur, Université d'Orléans	Président du jury
M. Patrick GROS	Directeur de Recherche, Inria	
M. Frédéric LERASLE	Professeur, Université Paul Sabatier de Toulouse	
M. Abdelmajid BEN HAMADOU	Professeur, Université de Sfax	
Mme Yousra BEN JEMAA	Maître de Conférences, Université de Sfax	
M. Mohamed HAMMAMI	Maître de Conférences, Université de Sfax	
Mme Sylvie TREUILLET	Maître de Conférences, Université d'Orléans	
M. Bruno EMILE	Maître de Conférences, Université d'Orléans	

Remerciements

Je tiens à remercier en tout premier lieu mes deux directeurs de thèse Mme. Sylvie TREUILLET et Mr. Abdelmajid BEN HAMADOU pour m'avoir permis de réaliser cette thèse au sein de leurs équipes respectives. Je tiens tout particulièrement à leur exprimer ma plus profonde gratitude pour leur encadrement de qualité, leur suivi quotidien, leurs encouragements et leurs conseils précieux.

Je remercie également Mr. Bruno EMILE et Mme. Yousra BEN JEMAA pour avoir co-encadré mes travaux. Un grand merci pour leur soutien, leur disponibilité et leur aide à surmonter les difficultés tout au long de cette thèse.

J'adresse toute ma gratitude à M. Karim ABDED MERAIM, Mr. Patrick GROS, Mr. Frédéric LERASLE et Mr. Mohamed HAMMAMI pour l'honneur qu'ils m'ont fait en acceptant de bien vouloir participer au jury de cette thèse.

Toute ma reconnaissance va aussi à l'Institut Français en Tunisie pour avoir financé une grande partie de mes travaux de recherche en France. Un merci particulier à Mme. Imen ANNEBI pour sa disponibilité et sa contribution au bon déroulement de mes stages. Je remercie aussi l'université d'Orléans pour ses contributions financières qui m'ont permis d'étendre mes séjours en France.

Je tiens à remercier tous les membres des deux laboratoires PRISME et MIRACL, le personnel de Polytech' Orléans et de l'Institut Supérieur d'Informatique et Multimédia de Sfax ainsi que mes chers collègues au sein des deux composantes pour leur bonne humeur et leur bonne ambiance.

Mes remerciements s'adressent également à tous mes amis au sein du laboratoire PRISME et MIRACL doctorants ou docteurs, que j'ai côtoyé quotidiennement tout au long de ces années. Merci aussi à tous mes amis qui m'ont aidé de près ou de loin et avec qui j'ai passé de bons moments.

Je voudrais exprimer ma gratitude et ma reconnaissance envers tous mes enseignants qui ont contribué à ma formation en Tunisie, commençant par l'école primaire El-Oued Sakiet-Eszzit SFAX passant par le lycée Monji Slim Sakiet-Eszzit SFAX et enfin l'Institut Supérieur d'Informatique et Multimédia de Sfax.

Je tiens à remercier, de tout mon cœur, ma famille de m'avoir soutenu pendant toutes mes études. Mes remerciements vont tout naturellement à mes parents pour leur confiance, leur sacrifice, leur soutien inconditionnel et leurs encouragements malgré la distance qui me sépare d'eux ces dernières années. Mes remerciements vont aussi à ma sœur unique ainsi qu'à son mari et ses deux enfants.

Résumé

Cette thèse s'intéresse à la reconnaissance automatique de piétons par l'apparence et la démarche dans des séquences vidéo issues de caméras dont les champs de vues sont non-chevauchants. Avec l'installation généralisée de réseaux de caméras de surveillance dans les zones urbaines, l'enjeu majeur consiste à savoir si une personne préalablement observée par une caméra est toujours présente dans le réseau. Cette reconnaissance nécessite des descripteurs robustes pour caractériser et identifier une personne au-delà des modifications d'apparence induites par un changement de point de vue, un changement d'éclairage ou des variations dans sa vitesse de déplacement.

Dans une première partie dédiée à l'identification par l'apparence, nous proposons une mesure de similarité basée sur des histogrammes régionaux. Sa particularité repose sur l'exploitation de l'information intra et inter-personnes pour s'adapter à la fois aux variations de luminosité et de pose. Une étude comparative montre l'intérêt de cette modélisation, puis une fusion pondérée des descripteurs d'apparence les plus performants permet d'améliorer le taux de ré-identification.

La seconde partie étudie la reconnaissance par la démarche et propose deux stratégies différentes pour traiter respectivement le point de vue latéral et frontal. Pour une observation latérale du piéton, la dynamique de la marche est modélisée par un nouveau descripteur spatio-temporel s'appuyant sur les points d'articulation des deux jambes. Tandis qu'en vue frontale, c'est la distribution des mouvements dans les silhouettes qui est analysée par deux représentations : une carte binaire (RED) et une carte en niveaux de gris (RFD) caractérisent respectivement l'enveloppe et la fréquence des mouvements. L'étude expérimentale démontre que les descripteurs proposés font preuve de bonnes performances face à l'état de l'art, et qu'une fusion de plusieurs descripteurs permet toujours d'améliorer notablement les taux de reconnaissance.

Enfin, la dernière partie de cette thèse introduit un système d'identification associant l'apparence et la démarche. Cette approche mixte est évaluée sur une base de séquences vidéo intégrant de fortes variations de point de vue et de conditions d'éclairage. Elle offre un bon compromis entre efficacité et temps de calcul pour une application en vidéo surveillance.

Abstract

This thesis focuses on automatic pedestrian's recognition by appearance and gait using video sequences from cameras having disjoint fields of view. With the wide-spread installation of surveillance camera networks in urban areas, the major issue is to determine whether a person previously observed by a camera is always present in the network. This recognition requires robust descriptors to characterize and identify a person despite the appearance changes caused by viewpoint, lighting changes or speed variations.

In a first part dedicated to identify people by appearance, we propose a similarity measure based on regional histograms. Its uniqueness consists in exploiting the intra and inter information between persons in order to deal with both the lighting and the pose variations. A comparative study shows the reliability of this modeling and proves that the fusion of the best appearance descriptors improves the re-identification rate.

The second part studies the recognition by gait and proposes two different strategies to deal with the lateral and the front view points respectively. For lateral observation of the pedestrian, the dynamics of walking is modeled by a new spatio-temporal descriptor based on the articulated points of the two legs. While in frontal view, the motion distribution in the silhouettes is analyzed by two representations : a binary (RED) and a grayscale (RFD) maps characterize the envelope and the frequency of the motion respectively. The experimental study proves that the proposed descriptors show good performances against the state of the art, and the fusion of several descriptors improves significantly the recognition rate.

Finally, the last part of this thesis introduces an identification system that combines appearance and gait. This merging approach is evaluated on a video sequences database including large viewpoints and lighting variations. It provides a good compromise between efficiency and computational time for application in video surveillance.

Table des matières

Introduction générale	1
1 État de l’art sur l’identification et le suivi des personnes par l’apparence	7
1.1 Introduction	7
1.2 Approches existantes	9
1.2.1 Suivi de personnes en mono-caméra	9
1.2.1.1 Approches basées sur l’estimation d’un modèle articulé . . .	9
1.2.1.2 Approches fusionnant un modèle prédictif et un modèle d’apparence	10
1.2.2 Identification des personnes dans un réseau de caméras	13
1.2.2.1 Ré-identification exploitant des connaissances a priori sur le réseau de caméra	13
1.2.2.2 Ré-identification basée sur l’apparence	15
1.3 Conclusion	24
2 Identification des personnes par l’apparence	25
2.1 Introduction	25
2.2 Descripteurs d’apparence existants	25
2.2.1 Descripteurs de couleur	26
2.2.1.1 Les histogrammes de couleur	26
2.2.1.2 Les spatiogrammes	26
2.2.2 Descripteurs de texture	27
2.2.2.1 Les matrices de co-occurrences	27
2.2.2.2 Les filtres de Schmid et de Gabor	28
2.2.3 Descripteurs de forme	29
2.2.3.1 Les histogrammes des gradients orientés (HOG)	30

2.2.3.2	Les moments de Zernike et de Hu	30
2.3	Mesure de similarité proposée	31
2.4	Etude comparative entre les descripteurs d'apparence	33
2.4.1	Performances des descripteurs sur la base VIPeR	34
2.4.1.1	La base VIPeR	34
2.4.1.2	Protocole d'évaluation	35
2.4.1.3	Analyse des Performances et interprétation des résultats	38
2.4.2	Performances des descripteurs sur les bases PETS09 et CVLAB	39
2.4.2.1	Les bases PETS09 et CVLAB	40
2.4.2.2	Extraction des silhouettes	41
2.4.2.3	Protocole d'évaluation	44
2.5	Fusion de descripteurs	47
2.5.1	L'algorithme AdaBoost	47
2.5.2	Protocole de test	49
2.6	Etude comparative avec des travaux existants	50
2.7	Conclusion	53
3	État de l'art sur l'identification des personnes par la démarche	54
3.1	Introduction	54
3.2	Identification des personnes par la démarche	55
3.2.1	Approches basées modèles	56
3.2.2	Approches directes	60
3.3	Conclusion	67
4	Identification des personnes par la démarche	68
4.1	Introduction	68
4.2	Approche modèle en point de vue latéral	68
4.2.1	Détection des points d'articulation du squelette	70
4.2.1.1	Techniques existantes avec marqueurs	70
4.2.1.2	Localisation des points d'intérêt	71
4.2.1.3	Erreurs de localisation des points d'intérêt	74
4.2.2	Caractérisation de la démarche à partir du squelette	78
4.2.2.1	Descripteurs existants	78
4.2.2.2	Descripteur <i>Pendule</i> proposé	79
4.2.2.3	Représentation fréquentielle	80

4.2.3	Etude comparative entre descripteurs	81
4.2.3.1	Protocole de test	81
4.2.3.2	CASIA (C) : variations des modes de déplacement en mono- vue :	82
4.2.3.3	CASIA(A) : variations de point de vue	83
4.2.3.4	Discussion	85
4.2.3.5	Impact de la localisation des points sur les descripteurs . . .	87
4.2.4	Fusion des descripteurs	90
4.2.4.1	Pondération des descripteurs retenus	90
4.2.4.2	Résultats	90
4.2.4.3	Comparaison avec d'autres approches	91
4.3	Identification des personnes en vue frontale	94
4.3.1	Analyse des mouvements en vue frontale	94
4.3.1.1	Représentation Fréquentielle de la Démarche (<i>RFD</i>)	95
4.3.1.2	Représentation de l'Enveloppe de la Démarche (<i>RED</i>) . . .	97
4.3.1.3	Caractérisation de la distribution des mouvements	100
4.3.2	Evaluation des descripteurs frontaux	100
4.3.2.1	Protocole de test et résultats	101
4.3.2.2	Apport de la fusion	101
4.3.3	Etude comparative avec des approches existantes	102
4.3.3.1	Travaux sur l'identification en vue frontale	102
4.3.3.2	Performances des approches comparées	104
4.4	Conclusion	104
5	Identification des personnes par l'association de l'apparence et de la dé- marche	105
5.1	Introduction	105
5.2	Approche proposée	105
5.2.1	Identification des personnes par l'apparence	107
5.2.2	Test d'ambiguïté	108
5.2.3	Consolidation par analyse de la démarche	109
5.3	Constitution d'une base de séquences vidéos	109
5.4	Résultats	111
5.4.1	Détermination du seuil de similarité	111
5.4.2	Taux de reconnaissance et temps de calcul	112

5.5 Conclusion	113
Conclusion générale	115
Bibliographie personnelle	120
Bibliographie	131

Table des figures

1.1	Système de vidéo surveillance en utilisant des moniteurs.	7
1.2	Illustration de la variation de l'apparence des piétons par : (a) changement de pose et (b) occultations partielles et (c) changement de couleurs (tiré de la thèse de Slawomir Bak[6]).	8
1.3	Modèle articulé [8].	9
1.4	(a) image originale, (b) image de contour et (c) histogramme de gradients orientés [16].	11
1.5	Processus de construction du modèle d'apparence « Rotation invariant apparence » : regroupement des pixels redondants dans les « cases » [20].	12
1.6	Processus itératif de création des liens spatio-temporels entre les caméras : les régions claires sont des zones d'entrée ou de sortie et les régions sombres sont des zones mortes [25].	15
1.7	(a) image originale, (b) détection des contours par l'algorithme de Canny et (c) patches et (d) les contraintes géométriques [30].	17
1.8	Exemple illustrant la projection du modèle triangulaire sur une image de piéton [44].	20
1.9	L'histogramme de couleur position : (a) image originale, (b) localisation de la silhouette, (c) la répartition des couleurs dans la silhouette [53].	22
1.10	Division des silhouettes en régions par les axes de symétrie et asymétrie [55].	23
2.1	Voisinage immédiat d'un pixel selon quatre directions.	28
2.2	Les filtres de Schmid.	28
2.3	Les filtres de Gabor.	29
2.4	Masques d'orientation du contour.	30
2.5	Descripteur proposé.	32

2.6	Quelques exemples de la base VIPeR : chaque colonne est une paire d'image représentant la même personne.	34
2.7	Courbes CMC pour les descripteurs de couleur (a, b, c), les descripteurs de texture (d), les descripteurs de forme (e) et les meilleurs descripteurs pour chaque catégorie (f).	37
2.8	(a) images originales de la même personne, (b) quelques blocs ayant des contours similaires malgré la grande variation de pose.	39
2.9	Extrait de la base PETS09.	40
2.10	Extrait de la base CVLAB.	41
2.11	Exemples d'extraction de l'arrière plan : (a) base PETS09 et (b) CVALAB.	42
2.12	Résultat des opérateurs morphologiques appliqués sur les mêmes images que celles de la figure 2.11.	43
2.13	Echantillons d'images de personnes extraites respectivement de la base PETS09 et CVLAB.	44
2.14	Calcul de la matrice d'autosimilarité.	45
2.15	Poids des descripteurs en pourcent en utilisant : (a) uniquement les descripteurs de couleur et (b) les descripteurs de couleur et de texture.	49
2.16	Apport de la fusion sur les taux de reconnaissance.	50
2.17	Courbes de précision CMC pour les cinq approches comparées.	52
3.1	Descripteurs utilisés dans [79].	56
3.2	Angles de la hanche et du genou utilisés dans [83] et [84].	57
3.3	Positions verticales des points d'articulation selon l'étude médicale [86].	58
3.4	Signatures spatio-temporelles de la démarche selon [89] : (au centre) la silhouette et les pourtours selon les quatre côtés de la boîte englobante, (au tour) les signatures temporelles cumulant les distances au cours du temps.	59
3.5	Modèle en trapèzes de la partie inférieure de la silhouette [91].	60
3.6	Les cinq positions de silhouettes les plus fréquentes dans un cycle de démarche [92].	61
3.7	Processus d'extraction du signal unidimensionnel [96].	62
3.8	Exemples de matrices SSP pour 6 personnes différentes calculées à partir d'images binaires des silhouettes [101].	63
3.9	Processus de construction du modèle de freize [102].	64
3.10	Extraction des images de différence [103].	65

3.11	Sélection de la région d'intérêt la plus discriminante : (a) par la méthode supervisée et (b) par la méthode non supervisée [105]	66
3.12	Présentation de l'image GEI (a) avec l'image EGEI (b) correspondante [106].	66
4.1	Modélisation des silhouettes : (a) boîte englobante, (b) modèle en squelette, (c) modèle en trapèzes et (d) modèle en ellipses.	69
4.2	Processus de détection des points d'intérêt selon l'attitude : (colonne de gauche) simple support, (colonne de droite) double support, avec deux points de vue différentes (b : fronto-parallèle et c : oblique).	73
4.3	Exemple de détection automatique des squelettes : (a) simple support et (b) double support.	74
4.4	Superposition des squelettes de la vérité de terrain et ceux détectés automatiquement pour la même personne et selon toutes les catégories de mouvement : (a) Normale, (b) Rapide, (c) Lente et (d) Avec Sac.	77
4.5	Descripteurs existants.	79
4.6	Descripteur <i>Pendule</i> : (a) Processus d'extraction et (b) variations durant deux cycles de démarche.	80
4.7	Exemples de signaux pour chaque descripteur.	86
4.8	Variations spatio-temporelles des descripteurs extraits à partir de la localisation automatique des points versus la vérité de terrain.	89
4.9	Poids attribués aux descripteurs retenus (en pourcentages).	90
4.10	Les zones d'intérêt pour (a) GEI et (b) Inf-GEI.	92
4.11	Silhouettes alignées et normalisées pour la même personne.	95
4.12	Séquence de silhouettes binaires (10 images régulièrement espacées dans la séquence) et les représentations <i>RFD</i> et <i>RED</i>	96
4.13	Représentation Fréquentielle de la Démarche (<i>RFD</i>) pour 3 personnes et 4 séquences différentes de la base CASIA(A).	97
4.14	Représentation de l'Enveloppe de la Démarche (<i>RED</i>) pour 3 personnes et 4 séquences différentes de la base CASIA(A) - même sélection que la figure 4.13.	99
5.1	Processus d'identification des personnes couplant l'apparence et la démarche.	107
5.2	Représentation du résultat sous forme de graphe.	108
5.3	Présentation de la base avec variations des conditions d'éclairage et de point de vue.	110
5.4	Organisation temporelle des séquences de la base.	111

5.5	F-mesure pour les différents seuils de ré-identification par l'apparence. . . .	112
-----	---	-----

Liste des tableaux

2.1	Différents configurations testées pour chaque descripteur d'apparence.	35
2.2	Performances des descripteurs retenus en termes d'identification mono-caméra et de ré-identification.	46
2.3	Confrontation de cinq approches de fusion.	51
4.1	Choix du modèle représentant les silhouettes.	69
4.2	Précision de la détection des points d'intérêt (en pixels).	76
4.3	Tests réalisés sur la base CASIA(C).	83
4.4	Taux de reconnaissances obtenus avec les 2 transformations fréquentielles pour chaque type de déplacement (en %).	83
4.5	Taux de reconnaissances en point de vue oblique (en %).	84
4.6	Taux de reconnaissances en point de vue fronto-parallèle (en %).	84
4.7	Taux de reconnaissances en multi-vues (en %).	85
4.8	Choix des transformations.	86
4.9	Impact de la détection automatique des points sur les descripteurs.	88
4.10	Taux de reconnaissance par la fusion des descripteurs face aux variations de modes de déplacement (en %).	91
4.11	Taux de reconnaissance par la fusion des descripteurs face aux variations de point de vue avec variations de vitesse (en %).	91
4.12	Taux de reconnaissances face aux variations de modes de déplacement (en %).	93
4.13	Taux de reconnaissances face aux variations de point de vue avec variations de vitesse (en %).	93
4.14	Taux de reconnaissance pour les descripteurs comparés (en %).	101
4.15	Taux de reconnaissance pour toutes les combinaisons de fusion (en %).	102
4.16	Comparaison des approches en vue frontale sur la base CASIA(A).	104
5.1	Descripteurs utilisés dans les différents modules d'identification.	106

5.2 Taux de reconnaissance et temps d'exécution. 113

Liste des algorithmes

2.1	Algorithme AdaBoost	48
4.1	Calcul de la Représentation Fréquentielle de la Démarche (<i>RFD</i>).	96
4.2	Calcul de la Représentation de l'Enveloppe de la Démarche (<i>RED</i>).	98
5.1	Processus d'identification des personnes par l'apparence.	108

Introduction générale

Garantir la sécurité dans les espaces publics est une préoccupation majeure que se soit pour les gouvernements comme pour les sociétés de transport afin d'offrir la qualité de service attendue par les usagers. Les systèmes de suivi de personnes se sont donc développés pour associer l'identité des personnes et leurs déplacements afin d'anticiper les comportements dangereux ou suspects (acte terroriste, violence, meurtre). Au-delà des objectifs sécuritaires, les systèmes de suivi de personnes sont aussi utilisés dans d'autres contextes visant à doter les bâtiments d'une certaine forme d'intelligence afin de faciliter la vie quotidienne des gens. Ils permettent, par exemple, d'offrir une assistance à la navigation des personnes handicapées dans les bâtiments ou de faciliter l'évacuation d'urgence en cas de sinistre.

Pour assurer un suivi automatique des personnes, différentes technologies embarquées ont été développées. Certains travaux utilisent le GPS « Global Positioning System » pour déterminer les positions des piétons [1]. D'autres travaux exploitent un réseau de capteurs ou de balises ultrasons [2], infrarouge [3] ou RFID [4] que ce soit seul ou en association avec des capteurs tels que les accéléromètres, les magnétomètres ou les gyroscopes [5]. Ces systèmes présentent toutefois des limitations : le GPS est inopérant en environnement intérieur, et il a tout comme les MEMS une faible précision en milieu urbain, et enfin, la sensibilité des balises à ultrasons, infrarouge et RFID face aux interférences reste importante.

La vision par ordinateur représente une solution alternative intéressante pour suivre et identifier les personnes aussi bien à l'intérieur qu' à l'extérieur. Le système peut être autonome et son coût reste faible. Cette démarche est motivée par l'installation généralisée des caméras de surveillance. Elle permet d'offrir une totale liberté de déplacements pour les personnes comparée aux technologies embarquées, tout en garantissant une observation plus ou moins complète de l'espace contrôlé.

Les travaux de thèse présentés dans ce mémoire contribuent au développement d'un système automatique d'identification de personne dans un réseau de caméras. L'objectif principal est de proposer un module de caractérisation fusionnant l'apparence et la démarche de chaque personne pour permettre une identification robuste face aux nombreuses difficultés inhérentes à cette problématique.

En ce qui concerne l'apparence, une personne représente un objet particulièrement difficile à suivre et à reconnaître dans la mesure où il s'agit d'un objet très articulé dont le mouvement est imprévisible. La phase de caractérisation doit prendre en compte cette variabilité. En mono-caméra, un système de reconnaissance de piéton doit être capable de gérer les faibles variations de pose car les positions des silhouettes varient légèrement entre deux images successives, ainsi que les occultations partielles. Dans le but d'élargir la zone contrôlée, plusieurs caméras à champs de vue non-chevauchants sont fréquemment installées. Les personnes sont alors invisibles lors de leur passage entre deux sites, d'où l'apparition du problème de ré-identification. La ré-identification consiste à déterminer si une personne a déjà été observée préalablement dans le réseau de caméras. Un système de reconnaissance de piéton multi-caméras doit palier aux fortes variations de pose, de point de vue et d'échelle, les occultations partielles et totales, les grands changements d'éclairage ou la variabilité du paramétrage des caméras.

La démarche est une caractéristique biométrique non-intrusive permettant l'identification des personnes à distance. Toutefois, plusieurs problématiques complexes doivent être gérées afin d'assurer une bonne phase de reconnaissance, notamment les fortes variations de pose, de point de vue, d'apparence (manière de s'habiller : jupe, pantalon, manteau, etc.) et la vitesse de déplacement. La façon de se déplacer représente également une contrainte difficile à gérer qui dépend de plusieurs facteurs comme la nature des chaussures et du terrain, le fait de porter ou non un objet et de l'état psychologique des personnes lors de leurs déplacements.

Dans le cadre de cette thèse, nous nous sommes principalement focalisés sur la phase de caractérisation en traitant les problématiques de l'identification des personnes que se soit par l'apparence ou par la démarche. Nous avons étudié l'impact des variations des conditions d'éclairage et de point de vue sur les performances des descripteurs d'apparence, ainsi que l'impact des variations de point de vue, de vitesse et de la façon de se déplacer sur les descripteurs de démarche. Les principales contributions de la thèse sont les suivantes :

- Une mesure de similarité basée sur les histogrammes régionaux couleur a été proposé pour identifier les personnes par l'apparence. Sa particularité consiste à intégrer l'information intra-personne pour réduire l'impact des variations des conditions d'éclairage. Ce travail est détaillé dans le chapitre 2.
- Une étude comparative a été réalisée pour évaluer les performances de cette modélisation face aux principaux descripteurs de couleur, de texture et de forme de la littérature. Deux variantes de fusion ont ensuite été testées avec les descripteurs ayant démontré les meilleures performances lors de la comparaison. La fusion des descripteurs de couleur seuls ou combinés avec des descripteurs de texture ont été finalement comparées à plusieurs travaux de référence dans la littérature. Ce travail est décrit dans le chapitre 2.
- Afin de réaliser l'analyse de la démarche, nous avons introduit une méthode automatique de détection des points d'articulation des silhouettes. Basée sur des règles morphologiques simples, elle s'applique à la fois en point de vue fronto-parallèle ou oblique. Sa fiabilité a été évaluée en mesurant les écarts introduits par la détection automatique par rapport à une réalité de terrain construite manuellement sur plusieurs séquences. L'impact de ces erreurs de localisation sur la phase de caractérisation a aussi été étudié. Les résultats obtenus confirment que la méthode de localisation des points d'intérêt proposée permet de gérer les occultations partielles, les variations de la vitesse et de la manière de se déplacer. Cette contribution sera présentée dans le chapitre 4.
- Un descripteur de la démarche d'une personne observée latéralement a été proposé. Il représente le déplacement de la personne comme un pendule oscillant en utilisant les positions de la hanche et des chevilles. La particularité de ce descripteur spatio-temporel consiste à exploiter la relation entre la force du mouvement (largeur de la foulée) et les variations de la position d'équilibre (point de la hanche). Ce travail est détaillé dans le chapitre 4.
- Une étude comparative entre le descripteur de démarche proposé et les principaux descripteurs de la littérature permet d'évaluer leurs performances faces aux variations de point de vue, de vitesse et de manière de se déplacer. Pour la représentation des descripteurs spatio-temporels de la démarche, nous avons appliqué la transformée de Fourier largement utilisée dans la littérature, mais aussi l'enveloppe de la transformée en ondelettes appliquée pour la première fois dans ce contexte. L'approche proposée consiste à

combiner les meilleurs descripteurs avec la transformée la plus appropriée afin d'identifier les personnes en déplacements latéraux. Ses performances sont comparées avec les principaux travaux de la littérature dans plusieurs contextes expérimentaux. Ce travail est détaillé dans le chapitre 4.

- Afin de pouvoir identifier les personnes en vue frontale, deux nouvelles modélisations de la distribution du mouvement à l'intérieur des silhouettes ont été proposées : la Représentation Fréquentielle de la Démarche (RFD) et la Représentation de l'Enveloppe de la Démarche (RED). La représentation RFD est caractérisée par deux nouveaux descripteurs modélisant respectivement l'entropie et la dispersion des mouvements dans les silhouettes. Nous avons aussi testé plusieurs descripteurs de forme classiques pour caractériser la représentation RED. Les résultats obtenus confirment que ces deux représentations sont très complémentaires et permettent d'assurer une bonne phase de caractérisation des personnes en vue frontale. Ce travail est présenté dans le chapitre 4.
- Enfin, un module d'identification complet, fusionnant l'apparence et la démarche, a été testé dans un contexte de vidéo surveillance intégrant de fortes variations de point de vue et d'éclairage. Les résultats obtenus ont montré que l'intégration des caractéristiques comportementales permet d'améliorer considérablement les performances en termes d'identification des personnes en multi-caméras. Cette partie est présentée dans le chapitre 5.

Plan de la thèse

Cette thèse porte sur l'identification des personnes en multi-caméras en fusionnant des descripteurs d'apparence et de comportement (démarche). Les chapitres sont organisés comme suit.

Le chapitre 1 présente un état de l'art des différentes approches d'identification des personnes par l'apparence. Il est particulièrement orienté vers les modèles de caractérisation de l'apparence et les différentes méthodes de classification utilisées.

Le chapitre 2 introduit une nouvelle mesure de similarité appliquée sur les histogrammes

régionaux ainsi que les descripteurs existants dans la littérature pour caractériser les personnes s'appuyant sur la couleur, la texture et la forme. Ces descripteurs sont dans un premier temps évalués dans un contexte de ré-identification en faisant varier plusieurs paramètres influents comme l'espace colorimétrique, la phase de quantification ou la taille des blocs. Puis une sélection des descripteurs les plus efficaces est évaluée dans un contexte de suivi mono-caméra et ré-identification selon un protocole de test basé sur la similarité moyenne pour la même personne, la similarité moyenne pour différentes personnes et le nombre de fausses alertes détectées. Nous clôturons ce chapitre par une approche de fusion combinant les descripteurs les plus robustes qui est finalement comparée à plusieurs travaux de la littérature.

Le chapitre 3 montre l'intérêt d'intégrer une caractéristique biométrique dans la phase de caractérisation et argumente le choix de la démarche. Le reste du chapitre est destiné à faire l'état de l'art sur les différentes approches permettant de caractériser les personnes par leur démarche.

Le chapitre 4 présente les différentes stratégies adoptées pour identifier les personnes par la démarche que ce soit en vue latérale (points de vue fronto-parallèle et oblique) et en vue frontale. Pour les points de vue latéraux, nous avons introduit un descripteur spatio-temporel s'appuyant sur une détection automatique des points d'articulation des membres inférieurs. Il sera comparé avec les principaux descripteurs de la littérature dans plusieurs contextes expérimentaux intégrant des variations de point de vue, de vitesse et de façon de se déplacer. Deux transformées ont été appliquées sur les descripteurs afin de caractériser leurs variations spatio-temporelles : une transformée fréquentielle et une autre temps-fréquence. La stratégie proposée consiste ensuite à combiner les descripteurs les plus robustes avec la transformée la plus appropriée pour caractériser la démarche. Pour les points de vue frontaux, nous avons adopté une stratégie différente basée sur un modèle de distribution du mouvement dans les silhouettes des personnes. Pour cela, deux représentations ont été construites (RFD et RED). La première est caractérisée par l'entropie et la dispersion du mouvement par rapport au centre de la silhouette et la deuxième par des descripteurs de forme classiques. Ces nouvelles représentations sont finalement comparées à plusieurs travaux existants dans la littérature.

Le chapitre 5 présente le système complet d'identification des personnes associant l'apparence et le comportement. Ce système est basé sur une identification image/image réalisée par le modèle d'apparence suivi par une étape de correction séquence/séquence assurée par le modèle de démarche. Une étude expérimentale dans un contexte réel de vidéosurveillance

intégrant des fortes variations de point de vue et de conditions d'éclairage met en valeur l'apport des descripteurs de la démarche sur les performances d'identification.

Enfin, la conclusion résume les différentes contributions de ce travail et s'ouvre sur des propositions d'améliorations et des perspectives.

Chapitre 1

État de l'art sur l'identification et le suivi des personnes par l'apparence

1.1 Introduction

La vidéosurveillance des personnes s'est beaucoup développée avec la généralisation de l'équipement en caméras des sites urbains. Ces caméras permettent une observation plus ou moins complète de scènes en direct ou à posteriori. Dans certains cas, cette tâche de vidéo surveillance est confiée à un agent de sécurité qui visualise un système de moniteurs (figure 1.1).



FIGURE 1.1 – Système de vidéo surveillance en utilisant des moniteurs.

En pratique, il est très difficile pour l'opérateur humain de surveiller continuellement un grand nombre de moniteurs. Des systèmes automatisés d'analyse de scènes ont alors été développés pour permettre de donner l'alerte en cas de besoin.

Plusieurs approches existent pour assurer le suivi et l'identification automatiques des personnes dans des séquences vidéo. Nous pouvons distinguer les approches mono et multi-caméras. En mono-caméra, un système de reconnaissance de personnes est soumis à de faibles variations de pose (les positions des silhouettes varient légèrement entre deux images successives), ainsi qu' à des occultations partielles. En multi-caméras, le problème d'identification des personnes devient plus complexe. Les variations de pose peuvent être plus brutales : de grands changements dans les couleurs causés par la modification des conditions d'éclairage ou par la variabilité au niveau du paramétrage des caméras dans le réseau sont possibles ainsi que les changements d'échelle (figure 1.2).



FIGURE 1.2 – Illustration de la variation de l'apparence des piétons par : (a) changement de pose et (b) occultations partielles et (c) changement de couleurs (tiré de la thèse de Slawomir Bak[6]).

1.2 Approches existantes

1.2.1 Suivi de personnes en mono-caméra

Les approches de suivi de personnes en mono-camera peuvent être classées en deux catégories : les approches basées sur l'estimation d'un modèle articulé et les approches fusionnant un modèle prédictif et un modèle d'apparence.

1.2.1.1 Approches basées sur l'estimation d'un modèle articulé

Ces approches sont basées sur l'évolution d'un modèle articulé. Le suivi est réalisé par la minimisation de la différence entre les modèles articulés des images à l'instant t et $t-1$. Ce modèle (figure 1.3) extrait des silhouettes un ensemble de rectangles correspondant aux différentes parties du corps et leurs points d'articulation remarquables (hanches, genoux, chevilles).

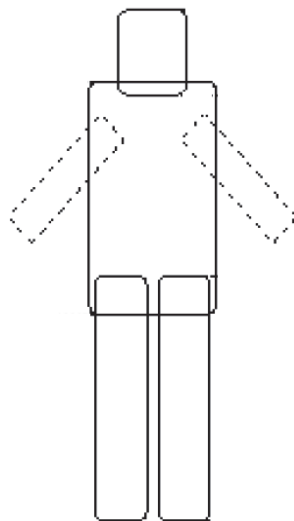


FIGURE 1.3 – Modèle articulé [8].

Ioffe and Forsyth [7] proposent de suivre les personnes en positionnant dans chaque image un modèle articulé nommé « Mixture of Tree » (racine de l'arbre : le torse, noeuds du premier niveau : les bras et les cuisses et noeuds du deuxième niveau : les avant-bras et les jambes). Une telle structure permet un assemblage hiérarchique des composants, recherche du torse puis des bras et des cuisses et enfin des avant-bras et des jambes, ce qui réduit considérablement le temps d'exécution. Cette structure hiérarchique offre aussi une grande capacité

à gérer les occultations partielles : par exemple, si la cuisse est cachée, le noeud de l'avant jambe reste attaché au noeud torse. Le suivi de personnes est réalisé en estimant l'évolution de ce modèle articulé par la maximisation de la probabilité d'assemblage des composants en tenant compte des contraintes spatiales (pose) et temporelles (positions des silhouettes). En effet, les variations de pose et de la position des silhouettes sont minimales entre deux images successives de la même personne.

Cheng and Moura [8] utilisent une représentation géométrique des silhouettes en un ensemble d'objets rigides articulés combiné à une représentation périodique du mouvement. Le suivi des personnes est réalisé en fusionnant les contraintes dynamiques calculées à partir de la phase et de la fréquence du mouvement avec les contraintes cinématiques destinées à l'estimation des angles reliant les différentes parties du corps humain.

Pour suivre les personnes, Haritaoglu et al. [9] combinent une estimation de la position des silhouettes réalisée par la méthode des moindres carrés récurrents à une mise en correspondance entre les contours des silhouettes successives. Cependant, dans certains cas, la forme des silhouettes varie considérablement entre deux images successives à cause des occultations partielles, du croisement ou de la séparation des personnes. Pour palier ce problème, les auteurs proposent de suivre séparément la tête, le torse, les mains et les jambes représentés par un modèle articulé nommé « Cardboard Model ». La phase de vérification et de raffinement des positions des différentes parties du corps humain est réalisée par un modèle d'apparence nommé « Temporal texture template » obtenu en moyennant les silhouettes des personnes.

1.2.1.2 Approches fusionnant un modèle prédictif et un modèle d'apparence

Le suivi de personnes en mono-caméra peut également être réalisé en combinant un modèle de mouvement prédictif linéaire de type « filtre de Kalman » [10] ou non linéaire « les filtres à particules » [11] avec un modèle d'apparence. Dans ce cas, le filtre prédictif permet d'estimer la position d'une personne en se référant à l'historique de son mouvement. Par exemple, TraKulPong et Bowden [12] proposent d'appliquer un modèle de mouvement basé sur le filtrage de Kalman pour prédire la position du centroïde, la hauteur et la largeur de la boîte englobant la personne. Ce modèle de mouvement est fusionné avec un modèle d'apparence représentant les silhouettes par un histogramme de couleur dans un espace colorimétrique réduit appelé « CCCM ». Adapté aux images de basses résolutions, cet espace de couleur est formé par les 11 couleurs de base de l'espace RVB sélectionnés selon une étude

physiologique du système visuel humain [13].

D'autres travaux combinent le filtrage de Kalman avec un histogramme de couleur en intégrant de l'information spatiale (poids et indicateur d'affectation de chaque classe de l'histogramme à une partie de la silhouette : la tête, les parties supérieures ou inférieures du corps humain) [14], ou avec un mélange de modèles gaussien représentant chacun la teinte et la saturation de chaque pixel de la silhouette [15].

Mais, la trajectoire rectiligne d'une personne dans le plan du sol n'est pas toujours linéaire : une personne peut imprévisiblement s'arrêter ou changer de direction. Ainsi, certains travaux ont choisi d'appliquer un modèle de prédiction non linéaire à savoir les filtres à particules.

Dans ce sens, Yang et al. [16] proposent de combiner le filtre à particules avec un modèle d'apparence couleur et de forme en cascade en utilisant l'algorithme AdaBoost [17]. Le modèle de couleur est représenté par la moyenne des trois composantes RVB pour chaque bloc horizontal de la silhouette. La forme est modélisée par l'histogramme de gradient orienté illustré dans la figure 1.4.

