ECOLE DES MINES DE DOUAI

UNIVERSITE LILLE 1
SCIENCES ET TECHNOLOGIES

THESE

Présentée en vue de l'obtention du grade de

DOCTEUR

Spécialité

Automatique, Génie Informatique, Traitement du Signal et des Images

Par

Lyes HAMOUDI

Doctorat délivré conjointement par
l'Ecole des Mines de Douai et l'Université Lille 1 Sciences et Technologies

---

Application de techniques d'apprentissage pour la détection et la reconnaissance d'individus

---

Soutenue le 7 juin 2011 devant le jury d'examen :

Président	M. François CABESTAING, Professeur Université Lille 1
Rapporteur	Mme Danielle NUZILLARD, Professeur Université Reims Champagne-Ardenne
Rapporteur	M. Jean-José ORTEU, Professeur Ecole des Mines d'Albi
Examinateur	M. Michel VACHER, Ingénieur de Recherche CNRS Université Grenoble 1
Directeur de thèse	M. Stéphane LECOEUCHE, Professeur Ecole des Mines de Douai
Co-directeur	M. Jacques BOONAERT, Maître-assistant Ecole des Mines de Douai

Laboratoire d'accueil : Département Informatique et Automatique de l'Ecole des Mines de Douai
Ecole Doctorale SPI 072 (Lille I, Lille III, Artois, ULCO, UVHC, Centrale Lille)

REMERCIEMENTS

C’est avec un grand plaisir que je réserve cette page, en signe de gratitude et de reconnaissance à tous ceux qui m’ont aidé à la réalisation de ce travail.

Je remercie tout d’abord mes encadrants, Messieurs Stéphane Lecoeuche et Jacques Boonaert, pour leurs confiances et leurs soutiens continuels. Qu’ils trouvent ici l’expression de ma gratitude pour toute l’aide qu’ils m’ont procurée durant ma thèse. Je tiens à les remercier bien vivement pour leurs encadrements, leurs conseils, leurs encouragements et aussi pour leurs précieuses contributions à l’amélioration de la qualité de ce mémoire.

Je remercie vivement les membres de jury d’avoir accepté de juger mon travail de thèse. Je suis très reconnaissant envers Madame Danielle Nuzillard et Monsieur Jean-José Orteu de s’être penchés avec rigueur et grand intérêt sur ce rapport et de m’avoir fait l’honneur d’en être les rapporteurs. Mes sincères remerciements vont également à Monsieur François Cabestaing et Monsieur Michel Vacher qui ont examiné ce travail de thèse et ont participé au jury.

Merci à Monsieur Philippe Hazebroucq, Chef du département IA de l’Ecole de Mines de Douai, pour m’avoir accueilli au sein de son département, et mis à ma disposition tout ce dont je pouvais avoir besoin pour mener à bien ce travail. Merci également à Monsieur Philippe Desodt, pour m’avoir confié, dès le début de ma thèse, des activités d’enseignement de façon régulière.

J’adresse mes sincères remerciements à l’ensemble des membres du département IA pour la bonne ambiance qu'ils ont su créer tout au long de ma présence au sein du département, et pour les qualités scientifiques et humaines de chacun d’eux. Merci tout particulièrement à mes fidèles camarades et amis Messieurs Khaled Boukharouba et Laurent Bako.

Pour terminer, mes remerciements s’adressent à ma famille. A ma mère, envers qui les mots ne sauraient exprimer ma reconnaissance et ma gratitude. Enfin, merci à mon épouse, qui m’a encouragé et épaulé dans tous les instants.

Table des matières

INTRODUCTION GENERALE

Contexte et motivations

La dernière décennie fut marquée par des événements dramatiques, tels que des attaques terroristes de grande envergure. A cela s'ajoutent les actes criminels, de délinquance ou de grand banditisme dont nos sociétés sont l'objet de façon récurrente. La médiatisation, conduisant à une très large et très rapide diffusion des informations liées à ces actes, a probablement contribué à l'augmentation du sentiment d'insécurité dans la population. Celui-ci a poussé au premier plan la problématique sécuritaire pour en faire une priorité dans les politiques des gouvernements. Afin de répondre à certains des besoins liés à la sécurité et à l'ordre public, la vidéosurveillance se présente alors comme une solution de choix. En effet, portée par des progrès technologiques très rapides, un ensemble d'applications très variées, un marché de la sécurité florissant et stimulé par des politiques d'état incitatives (en Angleterre au début des années 1990 et, plus récemment, en France), cette technologie s'est progressivement imposée comme un moyen incontournable pour contribuer à l'amélioration de la sécurité dans les villes. Aussi, depuis une dizaine d'années, le thème de la vidéosurveillance suscite-t-il un engouement considérable aussi bien auprès des scientifiques, des industriels, des particuliers qu'auprès des instances politiques. Ainsi, les caméras de surveillance qui étaient principalement utilisées dans des espaces privés (banques, centres commerciaux, résidences privées, *etc.*) sont désormais présentes dans nombre d'espaces accueillant du public (aéroports, gares, métros, stades, musées, *etc.*). En France, le ministère de l'Intérieur estime qu'à la fin de l'année 2007, 1522 communes étaient équipées d'un système de vidéosurveillance (contre 812 en 2005), avec au total 340 000 caméras dans les espaces publics. Aujourd'hui, rien que les installations de vidéosurveillance de la RATP comportent plus de 6500 caméras, et celles de la SNCF en comptent plus de 3300.

L'apparition de la vidéosurveillance a débuté avec les systèmes CCTV (Closed-Circuit TeleVision) analogiques. Ces systèmes consistent en un certain nombre de caméras placées dans de multiples endroits et connectées à un ensemble de moniteurs placés dans une salle de contrôle. Ainsi, ils permettent à l'opérateur humain chargé de regarder les écrans de contrôler et de surveiller en temps réel ce qui se passe sur l'ensemble des sites observés. L'un des inconvénients majeurs de ces systèmes est que la détection d'événements « d'intérêt » dépend entièrement et uniquement des opérateurs humains, qui ont une capacité et une durée d'attention limitées. Par la suite, les recherches menées dans ce domaine ont tenté de développer des systèmes automatisés prenant en charge une partie de l'analyse de la scène afin d'assister l'opérateur humain pour l'émission d'alertes relatives à des événements complexes, ainsi que pour attirer son attention en temps réel. Ceci a donné lieu aux systèmes de vidéosurveillance dite « intelligente », qui peuvent prendre en charge la surveillance temps-réel « d'objets » mobiles ou immobiles dans un environnement spécifique. Les objectifs principaux de ces systèmes sont de fournir une interprétation automatique des scènes filmées, d'analyser et de prédire les actions et interactions des objets observés.

Cependant, la diversité des besoins et des fonctionnalités en matière de vidéosurveillance impose une définition précise des objectifs assignés au système et de son cadre applicatif. Ainsi, le champ couvert peut être vaste, allant de la simple détection de mouvements à la reconnaissance d'événements complexes. Par conséquent, la compréhension du domaine ciblé requiert une expertise fine afin de spécifier l'ensemble des besoins dans une solution

technique performante et adaptée. Pour répondre à ces besoins, les recherches actuelles portent principalement sur le développement d'algorithmes de traitement et d'analyse d'images, développés sous forme de fonctions réutilisables, pouvant être exploités de façon modulaire dans diverses applications de vidéosurveillance. De la sorte, on peut construire des architectures de systèmes de surveillance par assemblage de fonctions autonomes réutilisables. Cela permet de bénéficier de la relative facilité d'adaptation aux différents cas d'usage que procure une telle démarche. De manière générale, un système de vidéosurveillance intelligente est composé de plusieurs blocs de traitement permettant une analyse et une interprétation automatique des scènes observées. Afin que celui-ci puisse être efficace et fonctionner de manière optimale, la construction et l'assemblage de ses blocs doivent dépendre du cadre applicatif du système et de ses objectifs précis.

Un des domaines d'applications les plus importants pour lesquels les systèmes de vidéosurveillance intelligente sont dédiés est la surveillance des lieux accueillant du public, tels que des stations de métro, aéroports, banques ou autres administrations. Dans un tel cadre, le système est conçu de manière à fournir une interprétation automatique des scènes filmées contenant plusieurs individus - généralement dans une installation multicaméra - et de générer des alertes en temps réel sur des événements ou des comportements suspects. Les exigences d'un tel système peuvent également comprendre l'ajout de la capacité d'apprentissage automatique pour fournir la possibilité de déterminer des modèles d'activités qui doivent être reconnues comme des événements potentiellement dangereux.

Les travaux effectués dans cette thèse s'inscrivent dans le cadre du projet CAnADA (Comportements Anormaux : Analyse, Détection, Alerte). Ce projet a pour objectif l'analyse de scènes et la génération d'alertes basées sur l'analyse comportementale d'individus dans un lieu accueillant du public, par le biais d'une installation multicaméra. Les différents blocs de traitement qui constituent un tel système sont la détection, le suivi, la génération et le suivi de trajectoires ainsi que l'analyse des actions et interactions des individus concernés. Ainsi, la première opération consiste en la détection et la segmentation des personnes cibles, afin de les isoler du reste de l'image pour des analyses ultérieures. Quand un individu est détecté, le processus de suivi est alors enclenché et les informations sur sa position, la vitesse et la direction de ses mouvements sont mises à jour tout au long des séquences d'images. Puis, l'ensemble du chemin parcouru par cet individu est extrait par une étape d'analyse de trajectoire. A partir des informations de position, de vitesse de déplacement, de direction et une analyse d'actions, des caractéristiques permettant une description sémantique pourront être extraites. On pourra alors effectuer des analyses de plus haut niveau portant sur la reconnaissance de comportements et d'activités. Le système pourra alors générer des alertes en cas de détection d'activité suspecte.

Dans une telle chaine de traitement, l'un des problèmes majeurs auquel est confronté le système est celui de la « disparition » des individus cibles et leur « réémergence » dans la scène. En effet, la disparition du champ de la caméra d'un individu cible survenant lors du processus de suivi conduira à la rupture de ce processus ainsi que de sa trajectoire, interrompant ainsi l'analyse en cours de son comportement ou de son activité. Par la suite, lorsqu'il reparaitra dans la scène (qui correspond au phénomène de « réémergence » évoqué plus haut), il conviendra de mettre en place l'ensemble des mécanismes permettant de réaliser la mise en correspondance avec les données antérieures le concernant. En effet, le problème de l'analyse comportementale nécessite que les individus soient suivis sur une fenêtre temporelle suffisamment large. En effet, les caractéristiques fréquentielles du mouvement humain au sein de l'environnement observé imposent une observation sur une durée

relativement longue. Par conséquent, il importe de définir une méthode de suivi robuste capable de s'accommoder de telles contraintes opérationnelles.

Contribution et plan du manuscrit

Les travaux effectués dans le cadre de cette thèse portent sur la reconnaissance de personnes, au sens où il s'agit d'identifier un individu cible au sein d'un ensemble. L'objectif est de fournir un bloc de traitement pouvant s'incorporer dans un système de surveillance plus global dédié à l'analyse comportementale. En effet, pouvoir reconnaitre (identifier) un même individu lors de ses différentes apparitions et déplacements dans l'environnement surveillé permettra de maintenir le processus de suivi (au sens temporel) de cet individu. De cette manière, l'ensemble de ses déplacements (et donc de ses trajectoires) ainsi que de ses activités pourra être regroupé afin de permettre de faire une analyse globale de son comportement.

Dans ce contexte, nous avons développé une approche de reconnaissance de personnes basée sur l'apparence. A l'inverse des méthodes de reconnaissance biométrique telles que la reconnaissance de visage ou la reconnaissance de la démarche, notre objectif est ici de distinguer des personnes entre elles, par le biais de leur « modèle d'apparence », plutôt que d'associer un identifiant unique à chaque individu. Ce modèle d'apparence permet de fournir des éléments pertinents pour réaliser le processus de « reconnaissance ». Contrairement aux méthodes qui existent dans la littérature, le modèle d'apparence que nous construisons est issu de la modélisation séparée des parties supérieure et inférieure (appelées respectivement éléments « *haut* » et « *bas* ») des corps des individus à reconnaitre. Bien sûr, ses éléments constitutifs sont avant tout conditionnés par les vêtements portés. Par conséquent, chaque individu observé est identifié par son modèle d'apparence qui est, en pratique, défini par une combinaison de vêtements (*haut* + *bas*) plutôt que par son apparence globale correspondant au corps tout entier. Une telle démarche se rapproche de la façon dont la description d'un individu est faite sous la forme d'un « signalement », lorsqu'il s'agit de l'identifier parmi une multitude de personnes (ce signalement comportant en général beaucoup d'indications relativement à la tenue).

Le manuscrit de la thèse est organisé autour de quatre chapitres, de la manière suivante :

- Le chapitre 1 expose la problématique générale des systèmes de vidéosurveillance. Ainsi, nous précisons l'étendue des usages de ces systèmes ainsi que les différentes manières dont ils peuvent être exploités pour l'analyse de scène et le déclenchement automatique d'alertes. Nous mettons alors en évidence la structure générique de la chaine de traitements d'un système de vidéosurveillance intelligente dédié à l'analyse de scène et l'analyse d'activité d'individus dans un lieu accueillant du public. Nous décrivons alors les différents blocs de traitements constituant un tel système.
- Le chapitre 2 s'intéresse aux indispensables étapes de détection, de localisation et de segmentation d'un individu dans la scène. Nos méthodes de reconnaissance de personnes étant développées pour des environnements en intérieur avec utilisation de caméras fixes et fond de scène statique, nous avons opté pour l'utilisation d'une méthode de soustraction de fond avec modèle de fond adaptatif pour segmenter les personnes avant de tenter de les reconnaitre. Pour faciliter cette tâche de segmentation, nous avons également développé une approche de détection de la peau afin de trouver des visages sur les blobs issus de la phase de soustraction de fond qui nous permettra de nous assurer que ces blobs correspondent bien à des humains. Afin d'améliorer les taux de détection de la peau et

obtenir une meilleure robustesse, nous avons mis en œuvre une méthode de classification (basée sur les *Support Vector Machines* ou *SVM*) des pixels-peau qui permet de s'adapter en ligne, notamment, aux changements d'éclairage.

- Nous présentons dans le chapitre 3 un état de l'art des techniques de reconnaissance de personnes. La reconnaissance ou l'identification des personnes peut s'effectuer soit par des méthodes biométriques ou par des méthodes basées sur l'apparence. Nous décrivons dans ce chapitre les trois grandes approches de reconnaissance de personne utilisées dans des applications de vidéosurveillance, à savoir la reconnaissance de visage, la reconnaissance de la démarche et la reconnaissance basée sur l'apparence. Nous évoquons les principales méthodes développées et mettons en évidence les difficultés inhérentes à chacune de ces trois approches.
- Dans le quatrième chapitre, nous présentons l'approche de reconnaissance basée sur l'apparence que nous avons développée. Comme nous l'avons évoqué, le « modèle d'apparence » de chaque individu à reconnaitre est constitué par le vêtement du haut et le vêtement du bas qu'il porte. Afin de modéliser l'apparence de chaque vêtement, une « signature », composée d'un ensemble de vecteurs de caractéristiques couleur et texture, est extraite. Afin d'effectuer les phases d'apprentissage et de reconnaissance, nous avons mis au point une stratégie de classification (basée sur la technique one-class *SVM*) nous permettant d'effectuer deux opérations. Nous appelons la première opération la « fusion des classes », qui consiste à détecter en amont (c'est-à-dire lors de l'apprentissage) les classes de vêtement similaires et les rassembler en une seule classe. Cette opération est effectuée afin d'anticiper et d'éviter des confusions entre les classes lors de la reconnaissance des individus. La deuxième opération est un « apprentissage en ligne ». Cette dernière permet au système, lors de la phase de reconnaissance, de reconnaitre que l'individu présent dans la scène correspond à une « nouveauté » (c'est-à-dire que cet individu n'a pas été appris), puis de l'incorporer dans la base d'apprentissage. Nous concluons ce chapitre en présentant les résultats des procédures d'apprentissage et de reconnaissance obtenues sur notre base de données contenant des séquences d'images de 54 individus.

CHAPITRE 1

1. LA VIDÉOSURVEILLANCE INTELLIGENTE

1.1 Introduction

Les événements récents, tels que les attaques terroristes ou les actes de banditismes de grande envergure ont conduit à l'augmentation de la demande de sécurité dans la société. Cela a incité les gouvernements à faire de la sécurité une priorité dans leurs politiques, ce qui a permis le développement et le déploiement de grands systèmes de vidéosurveillance. A titre d'exemple, le nombre de caméras de surveillance au Royaume-Uni est estimé à 4,5 millions en 2009 [Liberty09], ce qui représente environ une caméra pour 14 habitants, et fait ainsi du Royaume-Uni le leader mondial dans ce domaine. D'autres pays comblent leur retard par une adoption rapide des technologies de vidéosurveillance dont la France, les Etats-Unis, les Pays-Bas, la République d'Irlande et l'Italie. L'augmentation de l'utilisation de la vidéosurveillance est également croissante au Moyen-Orient, en Afrique du Sud, en Australie, en Inde, en Russie et en Europe de l'ouest.

Les systèmes de vidéosurveillance dite « intelligente » s'occupent de la surveillance temps-réel « d'objets » mobiles ou immobiles (généralement des personnes, des véhicules ou divers colis et bagages) dans un environnement spécifique. Les objectifs principaux de ces systèmes sont de fournir une interprétation automatique des scènes filmées et de comprendre et prédire les actions et interactions des objets observés. Un système de vidéosurveillance dispose de 3 fonctions importantes et interdépendantes (**Figure 1.1**) : (1) la fonction « réception », constituée par les caméras, qui assurent la collecte des données, essentiellement des images ou des séquences d'images, (2) la fonction « gestion », constituée par les serveurs, qui centralisent l'acquisition, l'analyse et le stockage de celles-ci et (3) la fonction « visualisation » constituée par les moniteurs ou terminaux qui permettent la visualisation des images et le pilotage du système.

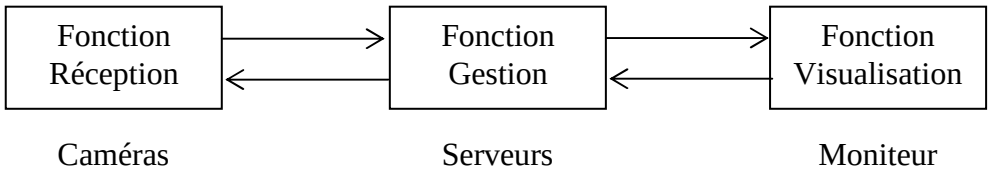


Figure 1.1 : Synoptique d'une installation de vidéosurveillance.

L'orientation des recherches sur le développement des systèmes de vidéosurveillance a considérablement évoluée dans le temps. Lors de la mise en place des premiers systèmes, les chercheurs tendaient à développer des outils (algorithmes et matériels) pour des applications spécifiques, telles que la détection d'intrusion et la détection d'objets abandonnés. Actuellement, les recherches menées tendent plus communément à développer et améliorer des algorithmes de traitement et d'analyse d'images (tels que la détection, le suivi, la reconnaissance, *etc.*) de manière à pouvoir être exploités de façon modulaire. De cette manière, on peut construire des architectures de systèmes de surveillance par assemblage de fonctions autonomes réutilisables. Cela permet de bénéficier de la « réutilisabilité » des

composants développés, ainsi que de la relative facilité d'adaptation aux différents cas d'usage que procure une telle démarche. Certaines recherches s'occupent également du développement de nouvelles solutions pour la communication vidéo dans les systèmes de surveillance distribuée ou de la conception de nouveaux capteurs d'image qui intègrent directement des algorithmes complexes de surveillance.

1.2 La vidéosurveillance : un marché en plein essor

Depuis une dizaine d'années, le thème de la vidéosurveillance suscite un engouement considérable aussi bien auprès des scientifiques, des industriels, des particuliers et des instances politiques. L'intérêt grandissant de la recherche scientifique dans ce domaine est notamment illustré par le nombre de conférences, workshops et revues qui y sont dédiés [AVSS+] [PETS+] [ICNS+] [BioSurv07] [Auto07ID] [BSYM+] [TIWDC08] [PASSIVE08] [ICDSC09] [ICCV09] [ICCST09] [RADAR09] [WIFS09] [ISSNIP09] [SPJ+]. La très forte augmentation en France, du chiffre d'affaires des quelques 200 entreprises travaillant dans ce secteur est très significative du rapide essor de cette technologie et de son succès commercial. Il est passé de 473 millions d'euros en 2000, à plus de 750 millions d'euros en 2006 et représente aujourd'hui 5% de la dynamique de marché de la sécurité privée. En témoigne également la politique incitative de l'état français qui valorise cet outil et cherche à assurer son développement dans les espaces publics par différents moyens (aides financières, réglementation).

Plusieurs raisons peuvent expliquer cet essor :

- **Les progrès technologiques très rapides**, notamment la mise au point de divers algorithmes puissants de vision par ordinateur (détection de mouvement, suivi de personnes, détection de visage, *etc.*). Il y a également le passage de l'analogique au numérique, qui facilite la connectique, autorise l'usage au travers de réseaux LAN ou WAN, améliore la robustesse des traitements et, globalement, la fiabilité des systèmes de vidéosurveillance.