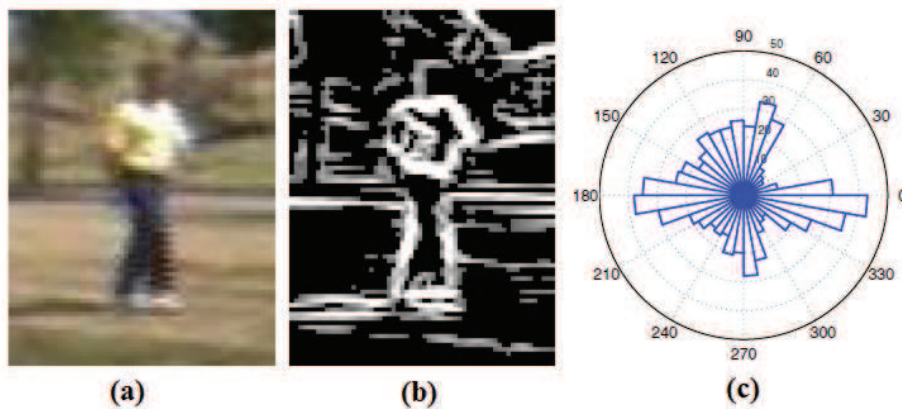


FIGURE 1.4 – (a) image originale, (b) image de contour et (c) histogramme de gradients orientés [16].

Tung et Matsuyama [18] proposent d'intégrer un modèle d'apparence couleur et un modèle de mouvement dans le filtrage à particules afin de suivre séparément la tête et les deux mains. L'apparence de ces 3 parties du corps humain (tête et les deux mains) est représentée

chacune par un histogramme de couleur dans l'espace HSV. Le modèle de mouvement est basé essentiellement sur une prédiction par le filtrage à particules en tenant compte aussi de la prédiction par l'algorithme de flot optique [19]. Le flot optique permet de prédire le nuage de particules estimant la distribution *a posteriori* de l'état. Ceci offre l'avantage de limiter les erreurs de prédiction des filtres à particules, garantissant ainsi que l'estimation ne s'écarte pas trop de la position réelle.

Etant donné que la fiabilité et le temps d'exécution des filtres particuliers dépend principalement du nombre de particules, le travail de Wang et al. [20] a été de réduire leur nombre tout en assurant une bonne phase de prédiction. Pour cela, une phase d'apprentissage hors ligne est réalisée pour construire une base de trajectoires de référence. L'estimation de la position de la personne se fait ensuite en choisissant un nombre de particules réduit et en utilisant la trajectoire de référence la plus proche de la trajectoire actuelle pour le suivi. La mise à jour des poids des particules utilise une modélisation par un mélange de gaussiennes pour représenter les régions homogènes des visages des personnes ainsi qu'un modèle de couleur invariant à la variation de pose pour caractériser les pieds (figure 1.5). Un pied est d'abord détecté à l'aide d'un module de segmentation puis placé dans un cercle de dimensions minimales. Pour chacune des directions présentées dans la figure 1.5, ce cercle est découpé en 4 régions concentriques de rayons différents. Les pixels redondants appartenant au même cercle concentrique pour les différentes directions sont cumulés dans des « cases ». Finalement, un modèle gaussien caractérise les propriétés colorimétriques de chaque « case ».

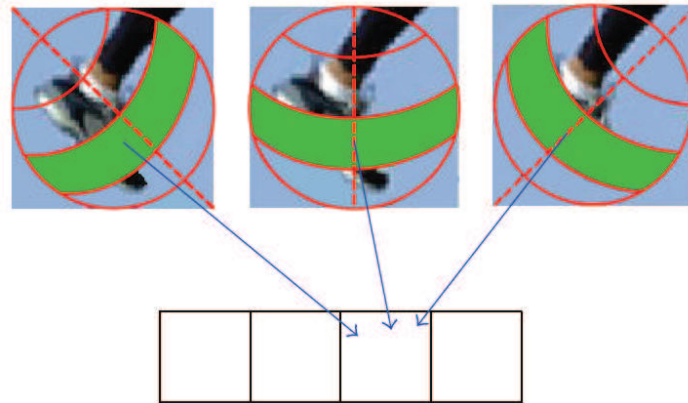


FIGURE 1.5 – Processus de construction du modèle d'apparence « Rotation invariant apparence » : regroupement des pixels redondants dans les « cases » [20].

1.2.2 Identification des personnes dans un réseau de caméras

Dans un système de vidéosurveillance, le nombre de caméras utilisé ainsi que leur topologie affecte énormément le coût et surtout la fiabilité du système. Un premier cas de figure consiste à utiliser des caméras dont les champs de vue sont chevauchants. Cette topologie fournit une description très riche de la scène surveillée dans le sens où les cibles mobiles seront toujours visibles dans le réseau, ce qui permet de gérer aussi les cas d'occultations et d'avoir la possibilité de calculer les positions dans l'espace contrôlé [21][22]. Mais, le besoin de couvrir de plus en plus d'espace implique l'installation de caméras à champs de vues non-chevauchant à moindre coût. Cette topologie est la plus probable dans un contexte réel de vidéosurveillance dans les grandes infrastructures. Cependant, l'utilisation de caméras à champs de vus disjoints nécessite la gestion des interruptions de suivi lorsque les personnes seront invisibles durant un intervalle de temps non défini. Un nouveau problème, appelé la ré-identification, est alors considéré. La ré-identification consiste à reconnaître les personnes qui sont filmées par une caméra i et savoir si elles ont été précédemment capturées par une autre caméra j du réseau. Ce problème présente un certain nombre de difficultés causées par la nature très articulée du corps humain, les changements de pose et d'échelle ainsi que la variabilité des réponses colorimétriques des différentes caméras du réseau, liée au paramétrage et aux changements des conditions d'éclairage.

Dans le reste de ce chapitre, nous nous intéressons aux approches traitant le problème de ré-identification. Ces approches peuvent être classées en deux catégories : approches fusionnant des connaissances *a priori* sur le réseau de caméra avec un modèle d'apparence et les approches basées uniquement sur un modèle d'apparence (sans connaissance *a priori* sur le réseau).

1.2.2.1 Ré-identification exploitant des connaissances a priori sur le réseau de caméra

Les connaissances *a priori*, collectées durant une phase d'apprentissage hors ligne, sont relatives aux différentes trajectoires possibles entre les caméras du réseau ainsi que le temps nécessaire pour les parcourir. Ces connaissances *a priori* sont utilisées quand les personnes sont invisibles durant le passage entre caméras.

Kettnaker et Zabih [23] proposent d'exploiter les contraintes liées au mouvement des personnes pour les suivre à travers un réseau de caméras. En effet, les positions et les vitesses

des piétons ainsi que les temps de transition inter-caméras sont utilisés pour estimer la probabilité de transition entre les caméras. Le suivi de personnes est réalisé en tenant compte des probabilités de transition déjà calculées et d'un modèle d'apparence formé par un histogramme de couleur dans l'espace HSV.

Javed et al. [24] présentent un système d'apprentissage permettant d'estimer la topologie des caméras ainsi que les variations des réponses colorimétriques entre elles dans le réseau. Les probabilités de passage sont calculées par un modèle Bayésien exploitant la redondance dans les parcours empruntés par les personnes. Il tient également compte de quelques paramètres spatio-temporels comme le temps de transition inter-caméras, les emplacements des zones d'entrée et de sortie des caméras et les directions des mouvements des personnes. Le changement d'apparence des personnes entre les caméras est représenté par une distribution gaussienne modélisant les variations colorimétriques entre leurs différentes observations dans le réseau. L'identification des personnes est finalement réalisée en fusionnant les probabilités de passage entre les caméras avec un histogramme de couleur.

Gilbert et Bowden [25] proposent d'appliquer une phase d'apprentissage pour extraire les liens spatio-temporels entre les différentes zones d'entrée et de sortie des caméras (trajectoires possibles) ainsi que les matrices de transfert de couleurs entre les caméras. Pour estimer les liens spatio-temporels, chaque champ de vision d'une caméra est initialement modélisé par un bloc. Ces blocs sont itérativement fusionnés, divisés, ou supprimés compte tenu de l'existence de passages des personnes entre ces zones (figure 1.6). Les passages sont déterminés par un suivi des personnes basé sur un histogramme de couleur dans l'espace CCCM. Après la phase d'apprentissage, l'identification des personnes est réalisée en combinant les liens spatio-temporels avec un histogramme de couleur dans l'espace RVB. Celui-ci est mieux adapté que l'espace CCCM après la phase d'étalonnage des couleurs par les matrices de transfert [25].

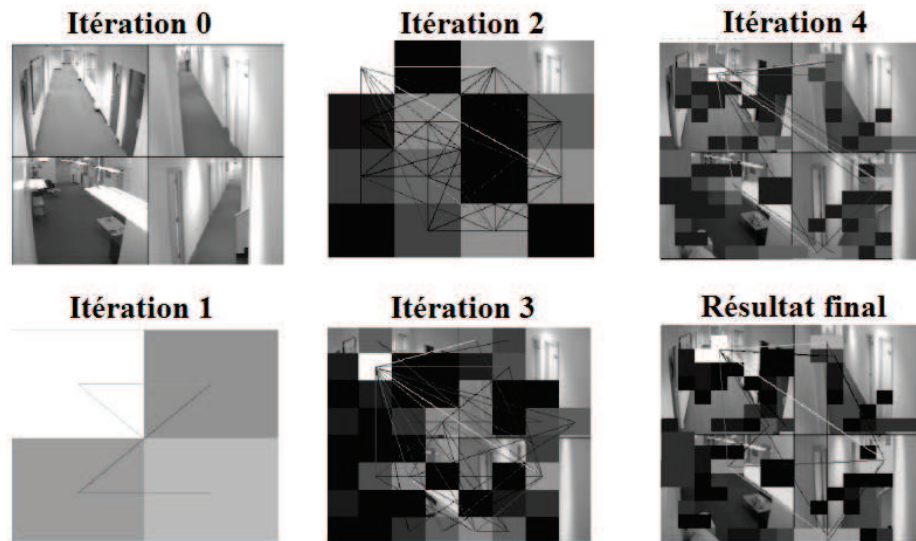


FIGURE 1.6 – Processus itératif de création des liens spatio-temporels entre les caméras : les régions claires sont des zones d’entrée ou de sortie et les régions sombres sont des zones mortes [25].

1.2.2.2 Ré-identification basée sur l’apparence

L’estimation des connaissances *a priori* du réseau de caméras (probabilités et temps de transition inter-caméra) nécessitent une phase d’apprentissage hors ligne qui doit être réalisée séparément pour chaque site et bien évidemment chaque fois que la topologie du réseau est modifiée. Plusieurs travaux ont alors opté pour une ré-identification de personnes basée uniquement sur un modèle d’apparence. Cette stratégie nécessite l’élaboration d’une signature non seulement discriminante, qui permet de bien séparer les différentes personnes, mais aussi invariante et robuste face aux différents problèmes liés au contexte multi-caméras (changement de pose, des conditions d’éclairage, variabilité de la réponse colorimétrique des caméras, les occultations...). Les travaux proposés dans la littérature pour caractériser l’apparence des personnes peuvent être divisés en deux grandes catégories : les approches «mono-image» visant à modéliser séparément chaque image de personnes et les approches «multi-images» tirant profit de la redondance d’information à partir d’un ensemble d’images de la même personne.

- **Approches mono-image** Les approches mono-image offrent l’avantage d’être les plus adaptées aux applications temps réel car la caractérisation est réalisée sur une seule image permettant ainsi de manipuler des vecteurs caractéristiques de petite taille. Toutefois, il est

difficile d'extraire une signature robuste à partir d'une seule observation.

Park et al. [26] proposent de caractériser les personnes par une modélisation colorimétrique des parties inférieures de leurs silhouettes couplée avec une estimation de la hauteur des personnes dans chaque vue par une transformation homographique. Chaque silhouette est divisée en 3 blocs proportionnellement à la hauteur ($1/5H$, $3/5H$ et $1/5H$). Les 2 blocs inférieurs sont modélisés par des histogrammes de couleurs dans un espace colorimétrique réduit (de 10 teintes) afin de limiter les variations des couleurs possibles entre caméras.

Au contraire, Gallagher et Chen [27] proposent d'identifier les personnes par la couleur de la partie supérieure de la silhouette couplée à quelques caractéristiques faciales. Les visages des personnes sont détectés par la méthode de Viola et Jones [28] puis projetés dans un espace de dimensions réduites « Fisherfaces » [29]. La phase d'extraction de la partie supérieure des silhouettes est basée d'une part, sur l'estimation d'un masque nommé « Global Clothing Masks » construit par un apprentissage automatique hors ligne en utilisant l'information mutuelle et, d'autre part, sur un « graph cuts » permettant de mettre à jour le « Global Clothing Masks » par la classification des pixels de l'avant et l'arrière plan en cas de variations de pose ou d'occultations. Le modèle d'apparence appliqué consiste à segmenter la partie supérieure des personnes en régions homogènes qui sont, par la suite, représentées séparément par un vecteur caractéristique de taille 5. Les 3 premières caractéristiques sont obtenues par une transformation linéaire de l'espace de couleur RVB et les deux autres caractéristiques représentent la texture en tenant compte respectivement de la direction horizontale et verticale.

Cai et al. [30] exploitent des patches de couleurs situés à proximité des contours des silhouettes pour modéliser l'apparence des personnes (figure 1.7). Les contours sont obtenus par l'algorithme de Canny [31]. Les patches (ou régions) sélectionnés sont caractérisés par les couleurs dominantes ainsi que leurs fréquences d'apparition dans ces régions. La similarité entre deux personnes dépend de la ressemblance colorimétrique entre les patches en considérant aussi des contraintes géométriques sur leur position relative au point supérieur de la tête (par la distance D normalisée par rapport à la hauteur de la silhouette et l'angle de déviation θ).

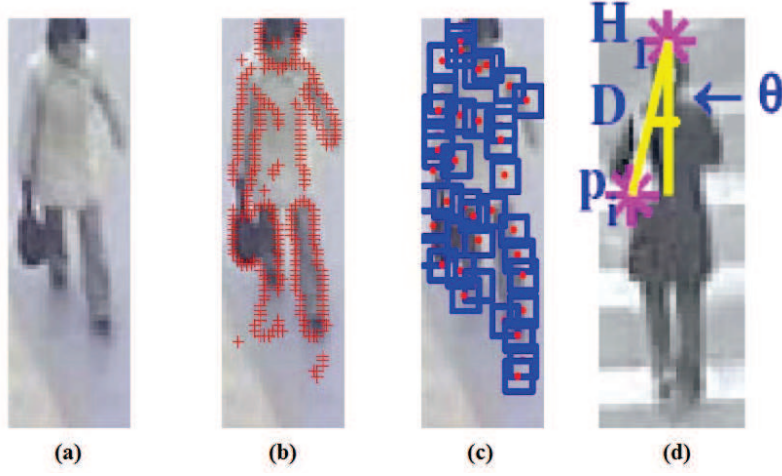


FIGURE 1.7 – (a) image originale, (b) détection des contours par l’algorithme de Canny et (c) patches et (d) les contraintes géométriques [30].

Lin et Davis [32] proposent de représenter chaque pixel de la personne par un vecteur caractéristique $s = [c, h]^t$ de dimension 4, où c est un vecteur formé par les 3 composantes RVB normalisées et h représente la coordonnée verticale du pixel dans l’image (par rapport à la hauteur de la silhouette). En effet, l’arrangement vertical des couleurs est plus stable que son homologue horizontal face aux variations de pose et de point de vue. Les vecteurs caractéristiques calculés sont ensuite agrégés en utilisant une fonction de densité de probabilité [33] (équation 1.1). Finalement, le classifieur KPPV avec la divergence de Kullback-Leibler [34] est appliqué pour l’identification des personnes.

$$f(u) = \frac{1}{N_s \sigma_1 \sigma_2 \sigma_3 \sigma_4} \sum_{i=1}^{N_s} \prod_{j=1}^4 k\left(\frac{u_j - s_{ij}}{\sigma_j}\right) \quad (1.1)$$

où $k(\cdot)$ est le noyau gaussien, les $\{\sigma_j\}_{j=1..4}$ sont les paramètres de lissage de chaque composante et N_s est le nombre de pixels utilisé dans la phase d’estimation de la densité de probabilité.

Plusieurs approches ont opté vers la fusion d’un ensemble de descripteurs afin d’améliorer la phase de caractérisation. Dans ce sens, Schwartz et Davis [35] proposent de partitionner l’image d’une personne en un ensemble de blocs non-chevauchants dont chacun est représenté par des descripteurs de couleur, de texture et de forme. Ils utilisent des histogrammes de couleur dans l’espace colorimétrique RVB normalisé, les matrices de co-occurrence et les histogrammes de gradients orientés. Dans le but de garantir les meilleures performances, ces

auteurs appliquent un outil statistique appelé « Partial Least Square », introduit dans [36], pour pondérer ces descripteurs en fonction de leur pouvoir de discrimination.

Bak et al. [37] ont introduit un modèle d'apparence basé sur les caractéristiques pseudo-Haar (Haar-like features en anglais) [28]. Une étape de filtrage est réalisée afin de sélectionner les caractéristiques les plus discriminantes à l'aide de l'algorithme AdaBoost [17]. Les caractéristiques retenues sont par la suite combinées avec des informations supplémentaires comme les couleurs dominantes des silhouettes pour renforcer la caractérisation des personnes.

Gray and Tao [38] proposent de fusionner plusieurs descripteurs en utilisant l'algorithme Adaboost [17] dans le but de produire une signature capable de discriminer les piétons. Les silhouettes des personnes sont initialement divisées en bandes horizontales qui sont caractérisées séparément par des histogrammes calculées sur des canaux de couleurs et de texture. Les canaux colorimétriques utilisés sont les composantes (R, V, B, Y, Cb, Cr, H, S) alors que les canaux de luminance sont le résultat de convolution par des filtres de Gabor [39] et de Schmid [40] appliqués sur une image en niveau de gris. L'algorithme Adaboost est finalement appliqué afin de sélectionner pour chaque bande horizontale l'ensemble des meilleurs descripteurs qui la caractérise.

Prosser et al. [41] ont étendu le travail de Gray and Tao en reformulant le problème de re-identification en un problème d'ordonnancement. Pour cela, la similarité entre une observation de référence (filmée par une caméra i) avec toutes les observations possibles (filmées par une caméra j) n'est pas mesurée par une distance comme dans [38], mais par un rang. En reprenant les mêmes descripteurs de couleur et de texture, Prosser a montré que cette approche offre de meilleures performances. Plutôt que de réaliser une phase d'apprentissage sur la totalité des exemples d'apprentissage (cas des RankSVM [42]), Prosser a opté pour l'utilisation d'un ensemble de classifieurs RankSVMs faibles appliqués chacun sur un ensemble réduit de données d'apprentissage, puis combinés pour construire un classifieur plus robuste. Cette stratégie permet d'augmenter la stabilité du classifieur, et de réduire significativement la consommation de la mémoire tout en conservant des performances élevées par rapport à l'approche RankSVM initiale.

● **Approches multi-images** Les approches de ré-identification multi-images consistent à extraire une signature à partir d'un ensemble d'observations de la même personne. Cette signature peut être calculée soit en utilisant toute la séquence, soit en sélectionnant un en-

semble « d'images clés » comme étant les plus représentatives de la séquence. Dans le cadre du suivi multi-cibles où plusieurs personnes peuvent être présentes dans le champ de vue de la même caméra, il est parfois possible que quelques observations soient attribuées par erreur à la mauvaise personne, ce qui affecte considérablement l'étape de caractérisation. Pour cela, la plupart des approches multi-images sont testées dans le cadre mono-cible où chaque personne est filmée séparément. Il est à noter aussi que des méthodes de réduction de la dimension sont généralement appliquées dans les approches multi-images afin de réduire la taille des vecteurs caractéristiques et d'accélérer le traitement.

Brid et al. [43] proposent un système d'identification des personnes par l'apparence destiné aux zones de transport public. Chaque silhouette de piéton est découpée en 10 bandes horizontales équidistantes. Chacune est représentée par les moyennes des composantes H, S et V. Afin de s'adapter aux grandes variations d'éclairage, la composante de luminance V de chaque bande est pondérée par le facteur 0.45. Les signatures sont finalement moyennées pour caractériser une séquence.

Gheissari et al. [44] proposent d'exploiter les régions des silhouettes présentant des variations spatio-temporelles similaires pour caractériser les piétons. La décomposition des silhouettes en régions est réalisée en projetant un modèle triangulaire, construit durant une phase d'apprentissage hors ligne, sur les images des personnes (figure 1.8). La correspondance entre les régions est basée sur l'information colorimétrique et structurelle. L'information colorimétrique est représentée par deux histogrammes sur les composantes H et S. Alors que l'information structurelle est basée, d'une part, sur l'orientation des contours dans les différentes régions et, d'autre part, sur les rapports entre les composantes de couleurs RVB entre deux régions de part et d'autre d'une arête du modèle triangulaire.

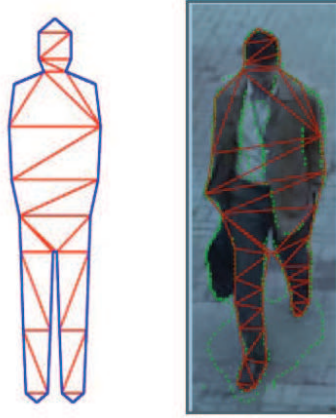


FIGURE 1.8 – Exemple illustrant la projection du modèle triangulaire sur une image de piéton [44].

Madden et al. [45] ont introduit un modèle d'apparence appelé MCSHR (Major Colour Spectrum Histogram Representation) modélisant chaque silhouette par un histogramme de ses couleurs principales. La détection des couleurs principales dans une image est réalisée en appliquant l'algorithme de segmentation K-means [46]. Pour caractériser le passage d'un piéton, n représentations MCSHR sont cumulées afin de générer une description globale nommée IMCASHR (Incrémental MCSHR). Finalement, une phase de compensation d'illumination par l'égalisation d'histogramme est appliquée sur les signatures pour palier les fortes variations d'éclairage.

Yu et al. [47] combinent des informations spatiales et colorimétriques pour décrire l'apparence d'une personne. Chaque pixel de la silhouette est modélisé par ses $\mathcal{3}$ composantes couleur dans l'espace RVB (vecteur x) et la distance qui le sépare d'un point de référence l . Une estimation par noyau basée sur la méthode de Parzen-Rozenblatt [33] est calculée pour représenter la distribution des vecteurs caractéristiques déjà extraits (équation 1.2) :

$$p(x, l) = \frac{1}{N} \sum_{i=1}^N k\left(\left\|\frac{x - x_i}{h_x}\right\|^2\right) k\left(\left\|\frac{l - l_i}{h_l}\right\|^2\right) \quad (1.2)$$

avec N le nombre de pixels dans l'image, $k(\cdot)$ une fonction de noyau, x et l respectivement les $\mathcal{3}$ composantes couleur et la position du pixel de référence et h est un paramètre de lissage.

Pour caractériser une séquence modélisant le passage d'une personne, Yu et al. [47] proposent une sélection des images clés présentant les variations de posture et de point de vue

maximales. L'appariement entre deux signatures représentant deux passages A et B dépend des distances entre les images clés comme illustré dans l'équation 1.3.

$$Similarité^{(A,B)} = \underset{i \in K^{(A)}}{median} \left[\underset{j \in K^{(B)}}{min} D(p_i^{(A)}, p_j^{(B)}) \right] \quad (1.3)$$

où $p_i^{(A)}$ représente la signature de l'image i appartenant au passage A , $p_j^{(B)}$ représente la signature de l'image j appartenant au passage B , $K^{(A)}$ et $K^{(B)}$ représentent respectivement l'ensemble des images clés relatives aux passages A et B et enfin D est la distance de Kullback-Leibler [34].

Hamdoun et al. [48] ont élaboré un système d'identification de personnes basé sur la mise en correspondance entre des points d'intérêt détectés par l'algorithme SURF [49]. L'approche proposée est composée de deux étapes : apprentissage et ré-identification. La phase d'apprentissage consiste à construire des modèles relatifs à chaque personne regroupant les points d'intérêt détectés. La ré-identification est réalisée par un système de vote où chaque point extrait est comparé aux points d'intérêts stockés dans les modèles d'apprentissage. Les points d'intérêts sont organisés selon une structure arborescente (KD-tree) afin d'accélérer la phase de mise en correspondance. Finalement, l'identité de la personne est attribuée en se référant au modèle d'apprentissage le plus sollicité.

Il est intéressant de noter que les descripteurs basés sur les points d'intérêts comme SIFT [50] ou SURF [49], qui sont largement utilisés dans la vision par ordinateur, n'ont pas trouvé autant d'intérêt dans la communauté de ré-identification de personnes. Gheissari [44] a prouvé que les approches à base de points d'intérêt sont nettement moins performantes que les approches basées couleurs.

Truong Cong et al. [51] ont testé plusieurs signatures basées sur les histogrammes de couleurs (intégrant ou non de l'information spatiale) avec différents types de normalisation de couleur pour identifier les personnes. Les vecteurs caractéristiques d'une séquence représentant le passage d'une personne dans le champ de la caméra sont projetés en 2D par une analyse spectrale [52]. L'identification des personnes est réalisée par une minimisation de la distance séparant les centres de gravités relatives aux différents passages. L'étude expérimentale réalisée dans [51] confirme que la représentation en spatiogrammes (histogrammes de couleurs intégrant de l'information spatiale) combinée avec une normalisation de couleurs

par une égalisation de l'histogramme permet d'assurer les meilleures performances.

Ce travail a été étendu dans [53] dans le but d'améliorer respectivement la phase de caractérisation et de classification. En ce qui concerne la phase de caractérisation, un nouveau modèle d'apparence nommé «color-position histogram» est élaboré. Ce modèle consiste à partitionner chaque image de personne en n blocs horizontaux chacun caractérisé par la moyenne des couleurs (figure 1.9). Afin d'améliorer la phase de classification, une nouvelle mesure de similarité est proposée exploitant, d'une part, la marge optimale séparant les différents passages de personnes et, d'autre part, les erreurs de classifications obtenues par SVM. Les résultats montrent également que la normalisation Greyworld [54] est la plus adaptée pour réduire les variations des conditions d'éclairage.

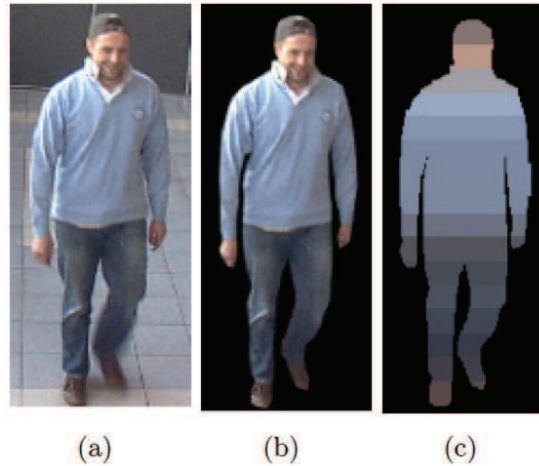


FIGURE 1.9 – L'histogramme de couleur position : (a) image originale, (b) localisation de la silhouette, (c) la répartition des couleurs dans la silhouette [53].

Farenzena et al. [55] s'intéressent aux parties saillantes dans les silhouettes extraites en appliquant les principes de perception visuelle de la symétrie et d'asymétrie. Deux axes horizontaux asymétriques séparent les trois principales régions du corps humain correspondant respectivement à la tête, le torse et les jambes. Un axe de symétrie vertical divise les régions relatives au torse et aux jambes en deux parties (figure 1.10).

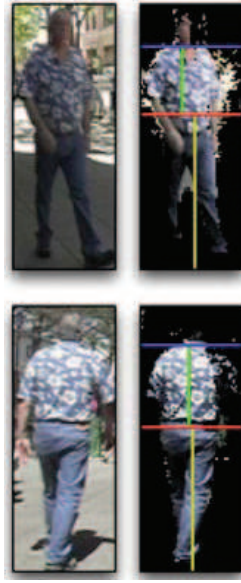


FIGURE 1.10 – Division des silhouettes en régions par les axes de symétrie et asymétrie [55].

Chaque région détectée est modélisée par i) une signature globale représentée par un histogramme HSV, ii) l'arrangement spatial des couleurs (surface, centre de gravité ainsi que la valeur moyenne pour chaque couleur) et enfin iii) l'ensemble des motifs récurrents locaux ayant une entropie élevée. Le modèle d'apparence final, permettant de caractériser un passage de personne, est obtenu en cumulant les signatures extraites à partir d'un ensemble d'images clés.

Dans le même sens, Bazzani et al. [56] présentent une nouvelle signature appelée « Histogramme Plus Epitome » (HPE) intégrant des descripteurs locaux et globaux calculés à partir d'un ensemble d'images clés. La signature globale est obtenue en moyennant les histogrammes de couleurs calculés dans l'espace HSV. Tandis que la signature locale est basée sur l'exploitation des motifs récurrents en utilisant la représentation en image d'epitomé [57]. Cette représentation est le résultat de partitionnement d'une image en un ensemble de patches intégrant les mêmes propriétés de couleur, de texture et de forme que l'image initiale (de référence). La description locale du passage d'une personne est obtenue en moyennant les patches déjà extraits.

Hirzer et al. [58] proposent un système de ré-identification de personnes fondé sur une architecture composée d'un modèle descriptif et d'un modèle discriminatif. La phase d'apprentissage du modèle discriminatif est réalisée sous assistance du modèle descriptif dont

le rôle est de produire les résultats intermédiaires. Le modèle descriptif utilisé est basé sur le partitionnement de l'image d'un piéton en 7 blocs horizontaux modélisés chacun par la matrice de covariance. Ce modèle d'apparence permet de générer pour chaque image requête une liste d'observations triées selon la similarité (les exemples présents dans la tête de la liste sont les plus proches de l'image requête). Dans ce travail, les auteurs ont opté pour la réalisation d'un apprentissage du modèle discriminatif basée sur une maximisation de la différence plutôt qu'une minimisation de la similarité. Ils utilisent alors les observations négatives présentes à la fin des listes pour sélectionner à l'aide de l'algorithme Adaboost [17] les descripteurs de Haar [28] les plus discriminants. Finalement, les descripteurs de Haar retenus sont maintenus pour caractériser l'ensemble d'image représentant un passage de personne.

1.3 Conclusion

Dans ce chapitre, nous avons présenté les différentes approches permettant le suivi et l'identification des personnes par l'apparence. Ces approches ont été divisées en deux grandes catégories : mono et multi-caméras. En mono-caméras, certains travaux se sont basés sur la minimisation de la différence entre les modèles articulés présents dans des images successives, alors que d'autres ont combiné un modèle de mouvement prédictif avec un modèle d'apparence. En multi-caméras, l'identification des personnes est principalement basée sur l'apparence en tenant compte d'une seule observation (approches mono-image) ou exploitant la redondance d'informations dans une série d'images de la même personne (approches multi-images). La fusion d'un modèle d'apparence avec des connaissances *a priori* sur le réseau est aussi utilisée pour l'identification.

Dans le chapitre suivant, nous allons introduire un nouveau descripteur de couleur exploitant l'information intra et inter-personnes à partir d'histogrammes régionaux pour caractériser l'apparence. Ce descripteur est ensuite comparé avec plusieurs descripteurs de la littérature en terme de performances. Les descripteurs les plus robustes seront finalement fusionnés afin de produire une signature discriminante capable de bien caractériser les personnes.

Chapitre 2

Identification des personnes par l'apparence

2.1 Introduction

Dans le but de proposer une signature robuste de l'apparence des personnes, ce chapitre débute par une présentation des descripteurs (couleur, texture et forme) les plus rencontrés dans la littérature. Par la suite, nous introduisons un nouveau descripteur élaboré à partir d'histogrammes régionaux par la fusion de l'information intra et inter personnes. Il permet d'assurer une invariance face aux grandes variations de pose et d'éclairage. Les différents descripteurs introduits seront ensuite comparés dans deux contextes expérimentaux : en multi-caméras (pour tester la robustesse à la ré-identification) et en mono/multi-caméras (pour valider le suivi). Les performances seront ensuite améliorées par la fusion des descripteurs sélectionnés comme les plus performants via deux variantes de l'algorithme AdaBoost. Les résultats obtenus seront finalement confrontés à l'état de l'art en terme de fiabilité pour l'identification des personnes.

2.2 Descripteurs d'apparence existants

Dans cette section, nous présentons les descripteurs les plus utilisés dans la littérature pour identifier les personnes par l'apparence. Ces descripteurs peuvent être classés en trois catégories à savoir les descripteurs de couleur, de texture et de forme.

2.2.1 Descripteurs de couleur

Les descripteurs de couleur sont sans doute les plus utilisés dans la littérature pour caractériser l'apparence des personnes [12] [25] [56] étant donné leur simplicité et leur robustesse face aux grandes variations de pose. Dans ce travail, nous avons retenu les histogrammes de couleur et les spatiogrammes présentés dans les sections suivantes.

2.2.1.1 Les histogrammes de couleur

Les histogrammes de couleur représentent la distribution des couleurs dans la région d'intérêt cadrée sur la personne. Pour réduire la dimension de l'espace des couleurs possibles (3 canaux RVB sur 8 bits), une phase de quantification est appliquée sur les images pour se limiter à N classes de couleurs. L'histogramme de couleur est finalement calculé en cumulant le nombre de pixels correspondant à chaque classe. Les histogrammes de couleur sont largement appliqués dans la littérature avec de nombreuses variantes selon l'espace de couleur utilisé (RVB, HSV, YCbCr,...) et le nombre de classes N retenu.

2.2.1.2 Les spatiogrammes

Les spatiogrammes contiennent l'information de la distribution des couleurs et leurs répartitions spatiales. Dans la littérature, ils apparaissent le plus souvent parmi les descripteurs les plus performants pour identifier les personnes [51] [59]. Dans notre étude comparative, nous avons choisi de tester le spatiogramme proposé par C. 'O Conaire vu son grand pouvoir représentatif prouvé dans [59]. A partir d'un histogramme de couleur (noté h), C. 'O Conaire calcule ensuite pour chaque classe d'intensité sa répartition spatiale par la moyenne μ et la covariance σ de la position des pixels appartenant à cette classe. Pour comparer deux spatiogrammes $S = \{h, \mu, \sigma\}$ et $S' = \{h', \mu', \sigma'\}$ contenant chacun N classes, la mesure de similarité présentée par l'équation 2.1 est appliquée.

$$similarité = \sum_{n=1}^N \psi_n \sqrt{h_n h'_n} \quad (2.1)$$

$$\psi_n = \eta \exp \left\{ -\frac{1}{2} \left(\mu_n - \mu'_n \right) \hat{\sigma}_n^{-1} \left(\mu_n - \mu'_n \right) \right\}$$

$$\hat{\sigma}_n^{-1} = \sigma_n^{-1} + \left(\sigma'_n \right)^{-1}$$

où N représente le nombre de classes d'intensité dans l'histogramme h , h_n le cardinal de la classe n , ψ représente la mesure de similarité spatiale et η est un facteur de normalisation.

2.2.2 Descripteurs de texture

Plusieurs travaux dans la littérature ont opté pour l'exploitation des descripteurs de texture pour identifier les personnes par l'apparence. Plus précisément, les matrices de co-occurrences [35] et les filtres de Schmid et de Gabor [38] [41] sont les plus fréquemment utilisés dans ce contexte. Les descripteurs sont appliqués sur une image en niveau de gris. L'image en niveau de gris est obtenue à partir des 3 canaux RVB de l'image couleur originale par l'équation 2.2. Les tests réalisés feront varier le niveau de quantification de l'image en niveau de gris ($N=8, 16$ ou 32).

$$Gris = 0.3R + 0.59V + 0.11B \quad (2.2)$$

2.2.2.1 Les matrices de co-occurrences

Les matrices de co-occurrences [35] représentent un descripteur de texture robuste comptabilisant le nombre de fois où deux pixels voisins ont une configuration identique (mêmes classes d'intensité et même déplacement). Ces matrices sont généralement calculées à partir d'images en niveaux de gris en considérant un voisinage immédiat (distance=1 et 4 directions) présentées dans la figure 2.1. Pour réduire les variations d'intensité, chaque image est quantifiée, puis les quatre matrices d'adjacences sont sommées formant une matrice générique. La phase de caractérisation des personnes consiste à calculer un histogramme à partir de cette matrice générique.

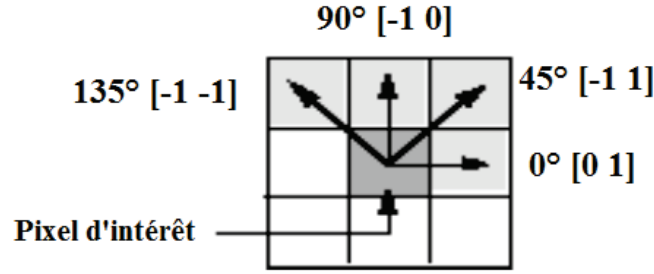


FIGURE 2.1 – Voisinage immédiat d'un pixel selon quatre directions.

2.2.2.2 Les filtres de Schmid et de Gabor

Les filtres de Schmid et de Gabor sont utilisés dans la littérature pour représenter les personnes en multi-caméras [38] [41].

L'équation paramétrique des filtres de Schmid est définie par la formule 2.3.

$$F(r, \sigma, \tau) = \frac{1}{z} \cos\left(\frac{2\pi \tau r}{\sigma}\right) e^{-\frac{r^2}{2\sigma^2}} \quad (2.3)$$

où r fixe le rayon des filtres, z est un facteur de normalisation et σ et τ deux paramètres fixés à $(2,1)$, $(4,1)$, $(4,2)$, $(6,1)$, $(6,2)$, $(6,3)$, $(8,1)$, $(8,2)$, $(8,3)$, $(10,1)$, $(10,2)$, $(10,3)$ et $(10,4)$ selon une étude expérimentale réalisée dans [38]. Le choix de ces treize filtres est motivé par leur invariance par rapport aux variations de pose. La figure 2.2 illustre les treize filtres de Schmid appliqués pour la caractérisation des personnes.

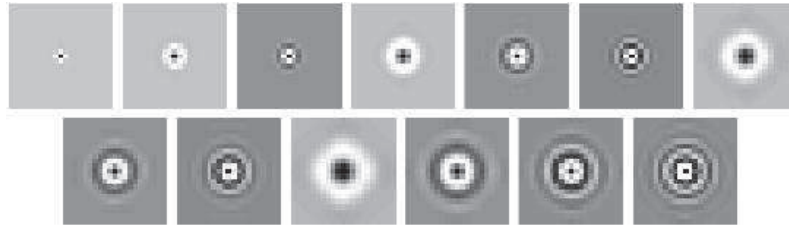


FIGURE 2.2 – Les filtres de Schmid.

L'équation paramétrique des filtres de Gabor est définie par la formule 2.4.

$$G(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp \left(-\frac{(x')^2 + \gamma^2 (y')^2}{2\sigma^2} \right) \cos \left(2\pi \frac{x'}{\lambda} + \psi \right) \quad (2.4)$$

$$x' = x \cos \theta + y \sin \theta$$

$$y' = -x \sin \theta + y \cos \theta$$

où ψ représente le déphasage, λ la longueur d'onde, θ l'orientation du filtre, σ la variance de la gaussienne et γ un paramètre spatial. Les paramètres $(\lambda, \theta, \sigma^2, \gamma)$ sont fixés à $(4, 0, 2, 0.3)$, $(8, 0, 2, 0.3)$, $(4, 0, 1, 0.4)$, $(8, 0, 1, 0.4)$, $(4, \pi/2, 2, 0.3)$, $(8, \pi/2, 2, 0.3)$, $(4, \pi/2, 1, 0.4)$ et $(8, \pi/2, 1, 0.4)$ selon la même étude expérimentale réalisée dans [38]. L'ensemble des filtres de Gabor utilisés est présenté dans la figure 2.3.

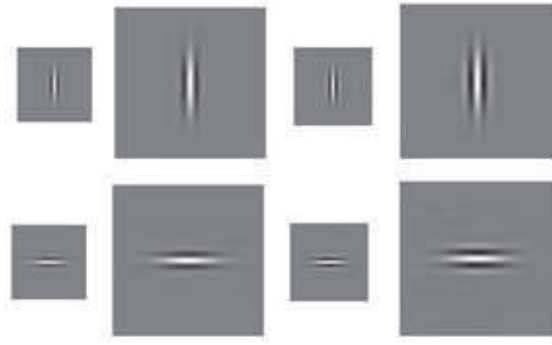


FIGURE 2.3 – Les filtres de Gabor.

Une matrice générique est obtenue en sommant les matrices générées par chaque filtre (Gabor ou Schmid). La signature utilisée pour caractériser les personnes par la texture est un histogramme d'intensité calculé à partir de la matrice générique.

2.2.3 Descripteurs de forme

Etant donné la nature très articulée du corps humain ainsi que les grandes variations de pose, les descripteurs de forme apparaissent beaucoup moins appropriés pour discriminer des personnes. Toutefois, certains descripteurs de forme globaux tels que les moments de Zernike [60] et de Hu [61] sont connus pour leur invariance face aux variations de pose. De plus, le descripteur de forme local HOG (histogrammes de gradients orientés) a été appliqué dans

le suivi des personnes en mono-caméra [16] et en multi-cameras [35]. Afin d'évaluer leurs performances, nous avons testé ces 3 descripteurs de forme.

2.2.3.1 Les histogrammes des gradients orientés (HOG)

Les histogrammes de gradients orientés analysent la distribution de l'orientation des contours régionaux dans l'image. Une phase de prétraitement est réalisée en appliquant un filtre Gaussien [62] pour réduire le bruit, avant d'extraire les contours de l'image par le filtre de Canny [31]. Dans ce travail, nous avons testé les HOG après découpage des images de personnes en cellules non-chevauchantes. Sur chacune d'elle, un histogramme d'orientation des contours est calculé selon 5 directions en appliquant les masques illustrés en figure 2.4.



FIGURE 2.4 – Masques d'orientation du contour.

2.2.3.2 Les moments de Zernike et de Hu

Les moments de Zernike [60] et de Hu [61] sont des descripteurs de forme globaux capables de caractériser les formes complexes en assurant une invariance par rapport aux transformations géométriques (translation, rotation et changement d'échelle).

Les moments de Zernike représentent les projections de fonctions polynomiales complexes formant un ensemble orthogonal complet sur le disque unitaire en coordonnées polaires. En général, les basses fréquences dans les images sont encodées dans les moments d'ordres inférieurs. Dans ce travail, nous nous sommes basés sur des études existantes [63] [64] montrant que les 12 premiers moments de Zernike sont les plus robustes pour caractériser la forme.

Les moments de Hu sont calculés à partir des moments normalisés jusqu'à l'ordre 3 pour caractériser les images des personnes.

2.3 Mesure de similarité proposée

Les histogrammes régionaux sont largement utilisés dans la littérature pour identifier les personnes par l'apparence [18] [35] [55]. Ils sont capables de s'adapter à la nature très articulée du corps humain. Ils sont aussi robustes aux occultations partielles et aux grandes variations de pose. Cependant, comme tout descripteur de couleur, les histogrammes régionaux restent sensibles aux fortes variations d'éclairage et ceci malgré l'application d'une étape de normalisation. Nous avons alors décidé de proposer une mesure de similarité adaptée aux histogrammes régionaux. Cette mesure de similarité intègre l'information intra-personne pour réduire l'impact des variations d'éclairage et assurer ainsi une bonne phase d'appariement entre personnes.

La première étape consiste à diviser horizontalement l'image de la personne en 4 blocs équidistants qui respectent les régions d'articulation les plus importantes du corps humain à savoir les épaules, la hanche et les genoux puis à calculer l'histogramme de couleur régional de chaque bloc (figure 2.5). Cette division permet de révéler la variabilité de l'apparence des différents blocs. Pour chaque personne, deux vecteurs sont calculés. Le premier vecteur de taille 6 contient les distances de Bhattacharyya [65] (équation 2.6) entre les histogrammes de couleurs des 4 bandes pour la même personne. Il représente la différence intra-classe (figure 2.5.a). Un second vecteur de taille 4 contient les distances de Bhattacharyya entre deux histogrammes de couleurs de la même bande pour deux personnes et deux poses différentes. Il représente la différence inter-classes (figure 2.5.b). L'information intra et inter personnes est destinée à gérer respectivement les variations d'éclairage et de pose.

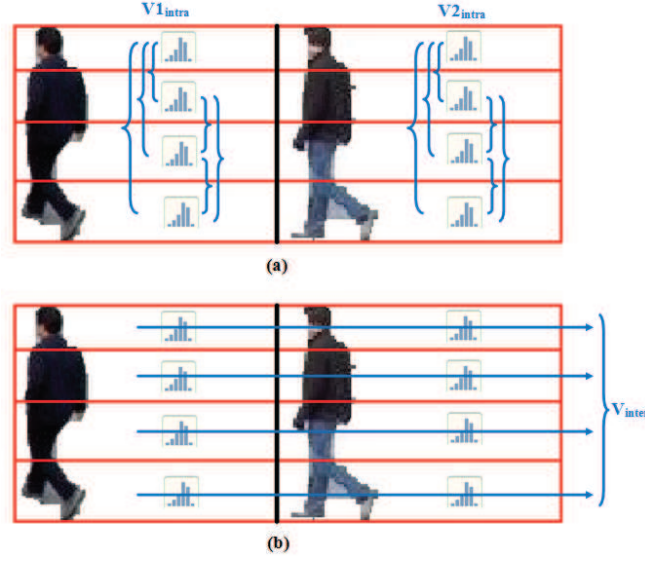


FIGURE 2.5 – Descripteur proposé.

La mesure de similarité entre deux silhouettes (1 et 2) dépend de l'information intra et inter personnes. Elle est calculée selon la formule 2.5.

$$Similarité = W_{inter} [1 - std(V_{inter})] + W_{intra} [1 - std(V1_{intra} - V2_{intra})] \quad (2.5)$$

$$Bhattacharyya = \sum_{i=1}^{Classes} \left(\sqrt{H1_i \times H2_i} \right) \quad (2.6)$$

avec V_{inter} est le vecteur inter-personnes, $V1_{intra}$ et $V2_{intra}$ sont les vecteurs intra-personnes, std est l'écart type, W_{inter} et W_{intra} représentent respectivement les poids assignés à l'information intra et inter personnes.

La mesure de similarité proposée (formule 2.5) offre l'avantage de s'adapter à plusieurs contextes expérimentaux en faisant varier les poids W_{inter} et W_{intra} . En effet, si les variations d'éclairage sont plus importantes que les variations de pose alors W_{intra} doit être supérieur à W_{inter} et vice-versa. Etant donné que nous étudions simultanément l'impact des variations de pose et des conditions d'éclairage, nous avons attribué des poids identiques à l'information intra et inter dans les différents tests réalisés soit $W_{intra} = W_{inter} = 0.5$.

Comme nous l'avons vu dans le paragraphe 2.2.1 consacré aux histogrammes de couleurs,

plusieurs espaces (RVB, HSV, YCbCr) et niveau de quantification (8, 16, 32) seront testés. Les transformations des espaces colorimétriques sont réalisées par les équations suivantes :

– De RVB à HSV :

$$\begin{aligned}
 V_{HSV} &= \max(R, V, B) \\
 S_{HSV} &= \frac{[V_{HSV} - \min(R, V, B)]}{V_{HSV}} \quad \text{si } V_{HSV} \neq 0, 0 \text{ sinon} \\
 H_{HSV} &= \begin{cases} \frac{V-B}{6 \times S_{HSV}} & \text{si } V_{HSV} = R \\ 0.5 + \frac{(B-R)}{6 \times S_{HSV}} & \text{si } V_{HSV} = V \\ 0.66 + \frac{(R-V)}{6 \times S_{HSV}} & \text{si } V_{HSV} = B \end{cases}
 \end{aligned} \tag{2.7}$$

– De RVB à YCbCr

$$\begin{aligned}
 Y &= \text{Arr} \left(\frac{(77 \times R + 150 \times V + 29 \times B)}{255} \right) + 16 \\
 Cb &= \text{Arr} \left(\frac{(-44 \times R - 87 \times V + 131 \times B)}{255} \right) + 128 \\
 Cr &= \text{Arr} \left(\frac{(131 \times R - 110 \times V - 21 \times B)}{255} \right) + 128
 \end{aligned} \tag{2.8}$$

avec Arr (x) est une fonction retournant l'entier le plus proche de x.

2.4 Etude comparative entre les descripteurs d'apparence

L'objectif de cette section est de sélectionner les descripteurs d'apparence les plus robustes en termes d'identification des personnes. Pour cela, nous avons réalisé deux tests. Le premier exploite la base VIPeR pour faire une sélection sur les descripteurs dans le contexte d'une ré-identification. Le second test considère les bases PETS09 et CVLAB pour prendre en compte un contexte expérimental plus large. Ce deuxième test n'est appliqué qu'aux meilleurs descripteurs de chaque catégorie issus du premier test.

2.4.1 Performances des descripteurs sur la base VIPeR

Dans cette section, nous présentons la base d'images VIPeR [66], puis le protocole d'évaluation des descripteurs. Tous les résultats obtenus sont analysés sous formes de courbes «cumulative matching characteristic» (CMC) comme dans [38] [41].

2.4.1.1 La base VIPeR

Pour évaluer la performance de chaque descripteur en termes de ré-identification des personnes, nous avons utilisé la base VIPeR [66] qui contient 632 personnes photographiées chacune selon deux points de vue avec de grandes variations d'éclairage entre les deux caméras. Toutes les images sont normalisées en dimensions sur un rectangle de 128×48 pixels. Quelques paires d'images de la base VIPeR sont illustrées dans la figure 2.6.



FIGURE 2.6 – Quelques exemples de la base VIPeR : chaque colonne est une paire d'image représentant la même personne.

Afin de réduire les fortes variations d'éclairage, nous avons opté pour une étape de normalisation des couleurs avant d'extraire les différents descripteurs de couleur et de texture. Plusieurs méthodes de normalisation de couleurs ont été proposées dans la littérature, à savoir Greyworld [54], Affine [67], RVB-rang [68], Coordonnées réduites [69] et Compréhensive [70]. Dans ce travail, nous nous sommes basés sur l'étude comparative réalisée dans [53] qui montre que la normalisation Greyworld est la plus adaptée. Cette normalisation consiste à diviser les coordonnées RVB de chaque pixel par la moyenne de la composante correspondante (équation 2.9) calculée sur toute l'image.

$$R = \frac{R}{\text{moyenne}(R)} \quad V = \frac{V}{\text{moyenne}(V)} \quad B = \frac{B}{\text{moyenne}(B)} \quad (2.9)$$

2.4.1.2 Protocole d'évaluation

La phase d'extraction des descripteurs est réalisée sur chaque image de personne avec différents paramètres. Pour chaque catégorie de descripteurs (couleur, texture et forme), différentes configurations ont été testées en faisant varier l'espace de couleur et le nombre de classes de quantification pour les descripteurs de couleurs ou pour les descripteurs de texture, et enfin la taille des blocs pour le descripteur HOG. Le protocole de test complet est décrit dans le tableau 2.1.

		Espace de couleur			Quantification			Taille des blocs			Nombre de tests
Descripteurs		RGB	YCbCr	HSV	8	16	32	4	16	32	
Couleur	Histogramme de couleur	☒	☒	☒	☒	☒	☒				9
	Spatigrammes	☒	☒	☒	☒	☒	☒				9
	Histogrammes régionaux	☒	☒	☒	☒	☒	☒				9
Texture	Matrices de co-occurrences				☒	☒	☒				3
	Filtres de Schmid				☒	☒	☒				3
	Filtres de Gabor				☒	☒	☒				3
Forme	HOG							☒	☒	☒	3
	Moments de Zernike										1
	Moments de Hu										1
											Total=41

Tableau 2.1 – Différents configurations testées pour chaque descripteur d'apparence.

En considérant toutes les configurations présentées dans le tableau 2.1, un ensemble de 41 vecteurs caractéristiques sont extraits de chaque image. Les vecteurs caractéristiques calculés sont normalisés en divisant par la somme du vecteur dans le but d'avoir des mesures de similarité entre 0 et 1. La mesure de similarité inter-images appliquée pour les descripteurs de la littérature est 1- la distance de Bhattacharyya [65], alors que pour les histogrammes régionaux, la mesure de similarité est celle présentée dans l'équation 2.5. Il est à noter que la distance de Bhattacharyya exprime la différence et que 1- la distance de Bhattacharyya représente la similarité. Pour réduire le temps de calcul, nous avons utilisé dans ce test que la moitié de la base VIPeR soit 316 paires de piétons tirées au hasard. Au total, 41×316^2 mesures de similarité ont été calculées pour comparer les descripteurs.

Les résultats sont présentés sous la forme de courbes de précision CMC, largement utilisées dans la littérature pour évaluer la ré-identification [38][41]. Elles calculent le taux de bonnes identifications en fonction du rang d'apparition des exemples d'apprentissage par rapport aux exemples de test. Une image de test est correctement classée à l'ordre 1 lorsque sa similarité est maximale avec un exemple d'apprentissage de même identité.

Pour cela, les mesures de similarité relatives à chaque configuration sont stockées dans une matrice (D) de taille 316×316 (les 316 images de la caméra 1 sont placées en lignes et les 316 images de la caméra 2 sont placées en colonnes). La diagonale représente la similarité entre une paire d'images correspondant à une même personne et devrait contenir les similarités maximales de chaque ligne. Pour chaque ligne (i) de la matrice (D) , $d(i)$ représente le nombre d'appariements ayant une similarité supérieure à la diagonale. Pour un rang n donné, le taux de reconnaissance TR , normalisé entre 0 et 1 , est le rapport entre le nombre de bonnes identifications (quand $d(i) \leq n$) et le nombre de lignes dans la matrice similarité (D) (équation 2.10).

$$TR = \frac{\sum_{i=1}^N [d(i) \leq n]}{N} \quad (2.10)$$

La figure 2.7 illustre les performances de chaque descripteur en faisant varier les paramètres décrits dans le tableau 2.1. Le descripteur le plus robuste est celui qui présente une courbe CMC se rapprochant le plus rapidement de 1 dès les premiers rangs (coin haut-gauche).

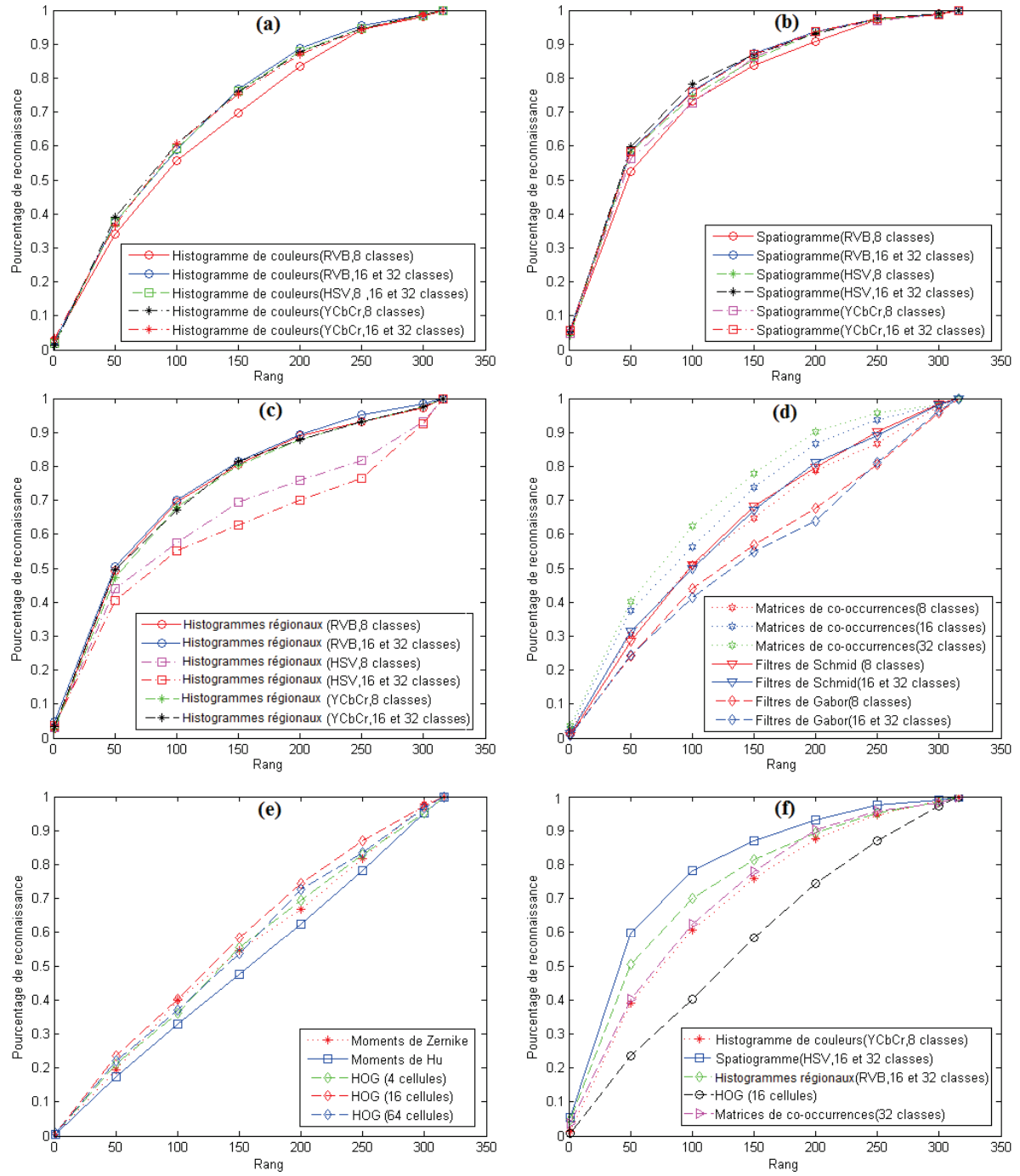


FIGURE 2.7 – Courbes CMC pour les descripteurs de couleur (a, b, c), les descripteurs de texture (d), les descripteurs de forme (e) et les meilleurs descripteurs pour chaque catégorie (f).

2.4.1.3 Analyse des Performances et interprétation des résultats

Pour les histogrammes de couleur et les spatiogrammes (Figure 2.7 (a) et (b)), nous pouvons constater que l'espace colorimétrique ainsi que la phase de quantification affectent peu les taux de reconnaissance. En effet, ces deux descripteurs sont invariants à l'espace de couleur utilisé puisqu'ils représentent la distribution des intensités dans l'intégralité de l'image (descripteurs globaux). Cela réduit considérablement l'impact des variations d'intensité locales dans les images des piétons et par conséquent nous obtenons des histogrammes de couleurs très similaires pour la même personne quelque soit l'espace de couleur choisi (RVB , YCbCr ou HSV). Ces descripteurs sont également invariants à la phase de quantification (avec 8, 16 ou 32 classes d'intensité) car les couleurs portées par une personne sont généralement homogènes et uniformes.

Toutefois, les performances des histogrammes régionaux avec la mesure de similarité proposée (figure 2.7 (c)) en utilisant l'espace de couleur HSV sont inférieures à celles obtenues avec les espaces RVB ou YCbCr. La grande variabilité de la distribution d'intensité dans cet espace colorimétrique (équation 2.7) augmente la différence entre les histogrammes régionaux de la même personne.

Le graphe (d) de la figure 2.7 montre que les matrices de co-occurrences sont les plus adaptées à l'identification des personnes car elles prennent en considération toutes les directions (horizontale, verticale et deux diagonales). Les performances des filtres Schmid et Gabor sont légèrement inférieures car ils exploitent un ensemble réduit de filtres afin d'assurer un bon compromis efficacité/temps de calcul.

Selon le graphe (e) de la figure 2.7, nous pouvons constater que les descripteurs de forme sont les moins performants étant donné les fortes variations de pose existantes dans la base VIPeR. Il est à noter aussi que le descripteur HOG est légèrement plus performant comparé aux moments de Zernike et de Hu. En effet, certains blocs dans les silhouettes des personnes peuvent garder la même forme de contour, et ceci malgré des changements de pose (figure 2.8). Nous pouvons conclure alors que les descripteurs de forme locaux ou régionaux (tel que HOG) sont les plus adaptées pour identifier les personnes en multi-caméra comparés aux descripteurs de forme globaux.



FIGURE 2.8 – (a) images originales de la même personne, (b) quelques blocs ayant des contours similaires malgré la grande variation de pose.

On peut noter sur le graphe (f) de la figure 2.7 un bon comportement des histogrammes régionaux avec la mesure de similarité proposée, dont les performances les placent en seconde position juste après les spatiogrammes.

2.4.2 Performances des descripteurs sur les bases PETS09 et CV-LAB

Dans cette section, nous présentons les deux bases de piétons utilisées pour tester la robustesse des meilleurs descripteurs aux changements de pose et conditions d'éclairage (Ré-identification), mais également leurs performances pour le suivi des personnes par une même caméra. Les descripteurs retenus (figure 2.7 graphe (f)) sont les histogrammes de couleurs (YCbCr, 8 classes), les spatiogrammes (HSV, 16 classes), les histogrammes régionaux (RVB, 16 classes), la matrice de co-occurrences (32 classes) et les histogrammes de gradients orientés (16 cellules).

2.4.2.1 Les bases PETS09 et CVLAB

Pour évaluer simultanément les performances des descripteurs retenus en termes de suivi mono-caméra et de ré-identification des personnes en multi-vues, nous avons utilisé les deux bases de piétons PETS2009 [71] et CVLAB [72]. Ces deux bases contiennent plusieurs séquences de personnes filmées par différentes caméras. Deux séquences filmées par 3 caméras différentes et extraites chacune à partir d'une base sont présentées dans les figures 2.9 et 2.10. Elles intègrent des grandes variations de pose, de condition d'éclairage, d'échelle, des occultations partielles importantes mais également une ressemblance d'apparence des personnes.



FIGURE 2.9 – Extrait de la base PETS09.

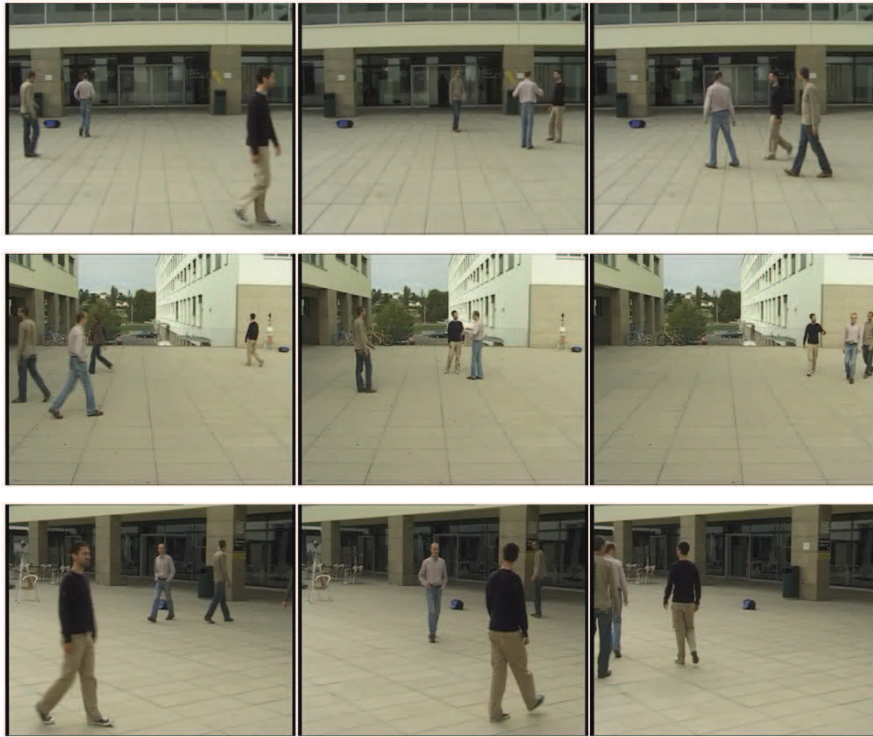


FIGURE 2.10 – Extrait de la base CVLAB.

2.4.2.2 Extraction des silhouettes

Avant d'évaluer les performances des descripteurs, nous avons procédé à une étape d'extraction des silhouettes. Ces deux bases PETS09 et CVLAB intègrent plusieurs vidéos où les piétons sont les seuls objets mobiles dans la scène et ils sont toujours en position debout. Afin de détecter les personnes, nous avons choisi d'éliminer l'arrière plan fixe. Le processus complet de localisation des personnes est composé des étapes suivantes :

Elimination de l'arrière plan Nous avons appliqué la méthode proposée par Kelly et al. [73] qui démontrent qu'elle est plus performante que le mélange de gaussiennes ou la différence d'images. Cette méthode construit un modèle d'arrière plan où chaque pixel est représenté par N couleurs qui sont itérativement mises à jour selon une fréquence F en se référant aux $\mathcal{B}$ images de différences les plus récentes. Un pixel est considéré comme faisant partie de l'arrière-plan si l'écart d'une de ses composantes à l'arrière-plan est inférieur à un seuil T . Pour initialiser l'algorithme, un paramètre P précise le pourcentage de pixels de l'arrière plan par rapport à la dimension de l'image. Sur la base de nos observations expérimentales, nous

avons choisi de fixer les paramètres N , F , T , P respectivement à 5, 0.99, 20 et 0.85.

Etant donné que les variations d'éclairage et des ombres sont assez fréquentes dans les vidéos filmées en extérieur, Kelly et al. [73] proposent de filtrer les pixels détectés sur les objets mobiles par un seuillage selon l'équation 2.11.

$$D = C \times Dist + \left| \log \left(\frac{V_{AP}}{V_{Actuel}} \right) \right| \quad (2.11)$$

où $C = 18$ est une constante définie expérimentalement dans [73], $Dist$ est la distance euclidienne entre le pixel courant et le pixel d'arrière-plan le plus fréquent, V_{AP} est la luminosité de l'arrière-plan et V_{Actuel} est la luminosité du pixel courant. Si $D \leq 0.5$, le pixel est considéré comme marqué par un changement d'éclairage (ombre/lumière) et par conséquent il sera éliminé de l'avant plan. La figure 2.11 illustre les images originales ainsi que les résultats d'extraction de l'arrière plan respectivement pour les deux bases utilisées.

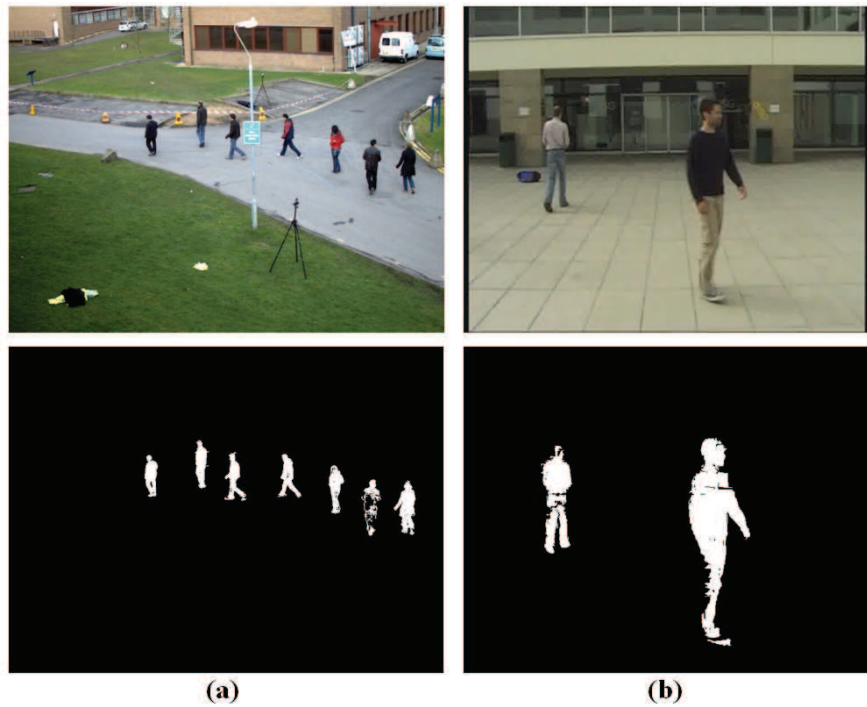


FIGURE 2.11 – Exemples d'extraction de l'arrière plan : (a) base PETS09 et (b) CVALAB.

Application des opérateurs morphologiques Comme les silhouettes ainsi obtenues sont incomplètes et trouées (figure 2.11), nous avons appliqué des opérateurs morphologiques :

dilatation puis érosion (fermeture) avec un filtre de taille 9×9 pixels sur les images binaires. Le résultat est présenté dans la figure 2.12.

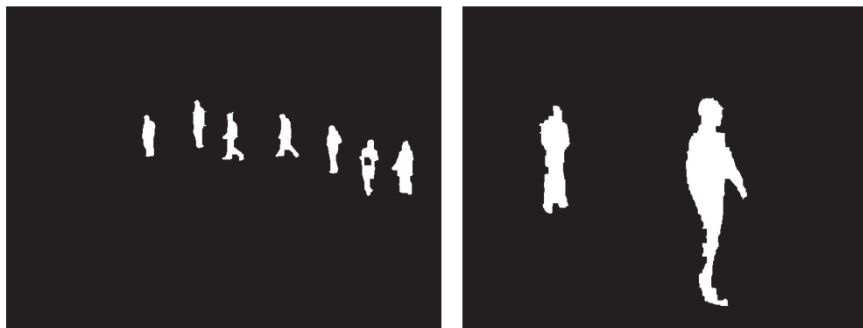


FIGURE 2.12 – Résultat des opérateurs morphologiques appliqués sur les mêmes images que celles de la figure 2.11.

Détection des personnes Dans les deux bases PETS09 et CVLAB, les personnes filmées sont toujours en position debout. Pour cette raison, nous avons adopté une règle morphologique simple pour les détecter : une région appartenant à l'avant plan dont la hauteur est supérieure au double de sa largeur est considérée comme un piéton.

Normalisation spatiale et colorimétrique La dernière étape de prétraitement consiste à effectuer une normalisation spatiale des silhouettes pour palier aux variations d'échelle, ainsi qu'une normalisation des couleurs pour assurer une invariance face aux changements d'éclairage. La normalisation spatiale des silhouettes en 128×48 pixels est réalisée par une interpolation bicubique. Alors que la normalisation des couleurs est basée, comme dans le premier test (section 2.4.1.1), sur la méthode Greyworld [54].

Nous avons sélectionné un total de 606 images de 4 personnes dans la base PETS09 et 900 images de 3 personnes dans la base CVLAB. Pour une même caméra, les images successives permettent d'évaluer les descripteurs en termes d'identification mono-caméra. La ré-identification est évaluée en appariant deux images de la même personne filmées par deux caméras différentes, mais également deux images de deux personnes différentes. La figure 2.13 illustre quelques images de personnes extraites des deux bases PETS09 et CVLAB suite à l'application du processus de prétraitement.

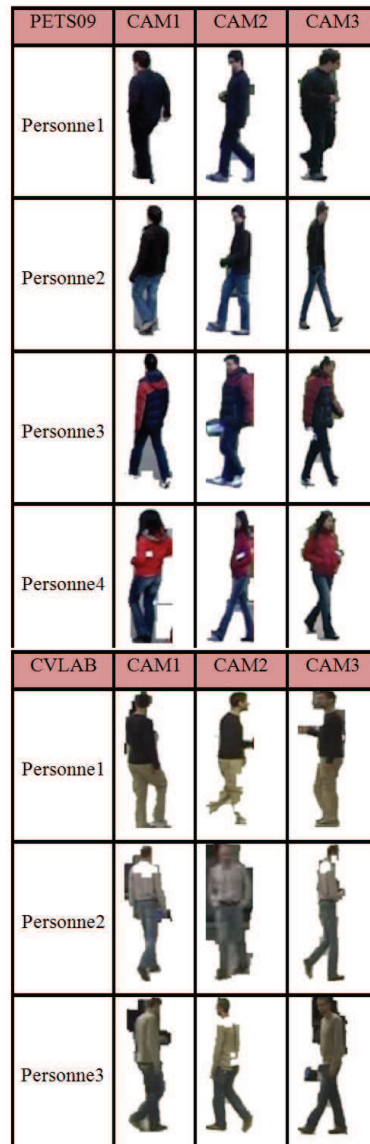


FIGURE 2.13 – Echantillons d'images de personnes extraites respectivement de la base PETS09 et CVLAB.

2.4.2.3 Protocole d'évaluation

Le pouvoir de discrimination des 5 descripteurs sélectionnés (figure 2.7 graphe (f)) est mesuré par appariements entre deux images successives à l'instant t et $t+1$ issues de la même caméra (identification mono-caméra) ou par appariements des images issues de caméras différentes (ré-identification). Le protocole de test complet est composé des étapes suivantes.

Pour chaque descripteur, une matrice de similarité (S) est d'abord calculée en appariant les mêmes séquences de silhouettes placées respectivement en lignes et en colonnes (figure 2.14). Etant donné que cette matrice est symétrique, nous n'avons considéré que les valeurs qui sont inférieures à la diagonale. La matrice de similarité (S) est calculée par l'équation 2.12.

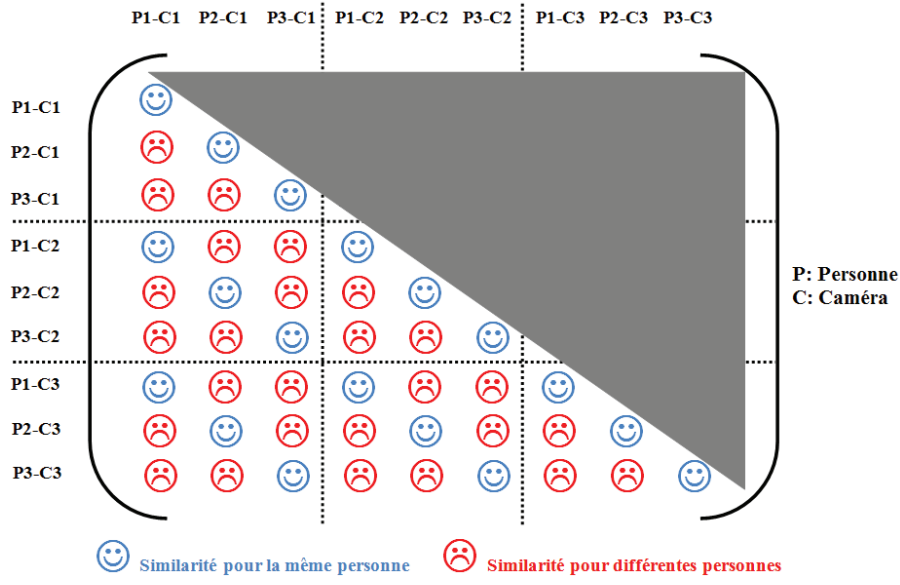


FIGURE 2.14 – Calcul de la matrice d'autosimilarité.

$$S(i, j) = \frac{1}{T} \sum_{t=1}^T [1 - \text{Dist}(V_{i,t}, V_{j,t+1})] \quad (2.12)$$

$$T = \min(nb_i, nb_j)$$

où i et j représentent les séquences placées en lignes respectivement en colonnes, $V_{i,t}$ représente le vecteur caractéristique calculé à partir de la silhouette apparaissant à l'instant t de la séquence i , $V_{j,t+1}$ représente le vecteur caractéristique calculé à partir de la silhouette apparaissant à l'instant $t + 1$ de la séquence j , Dist est la distance de Bhattacharyya [65] entre deux vecteurs caractéristiques, nb_i et nb_j représentent le nombre d'images dans la séquence i respectivement j et T est le nombre d'appariement réalisé pour comparer deux séquences.