- **La multiplication des finalités assignées à la vidéosurveillance**. Elle n'est plus utilisée uniquement à des fins sécuritaires, mais est aussi mobilisée pour d'autres tâches telles que la régulation du trafic routier, la gestion urbaine, l'assistance aux personnes, l'analyse de comportement de clients (dans les centres commerciaux), *etc.*

- **Un contexte international instable marqué par des actes terroristes,** en particulier depuis les attentats du 11 septembre 2001, qui a conduit à l'émergence d'une exigence sécuritaire dans les politiques gouvernementales. C'est d'ailleurs en raison de ce contexte international que l'état français a engagé une politique de développement de la vidéosurveillance depuis 2003.

- **L'évolution de l'image de la vidéosurveillance qui est devenue aujourd'hui plus positive auprès de l'opinion publique**. Le discours sur son efficacité pour lutter contre la délinquance, véhiculé par les séries télévisées policières et par les industriels et distributeurs de matériel de vidéosurveillance, a sans doute participé à la valorisation de cette technologie. De même, la médiatisation du rôle déterminant joué par la vidéosurveillance dans diverses affaires criminelles, notamment dans l'identification des auteurs des attentats terroristes de Londres de juillet 2005, a contribué à son succès. Par ailleurs, l'introduction du mot « vidéoprotection » en substitution à celui de

« vidéosurveillance » illustre la volonté des instances à le faire mieux accepter par l'opinion publique.

1.3 Politique incitative pour le développement et le déploiement des systèmes de vidéosurveillance

En l'espace d'une dizaine d'années, les caméras de surveillance qui étaient principalement utilisées dans des espaces privés (banques, centres commerciaux, résidences privées, *etc.*) sont aujourd'hui présentes dans nombre d'espaces accueillant du public. Portée par un marché de la sécurité florissant et stimulée par des politiques d'état incitatives (en Angleterre au début des années 1990 et, plus récemment, en France), cette technologie s'est progressivement imposée comme un moyen incontournable pour contribuer à l'amélioration de la sécurité dans les villes. En France, elle séduit nombre de municipalités qui en font un élément important de leur stratégie de lutte contre la délinquance. Le ministère de l'Intérieur (Direction des Libertés Publiques et des Affaires Juridiques - DLPAJ) estime ainsi qu'à la fin de l'année 2007, 1522 communes étaient équipées d'un système de vidéosurveillance (contre 1142 en 2006 et 812 en 2005) [DLPAJ07]. Cette croissance a été notamment expliquée par les possibilités de financement des dispositifs de vidéosurveillance qui ont été offertes par une circulaire du 6 juin 2006 [DLPAJ07]. Par ailleurs, la DLPAJ estime également à 90175 le nombre d'autorisations préfectorales délivrées pour l'installation de systèmes de vidéosurveillance entre 1997 et 2007, avec 9762 rien qu'en 2007, contre 3607 en 2000. La principale raison avancée par le gouvernement est, en premier lieu, d'assurer la sécurité et l'ordre publics. De manière significative, dans un sondage IPSOS réalisé en mars 2008 auprès d'un échantillon de 972 individus, 71% des personnes interrogées se disent favorables à la présence de la vidéosurveillance dans les lieux publics, 43% pensent qu'il n'y a pas assez de caméras dans les espaces publics et 65% considèrent que la vidéosurveillance permettra de lutter efficacement contre la délinquance et le terrorisme [IPSOS08]. Par ailleurs, le développement de la vidéosurveillance dans les espaces publics s'explique aussi par une raison plus symbolique : rassurer les populations en exhibant une preuve « visible » de la prise en considération de la sécurité dans la politique menée par le gouvernement.

Ainsi, le recours à la vidéosurveillance par les autorités repose sur quatre présupposés **:**

1 - La vidéosurveillance jouerait un rôle dissuasif permettant de diminuer la délinquance. Les délinquants potentiels reconsidèreraient leurs actes devant cette surveillance technique d'un espace et préfèreraient soit ne pas commettre de délit, soit le commettre ailleurs. Dans cette perspective, la vidéosurveillance est considérée comme un « gardien compétent ».
2 - La vidéosurveillance aide à l'élucidation des délits et des désordres en constituant une preuve à charge dans les enquêtes judiciaires et en aidant à l'arrestation des auteurs de l'acte.
3 - La vidéosurveillance permet un déploiement approprié des forces de police. Elles pourraient ainsi ne se déplacer que lorsque cela est nécessaire et adapter les effectifs envoyés sur le terrain.
4 - La vidéosurveillance a un impact positif auprès de la population, et permet de diminuer le sentiment d'insécurité.

1.4 Evolution des systèmes de vidéosurveillance intelligents

L'évolution technologique des systèmes de vidéosurveillance a débuté avec les systèmes CCTV (Closed-Circuit TeleVision) analogiques. Ces systèmes consistent en un certain nombre de caméras situées dans de multiples endroits relativement distants et connectées à un

ensemble de moniteurs, généralement placés dans une seule et même salle de contrôle. Actuellement, la majorité des systèmes de vidéosurveillance utilisent des caméras à capteurs CCD (Charge-Coupled Device, ou dispositif à transfert de charge) pour capturer des images discrétisées, et des techniques analogiques pour la distribution et le stockage de ces images. Les images sont converties en un signal vidéo analogique composite, qui est relié aux écrans et au matériel d'enregistrement. L'amélioration technologique apportée par ces nouveaux systèmes a conduit au développement de dispositifs semi-automatiques, connus sous le nom de « systèmes de surveillance de deuxième et de troisième génération ». Nous regroupons donc ci-après les systèmes de vidéosurveillance en trois générations :

- **1ère génération (1960-1980) :** ces systèmes sont basés sur des sous-systèmes analogiques pour l'acquisition, la transmission et le traitement des images. Ils étendent l'œil humain - au sens spatial - en transmettant les sorties de plusieurs caméras surveillant un ensemble de sites à des écrans dans une salle de contrôle centrale. Ces systèmes ont des inconvénients majeurs tels que la nécessité de disposer d'une large bande passante pour la transmission des données, la difficulté d'archivage et de récupération d'événements en raison du besoin d'une grande quantité de bandes vidéo, et la difficulté de détection d'événements en temps réel qui dépend uniquement des opérateurs humains, qui ont une capacité et une durée d'attention limitées.

- **2ème génération (1980-2000) :** les systèmes de deuxième génération sont « hybrides » dans le sens où ils utilisent des sous-systèmes analogiques et numériques pour pallier certains inconvénients de leurs prédécesseurs. Ils mettent en œuvre les premières avancées des méthodes de traitement de vidéos numériques qui fournissent une assistance à l'opérateur humain en filtrant les faux événements. La plupart des travaux durant la deuxième génération a été axée sur la détection d'événements en temps réel.

- **3ème génération (2000-) :** la troisième génération de systèmes de surveillance gèrent l'acquisition et le traitement des images au niveau du capteur, la communication par réseaux hétérogènes large bande fixe et mobile, et le stockage d'images au niveau des serveurs centraux bénéficiant d'infrastructures numériques à bas coût. Contrairement aux générations précédentes, la troisième génération permet le traitement et la distribution d'images au niveau du capteur en utilisant des caméras intelligentes capables de numériser et compresser le signal acquis ainsi que l'application d'algorithmes de traitement d'images. Elle gère également le stockage distribué et la récupération de données vidéo basées sur le contenu. Cette génération de systèmes propose un panel de fonctionnalités lui permettant de passer du stade de recueil de l'information à celui de l'analyse en ligne des situations observées. En définitive, comparée aux systèmes de génération précédente et à la simple fonction « d'enregistrement » et de constat *a posteriori* auquel ils étaient limités, les systèmes de troisième génération se voient attribuer un réel rôle d'analyse de la scène. Ainsi, un des principaux objectifs d'un système de vidéosurveillance de troisième génération est de permettre une bonne compréhension de la scène, afin d'assister l'opérateur humain pour l'émission d'alertes relatives à des événements complexes et pour attirer son attention en temps réel, notamment dans un environnement proposant des informations multicapteur. Les exigences de ces systèmes peuvent également comprendre l'ajout de la capacité d'apprentissage automatique afin de modéliser les scènes et détecter automatiquement des événements potentiellement dangereux qui s'y rapportent.

1.5 La vidéosurveillance sur le terrain et les problèmes rencontrés par les opérateurs

Dans la majorité des grandes installations de vidéosurveillance qui comprennent des centaines de caméras, seule une petite fraction est effectivement surveillée. En effet, Dee et al. [DeeV08] rapportent que, dans une étude de quatre installations de collectivités locales au Royaume-Uni en 2007, le rapport écran / caméra est situé entre 1 pour 30 et 1 pour 4 tandis que le rapport opérateur / écran peut descendre jusqu'à 1 pour 16. Ainsi, alors qu'en théorie toutes les caméras devraient être scrutées, seule une petite fraction peut être suivie en temps réel par un opérateur. Le reste n'est éventuellement regardé qu'à la suite d'un incident (visionnage de vidéos enregistrées). En pratique, il est reconnu que chaque opérateur ne peut efficacement surveiller que 1 à 4 écrans à la fois [WD88]. L'attention de celui-ci ne peut être sollicitée de façon continue que sur une durée limitée et il est recommandé de faire une pause de 5 à 10 minutes toutes les heures pour des raisons de santé et de sécurité [WD88].

Généralement, les systèmes de vidéosurveillance actuels laissent l'initiative des caméras à regarder aux opérateurs eux-mêmes, ce qui peut conduire à une exploitation sous-optimale, voire à l'émergence des soucis d'ordre déontologique. Par exemple, un des problèmes signalés par des études sociologiques [MN03] [Sta03] est qu'il n'est pas rare que les opérateurs se basent sur l'apparence des personnes filmées plutôt que sur leur comportement. De fait, devoir gérer et trier un grand nombre d'informations visuelles amène les opérateurs à se focaliser sur « une gamme étroite de caractéristiques facilement repérables plutôt que sur les comportements suspects qui le sont moins », pratique également appelé phénomène de « tri social ». Ainsi, dans une étude réalisée en 1999, sur la base d'une observation du travail des opérateurs d'une salle de contrôle, Norris et Armstrong [NorA99] ont mis en évidence que 86 % des personnes surveillées avaient moins de trente ans, 93% étaient de sexe masculin, et que 68 % des gens de couleur soumis à une attention particulière de la part des opérateurs le sont « sans raison apparente ». Ceci a notamment attiré l'attention des défenseurs des droits de l'homme et des groupes anti-surveillance [Liberty].

Les opérateurs humains souffrent également du problème récurrent de l'ennui ; dans la grande majorité des situations de surveillance, rien ne se passe. La lecture de journaux ou les fréquentes pauses contribuent à atténuer celui-ci. Pour l'anecdote, un opérateur a même admis avoir exclusivement porté son attention pendant toute une nuit sur la caméra qui filmait sa propre voiture [Smi04]. Enfin, tous les opérateurs ne sont pas toujours forcément fiables, en témoigne le cas des opérateurs « Sefton Conseil » accusés de voyeurisme [BBCnews05].

L'un des enjeux est de permettre une utilisation plus efficace de l'expertise humaine en la réservant aux cas les plus anormaux. En traitant de façon automatique, un premier niveau d'alerte, le système de vidéosurveillance intelligent limitera la visualisation uniquement aux scènes les plus complexes, là où la connaissance humaine sera indispensable et là où l'opérateur aurait un intérêt intellectuel et une motivation dans l'accomplissement de sa mission.

1.6 Domaines d'application de la vidéosurveillance intelligente

Comme nous l'avons évoqué, la demande croissante de sécurité dans la société conduit à un besoin accentué d'applications de surveillance dans de nombreux environnements. Le but des paragraphes qui vont suivre est de dresser un rapide panorama d'applications parmi les plus emblématiques de la vidéosurveillance. En particulier, la demande de surveillance à des fins de sûreté et de sécurité a reçu une attention particulière, notamment dans les domaines suivants :

1.6.1 Sécurité de sites publics et commerciaux

- Surveillance des banques, centres commerciaux, aéroports, ports, gares, métros, musées, stades, propriétés privées et parkings pour la prévention et la détection d'intrusions, de crimes ou d'actes de dégradation de biens [DimSG09] [KraTYP09].
- Surveillance des autoroutes et des chemins de fer pour la détection d'accidents [YonCJ04] [Kos09].
- Surveillance des forêts pour la détection d'incendies [Adelie].
- Observation des personnes âgées et des personnes à mobilité réduite pour déclencher des alarmes en cas de problème, ainsi que pour mesurer l'efficacité de traitements médicaux [ZouBT09].
- Contrôle d'accès à des sites privés.

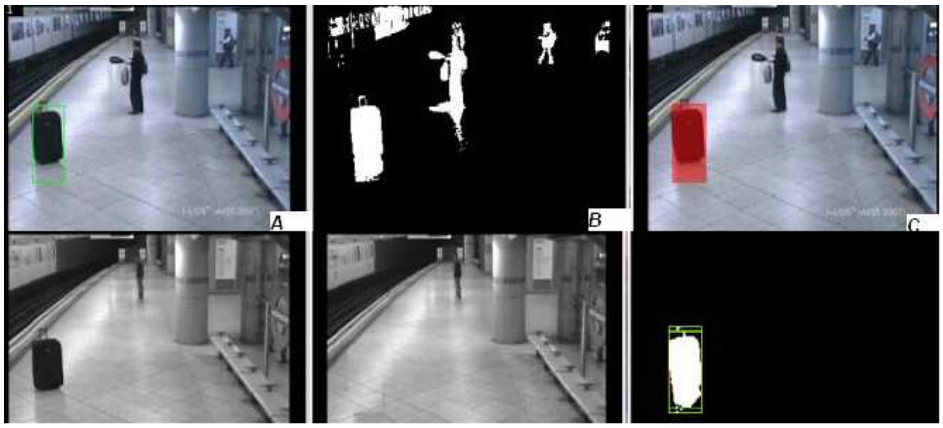


Figure 1.2 : Exemple de détection d'objet abandonné [SinSMM09].

1.6.2 Extraction d'informations

- Mesure du débit de la circulation routière et d'encombrement de piétons [LiTCW08] [WasLCW09].
- Etablissement de profils de consommateurs dans des centres commerciaux [Cliris].
- Comptage de personnes dans des espaces publics [HarBD05] [ZhaDC09].
- Extraction de statistiques dans des activités sportives [WanP03] [XioRD03].

Figure 1.3 : Exemple de détection de personne dans une zone interdite [DeeV08].

1.6.3 Applications militaires

- Surveillance des frontières.
- Surveillance de sites sécurisés.
- Assistance du contrôle et du commandement sur le terrain.

1.6.4 Applications civiles

- Mesure de la vitesse des véhicules [ZhuHLL09].
- Détection de franchissement de feu rouge et de franchissement de ligne continue [ZhaH09].

Figure 1.4 : Exemple de surveillance du trafic routier [Kos09].

1.7 Exploitation de la vidéosurveillance pour l'analyse de scènes et le déclenchement d'alertes

L'analyse de scène permet l'interprétation d'événements et le déclenchement d'alertes, qui sont des objectifs majeurs d'un système de vidéosurveillance. En effet, ces traitements permettent *a minima* d'assister et de suppléer l'opérateur humain dans la reconnaissance d'événements caractéristiques (et potentiellement dangereux) et ainsi déclencher automatiquement des alertes. Plus le système est apte à effectuer cette tâche, moins l'intervention de l'homme est nécessaire, soulageant ainsi l'opérateur de la nécessité d'observer en permanence tous les flux vidéo issus des caméras. Des environnements différents auront des besoins de surveillance différents, allant de la simple détection de mouvements à la reconnaissance d'événements complexes [KirRIR08] [BenJSR09]. Alors que des systèmes capables d'effectuer la détection de mouvement et le suivi de cibles sont développés depuis deux décennies, les systèmes traitant de la détection d'anomalies survenant dans l'environnement surveillé sont beaucoup plus récents.

Un événement dit « anormal » peut être automatiquement détecté comme étant un écart par rapport à des modèles communs d'activité [IvaDHE09] [WanAR09]. Par conséquent, savoir comment définir et représenter un comportement « normal » est un point crucial. Des recherches sur la reconnaissance d'événements complexes ont été menées au cours des quelques dernières années. Dans ce qui suit, nous décrivons certains cas d'usage des systèmes de vidéosurveillance pour l'aide au déclenchement d'alertes. Nous pouvons définir deux types d'alertes qui peuvent être générés par un système de surveillance intelligente :

- Des alertes d'événements définis par l'utilisateur (nécessitant une modélisation « explicite » des événements),

- Des alertes d'événements non-définis par l'utilisateur (avec une modélisation « implicite » des événements).

1.7.1 Alertes d'événements définis – modélisation explicite

Dans cette approche, on trouve tous les systèmes qui nécessitent la définition explicite de ce qui constitue des événements normaux ou anormaux [KovUSHP09]. Ici, le système détecte une variété d'événements définis par l'utilisateur qui se produisent dans l'espace surveillé et avertit l'opérateur en temps réel. Ainsi, il appartient par la suite à l'opérateur d'évaluer la situation et de prendre des mesures préventives. Quelques événements typiques sont présentés ci-après.

1.7.1.1 Alertes basées sur le mouvement d'objets

Ces alertes dépendent des propriétés des mouvements des objets dans l'espace surveillé. Voici quelques exemples courants:

- Détection de mouvement de tout objet dans une zone spécifiée [BaySM09].
- Détection de caractéristiques de mouvement d'objets, telles que la direction spécifique du mouvement (direction vers une zone interdite), ou la vitesse (objet en mouvement trop rapide) [PioNC09].
- Détection d'objets abandonnés, par exemple, un bagage sans surveillance dans un aéroport, ou une voiture garée dans une zone interdite [BhaCRA07] [MagTBS09] [SinSMM09].
- Détection d'enlèvement d'objets qui ne devraient pas se déplacer, par exemple, un tableau dans un musée [MigM08] [MagTBS09].

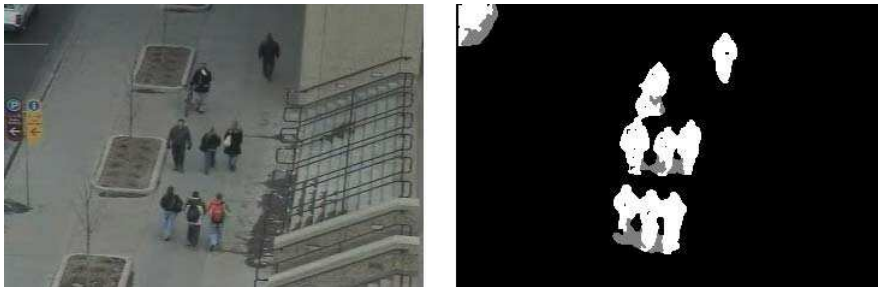


Figure 1.5 : Exemple de comptage de personnes [FehSLY09].

1.7.1.2 Alertes basées sur la classe d'objets

Ces alarmes prennent en compte le type d'objets en plus des propriétés de leurs mouvements. On peut citer à titre d'exemple :

- Détection de mouvement de classes spécifiques. Par exemple, lors de la surveillance des pistes d'atterrissage d'un aéroport, le système déclencherait une alerte en cas de présence ou de mouvement spécifiques d'individus sur le tarmac, mais pas celles des avions [DimSG09].

- Comptage du nombre de personnes. Par exemple, passage de plus d'une personne dans un portique de sécurité, ou sur la densité d'une foule dans une discothèque [FehSLY09] [ZhaDC09].

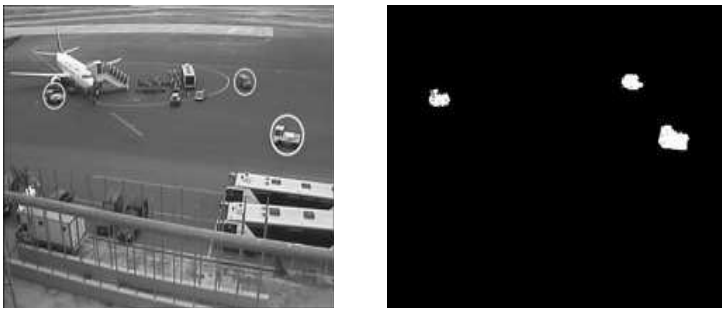


Figure 1.6 : Exemples de surveillance d'un aéroport [DimSG09].

1.7.1.3 Alertes basées sur le comportement

Ces alarmes sont généralement basées sur l'apprentissage et l'analyse des habitudes de déplacement d'objets sur de longues périodes de temps. Ces alarmes sont utilisées dans des applications spécifiques et utilisent une quantité importante d'informations de contexte, par exemple:

- Détection de comportements suspects dans les parkings de stationnement, par exemple, une personne essayant d'ouvrir plusieurs voitures [HamBBD09].
- Détection de va-et-vient, par exemple une même personne faisant plusieurs va-et-vient aux alentours d'un site surveillé [MorT08].