Ensuite, nous calculons à partir de chaque matrice de similarité, 3 critères d'évaluation qui sont la similitude moyenne pour la même personne ($\hat{S}P$), la similitude moyenne pour

différentes personnes ($\hat{DP}$) et le nombre de fausses alarmes FA :

- ($\hat{SP}$) est la moyenne des valeurs de similarité quand les séquences i et j correspondent à la même personne (les appariements en *bleu* dans la figure 2.14).
- ($\hat{DP}$) est la moyenne des valeurs de similarité quand les séquences i et j représentent des personnes différentes (les appariements en *rouge* dans la figure 2.14).
- Pour calculer le nombre de fausses alarmes FA , nous distinguons les sous-blocs de la matrice de similarité associés à un couple de caméras (C_i , C_j). Puis, un test booléen vérifie si la similarité pour deux personnes différentes est toujours inférieure à celle correspondant à la même personne. Dans chaque bloc, il existe un seul appariement par ligne relatif à la même personne qui devrait correspondre à la similarité maximale. Le test sera donc négatif chaque fois que la similarité pour deux personnes différentes sera supérieur à cet appariement. Le nombre de fausses alarmes FA est le cumul des résultats de tous les tests booléens négatifs.

Les trois critères d'évaluation ($\hat{SP}$), ($\hat{DP}$) et FA sont utilisés pour mesurer la fiabilité des descripteurs. Un descripteur robuste doit accentuer l'écart entre ($\hat{SP}$) et ($\hat{DP}$) et réduire le nombre de fausses alarmes. Le tableau 2.2 résume tous les résultats obtenus.

Descripteurs	$\hat{SP}$		$\hat{DP}$		Fausses alarmes	
	PETS09	CVLAB	PETS09	CVLAB	PETS09	CVLAB
Histogramme de couleur (YCbCr, 8 classes)	98.91	99.54	97.73	98.24	2	2
Spatiogramme (HSV, 16 classes)	87.42	93.87	76.54	87.37	2	2
Histogrammes régionaux (RVB, 16 classes)	97.38	95.67	91.94	91.33	0	1
Matrices de co-occurrences (32 classes)	98.00	99.05	93.23	96.65	3	2
HOG (16 cellules)	73.49	80.11	69.97	70.65	11	4

Tableau 2.2 – Performances des descripteurs retenus en termes d'identification mono-caméra et de ré-identification.

Les résultats présentés dans le tableau 2.2 confirment ceux obtenus dans la première étude comparative (figure 2.7 (f)). Les descripteurs de couleur sont les plus adaptés pour caractériser l'apparence des personnes. Les couleurs sont peu affectées par les changements de pose et la phase de normalisation permet de minimiser les effets des changements d'éclairage. De plus, les descripteurs de couleur qui intègrent également de l'information spatiale, comme les

spatiogrammes et les histogrammes régionaux sont plus robustes. Pour ces deux descripteurs l'écart entre $(\hat{SP})$ et $(\hat{DP})$ est respectivement de 11% et 6% pour la base PETS09 et de 7% et 5% pour la base CVLAB. Tandis que pour les histogrammes de couleur, cet écart est de seulement de 1% pour les deux bases.

Etant donné que les descripteurs de couleur et de texture sont les plus performants pour caractériser les personnes par l'apparence, nous proposons dans la section suivante de les fusionner par l'algorithme AdaBoost pour améliorer les performances.

2.5 Fusion de descripteurs

Les deux études comparatives précédentes montrent qu'un seul descripteur reste toutefois insuffisant pour caractériser les personnes par l'apparence. Deux stratégies sont envisageables pour essayer d'améliorer les performances. La première consiste à renforcer la classification en appliquant un classifieur de type SVM ou réseau de neurones. La seconde consiste à améliorer la phase de caractérisation en fusionnant des descripteurs robustes. La stratégie de fusion par l'algorithme AdaBoost a déjà été fréquemment utilisée pour identifier les personnes par l'apparence [37][38][58]. Nous avons donc choisi de la tester.

Deux configurations de descripteurs ont été testées à savoir la fusion des descripteurs de couleurs seuls et la fusion des descripteurs de couleurs et de texture. Ces deux configurations sont évaluées sur la base VIPeR. Le choix de repartir avec cette base est principalement dû à la complexité du contexte expérimental qu'elle offre : appariements entre deux observations temporellement éloignées et en présence de fortes variations de pose et d'éclairage. En effet, si un modèle d'apparence est performant dans ce contexte expérimental, il est évident qu'il le sera aussi dans le cadre d'une identification mono-caméra. De plus, plusieurs travaux de la littérature sont évalués sur la base VIPeR, ce qui nous offre la possibilité de se comparer avec eux.

2.5.1 L'algorithme AdaBoost

AdaBoost (Adaptive Boosting) [17] est la variante la plus utilisée du Boosting. Le principe est de construire un classifieur fort par une combinaison linéaire de classifieurs faibles. L'algorithme d'Adaboost s'appuie sur un apprentissage pour définir les poids affectés à chaque descripteur (classifieur faible). A la fin, Adaboost combine les classifieurs faibles par un vote

pondéré pour en déduire un classifieur fort. L'algorithme Adaboost est détaillé dans l'algorithme 2.1.

Algorithme 2.1 Algorithme AdaBoost

Initialisation

- N , exemples d'apprentissage labélisés (x_i, y_i) , avec x_i une paire d'images de personne et $y_i \in \{-1, 1\}$; 1 si les deux images appartiennent à la même personne, -1 sinon.
- Initialisation de la distribution des poids pour les exemples d'apprentissage $D_i = \frac{1}{N}$ pour $i = 1..N$.

Processus itératif d'apprentissage

Pour $t = 1..T$ % parcourt des classifieurs faibles

- Trouver le classifieur (descripteur) faible c_t minimisant l'erreur :

$$\gamma_t : c_t = \underset{c_t \in C}{\operatorname{argmin}} \gamma_t = \sum_{i=1}^N D_t(i) [y_i c(x_i)]$$

% Avec $c(x_i)$ la réponse du classifieur sélectionné concernant l'exemple x_i et $[y_i c(x_i)]$ un test booléen qui retourne vrai (1) si y_i et $c(x_i)$ sont identiques et faux (-1) sinon. %

- Si $\gamma_{1..t} < 0$ Arrêt du processus itératif
% Arrêt du processus itératif quand l'erreur de tous les classifieurs restants est < 0 .
- Calculer le poids du descripteur sélectionné $\alpha_t = \frac{1}{2} \ln \frac{1+\gamma_t}{1-\gamma_t}$
- Mettre à jour de la distribution des poids des exemples d'apprentissage

$$D_{t+1}(i) = \frac{1}{Z_t} D_t(i) \exp(-\alpha_t [y_i c(x_i)])$$
 % Avec Z_t est un facteur de normalisation assurant que $\sum D_t(i) = 1$.
- Ajouter la paire {classifieur : c_t poids : α_t } dans l'ensemble final.

Fin

La base VIPeR est aléatoirement divisée en deux portions égales contenant chacune 316 paires de personnes : une pour l'apprentissage et l'autre pour le test. L'algorithme AdaBoost permet d'estimer les poids relatifs aux descripteurs sur les données d'apprentissage. Ces poids serviront à pondérer les descripteurs retenus sur les données de test afin de mesurer les performances.

En se référant à la figure 2.7, nous avons retenu pour la fusion les descripteurs de couleurs : les histogrammes de couleurs (RVB, 16) (YCbCr, 8) et (HSV, 8), les spatiogrammes (RVB, 16) (YCbCr, 16) et (HSV, 16), et les histogrammes régionaux (RVB, 16) (YCbCr, 16). Puis, nous avons fait un nouveau test en adjoignant les deux descripteurs de textures les plus performants à savoir les matrices de co-occurrences (32) et les filtres de Schmid (16).

La figure 2.15 présente les poids, normalisés à 100%, assignés aux descripteurs par l'algorithme AdaBoost respectivement pour la fusion des descripteurs de couleur uniquement (a) et pour la fusion des descripteurs de couleur et de texture (b).

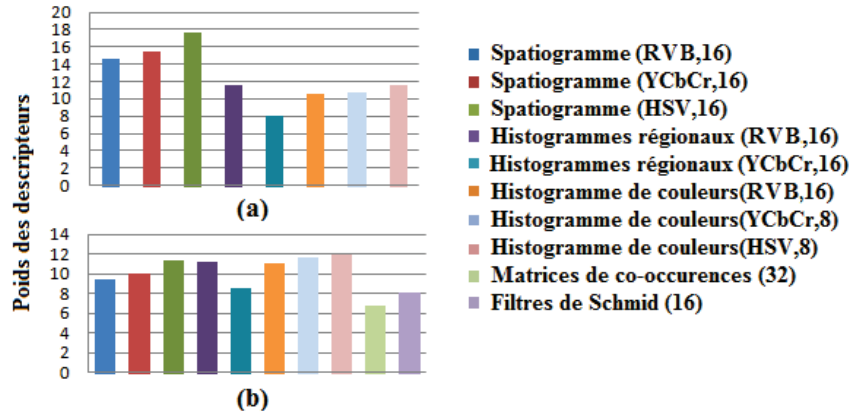


FIGURE 2.15 – Poids des descripteurs en pourcent en utilisant : (a) uniquement les descripteurs de couleur et (b) les descripteurs de couleur et de texture.

2.5.2 Protocole de test

Pour évaluer les performances de la fusion, nous avons appliqué le protocole de test déjà introduit dans la section 2.4.1.2, basé sur la mesure de précision CMC. La seule différence concerne la matrice de similarité finale obtenue par combinaison linéaire des matrices de similarité de chaque descripteur pondérées par leur poids respectif (équation 2.13). Les performances des deux variantes de fusion (couleur, couleur+texture) sont comparées en premier lieu à ceux obtenus par le meilleur descripteur (figure 2.16), puis à quelques travaux de la littérature dans la section suivante.

$$MSF = \sum_{i=1}^N \alpha_i \times D_i \quad (2.13)$$

où MSF est la matrice de similarité finale, D_i et α_i représentent respectivement la matrice de similarité ainsi que le poids relatif au $i^{ème}$ descripteur sélectionné.

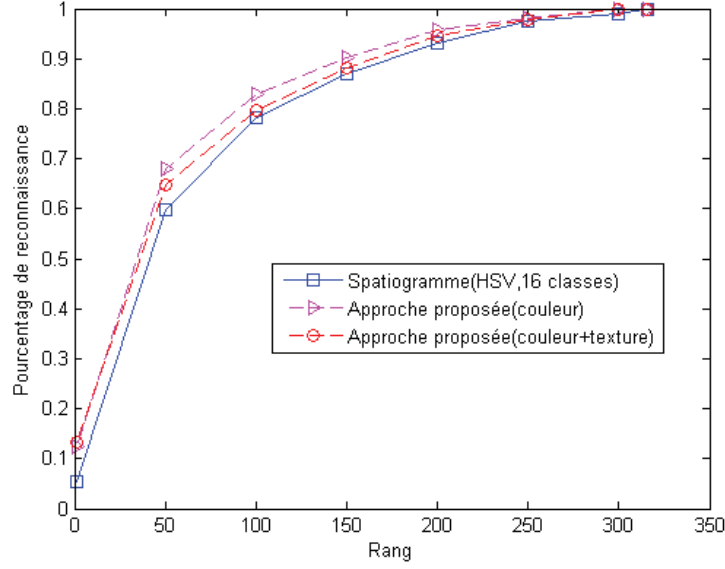


FIGURE 2.16 – Apport de la fusion sur les taux de reconnaissance.

La figure 2.16 montre que la fusion des descripteurs améliore les performances par rapport au meilleur descripteur : spatiogrammes (HSV,16). Pour le rang 1, les taux de reconnaissance sont améliorés de 7%, soit à peu près 22 personnes correctement classées de plus. Nous pouvons constater que l'intégration des descripteurs de texture dégrade légèrement les performances par rapport à l'utilisation des seuls descripteurs de couleur. Par conséquent, nous ne retenons que la variante couleur pour la comparer à d'autres approches existantes sur la base VIPeR.

2.6 Etude comparative avec des travaux existants

L'approche proposée de fusion des meilleurs descripteurs couleur est comparée à 4 variantes existantes dans la littérature testées dans [41]. Ces variantes s'appuient sur un partitionnement des images des personnes en six bandes horizontales de même taille représentant la tête, la partie supérieure et inférieure du torse respectivement des jambes. Les descripteurs

d'apparence utilisés pour caractériser les différentes bandes sont les histogrammes de couleurs calculés sur les 8 composantes (R, V, B, H, S, Y, Cb et Cr) et les descripteurs de texture (Filtres de Gabor et de Schmid). Ces approches peuvent être divisées en deux catégories selon l'utilisation ou non de l'algorithme AdaBoost dans la phase de fusion des descripteurs en : approches avec-apprentissage et approches sans-apprentissage.

Pour les méthodes sans-apprentissage, la phase de classification est basée sur la distance de Bhattacharyya (Bhattacharyya [41]) et la Norme L1 (Norme L1 [41]).

Les approches avec-apprentissage testées dans ce travail sont RangBoost [74] et ELF [38]. RangBoost reformule le problème d'identification des personnes en un problème d'arrangement en utilisant les rangs au lieu des distances pour classer les images de piétons. L'algorithme AdaBoost est appliqué afin d'attribuer un poids à chaque descripteur en fonction des rangs pour améliorer les performances. L'approche ELF consiste à appliquer l'algorithme AdaBoost dans le but de sélectionner pour chaque bande horizontale les meilleurs descripteurs qui la caractérise. La mesure de similarité est un ensemble pondéré de tests de vraisemblance entre les instances du même descripteur par la Norme L1. Les cinq approches comparées sont synthétisées dans le tableau 2.3.

Approches	Descripteurs utilisées	Mesure de similarité	Apprentissage (AdaBoost)
Bhattacharyya	Histogrammes de couleurs Filtres de Gabor et de Schmid	Bhattacharyya	
Norme L1		Norme L1	
ELF		Norme L1	☒
RangBoost		Rang	☒
Approche proposé (couleur)	Histogrammes de couleurs Spatigrammes Histogrammes régionaux	Bhattacharyya	☒

Tableau 2.3 – Confrontation de cinq approches de fusion.

La figure 2.17 illustre les performances des différentes approches en se limitant aux 20 premiers rangs qui sont les plus significatifs dans la mesure de précision CMC [75]. En effet, une approche est considérée robuste si elle assure des taux d'identification maximales dans les premiers rangs [76].

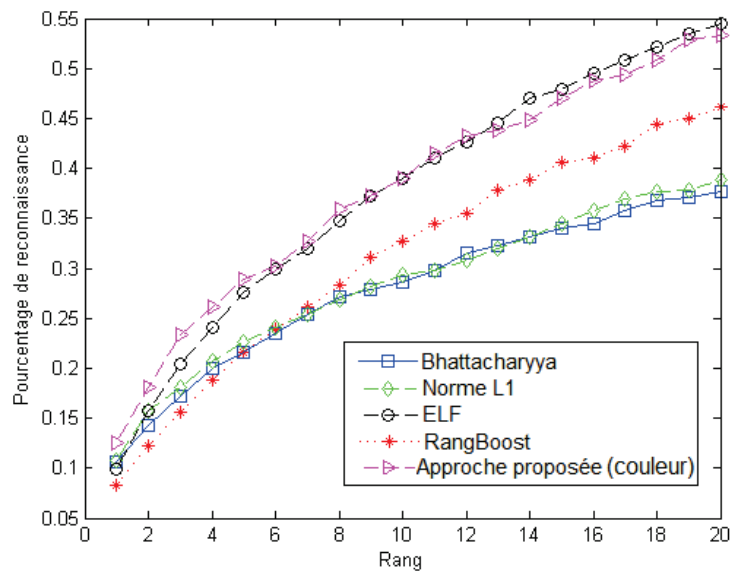


FIGURE 2.17 – Courbes de précision CMC pour les cinq approches comparées.

La figure 2.17 montre que les approches sans-apprentissage, qui utilisent une simple mesure de distance (Bhattacharyya ou Norme L1) ne sont pas performantes pour identifier des personnes dans un contexte multi-caméras. Ceci est relatif à la complexité du test réalisé dans le sens où le nombre total d'appariements est de $(316 \times 316 = 99856)$ dont uniquement 316 appariements sont pour la même personne, avec de fortes variations de pose et d'éclairage.

La figure 2.17 montre aussi que l'approche proposée est largement plus robuste que l'approche RangBoost et légèrement plus performante que l'approche ELF. Pour le rang 1, l'approche proposée permet d'améliorer respectivement les performances d'environ 3% et 6% (soit 10 et 20 personnes correctement classées de plus) par rapport à ELF et RangBoost. Cette amélioration repose sur le choix de descripteurs robustes plus notamment les spatio-grammes et les histogrammes régionaux ; qui représente une stratégie très appropriée pour identifier les piétons en multi-caméra. La légère supériorité de l'approche ELF pour les rangs supérieurs à 12 est relative à l'utilisation de la valeur absolue de la différence (Norme L1) comme mesure de similarité ce qui permet une meilleure séparation entre les exemples par rapport à la distance de Bhattacharyya utilisé dans notre travail.

2.7 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle mesure de similarité utilisant l'information intra et inter-personnes basée sur les histogrammes de couleurs régionaux. Cette mesure de similarité permet de gérer les variations de pose et d'éclairage par l'exploitation de la différence d'intensité intra-personne. Nous avons aussi réalisé une étude comparative entre les principaux descripteurs de couleur, de texture et de forme proposés dans la littérature. Les résultats obtenus confirment que les descripteurs de couleurs et de texture sont les plus appropriés pour caractériser et identifier les personnes par l'apparence.

Nous avons ensuite montré que la caractérisation pourrait être notablement améliorée par une fusion des meilleurs descripteurs avec l'algorithme AdaBoost. Pour cela, deux variantes ont été testées à savoir la fusion des descripteurs de couleurs seuls ou combinés avec des descripteurs de texture. Les résultats obtenus montrent clairement que l'intégration des descripteurs de texture dégrade légèrement les performances. La variante couleur proposée démontre d'excellents résultats face aux travaux les plus reconnus de la littérature.

Dans le but d'améliorer encore la phase de caractérisation des personnes, nous proposons de combiner l'apparence à une caractéristique biométrique modélisant la marche des piétons. Le chapitre suivant présente un état de l'art sur la reconnaissance des personnes par la démarche.

Chapitre 3

État de l’art sur l’identification des personnes par la démarche

3.1 Introduction

Les chapitres précédents se sont attachés à la reconnaissance d’une personne par son apparence en utilisant essentiellement des descripteurs de couleur. Cependant, une telle reconnaissance reste fragile, notamment dans le cas de parures vestimentaires proches qui peuvent amener des ambiguïtés lors des ré-identifications. Afin d’améliorer la phase de caractérisation et assurer une meilleure identification des personnes dans un réseau de caméra, nous avons souhaité combiner l’apparence avec une caractéristique biométrique.

Les techniques biométriques qui se sont largement développées ces dernières années, peuvent être classées en deux grandes catégories selon le degré de coopération de la personne dans la phase d’identification : intrusives ou non-intrusives. Les techniques biométriques intrusives nécessitent que l’individu soit en contact direct avec le matériel d’acquisition (cas des empreintes digitales) ou situé à une distance raisonnablement limitée par rapport à la caméra (cas du visage ou l’iris). Ces techniques intrusives restent donc inadaptées au domaine de la vidéosurveillance où les piétons évoluent librement dans un espace urbain.

Une des caractéristiques biométriques intéressante dans ce contexte est la démarche. Chaque personne a une façon bien à lui de se mouvoir qui peut être observée par une caméra et permettre une identification automatique.

Ce chapitre dresse un état de l'art sur l'identification des personnes par la démarche.

3.2 Identification des personnes par la démarche

En 1973, une étude psychologique réalisée par Johansson [77] a démontré que la façon de marcher est unique pour chaque personne. De nombreux travaux se sont alors intéressés à la démarche comme caractéristique biométrique pour identifier les personnes. Extraire une signature robuste pour identifier les personnes par la démarche à partir d'une vidéo reste une tâche difficile si on veut s'affranchir des variations de point de vue, d'apparence (liée à la manière de s'habiller : jupe, pantalon, manteau, etc.) ou de vitesse de déplacement des piétons. La démarche peut aussi être affectée par la nature des chaussures, la topologie du terrain, le fait de porter un objet ou non et selon l'état psychologique des personnes. Dans ce sens, Sarkar et al. [78] proposent un travail de synthèse qui consiste à étudier l'impact de différentes problématiques liées à ce domaine de recherche à savoir les variations du point de vue, de la manière de déplacement (type de chaussure, la nature du terrain, le fait de porter ou non un objet) ainsi que le problème de synchronisation entre les séquences de marche, sur les taux de reconnaissance. Cette étude a montré que la nature du terrain et les problèmes de synchronisation ont le plus grand impact sur les taux de reconnaissance.

La majorité des approches existantes est focalisée sur l'observation des déplacements en vue latérale, un petit nombre l'aborde en vue frontale. Par souci d'homogénéité, ce chapitre se consacre donc sur les principaux travaux d'identification de personnes en vue latérale. Les approches en vue frontale seront détaillées dans le chapitre 4 lors de la comparaison avec l'approche frontale proposée.

Dans les sections suivantes, nous séparons les approches existantes pour caractériser les personnes par la démarche en vue latérale en deux catégories : les approches basées modèles et les approches directes (sans modèle). Les premières s'appuient sur un modèle morphologique des silhouettes (boîte englobante, squelette articulée, etc.) pour extraire des descripteurs de démarche, tandis que les secondes calculent directement une signature de démarche à partir des images successives des piétons, sans passer par un modèle morphologique.

3.2.1 Approches basées modèles

Dans ces approches basées modèles, les descripteurs spatio-temporels qui caractérisent la démarche sont calculés à partir des variations du modèle morphologique. Différents types de modèles ont été proposés dans la littérature. Un premier groupe de travaux utilise un squelette de la personne pour calculer des distances ou des angles relatifs aux articulations du corps humain (figure 3.1). Bobick et Johnson [79], par exemple, exploitent quatre paramètres du squelette : la hauteur de la personne ($D1$), la distance entre la tête et la hanche ($D2$), la distance entre la hanche et la cheville la plus éloignée ($D3$) et enfin la distance entre les deux chevilles ($D4$). Ces quatre paramètres ne sont calculés que lorsque la foulée est maximale pour avoir des vecteurs caractéristiques synchrones lors de la phase d'appariement. L'identification des personnes est réalisée par une simple mesure de similarité basée sur l'information mutuelle. D'après cette étude, Bobick et Johnson ont montré que l'utilisation des deux paramètres $D1$ et $D3$ est plus robuste que la fusion de tous les paramètres.

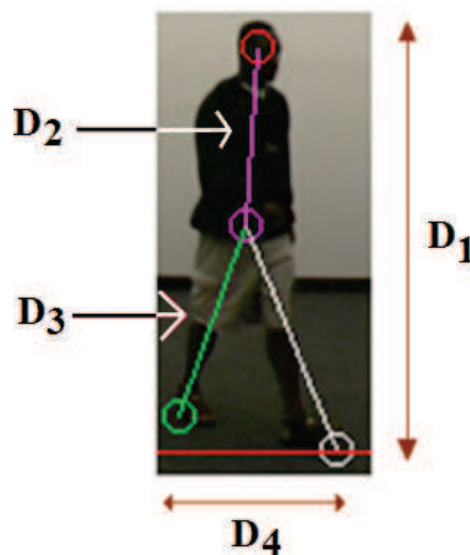


FIGURE 3.1 – Descripteurs utilisés dans [79].

La démarche peut également être caractérisée par la moyenne et l'écart-type des variations des angles extraits du squelette [80] durant un cycle de démarche. D'autres travaux préfèrent un modèle en ellipses [81] [82]. La phase de reconnaissance est réalisée en utilisant un classifieur par réseau de neurones dans [80] ou par KPPV dans [81] [82].

Certains auteurs [83] [84] se sont focalisés sur la partie inférieure du corps en utilisant

uniquement les variations spatio-temporelles des angles de la hanche et du genou (figure 3.2). Afin de reconnaître les personnes, Yam et al. [84] proposent d'appliquer la transformée de Fourier sur les signaux extraits (variations des angles de la hanche et du genou) et de pondérer la phase par l'amplitude du signal pour assurer une bonne caractérisation. Les résultats obtenus montrent que les déplacements rapides permettent de mieux discriminer les personnes par la démarche comparés aux déplacements lents. ChewYean et al. [83] proposent de palier aux changements de vitesse de déplacement en reliant les variations des angles lorsque la personne marche aux variations des mêmes angles lorsque la personne court. Cette transformation est basée sur un simple ajustement de la phase et de l'amplitude du signal original.

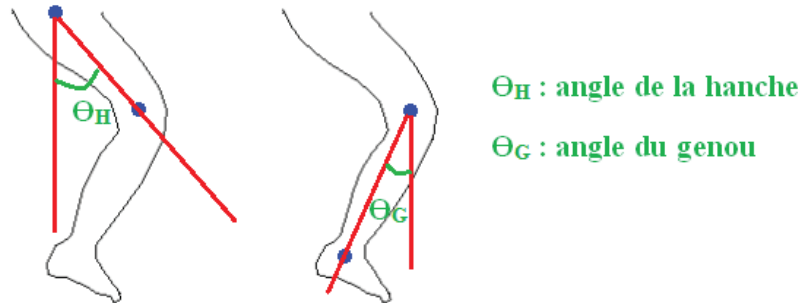


FIGURE 3.2 – Angles de la hanche et du genou utilisés dans [83] et [84].

Yoo et Nixon [85] proposent de fusionner plusieurs paramètres calculés durant un cycle de marche à savoir la hauteur de la silhouette, la durée d'un cycle (nombre d'images), la largeur du pas, la vitesse de déplacement, la moyenne des angles de la silhouette, la variation de l'angle de la hanche et la corrélation entre les mouvements du pied droit et gauche. Avant d'extraire ces descripteurs, une phase d'initialisation est réalisée pour localiser les points d'articulation sur la première image de la séquence (figure 3.3). Les coordonnées verticales de ces points d'intérêt sont estimées proportionnellement à la hauteur de la silhouette selon une étude médicale proposée dans [86], leur coordonnées horizontales au milieu de la silhouette. Pour le reste de la séquence, les positions évoluent selon une fonction paramétrique modélisant leurs variations durant la marche.

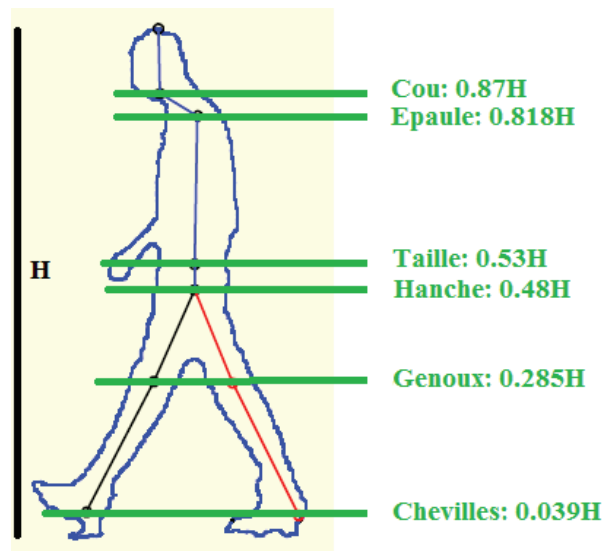


FIGURE 3.3 – Positions verticales des points d'articulation selon l'étude médicale [86].

Tanawongsuwan et Bobick [87] utilisent les variations de quatre angles appartenant à la partie inférieure du corps humain, mais dans ce cas, la détection des points d'intérêts relatifs aux hanches et aux genoux est réalisée en plaçant plusieurs capteurs sur le corps du piéton. Avant de faire l'appariement, une double normalisation est appliquée sur chaque signal pour assurer une invariance temporelle par rapport à la taille du signal et une invariance par rapport à l'amplitude. Ces quatre signaux normalisés sont ensuite concaténés pour former le vecteur caractéristique.

Les approches [83] [84] [85] [87] utilisent toutes l'algorithme KPPV pour la classification.

Compte tenu de la difficulté de l'extraction automatique des squelettes, d'autres approches exploitent un modèle morphologique beaucoup plus simple comme la boîte englobante. Guillen et al. [88] ont testé plusieurs descripteurs spatio-temporels calculés à partir des variations des dimensions de la boîte englobante durant un cycle de marche, à savoir, la hauteur, la largeur, l'angle de la diagonale, la surface ainsi que la transformée de Fourier appliquée sur la variation de la surface. Avant de procéder à l'identification des personnes, une phase d'apprentissage est effectuée, au cours de laquelle ils calculent pour chaque personne la valeur moyenne ainsi que le minimum et le maximum relatif à chaque descripteur utilisé. La phase de test consiste à identifier une personne en se référant aux intervalles déjà calculés dans la phase d'apprentissage. Les résultats obtenus dans ce travail montrent que les variations

spatio-temporelles de la hauteur offrent les meilleures performances comparé aux autres descripteurs pour identifier les personnes.

Ekinci [89] introduit une approche qui combine les aspects structurel et dynamique des silhouettes calculés à partir de la boîte englobante (figure 3.4). Pour chaque image de la séquence, il calcule les distances entre le pourtour de la silhouette et les quatre cotés respectifs de sa boîte englobante (haut, bas, gauche et droite). Ces 4 vecteurs de distances sont ensuite cumulés au cours du temps afin d'obtenir une signature temporelle. Finalement, la dimension de ces signatures temporelles est réduite par une analyse en composantes principales (ACP). L'algorithme KPPV est appliqué pour classifier les personnes.

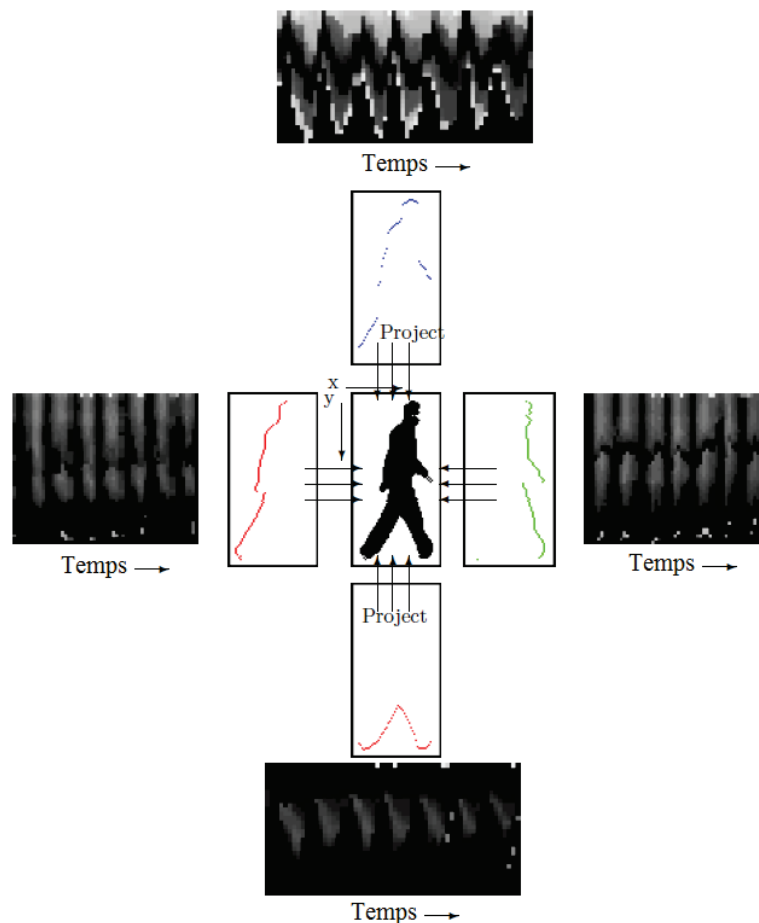


FIGURE 3.4 – Signatures spatio-temporelles de la démarche selon [89] : (au centre) la silhouette et les pourtours selon les quatre côtés de la boîte englobante, (au tour) les signatures temporelles cumulant les distances au cours du temps.

Dans [90], la moyenne des contours des silhouettes durant un cycle de marche est associée aux variations des angles droites et gauches de la hanche et des deux genoux. Tandis que dans [91], le gabarit des personnes est représenté par les longueurs de la cuisse et du tibia modélisés en trapèzes (figure 3.5) alors que la partie dynamique est composée par les variations des angles de la hanche et des genoux complétée par le mouvement spatial de la hanche (x et y). Pour s'affranchir des variations de vitesse de déplacement, les auteurs [91] proposent d'adapter les paramètres dynamiques selon la vitesse. Cette adaptation repose sur une synchronisation des séquences de marche, puis un ajustement de la foulée, en effectuant une étude préliminaire sur la base de séquences d'apprentissage.

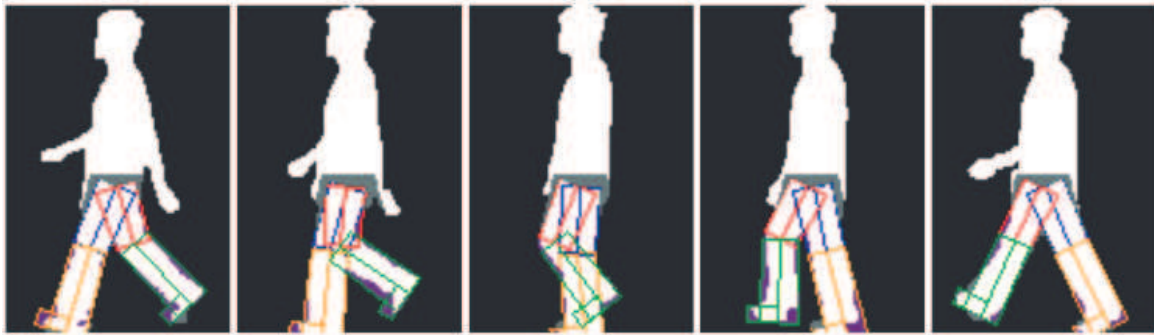


FIGURE 3.5 – Modèle en trapèzes de la partie inférieure de la silhouette [91].

3.2.2 Approches directes

Ces approches extraient des caractéristiques de la démarche directement de la séquence des images des silhouettes sans considérer aucun modèle morphologique.

– Graphes HMM :

Une première catégorie d'approches utilise les modèles de Markov cachés (HMM) pour représenter la démarche sous la forme d'un graphe dont les nœuds constituent les attitudes (les silhouettes) et les arcs modélisent les transitions possibles entre ces états. Pour limiter la complexité liée au nombre de nœuds, la plupart des approches ne considèrent que quelques attitudes clés dans le cycle de marche.

Dans ce sens, Kale et al. [92] [93] proposent d'utiliser les cinq attitudes les plus fréquentes dans un cycle de marche (figure 3.6) sélectionnées par l'application de l'algorithme K-means

[46] sur la totalité des exemples d'apprentissage. Chaque état du modèle HMM est représenté par le vecteur des largeurs de la silhouette ligne-à-ligne [92]. Ce travail a été étendu dans [93] en combinant le vecteur des largeurs avec l'intégralité de l'image de la silhouette dans le modèle d'état du HMM. La phase de reconnaissance est réalisée par l'algorithme de Viterbi [94] qui consiste à maximiser la probabilité de pouvoir générer un HMM identifié à partir d'un HMM de test (anonyme).



FIGURE 3.6 – Les cinq positions de silhouettes les plus fréquentes dans un cycle de démarche [92].

He et Debrunner [95] utilisent les moments de Hu [61] pour caractériser chaque état dans les HMM, pour profiter de leurs propriétés d'invariance à la translation, rotation et changement d'échelle. Avant d'apparier les séquences, une étape préliminaire sépare les séquences représentant les mouvements lents (marche) et les mouvements rapides (course) en se référant au nombre d'images constituant le cycle de déplacement. La phase de reconnaissance est réalisée par l'algorithme de Viterbi [94] pour une même vitesse (appariement marcher/marcher ou courir/courir).

– Variations du contour de la silhouette :

Une deuxième catégorie d'approches exploite les variations spatio-temporelles de la forme des silhouettes durant un cycle de marche pour caractériser les personnes. Wang et al. [96] utilisent un signal représentant le contour de la silhouette par l'ensemble des distances qui le séparent de son centroïde (figure 3.7).

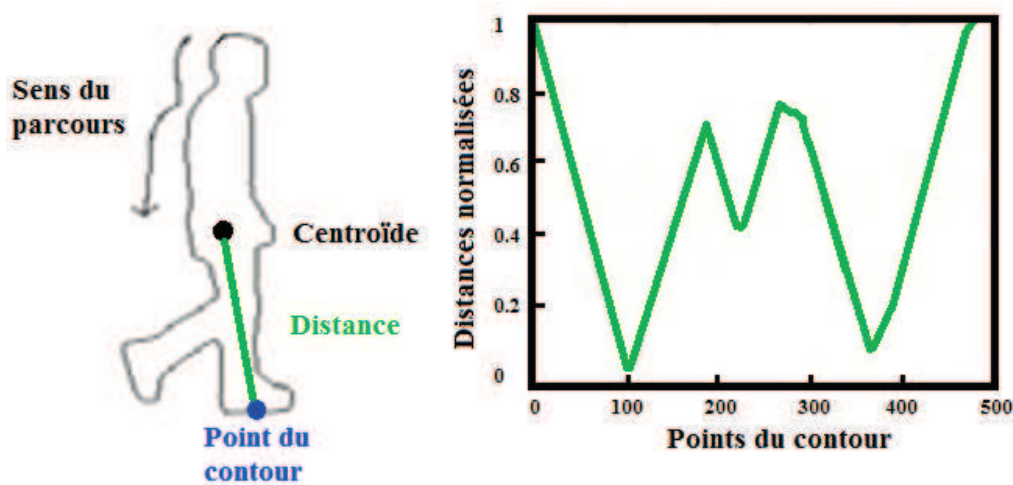


FIGURE 3.7 – Processus d'extraction du signal unidimensionnel [96].

Les auteurs ont ensuite étendu leur approche en proposant de moyenniser les signaux mono-dimensionnels durant un cycle de démarche [97]. D'autres auteurs ont appliqué une transformée en ondelette (DWT) sur ces signaux dans le but d'améliorer la phase de caractérisation [98]. Finalement, une ACP est appliquée pour limiter la dimension, avant d'utiliser l'algorithme KPPV pour la classification. Ces travaux ont prouvé que l'intégration de certaines propriétés physiques, comme la largeur de la foulée et la hauteur de la silhouette, permet d'améliorer le taux de reconnaissance.

Boulgouris et al. [99] exploitent également la forme de la silhouette binaire. Après découpage de la silhouette en plusieurs secteurs angulaires à partir du centroïde, ils construisent une signature constituée des distances entre le centre de gravité et les pixels de chaque secteur. L'identification s'effectue ensuite par distance euclidienne entre les signatures modélisant une séquence de mouvement.

– Paramètres génériques (foulée et cadence) et matrice d'autosimilarité :

En considérant une vitesse de déplacement constante, BenAbdelkader et al. [100] proposent de caractériser la démarche en utilisant la largeur de la foulée et la cadence du mouvement. En effet, la largeur de la foulée dépend de la distance parcourue durant un intervalle de temps, alors que la cadence est calculée en fonction de la périodicité du mouvement. BenAbdelkader et al. [101] introduisent aussi la matrice d'autosimilarité (Self Similarity Plot : SSP) calculée par corrélation entre les silhouettes de la même séquence (figure 3.8). Deux variantes de

la matrice SSP ont été testées dans ce travail en considérant des silhouettes en couleur et binaires. Les résultats montrent que la variante binaire est la plus appropriée pour identifier les personnes.

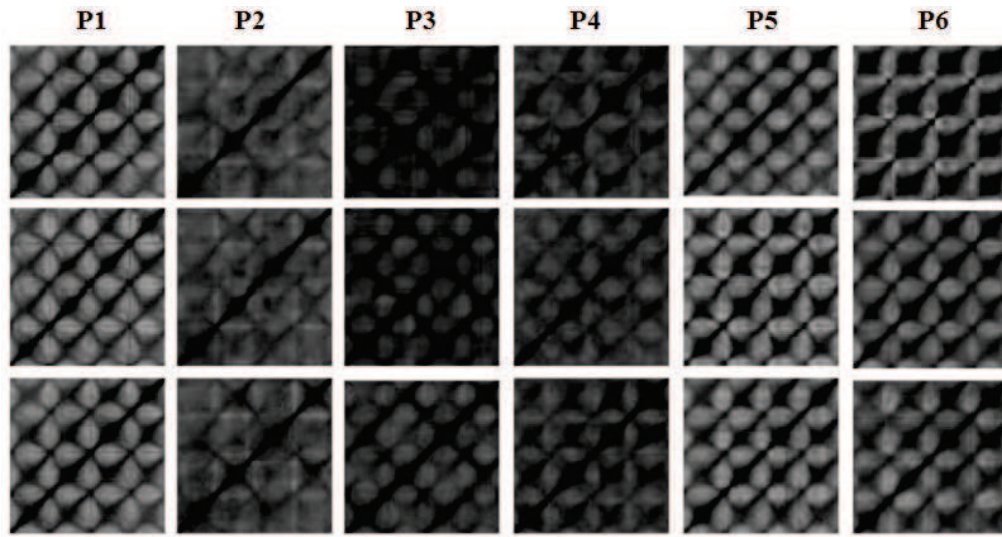


FIGURE 3.8 – Exemples de matrices SSP pour 6 personnes différentes calculées à partir d'images binaires des silhouettes [101].

La phase de caractérisation des personnes consiste à extraire un ensemble d'unités d'autosimilarité (USS) à partir du triangle supérieur de la matrice symétrique SSP et qui représentent un cycle de déplacement. Après une réduction de dimensionnalité par s-LDA, une classification est effectuée par KPPV avec la distance euclidienne comme métrique sur les USS.

– Modèle de frise (motifs répétitifs) :

Une quatrième famille de méthodes exploite la séquence des silhouettes sous forme de frise. Ce modèle introduit par Liu et al. [102], consiste à projeter horizontalement et verticalement le nombre de pixels appartenant à la silhouette. En cumulant ces projections au cours du temps, ils mettent en évidence les motifs répétitifs du cycle des mouvements comme une frise. La similitude entre frises est évaluée par corrélation et utilisée par un algorithme KPPV pour la phase de reconnaissance.

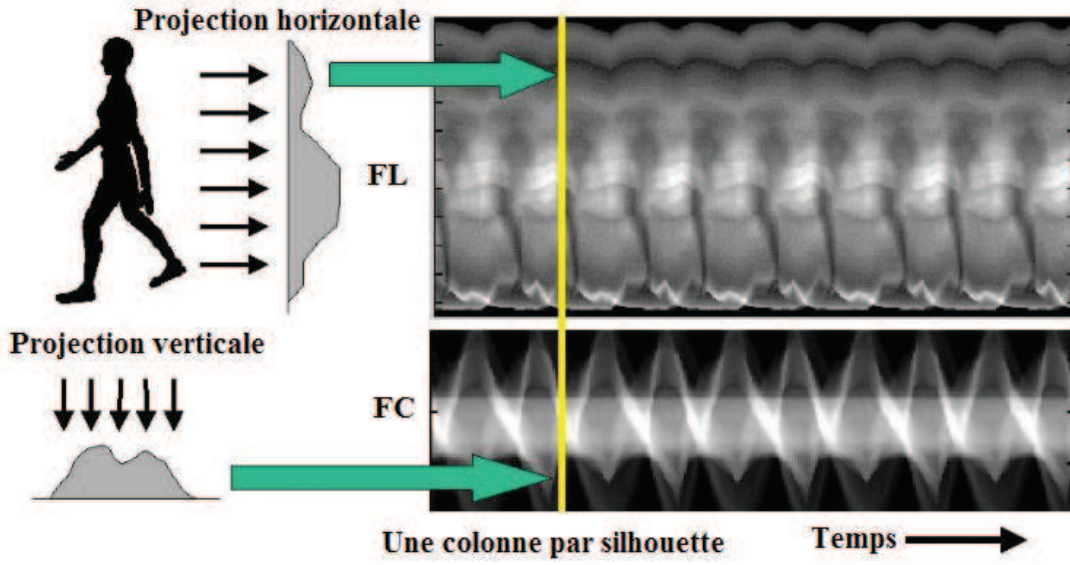


FIGURE 3.9 – Processus de construction du modèle de freize [102].

Lee et al. [103] ont proposé une variante de frise en soustrayant à chaque image du cycle une image clé correspondant à la position d'écart maximal entre les jambes, comme illustré dans la figure 3.10. Le processus d'identification entre 2 piétons (i et j) est réalisé par la minimisation de l'expression (équation 3.1) :

$$C(i, j) = \alpha \times \phi_{FP_h} + \beta \times \phi_{FP_v} + \gamma \times (\phi_{SM_H} + \phi_{SM_V}) \quad (3.1)$$

où ϕ_{FP_h} et ϕ_{FP_v} sont les corrélations entre deux frises horizontales respectivement verticales, ϕ_{SM_H} et ϕ_{SM_V} les corrélations entre deux moitiés frise horizontales respectivement verticales du même cycle et α , β et γ des coefficients de pondération. Cette différence permet de mieux prendre en compte des variations de forme au cours du cycle.

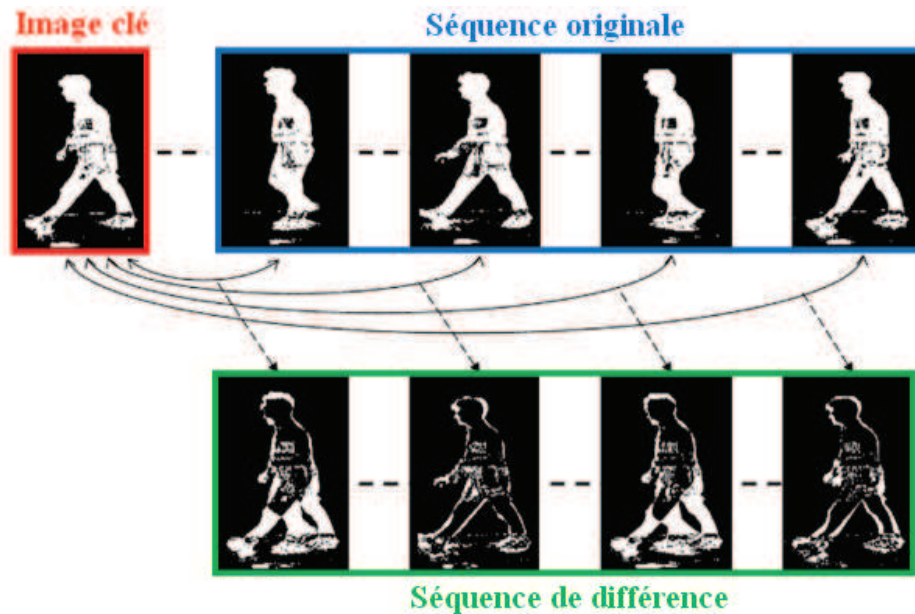


FIGURE 3.10 – Extraction des images de différence [103].

– Images d'énergie : GEI, EGEI, FDEI, AEI :

Enfin, une cinquième famille regroupe les travaux autour de la signature 2D sous forme d'images d'énergie. En 2006, Han and Bhanu [104] introduisent le concept de GEI (Gait Energy Image) en moyennant les silhouettes des personnes durant un cycle de marche (figure 3.12 (a)). De part sa simplicité et son efficacité, cette signature a été beaucoup réutilisée dans la littérature avec des variantes d'amélioration.

Bashir et al. [105] proposent de limiter la signature à la zone la plus discriminante. Une première méthode supervisée consiste à se concentrer sur la moitié inférieure de la GEI et itérativement éliminer les lignes de haut en bas qui n'améliorent pas les taux de reconnaissance. Ainsi le processus itératif s'arrête lorsque l'on obtient la bande la plus étroite pour le meilleur taux de reconnaissance sur l'ensemble d'apprentissage. Une variante non-supervisée consiste à ne garder que les pixels ayant un taux de variation élevé par un simple seuillage de l'image GEI. Les résultats de ces deux méthodes sont illustrés par la figure 3.11.

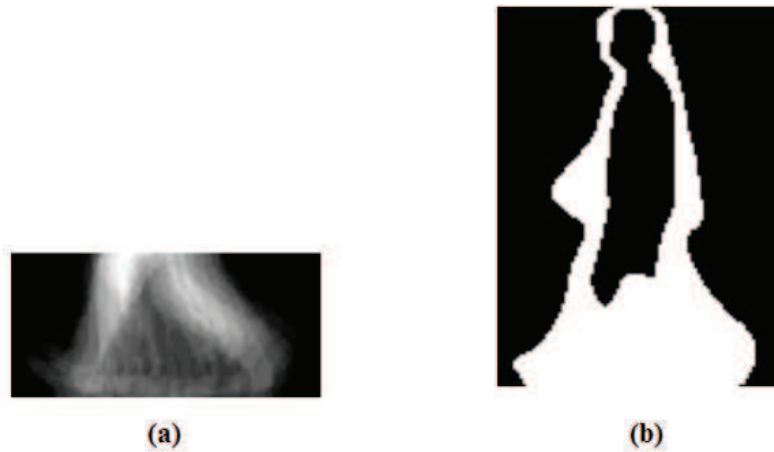


FIGURE 3.11 – Sélection de la région d'intérêt la plus discriminante : (a) par la méthode supervisée et (b) par la méthode non supervisée [105]

Dans le même sens, Lin et Wang [106] proposent une signature nommée EGEI (Enhanced Energy Gait Image) ne retenant que les pixels à haute variabilité de l'image GEI. Pour cela, un masque de poids dynamiques appelé DWM (Dynamic Weight Mask) est construit à partir des images GEI d'apprentissage par une analyse de la variance. Ce masque est ensuite appliqué à toutes les images GEI afin d'obtenir les représentations EGEI. La figure 3.12 présente l'image GEI et l'image EGEI correspondante.

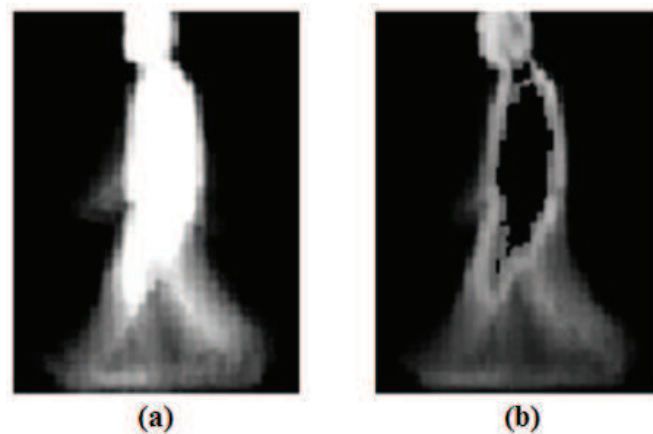


FIGURE 3.12 – Présentation de l'image GEI (a) avec l'image EGEI (b) correspondante [106].

Afin de s'affranchir de la variance selon les points de vue, Huang et Boulgouris [107] proposent de pondérer les images GEI selon les points de vue. Face aux variations d'échelle, Yu et al. [108] préconisent une phase de normalisation sur l'image GEI en divisant chaque

valeur par la somme de l'image, puis d'effectuer une décomposition quadtree pour extraire une signature multi-résolution de la démarche.

Dans certains cas, les erreurs de segmentation relatives à l'étape de soustraction de l'arrière plan génèrent des trous à l'intérieur des silhouettes, et donc une image GEI incomplète. Pour surmonter ce problème, Chen et al. [109] proposent une nouvelle représentation appelée FDEI (Frame Difference Energy Image) combinant l'image GEI avec la différence entre silhouettes successives. Le cumul des images de différence entre silhouettes successives est également utilisé par Zhang et al. [110] pour extraire les zones à fortes variations de mouvement dans les GEI dans une représentation nommée AEI (Active Energy Image).

Même en limitant la région d'intérêt, les variantes de GEI produisent des vecteurs caractéristiques de grandes tailles (signature image 2D), elle elles impliquent une phase de réduction de dimensionnalité par différentes méthodes (PCA, LDA, LPP, 2DPCA, 2DLDA et 2DLP) [109]. Le vecteur caractéristique réduit est généralement classifié par l'algorithme KPPV avec la métrique euclidienne.

3.3 Conclusion

Dans ce chapitre, nous nous sommes particulièrement intéressés à l'identification des personnes par la marche étant donné son aspect non-intrusif permettant même l'identification à distance. Dans cette thématique de recherche, plusieurs travaux ont été proposés dans la littérature dont la plus part d'entre eux ont été présentés dans ce chapitre. En effet, les travaux existants dans ce domaine peuvent être classifiés en deux grandes catégories à savoir les approches basées modèle et approches sans modèle. La seule différence entre ces deux catégories d'approches est que la première exploite les variations d'un modèle morphologique représentant les silhouettes des personnes, tandis que la deuxième caractérise la démarche en analysant les variations au niveau des images des personnes.

Dans le chapitre suivant, nous introduirons la stratégie adoptée pour assurer une bonne phase de caractérisation des personnes par la démarche. Nous présentons par la suite, les différents tests réalisés ainsi que tous les résultats obtenus face aux variations de point de vue, de la vitesse et de la façon de marcher.

Chapitre 4

Identification des personnes par la démarche

4.1 Introduction

Dans le chapitre précédent, nous avons présenté les différentes approches de la littérature pour identifier les personnes par la démarche. Celles ci sont classées en deux catégories : les approches basées-modèles et les approches directes (sans-modèle). Des études existantes [111] [112] montrent que les approches basées-modèles sont plus adaptées dans un contexte non-contrôlé intégrant des changements de point de vue, mais elle nécessitent un temps de traitement plus long pour construire les modèles. Dans ce chapitre, nous présentons deux stratégies différentes pour reconnaître une personne par la démarche selon le point de vue d'observation de la caméra. En point de vue latéral, nous proposons une approche basée-modèle fusionnant des descripteurs spatio-temporels extraits à partir d'un modèle articulé. En point de vue frontal, nous avons opté pour une approche directe exploitant la distribution du mouvement dans les silhouettes.

Les deux sections suivantes introduisent la chaîne de traitements ainsi que les résultats obtenus pour chaque stratégie.

4.2 Approche modèle en point de vue latéral

Par point de vue latéral s'entend, en fait, une variété de points de vue observant la personne plus ou moins de biais selon sa direction de déplacement par rapport à la caméra.

La signature doit donc être assez robuste aux changements de point de vue. Plusieurs modèles ont été proposés dans la littérature à savoir la boîte englobante [88] [89], le modèle en squelette [80] [85] [87], le modèle en ellipses [81] [82] ou en trapèzes [91] (figure 4.1).

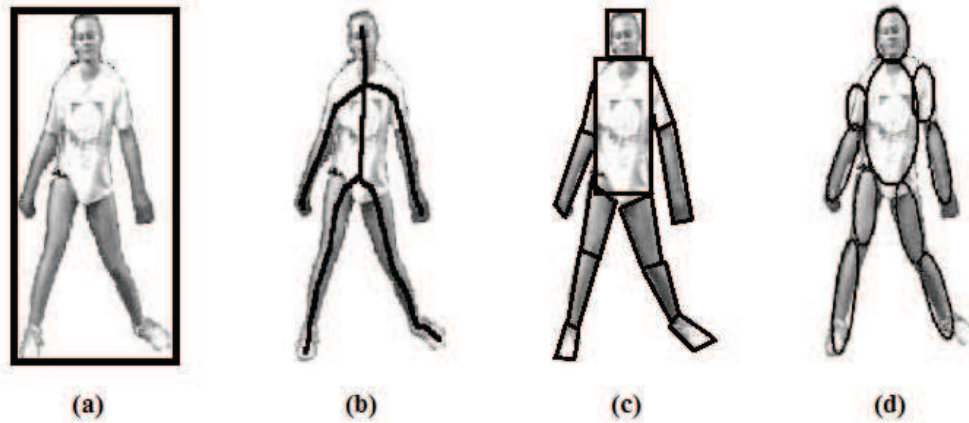


FIGURE 4.1 – Modélisation des silhouettes : (a) boîte englobante, (b) modèle en squelette, (c) modèle en trapèzes et (d) modèle en ellipses.

Le modèle doit tenir en compte de la forme du corps humain (pouvoir représentatif) et minimiser le temps nécessaire pour son extraction (simplicité). Le tableau 4.1 présente la qualité des 4 types de modèles selon ces 2 critères.

	Pouvoir représentatif	Simplicité
Modèles		
Boîte englobante	Faible	Simple
Squelette	Fort	Moyenne
Trapèzes	Fort	Difficile
Ellipses	Fort	Difficile

Tableau 4.1 – Choix du modèle représentant les silhouettes.

Le modèle en squelette permet d'assurer le meilleur compromis. Il permet de prendre en considération la structure du corps humain avec un temps d'extraction acceptable. Nous l'avons donc utilisé dans ce travail pour représenter les silhouettes des personnes. Ce modèle en squelette est obtenu en reliant certains points d'articulation ou d'intérêt du corps humain.

Dans les sections suivantes, nous commençons par introduire le processus utilisé pour détecter automatiquement les points d'intérêt après un court état de l'art sur les techniques existantes. Nous aborderons ensuite la phase de caractérisation de la démarche. Un nouveau descripteur de type *Pendule* sera introduit puis comparé aux descripteurs existants. Finalement, les meilleurs descripteurs seront fusionnés par l'algorithme AdaBoost.

4.2.1 Détection des points d'articulation du squelette

L'analyse de la démarche est basée sur les variations de certains points d'articulations du corps humain comme les épaules, les coudes, les hanches, les genoux ou les chevilles. Plusieurs stratégies ont été adoptées dans la littérature pour les localiser sur les silhouettes.

4.2.1.1 Techniques existantes avec marqueurs

La première technique consiste à placer des marqueurs passifs ou actifs sur les régions clés du corps humain afin de faciliter leur détection. Les marqueurs passifs appelés aussi «point lumineux» [113] [114] [115] sont fabriqués avec des matériaux réfléchissants pour faciliter leur localisation et ils sont généralement placés directement sur la peau de la personne ou sur une combinaison élastique pour réduire les variations de forme provoquées par les vêtements. Cette technique est sensible aux occultations et elle nécessite un prétraitement pour détecter et suivre les marqueurs. Les marqueurs actifs (par exemple des capteurs électromagnétiques, capteurs sans fil, etc.) permettent de remédier à ces problèmes [87] [116] [117], mais leur utilisation reste confiée aux espaces contrôlés (par exemple, dans l'analyse de la performance des athlètes) où la détection reste possible.

Une fois les points initialisés, le suivi s'appuie sur un modèle de mouvement [85] [118] [119] qui reste sensible aux changements de la vitesse et de mode de déplacement (avec port d'un objet ou non, selon la nature des chaussures ou du sol, etc.). Ces variations doivent donc être détectées immédiatement afin de mettre à jour le modèle de mouvement, ce qui est difficile et coûteux en termes de temps d'exécution.

Si l'utilisation des marqueurs passifs ou actifs est indispensable à des fins expérimentales de haute précision, elle est inadaptée à l'analyse de la marche dans les espaces urbains par vidéosurveillance. Une méthode automatique et sans marqueurs semble mieux adaptée dans ce cas.

Un processus fiable de détection des points d'intérêt doit assurer un bon compromis entre temps de calcul (vitesse d'exécution) et précision. Pour accélérer les traitements, nous proposons une technique basée sur des règles morphologiques simples pour détecter automatiquement les points d'articulation. Ces règles sont appliquées de manière itérative image par image afin de prendre en compte chaque attitude de la silhouette, sans utiliser de modèle de mouvement pour s'affranchir des variations de la vitesse et du mode de déplacement.

4.2.1.2 Localisation des points d'intérêt

Plusieurs travaux pointent la partie inférieure du corps humain comme la région la plus appropriée et la plus représentative pour la reconnaissance des personnes par la démarche [83] [91] [105]. Nous nous sommes donc limités à localiser automatiquement les points d'intérêts situés sur la hanche, les genoux et les chevilles en fixant des règles morphologiques simples s'appliquant à la fois en points de vue fronto-parallèle et oblique.

L'étude médicale [86] estime les positions verticales de la hanche, des genoux et des chevilles proportionnellement à la hauteur H de la silhouette, respectivement à $0.48H$, $0.285H$ et $0.039H$ (figure 4.2.a). Toutefois, le niveau vertical des chevilles n'est pas le même pour les deux pieds et il change considérablement durant le cycle de marche. En outre, dans certaines positions du corps, les coordonnées horizontales des genoux et des chevilles sont difficiles à détecter (même visuellement) en raison des occultations partielles entre les jambes.

Nous proposons donc des améliorations pour détecter les genoux et les chevilles. La détection de la hanche restera fixée verticalement à $0.48H$ et horizontalement au milieu du contour de la silhouette. Pour la détection des genoux et des chevilles, deux cas de figure sont considérés (figure 4.2(a)) : **i**) l'attitude en simple support où la personne a un seul pied d'appui et les jambes sont chevauchantes et **ii**) l'attitude en double support où les deux pieds sont en contact avec le sol et les jambes écartées.

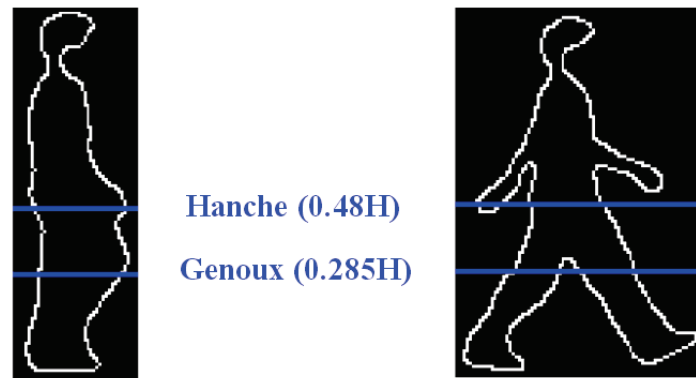
Ces deux attitudes sont distinguées par une analyse des lignes inférieures de la silhouette au niveau des genoux ($\leq 0.285H$). Si la majorité des lignes ont quatre intersections avec le contour de la silhouette alors c'est une attitude en double support. Sinon, il s'agit d'une attitude simple support. Pour chaque attitude, les règles simples suivantes sont alors appliquées :

● **Attitude en simple support (figure 4.2 (b) et (c)) :**

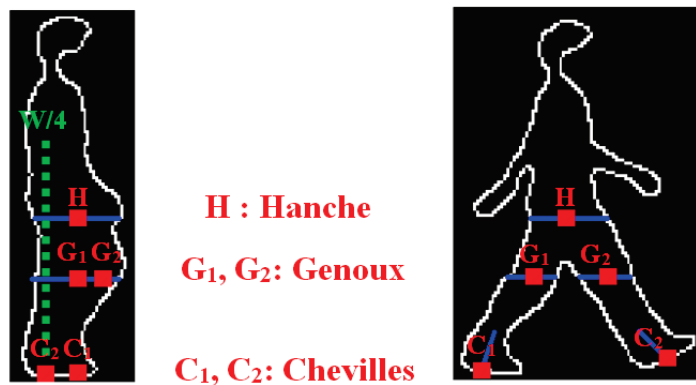
- la position verticale des 2 genoux (G_1 et G_2) est fixée à $0.285H$.
- la position verticale de la cheville C_1 (pied d'appui) est fixée comme le point le plus bas du contour de la silhouette.
- les positions horizontales du genou G_1 et de la cheville C_1 sont alignées avec la hanche H .
- les positions horizontales du genou en mouvement G_2 est situé au milieu du segment reliant le point de genou déjà détecté (G_1 du pied d'appui) au contour de la silhouette.
- la position verticale de la cheville mobile C_2 est située à l'intersection du contour et de $y = W/4$ (où W représente la largeur de la silhouette).

● **Attitude en double support (figure 4.2 (b) et (c)) :**

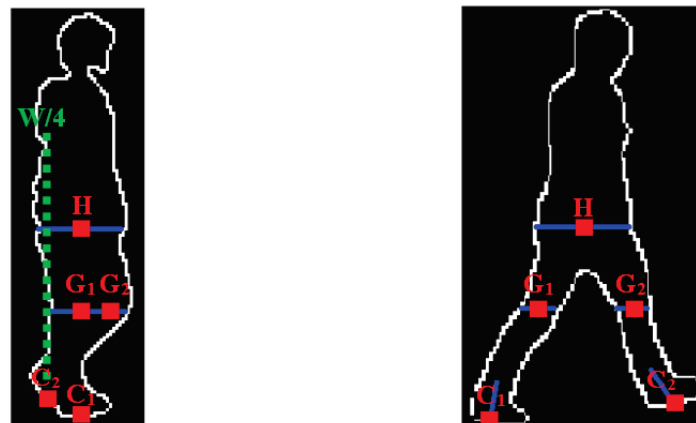
- Les deux genoux (G_1 et G_2) sont situés verticalement à $0.285H$ et horizontalement au milieu des deux segments correspondant aux jambes de la silhouette. Comme, il est parfois impossible d'obtenir deux segments à cette hauteur en raison du croisement entre les cuisses ou encore de la nature des vêtements (un long manteau par exemple), nous commençons par chercher le niveau vertical le plus haut dans la silhouette où il existe quatre intersections avec le contour, en balayant l'image ligne par ligne depuis la hauteur $0.285H$. Finalement, les positions horizontales des genoux sont estimées par la projection orthogonale du milieu de chaque segment sur le niveau vertical $0.285H$.
- Pour positionner les deux chevilles (C_1 et C_2), on détermine d'abord l'axe médian de chaque jambe en appliquant la méthode des moindres carrés (K-means [46]) sur l'ensemble des points milieux de la partie inférieure de chaque jambe ($\leq 0.07H$). Puis, les chevilles sont situées le long de cet axe sur le point de contour le plus bas (extrême).



(a)



(b)



(c)

FIGURE 4.2 – Processus de détection des points d'intérêt selon l'attitude : (colonne de gauche) simple support, (colonne de droite) double support, avec deux points de vue différentes (b : fronto-parallèle et c : oblique).

En appliquant ces règles morphologiques simples, nous pouvons localiser automatiquement les points d'intérêt sur la partie inférieure du corps humain au cours du cycle de marche malgré les occultations partielles entre les jambes. Un modèle squelettique est obtenu en reliant les points détectés (figure 4.3). Il est à préciser que le passage d'une attitude double support à une attitude simple support nécessite le croisement des jambes. Dans ce cas, le genou avant sera relié à la cheville arrière et vice versa.

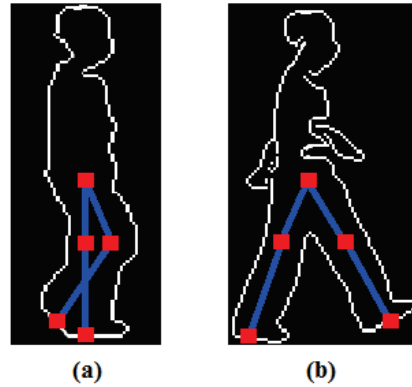


FIGURE 4.3 – Exemple de détection automatique des squelettes : (a) simple support et (b) double support.

4.2.1.3 Erreurs de localisation des points d'intérêt

Afin de valider les règles précédentes, nous proposons de mesurer les déviations spatiales des points détectés automatiquement par rapport à une vérité de terrain construite manuellement à partir de la base CASIA (C) [120].

Cette base contient 153 personnes filmées par une caméra infrarouge dans un environnement extérieur en vue fronto-parallèle avec des variations de vitesses et de modes de déplacements. Chaque personne est représentée par 10 séquences : 4 marches normales, 2 marches normales avec un sac, 2 marches lentes et 2 marches rapides. Chaque séquence contient deux cycles de marche.

La vérité de terrain comprend les positions horizontales et verticales de la hanche, des genoux et des points de chevilles manuellement extraites sur 535 images, soit un total de 2675 points d'intérêt. La base d'image correspond à un cycle de marche de chaque catégorie de mouvement (soit 4 séquences) pour 5 personnes choisies aléatoirement. Le nombre d'images constituant un cycle de marche varie en fonction de la vitesse de déplacement, soit 23, 27 et

30 images respectivement pour les séquences de marche rapide, normale et lente.

L'erreur $E_{seq}(p, s)$ mesurant la déviation spatiale entre les positions successives d'un point p détectées automatiquement tout au long de la séquence s et la vérité de terrain est estimée par la distance euclidienne (équation 4.1) :

$$E_{seq}(p, s) = \frac{\sum_{t=1}^T \sqrt{(x_{vt}(t) - x(t))^2 + (y_{vt}(t) - y(t))^2}}{T} \quad (4.1)$$

où (x_{vt}, y_{vt}) et (x, y) sont respectivement les coordonnées des points d'intérêt de la vérité de terrain et du processus de détection automatique et T est le nombre d'images dans la séquence s .

Puis, nous calculons la déviation moyenne $E_{mov}(p, m)$ d'un point p pour chaque catégorie de mouvement $m = \{N \text{ Normale}, R \text{ Rapide}, L \text{ Lente et } S \text{ normale avec Sac}\}$ selon l'équation (4.2) :

$$E_{mov}(p, m = \{N, R, L, S\}) = \frac{\sum_1^R (E_{seq}(p, s))}{N} \quad (4.2)$$

où N est le nombre de séquences appartenant à la catégorie de mouvement m pour toutes les personnes.

Le tableau 4.2 présente la moyenne et l'écart type de l'erreur $E_{mov}(p, m)$ pour chaque point d'intérêt p en considérant toutes les personnes et toutes les catégories de mouvements, calculés respectivement par les équations 4.3 et 4.4.

$$M(p) = \frac{\sum_{m \in \{N, R, L, S\}} E_{mov}(p, m)}{P} \quad (4.3)$$

$$STD(p) = \left[\frac{\sum_{m \in \{N, R, L, S\}} (E_{mov}(p, m) - M(p))^{\frac{1}{2}}}{P} \right]^2 \quad (4.4)$$

où P est le nombre de catégories de mouvement dans la base.

	Hanche	Genou gauche	Genou droite	Cheville gauche	Cheville droite
Précision (en pixels)					
Moyenne	3.22	4.07	3.76	3.86	3.95
Ecart type	0.25	0.48	0.57	0.31	0.45

Tableau 4.2 – Précision de la détection des points d'intérêt (en pixels).

Les résultats présentés dans le tableau 4.2 confirment que les positions des points d'intérêt détectés automatiquement sont proches de la vérité de terrain malgré les variations de mode de déplacement. En effet, la moyenne de l'erreur ne dépasse pas 4 pixels (avec un écart type maximum de 0.6 pixel) sur des silhouettes de taille moyenne allant de 120×40 pixels dans le cas de simple support à 120×80 pixels dans le cas de double support.

Pour visualiser les performances obtenues, nous superposons les squelettes de la détection automatique et de la réalité de terrain pour la même personne et selon toutes les catégories de mouvement sur la figure 4.4. Cette figure montre clairement que les deux squelettes sont généralement proches pour toutes les catégories de mouvement. Nous obtenons des résultats similaires pour toutes les personnes testées.

Ces premiers tests de validation seront complétés en section 4.2.3.5 en observant l'impact des erreurs de localisation des points sur les phases de caractérisation de la démarche et d'identification des personnes.

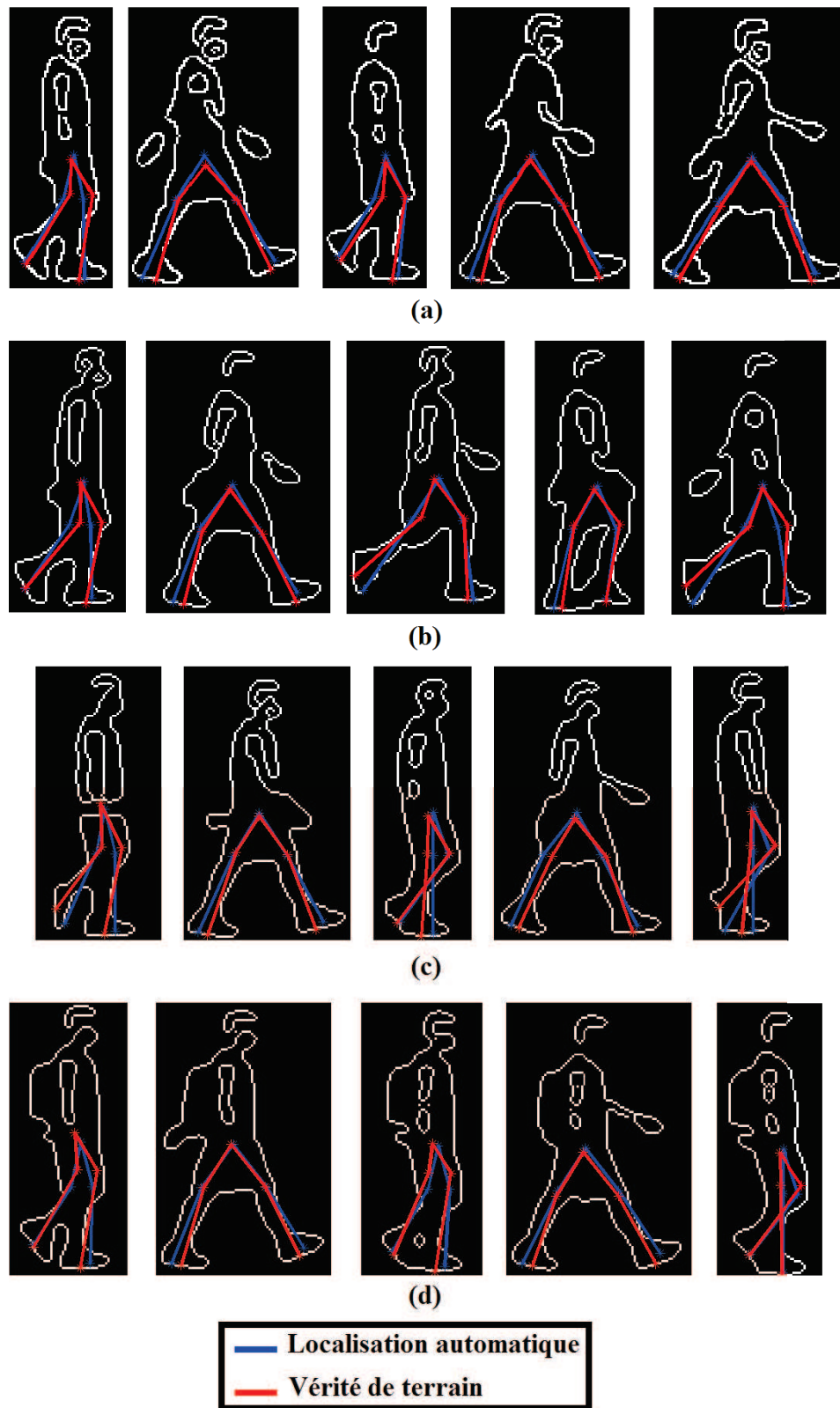


FIGURE 4.4 – Superposition des squelettes de la vérité de terrain et ceux détectés automatiquement pour la même personne et selon toutes les catégories de mouvement : (a) Normale, (b) Rapide, (c) Lente et (d) Avec Sac.

4.2.2 Caractérisation de la démarche à partir du squelette

Une fois les points d'intérêt détectés, nous entamons la phase de caractérisation qui permettra de différencier les personnes par leur démarche. Nous commençons par présenter les descripteurs de la littérature, puis le descripteur proposé.