1.7.2 Alertes d'événements non-définis – modélisation implicite

Contrairement aux alertes d'événements définis, ici aucune connaissance *a priori* n'est fournie, le système génère des alertes quand il détecte une activité qui s'écarte de la norme ou dite « activité habituelle » [IvaDHE09]. Par exemple, lors de la surveillance d'une rue, le système apprend que les véhicules se déplacent sur la route et les personnes se déplacent sur le trottoir. Basé sur ce modèle, le système fournira une alerte si une voiture roule sur le trottoir, ou si un piéton se déplace sur la route. Ce type de démarche sous-tend des questions complexes : En premier lieu, quelles sont les données pertinentes à recueillir pour faire l'apprentissage des différents modèles utiles ? Il est intuitivement clair que le choix de celles-ci dépendra des applications envisagées. Ensuite, cet apprentissage doit il être réalisé « hors ligne » ou « en ligne » ? Cela revient à savoir si l'on dispose ou non *a priori* d'une collection de données étiquetées représentatives des situations d'intérêt. Par ailleurs, on peut concevoir que cet étiquetage ne soit pas nécessairement connu (occurrence d'une situation originale), ce qui pose alors la question de la mise en œuvre d'un apprentissage supervisé, semi-supervisé (connaissance partielle des classes d'intérêt avec détection de la nouveauté) ou non-supervisé. Dans le cas d'un apprentissage supervisé hors ligne, on devra supposer que les bases d'apprentissage et de test relatent exhaustivement les situations que le système sera susceptible de rencontrer lors de son exploitation. De telles bases peuvent être fastidieuses à générer, pour peu que leur constitution soit possible. Enfin, qu'est-ce en définitive qu'une situation « anormale » ? Il est sans doute plus facile de l'associer à une situation « non habituelle », au sens ou elle s'écarte (d'une façon qu'il faut pouvoir quantifier) de la situation

habituellement constatée dans des conditions comparables. Cette notion de « conditions comparables » fait elle-même intervenir un panel de facteurs, dépendants (tout comme les caractéristiques utilisées pour la modélisation) du contexte applicatif. La sélection de ces derniers peut s'avérer une opération délicate. L'ensemble de ces questions a naturellement attiré l'attention de chercheurs qui se sont attachés à proposer des solutions.

Les caractéristiques utilisées pour la construction des modèles d'activités sont principalement des caractéristiques de mouvement telles que la vitesse ou la trajectoire. Par exemple, une approche de reconnaissance d'événements basée sur l'apprentissage de signatures de trajectoires dans un système multicaméra est présentée dans [SniPF06]. Ivanov et al. [IvaDHE09] définissent un système de détection d'événements « inhabituels » en terme de vitesse et d'accélération des mouvements des individus. Krats et al. [KraN09] proposent une approche statistique de modélisation de comportement en termes de mouvement de foule. Micheloni et al. [MicSF06] présentent un système capable d'apprendre les statistiques de base sur les évènements en cours dans l'environnement surveillé, et étendent cette approche dans [MicSF09] et proposent l'analyse d'événements composés en ré échantillonnant tous les événements plus simples qui les composent.

De manière générale, la modélisation implicite d'événements rend le système très adaptable à différents scénarios et situations, mais devient inadaptée pour la détection d'événements complexes. D'un autre côté, la modélisation explicite apporte généralement de meilleurs résultats en termes de nombre de fausses alertes et d'alertes manquées que la modélisation implicite. Cependant, elle n'est pas auto-adaptable puisque tout l'ensemble des connaissances doit être apporté par l'opérateur.

1.8 Chaine de traitements d'un système de vidéosurveillance intelligente pour l'analyse de scène

La diversité des besoins et des fonctionnalités en matière de vidéosurveillance impose une définition précise des objectifs assignés au système et de son cadre applicatif. Ainsi, le champ couvert peut être vaste : un système de vidéosurveillance pouvant intervenir en observation d'une zone en continu ; en analyse d'événements ponctuels, en surveillance contre le vol, les agressions ; en surveillance du bon fonctionnement d'activités ou de processus, en collecte d'informations et extraction automatique de la connaissance, *etc.* La compréhension du domaine ciblé nécessite donc une expertise fine afin de spécifier l'ensemble des besoins dans une solution technique performante et adaptée. En effet, si la vidéosurveillance peut servir à plusieurs types d'activités (sécurité, gestion urbaine, trafic routier, analyse d'activité, aide au déploiement des forces de police, *etc.*), il est difficilement pensable de considérer que tous ces objectifs peuvent être atteints au même moment, avec un même dispositif et les mêmes opérateurs. Le risque est de considérer la vidéosurveillance comme une « machine à tout faire » et de ne pas lui donner d'objectif précis, ce qui mènerait à disperser les efforts et par conséquent de ne pas être efficace. En outre, pour traiter les quantités gigantesques d'informations issues de grandes installations de vidéosurveillance (on peut citer celle de la RATP qui comporte 6500 caméras, celle de la SNCF qui en compte 3300, le métro de Londres et l'aéroport d'Heathrow qui sont équipés de plus de 6000 caméras chacun) des questions relatives à la réactivité (quand et pourquoi le système doit-il émettre une alerte) et à la diffusion de données (comment, quand et à qui l'information doit être transmise) deviennent primordiales.

Pour répondre à ces besoins, des recherches sont constamment menées dans les milieux académiques et industriels pour trouver des solutions innovantes. A cet effet, les recherches actuelles portent principalement sur le développement d'algorithmes (développés sous forme de fonctions réutilisables) pouvant être utilisés de façon modulaire dans diverses applications de vidéosurveillance. De manière générale, un système de vidéosurveillance intelligente est composé de plusieurs blocs de traitement permettant une analyse et une interprétation automatique des scènes observées. La **Figure 1.7** illustre la structure générique d'un système de vidéosurveillance intelligente multicaméra pour l'analyse automatique de scène. Des traitements et des analyses sont d'abord effectués au niveau de chaque caméra. Généralement, des opérations de bas niveau tels que la détection d'objets sont effectuées afin d'isoler les éléments d'intérêt du reste de l'image. A un niveau intermédiaire, l'extraction de caractéristiques de ces objets, telles que des caractéristiques spatiotemporelles (mouvement et trajectoire) ou d'apparence (forme, couleur) sont extraites. Puis, des descriptions sémantiques et des étapes de plus haut niveau sont mobilisées pour l'analyse et l'interprétation d'actions, d'activités ou de comportements. Eventuellement, une étape d'évaluation au niveau local est effectuée avant la fusion des données de toutes les caméras. Enfin, les données issues de l'ensemble des caméras sont fusionnées pour permettre globalement une analyse et une évaluation sur l'ensemble du réseau.

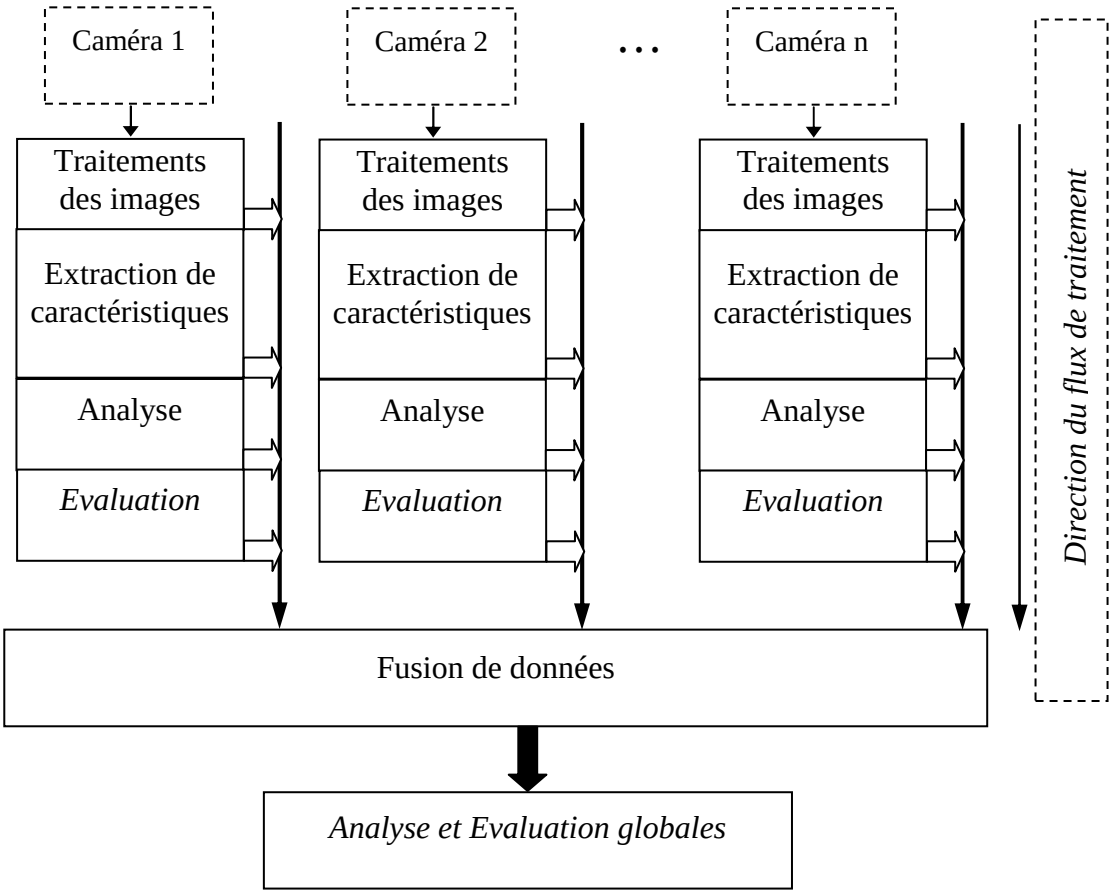


Figure 1.7 : Schéma d'un système de vidéosurveillance intelligent multicaméra pour l'analyse de scène.

Afin que le système de vidéosurveillance puisse être efficace et fonctionner de manière optimale, la construction et l'assemblage de ses blocs doivent dépendre du cadre applicatif du système et de ses objectifs précis. En effet, à titre d'exemple, un système de vidéosurveillance dédié à la surveillance du trafic routier ne sera pas conçu de la même manière qu'un système dédié à la surveillance d'un espace public. Un des domaines d'applications les plus importants pour lesquels les systèmes de vidéosurveillance intelligente sont dédiés est la surveillance des lieux accueillant du public, tels que des stations de métro, aéroports, banques, administrations (préfecture, mairie) *etc.* Dans un tel cadre, le système est conçu de manière à fournir une interprétation automatique des scènes filmées contenant plusieurs individus - généralement dans une installation multicaméra - et de générer des alertes en temps réel sur des événements ou des comportements suspects. Les exigences d'un tel système peuvent également comprendre l'ajout de la capacité d'apprentissage automatique pour fournir la possibilité de déterminer des modèles d'activités qui doivent être reconnues comme des événements potentiellement dangereux.

Ainsi, les travaux effectués dans cette thèse s'inscrivent dans le cadre du projet CAnADA (Comportements Anormaux : Analyse, Détection, Alerte). Ce projet a pour objectif l'analyse de scènes et la génération d'alertes basées sur l'analyse comportementale dans un lieu accueillant du public, par le biais d'une installation multicaméra. La **Figure 1.8** illustre le prototype d'une chaine de traitement dans un tel cadre applicatif. Le premier bloc de traitement consiste en la détection et la segmentation des personnes au premier plan, afin de les isoler du reste de l'image pour des analyses ultérieures. Quand un individu est détecté, le processus de suivi est alors enclenché et les informations sur sa position, la vitesse et la direction de ses mouvements sont mises à jour tout au long des séquences d'images. Puis, l'ensemble du chemin parcouru par cet individu est extrait par une étape d'analyse de trajectoire. A partir des informations de position, de vitesse de déplacement, de direction et une analyse d'actions, des caractéristiques permettant une description sémantique pourront être extraites. On pourra alors effectuer des analyses de plus haut niveau portant sur la reconnaissance de comportements et d'activités. Le système pourra alors générer des alertes en cas de détection d'activité suspecte. En plus des différentes étapes de traitements et d'analyse citées, une modélisation de l'environnement peut également être ajoutée afin de fournir une connaissance supplémentaire pouvant être exploitée pour une meilleure analyse de scène.

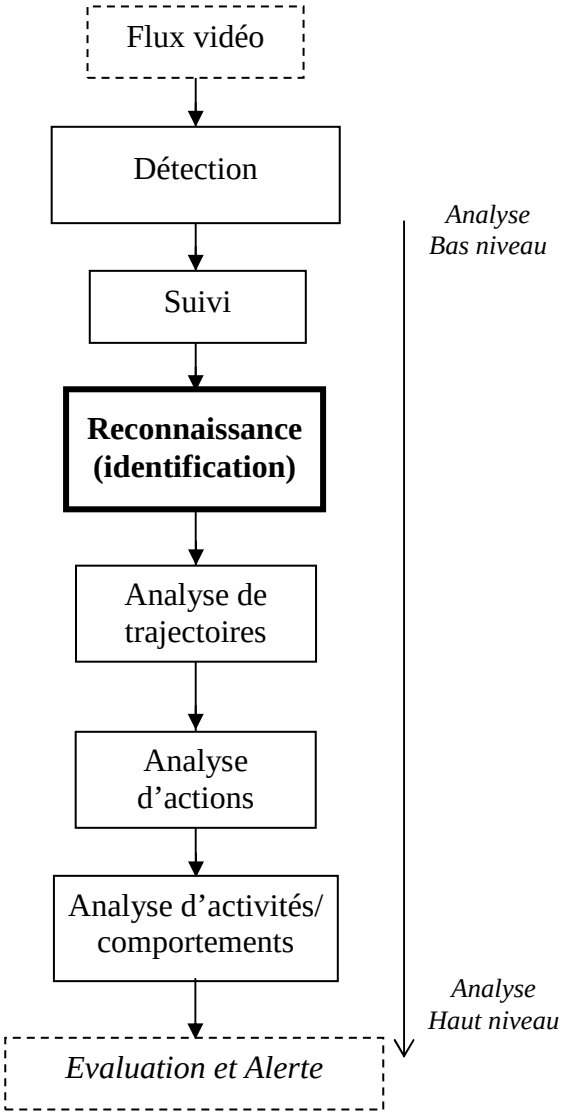


Figure 1.8 : Prototype d'une chaine de traitement pour l'analyse d'activités d'individus.

Dans la suite de ce chapitre, dans le but de positionner les travaux de la thèse, nous présentons brièvement les différents blocs de traitement que comporte un système de vidéosurveillance dédié à l'analyse de scène et l'analyse d'activité d'individus.

1.8.1 La détection

Comme nous venons de le constater, il y a 3 points clés pour l'analyse de scène et l'interprétation d'activités dans des vidéos de surveillance :

- la détection d'individus d'intérêt,
- le suivi de ces individus tout au long des séquences,
- l'analyse de leurs déplacements (trajectoires) ainsi que de leurs actions.

Aussi, la détection est une étape décisive et incontournable. En effet, avant de suivre et d'analyser les caractéristiques d'un individu cible (mouvement, vitesse, direction, trajectoire, taille, forme, *etc.*), il est nécessaire de le détecter et de le segmenter du reste de l'image. Ainsi,

l'objectif de la détection est de localiser les individus au premier plan dans la scène pour des analyses ultérieures. L'opération de détection peut s'effectuer sur chaque image de la séquence ou en début de séquence pour initialiser le mécanisme de suivi. La procédure de détection utilise généralement tout ou partie des informations disponibles jusqu'à l'instant courant (dans le cadre d'une démarche « causale »). Ainsi, certaines méthodes utilisent des informations temporelles calculées à partir de la séquence d'images afin de réduire le nombre de fausses détections, comme par exemple, réduire la zone de recherche dans l'image. Ces informations temporelles sont généralement obtenues en effectuant la différence entre deux images successives, pour faire ressortir les zones de changement entre ces images. Une fois que les régions correspondant aux individus d'intérêt sont détectées dans l'image, il est ensuite du rôle du suivi de maintenir la correspondance de chaque individu pour chaque image de la séquence afin de générer sa trajectoire. Les questions liées aux occultations, variations d'apparence, changements d'éclairage, vitesse de déplacement doivent être prises en compte. De nombreuses méthodes ont été développées pour la détection. Des états de l'art des méthodes existantes peuvent être trouvées dans [RadAAR05], [AlpOM06] et [BugP09].

Dans les applications où l'on utilise des caméras fixes avec fond statique, l'approche de détection la plus populaire et la plus utilisée est la soustraction de fond [BaySM09]. Cette approche consiste à estimer une représentation appropriée de la scène appelée modèle de fond (que celui-ci soit préenregistré ou construit de façon dynamique), puis à chercher tout changement ou déviation par rapport au modèle dans chaque image d'entrée. Il existe diverses méthodes de soustraction de fond dans la littérature et des études comparatives peuvent être trouvées dans [Pic04] et [BaySM09]. De manière générale, ces approches peuvent être divisées en deux catégories : la première utilisant un seul modèle de fond et l'autre utilisant plusieurs modèles de fond [LiaCC08] [PorIH08]. Une taxonomie des méthodes de détection par soustraction de fond pourra être trouvée dans [BaySM09].

D'autres méthodes de détection dans des vidéos ont été proposées telles que celles basées sur la segmentation en régions (Mean-Shift [CamM99]) ou celles basées sur le mouvement (flot optique [CavE00] [BroBM09], contours actifs [VeiB05] [ZheLSZ09]). Ces méthodes sont principalement utilisées pour effectuer la détection d'individus ou d'objets « en mouvement » et relève également du suivi de cible. Elles seront brièvement présentées dans la section suivante dédiée au suivi.

Des méthodes de détection basées sur un apprentissage supervisé ont également été développées. Les approches d'apprentissage les plus utilisées sont : les réseaux de neurones [RowBA98], le boosting adaptatif [VioJS03], les séparateurs à vaste marge (Support Vector Machines) [PapOP98] et les arbres de décision [GreK95]. Les méthodes de détection par apprentissage supervisé requièrent généralement un grand nombre d'exemples d'apprentissage pour chaque classe « d'objets ». De plus, l'ensemble d'apprentissage devra être manuellement rassemblé et étiqueté. Par conséquent, ces méthodes, assez performantes, sont beaucoup plus laborieuses à mettre en œuvre que les méthodes de soustraction de fond. La soustraction de fond reste l'approche la plus utilisée pour la détection et ceci est dû notamment au fait que, en plus d'être plus rapide et plus simple, elle exploite certaines techniques permettant de modéliser aisément les changements d'éclairage, le bruit, et les mouvements périodiques de régions dans l'image de fond. Ces capacités permettent d'effectuer une détection plus précise et plus adaptée dans une grande variété de situations. Bien entendu, pour qu'une telle méthode puisse convenablement fonctionner, il est nécessaire :

- soit de disposer d'une correspondance entre la position et l'orientation de la caméra avec l'image de référence associée (auquel cas les objets d'intérêt se manifestent par les modifications qu'ils entraînent sur cette image de référence). Une caméra fixe n'est alors qu'un cas particulier.
- soit d'avoir la certitude que les mouvements propres de la caméra entraînent des modifications dotées d'une dynamique moins rapide que celle issue de l'apparition dans le champ de potentiels objets d'intérêt.

Ainsi, une fois qu'un individu d'intérêt a été détecté dans la scène, la procédure de suivi est déclenchée.

1.8.2 Le suivi

L'objectif du suivi d'une cible est de déterminer sa position de manière continue et fiable tout au long du flux vidéo. Dans les systèmes de vidéosurveillance, le suivi (notamment de multiples cibles) en temps réel représente une étape primordiale pour des applications d'analyse d'activités et de compréhension d'événements. En effet, le processus de suivi d'un individu cible permettra de générer son chemin parcouru (appelé « trajectoire ») qui représente un des moyens les plus utilisés pour décrire l'activité d'un personne. Pour des applications de surveillance, le système doit généralement suivre de manière continue tous les éléments impliqués dans la scène, et cela même lorsqu'ils sont partiellement occultés par d'autres objets ou interagissent avec. En outre, il est indispensable de pouvoir suivre le comportement des individus, qu'ils soient en interaction ou qu'ils agissent de façon individuelle et séparée.

Dans certains cas, l'information spatiale en 3D d'une cible suivie peut être nécessaire, requérant ainsi l'utilisation de plusieurs caméras. La coopération entre les caméras devient alors essentielle dans un réseau multicapteur. Ceci a été exploité dans plusieurs applications [MonMK09]. Dans de tels systèmes, des techniques comme la fusion des données, l'étalonnage des caméras [VelW05] et le traitement multitâche sont indispensables [ColLK00] [MonMK09].

Diverses stratégies de suivi de cibles ont été développées. La **Figure 1.9** présente la taxonomie de ces méthodes. On peut voir qu'on peut les regrouper en trois grandes familles :

- les méthodes de suivi basées point, où les objets détectés dans des images consécutives sont représentés par des points, et l'association de ces points est basée sur l'état précédent de l'objet en terme de position et de mouvement. Cette approche nécessite l'addition d'un mécanisme de détection afin de détecter les objets dans chaque image.
- celles basées noyau, où le terme « noyau » se réfère à la forme et à l'apparence de l'objet. Par exemple, le noyau peut être un motif rectangulaire ou une forme elliptique avec un histogramme associé. Les objets sont suivis en calculant le mouvement du noyau dans des images consécutives.
- celles basées silhouette, qui effectuent le suivi d'un objet en estimant sa région dans chaque image puis utilisent les informations extraites de cette région pour le suivi. Les informations peuvent être sous forme de densité d'apparence ou de modèle de forme qui sont généralement représentés par des cartes de contours. Pour un modèle d'objet, la silhouette est suivie soit par un appariement de forme ou par une évolution des contours.