4.2.2.1 Descripteurs existants

Plusieurs descripteurs spatio-temporels ont été proposés dans la littérature (figure 4.5). Dans [79], les variations de la *Hauteur* de la silhouette, de la distance entre la tête et la hanche ($D1$), de la distance entre la hanche et la cheville la plus éloignée ($D2$) que nous l'appellerions *Jambe* et enfin de la distance entre les deux chevilles ($D3$) ont été utilisées. Les résultats obtenus dans [79] confirment que les descripteurs *Hauteur* et *Jambe* sont les plus performants. Dans [88], un ensemble de descripteurs sont extraits à partir de la boîte englobante tels que la *Hauteur*, la *Largeur*, l'*Angle de diagonale* (α) et de la *Surface*. Les résultats montrent que la *Hauteur* représente le descripteur le plus discriminant comparé à tous les autres. Enfin, plusieurs approches [80] [83] [90] se sont basées sur les variations spatio-temporelles des angles de la *Hanche* et du *Genou* (θ_H , θ_G). À la lumière de tous ces travaux, nous avons sélectionné les descripteurs les plus robustes de la littérature pour notre étude expérimentale qui sont les deux angles (θ_H , θ_G) et les deux distances (*Hauteur*, *Jambe*)

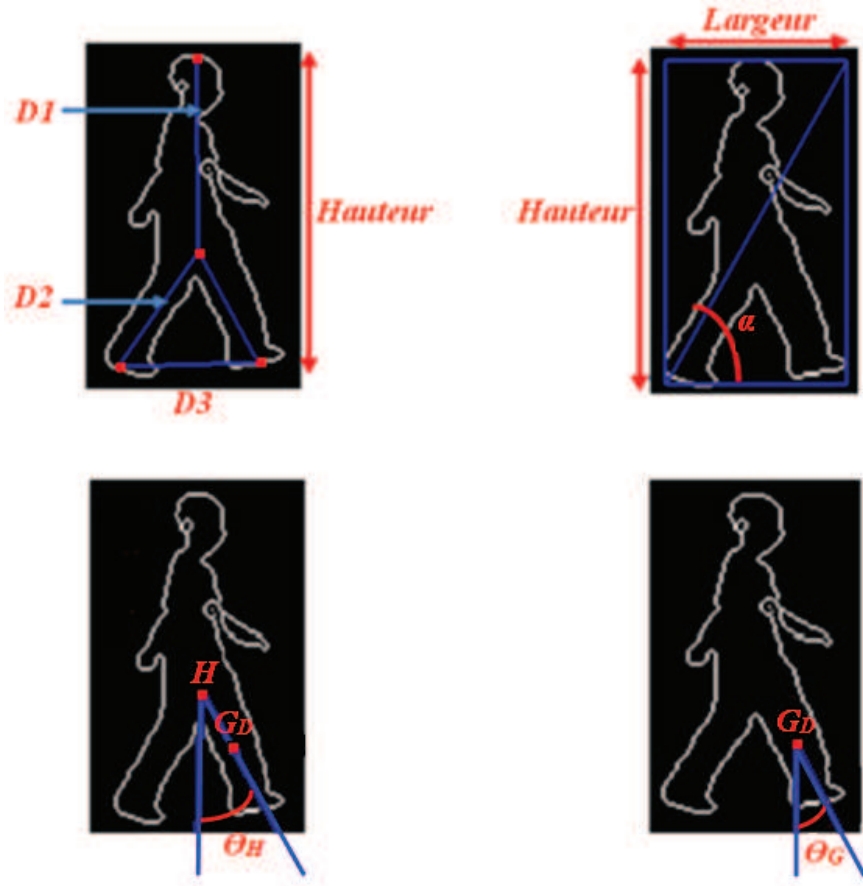


FIGURE 4.5 – Descripteurs existants.

4.2.2.2 Descripteur *Pendule* proposé

Généralement, la démarche est basée sur deux concepts fondamentaux et interdépendants que sont le mouvement et l'équilibre. Tandis que les chevilles caractérisent la face du mouvement vers l'avant (largeur de la foulée), les variations de la hanche maintiennent l'équilibre. Sur la base de ces remarques, nous proposons un nouveau descripteur représentant la démarche comme un pendule oscillant basé sur les positions de la hanche et des chevilles. Le descripteur proposé nommé *Pendule* exploite les variations spatio-temporelles de la distance entre les points PH et M (figure 4.6), où

- PH est la projection orthogonale de la hanche sur le segment reliant les deux chevilles $[C_1, C_2]$.

- M est le milieu du segment formé par les deux chevilles $[C_1, C_2]$.

Le descripteur *Pendule* est estimé pour chaque image t en calculant la distance euclidienne entre les deux points M et PH (équation 4.5).

$$Pendule(t) = \sqrt{[PH_x(t) - M_x(t)]^2 + [PH_y(t) - M_y(t)]^2} \quad (4.5)$$

La figure 4.6 illustre le processus d'extraction du descripteur *Pendule* et montre un exemple de variations du signal au cours de deux cycles de démarche.

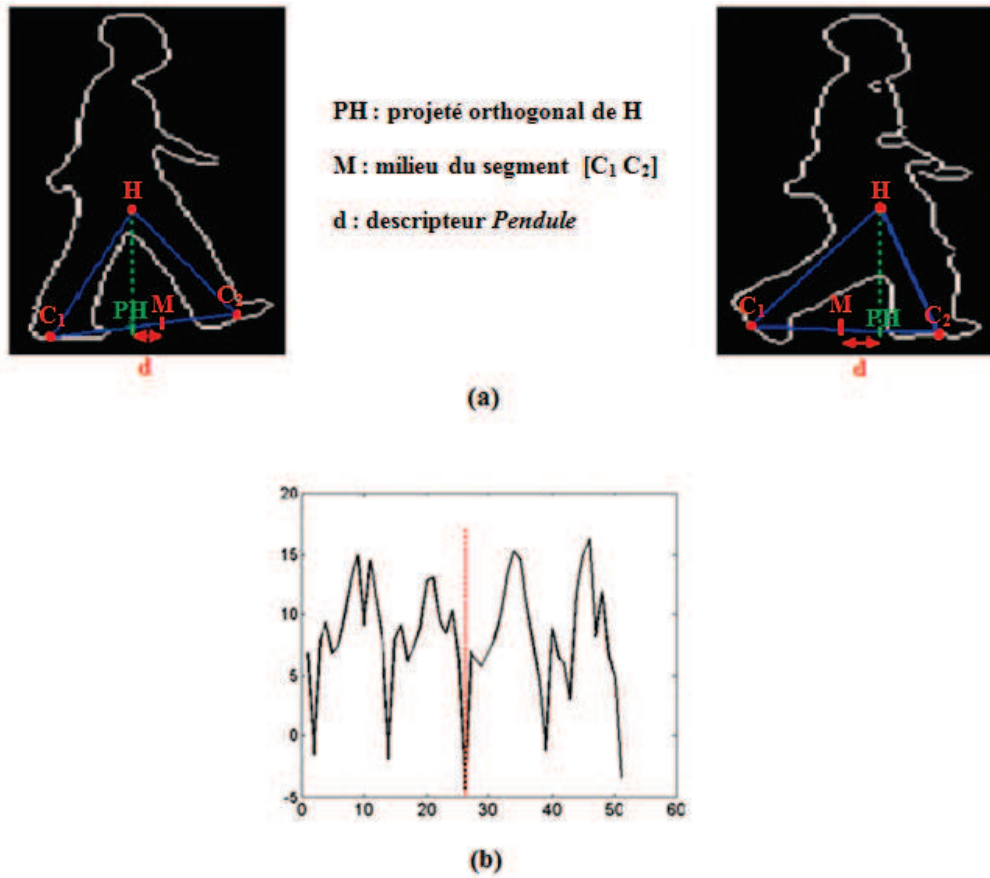


FIGURE 4.6 – Descripteur *Pendule* : (a) Processus d'extraction et (b) variations durant deux cycles de démarche.

4.2.2.3 Représentation fréquentielle

Les descripteurs brutes sont généralement sensibles face aux variations d'échelle (personne proche ou loin de la caméra) et face aux erreurs de localisation des points d'intérêt. Pour

palier à ça, plusieurs travaux ont appliqué une transformation fréquentielle sur les signaux spatio-temporels caractérisant la démarche. La transformée de Fourier (FT) est la plus utilisée dans ce contexte [80] [121] [122]. Cette transformation génère un spectre de fréquences sans conserver l'information temporelle. La transformée en ondelettes discrète (DWT) permet de fournir une représentation temps-fréquence qui peut s'avérer plus descriptive pour les signaux non stationnaires ou non périodiques. Cependant, l'inconvénient majeur de la DWT réside dans sa sensibilité à la translation temporelle des signaux qui reste présente dans les séquences de démarche, même après avoir effectué une synchronisation. Nous avons donc choisi d'appliquer la transformée de Hilbert (HT) qui décrit l'enveloppe complexe du signal DWT pour limiter l'impact de la translation [123]. Afin de montrer l'intérêt de cette transformation, tous les descripteurs sélectionnés sont comparés dans la section suivante en utilisant les deux transformations : transformée de Fourier (T1) et transformée de Hilbert appliquée sur le signal DWT (T2).

4.2.3 Etude comparative entre descripteurs

Les tests sont réalisés sur les bases CASIA (C) et CASIA (A) pour analyser l'impact des variations de vitesse, de mode de déplacement et de point de vue sur les différents descripteurs.

4.2.3.1 Protocole de test

La base CASIA (C) a déjà été présentée dans la section 4.2.1.3. Elle intègre des personnes filmées selon un seul point de vue fronto-parallèle avec de grandes variations de vitesse (marche normale, rapide et lente) ou de modes de déplacement (marche avec sac).

La base CASIA (A), connue aussi sous le nom de NLPR [124], contient 20 personnes filmées selon 3 points de vue (frontale, fronto-parallèle et oblique) alors qu'elles se déplacent à vitesse libre le long d'un trajet rectiligne dans un environnement extérieur. Pour chaque point de vue, quatre séquences sont capturées (deux en s'approchant de la caméra et deux en s'éloignant) soit 12 séquences par personne.

Ces deux bases d'images sont déjà prétraitées pour extraire les silhouettes des personnes de l'arrière-plan. Cependant, des erreurs de segmentation peuvent causer des déformations ou des trous dans certaines régions clés des silhouettes qui pénalisent la détection des points d'intérêt. Pour cette raison, nos différentes expérimentations ont été réalisées sur un sous-ensemble de personnes de ces deux bases où la segmentation est correctement réalisée (30

personnes de la CASIA (C) et 13 personnes de la CASIA (A)).

Afin de comparer le descripteur *Pendule* proposé par rapport aux descripteurs de la littérature (*Hanche*, *Genou*, *Jambe*, *Hauteur*), l'ensemble des séquences vidéo a été synchronisé en utilisant la largeur du pas. Le début et fin du cycle de la marche de chaque séquence est détectée par la plus large foulée (distance entre les deux chevilles). Les séquences synchronisées sont ensuite divisées en deux parts : les séquences requêtes et les séquences références. Les tests consistent à balayer toutes les séquences références afin d'attribuer une identité à une séquence requête en maximisant une mesure de similarité (1- distance euclidienne).

Pour chaque descripteur, une matrice de similarité (D) est construite en appariant toutes les séquences requêtes et références placées respectivement en lignes et en colonnes. Pour chaque ligne, une séquence requête est correctement identifiée si la similarité maximale correspond à une séquence référence ayant la même identité. Le taux de reconnaissances (TR) d'un descripteur est évalué en calculant le rapport entre le nombre de séquences requêtes correctement identifiées et le nombre total de séquences requêtes (exprimé en %).

Dans les sections suivantes, deux expérimentations sont présentées :

- La première réalisée sur la base CASIA (C), analyse l'impact des variations de vitesse et de modes de déplacement.
- La deuxième réalisée sur la base CASIA (A), ajoute les problèmes liés aux différences de point de vue.

Les valeurs indiquées en rouge dans les tableaux de résultats qui suivent représentent les meilleurs taux de reconnaissance obtenus pour chaque descripteur.

4.2.3.2 CASIA (C) : variations des modes de déplacement en mono-vue :

Les tests réalisés sont détaillés dans le tableau 4.3. Ils concernent une sélection de 30 personnes. Les performances de chaque descripteur sont analysées séparément pour les quatre modes de déplacement (Normale, Rapide, Lente et avec Sac), puis avec une base mixte destinée à évaluer la robustesse des descripteurs dans un scénario plus complexe en mélangeant

des séquences intégrant tous les types de déplacements.

	Séquences références	Séquences requêtes
Tests		
Normale	60 normales	60 normales
Rapide	30 rapides	30 rapides
Lente	30 lentes	30 lentes
Sac	30 avec sac	30 avec sac
Mixte	30 normales 30 avec sac 30 lentes 30 rapides	90 normales 30 avec sac 30 lentes 30 rapides

Tableau 4.3 – Tests réalisés sur la base CASIA(C).

Le tableau 4.4 donne les taux de reconnaissances exprimés en % pour tous les descripteurs avec les deux transformées : la Transformée de Fourier (T1) et la transformée de Hilbert sur le signal DWT (T2).

Tests Descripteurs	Normal		Rapide		Lente		Sac		Mixte	
	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
Pendule	29	22	27	27	20	23	37	40	28	22
Hanche	32	35	47	37	30	27	57	23	34	29
Genou	15	10	14	30	10	24	4	7	11	13
Jambe	32	22	7	43	24	43	30	30	20	35
Hauteur	7	22	7	33	7	23	13	40	17	31

Tableau 4.4 – Taux de reconnaissances obtenus avec les 2 transformations fréquentielles pour chaque type de déplacement (en %).

4.2.3.3 CASIA(A) : variations de point de vue

Trois tests ont été réalisés sur une sélection de 13 personnes de la base CASIA (A). Les deux premiers tests évaluent les descripteurs en appariant des séquences requêtes et références

de même point de vue (oblique ou fronto-parallèle). Le troisième test traite le problème du multi-vues en appariant des séquences requêtes en point de vue oblique avec des séquences référence en point de vue fronto-parallèle.

Point de vue oblique : Ce test est réalisé avec 13 séquences requêtes et 13 séquences de référence en sélectionnant les séquences obliques de la base CASIA(A) pour lesquelles les personnes sont de dos et s'éloignant de la caméra. Les résultats sont présentés dans le tableau 4.5.

Oblique Descripteurs	T1	T2
Pendule	27	10
Hanche	20	13
Genou	17	15
Jambe	7	23
Hauteur	7	23

Tableau 4.5 – Taux de reconnaissances en point de vue oblique (en %).

Point de vue fronto-parallèle : Par similitude avec le test précédent, ce test est réalisé avec 13 séquences requêtes et 13 séquences de référence en sélectionnant les séquences où les personnes se déplacent de gauche à droite (même sens que dans les séquences obliques). Les résultats sont présentés dans le tableau 4.6.

Fronto/parallèle Descripteurs	T1	T2
Pendule	27	23
Hanche	23	20
Genou	10	17
Jambe	13	25
Hauteur	7	27

Tableau 4.6 – Taux de reconnaissances en point de vue fronto-parallèle (en %).

Multi-points de vue : Dans ce test, 26 séquences requêtes en point de vue oblique (2 séquences / personne s'approchant de la caméra) ont été affrontées à 26 séquences de référence de point de vue fronto-parallèle (2 séquences / personne se déplacent de gauche à droite dans l'image). Les résultats de cette expérimentation sont présentés dans le tableau 4.7.

Multi-vues Descripteurs	T1	T2
Pendule	33	29
Hanche	29	15
Genou	12	8
Jambe	12	25
Hauteur	8	27

Tableau 4.7 – Taux de reconnaissances en multi-vues (en %).

4.2.3.4 Discussion

En considérant les taux de reconnaissance de tous les tests effectués, un choix peut déjà être fait pour la représentation fréquentielle la plus adaptée à chaque descripteur. En effet, la transformée de Fourier est clairement plus adaptée aux descripteurs de l'angle de la *Hanche* et du *Pendule*. Elle apparaît respectivement sept fois et cinq fois avec des taux de reconnaissance supérieurs dans les 8 tests. Tandis que le descripteur *Jambe* et la *Hauteur* obtiennent de meilleurs résultats en utilisant la transformée de Hilbert appliquée sur les signaux DWT (respectivement 6 et 8 fois supérieurs sur les 8 tests).

En observant la nature des signaux présentés dans la figure 4.7, l'angle de la *Hanche* et le descripteur *Pendule* sont plutôt périodiques et assez réguliers. Par conséquent, la transformée de Fourier est une transformation fréquentielle pertinente pour ces signaux. Par contre, les variations des descripteurs *Jambe* et *Hauteur* sont plutôt irrégulières (figure 4.7.b) et nécessitent donc une représentation plus complexe en temps-fréquence. La transformée de Hilbert appliquée sur le signal DWT est donc plus adéquate.

En ce qui concerne l'angle du *Genou*, les résultats sont un peu moins marqués, mais la transformée de Hilbert appliquée sur le signal DWT est tout de même supérieures dans 5

tests sur 8. Le choix de la transformation la plus adaptée augmente notablement les taux de reconnaissance de chaque descripteur et représente en cela un résultat très intéressant.

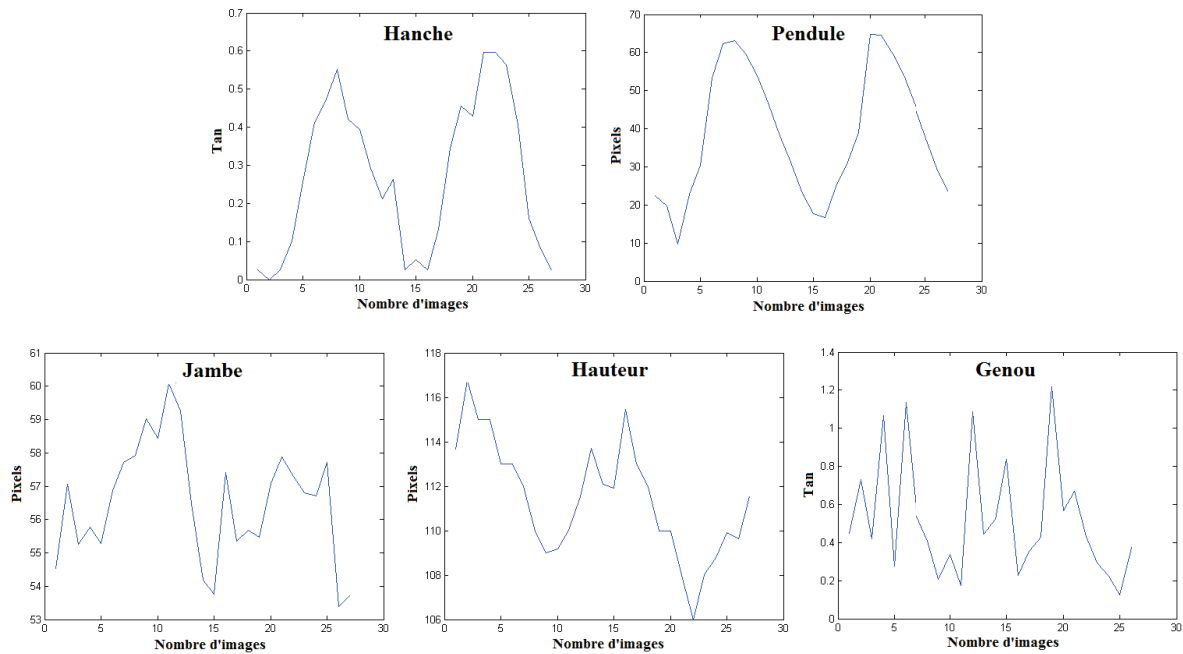


FIGURE 4.7 – Exemples de signaux pour chaque descripteur.

En conclusion, nous retiendrons pour les tests ultérieurs la transformation la plus adaptée à chaque descripteur (tableau 4.8).

	Transformation retenue	Nombre de fois supérieure sur les 8 tests
Descripteurs		
Pendule	Fourier	5
Hanche	Fourier	7
Genou	DWT + Hilbert	5
Jambe	DWT + Hilbert	6
Hauteur	DWT + Hilbert	8

Tableau 4.8 – Choix des transformations.

Globalement, les taux de reconnaissance du descripteur *Pendule* proposé sont satisfaisants comparés à ceux des descripteurs existants. En considérant la transformation fréquentielle la

plus appropriée pour chaque descripteur, il apparaît même en première position sur les test multi-points de vue.

4.2.3.5 Impact de la localisation des points sur les descripteurs

L'objectif de ce test est de comparer la détection automatique et la détection manuelle sur les performances des descripteurs afin de valider définitivement le processus de localisation automatique. Pour ce test, le descripteur *Hauteur* n'est pas considéré car il représente la distance entre le point le plus haut et le point le plus bas de la silhouette indépendamment des points d'intérêt détectés.

Nous avons considéré les mêmes données que celles de la section 4.2.1.3 (5 personnes chacune représentée par 4 séquences de marches différentes). Etant donné qu'il est difficile d'annoter manuellement un grand nombre de séquences, les performances des descripteurs sont calculées en appliquant une validation croisée (Leave One Out Cross Validation) qui est une méthode d'évaluation adaptée aux bases de données de taille réduite.

Pour chaque descripteur, deux matrices de similarité sont construites : une en utilisant les points d'intérêt détectés automatiquement et l'autre en utilisant les points de la vérité de terrain pointés manuellement. Chaque matrice de similarité contient les mesures de similarité (1- distance euclidienne) entre les 20 séquences placées respectivement en lignes (requêtes) et en colonnes (références). Les diagonales ne sont pas considérées dans l'évaluation (même séquence et même personne). Chaque ligne de matrice contient 3 appariements de la même personne sur des séquences différentes (3 types de marches différentes de celle de la requête) et 16 appariements de différentes personnes. Soit, au total, $3 \times 20 = 60$ appariements de la même personne et $16 \times 20 = 320$ appariements de différentes personnes dans chaque matrice.

Pour mesurer l'impact des erreurs de localisation des points sur les performances des descripteurs, trois critères sont analysés :

- Nombre de Bonnes Classifications (*BC*) : une séquences requête est correctement classée si la similarité maximale par ligne correspond à une séquence référence appartenant à la même personne.
- Nombre de Vrais Positifs (*VP*) : nombre de fois où les similarités correspondant à la

même personne sont supérieures à celles correspondant à des personnes différentes. Ce critère est maximisé par 60 puisque le total des appariements de la même personne est égal à 60 dans notre cas.

- Nombre de similarités Ambigües (*SA*) : nombre de fois où les similarités correspondant à des personnes différentes sont supérieures à celles correspondant à la même personne. Ce critère est maximisé par 320 puisque le total des appariements de différentes personnes est égal à 320 dans notre cas.

Descripteurs	Détection automatique			Détection manuelle		
	BC	VP	SA	BC	VP	SA
Hanche	8	12	253	7	13	164
Genou	2	3	264	3	3	253
Pendule	4	5	267	6	6	271
Jambe	10	16	112	7	9	176

Tableau 4.9 – Impact de la détection automatique des points sur les descripteurs.

D’après le tableau 4.9, nous pouvons constater que la détection automatique des points affecte légèrement le processus de caractérisation de la démarche. Pour tous les descripteurs, le nombre de bonnes classifications (*BC*) et l’ordonnancement des séquences référence par rapport aux séquences requêtes (*VP* et *SA*) sont proches des valeurs obtenues avec la réalité de terrain. Il est à noter que le descripteur *Jambe* est le plus sensible à la position des points. En effet, l’angle de la *Hanche* et du *Genou* sont représentés par les variations de leur tangente ce qui réduit considérablement la sensibilité à la localisation des points. De même, le descripteur *Pendule* proposé exploite une distance entre deux points spatialement proches (la projection orthogonale du point de la hanche sur le segment des chevilles et le milieu de ce segment), ce qui le rend assez robuste. Alors que le descripteur *Jambe* qui représente la distance entre la hanche et la cheville la plus éloignée est directement impacté par les erreurs de localisation de ces deux points d’articulation spatialement éloignés.

Les performances du descripteur *Jambe* avec les points automatiques sont légèrement plus élevées que celles de la vérité de terrain. Ceci peut s’expliquer visuellement par l’analyse des variations spatio-temporelles de ce descripteur. En effet, les erreurs générées par le processus de détection automatique ajoutent quelques irrégularités dans le signal permettant ainsi de mieux séparer des personnes différentes.

La figure 4.8 illustre les variations spatio-temporelles des descripteurs pour la même personne selon les différentes catégories de mouvement. Pour cette figure, nous avons choisi de représenter les descripteurs bruts (sans appliquer les transformations fréquentielles) pour faciliter l'interprétation des courbes.

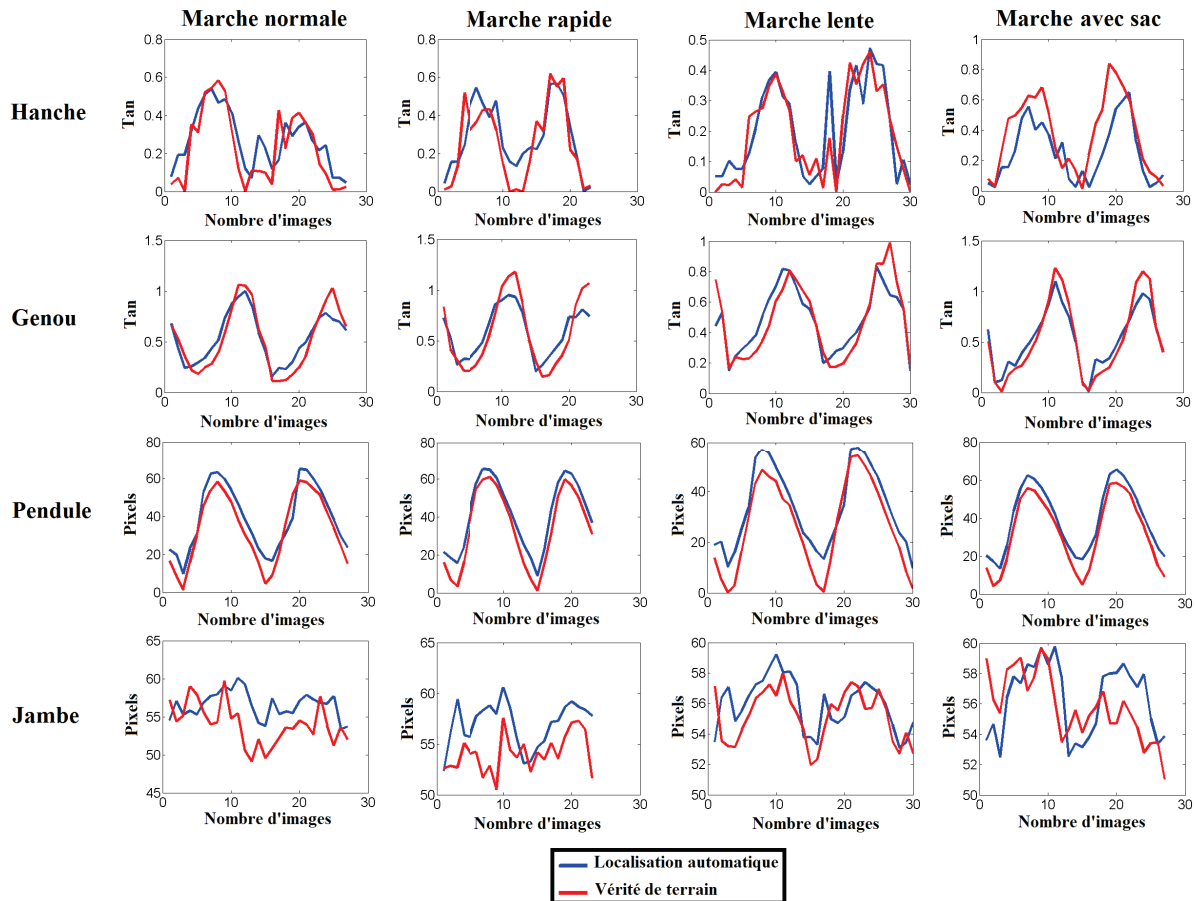


FIGURE 4.8 – Variations spatio-temporelles des descripteurs extraits à partir de la localisation automatique des points versus la vérité de terrain.

Les résultats attestent clairement la fiabilité du processus de détection automatique proposé. Les points détectés automatiquement sont spatialement proches de ceux de la vérité de terrain (section 4.2.1.3) et ils ne perturbent que légèrement le processus de caractérisation de la démarche.

Afin de renforcer encore la phase de caractérisation, la section suivante propose de fusion-

ner les descripteurs les plus performants.

4.2.4 Fusion des descripteurs

4.2.4.1 Pondération des descripteurs retenus

Compte tenu de la complexité, l'utilisation d'un seul descripteur reste souvent insuffisante pour bien caractériser les piétons. La fusion de plusieurs descripteurs est une stratégie intéressante pour améliorer les performances.

La fusion est réalisée avec l'algorithme AdaBoost, déjà présenté dans le chapitre 2 de la thèse dans le cadre de la fusion des descripteurs d'apparence (section 2.5.1). La figure 4.9 présente les valeurs des poids, normalisées à 100%, qui sont attribuées par l'algorithme AdaBoost aux différents descripteurs retenus en appliquant la transformée la plus appropriée pour chacun.

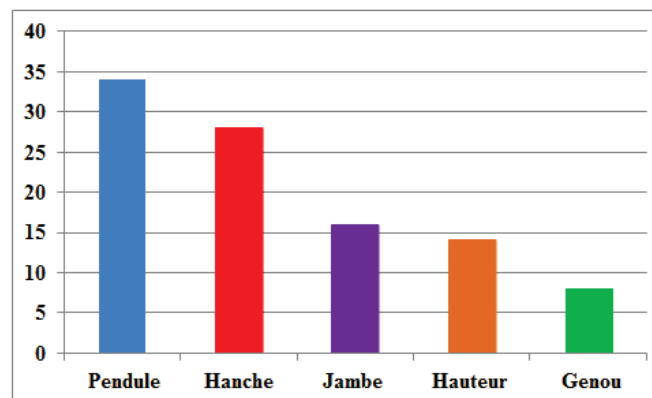


FIGURE 4.9 – Poids attribués aux descripteurs retenus (en pourcentages).

La section suivante présente les taux de reconnaissance obtenus par la fusion des descripteurs sur les bases CASIA(C) et CASIA(A) dans les mêmes conditions que lors de l'étude comparative précédente. Pour chaque test, une matrice de similarité générique (*MSG*) est calculée par la combinaison linéaire des matrices de similarité propres à chaque descripteur pondérées par leur poids relatif (équation 2.13).

4.2.4.2 Résultats

Les tableaux 4.10 et 4.11 présentent les taux de reconnaissance après l'étape de fusion comparés à ceux obtenus par le meilleur descripteur pour chaque test.

TR	CASIA(C)				
	Normale	Rapide	Lente	Sac	Mixte
Fusion	57	64	60	77	54
Meilleur descripteur	<i>Hanche:32</i>	<i>Hanche:47</i>	<i>Jambe:43</i>	<i>Hanche:57</i>	<i>Jambe:35</i>

Tableau 4.10 – Taux de reconnaissance par la fusion des descripteurs face aux variations de modes de déplacement (en %).

TR	CASIA(A)		
	Oblique	Fronto/parallèle	Multi-vues
Fusion	61	63	54
Meilleur descripteur	<i>Pendule:27</i>	<i>Pendule:27</i>	<i>Pendule:33</i>

Tableau 4.11 – Taux de reconnaissance par la fusion des descripteurs face aux variations de point de vue avec variations de vitesse (en %).

Les performances sont considérablement améliorées par la fusion de plusieurs descripteurs. Face aux variations des modes de déplacement, les taux de reconnaissances sont rehaussés d'environ *20%* dans la première expérimentation et de *35%* dans la seconde si on considère un seul point de vue. En combinant des forts changements de point de vue et de vitesses de déplacement (expérimentation multi-vues), l'amélioration est de *20%* malgré la difficulté du contexte expérimental. Cela prouve que la fusion des descripteurs spatio-temporels représente une stratégie très appropriée pour identifier les piétons par la démarche.

4.2.4.3 Comparaison avec d'autres approches

Jusqu'à présent nous avons effectué une étude comparative entre les descripteurs issus du squelette et nous avons mesuré l'apport de la fusion sur les taux de reconnaissance. Nous allons maintenant comparer les performances de l'approche proposée à trois approches existantes : deux approches sans-modèle *GEI* [104], qui sert de référence, et une version améliorée basée sur la partie inférieure du *GEI* [105] que nous appelons *Inf-GEI*, et une approche basée-modèle que l'on nomme *Projections* [89]. Ces approches ont déjà été présentées dans l'état de l'art du chapitre 3.

L'approche GEI est basée sur l'image moyenne calculée à partir d'un cycle de déplacement. La méthode Inf-GEI qui en découle, sera utilisée dans un mode supervisé pour extraire la zone d'intérêt à partir de GEI (figure 4.10) car il offre les meilleurs performances. Compte tenu de la taille du vecteur caractéristique extrait des images, les deux approches GEI et Inf-GEI nécessitent une phase de réduction des données par l'analyse en composantes principales (ACP).

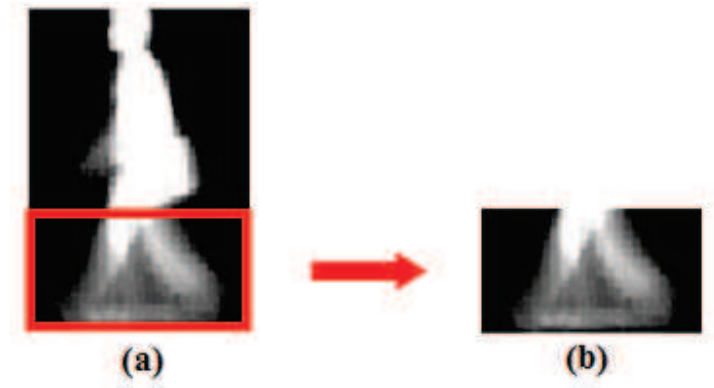


FIGURE 4.10 – Les zones d'intérêt pour (a) GEI et (b) Inf-GEI.

L'approche Projection est basée sur l'extraction de signaux 1D représentant les distances entre les silhouettes et les 4 cotés de leur boîte englobante. Les vecteurs distances de chaque direction sont ensuite cumulés temporellement dans des vecteurs, puis réduits par ACP. Ces descripteurs contiennent donc à la fois les caractéristiques structurelles de forme et des caractéristiques dynamiques de mouvement pour identifier les personnes par la démarche.

Nous avons comparé ces 3 approches (GEI, Inf-GEI et Projection) à l'approche proposée selon les mêmes protocoles de test que précédemment. Les tableaux 4.12 et 4.13 illustrent respectivement les taux de reconnaissance obtenus face aux changements de modes de déplacements (CASIA(C)), et ceux face aux fortes variations de point de vue avec variations de vitesse (CASIA(A)).

TR	CASIA(C)				
	Normale	Rapide	Lente	Sac	Mixte
GEI	98	97	100	93	98
Inf-GEI	97	93	97	93	97
Projection	92	93	83	83	89
Fusion	57	64	60	77	54

Tableau 4.12 – Taux de reconnaissances face aux variations de modes de déplacement (en %).

TR	CASIA(A)		
	Oblique	Frontal/parallèle	Multi-vues
GEI	100	100	15
Inf-GEI	77	69	15
Projection	40	65	8
Fusion	61	63	54

Tableau 4.13 – Taux de reconnaissances face aux variations de point de vue avec variations de vitesse (en %).

D'après les résultats du tableau 4.12, nous pouvons constater que les approches *GEI* et *Inf-GEI* offrent des taux de reconnaissance supérieurs à 90% pour tous les tests mono-vue. Sur la première expérimentation de la base CASIA(C), le très bon comportement de *Inf-GEI* par rapport à *GEI* confirme l'intérêt de se concentrer sur la partie inférieure en vue latérale pour analyser la démarche [83] [91] [105]. Dans le même contexte expérimental, l'approche *Projection* offre des performances légèrement en-deçà et l'approche proposée de fusion de descripteurs issus du squelette ne dépasse pas les 60% de taux de reconnaissance. Ceci prouve que les approches directes (sans-modèle) sont les plus adaptées pour identifier les personnes quand les séquences sont filmées en mono-vue.

Par contre, dans la seconde expérimentation qui présente un contexte multi-vues (dernière colonne du tableau 4.13), l'ordre des approches est inversé. En effet, les performances des approches *GEI* et *Inf-GEI* ont largement chuté jusqu'à atteindre un taux de reconnaissance

de 15%. Cela prouve que les approches sans-modèle sont très sensibles aux variations importantes de point de vue. Dans le même sens, les performances de l'approche projection sont fortement dégradées et elle permet de classer uniquement 8% des exemples requêtes, alors que l'approche fusion proposée permet d'assurer un taux de reconnaissance encourageant de 54%, et montre des performances assez stables entre mono et multi-vues.

Etant donné que l'approche proposée est basée sur la partie inférieure du corps humain (4/5 descripteurs sont extraits de cette région), les performances obtenues de 60% dans les tests mono-vue de la seconde expérimentation peuvent être considérées comme assez bonnes comparées à ceux de l'approche Inf-GEI qui utilise aussi la partie inférieure.

D'après les résultats obtenus, nous pouvons conclure que :

- Les approches sans-modèle restent les plus appropriées pour identifier les personnes par la démarche en mono-vue avec des variations de modes de déplacement,
- l'approche proposée, utilisant le squelette, assure quant à elle un bon compromis en termes de taux de reconnaissance mono/multi-vues, ce qui prouve l'intérêt de son application dans un contexte non-contrôlé.

4.3 Identification des personnes en vue frontale

La majorité des approches de la littérature se sont focalisées sur l'identification des personnes par la démarche en vue latérale (fronto-parallèle ou oblique). Très peu se sont intéressées à la reconnaissance des personnes en vue frontale car dans ce contexte, le faible mouvement des silhouettes observées rend la phase de caractérisation plus délicate.

Dans cette partie, nous introduisons une nouvelle approche d'identification en vue frontale basée sur la distribution du mouvement dans les silhouettes. Les performances de l'approche proposée sont ensuite comparées à plusieurs travaux existants de la littérature.

4.3.1 Analyse des mouvements en vue frontale

En vue frontale, les mouvements des silhouettes sont assez réduits étant donné que les personnes se déplacent en un bloc compact. La phase de caractérisation est alors assez diffi-

cile. Pourtant le cerveau humain est capable de reconnaître une personne en vue frontale de loin. Comment fait-il ?

En analysant les mouvements, nous sommes capable de détecter quelques particularités qui diffèrent d'une personne à l'autre. Par exemple, une personne bouge un peu, trop ou jamais ses bras ou plus ou moins sa tête pendant le déplacement, une autre aura la plus grande foulée. En s'appuyant sur cette constatation, nous proposons de fonder notre approche sur l'hypothèse suivante : nous pouvons correctement caractériser une personne par la démarche en point de vue frontale si nous parvenons à bien modéliser la distribution des mouvements de la personne dans toute sa silhouette. Pour assurer cet objectif, nous introduisons dans la section suivante deux nouvelles représentations appelées Représentation Fréquentielle de la Démarche (RFD) et Représentation de l'Enveloppe de la Démarche (RED). Ces deux représentations sont calculées à partir des images binaires des silhouettes alignées et normalisées. Toutes les boîtes englobantes sont redimensionnées en 128×48 pixels par interpolation bicubique. Ces silhouettes sont ensuite translatées pour aligner leur axe de symétrie vertical au milieu de la boîte englobante. La figure 4.11 illustre quelques silhouettes prétraitées pour la même personne.



FIGURE 4.11 – Silhouettes alignées et normalisées pour la même personne.

4.3.1.1 Représentation Fréquentielle de la Démarche (*RFD*)

La Représentation Fréquentielle de la Démarche est un modèle 2D en niveaux de gris illustrant la distribution des mouvements des silhouettes. Cette représentation est basée sur les images de différences entre images successives de la séquence. Pour chaque pixel, nous cumulon le nombre de fois où il a changé d'état de 0 à 1 ou vice versa entre 2 images successives (0 pixel inactif, 1 pixel actif). Une étape de normalisation est ensuite appliquée en divisant par la valeur maximale. Le processus de calcul de la *RFD* est présenté dans l'algorithme 4.1.

Algorithme 4.1 Calcul de la Représentation Fréquentielle de la Démarche (RFD).

Initialisation

T : le nombre d'images dans la séquence de marche

H et L : hauteur et largeur de la boîte englobant les silhouettes (en pixels)

$$RFD(i, j) = 0; \quad \forall i = 1..L \text{ et } \forall j = 1..W$$

Processus

Pour $t = 1..T$

Si le pixel $I(i, j)$ a changé d'état entre 2 images successives alors $GFR(i, j) + = 1$; avec I l'image de la silhouette.

Fin

$$RFD(i, j) = \frac{RFD(i, j)}{M} \text{ où } M \text{ est le maximum de tous les } GFR(i, j)$$

Un exemple de RFD est présenté dans la figure 4.12.



FIGURE 4.12 – Séquence de silhouettes binaires (10 images régulièrement espacées dans la séquence) et les représentations RFD et RED .

La figure 4.13 illustre quelques exemples de la RFD pour 3 personnes et 4 séquences différentes de la base CASIA(A).

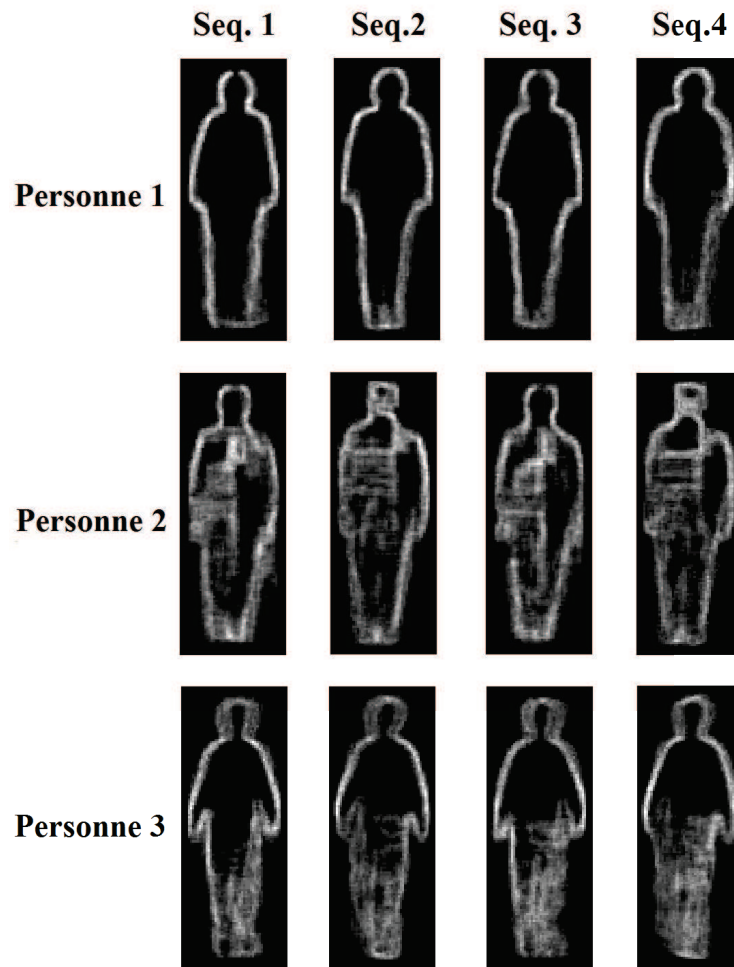


FIGURE 4.13 – Représentation Fréquentielle de la Démarche (*RFD*) pour 3 personnes et 4 séquences différentes de la base CASIA(A).

4.3.1.2 Représentation de l'Enveloppe de la Démarche (*RED*)

La Représentation de l'Enveloppe de la Démarche est un modèle 2D illustrant la distribution du gradient du mouvement. Elle permet d'extraire une enveloppe binaire indiquant l'existence ou non de mouvement pour chaque pixel de la silhouette durant une séquence de déplacement. Pour construire cette représentation, nous commençons par sommer toutes les images de différence entre les images successives de la séquence. Ensuite, un test booléen est effectué pour chaque valeur de la matrice, afin de détecter les pixels modifiés ou non (pixels modifiés=1 et 0 sinon). Le processus de calcul de la *RED* est détaillé par l'algorithme 4.2.

Algorithme 4.2 Calcul de la Représentation de l'Enveloppe de la Démarche (*RED*).

Initialisation

T : le nombre d'images dans la séquence de marche

H et **L** : hauteur et largeur de la boîte englobant les silhouettes (en pixels)

$S(i, j) = 0$; et $RED(i, j) = 0$; $\forall i = 1..L$ et $\forall j = 1..W$

Processus

Pour $t = 1..T - 1$

$S(i, j) += |I(i, j, t + 1) - I(i, j, t)|$ avec I l'image de la silhouette.

Fin

$$RED(i, j) = \begin{cases} 1 & \text{Si } S(i, j) \neq 0 \\ 0 & \text{sinon} \end{cases}$$

La figure 4.13 illustre un exemple de la *RED* et la figure 4.14 présente les *RED* pour les mêmes personnes et séquences que la figure 4.13.

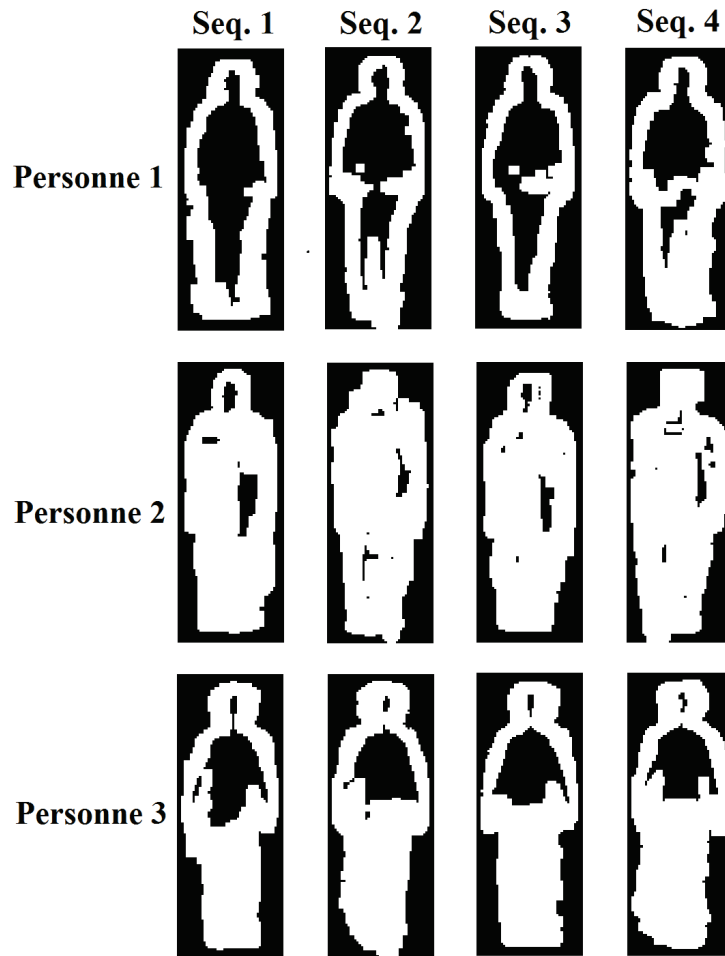


FIGURE 4.14 – Représentation de l’Enveloppe de la Démarche (*RED*) pour 3 personnes et 4 séquences différentes de la base CASIA(A) - même sélection que la figure 4.13.

Ces deux représentations mettent en évidence la distribution du mouvement dans les silhouettes qui est invisible quand on regarde les séquences originales. Elles offrent aussi l’avantage d’être rapides et faciles à extraire.

En observant les figures 4.13 et 4.14 nous pouvons remarquer que les représentations *RFD* et *RED* présentent de grandes similitudes intra-personnes et des différences inter-personnes. En effet, la personne 1 a un mouvement uniformément réparti sur toutes les parties de son corps (tête, bras et jambes), la personne 2 marche en agitant fortement tout son corps, et enfin la personne 3 oscille la tête et les jambes et légèrement les bras durant son déplacement. Cela confirme que la distribution du mouvement dans les silhouettes peut, si elle est bien

caractérisée, être un marqueur d'identification des personnes.

4.3.1.3 Caractérisation de la distribution des mouvements

Etant donné que nous avons une représentation en niveaux de gris (*RFD*) et une représentation binaire (*RED*) deux caractérisations différentes sont appliquées.

Pour décrire une signature de démarche par la dispersion du mouvement, deux alternatives ont été adoptées dans la littérature. La première consiste à considérer les projections horizontales et verticales (respectivement sur les lignes et colonnes) des silhouettes comme dans le modèle de frise [102] [103]. La deuxième tient compte de la dispersion des contours en se référant au centroïde des silhouettes [96] [97] [98].

Pour caractériser la *RFD*, nous proposons deux descripteurs exploitant les mêmes alternatives. Le premier calcule l'entropie des intensités pour chaque ligne (respectivement colonne) de la *RFD* sans considérer les pixels noirs (absence de mouvement). Toutes les valeurs d'entropies sont stockées dans un seul vecteur caractéristique de taille $128 + 48 = 176$ éléments (appelé *2D-Entropie*). Le deuxième descripteur modélise la distribution du mouvement par rapport au centroïde de la silhouette. Les pixels sont d'abord organisés en n classes en fonction de leur intensité. Chaque classe est ensuite représentée par un histogramme de distribution des distances euclidiennes entre les pixels de la classe et le centroïde de la silhouette. Finalement, un vecteur caractéristique (nommé *Hist-distances*) de taille $n \times b$ est obtenu. Nous avons fixé $n = b = 5$ pour assurer le meilleur compromis entre la taille du vecteur caractéristique et le taux de reconnaissance.

Comme la *RED* est un modèle binaire, une analyse basée sur la forme est bien adaptée pour la caractérisation. Nous avons alors exploité les descripteurs de forme les plus utilisés dans la littérature qui sont l'histogramme de gradients orientés [35], les moments de Zernike [60] et de Hu [61]. Ces descripteurs ont été déjà introduits dans le chapitre 2. Ils sont considérés parmi les plus robustes pour caractériser les formes complexes.

4.3.2 Evaluation des descripteurs frontaux

Dans cette section, nous commençons par réaliser une étude comparative entre tous les descripteurs introduits dans la section précédente. Ensuite, nous évaluons l'apport de la fusion des plus robustes d'entre eux sur les taux de reconnaissance.

4.3.2.1 Protocole de test et résultats

Nous avons considéré les séquences de la base CASIA (A) en vue frontale. Au total 80 séquences sont utilisées pour les tests (20 personnes différentes, chacune est représentée par 4 séquences de marche).

Pour comparer les performances des descripteurs frontaux, nous avons appliqué le même protocole de test que dans la section 4.2.3. Les 80 séquences frontales ont été divisées en deux groupes distincts : requêtes et références. Le test consiste à attribuer une identité à chaque séquence requête en la comparant à toutes les séquences références en maximisant la similarité. Le tableau 4.14 représente les taux de reconnaissance obtenus pour tous les descripteurs caractérisant respectivement les représentations *RFD* et *RED*.

Représentations	Descripteurs	TR
RFD	2D-Entropie (RFD)	62.5
	Hist-distances (RFD)	62.5
RED	HOG (RED)	43.75
	Zernike (RED)	62.5
	Hu (RED)	25

Tableau 4.14 – Taux de reconnaissance pour les descripteurs comparés (en %).

Les résultats présentés dans le tableau 4.14 montrent que les deux descripteurs proposés (*2D-Entropie* et *Hist-distances*) pour la *RFD* ainsi que les moments de Zernike pour la *RED* offrent les meilleurs taux de reconnaissance. Ces trois descripteurs seront par conséquent retenus pour l'étape de fusion.

4.3.2.2 Apport de la fusion

Etant donné que les descripteurs sélectionnés offrent tous le même taux de reconnaissance (62.5%), nous avons opté pour une méthode de fusion simple en testant toutes les combinaisons possibles de descripteurs. En effet, les taux de reconnaissances pour chaque test sont calculés sur une matrice générique sommant les matrices de similarité des descripteurs rete-

nus. Les performances de la fusion pour toutes les combinaisons testées sont illustrées dans le tableau 4.15.

Tests	TR
2D-Entropie + Zernike	77.5
Hist-distances + Zernike	82.5
2D-Entropie + Hist-distances	80
2D-Entropie + Hist-distances + Zernike	90

Tableau 4.15 – Taux de reconnaissance pour toutes les combinaisons de fusion (en %).

D’après le tableau 4.15, le meilleur taux de reconnaissance de *90%* est obtenu en fusionnant les trois descripteurs : *2D-Entropie* et *Hist-distances* modélisant la *RFD* et les moments de Zernike caractérisant la *RED*. On peut alors conclure que :

- La distribution du mouvement dans l’image de la silhouette représente une signature robuste pour caractériser les personnes en vue frontale,
- la Représentation Fréquentielle de la démarche (*RFD*) ainsi que la Représentation de l’Enveloppe de la Démarche (*RED*) proposées sont très complémentaires,
- les deux descripteurs introduits (*2D-Entropie* et *Hist-distances*) ainsi que les moments de Zernike sont appropriés pour modéliser les deux représentations.

4.3.3 Etude comparative avec des approches existantes

Dans cette section, nous comparons les performances de l’approche de fusion proposée à d’autres travaux de la littérature consacrés à l’identification des personnes par la démarche en point de vue frontal. Nous allons tout d’abord introduire les approches existantes, puis nous présentons les résultats.

4.3.3.1 Travaux sur l’identification en vue frontale

L’identification des personnes par la démarche en vue frontale n’a été que très peu abordée par rapport aux nombreux travaux existants en point de vue latéral. Quelques travaux

en multi-vues testent des vues frontales [126] [127], et on pointe deux approches spécialement conçues pour la vue frontale [97] [125].

Wang et al. [97] exploitent les variations temporelles de la forme des silhouettes pour caractériser une personne par la démarche. La première étape consiste à modéliser chaque silhouette par un vecteur complexe qui représente les distances entre le centroïde de la silhouette et des points du contour. Puis, les vecteurs représentant un passage de personne sont moyennés après avoir été centrés afin de limiter l'impact de la translation. L'identification des personnes est réalisée en utilisant les vecteurs propres relatifs aux plus grandes valeurs propres de chaque signature moyenne.

Chen et Gao [125] projettent les silhouettes sur un disque dans le plan polaire. Le disque est uniformément divisé en plusieurs régions d'angle et distances variables par rapport au point central. Finalement, la phase de caractérisation consiste à calculer le nombre de pixels de la silhouette situés dans chaque région polaire.

L'approche proposée par Goffredo et al. [126] consiste à construire une signature volumique de la démarche en plaçant les points de contour des silhouettes dans un espace 3D (x, y, t). La reconnaissance des personnes est réalisée en fusionnant les moments de Hu extraits de cette représentation volumique avec d'autres caractéristiques telles que le nombre d'images dans un cycle, la hauteur de la silhouette et la variation maximale de la largeur de la silhouette durant le déplacement.

Tous les travaux précédents appliquent une classification par KPPV.

Barnich et Droogenbroeck [127] ont proposé de couvrir la surface intérieure des silhouettes par des rectangles adaptatifs. Une étape de filtrage est réalisée afin d'assurer un bon compromis entre l'espace couvert et le nombre de rectangles utilisé. Pour chaque image, deux histogrammes, qui représentent les distributions respectives des largeurs et des hauteurs des rectangles, sont concaténés. Finalement, un vote majoritaire sur l'ensemble des signatures est appliqué afin d'identifier une séquence requête par rapport aux séquences références.

Pour la comparaison qui suit, nous avons retenu les approches [97] [125] [126] qui ont été évaluées sur la base CASIA(A).

4.3.3.2 Performances des approches comparées

Les performances en termes de taux de reconnaissance des quatre approches comparées sont présentées dans le tableau 4.16.

Approches	Direction de la marche	TR
Wang et al., 2003 [88]	Toutes	81.25
Chen et Gao, 2007 [117]	Se rapprochant de la caméra	75
	S'éloignant de la caméra	65
Goffredo et al., 2008 [118]	Toutes	91
Approche proposée	Toutes	90

Tableau 4.16 – Comparaison des approches en vue frontale sur la base CASIA(A).

D'après le tableau 4.16, nous pouvons constater que l'approche proposée est performante car elle offre un taux de reconnaissance similaire à l'approche de Goffredo [126] (qui est la plus robuste) en utilisant des descripteurs plus simples. La modélisation de la distribution du mouvement proposée représente donc une stratégie appropriée pour identifier les personnes par la démarche en vue frontale.

4.4 Conclusion

Dans ce chapitre, nous avons présenté deux approches différentes pour identifier les personnes par la démarche respectivement en point de vue latéral (fronto-parallèle et oblique) et frontal. En vues latérales, nous avons exploité les variations spatio-temporelles calculées à partir d'un squelette de la partie inférieure des silhouettes. Les résultats obtenus montrent clairement que l'approche proposée, si elle n'est pas la plus performante en mono-vue, se révèle très efficace dans un contexte multi-vues comparée aux approches existantes. En vue frontale, nous avons élaboré deux modèles de distribution du mouvement dans les silhouettes. Les performances des descripteurs proposés offrent des taux de reconnaissances très similaires par rapport aux meilleurs approches de la littérature.

Dans le chapitre suivant, nous présentons un système complet fusionnant l'apparence et la démarche pour identifier les personnes dans un contexte multi-caméras.

Chapitre 5

Identification des personnes par l'association de l'apparence et de la démarche

5.1 Introduction

Dans les chapitres précédents, nous avons présenté trois approches destinées à identifier les personnes par l'apparence ou par la démarche. Les différentes approches ont été testées et comparées dans plusieurs contextes expérimentaux variant les modes de déplacement et/ou les points de vue (fronto-parallèle, oblique, frontal). Le but de ce chapitre est de proposer un système complet associant l'apparence et la démarche en vue d'améliorer les performances de reconnaissance des personnes en multi-caméras. Ce système sera testé dans un contexte réel de vidéosurveillance sur une base d'images intégrant les principales difficultés liées à ce domaine.

La section suivante introduit le processus d'identification des personnes proposé couplant l'apparence et la démarche. La troisième section est réservée à la description de la base d'images. Finalement, tous les résultats sont présentés et discutés dans la section 5.4.

5.2 Approche proposée

L'identification par l'apparence est basée sur la fusion des descripteurs de couleurs ayant donné les meilleurs résultats (section 2.5.2). Pour la démarche, l'identification en vue latérale est réalisée en fusionnant des descripteurs spatio-temporels les plus robustes (section 4.2.4.2),

et l'identification en vue frontale est effectuée par fusion des descripteurs de distribution des mouvements dans la silhouette (section 4.3.2.2). L'ensemble des descripteurs utilisés dans les différents modules est présenté dans le tableau 5.1 .

Approches	Point de vue	Descripteurs
Apparence	Multiple	• Histogrammes de couleurs (RGB, HSV, YCbCr) • Spatiogrammes (RGB, HSV, YCbCr) • Histogrammes régionaux (RGB, YCbCr)
Démarche	Latéral	• Pendule (FT) • Hanche (FT) • Genou (DWT+HT) • Jambe (DWT+HT) • Hauteur (DWT+HT)
	Frontal	• 2D-Entropie (RFD) • Hist-distances (RFD) • Moments de Zernike (RED)

Tableau 5.1 – Descripteurs utilisés dans les différents modules d'identification.

Le processus d'identification proposé est illustré dans la figure 5.1. Il est composé de trois étapes :

- Un module primaire se charge de l'extraction des personnes dans les séquences vidéos pour réaliser l'identification des personnes par l'apparence,
- Un test detecte les passages ambigus entre caméras par seuillage de la similarité,
- Un module secondaire propose de renforcer les ré-identifications par une analyse de la démarche.

Les étapes sont détaillées dans les sections suivantes. La démarche n'est pas appliquée en même temps que l'apparence dans ce travail pour deux raisons. La première est directement liée au mode de traitement car le modèle d'apparence opère image-par-image, alors que le modèle de démarche travaille par séquence, et la prise de décision sur l'identité des personnes

est différée après analyse de toute la séquence. La deuxième raison repose sur les performances des deux modules. En effet, l'apparence est généralement plus discriminante surtout que pour les déplacements en vues latérales les taux de reconnaissances par la démarche étant généralement limités à 60% (section 4.2.4.2).

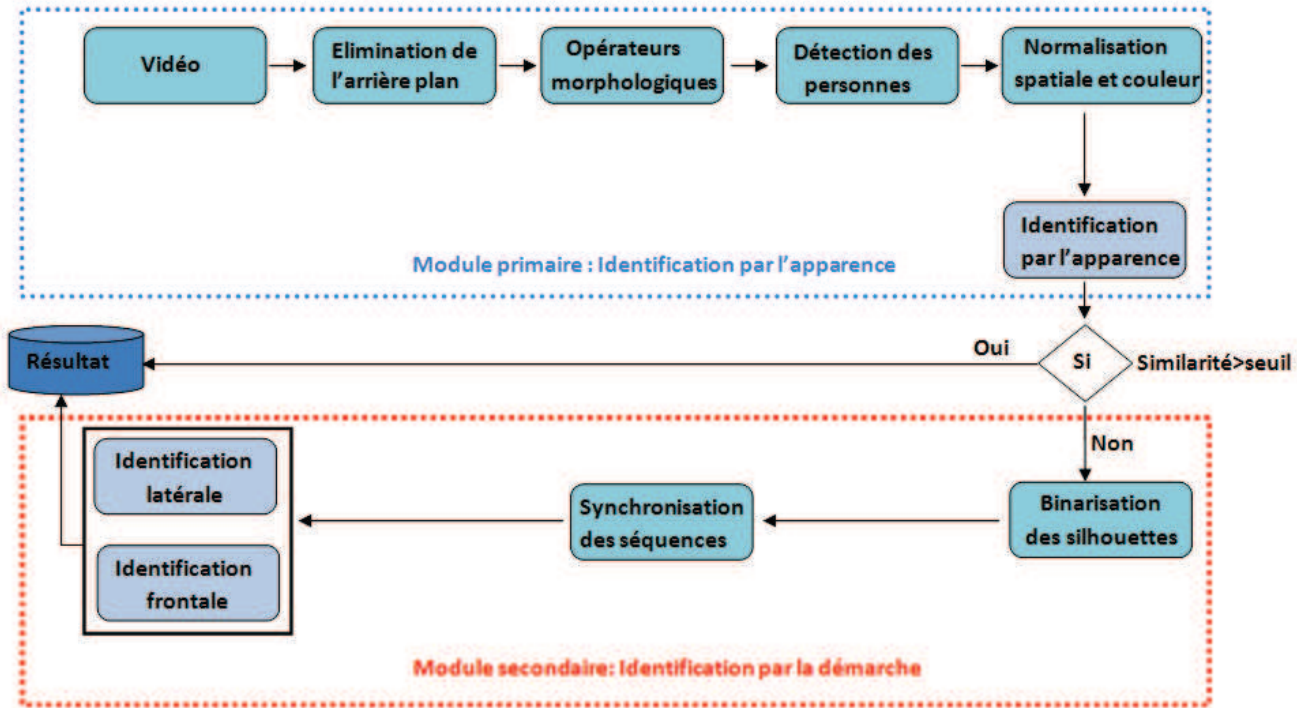


FIGURE 5.1 – Processus d'identification des personnes couplant l'apparence et la démarche.

5.2.1 Identification des personnes par l'apparence

Cette étape consiste à conserver l'identité des personnes entre des images successives. Le modèle d'apparence est appliqué à la fois dans le cas d'une identification mono-caméra (quand les images sont prises par la même caméra) ou encore dans le cas de ré-identification (quand les images sont prises par deux caméras différentes en changeant de point de vue).

Le processus d'identification des personnes par l'apparence est basé sur une maximisation de la similarité sur l'ensemble des signatures $A = \{APP_1, APP_2, \dots, APP_n\}$ représentant les n personnes présentes dans l'image à l'instant t et l'ensemble des signatures $A' = \{APP'_1, APP'_2, \dots, APP'_n\}$ modélisant les m personnes dans l'image à l'instant $t+1$ (voir algorithme ci-après).

Algorithme 5.1 Processus d'identification des personnes par l'apparence.

Tant que $A \neq \emptyset$ ou $A' \neq \emptyset$

$\{APP, APP'\} = \max(APP_{1..n}, APP'_{1..m})$: sélection du couple de signature $\{APP, APP'\}$ maximisant la similarité

$A = A - \{APP\}$ et $A' = A' - \{APP'\}$: retirer les signatures sélectionnées à partir des ensembles A et A'

Fin

Finalement, le résultat est un graphe (figure 5.2) où chaque nœud représente l'ensemble d'images attribué à la même identité en mono-caméra et les arcs représentent la conservation de l'identité en multi-caméra (cas des ré-identifications).

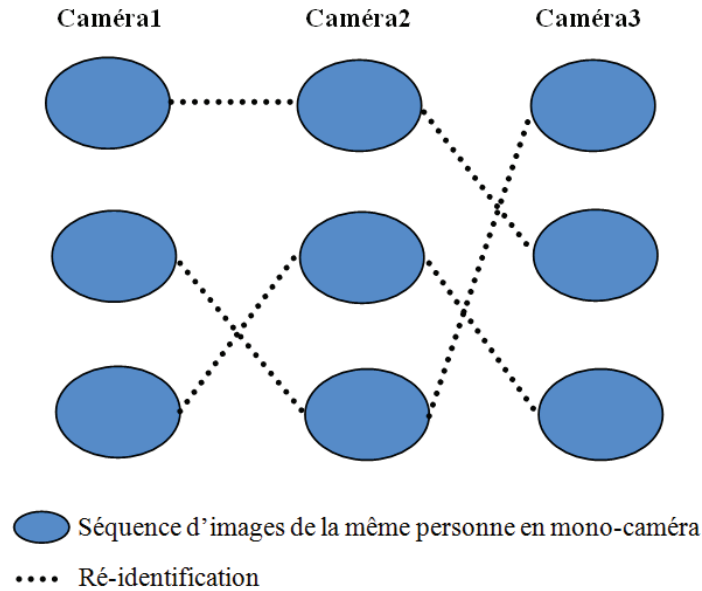


FIGURE 5.2 – Représentation du résultat sous forme de graphe.

5.2.2 Test d'ambigüité

Généralement, les modèles d'apparence sont plus robustes et efficaces dans un contexte d'identification mono-caméra que dans un contexte de ré-identification multi-caméras. Ceci est dû aux variations de pose, de conditions d'éclairage et d'échelle ainsi que la variabilité des réponses colorimétriques entre les caméras du réseau. Afin de renforcer éventuellement l'identification par la démarche, nous proposons un test pour détecter les passages ambigus

par seuillage de la similarité. En effet, un passage est jugé « ambigu » si la similarité d'un arc de ré-identification est inférieure à un seuil. La valeur de ce seuil sera discutée lors de l'étude expérimentale dans le paragraphe 5.4.1.

5.2.3 Consolidation par analyse de la démarche

Cette étape consiste à corriger les passages jugés « ambigu » en mettant à jour les arcs de ré-identification par l'analyse de la démarche. Le processus d'identification des personnes est basé sur une maximisation de la similarité entre deux ensembles de signatures temporelles de démarche $A = \{D_1, D_2, \dots, D_n\}$ et $A' = \{D'_1, D'_2, \dots, D'_m\}$ (algorithme 5.1) .

Etant donné que nous traitons différemment les vues latérale et frontale pour caractériser la démarche, les signatures dans les deux ensembles A et A' doivent correspondre à un même point de vue. Donc, avant de faire l'appariement, nous devons construire l'ensemble des signatures déjà labélisées A ayant le même point de vue que l'ensemble des signatures anonymes A' . Pour cela, nous utilisons des connaissances *a priori* sur le réseau de caméras, pour estimer à chaque intervalle l'orientation de la personne (latérale ou frontale).

5.3 Constitution d'une base de séquences vidéos

La plupart des bases d'images de la littérature sont destinées à évaluer l'identification de personnes par l'apparence ou par la démarche et non les deux à la fois. En effet, dans les bases d'images couleurs, les personnes ne sont pas filmées selon les différents points de vue, et dans les bases conçues pour analyser la démarche, les images des silhouettes sont généralement binaires et ne donnent pas accès à la couleur. Pour ces raisons, nous avons choisi de construire une base de séquences vidéo où les personnes sont filmées en couleur selon différents points de vue.

La construction de la base a été réalisée avec 3 personnes différentes dans un environnement intérieur. Chaque personne est capturée suivant 3 points de vues différents (frontal, oblique et fronto-parallèle). Chaque scène est réalisée 4 fois dans la même salle mais avec deux conditions d'éclairage différentes. La première prise est faite avec une source de lumière blanche et la deuxième prise avec une source de lumière jaune. La base de séquences représente ainsi plus de mille images de test. La figure 5.3 illustre quelques images de la base représentant les différentes problématiques citées.



FIGURE 5.3 – Présentation de la base avec variations des conditions d'éclairage et de point de vue.

Comme le montre la figure 5.3, chaque personne est filmée séparément afin d'éviter les occultations partielles qui peuvent apparaître dans le cas où plusieurs piétons sont présents dans le champ de vision de la même caméra. Il faut préciser aussi que les séquences sont préalablement synchronisées en fixant l'image initiale des piétons sur une attitude double support. Les personnes se déplacent en marche normale avec une vitesse constante. Chaque séquence représente un cycle de démarche composé de 28 images.

Les séquences sont organisées temporellement comme illustré dans la figure 5.4. En lignes, $P1$ à 3 représentent les 3 personnes. En colonnes, le couple $V - E$ représente respectivement le point de vue et les conditions d'éclairage pour chaque séquence : $V = \{f : \text{frontal}, ob : \text{oblique}, fp : \text{fronto-parallèle}\}$ et $E = \{b : \text{lumière blanche}, j : \text{lumière jaune}\}$.

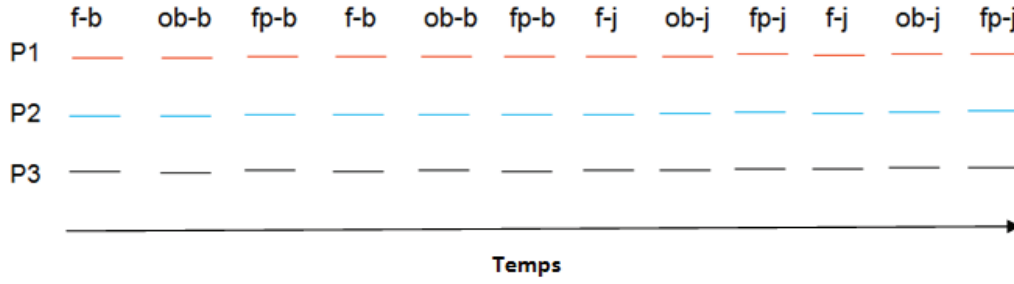


FIGURE 5.4 – Organisation temporelle des séquences de la base.

5.4 Résultats

Dans cette section, nous allons, dans un premier temps, détailler la manière dont nous avons déterminé le seuil de similarité pour détecter les passages ambigus. Par la suite, nous évaluerons les performances du système complet d'identification de personnes proposé.

5.4.1 Détermination du seuil de similarité

Ce seuil de similarité doit être ajusté correctement afin d'assurer une bonne détection des cas ambigus qui doivent être analysés par la démarche. En effet, si ce seuil est trop réduit, plusieurs passages correctement gérés par le modèle d'apparence vont être revus par le modèle de démarche qui risque alors de dominer la phase d'identification. Dans le cas contraire, c'est le modèle d'apparence qui va dominer la phase l'identification finale.

Nous avons réalisé une étude expérimentale sur notre base d'images en augmentant progressivement le seuil de 1% à partir de 90% (100% correspond à la similarité maximale). Les performances liées à chaque valeur de seuil sont évaluées en termes de rappel, précision et F-mesure en se référant à la réalité de terrain (formules 5.1, 5.2 et 5.3).

$$Rappel = \frac{\text{nombre de fausses réidentifications correctement classées}}{\text{nombre de fausses réidentifications}} \quad (5.1)$$

$$Précision = \frac{\text{nombre de fausses réidentifications correctement classées}}{\text{nb de réidentification classées fausses}} \quad (5.2)$$

$$F - mesure = \frac{2 \times Rappel \times Précision}{Rappel + Précision} \quad (5.3)$$

Les résultats obtenus pour les différents seuils ont donné la courbe F-mesure présentée dans la figure 5.5.

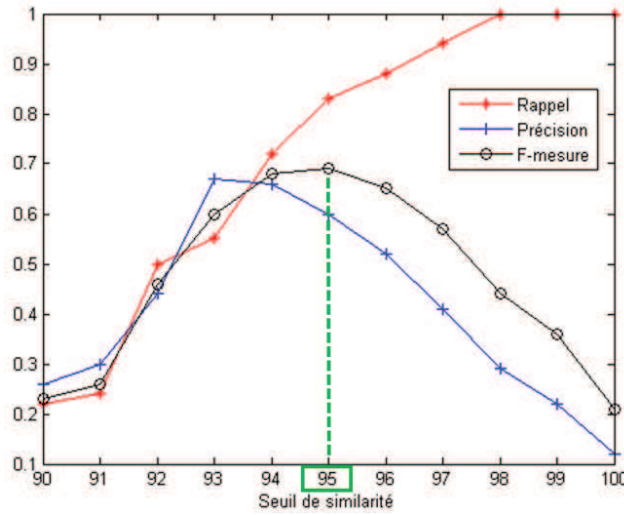


FIGURE 5.5 – F-mesure pour les différents seuils de ré-identification par l'apparence.

D'après la figure 5.5, nous pouvons conclure que le seuil de 95% est le plus approprié pour détecter les mauvaises ré-identifications puisqu'il maximise la F-mesure. Par conséquent il sera appliqué dans le processus final d'évaluation. Un passage sera alors reconnu ambigu s'il contient au moins une mauvaise ré-identification dont la similarité est inférieure au seuil de 95%. Dans ce cas, ce passage sera réanalysé par la démarche.

5.4.2 Taux de reconnaissance et temps de calcul

Afin d'évaluer l'apport du module secondaire d'identification par la démarche dans le système complet, nous avons réalisé deux tests. Le premier est basé uniquement sur le modèle d'apparence, tandis que le second couple l'apparence et la démarche. Les performances de chaque test sont évaluées en termes de taux de reconnaissance des personnes TR (équation 5.4) et temps de traitement total (en images/seconde). Ce temps comprend toute la chaîne incluant la phase de prétraitement, l'extraction des descripteurs et l'appariement des signatures. Les résultats sont présentés dans le tableau 5.2.

$$TR = \frac{\text{Nombre d'images correctement identifiées}}{\text{Nombre total d'images dans la base}} \quad (5.4)$$

	Taux de reconnaissance	Temps
Tests		
Apparence	66	15 images/s
Apparence + démarche	78	7 images/s

Tableau 5.2 – Taux de reconnaissance et temps d'exécution.

D'après le tableau 5.2, l'association de l'apparence et la démarche permet d'améliorer nettement les taux de reconnaissance par rapport à l'utilisation d'un modèle d'apparence seul. En effet, les 12% supplémentaires représentent $(12 \times 1008) / 100 = 121$ images de personnes, soit à peu près 4 séquences correctement classées suite à l'application du modèle de la démarche. Ce qui prouve que combiner l'apparence et la démarche représente une stratégie pertinente pour identifier les personnes dans un contexte de vidéosurveillance multi-caméras.

L'intégration d'une analyse de la démarche sur les passages ambigus engendre un temps d'exécution plus important. L'approche proposée permet tout de même d'offrir un bon compromis entre efficacité et temps de calcul, avec un taux de reconnaissance de 78% et vitesse d'exécution de 7 images/seconde. Ces performances rendent tout à fait réaliste l'application de cette stratégie dans un contexte réel d'identification des personnes dans un réseau de caméras.