Parmi les algorithmes de suivi, les plus populaires sont le filtre de Kalman, l'algorithme de condensation (les filtres particulaires), le Mean-Shift, le flot optique et le suivi de contour. Ci-après, nous trouverons une brève présentation de chacun de ces algorithmes.

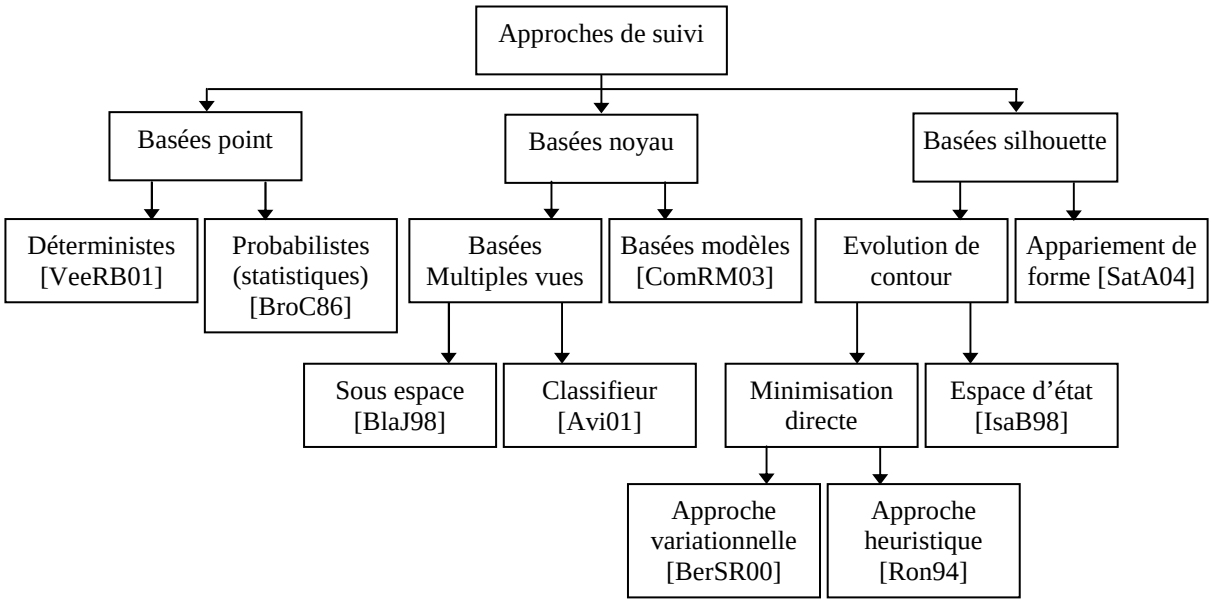


Figure 1.9 : Taxonomie des approches de suivi d'objets.

- **Filtre de Kalman**

Dans le cadre du suivi de cibles dans une vidéo, le filtrage de Kalman est généralement utilisé afin de prédire la position de la cible suivie étant données ses positions précédentes. Il est particulièrement employé pour réduire la zone de traitement et d'analyse dans l'image. Cet algorithme de suivi a énormément été utilisé et notamment dans [RemBGH97] [NVWB03] [HanYJ08].

- **Algorithme de Condensation - Filtres particulaires**

Une limitation du filtre de Kalman est l'hypothèse que les variables d'état possèdent une distribution gaussienne. Par conséquent, il donnera une estimation pauvre des variables qui ne suivent pas celle-ci [WelB01]. Cette limitation peut être contournée en utilisation l'algorithme de « Condensation ». L'algorithme de « Condensation », qui appartient au groupe des filtres particulaires, a été introduit par Isard et Blake [IsaB96] pour le suivi de personnes. Cet algorithme est un estimateur de l'état d'un système non linéaire Markovien soumis à des excitations aléatoires possiblement non Gaussiennes. Cet algorithme de suivi a été également beaucoup utilisé et notamment dans [LlaZ08] [WanMY08] [BouBPB09] [BlaGSJ09] [HesF09].

- **Mean-Shift**

L'algorithme Mean-Shift, au même titre que l'algorithme de condensation ou le filtre de Kalman, est très populaire et très utilisé par la communauté de vision par ordinateur. Initialement introduit par Comaniciu et Meer [CamM99], le Mean-Shift est une approche

récente conçue pour la détection et le suivi d'objets non-rigides. Chaque objet d'intérêt est modélisé par la distribution statistique de ses caractéristiques telles que la couleur, la position, la texture, *etc.* Le principe est de rechercher, dans l'image courante, le maximum local de la densité de probabilité du vecteur d'observation permettant ainsi de faire évoluer les modèles. Cet algorithme a notamment été utilisé dans [ComRM03] [XiaL08] [ZhaKR08]. La méthode du Mean-Shift est surtout efficace lorsque les cibles suivies ont une apparence bien distincte du reste de l'image (forme, couleur, texture, *etc.*).

- **Flot optique**

La méthode du flot optique est également très utilisée pour la détection et le suivi de cibles [CavE00] [BroBM09]. Cette approche fait une approximation du mouvement des objets suivis par l'estimation des vecteurs origines des pixels dans des séquences d'images, de sorte à représenter le champ de vitesse qui déforme une image dans un autre espace de haute dimension. Ainsi, la technique du calcul du flot optique est une méthode différentielle où l'on cherche à estimer, en tout point, le vecteur déplacement comme étant fonction du changement d'intensité du pixel ainsi que de son voisinage [HorS81]. Cet algorithme a notamment été utilisé dans [ShiKKL05] [LeiY09] [BroBM09]. Ces méthodes peuvent détecter avec précision le mouvement dans la direction de l'intensité du gradient, mais le mouvement qui est tangent à l'intensité du gradient ne peut pas être bien représenté. En outre, les méthodes de flot optique sont sensibles au changement d'éclairage.

- **Contours Actifs**

Le suivi de contours (aussi appelés contours actifs ou snakes) a été introduit par Kass et al. [KasWT88]. La principale hypothèse de cette méthode est que le contour de l'objet suivi présente des caractéristiques particulières (contraste de texture ou discontinuité de mouvement), ce qui sert à recaler la position du contour d'une image à l'image suivante en faisant converger celui-ci vers les points ayant les mêmes caractéristiques [KasWT88] [XinLS04]. Aussi, les contours actifs sont définis par une courbe (contour) continue, fermée ou non, à extrémité fixe ou non. L'objectif est alors, à partir d'une position d'initialisation située près de la cible, de déformer les contours actifs afin que ceux-ci s'ajustent aux contours de la cible tout au long de ses déplacements. Cet algorithme a notamment été utilisé dans [ShiT94] [YilXS04] [VeiB05] [ZheLSZ09]. Ces méthodes peuvent suivre des cibles en mouvement avec des tailles et de formes différentes, et se veulent insensibles aux changements d'éclairage [ToyKBM99]. Cependant, elles gèrent assez mal des déplacements rapides des cibles et sont en outre coûteuses en calcul.

Comme nous l'avons évoqué précédemment, le processus de suivi robuste d'un individu cible au long d'une séquence d'images permet de générer sa trajectoire qui représente une source d'information riche pour décrire et interpréter son activité, notamment au sein d'un réseau de caméras.

1.8.3 Analyse de trajectoire

L'analyse de trajectoire est un des moyens les plus utilisés dans les systèmes de surveillance pour décrire l'activité d'une personne. La trajectoire que suit un individu représente une source d'information riche, pouvant être extraite à partir d'une seule ou de plusieurs caméras. Morris et al. [MorT08] ont récemment proposé un état de l'art des techniques de modélisation et de classification de trajectoires dans des applications de vidéosurveillance. La

représentation de la trajectoire permet de décrire le mouvement d'un individu dans l'espace surveillé et peut être considérée comme une caractéristique déterminante qui aide à la discrimination entre des activités « normales » et « anormales » (dans un sens dépendant de l'application envisagée). Aussi, les caractéristiques de la trajectoire sont extraites puis comparées à des modèles prédéfinis ou appris. D'autres caractéristiques telles que la vitesse et l'accélération des mouvements des individus suivis peuvent être ajoutées pour améliorer l'analyse. Les algorithmes utilisés pour décrire et comparer les trajectoires peuvent être divisés en deux principales catégories: l'appariement statistique et l'appariement de vecteurs [MorT08].

- L'appariement statistique se réfère à des méthodes qui traitent conjointement un ensemble de trajectoires afin d'estimer la distribution des paramètres pertinents (par exemple, localisation spatiale, direction du déplacement, vitesse, *etc.*). De cette manière, des événements similaires peuvent être caractérisés par des distributions comparables, tandis que les comportements inhabituels peuvent être isolés. Johnson et al. [JohH95] ont développé une séquence de vecteurs de flot pour représenter la trajectoire de la cible suivi. Une estimation de la répartition spatiale statistique de ces vecteurs est réalisée en appliquant une quantification vectorielle. Deux réseaux de neurones sont utilisés pour d'abord identifier la séquence de vecteurs qui représente le mieux une trajectoire cible et, successivement, pour construire des clusters de déplacements similaires. Une version améliorée de cet algorithme est présentée dans [MecP05], où les auteurs proposent un système entièrement autonome capable d'apprendre les déplacements « anormaux ». Leur méthode permet une configuration automatique des prototypes de trajectoire. Chaque prototype est supposé avoir une distribution gaussienne, et la détection des anomalies est effectuée en faisant une comparaison statistique du prototype entrant. Les techniques de clustering peuvent également être incluses dans cette catégorie d'approches. Les travaux présentés dans [Por04] décrivent une stratégie de mesure de distance de trajectoire utilisant un modèle de Markov caché (HMM) pour le clustering. Le HMM est utilisé pour déterminer une mesure de similarité entre les trajectoires, dont l'évolution et les propriétés dynamiques sont calculées à partir d'une matrice de transition d'état. Plus récemment, l'approche dans [AnjC07] propose un algorithme de clustering non supervisé utilisant le Mean-Shift pour détecter des groupes. Puis, une procédure de fusion est exécutée dans laquelle les régions détectées adjacentes sont regroupées et les trajectoires des valeurs aberrantes sont détectées et supprimées.

- La deuxième catégorie de méthodes d'analyse de trajectoires repose sur l'appariement dit vectoriel. L'objectif est d'identifier les similitudes entre les trajectoires en se basant sur la distance entre les vecteurs caractéristiques associés [JunJS04]. Le principe de ces méthodes est de faire correspondre à chaque trajectoire un ensemble de fonctionnalités, et d'appliquer successivement une métrique (Euler, Minkovsky, distance de Hausdorff) afin de déterminer la mesure de similarité. Parmi les techniques les plus intéressantes, [CheOO04] propose un système d'extraction de trajectoire utilisant une représentation symbolique appelée chaîne de modèle de mouvement. Cette méthode fait une approximation de la trajectoire réelle et utilisent des symboles spécifiques afin de caractériser les modèles de mouvement. Zheng et al. présentent dans [ZheFZ05] un autre système qui compare les séquences vidéo en fonction de la similitude des trajectoires des objets en mouvement. De même que pour [CheOO04], les auteurs proposent une méthode hybride pour saisir des caractéristiques géométriques et une description sémantique de chaque trajectoire, la comparaison est ensuite effectuée grâce à un appariement de chaînes.

La principale difficulté que doivent gérer les méthodes de suivi de cibles et d'analyse de trajectoires concerne les occultations [WuN07] [WanHY09]. En effet, une occultation survenant lors du processus de suivi d'un individu cible conduira à la rupture de ce processus ainsi que de sa trajectoire, rompant ainsi toute analyse de son comportement ou de son activité. Les occultations qui peuvent survenir lors du suivi d'un individu peuvent être classées en trois catégories. La première est celle où l'occultation est provoquée par des éléments de la scène (objets appartenant au « fond »), par exemple une table, une chaise, un pilier, *etc.* La deuxième catégorie est celle où l'occultation est provoquée par l'intersection entre deux (ou plusieurs) individus suivis. La dernière catégorie est celle où l'individu suivi sort complètement du champ de la caméra, pour ensuite réapparaitre plus tard (soit après quelques secondes ou plusieurs minutes ou même plusieurs heures). On parlera alors de « disparition » et « réémergence ». Les occultations dans les deux premières catégories peuvent être partielles ou totales. Celles de la troisième sont bien entendu toujours totales.

L'approche la plus commune pour gérer les occultations des deux premières catégories, qu'elles soient totales ou partielles, est de décrire le mouvement de la cible par des modèles dynamiques (linéaires ou non-linéaires), puis, quand une occultation se produit, de continuer à estimer (prédire) la position de la cible jusqu'à sa réapparition complète. A titre d'exemple, les auteurs dans [BeyK99] construisent un modèle linéaire de vélocité et utilisent un filtre de Kalman afin d'estimer la position et le mouvement des individus. Dans [IsaM01], les auteurs utilisent un modèle dynamique non-linéaire et un filtrage particulaire pour l'estimation d'état (état constitue par des caractéristiques de position). Des méthodes utilisent également d'autres caractéristiques pour gérer les occultations, comme par exemple des projections de silhouette dans [HarHD00] pour localiser la tête de l'individu pendant une occultation partielle, ou le flot optique dans [DocT01], supposant que deux individus suivis se déplacent dans deux directions opposées. D'autres méthodes utilisent des modèles de forme construits au préalable [CreKS002] ou en ligne [YilXS04] afin de reconstituer les contours des parties occultées.

Bien que ces méthodes soient plus au moins efficaces pour gérer des occultations partielles ou des occultations totales de très courte durée (lorsque le mouvement de l'individu suivi reste continu avant, pendant et après occultation), elles sont totalement inefficaces lorsque l'individu disparait totalement du champ de la caméra pour ne réapparaitre qu'un certain temps plus tard, sans aucune connaissance de ses mouvements. Dans ces cas, les processus de suivi et d'analyse de trajectoire sont complètement rompus. Par la suite, lorsqu'il reparaitra dans la scène, cet individu sera détecté et suivi mais en tant que « nouvelle cible ». Aussi, sa trajectoire ainsi que d'autres paramètres seront recalculés, mais aucune correspondance ne sera faite avec les analyses effectuées (position, chemin parcouru, temps de présence, activité, *etc.*) avant qu'il ne « disparaisse » des champs des caméras. Or, à ce stade, le problème de l'analyse comportementale nécessite que les individus soient suivis sur une fenêtre temporelle suffisamment large. En effet, les caractéristiques fréquentielles (disparition et réémergence) du mouvement humain au sein de l'environnement observé imposent une observation sur une longue durée. En l'occurrence, la structure des réseaux de caméras utilisés conduisent nécessairement à des problèmes de rupture du suivi. Par conséquent, il importe de définir une méthode de suivi robuste capable de s'accommoder de telles contraintes opérationnelles. Dans ce contexte, la reconnaissance de personnes peut constituer une solution efficace. En effet, pouvoir identifier une même personne sur différentes séquences d'images acquises à des moments ou à endroits différents peut permettre de faire le lien entre l'ensemble de ses déplacements et de ses actions, et donc de permettre une analyse globale de son activité. A titre d'exemple, un individu se trouvant dans une station de métro (ou se déplaçant au sein la station) durant plusieurs heures (sans ou avec interruptions) indiquerait plus qu'il s'agit plus

d'un pickpocket que d'un simple voyageur. Car le comportement « normal » d'un voyageur serait de se trouver là juste le temps d'attendre le métro. Aussi, pouvoir suivre la position d'un individu durant une longue période de temps dans la même journée et de faire le lien entre l'ensemble de ses déplacements dans une scène ou au sein d'un réseau de caméras, et cela même lorsqu'il disparait temporairement du champ des caméras, permettrait une analyse plus pertinente de son comportement.

1.8.4 La reconnaissance (identification)

La reconnaissance de personnes est le processus qui consiste à identifier un individu cible parmi un ensemble d'individus. Dans les systèmes actuels, les principales approches de reconnaissance de personnes utilisées sont la reconnaissance de visage [Zha03], la reconnaissance de la démarche [NixC04] et la reconnaissance basée sur l'apparence [NakPHP03]. Les deux premières catégories d'approches font partie des méthodes de reconnaissance biométrique. Ces dernières ont l'avantage d'être non invasives et d'utiliser des caractéristiques uniques à chaque personne. Les méthodes basées sur l'apparence font référence à l'apparence extérieure des personnes, et utilisent généralement des caractéristiques décrivant leur forme et leur taille ou leurs vêtements telles que la couleur et la texture. Ces méthodes ont également l'avantage d'être non invasives et sont particulièrement adaptées au contexte de la vidéosurveillance du point de vue des modalités de prise de vue des séquences d'images, ce qui n'est pas toujours les cas des méthodes biométriques qui imposent généralement des contraintes en termes de résolution et de conditions de prises de vue. Ayant comme motivation de disposer de la capacité de suivre des activités humaines au travers d'un réseau multicaméra et en dépit des périodes hors du champ d'acquisition, le recours aux méthodes de reconnaissance est indispensable. Ainsi, les méthodes de reconnaissance de personnes feront l'objet d'un chapitre spécifique (cf chapitre 3).

1.8.5 La reconnaissance d'actions

L'objectif ici est de reconnaitre les actions des individus cibles dans des séquences vidéo afin de permettre une description sémantique ultérieure, de sorte que le système puisse comprendre le comportement et l'activité de chaque personne. Des états de l'art concernant les travaux de recherche sur la reconnaissance d'actions peuvent être trouvés dans [HuTWM04], [MoeHK06] et [Pop07]. Les approches existantes peuvent être divisées en trois catégories : basée modèle humain, basée état-espace et basée espace-temps.

- Dans les approches basées sur un modèle humain, les chercheurs construisent des modèles cinématiques et d'apparence et proposent de reconnaître les actions d'une personne en analysant les éléments de son corps. Certaines approches suivent les méthodes du système W^4 [HarHD00] pour détecter les silhouettes et les parties du corps. Le système W^4 est principalement conçu pour la détection et le suivi de multiples personnes et l'analyse de leurs actions. Ce système utilise l'analyse de forme pour localiser les personnes ainsi que leurs membres (tête, mains, buste, pieds) et crée des modèles d'apparence pour gérer les occultations partielles lorsque leurs trajectoires se croisent. Davis et al. [DavT02] ont proposé un système qui permet de reconnaitre l'action « marcher » en utilisant le système W^4 [HarHD00]. Plusieurs parties du corps sont détectées à partir de la silhouette puis quatre propriétés de mouvement sont extraites en se basant sur la position des pieds. Wang et al. [WanTNH03] ont calculé un contour moyen pour représenter l'information de silhouette statique. Quatorze parties rigides du corps sont utilisées pour construire un modèle dynamique, chaque partie étant représentée par un cône. Un filtre particulaire et

un classifieur du k-plus-proches-voisins sont ensuite appliqués pour la classification d'actions.

- L'approche basée état-espace est une autre approche populaire pour la reconnaissance d'actions, où une action est modélisée comme étant un ensemble d'états et de connexions dans l'espace d'état en utilisant un réseau probabiliste dynamique. Dans cette catégorie, le modèle caché de Markov (Hidden Markov Model (HMM)) est le plus couramment utilisé et présente des avantages dans la modélisation des caractéristiques variables dans le temps. Xiang et al. [XiaG06] ont développé méthode basée sur un modèle caché de Markov (appelé DML-HMM) pour la modélisation des activités humaines. Le nombre de processus temporels dans le DML-HMM est le même que le nombre d'actions détectées dans la scène. Ahmad et al. [AhmL08] ont traité le problème de la reconnaissance d'actions en utilisant une combinaison de flots de forme et de flots de mouvement. En se basant sur des caractéristiques combinées, un ensemble de HMM multidimensionnels ont été construits pour représenter chaque action à partir de vues multiples.

- L'approche basée espace–temps devient de plus en plus populaire du fait qu'elle ne nécessite pas de procédure de segmentation ou de suivi. Gorelick et al. [GorBSIB07] ont proposé une méthode de reconnaissance d'actions basée sur la corrélation de patchs spatio-temporels. De petits volumes spatio-temporels de référence sont corrélés avec les volumes cibles de la séquence vidéo. La valeur du maximum de corrélation global indique les actions les plus ressemblantes. Les méthodes d'extraction de points d'intérêt spatio-temporels ont été introduites pour reconnaître les actions humaines, avec moins de complexité de calcul que les méthodes précédentes. Cependant, ces méthodes ne permettent de détecter qu'un petit nombre de points d'intérêt stables, qui peuvent ne pas être suffisants pour caractériser des événements complexes.

1.9 Conclusion

Nous avons présenté dans ce chapitre une vue d'ensemble de ce qu'est la vidéosurveillance intelligente. Nous avons également pu constater que le marché de la vidéosurveillance était en plein essor en raison d'un contexte international instable marqué par des actes terroristes et, en réponse aux besoins de sécurité des personnes, une politique gouvernementale incitative pour le développement et le déploiement de systèmes de vidéosurveillance. Par ailleurs, cet essor est également facilité par des progrès technologiques très rapides et par un ensemble d'applications très variées. Ainsi, nous avons pu découvrir l'étendue des usages de ces systèmes ainsi que les différentes manières dont ils pouvaient être exploités pour l'analyse de scène et le déclenchement automatique d'alertes. Nous avons également évoqué l'évolution des recherches sur le développement de tels dispositifs. En effet, les nouvelles avancées technologiques réalisées permettent aujourd'hui une bonne intégration d'un système de vidéosurveillance au sein d'un système plus global de sécurité. Outre le passage à l'ère numérique ainsi qu'une capacité accrue de stockage des images, le développement d'algorithmes d'analyse vidéo puissants permet notamment une surveillance plus intelligente ainsi qu'un pilotage et une surveillance à distance. Nous avons par ailleurs pu mettre en évidence la structure générique de la chaine de traitements d'un système de vidéosurveillance intelligente dédié à l'analyse de scène et l'analyse d'activité d'individus dans un lieu accueillant du public, et dans une installation multicaméra. Nous avons ainsi décrit les différents blocs de traitements constituant un tel système. Nous avons alors identifié une difficulté majeure inhérente à une telle application. En effet, l'occultation totale d'un individu cible causée par sa « disparition » des champs des caméras rompt complètement les processus

de suivi, d'analyse de trajectoire et ainsi de l'analyse de l'activité et du comportement de cet individu. Nous avons alors proposé une piste exploitant la reconnaissance de personnes comme une solution efficace à ce problème.

Dans le cadre de cette thèse, nous avons développé une approche de reconnaissance de personnes basée sur l'apparence. Notre méthode se base sur la reconnaissance de la partie supérieure et de la partie inférieure de chaque individu cible séparément puis de la combinaison de ces deux parties, chacune d'entre elles étant décrite par un ensemble de descripteurs couleur et texture. Nous utilisons pour cela une stratégie de classification basée sur des « classifieurs à vaste marge » (« SVM »). Nous présentons dans le chapitre 3 les méthodes actuelles de reconnaissance de personnes, puis les détails de notre approche seront présentés dans le chapitre 4. Nos méthodes de reconnaissance de personnes ont été développées pour des environnements intérieurs avec utilisation de caméras fixes et fond de scène statique et invariant. Aussi, nous avons opté pour l'utilisation d'une méthode de soustraction de fond avec modèle de fond adaptatif pour segmenter les personnes avant de tenter de les reconnaitre. Pour faciliter cette tâche de segmentation, nous avons également développé une approche de détection de la peau afin de trouver des visages sur les blobs issus de la phase de soustraction de fond afin de nous permettre de nous assurer que ces blobs correspondent bien à des humains. Les détails de cette méthode seront présentés au chapitre suivant.

CHAPITRE 2

2. DETECTION ET SEGMENTATION D'INDIVIDUS

2.1 Introduction

Comme nous l'avons mentionné dans le chapitre précédent sur la détection (section 1.8.1), il est nécessaire, avant d'entamer toute opération de suivi et d'analyse d'individus cible (mouvement, vitesse, direction, trajectoire, apparence, *etc.*), de les détecter et de les segmenter du reste de l'image. Nous présentons dans ce chapitre les étapes de traitements appliquées à nos séquences d'images pour détecter et isoler les individus présents dans une scène afin de tenter de les reconnaitre ultérieurement. La première étape consiste à isoler (segmenter) tout élément (humain ou autre) se distinguant du contenu présent de façon préalable dans l'image (ce contenu préalable correspondant à la notion de « fond »). Du fait que les séquences que nous utilisons proviennent d'une caméra fixe filmant une scène en intérieur avec un fond statique et dans des conditions supposées lentement variables (par rapport à la dynamique des entités d'intérêt), nous avons choisi d'utiliser une méthode de soustraction de fond afin d'effectuer la segmentation. Ces méthodes de soustraction de fond présentent l'intérêt d'être relativement simples et efficaces dans des environnements en intérieur pour lesquels le fond présente des caractéristiques suffisamment stables. Le résultat est alors une image dans laquelle seuls apparaissent les éléments dans l'image se distinguant du contenu du fond (éléments dits « de premier-plan »), tout le reste de l'image étant alors supprimé (car correspondant à des « éléments d'arrière-plan », présents aux mêmes emplacements dans l'image de fond). Même si cette méthode est efficace, des erreurs de segmentation peuvent subsister. Les principales causes de ces erreurs sont les zones d'ombres qui correspondent aux régions de l'image dont l'éclairage (et donc l'intensité lumineuse) a été modifiée. Ce changement d'intensité pourrait alors amener l'opération de segmentation à classer cette zone comme étant un nouvel élément ou un élément manquant du fond. Pour cela, des méthodes de soustraction d'ombres ont été développées [CucGPP01] [Jun09]. Pour contourner ce problème, nous proposons de rajouter à l'étape de soustraction de fond, une étape dite de « soustraction du contour du fond ». Nous avons privilégié l'utilisation de cette technique à celles proposées dans la littérature en raison de sa simplicité et son adéquation à notre contexte applicatif. Le résultat de cette étape sera alors d'isoler dans une zone rectangulaire (dite « zone de contour ») l'élément ayant provoqué un changement de contour dans l'image. La fusion de ces deux premières étapes consistera alors à supprimer de chaque blob issu de la soustraction de fond, tous les pixels de ce blob qui ne sont pas à l'intérieur de la zone de ces contours. Le résultat de ces deux étapes sera alors l'image dans laquelle seuls apparaissent les blobs correspondant aux nouveaux éléments (avec certaines zones d'ombre en moins).

Une fois les blobs détectés, la prochaine étape est de s'assurer que les éléments auxquels ils correspondent sont bien des humains et non pas des objets. Dans un premier temps, nous utilisons notre connaissance des modalités de prise de vue pour poser des conditions sur la forme (distribution spatiale des pixels formant le blob) et l'orientation que doit avoir un blob qui correspond à un humain dans les conditions d'observation associées au type d'applications envisagées. Une fois ce filtrage effectué, il faut trouver dans les blobs restants une caractéristique qui permette de savoir de manière univoque que chaque blob correspond bien à une personne.

Pour cela, l'une des caractéristiques utilisées dans la littérature est la présence d'un visage. En effet, détecter un visage dans l'image conforte bien la présence d'une personne. Afin de détecter le visage, l'une des méthodes usitées est la détection de la peau du visage. Nous utilisons donc cette technique pour nous assurer que les blobs segmentés correspondent bien à des personnes. Pour la détection de la peau dans une image, l'une des approches couramment utilisées est de transformer l'image dans un espace couleur particulier puis d'utiliser des seuils fixes pour classer les pixels de l'image en « pixels-peau » et « pixels-non-peau ». Cependant, il arrive que des changements d'éclairage se produisent sur les visages au cours d'une séquence d'images. Outre la source d'éclairage, la position, la posture du visage et son orientation par rapport à la source d'éclairage créent des zones éclairées et des zones d'ombre qui peuvent changer l'apparence du visage. Dans ces cas, les seuils fixes prédéfinis ne sont plus adéquats et ne définissent plus les limites de la classe peau. Notre contribution a consisté alors à proposer une méthode de classification qui permet d'adapter en ligne les contours de la classe peau [HamBBL09]. Une fois le visage détecté sur chaque blob, ce blob sera alors confirmé comme correspondant à une personne.

Comme l'illustre la **Figure 2.1**, l'approche que nous proposons pour la détection des « blobs d'intérêt » censés correspondre à des personnes comporte une première étape de soustraction de fond, suivie par une sélection des blobs candidats sur la base de contraintes géométriques portant sur leur forme. Ces derniers sont finalement sélectionnés grâce à l'application d'une méthode de détection de visage, s'appuyant elle-même sur une technique de détection de pixel d'intérêt (pixel de type « peau »). La suite de ce chapitre se propose de détailler l'ensemble de ces étapes.

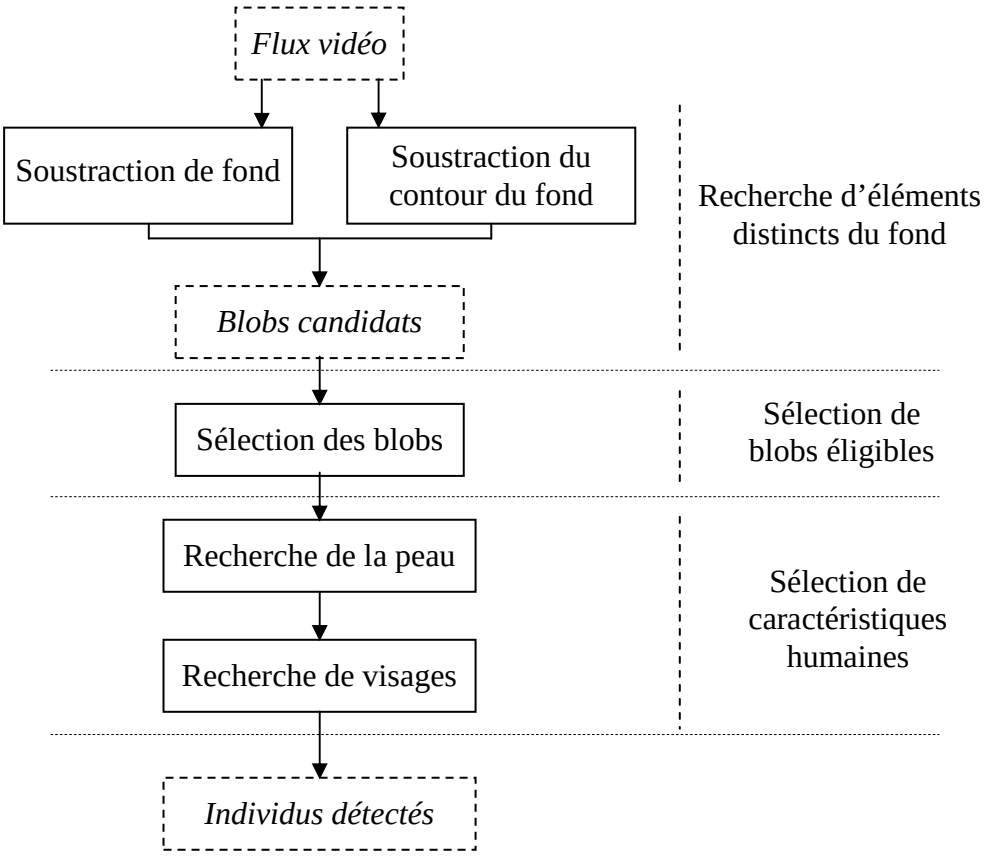


Figure 2.1 : Procédure de traitements pour la détection de personnes.

2.2 Détection de personnes par soustraction de fond

2.2.1 Introduction

L'étape de détection est indispensable afin de détecter et segmenter les individus présents dans la scène avant de tenter de les caractériser et de les suivre. Diverses méthodes ont été proposées dans la littérature afin d'effectuer cette tâche (voir à ce titre dans le chapitre 1, section 1.8.1). Comme mentionné précédemment, dans des applications où l'on utilise des caméras fixes avec fond statique invariant, l'approche de détection la plus commune est la soustraction de fond [BaySM09]. Du fait que l'environnement de notre cadre applicatif (reconnaissance de personnes en intérieur avec utilisation de caméras fixes) corresponde à ces conditions, nous avons opté pour ce type de méthode pour détecter et segmenter les individus.

La soustraction de fond est une technique qui consiste à estimer une représentation appropriée de la scène appelée « modèle de fond », puis à chercher tout changement ou déviation par rapport au modèle dans chaque image d'entrée traitée. Les régions de l'image d'entrée où il y aura des changements significatifs par rapport au modèle de fond correspondront aux éléments d'intérêt (éléments en mouvement ou nouveaux éléments au premier-plan). Les pixels constituant ces régions sont marqués pour des traitements ultérieurs. Usuellement, des étapes de post-traitement telles qu'une procédure de regroupement en composantes connexes et des traitements de morphologie mathématique sont appliquées afin d'éliminer le bruit (dû à une soustraction de fond imparfaite) et de constituer les régions (blobs) correspondant aux éléments. La **Figure 2.2** illustre le schéma du processus de détection par soustraction de fond.

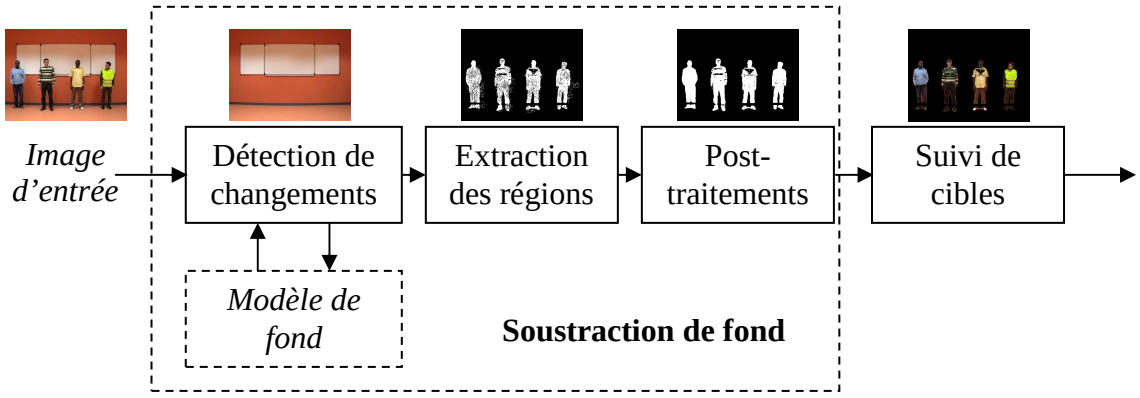


Figure 2.2 : Schéma de détection par soustraction de fond.

Les méthodes de soustraction de fond sont très efficaces lorsque le fond de la scène est statique et reste invariant ou légèrement changeant (pas ou peu de changement d'éclairage, pas d'objets déplacés ou mouvements très lents, *etc.*) mais s'avèrent inopérantes autrement. Aussi, pour une meilleure tolérance aux variations légères du fond, des méthodes de soustraction de fond avec modèle adaptatif ont été proposées. [MatYZ05] [Por07]. Dans ces méthodes, le modèle de fond est continuellement mis à jour à l'aide des nouvelles images de la séquence traitée. Elles permettent d'adapter le modèle de fond en incorporant de nouveaux éléments stationnaires apparus dans la scène et qui demeurent statiques durant une période suffisamment longue.

La méthode que nous avons choisie d'utiliser est inspirée de celle proposée par François et Medionai [FraM99] et reprise par Seitner et Hanbury [SeiH06]. La différence apportée par notre méthode concerne le fait que celle-ci travaille directement dans l'espace RGB et non pas l'espace couleur HSV. Ces auteurs ont justifié leur utilisation de l'espace HSV par le fait que cet espace sépare la composante intensité (V) des composantes chromatiques (H et S). Cependant, lorsque la valeur saturation S d'un pixel est au dessous d'un certain seuil, ce pixel est considéré comme « achromatique », et ni sa composante saturation S ni sa composante teinte H ne sont prises en compte, et ils se basent alors uniquement sur la composante intensité V pour décider si ce pixel appartient à un nouvel élément ou à l'arrière-plan. Par ailleurs, lors de l'enregistrement de nos images de test, nous avons constaté que beaucoup de pixels dans les images tombaient dans le cas achromatique. Nous avons également constaté que l'utilisation du simple espace RGB donnait, dans nos conditions opératoires, des résultats de segmentation satisfaisants. De plus, la transformation des images de l'espace RGB (qui est le format d'acquisition de nos images) vers l'espace HSV rajoute des traitements et donc un léger coût en temps de calcul. Aussi, avons nous choisi de travailler dans l'espace RGB.

Notre méthode de détection utilise un modèle de fond adaptatif pour détecter des changements dans l'image et ainsi détecter les nouveaux éléments d'intérêt survenus. Nous construisons le modèle de fond au niveau du pixel (apprentissage et classification), c'est-à-dire que chaque pixel est présumé être le résultat d'un processus indépendant. La valeur de chaque pixel du fond est modélisée par une distribution gaussienne multidimensionnelle dans l'espace RGB. Quand une nouvelle image est traitée, la valeur observée de chaque pixel est comparée à la distribution correspondante afin de décider si cette valeur correspond à l'arrière-plan ou à un nouvel élément au premier-plan. Les pixels marqués comme premier-plan sont ensuite regroupés en composantes connexes qui subissent ensuite des traitements de morphologie mathématique. Le résultat est une liste de zones (blobs) dans l'image qui représentent les régions des éléments d'intérêt. A la fin du processus de détection sur chaque image, la distribution de chaque pixel d'arrière-plan est mise à jour en utilisant la dernière observation (mesure), afin de prendre en compte des changements dans l'image de fond.

La première étape de détection d'éléments de premier-plan est donc de générer le modèle de fond. La section suivante décrit le processus de génération de ce modèle.

2.2.2 Génération du modèle de fond

Une séquence d'images de la scène de fond est tout d'abord enregistrée. Cette séquence contient 200 images (au format couleur RGB) correspondant à huit secondes d'acquisition avec une cadence de 25 images par seconde. Ces 200 images représentent un échantillon représentatif pour la constitution d'un modèle gaussien au vu de la rapidité des phénomènes censés être observés par la suite (mouvements d'individus). Chaque pixel de chaque canal couleur (R, G et B) de l'image est traité séparément. Nous détaillons ici les opérations appliquées à un pixel, sachant que celles-ci seront appliquées à tous les pixels de l'image. Chaque pixel est modélisé par trois distributions gaussiennes (une distribution sur chaque canal), chacune caractérisée par sa moyenne μ et sa variance σ. La **Figure 2.3** illustre l'exemple de la distribution des couleurs d'un pixel de l'image. Il est à noter que même si les couleurs ne sont pas en réalité distribuées de façon gaussienne [SebL00], une telle approximation peut être appliquée en pratique [FraM99]. Pour un fond statique unimodal, nous pouvons présumer que les changements se produisent de manière progressive et modérée. Un pixel de fond de l'image à la position (i,j) est modélisé par trois distributions gaussiennes $[\mu_R(i,j), \sigma_R(i,j)]$, $[\mu_G(i,j), \sigma_G(i,j)]$ et $[\mu_B(i,j), \sigma_B(i,j)]$, qui correspondent à la

modélisation de ses trois composantes couleurs, respectivement rouge (R), verte (G) et bleue (B). Initialement, la moyenne de chaque pixel sur chaque canal est initialisée à la première trame de la séquence d'images par les valeurs respectives de chaque canal pour ce pixel. La valeur de la variance est toujours maintenue au-dessus d'une valeur minimale σ_{min} pour tolérer le bruit dans l'image. Après initialisation, deux tâches sont continuellement exécutées : Premièrement, une image de détection est générée en comparant l'image actuelle à l'image de référence (modèle du fond). Deuxièmement, la distribution du modèle du fond est mise à jour en utilisant des informations issues de l'image actuelle. Ces deux étapes seront détaillées dans les deux paragraphes suivants.

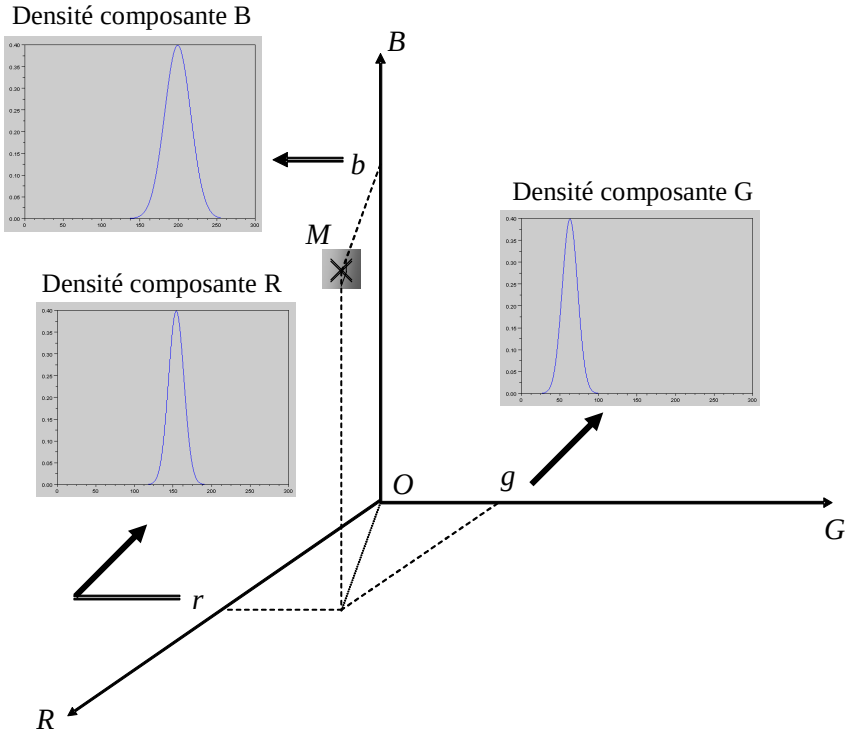


Figure 2.3 : Caractérisation de la couleur d'un pixel (*M*) de l'image à l'aide de distributions gaussienne dans l'espace de représentation couleur (ici RGB).

2.2.3 Génération de l'image de détection

L'image de détection est l'image qui résulte de la comparaison de l'image courante avec le modèle de fond. Elle contient donc les éléments de différence entre ces deux images. Aussi, chaque pixel de l'image doit être classé comme faisant partie de l'arrière plan ou d'un élément de premier plan. Les pixels classés dans la première catégorie auront la valeur 0 (noir) dans une image binaire de détection, et les pixels de la deuxième catégorie auront la valeur 1 (blanc) (cf. **Figure 2.4** (i), (j) et (k)).