5.5 Conclusion

Dans ce chapitre, nous avons introduit une approche combinée qui consiste à associer l'apparence et la démarche pour identifier les personnes en multi-caméras. Cette stratégie a été évaluée dans un contexte réel de vidéosurveillance intégrant des variations de point de vue et de conditions d'éclairage qui sont parmi les plus difficiles dans ce domaine. Afin de consolider l'identification basée sur l'apparence, un test d'ambiguïté est mis en place pour déclencher ou non le module secondaire d'analyse de la démarche. L'apport d'une analyse de la démarche est clairement mise en valeur dans les tests réalisés. Les résultats prouvent que l'intégration de la démarche à l'apparence permet d'améliorer les performances globales

du système d'identification des personnes en multi-caméras en offrant un bon compromis efficacité/temps de calcul.

Conclusion générale et perspectives

Placée dans un contexte de vidéo surveillance, cette thèse s'est principalement centrée sur la reconnaissance de piétons.

L'objectif à terme est d'identifier et localiser des personnes se déplaçant librement dans un espace urbain ou un bâtiment, en exploitant un réseau de caméras dont les champs de vues ne se chevauchent pas. Un tel réseau de caméras est le seul permettant de contrôler de larges espaces au moindre coût mais il complexifie la tâche de reconnaissance qui doit s'affranchir des discontinuités dans l'observation des personnes et des larges variations de point de vue et d'éclairage. Nous avons donc axé nos travaux sur la recherche de descripteurs robustes dans ce contexte de ré-identification. S'appuyant sur un état de l'art déjà dense, nos travaux se démarquent de l'existant en proposant une approche associant l'apparence et la démarche d'une personne pour renforcer son identification selon plusieurs points de vue.

Principales contributions

S'intéressant d'abord à l'apparence, nous avons proposé une mesure de similarité exploitant l'information intra et inter-personnes par l'analyse d'histogrammes régionaux afin de s'adapter à la fois aux fortes variations d'éclairage et de pose. Comparé aux principaux descripteurs d'apparence de la littérature (couleur, texture et forme) dans des situations de test variées, les histogrammes régionaux avec la mesure de similarité proposée se montrent parmi les plus performants pour caractériser une personne par l'apparence. Suite à cette analyse individuelle des descripteurs d'apparence, une fusion des meilleurs d'entre eux par Adaboost permet d'améliorer les performances d'environ 3% par rapport à la méthode ELF de D. Gray qui fait référence dans ce domaine.

Dans un second temps, l'étude de la démarche comme indicateur biométrique, nous a

amené à proposer deux stratégies différentes selon le point de vue d'observation de la caméra. La vue de profil (observations selon un angle variable entre fronto-parallèle et oblique) permet d'exploiter les points d'articulation du squelette de la personne et de caractériser ainsi le mouvement en ciseaux des jambes, les enjambées et les appuis bipodal ou unipodal. Après avoir mis en place une extraction automatique des points d'articulation (hanche, genoux, chevilles) à partir de l'analyse de la silhouette binaire, un nouveau descripteur, modélisant le mouvement comme un pendule oscillant a été proposé. Une étude comparative réalisée sur plusieurs bases de test intégrant des changements de point de vue et de vitesses de déplacement démontre un bon comportement du descripteur proposé vis-à-vis de l'état de l'art. La fusion de différents descripteurs permet là aussi une nette amélioration du taux de reconnaissance (environ 20 à 30%).

En vue de face, les enjambées ne sont plus visibles et notre stratégie se base alors sur une modélisation de la distribution du mouvement sur la silhouette entière, en proposant deux nouvelles représentations : l'une binaire sur les variations de la forme (RED) et l'autre en niveaux de gris sur la fréquence des mouvements (RFD). Une étude expérimentale a montré que la fusion des descripteurs associés à ces deux représentations permet de rivaliser avec les meilleures approches de la littérature.

Enfin, cette thèse ouvre la voie d'une exploitation combinée de l'apparence et de la démarche pour renforcer l'identification dans un réseau de caméras variant fortement les points de vue. Cette approche mixte a été expérimentée sur une base de séquences vidéo élaborée dans le but de contenir les différents aspects recherchés. L'approche proposée offre un taux de classification de 78% avec une vitesse d'exécution de 7 images/secondes, ce qui favorise son application en vidéosurveillance.

Limitations et perspectives

Le système d'identification proposé présente un certain nombre de limitations.

Détection des personnes : L'étape préalable à la reconnaissance qui consiste à détecter dans chaque image d'une séquence vidéo les silhouettes ou boîtes englobantes correspondant aux personnes présentes joue un rôle prépondérant dans la chaîne de traitements. Sa qualité influe considérablement sur les performances d'identification. Dans ces travaux de thèse, nous

avons adopté une approche assez classique basée sur deux hypothèses : les personnes sont les seuls objets mobiles dans la scène et elles sont toujours en position debout, ce qui permet de les détecter par élimination de l'arrière plan et application d'une règle morphologique simple.

Qualité des images : Les images à basse résolution et les occultations importantes perturbent considérablement la caractérisation de l'apparence et de la démarche. Les déformations à l'intérieur des silhouettes affectent notamment les descripteurs de la démarche au niveau de la détection des points d'intérêt ou lors du calcul des représentations RFD et RED.

Plusieurs extensions possibles à ces travaux peuvent être envisagées à plus ou moins long terme.

Extension du processus de localisation des points d'intérêt pour caractériser la démarche : Etant donné l'importance de cette étape dans le processus de caractérisation, nous envisageons d'associer notre approche à d'autres techniques de détection. Une première approche consisterait à associer des méthodes classiques de mise en correspondance, telles que SIFT et SURF, entre des images successives. Une seconde approche associerait des modèles prédictifs tels que le filtre de Kalman ou à particules qui permettront d'estimer les positions futures. Ces associations garantiraient que la détection automatique ne s'écarte pas trop de la position réelle en cas de déformations ou de changement brutal du mode de déplacement.

Amélioration de la phase caractérisation : Développer un processus d'appariement basé sur la mise en correspondance entre profils personnalisés. Chaque personne sera alors caractérisée par un sous ensemble de descripteurs qui la différencie plus des autres. Ceci offre l'avantage de consolider la phase de caractérisation et permet aussi d'accélérer les traitements en éliminant les profils non conformes à une personne requête.

Amélioration de la phase de classification : Tester l'apport de quelques classifieurs sophistiqués (comme SVM, ANN, ...) sur les performances du système.

Intégration du suivi : Intégrer les méthodes de suivi afin d'associer chaque personne à sa position réelle dans un plan 2D du bâtiment. Ceci permettra de définir les trajectoires des personnes à partir desquelles nous pouvons soit résoudre le problème de croisement et la séparation des personnes, soit extraire des paramètres de caractérisation tel que la vitesse et l'accélération, soit définir l'orientation de la personne. En utilisant des connaissances *a*

priori sur la topologie des caméras dans le bâtiment, l'orientation de la personne permet d'estimer les prochains champs de vue dans lesquelles elle sera présente.

Détection des personnes : Afin d'améliorer le processus de détection des personnes, l'approche introduite par Dollar et al. [128] peut être intégrée. Elle est basée sur une décomposition pyramidale robuste et rapide permettant de générer à partir d'une image initiale des images à différentes résolutions conservant l'orientation du gradient. Ceci leur a permis d'utiliser des HOG à plusieurs échelles afin de consolider l'étape de détection. En cas d'occultations importantes, nous envisageons une localisation par partie en utilisant le template matching. Une réduction de l'espace de recherche par une étape d'élimination de l'arrière plan ainsi que l'utilisation de quelques caractéristiques morphologiques sont alors nécessaires pour accélérer les traitements dans ce cas.

Bibliographie personnelle

Conférences internationales

- A. Derbel, A. Chetouani, S. Treuillet, B. Emile, N. Mansouri et Y. Ben Jemaa, "*Interest Lower Body Point's Detection For Markerless Gait Analysis*". publié dans le proceeding de International Conference on Image Processing Theory, Tools and Applications (IPTA), Paris, France, 2014.
- A. Derbel, Y. Ben Jemaa, S. Treuillet, B. Emile, R. Canals et A. Ben Hamadou, "*Robust Descriptors Fusion for Pedestrians' Re-identification and Tracking Across a Camera Network*". publié dans le proceeding de International Conference on Computer Vision Theory and Applications (VISAPP), Barcelone, Espagne, 2013.
- A. Derbel, N. Mansouri, Y. Ben Jemaa, , B. Emile et S. Treuillet, "*Comparative Study between Spatio/Temporal Descriptors for Pedestrians Recognition by Gait*". publié dans le proceeding de International Conference on Image Analysis and Recognition (ICIAR), Póvoa de Varzim, Portugal, 2013.
- A. Derbel, Y. Ben Jemaa, R. Canals , B. Emile, S. Treuillet et A. Ben Hamadou, "*Comparative study between Color Texture and Shape descriptors for multi-camera pedestrians' identification*". publié dans le proceeding de International Conference on Image Processing Theory, Tools and Applications (IPTA), Istanbul, Turquie, 2012.

Conférences nationales

- A. Derbel, S. Treuillet, B. Emile, Y. Ben Jemaa et A. Ben Hamadou, *"Proposition et test de descripteurs pour l'identification de personnes dans un réseau de caméras"*. publié dans le proceeding du Congrès National GRETSI, Bordeaux, France, 2011.
- A. Derbel, S. Treuillet, Y. Ben Jemaa, B. Emile et A. Ben Hamadou, *"Identification de personnes en vue frontale par analyse de mouvement"*. publié dans le proceeding du Congrès National sur la Reconnaissance de Formes et l'Intelligence Artificielle (RFIA), Rouen, France, 2014.

Bibliographie

- [1] K. Konolige, M. Agrawal, M. R. Blas, R.C. Bolles, B. Gerkey, J. Solà et A. Sundaresan, "Mapping, navigation, and learning for off-road traversal". Journal of Field Robotics, Vol. 26, No. 1, pp. 88–113, 2009.
- [2] L. Ran, S. Helal et S. Moore, "*Drishti : An Integrated Indoor/Outdoor or Blind Navigation System and Service*". International Conference on Pervasive Computing and Communications, pp. 23–30, 2004.
- [3] S. Ertan, C. Lee, A. Willets, H. Tan et A. Pentland, "*A wearable haptic navigation guidance system*". International Symposium on Wearable Computers, pp. 164–165, 1998.
- [4] V. Kulyukin, C. Gharpure, J. Nicholson et S. Pavithran, "*RFID in robot-assisted indoor navigation for the visually impaired*". International Conference on Intelligent Robots and Systems, pp.1979–1984, 2004.
- [5] V. Renaudin, O. Yalak, P. Tomé et B. Merminod, "Indoor Navigation of Emergency Agents". European Journal of Navigation, Vol. 5, No. 3, pp. 36–45, 2007.
- [6] Slawomir Bak, "Human Re-identification Through a Video Camera Network", Thèse de doctorat en informatique, sous la direction de François Brémond, Université de Nice-Sophia Antipolis, pp. 4–5, 2012.
- [7] S. Ioffe et D. Forsyth, "*Human tracking with mixtures of trees*". International Conference on Computer Vision, pp. 690–695, 2001.
- [8] J.C. Cheng et J.M.F. Moura, "*Tracking human walking in dynamic scenes*". International Conference on Image Processing, pp. 137–140, 1997.
- [9] I. Haritaoglu, D. Harwood et L.S. Davis, "*W4 : A Real Time System for Detecting and Tracking People*". International Conference on Computer Vision and Pattern Recognition, pp. 962–962, 1998.
- [10] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems". Journal of Basic Engineering, Vol. 82, Series D, pp. 35–45, 1960.

- [11] W. K. Hastings, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". *Biometrika*, Vol. 57, No. 1, pp. 97–109, 1970.
- [12] P. K. TraKulPong et R. Bowden, "*An Adaptive Visual System for Tracking Low Resolution Colour Targets*". *British Machine Vision Conference*, pp. 243–252, 2001.
- [13] J. Sturges et T. W. A. Whitfield, "Locating basic colours in the Munsell space". *Journal of Color Research and Application*, Vol. 20, No. 6, pp. 364–376, 1995.
- [14] E. Monari , J. Maerker et K. Kroschel, "*A Robust and Efficient Approach for Human Tracking in Multi-camera Systems*". *International Conference on Advanced Video and Signal-Based Surveillance*, pp. 134–139, 2009.
- [15] T. H. Chang et S. Gong, "*Bayesian modality fusion for tracking multiple people with a multi-camera system*". *European Workshop on Advanced Video-based Surveillance Systems*, 2001.
- [16] C. Yang, R. Duraiswami et L. Davis, "*Fast multiple object tracking via a hierarchical particle filter*". *International Conference on Computer Vision*, pp. 212–219, 2005.
- [17] Y. Freund et R. E. Schapire, "*A decision-theoretic generalization of on-line learning and an application to boosting*". *European Conference on Computational Learning Theory*, pp. 23–37, 1995.
- [18] T. Tung et T. Matsuyama, "*Human Motion Tracking using a Color-Based Particle Filter Driven by Optical Flow*". *International Workshop on Machine Learning for Vision-based Motion Analysis*, 2008.
- [19] B. Lucas et T. Kanade, "*An iterative image registration technique with an application to stereo vision*". *International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.
- [20] J. Wang, Y.G. Yin et H. Man, "Multiple Human Tracking Using Particle Filter with Gaussian Process Dynamical Model," *Journal on Image and Video Processing*, Vol. 2008, 10 pages, 2008.
- [21] A. Mittal et L.S. Davis, "M2 tracker : a multi-view approach to segmenting and tracking people in a clustered scene". *Journal of Computer Vision*, Vol. 53, No. 3, pp. 189–203, 2002.
- [22] S.L. Dockstader et A.M. Tekalp, "Multiple camera tracking of interacting and occluded human motion". *Proceedings of the IEEE*, Vol. 89, No. 10, pp. 1441–1455, 2002.
- [23] V. Kettner et R. Zabih, "*Bayesian multi-camera surveillance*". *International Conference on Computer Vision and Pattern Recognition*, pp. 2253–2259, 1999.

- [24] O. Javed, Z. Rasheed, K. Shafique et M. Shah, "*Tracking Across Multiple Cameras with Disjoint Views*". International Conference on Computer Vision, pp. 952–957, 2003.
- [25] A. Gilbert et R. Bowden, "*Tracking Objects Across Cameras by Incrementally Learning Inter-camera Colour Calibration and Patterns of Activity*". European Conference on Computer Vision, pp.125–136, 2006.
- [26] U. Park, A.K. Jain, I. Kitahara, K. Kogure et N. Hagita, "*ViSE : Visual Search Engine Using Multiple Networked Cameras*". International Conference on Pattern Recognition, pp.1204–1207, 2006.
- [27] A. C. Gallagher et T. Chen, "*Clothing cosegmentation for recognizing people*". International Conference on Computer Vision and Pattern Recognition, pp. 1–8, 2008.
- [28] P. Viola et M. Jones, "*Fast Multiview Face Detection*". International Conference on Computer Vision and Pattern Recognition, 2003.
- [29] P. N. Belhumeur, J. P. Hespanha et D. J. Kriegman, "Eigenfaces vs. Fisherfaces : recognition using class specific linear projection". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, pp. 711–720, 1997.
- [30] Y. Cai, K. Huang et T. Tan, "*Human appearance matching across multiple non-overlapping cameras*". International Conference on Pattern Recognition, pp.1–4, 2008.
- [31] J. Canny, "A computational approach to edge detection". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, No. 6, pages 679–698, 1986.
- [32] Z. Lin et L.S. Davis, "*Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance*". International Symposium on Advances in Visual Computing, pages 23–34, 2008.
- [33] E. Parzen, "On the estimation of a probability density function and mode". Annals of Mathematical Statistics, Vol. 33, pp.1065–1076, 1962.
- [34] S. Kullback et R.A. Leibler, "On information and sufficiency". Annals of Mathematical Statistics, Vol. 22, No. 1, pp. 79–86, 1951.
- [35] W.R. Schwartz et L.S. Davis, "*Learning discriminative appearancebased models using partial least squares*". Brazilian Symposium on Computer Graphics and Image Processing, pp. 322–329, 2009.
- [36] A. Hoskuldsson, "Partial least squares regression methods". Journal of Chemometrics, Vol. 2, No. 3, pp. 211–228, 1988.

- [37] S. Bak, E. Corvee, F. Bremond et M. Thonnat, "*Person reidentification using Haar-based and DCD-based signature*". International Conference on Advanced Video and Signal Based Surveillance, pp.1–8, 2010.
- [38] D. Gray et H. Tao, "*Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features*". European Conference on Computer Vision, pp. 262–275, 2008.
- [39] I. Fogel et D. Sagi, "Gabor filters as texture discriminator". Journal of Biological Cybernetics, Vol. 61, No. 2, pp. 103–113, 1989.
- [40] C. Schmid, "*Constructing models for content-based image retrieval*". International Conference on Computer Vision and Pattern Recognition, pp. 39–45, 2001.
- [41] B. Prosser, W. S. Zheng, S. Gong et T. Xiang, "*Person Re-Identification by Support Vector Ranking*". British Machine Vision Conference, pages 1–11, 2010.
- [42] T. Joachims, "*Optimizing search engines using clickthrough data*". International conference on Knowledge discovery and data mining, pp. 133–142, 2002.
- [43] N. D. Bird, O. Masoud, N. P. Papanikolopoulos et A. Isaacs, "Detection of loitering individuals in public transportation areas". IEEE Transactions on Intelligent Transportation Systems, Vol. 6, No. 2, pp. 167–177, 2005.
- [44] N. Gheissari, T. B. Sebastian et R. Hartley. "*Person Reidentification Using Spatiotemporal Appearance*". International Conference on Computer Vision and Pattern Recognition, pp. 1528–1535, 2006.
- [45] C. Madden, E.D. Cheng et M. Piccardi, "Tracking people across disjoint camera views by an illumination-tolerant appearance representation". Journal of Machine Vision and Applications, Vol. 18, No. 3, pp.233–247, 2007.
- [46] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman et Y. We, "An Efficient k-Means Clustering Algorithm : Analysis and Implementation". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, pp. 881–892, 2002.
- [47] Y. Yu, D. Harwood, K. Yoon et L.S. Davis, "Human appearance modeling for matching across video sequences". Journal of Machine Vision and Applications, Vol. 18, No. 3, pp.139–149, 2007.
- [48] O. Hamdoun, F. Moutarde, B. Stanciulescu et B. Steux, "*Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences*". International Conference on Distributed Smart Cameras, pp. 1–6, 2008.

- [49] H. Bay, A. Ess, T. Tuytelarrs et L. Van Gool, "Speeded-Up Robust Features (SURF)". Computer Vision and Image Understanding , Vol. 110, No. 3, pp. 346–359, 2008.
- [50] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints". International Journal of Computer Vision, Vol. 60, No.2, pp. 91–110, 2004.
- [51] D. N. T. Cong, C. Achard, L. Khoudour et L. Douadi, "*Video Sequences Association for People Reidentification across Multiple Non-overlapping Cameras*". International Conference on Image Analysis and Processing, pp.179–189, 2009.
- [52] U. V.Luxburg, "A tutorial on spectral clustering". Journal of Statistics and Computing, Vol. 17, No. 4, pp. 395–416, 2007.
- [53] D. N. Truong Cong, L. Khoudour, C. Achard, C. Meurie et O. Lezoray, "People re-identification by spectral classification of silhouettes". Signal Processing, Vol. 90, No. 8, pp. 2362–2374, 2010.
- [54] G. Buchsbaum, "A spatial processor model for object color perception". Journal of the Franklin Institute, Vol. 310, No. 1, pp. 1–26, 1980.
- [55] M. Farenzena, L. Bazzani, A. Perina, V. Murino et M. Cristani, "*Person re-identification by symmetry-driven accumulation of local features*". International Conference on Computer Vision and Pattern Recognition, pp. 2360–2367, 2010.
- [56] L. Bazzani, M. Cristani, A. Perina, M. Farenzena et V. Murino, "*Multiple-Shot Person Re-identification by HPE Signature*". International Conference on Pattern Recognition, pp. 1413–1416, 2010.
- [57] N. Jojic, B. J. Frey et A.Kannan, "*Epitomic analysis of appearance and shape*". International Conference on Computer Vision, pp. 34–41, 2003.
- [58] M. Hirzer, C. Beleznaï, P. M. Roth et H.Bischof, "*Person Re-identification by Descriptive and Discriminative Classification*". Scandinavian Conference on Image Analysis, pp. 91–102, 2011.
- [59] C. ÓConaire, N. E. OConnor et A.F. Smeaton, "*An improved spatiogram similarity measure for robust object localisation*". International Conference on Acoustics, Speech, and Signal Processing, pp. 1069–1072. 2007.
- [60] F. Zernike, "Diffraction theory of the cut procedure and its improved form, the phase contrast method". Physica, Vol. 1, pp. 689–704, 1934.
- [61] M. K. Hu, "Visual pattern recognition by moment invariants". IEEE Transactions on Information Theory, Vol. 8, pp. 179–187, 1962.

- [62] R.A. Haddad et A.N. Akansu, "A Class of Fast Gaussian Binomial Filters for Speech and Image Processing". IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 39, No. 3, pp 723–727, 1991.
- [63] M. Al-Rawi et Y. Jie, "Practical fast computation of Zernike moments". Journal of Computer Science and Technology, Vol. 17, No. 2, pp.181–188, 2002.
- [64] G. Amayeh, S. Kasaei, G. Bebis, A. Tavakkoli et K. Veropoulos, "*Improvement of Zernike moment descriptors on affine transformed shapes*". International Symposium on Signal Processing and Its Applications, pp.1–4, 2007.
- [65] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by probability distributions". Bulletin of the Calcutta Mathematical Society, Vol. 35, pp. 99–109, 1943.
- [66] D. Gray, S. Brennan et H. Tao, "*Evaluating appearance models for recognition, reacquisition, and tracking*". International Workshop on Performance Evaluation of Tracking and Surveillance, 2007.
- [67] S. Obdržálek, J. Matas et O. Chum, "*On the Interaction between Object Recognition and Colour Constancy*". International Workshop on Color and Photometric Methods in Computer Vision, 2003.
- [68] G. D. Finlayson, S. Hordley, G. Schaefer et G. Y. Tian. "Illuminant and device invariant colour using histogram equalisation". Pattern Recognition, Vol. 38, No. 2, pp. 179–190, 2005.
- [69] N. Vandenbroucke et L. Macaire, "Représentation de la couleur en analyse d'images". Techniques de l'Ingénieur, numéro S7602, pages 1–22, 2005.
- [70] G. D. Finlayson, B. Schiele et J. Crowley. "*Comprehensive color image normalization*". European Conference on Computer Vision, pp. 475–490, 1998.
- [71] P. K. Sharma, C. Huang et R. Nevatia, "*Evaluation of people tracking, counting and density estimation in crowded environments*". International Workshop on Performance Evaluation of Tracking and Surveillance, pages 39–46, 2009.
- [72] F. Fleuret, J. Berclaz, R. Lengagne et P. Fua, "Multi-camera people tracking with a probabilistic occupancy map". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, No. 2, pp. 267–282, 2008.
- [73] P. Kelly, C. ÓConaire, D. Monaghan, J. Kuklyte, D. Connaghan, A. P. Moneo, D. Juan et P. Daras, "*Performance analysis and visualisation in tennis using a low-cost camera network*". Multimedia Grand Challenge Track, pp. 25–29, 2010.

- [74] Y. Freund, R. Iyer, R. E. Schapire et Y. Singer, "An efficient boosting algorithm for combining preferences". *Journal of Machine Learning Research*, Vol. 4, pp. 933–969, 2003.
- [75] D. R. Kisku, P. Gupta et J. K. Sing, "Offline Signature Identification by Fusion of Multiple Classifiers using Statistical Learning Theory", *International Journal of Security and Its Applications*, Vol. 4, No. 3, 2010.
- [76] K. Popplewel, A. Alford, G. Dozier, K. Brayant, J. Kelly, J. Adams, T. Abegaz, K. Purrington et J. Shelton, "*A comparison of genetic feature selection and weighting techniques for multio-biometric recognition*". *ACM southeast Regional Conference*, pp.205–208, 2011.
- [77] G. Johansson, "Visual perception of biological motion and a model for its analysis". *Perception and Psychophysics*, Vol. 14, pp.201–211, 1973.
- [78] S. Sarkar, P.J. Phillips, Z. Liu, I. Robledo, P. Grother et K. Bowyer, "The humanoid gait challenge problem : data sets, performance, and analysis". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 2, pp. 162–177, 2005.
- [79] A. Bobick et A. Johnson, "*Gait recognition using static activity-specific parameters*". *International Conference on Computer Vision and Pattern Recognition*, pp. 423–430, 2001.
- [80] J.H. Yoo, D. Hwang, K.Y. Moon et M. S. Nixon, "*Automated human recognition by gait using neural network*". *International Workshops on Image Processing Theory, Tools and Applications*, pp. 1–6, 2008.
- [81] L.Lee et W.E.L.Grimson, "*Gait analysis for recognition and classification*". *International Conference on Automatic Face and Gesture Recognition*, pp. 148–155, 2002.
- [82] S. Sivapalan, D. Chen, S. Denman, S. Sridharan et C. Fookes, "*3D ellipsoid fitting for multi-view gait recognition*". *International Conference on Advanced Video and Signal Based Surveillance*, pp. 355–360, 2011.
- [83] Y. ChewYean, M.S. Nixon et J.N. Carter, "*On the Relationship of Human Walking and Running :Automatic Person Identification by Gait*". *International Conference on Pattern Recognition*, pp 287–290, 2002.
- [84] C.Y. Yam, M.S. Nixon et J.N. Carter, "*Gait Recognition By Walking and Running : A Model-Based Approach*". *Asian Conference on Computer Vision*, pp 23–25, 2002.
- [85] J. H. Yoo et M. S.Nixon, "Automated Markerless Analysis of Human Gait Motion for Recognition and Classification". *ETRI Journal*, Vol. 33, No. 2, pp. 259–266, 2011.

- [86] W. Dempster et G. Gaughran, "Properties of body segments based on size and weight". American Journal Anatomy, Vol. 120, pp. 33–54, 1967.
- [87] R. Tanawongsuwan et A. Bobick, "*Gait recognition from time-normalized joint-angle trajectories in the walking plane*". International Conference on Computer Vision and Pattern Recognition, pp 726–731, 2001.
- [88] E. Guillen, D. Padilla, A. Hernandez et K. Barner, "Gait recognition system : Bundle rectangle approach". World Academy of Science Engineering and Technology , Vol. 37, pp. 696–702, 2009.
- [89] M. Ekinici, "*A new attempt to silhouette-based gait recognition for human identification*". Conference of the Canadian Society for Computational Studies of Intelligence, pp. 443–454, 2006.
- [90] L. Wang, H. Ning, T. Tan et W. Hu, "Fusion of static and dynamic body biometrics for gait recognition". IEEE Transactions on Circuits and Systems for Video Technology, Vol. 14, No. 2, pp. 149–158, 2004.
- [91] A. Tsuji, Y. Makihara et Y. Yagi, "*Silhouette transformation based on walking speed for gait identification*". International Conference on Computer Vision and Pattern Recognition, pp. 717–722, 2010.
- [92] A. Kale, A. N. Rajagopalan, N. Cuntoor et V. Kruger, "*Gait-based recognition of humans using continuous hmms*". International Conference on Automatic Face and Gesture Recognition, pp. 321–326, 2002.
- [93] A. Kale, A. Sundaresan, A. N. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger et R. Chellappa, "Identification of humans using gait". IEEE Transactions on Image Processing, Vol. 13, No. 9, pp. 1163–1173, 2004.
- [94] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". IEEE Transactions on Information Theory, Vol. 13, No. 2, pp. 260–269, 1967.
- [95] Q. He et C. Debrunne, "*Individual recognition from periodic activity using hidden markov models*". Workshop on Human Motion, pp. 47–52, 2000.
- [96] L. Wang, W. Hu et T. Tan, "*A new attempt to gait-based human identification*". International Conference on Pattern Recognition, pp. 115–118, 2002.
- [97] L. Wang, T. Tan, W. Hu et H. Ning, "Automatic gait recognition based on statistical shape analysis". IEEE Transactions on Image Processing, Vol. 12, No. 9, pp. 1120–1131, 2003.

- [98] F. Dadashi, B. N. Araabi et H. S. Zadeh, "*Gait recognition using wavelet packet silhouette representation and transductive support vector machines*". International Conference on Image and Signal Processing, pp. 1–5, 2009.
- [99] N. V. Boulgouris, K. N. Plataniotis et D. Hatzinakos, "*An angular transform of gait sequences for gait assisted recognition*". International Conference on Image Processing, pp. 857–860, 2004.
- [100] C. BenAbdelkader, R. Cutler et L. Davis, "*Stride and cadence as a biometric in automatic person identification and verification*". International Conference on Automatic Face and Gesture Recognition, pp. 372–377, 2002.
- [101] C. BenAbdelkader, R. Cutler et L. Davis, "Gait recognition using image self similarity". Journal on Advances in Signal Processing, Vol. 20, No. 4, pp. 572–585, 2004.
- [102] Y. Liu, R. Collins et Y. Tsin, "*Gait sequence analysis using frieze patterns*". European Conference on Computer Vision, pp. 657–671, 2002.
- [103] S. Lee, Y. Liu et R. Collins, "*Shape variation- based frieze pattern for robust gait recognition*". International Conference on Computer Vision and Pattern Recognition, pp. 1–8, 2007.
- [104] J. Han et B. Bhanu, "Individual recognition using gait energy image". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, No. 2, pp. 316–322, 2006.
- [105] K. Bashir, X. Tao et G. Shaogang, "*Feature selection on gait energy image for human identification*". International Conference on Acoustics, Speech and Signal Processing, pp. 985–988, 2008.
- [106] C. Lin et K. Wang, "*A behavior classification based on enhanced gait energy image*". International Conference on Networking and Digital Society, pp. 589–592, 2010.
- [107] X. Huang et N. V. Boulgouris, "Human gait recognition based on multiview gait sequences". Journal on Advances in Signal Processing, Vol. 2008, No. 1, pp. 1–9, 2008.
- [108] C. C. Yu, H. Y. Cheng, C. H. Cheng et K. C. Fan, "Efficient human action and gait analysis using multiresolution motion energy histogram". Journal on Advances in Signal Processing, Vol. 2010, pp. 1–13, 2010.
- [109] C. Chen, J. Liang, H. Zhao, H. Hu et J. Tian, "Frame difference energy image for gait recognition with incomplete silhouettes". Pattern Recognition Letters, Vol. 30, No. 11, pp. 977–984, 2009.

- [110] E. Zhang, Y. Zhao et W. Xiong, "Active energy image plus 2dlpp for gait recognition". Signal Processing, Vol. 90, No. 7, pp. 2295–2302, 2010.
- [111] H. Ng, H. L. Tong, W. H. Tan , T. V. Yap, P. F. Chong et J. Abdullah, "Human Identification Based on Extracted Gait Features". International Journal on New Computer Architectures and Their Applications, Vol. 1, No. 2, pp. 358–370, 2011.
- [112] J. Wang, M. She, S. Nahavandi et A. Kouzani, "A review of vision-based gait recognition methods for human identification". Digital Image Computing : Techniques and Application, pp. 320–327, 2010.
- [113] G. Bingham, R. Schmidt et L. Rosenblum, "Dynamics and the orientation of kinematic forms in visual event recognition". Experimental Psychology : Human Perception and Performance, Vol. 21, No. 6, pp. 1473–1493, 1995.
- [114] M. Silaghi, R. Plankers, R. Boulic, P. Fua et D. Thalmann, "*Local and global skeleton fitting techniques for optical motion capture*". Workshop on Modelling and Motion Capture Techniques for Virtual Environments, pp. 26–40, 1998.
- [115] A. Fernandez-Baena, A. Susin et X. Lligadas, "*Biomechanical validation of upper-body and lowerbody joint movements of kinect motion capture data for rehabilitation treatments*". International Conference on Intelligent Networking and Collaborative Systems, pp. 656–661, 2012.
- [116] T. Weijun, L. Tao, Z. Rencheng et F. Hutian, "Gait analysis using wearable sensors". Sensors, Vol. 12, pp. 2255–2283, 2012.
- [117] S. Mondal, A. Nandy, P. Chakraborty et G. Nandi, "Gait based personal identification system using rotation sensor". Journal of Emerging Trends in Computing and Information Sciences, Vol. 3, No. 3, pp. 395–402, 2012.
- [118] S. Corazza, L. Muendermann, T. Chaudhari, T. Demattio, C. Cobelli et T. Andriacchi, "A markerless motion capture system to study musculoskeletal biomechanics : visual hull and simulated annealing approach". Annals of Biomedical Engineering, Vol. 34, No. 6, pp. 1019–1029, 2006.
- [119] K. Cheung, S. Baker et T. Kanade, "Shape from silhouette across time : Theory and algorithms". International Journal of Computer Vision, Vol. 63, No. 3, pp. 225–245, 2005.
- [120] S. Yu, D. Tan et T. Tan, "*A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition*". International Conference on Pattern Recognition, pp. 441–444, 2006.

- [121] C. Y. Yam, M. S. Nixon et J. N. Carter, "Automated person recognition by walking and running via model-based approaches". *Pattern Recognition*, Vol. 37, No. 5, pp. 1057–1072, 2004.
- [122] H. Mu, X. Yu et C. Wang, "*A frequency domain analysis based gait feature representation*". *International Conference on Computational Intelligence and Software Engineering*, pp. 1–6, 2009.
- [123] I. W. Selesnick, "Hilbert transform pairs of wavelet bases". *Signal Processing Letters*, Vol. 8, No. 6, pp. 170–173, 2001.
- [124] L. Wang, T. Tan, H. Ning et W. Hu, "Silhouette analysis-based gait recognition for human identification". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No.12, pp. 1505–1518, 2003.
- [125] S. Chen et Y. Gao, "*An Invariant Appearance Model for Gait Recognition*". *International Conference on Multimedia and Expo*, pp.1375–1378, 2007.
- [126] M. Goffredo, J. Carter et M. Nixon, "*Front-view Gait Recognition*". *International Conference on Biometrics : Theory, Applications and Systems*, pp. 1–6, 2008.
- [127] O. Barnich et M. V. Droogenbroeck, "Frontal-view gait recognition by intra-and inter-frame rectangle size distribution". *Pattern Recognition Letters*, Vol. 30, No. 10, pp. 893–901, 2009.
- [128] P. Dollar, S. Belongie et P. Perona, "*The fastest pedestrian detector in the west*". *British Machine Vision Conference*, pp. 1–11, 2010.

Ahmed DERBEL

Identification de personnes dans un flux vidéo par l'apparence et la démarche

Avec l'installation généralisée de caméras de surveillance dans les zones urbaines, l'enjeu de cette thèse est la reconnaissance automatique de piétons par l'apparence et la démarche. Cette reconnaissance nécessite des descripteurs robustes pour caractériser et identifier une personne au-delà des modifications d'apparence induites par un changement de point de vue, d'éclairage ou des variations de déplacement. La première partie propose une nouvelle mesure de similarité exploitant l'information intra et inter-personnes dans les histogrammes régionaux couleur. Une étude comparative montre l'intérêt de cette modélisation pour s'affranchir des variations de luminosité et de pose. Une fusion pondérée des descripteurs d'apparence les plus performants permet ensuite d'améliorer le taux d'identification. La seconde partie étudie la reconnaissance par la démarche selon un point de vue latéral puis frontal. Pour une observation latérale du piéton, la dynamique de la marche est modélisée par les variations spatio-temporelles des points d'articulation des deux jambes. Tandis qu'en vue frontale, on analyse la distribution des mouvements dans les silhouettes soit par une carte binaire de l'enveloppe (RED), soit par une carte en niveaux de gris des fréquences (RFD). L'étude expérimentale démontre que les descripteurs proposés font preuve de bonnes performances face à l'état de l'art, et qu'une fusion de plusieurs descripteurs permet toujours d'améliorer notablement les taux de reconnaissance. Enfin, la dernière partie de la thèse introduit un système d'identification complet associant l'apparence et la démarche. Cette approche mixte est évaluée sur une base de séquences vidéo intégrant de fortes variations de point de vue et d'éclairage. Elle offre un bon compromis entre efficacité et temps de calcul pour une application en vidéo surveillance.

Mots-clés : vision par ordinateur, identification de personnes, apparence, démarche

People identification in video sequences by appearance and gait

With the wide-spread installation of cameras in urban areas, this thesis deals with an automatic people identification by appearance and gait. This recognition requires robust descriptors to represent and discriminate each person despite the appearance variations caused by changing view point, lighting or way of moving (speed, carrying a bag...). The first part, dedicated to appearance identification, proposes a new similarity measure using intra and inter-person information with regional color histograms. A comparative study shows the efficiency of this representation to overcome the lighting and pose variations. The recognition rate is then improved by merging the most robust appearance descriptors. The second part is focused on gait recognition by distinguishing lateral and frontal points of view. For lateral observation, a pedestrian is characterized by several spatio-temporal variations of some articulated points of the legs. While, the motion of global silhouette is analyzed using a binary envelope map (RED) and a gray levels frequency map (RFD) in frontal point of view. Experimental test prove that those descriptors provide good performances compared to other state-of-art approaches, and that merging descriptors always improves the recognition rate. The last part introduces an identification system coupling appearance and gait. This merging approach is evaluated on a video sequences database including large points of view and lighting variations. It provides a good compromise between efficiency and processing time for application in video surveillance.

Key-words : computer vision, people identification, appearance, gait



Laboratoire PRISME, Université d'Orléans

Laboratoire MIRACL, Université de Sfax