Le système décide si un pixel occupant la position (*i,j*) dans l'image et ayant pour valeurs (pour ses trois composantes couleurs) $I(i,j) = \left[R(i,j) \;\; G(i,j) \;\; B(i,j)\right]^T$ appartient à l'arrière-plan ou à un élément de premier-plan en calculant la distance de type Manhatan (L^1), pour chaque canal pris séparément, entre ses valeurs et les moyennes des distributions correspondantes pour cette même position, soit $\mu(i,j) = \left[\mu_R(i,j) \;\; \mu_G(i,j) \;\; \mu_B(i,j)\right]^T$ dans le modèle de fond. Nous devons donc calculer $d_R(i,j) = \left|R(i,j) - \mu_R(i,j)\right|$,

$d_G(i,j) = |G(i,j) - \mu_G(i,j)|$ et $d_B(i,j) = |B(i,j) - \mu_B(i,j)|$. Si, pour ce pixel, la distance pour un des canaux est supérieure à un seuil $\lambda_{Canal}(i,j)$ avec $Canal \in \{R,G,B\}$, alors ce pixel est marqué comme pixel de premier-plan (valeur 1), sinon il sera marqué comme pixel d'arrière-plan (valeur 0). Ce seuil $\lambda_{Canal}(i,j)$ est déterminé en fonction de la variance de la distribution gaussienne pré-calculée : $\lambda_{Canal}(i,j) = 2\sigma_{Canal}(i,j)$

La valeur 2σ donne 95% de l'intervalle de confiance pour une véritable distribution gaussienne [SebL00]. Cependant, comme nous l'avons évoqué, la distribution des informations couleur n'est pas rigoureusement gaussienne [SebL00], aussi il convient d'être plus pessimiste quant au fait que le pixel se situe dans l'intervalle dans l'intervalle $[\mu - 2\cdot\sigma, \mu + 2\cdot\sigma]$, ce qui conduit à réduire l'intervalle de confiance à 75 % (valeur résultant de l'application du théorème d'iniquité de Tchebychev [SebL00]).

Pour résumer, un pixel à la position (*i*,*j*) admettant pour composantes couleur $I(i,j) = [R(i,j) \quad G(i,j) \quad B(i,j)]^T$ est marqué (classé) comme pixel de premier-plan si :

$$|R(i,j) - \mu_R(i,j)| > 2\cdot\sigma_R(i,j) \text{ ou}$$
$$|G(i,j) - \mu_G(i,j)| > 2\cdot\sigma_G(i,j) \text{ ou}$$
$$|B(i,j) - \mu_B(i,j)| > 2\cdot\sigma_B(i,j).$$

La **Figure 2.4** illustre la segmentation d'éléments de premier-plan. On voit que les pixels sont classés sur chaque canal couleur séparément. Les résultats de la classification des trois canaux sont fusionnés pour obtenir une image binaire de détection générale (**Figure 2.4** (l)). Cette image binaire subit ensuite des traitements de morphologie mathématique pour réduire le bruit (pouvant être le résultat d'une détection imparfaite) et un regroupement en composantes connexes afin d'agréger les pixels de premier-plan en blobs entiers. Après cette opération, le modèle de fond doit être mis à jour et prendre en compte d'éventuelles variations de l'arrière-plan. En effet, même si le fond est censé être invariant, de faibles variations peuvent toujours survenir (variations d'éclairage, apparition d'ombres ou de reflets, présence de nouveaux objets stationnaires, *etc.*). Aussi, une mise à jour du modèle de fond est effectuée à chaque trame.

2.2.4 Mise à jour du modèle de fond

Après que chaque pixel de l'image actuelle *k* ait été classé comme premier-plan ou arrière-plan, les distributions couleur (moyenne et variance) de tous les pixels ayant été marqués comme pixels d'arrière-plan sont mises à jour par :

$$\mu_{Canal}(i,j)_k = (1-\alpha)\cdot\mu_{Canal}(i,j)_{k-1} + \alpha\cdot I(i,j)$$
$$\sigma^2_{Canal}(i,j)_k = (1-\alpha)\cdot\sigma^2_{Canal}(i,j)_{k-1} + \alpha\cdot\left(\mu_{Canal}(i,j)_k - I(i,j)\right)^2$$

Ici, α représente le taux d'apprentissage, sous forme d'un facteur d'oubli qui limite la modélisation à un certain nombre d'images récentes.

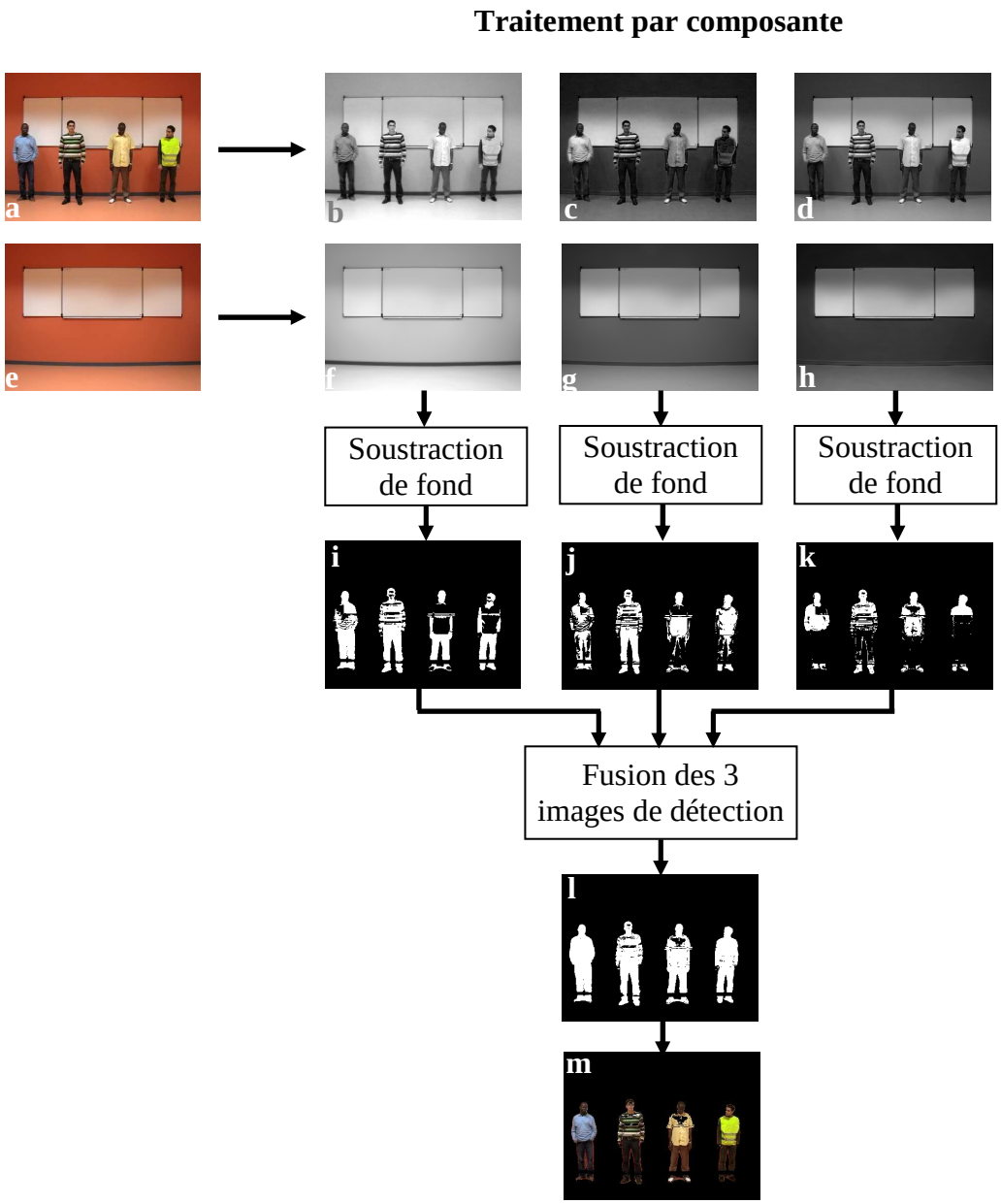


Figure 2.4 : Segmentation d'éléments de premier-plan. (a) et (e) : image d'entrée et modèle de fond (RGB) respectivement. (b), (c) et (d) : images d'entrée des canaux R, G et B respectivement. (f), (g) et (h) : les images du fond (R, G et B). (i), (j) et (k) : les images segmentées (R, G et B). (l) : résultat de la segmentation. (m) : les éléments de premier plan détectés.

La soustraction de fond fonctionne correctement dans les conditions que nous avons évoquées (caméra fixe et fond quasi-statique avec éclairage quasi-invariant). Cependant, des erreurs de segmentation peuvent subsister. Comme évoqué en introduction, elles peuvent correspondre à des ombres (sur le sol ou sur les murs) ou des changements dans le fond tels que des reflets. Ces zones peuvent alors être détectées comme des éléments de premier-plan. Les principales causes de ces erreurs restent cependant les zones d'ombres. A des fins d'illustration, la **Figure 2.5** présente un exemple de segmentation par soustraction de fond. L'individu a bien été segmenté mais on voit que son ombre sur le sol a également été segmentée avec lui. Pour pallier ce problème, nous proposons de rajouter à l'étape de soustraction de fond, une étape de

soustraction du contour du fond. Cette opération est destinée à détecter uniquement les zones de l'image où il y a des changements au niveau des contours. La fusion du résultat de ce processus avec celui de la soustraction de fond permettra alors d'éliminer les pixels détectés comme pixels de premier-plan engendrés par des ombres. Les détails de la segmentation par soustraction du contour du fond sont présentés ci-après.

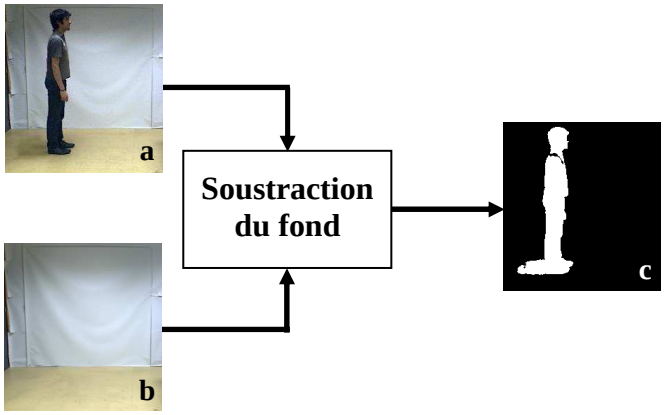


Figure 2.5 : Résultat de la segmentation par soustraction de fond, avec présence de zone d'ombre. (a) Image courante. (b) Image du fond. (c) Segmentation de l'objet.

2.2.5 Soustraction du contour du fond

La soustraction du contour du fond commence par calculer l'image des contours de l'image du fond et l'image des contours de l'image d'entrée et faire la différence de ces deux images. Le filtre utilisé pour calculer les contours est le filtre de Sobel. Le filtrage de Sobel consiste à calculer le gradient de l'intensité de chaque pixel, indiquant ainsi la direction de la plus forte variation d'intensité ainsi que le taux de changement dans cette direction. Cet algorithme a été choisi car il obtient de bons résultats tout en étant très simple. Le résultat de la soustraction sera alors une image dans laquelle il y aura uniquement les contours des nouveaux éléments, les autres contours (en commun) seront supprimés par la soustraction. Même si les ombres présentent des changements d'intensité dans certaines zones de l'image, elles présentent peu ou pas de contours dans les conditions d'éclairage diffus qui sont les nôtres. Par conséquent, ces zones ne seront alors pas détectées par la soustraction du contour du fond. La finalité de cette étape sera alors d'isoler chaque élément ayant provoqué un changement de contour dans l'image dans une zone rectangulaire (boite englobante).

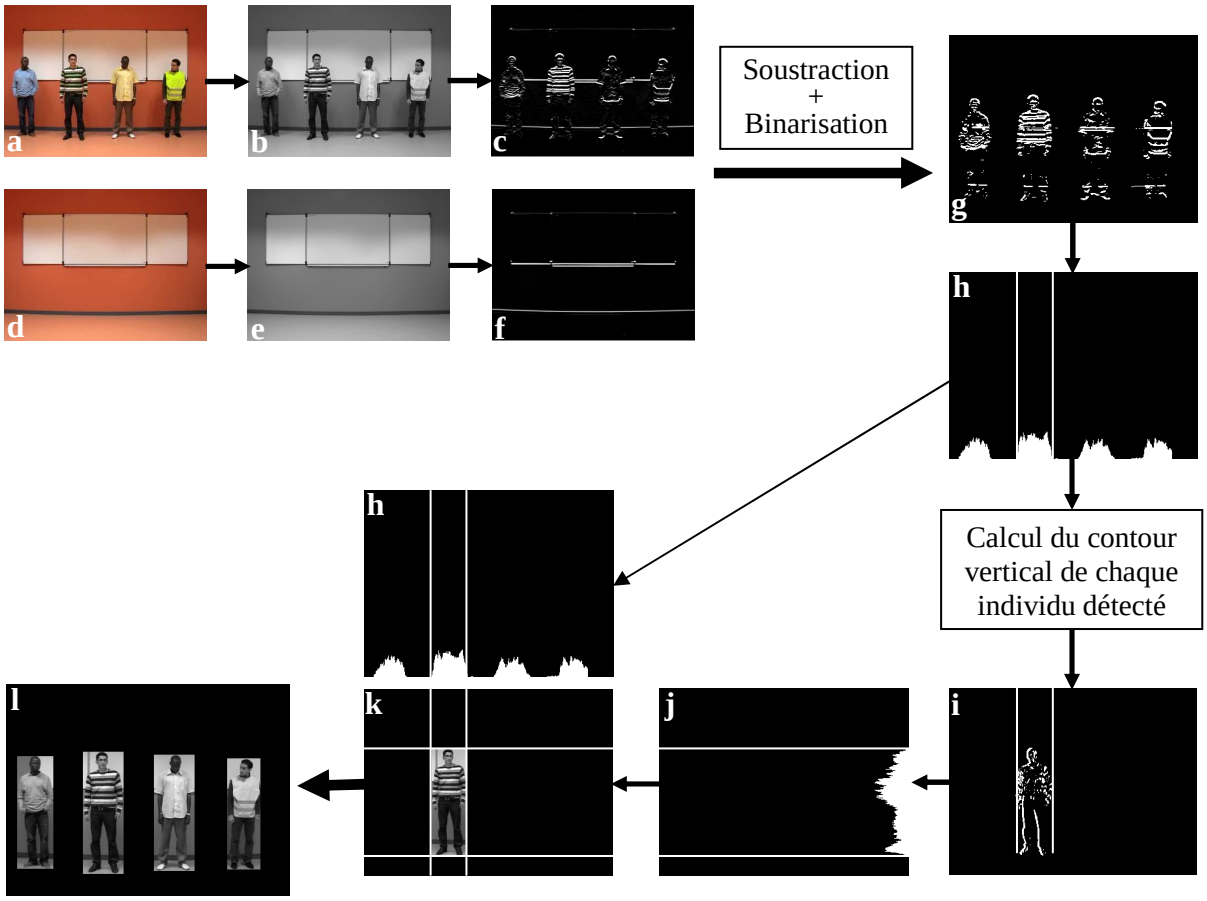


Figure 2.6 : Détection des individus par la soustraction du contour du fond. (a) et (b) Images d'entrée en RGB et en niveaux de gris respectivement. (d) et (e) image de fond en RGB et en niveaux de gris. (c) et (f) Images des contours horizontaux de l'image d'entrée et de l'image de fond. (g) Contours horizontaux des éléments au premier-plan. (h) Profil vertical de l'image (g). (i) Un des éléments détectés isolé. (j) Contours vertical de l'image (k) Zone de l'élément détecté encadrée. (l) Résultat final ; tous les éléments au premier-plan sont détectés et encadrés

Ce processus commence donc par la transformation de l'image courante et l'image du fond en couleur en images en niveaux de gris. Ces images ensuite passent par un filtrage passe bas. Cette opération de filtrage permet de lisser les images et de diminuer les faibles contours (notamment les contours d'éventuelles ombres qui sont généralement assez faibles) et donc de ne garder que les contours les plus importants. Cela permet d'être plus robuste aux fluctuations du calcul des contours dues notamment au bruit d'acquisition. Puis, à partir de ces deux images en niveaux de gris filtrées, on calcule les images des contours horizontaux de l'image courante (**Figure 2.6** (c)) et de l'image du fond (**Figure 2.6** (f)) respectivement. On génère ensuite une image binaire de détection sur laquelle apparaitront uniquement les pixels des contours présents dans l'image d'entrée mais qui ne sont pas présents dans les contours de l'image de fond. Cela peut s'assimiler à une opération de logique booléenne « inhibition » notée $a \cdot \bar{b}$ où $(a,b) \in \{0,1\} \times \{0,1\}$ représentent les valeurs des pixels des contours de l'image d'entrée et de l'image du fond respectivement pour une position (i, j) donnée. On note qu'on effectue au préalable une opération de dilation de l'image des contours du fond afin d'être plus robuste au bruit pouvant influer sur le calcul des contours. Ainsi, le résultat de cette opération est une image binaire sur laquelle seuls les contours des éléments au premier-plan

apparaissent (**Figure 2.6** (g)). A partir de cette image, on calcule le nombre de pixels blancs (valeur 1) sur chaque colonne, obtenant ainsi le profil vertical (**Figure 2.6** (h)). Sur ce profil, apparaissent clairement les colonnes de l'image où il y a eu des changements de contours, indiquant la présence d'éléments au premier-plan. Plusieurs modes peuvent alors être présents dans ce profil, chaque mode correspondant à un élément détecté (ici la **Figure 2.6** (h) indique la présence de quatre modes donc d'un minimum de quatre éléments au premier plan). Par la suite, on calcule les contours verticaux de l'image courante (**Figure 2.6** (i)) et on lui soustrait (par opération « inhibition ») les contours verticaux de l'image du fond (après lissage puis dilatation), obtenant ainsi uniquement les contours verticaux des éléments présents dans l'image courante. A l'aide des modes du profil vertical, les contours de chaque élément d'intérêt sont isolés. Pour chaque élément, on calcule le profil horizontal de ses contours verticaux (**Figure 2.6** (j)), pour trouver les lignes de l'image sur lesquelles l'élément est présent. Ainsi, à l'aide du profil vertical et du profil horizontal, on arrive à délimiter la zone (rectangulaire) dans laquelle chaque élément est situé (**Figure 2.6** (k)). Il est à noter ici qu'on postule que les éléments détectés sont dans une position verticale et suffisamment distants. Ainsi, les pixels détectés comme pixels de premier plan (lors de la phase de segmentation par soustraction de fond) qui ne se situent pas à l'intérieur d'un des rectangles dans lesquels se situent les éléments d'intérêt seront automatiquement mis à zéro et reclassés comme pixels d'arrière plan. La **Figure 2.7** illustre ce procédé.

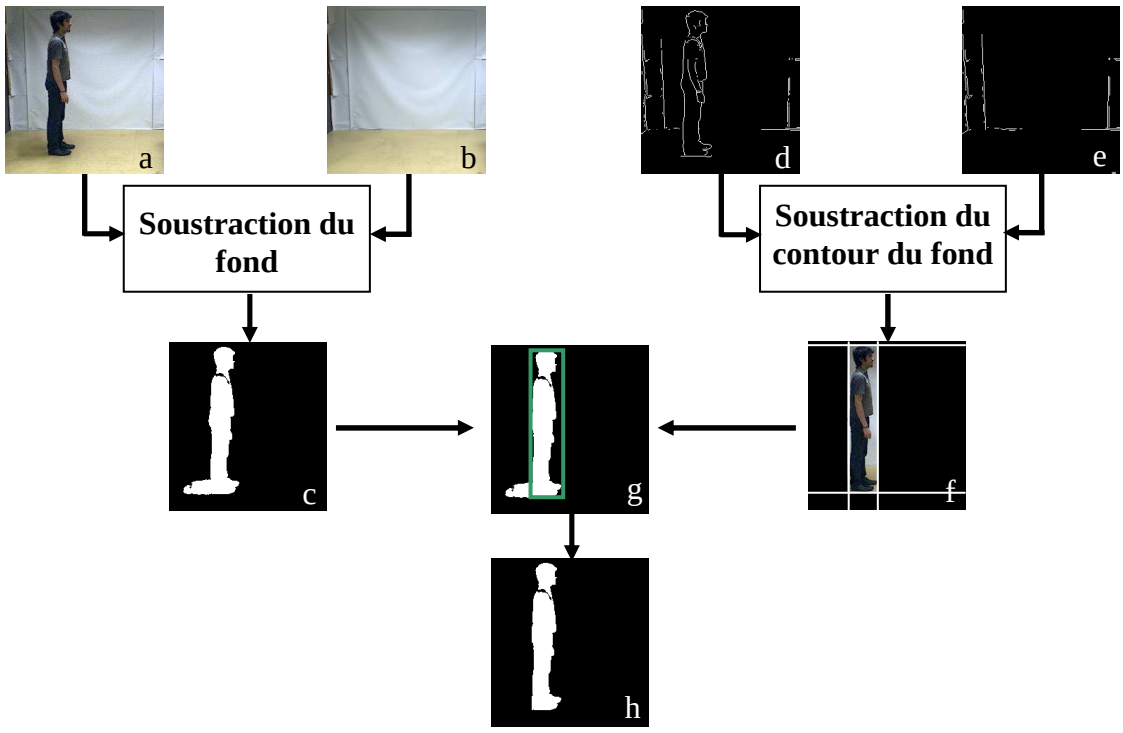


Figure 2.7 : Résultat des deux processus de segmentation (soustraction de fond + soustraction du contour du fond). (a) Image courante. (b) Image du fond. (c) Image de détection. (d) Contour de l'image courante. (e) Contour de l'image du fond. (f) Objet isolé dans le rectangle. (g) Suppression de la zone d'ombre. (h) Résultat de la détection.

Le résultat final de ces deux processus de détection (soustraction de fond + soustraction des contours du fond) est une image dans laquelle apparaissent uniquement les blobs correspondant aux éléments d'intérêt. A ce stade, la prochaine étape est de s'assurer que les éléments détectés sont bien des humains (plus précisément, il s'agit de vérifier que les blobs candidats présentent des caractéristiques spécifiques permettant de renforcer cette hypothèse de façon importante). Celles-ci se doivent d'être suffisamment discriminantes. Nous décrivons dans ce qui suit les caractéristiques choisies et les procédures relatives à leur extraction.

2.3 Détection de personnes par détection d'attributs humains

2.3.1 Introduction

Comme indiqué, à ce stade du processus de détection des personnes, l'étape suivante est de s'assurer que les blobs segmentés correspondent bien à des humains afin de lancer par la suite le processus de reconnaissance proprement dit. Dans un premier temps la prise en compte des conditions opérationnelles permet d'éliminer d'emblée certains blobs ne respectant pas un ensemble de critères simples, tels que la taille ou la forme géométrique. Nous appellerons cette procédure d'élimination la « sélection des blobs ». Pour cela, nous utilisons des informations sur les modalités de prise de vue des séquences d'images sur lesquelles sont appliqués nos différents algorithmes. Nous pouvons ainsi poser des conditions sur la taille, la forme et l'orientation que doit avoir un blob correspondant à une personne. Une fois cette sélection effectuée, il nous faut trouver dans les blobs restant une caractéristique qui nous permette d'affirmer, avec une confiance suffisante, que chaque blob correspond bien à une personne. La présence d'un visage dans le blob testé est l'un des indices les plus utilisés dans la littérature. En effet, détecter un tel élément dans l'image conforte bien la présence d'une personne (hormis le cas de la présence d'une représentation de visage dans l'image elle-même, comme ce peut être le cas si la scène comporte une affiche). Afin de détecter le visage, nous utilisons une technique de détection de la peau appliquée aux blobs segmentés. Pour la détection de la peau dans une image, l'une des approches couramment utilisées est de transformer l'image dans un espace couleur adapté puis d'utiliser des seuils fixes sur les canaux couleur pour classer les pixels de l'image en « pixels-peau » et « pixels-non-peau ». Cependant, il arrive que des changements d'éclairage se produisent sur les visages au cours d'une séquence d'images. En plus des variations qu'impose la source par ses possibles fluctuations, la position, la posture et l'orientation du visage par rapport à celle-ci créent des zones éclairées et des zones d'ombre qui peuvent changer l'apparence du faciès. Dans ces cas, les seuils fixes prédéfinis ne sont plus adéquats et ne définissent plus les limites de la classe pixels-peau. Nous proposons alors une méthode de classification qui permet d'adapter en ligne les contours de la classe pixels-peau [HamBBL09]. Une fois le visage détecté sur chaque blob, ce blob sera alors confirmé comme correspondant à une personne. Nous présentons dans ce qui suit les détails des procédures de sélections des blobs éligibles, de détection de la peau et la détection du visage.

2.3.2 Sélection des blobs candidats à une recherche de visage

Les conditions que nous allons poser sur chaque blob détecté pour le retenir ou l'éliminer concerne sa taille minimum (nombre de pixels), sa forme (distribution spatiale des pixels) et son orientation (angle entre le grand axe du blob et l'axe horizontal).

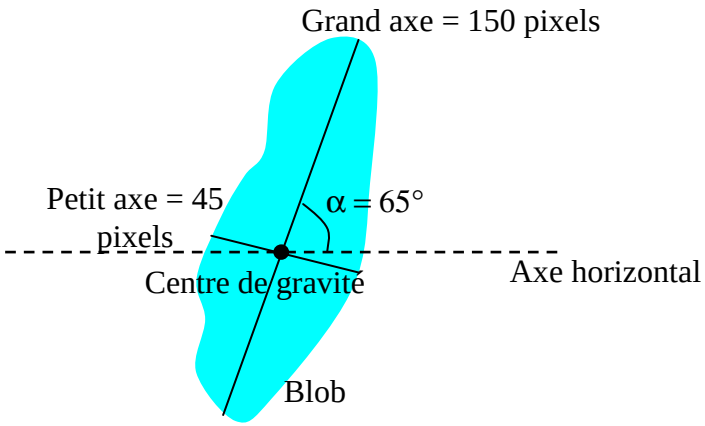


Figure 2.8 : Exemple d'un blob et de ses propriétés.

La caméra enregistrant nos séquences d'images a été montée sur un trépied à 1,5 mètres de hauteur et placée de manière à filmer l'intérieur d'une salle avec un mur en arrière plan à une distance de 5 mètres, devant lequel des individus se déplaceront. La résolution des images a été fixée à 240x320 pixels, au format RGB. Connaissant la distance entre la caméra et les emplacements où se trouvent les individus filmés ainsi que la résolution, nous pouvons savoir quelle taille devrait avoir un individu sur l'image. Sur une image de taille $N1 \times N2$ une personne située entre 4 et 5 mètres de la caméra aura entre $(1/2) \cdot N1$ et $(7/10) \cdot N1$ pixels de hauteur et entre $(1/8) \cdot N2$ et $(1/5) \cdot N2$ pixels de largeur. Un individu représentera donc entre $(1/16) \cdot N1 \cdot N2$ pixels et $(7/50) \cdot N1 \cdot N2$ pixels, ce qui équivaut à un ratio de 6% à 14% de l'image entière. Ainsi, nous décidons d'éliminer tous les blobs trop petits. On élimine alors tout blob dont la taille est inférieure à 5% de la taille de l'image (au lieu de 6% pour avoir un de marge). Ici, on élimine tout blob dont la taille est inférieure à 4000 pixels (c'est-à-dire $240 \times 320 \times 5\%$). En ce qui concerne la forme et l'orientation, nous savons que les séquences d'images sont acquises par des caméras fixes placées légèrement en hauteur dans une pièce dans laquelle des individus viendront marcher, se mettre debout et ressortir. Nous savons donc que ceux-ci seront constamment placés de manière verticale (position debout) par rapport à la caméra. Aussi, nous éliminerons tous les blobs qui ne sont pas de forme notoirement rectangulaire (grand axe / petit axe < 1,5) et dont l'orientation ne sera pas verticale (l'angle entre le grand axe et l'axe horizontal < 45° ou > 135°). La **Figure 2.8** illustre un exemple de blob qui réunit toutes les conditions citées : le rapport entre la taille de son grand axe et son petit axe est de 3,3 et l'angle entre son grand axe et l'axe horizontal est de 65°. En application de notre règle de décision, un tel blob pourra donc être retenu comme pouvant correspondre à une personne. La **Figure 2.9** présente un exemple d'image de segmentation. On peut voir que quatre blobs ont été détectés, mais qu'après filtrage par la technique décrite ici, un seul a été retenu comme pouvant correspondre à une personne. La **Figure 2.10** illustre des exemples de résultat final de segmentation.

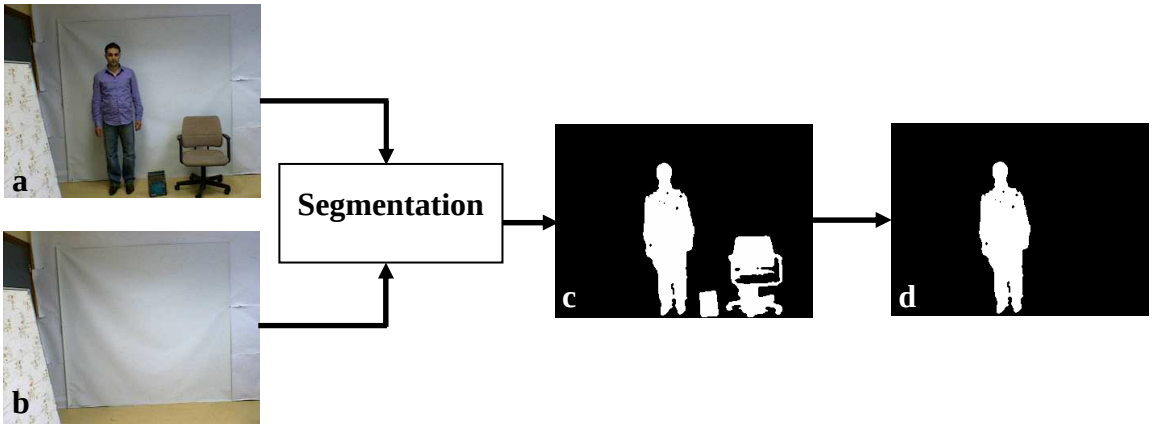


Figure 2.9 : Exemple de filtrage des blobs. (a) Image courante. (b) Image du fond. (c) Image de segmentation. (d) Image de segmentation après filtrage.

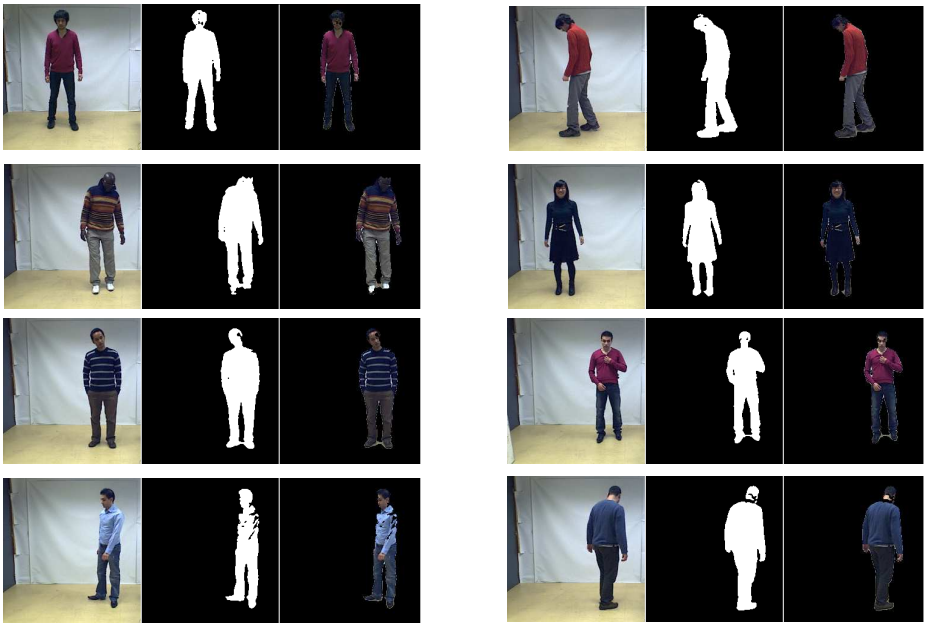


Figure 2.10 : Exemple d'images de segmentation.

Maintenant que nous avons gardé uniquement les blobs dont la taille et les propriétés géométriques pourraient correspondre à des personnes, il convient de rechercher dans chaque blob candidat un visage, ce qui renforcerait considérablement l'hypothèse selon laquelle on peut bien lui associer une personne. La détection de visage est utilisée généralement en vidéosurveillance pour soit trouver des personnes dans l'image, soit pour une opération de reconnaissance de visage ultérieure. Nous présentons ci-après un aperçu des approches de détection de visage existantes ainsi que les difficultés liées à cette problématique. Puis, nous présenterons la méthode - basée sur la détection de la peau - que nous avons choisie d'utiliser. Nous présenterons également une méthode de classification adaptative permettant une robustesse aux variations d'éclairage du visage.

2.4 Détection de visage

2.4.1 Introduction

Comme nous l'avons précédemment précisé, nous utiliserons la détection de visage afin de nous assurer que les blobs que nous détectons dans l'image correspondent à des humains. Les systèmes de détection de visage actuels peuvent être classés selon qu'ils se basent sur le visage entier ou sur des traits caractéristiques du visage [YanKM02]. Dans la première approche, on génère une base d'exemples à partir de laquelle un classifieur va apprendre ce qu'est un visage (réseaux de neurones, machines à vecteurs supports (SVM), analyse en composantes principales, eigenfaces, *etc*.). Ces systèmes peuvent obtenir des scores de détection élevés mais requièrent de grandes bases d'apprentissage. Dans la seconde approche, on peut distinguer trois niveaux d'analyse. Au niveau le plus rudimentaire, on ne prend en compte que la couleur ou les niveaux de gris pour détecter des régions ressemblant à un visage, généralement de face. A un niveau d'analyse intermédiaire, on cherche à détecter des caractéristiques indépendantes des conditions lumineuses et de l'orientation des visages. Enfin, à un haut niveau d'analyse, on recherche des traits caractéristiques du visage tels que les yeux, les contours extérieurs, le nez et la bouche que l'on associe à des configurations (« *templates* ») connues *a priori* ou apprises. Des modèles déformables, des snakes ou des modèles de points distribués (Point Distributed Models PDM) peuvent alors être utilisés. Ces derniers modèles requièrent notamment une bonne résolution de l'image. Des exemples de détections de visages dans des images sont illustrés par la **Figure 2.11**. Quelque soit l'approche utilisée, détecter les visages dans des images grand public est une tâche de haut-niveau qui doit faire face à des difficultés liées à divers facteurs, que nous nous proposons d'expliciter.

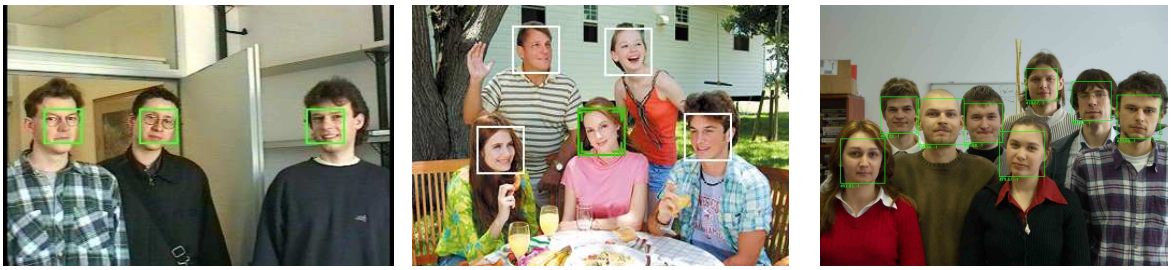


Figure 2.11 : Exemples de détection de visages dans des images.

2.4.2 Les difficultés liées à la détection de visage

Les difficultés liées à la détection de visage peuvent être attribuées aux facteurs suivants :

- **Posture.** L'image d'un visage change en raison de la position relative caméra-visage (de face, de profil ou dans une position intermédiaire), et certains attributs faciaux tels que les yeux ou le nez peuvent être partiellement ou complètement occultés.
- **Présence ou absence de composantes structurelles.** Les attributs faciaux tels que la barbe, la moustache et ou des lunettes peuvent être présents ou pas et cela avec une forte variabilité. De plus, ces attributs peuvent eux-mêmes revêtir des formes très différentes d'un individu à l'autre : géométrie, couleur, taille, *etc*.

- **Expression faciale.** L'expression faciale d'une personne affecte directement l'aspect de son visage.
- **Occultation.** Des visages peuvent être partiellement cachés par d'autres objets. Dans une image avec un groupe de personnes, certains visages peuvent partiellement ou entièrement en cacher d'autres.
- **Orientation de l'image.** Les images de visage changent directement pour différentes rotations autour de l'axe optique de la caméra.
- **Conditions de prise de vue.** Des facteurs tels que l'éclairage (distribution, orientation et intensité de la source) et les caractéristiques de la caméra (capteur, optique) affectent l'aspect d'un visage dans l'image.

Les difficultés citées ici représentent les plus gros défis pour les algorithmes de détection de visage. Les méthodes développées tendent à augmenter les performances de détection et à être assez robustes à ces dernières, c'est-à-dire à augmenter les taux de bonne détection et à diminuer les fausses détections (faux positifs et faux négatifs). Selon l'application visée (suivi d'individus, reconnaissance biométrique, *etc*.) et l'environnement d'utilisation (extérieur ou intérieur, fond de scène maitrisé ou non, éclairage variant ou invariant, *etc*.), certains écueils peuvent être directement évités. Il convient donc, avant de développer une méthode de détection de visage, de savoir pour quelle application et dans quel environnement celle-ci sera utilisée. De la sorte, son développement pourra être orienté pour pallier telle ou telle difficulté. Un autre facteur qui influence directement la manière dont un algorithme de détection de visage est développé et utilisé est lié aux contraintes associées au temps de réponse requis pour l'application (nécessité éventuelle d'une réaction en « temps réel »). Certains algorithmes peuvent être assez robustes mais lents en exécution, et d'autres peuvent être très rapides mais moins robustes. Le choix de la méthode à utiliser doit donc se faire principalement en fonction de l'environnement et du cahier des charges de l'application visée.

En ce qui nous concerne et comme nous l'avons précisé dans le cadre du paragraphe précédent, nous exploitons la détection de visage afin de « certifier » les blobs issus de notre phase de segmentation en tant qu'éléments associés à des personnes. La méthode sélectionnée à cet effet doit donc satisfaire les contraintes correspondant à l'application de suivi robuste et s'intégrer au mieux dans la chaîne de traitements associée. En particulier, l'approche mise en œuvre devra être robuste aux variations de certains paramètres tels que la posture, l'orientation et l'expression du visage. En effet, si les conditions de prise de vue peuvent être assez maîtrisées (fond fixe et éclairage invariant), d'autres, et spécialement ceux cités plus haut, restent complètement incontrôlables car à la libre initiative des personnes observées par le système. Afin de sélectionner la méthode la plus appropriée, nous présentons ci-après un bref aperçu des principaux algorithmes de détection de visage et nous concluons cette partie en précisant celui que nous avons choisi d'utiliser.

2.4.3 Les méthodes de détection de visage

Les méthodes de détection de visage peuvent être classées en quatre catégories [YanKM02], à savoir :

- les méthodes basées sur la connaissance,

- les méthodes basées sur des caractéristiques invariantes,
- les méthodes d'appariement de modèles,
- les méthodes basées sur l'apprentissage de modèles.

Dans ce qui suit, nous présentons la « philosophie » associée à chacune de ces approches.

2.4.3.1 Les méthodes basées sur la connaissance

Dans cette catégorie, les méthodes de détection de visage sont développées à partir de règles dérivées de la connaissance que l'on a des visages humains [YanH94]. En effet, il est assez aisé de proposer des règles simples pour décrire les caractéristiques d'un visage et leurs relations. Par exemple, un visage de face apparaît dans une image avec un nez, une bouche et deux yeux symétriques. Les relations entre les caractéristiques peuvent être représentées par leurs positions et leurs distances relatives. Des caractéristiques faciales dans l'image d'entrée sont d'abord extraites, puis les visages candidats sont identifiés en se basant sur des règles prédéfinies [KorP97]. Un procédé de vérification est habituellement appliqué pour réduire les fausses détections. L'un des problèmes de cette approche est la difficulté de traduire la connaissance humaine en règles bien définies. En effet, si celles-ci sont trop strictes, elles peuvent ne pas détecter les visages qui ne satisfont pas toutes les règles. A l'inverse, si elles sont trop « permissives », elles peuvent prendre à tort divers objets de la scène qui les respectent pour des visages. De plus, il est difficile d'étendre cette approche pour détecter des visages dans différentes postures puisqu'il est fastidieux d'énumérer tous les cas possibles. Quoiqu'il en soit, ces méthodes peuvent bien fonctionner sur des scènes simples et où la vue des visages est plus au moins frontale.

2.4.3.2 Les méthodes basées sur des caractéristiques invariantes

En réponse aux « points faibles » de l'approche précédente, on essaye ici de trouver des caractéristiques invariantes qui existent quels que soient la posture, le point de vue, et les conditions d'éclairage. On emploie ensuite ces dernières pour localiser les visages [YowC97]. Partant du constat que les humains peuvent sans effort détecter des visages et d'autres objets en différentes postures et conditions d'éclairage, on fait l'hypothèse qu'il doit exister des propriétés ou des caractéristiques qui sont invariantes malgré les facteurs d'influence précédemment cités. De nombreuses méthodes ont été proposées pour d'abord détecter les caractéristiques faciales qui ensuite indiquent la présence d'un visage [YowC97]. Ces dernières, telles que les sourcils, les yeux, le nez, la bouche, et les cheveux sont généralement extraites en utilisant des détecteurs de contour. En se basant sur ces éléments, un modèle statistique est construit pour décrire leurs relations et vérifier l'existence d'un visage. Cependant, du fait de la nature même de l'information de départ (une image), il demeure particulièrement difficile de s'affranchir des variations d'éclairage et des phénomènes d'occultation à l'aide des primitives précédentes. En effet, les contours de celles-ci peuvent être estompés, alors que les ombres peuvent créer de nombreux artefacts qui compliquent considérablement les tâches de ces algorithmes. Aussi, utilise-t-on des éléments moins sensibles à ces phénomènes. A ce titre, les propriétés colorimétriques de la peau (qu'on appellera ici « couleur de la peau ») ainsi que sa texture ont largement été utilisées [ForF99] [LeeK07] [Sun10]. En effet, celles-ci peuvent présenter une invariance suffisante malgré la posture du visage, son orientation ou son expression.

2.4.3.3 Les méthodes d'appariement de modèles

Dans cette catégorie de méthodes, plusieurs modèles standards de visage sont appris et stockés pour décrire ce dernier dans son ensemble ou certaines caractéristiques faciales

séparément [CraTB92] [LanTC95]. Pour une image d'entrée donnée, les valeurs de corrélation avec les modèles appris sont calculées indépendamment pour le contour du visage, les yeux, le nez et la bouche. La décision relative à la présence d'un visage dans l'image présentée est alors déterminée en se basant sur ces valeurs de corrélation. Cette approche a l'avantage d'être simple à implémenter mais est très peu robuste aux variations d'échelle, de posture et de forme, et surtout plus coûteuse à exécuter.

2.4.3.4 Les méthodes basées sur l'apprentissage de modèles

Dans cette catégorie, et contrairement aux méthodes d'appariement de modèles où ceux-ci sont prédéfinis par des experts, les modèles exploités ici sont appris à partir d'exemples d'images [SchK98] [SunP98]. En général, les méthodes basées sur l'apparence se basent sur des techniques d'analyse statistique et d'apprentissage pour trouver les caractéristiques appropriées des images de visage et de « non-visage ». Les caractéristiques apprises sont exprimées sous forme de modèles de distribution ou de fonctions discriminantes qui sont employés ensuite pour la détection. Les méthodes de cette catégorie obtiennent généralement les meilleures performances (taux de détection) mais ont l'inconvénient d'être coûteuses en calcul et assez laborieuses à mettre en œuvre.

2.4.4 Conclusion

Au vue des quatre catégories de méthodes de détection de visage et des difficultés liées à chacune d'entre elles, nous avons choisi d'utiliser une méthode basée sur l'exploitation de caractéristiques invariantes et plus particulièrement sur l'utilisation des propriétés colorimétriques du visage. Comme évoqué, cette approche a l'avantage d'être simple et intrinsèquement robuste à beaucoup des facteurs d'influence associés à notre cadre applicatif. Cependant, la difficulté principale que cette méthode doit gérer réside dans les variations d'éclairage. A des fins de précisions, nous faisons dans ce qui suit un rappel sur le principe de la détection de la peau. Nous décrivons ensuite la méthode que nous avons choisie d'utiliser. Puis, nous présentons un algorithme de classification adaptative que nous avons développé afin d'obtenir une certaine robustesse vis-à-vis des variations d'éclairage.

2.5 Détection de la peau

2.5.1 Introduction

Détecter de la peau dans une image revient à trouver les pixels et les régions de l'image correspondant à de la peau. Ce processus est généralement utilisé comme une étape de prétraitement pour trouver des régions qui pourraient correspondre à certaines parties du corps humain. Aussi, la présence de peau dans une image est une indication de la présence possible d'humains. La **Figure 2.12** illustre des exemples de détection de la peau. La plupart des travaux de recherche dans ce domaine s'appuie sur la couleur [KakMB07], et quelques approches y ajoutent également l'utilisation d'informations de texture [ForF99]. La détection des pixels de la peau est une opération relativement simple, peu coûteuse en calcul tout en étant efficace. Ces qualités ont encouragé son utilisation et ont prouvé son efficacité dans des applications telles que la détection et le suivi du visage et des mains, l'analyse de gestes et dans divers systèmes d'interaction homme-machine. Par exemple, dans l'une des premières applications, la détection de régions « peau » a été utilisée pour identifier des photos de nu sur internet pour pouvoir filtrer le contenu [FleFB96]. Dans une autre application, la détection de la peau a été utilisée sur des vidéos de journaux télévisés pour détecter la présence du présentateur (ou de la présentatrice), pour ensuite pouvoir faire un archivage automatique [AbdE99]. Dans une telle application, le visage et les mains du présentateur sont les plus grands éléments de l'image contenant des régions de couleur peau, puisque, généralement, le journal télévisé est enregistré dans un environnement contrôlé en intérieur (studio) avec un fond statique contenant peu ou pas d'objets présentant des caractéristiques colorimétriques comparables à celles de la peau. Dans un tel contexte, la démarche se montre particulièrement efficace.

Comme nous l'avons évoqué, nous avons dans nos travaux de recherche opté pour l'utilisation de la détection de la peau en nous basant sur des caractéristiques couleur et texture afin de trouver les visages présents dans l'image qui étayent l'hypothèse de la présence de personnes. La raison de ce choix est que cette approche s'affranchit de la plupart des difficultés liées à la détection de visage précédemment citées (posture, expression faciale, orientation, *etc.*). Par ailleurs, cette approche présente une relative simplicité de mise en œuvre et exhibe une vitesse d'exécution compatible avec le contexte applicatif de nos travaux.

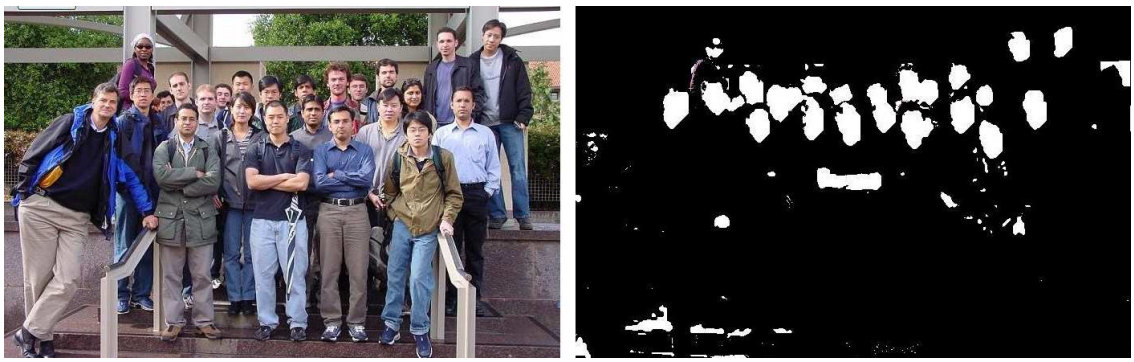


Figure 2.12 : Exemples de détection de la peau sur des images.

Même si le principe d'une telle approche est assez simple, certaines difficultés qu'elle doit gérer méritent d'être rappelées. En effet, l'apparence de la peau dans une image est très

influencée par le facteur source d'éclairage (intensité, direction, teinte, température). Les êtres humains ont la faculté de pouvoir identifier la couleur des objets dans une large gamme d'éclairage, qui est appelé « la constance couleur ». Aussi, un défi important dans la détection de la peau est de pouvoir caractériser les couleurs en cherchant des éléments invariants ou, du moins, peu sensibles aux variations d'éclairage. Le choix de l'espace couleur dans lequel sont représentées les images affecte considérablement les performances d'un détecteur de peau ainsi que sa sensibilité à cette grandeur perturbatrice. Un autre point crucial est que de nombreux objets dans le monde réel peuvent avoir des couleurs proches ou identiques à la « couleur peau » (le cuir, le bois, le sable, certains vêtements, *etc.*). La présence de tels objets dans la scène peut donc provoquer de fausses détections. Nous présentons dans ce qui suit les étapes d'un processus de détection de la peau prenant en compte ces considérations.

2.5.2 Les étapes d'un processus de détection de la peau

La détection de la peau est un processus comportant deux phases : une phase d'apprentissage et une phase de classification.
La première phase, concernant l'apprentissage, est effectuée en trois étapes :

1. Choisir un espace couleur approprié.
2. Collecter une base de données d'imagettes (« patches ») de peau à partir d'exemples d'images de visage. Une telle base de données contient généralement des patches de peau à partir d'une grande variété de personnes et dans différentes conditions d'éclairage.
3. Apprendre les paramètres du classifieur de peau.

Après l'apprentissage, la phase de classification des pixels de la peau dans une image donnée implique :

1. De convertir l'image dans le même espace couleur que celui qui a été utilisé dans la phase d'apprentissage.
2. De classifier chaque pixel en pixel-peau ou pixel-non-peau, en utilisant le classifieur entraîné.
3. Puis, généralement, d'appliquer des post-traitements sur les régions détectées comme de la peau en utilisant la morphologie mathématique, et ceci en imposant des critères géométriques et spatiaux à ces régions.

Un pixel donné est classé comme « pixel-peau » ou « pixel-non-peau » à partir du modèle de peau construit grâce aux images d'apprentissage dans un espace couleur particulier. En effet, en fonction de ce dernier, la couleur de la peau occupe une région plus ou moins compacte de cet espace (le « cluster » de couleur peau). Aussi, le choix de l'espace couleur est d'une grande importance et doit se faire de manière à ce que le cluster soit le plus compacte possible, permettant une représentation aisée avec le maximum d'exactitude. Le classifieur sera d'autant moins enclin aux erreurs (faux positifs ou faux négatifs) que cette condition sera satisfaite. Les faux positifs sont les pixels non-peau de l'image que le classifieur classifie comme de la peau. Les faux négatifs sont les pixels-peau de l'image que le classifieur classifie comme non-peau. Dans la section suivante, nous présentons brièvement les principaux espaces couleur utilisés dans la détection de la peau.

2.5.3 La détection de la peau et les espaces couleur

Il a été mis en évidence par Forsyth et Fleck [FleFB96] que la couleur de la peau des êtres humains est peu saturée et présente une gamme restreinte de couleurs, du fait qu'elle est

constituée par une combinaison de sang (rouge) et de mélanine (brun, jaune). Par conséquent, la distribution correspondante n'est pas aléatoire dans un espace couleur donné, mais est regroupée dans une zone de cet espace. Aussi, les chercheurs ont essayé d'utiliser un espace dans lequel la couleur peau resterait idéalement invariante aux conditions d'éclairage. On peut rassembler les espaces couleur utilisés pour la détection de la peau en quatre grandes familles. La première est celle de l'espace couleur RGB. La deuxième et troisième famille sont respectivement les espaces couleur TV et les espace couleur perceptuels. Enfin la quatrième famille est celle des espaces couleur dits « colorimétriques ». Un autre espace, conçu spécialement pour la détection de la peau, a été introduit par Forsyth et Fleck [ForF99]. Cette espace a la particularité d'inclure l'information de texture.

2.5.3.1 L'espace couleur RGB

L'espace couleur RGB est l'espace colorimétrique le plus couramment utilisé en imagerie numérique. Il permet d'encoder les couleurs comme une combinaison additive de trois couleurs primitives: le rouge (R), le vert (G) et le bleu (B). L'espace RGB peut être visualisé comme un cube en 3D où R, G et B sont les trois axes perpendiculaires. Un des principaux avantages de cet espace est sa simplicité. Toutefois, il ne correspond pas à la perception humaine (les distances dans l'espace RGB ne correspondent pas linéairement à la perception des êtres humains). Par ailleurs, cet espace ne sépare pas la luminance de la chrominance, et les composantes R, G et B sont corrélées. La luminance d'un pixel donné est une combinaison linéaire des valeurs de R, G et B. Par conséquent, le changement de luminance d'un patch de peau donné affecte toutes les composantes R, G et B. En d'autres termes, l'emplacement d'un patch de peau donnée dans le cube RGB change en fonction de l'intensité de l'éclairage (luminosité) sous lequel les images de ce patch ont été acquises. Cela se traduit par un cluster peau très étendu dans le cube RGB. En dépit de ces limites fondamentales, l'espace RGB est largement utilisé dans la littérature pour la détection de la peau et cela en raison de sa simplicité [JonR02].

Afin de réduire la dépendance à l'éclairage, les composantes couleur RGB peuvent être normalisées de sorte que la somme des composantes normalisées soit égale à 1 ($r + g + b = 1$). L'espace RGB normalisé est une représentation, qui est facilement obtenue à partir des valeurs RGB par une procédure de normalisation simple: $r = R / (R + G + B)$, $g = (G / R + G + B)$ et $b = B / (R + G + B)$. Comme la somme des trois éléments normalisés est connu ($r + g + b = 1$), la troisième composante b ne contient plus d'information supplémentaire et peut être omise (car elle peut être calculée à partir des deux autre, c-à-d : $b = 1 – (r + g)$), cela permet ainsi de réduire la dimension de l'espace. On note cependant qu'on perd l'information intensité lumineuse. . Les composants r et g sont souvent appelées "couleurs pures", car la dépendance à la luminosité de r et de g est diminuée par rapport à l'espace RGB. La simplicité de la transformation (normalisation) a permis à l'espace RGB normalisé d'être populaire et largement utilisé [OliPB97] [BroCL01].

2.5.3.2 Les espaces couleur TV

Parmi les espaces couleur orthogonaux utilisées dans les transmissions TV, l'espace YCbCr est l'un des espaces les plus apprécié pour la détection de la peau [HsuAJ02] [WonLS03]. Les espaces YUV et YIQ font également partie de cette famille. L'espace YIQ est utilisé en transmission TV NTSC, alors que l'espace YCbCr est utilisé dans la compression d'image JPEG et la compression vidéo MPEG. Un des avantages d'utiliser ces espaces couleur est que la plupart des médias vidéo sont déjà encodés en les utilisant. De plus, la transformation de l'espace RGB en un de ces espaces est une transformation linéaire [BurB08]. Par exemple,

dans l'espace YCbCr, la couleur est représentée par la quantité nommée *luma* (qui est la luminance, calculée à partir de RGB [Poy95]), construit comme une somme pondérée des valeurs R, G et B, et de deux valeurs de différence couleur Cb Cr et qui sont calculées en soustrayant la valeur *luma* des composantes B et R respectivement. Aussi, ces trois espaces couleur (YUV, YIQ, et YCbCr) séparent le canal d'intensité (luminosité) (Y) des deux canaux de chrominance orthogonaux (UV, IQ et CbCr respectivement). Par conséquent, contrairement à l'espace RGB, l'emplacement de la couleur peau dans les deux canaux de chrominance ne sera pas affecté par le changement de l'intensité (luminosité) de l'éclairage de la scène. Dans les canaux de chrominance, la couleur peau constitue généralement un cluster compact. Cela facilite la construction des détecteurs de peau qui sont invariantes à la luminosité et qui utilisent des classifieurs simples. La densité de la couleur peau dans les canaux de chrominance peut être assez facilement approximée en utilisant une distribution gaussienne.

La simplicité de la transformation et la séparation explicite des composantes luminance et chrominance a rendu ces espaces couleur très intéressants pour la modélisation et la détection de la peau [ShiCT02] [HsuAJ02] [WonLS03] [ZheZW04]. Une variante de YCbCr appelée YCgCr a été également utilisée [DioG03]. Cet espace diffère de l'espace YCbCr par l'utilisation de la composante couleur Cg au lieu de la composante Cb. D'autres espaces couleur similaires de cette catégorie ont également été utilisés pour la détection de la peau, et sont YIQ [DaiN96], YUV [MarV00] et YES [GomSS02].

2.5.3.3 Les espaces couleur perceptuels

Les espaces couleur perceptuels tels que l'espace HSV (et également HSI et HSL) ont également été populaires pour la détection de la peau. L'espace HSV sépare trois composantes: la teinte H (pour *Hue*), la saturation S et la luminosité appelée également intensité V (pour *Value*). Essentiellement, ces espaces couleur sont des déformations de l'espace (cube) RGB et peuvent être calculés à partir de l'espace RGB via une transformation non linéaire. Un des avantages de ces espaces dans la détection de la peau, est qu'ils permettent aux utilisateurs de spécifier de manière intuitive la limite de la classe couleur peau en fonction de la teinte et de la saturation. La composante V donne les informations de luminosité (intensité), et est souvent utilisée pour réduire la dépendance à la valeur intensité des pixels-peau. Ces espaces ont été notamment utilisés dans [AlbTD01] et [ShiCT02].

Le passage de l'espace RGB à un des ces espaces perceptuels s'effectue par une transformation non linéaire, et plusieurs transformations possibles existent. L'espace HSV définit la couleur comme suit : la teinte (H) qui définit la couleur dominante (comme le rouge, vert, violet et jaune) d'une région. La saturation (S) qui mesure le degré de coloration d'une région proportionnellement à sa luminosité [Poy95]. L'intensité (V) qui est liée à la luminance. L'intuitivité des composantes couleur de ces espaces et la discrimination explicite entre les propriétés luminance et chrominance fait que ces espaces sont populaires dans les travaux sur la segmentation couleur peau [ZarSQ99]. Plusieurs propriétés intéressantes de la composante H ont été constatées : elle est invariante aux reliefs dans des sources de lumière blanche, et aussi, pour les surfaces mates, à la lumière ambiante et à l'orientation de surface par rapport à la source de lumière. Cependant, plusieurs points négatifs ont été mis en évidence [Poy95]. A titre d'exemple, lorsque, au niveau d'un pixel, la valeur de la composante S est trop faible, cette composante et la composante H deviennent dépourvues d'informations ''colorimétriques'' et sont donc inutilisables. Il est également à noter que la

transformation de l'espace RGB vers un des ces espaces perceptuels est relativement coûteux en calcul.

2.5.3.4 Les espaces couleur « colorimétriques »

La séparation des composantes chromaticité et luminance est également réalisée dans des espaces couleur dits « colorimétriques », tels que les espaces CIE-XYZ, CIE-xy et CIE-Lab définis par la Commission Internationale d'Eclairage (CIE). Ces derniers sont détaillés dans [BurB08]. Un des inconvénients des espaces couleur XYZ et XY est que les différences de couleur ne sont pas perçues de manière égale dans différentes régions de l'espace de représentation. Malgré tout, ces espaces couleur sont moins utilisés dans la détection de la peau que les autres espaces présentés. Cela est principalement dû au fait que la transformation à partir de l'espace RGB est plus coûteuse en calcul que vers d'autres espaces. Néanmoins, l'espace CIE-XYZ a été utilisé par Shin et al. [ShiCT02] à des fins de comparaison avec d'autres.

2.5.3.5 L'espace THS

Forsyth et Fleck [ForF99] ont proposé un espace original à trois composantes conçu spécialement pour la détection de la peau. À l'instar de l'espace perceptuel HSV, cet espace se compose des composantes *Teinte* (*Hue*) et *Saturation* mais délaisse la composante d'intensité (*Luminance*) (V) et la remplace par une composante *Texture*. Comme son nom l'indique, cette nouvelle composante fournit des informations de texture d'une image. Elle apporte donc une information de type nouveau qui peut s'avérer particulièrement pertinente pour la détection de la peau.

Une image au format THS (*Texture*, *Hue* (*Teinte*), *Saturation*) est obtenue par une transformation non linéaire de l'image originale au format RGB. Celle-ci s'opère en deux étapes. En premier lieu, à partir des valeurs R, G et B de l'image originale, on calcule les valeurs log-opponent *I*, *Rg*, and *By*. C'est dans un second temps, à l'aide de ces valeurs, que les valeurs *Texture*, *Hue*, et *Saturation* sont calculées. La conversion des valeurs RGB en valeurs log-opponent est effectuée par :

$$I = \frac{\zeta(R)+\zeta(G)+\zeta(B)}{3}$$
$$Rg = \zeta(R)-\zeta(G)$$
$$By = \zeta(B)-\frac{\zeta(R)+\zeta(G)}{2}$$

Avec : $\zeta(x)=105\times\log_{10}(x+1+n)$ (le rôle de n est explicité dans ce qui suit).

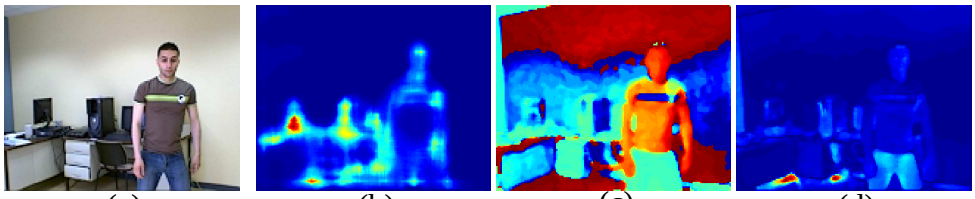
L'image *I* est constituée par la moyenne des trois composantes R, G et B. Dans la transformation logarithmique, la valeur 105 est une constante d'échelle pratique et *n* est une valeur de bruit aléatoire générée à partir d'une distribution uniforme dans l'intervalle [0,1]. Ce bruit aléatoire est ajouté pour éviter les artefacts dans les régions sombres de l'image. La transformation logarithmique contribue à rendre les valeurs *Rg* et *By* moins dépendantes de l'intensité.

Comme nous l'avons déjà évoqué, le visage humain a une texture distincte qui peut servir à le différencier des autres objets de l'image. De plus, la peau dans les images est d'ordinaire peu texturée. Aussi, une image *Texture* est calculée et utilisée pour trouver les zones de l'image de

faible texture. Pour générer l'image *Texture*, l'image *I* précédemment calculée est filtrée par un filtre médian Ψ, l'image résultante est soustraite de *I*, puis la valeur absolue de la différence est filtrée par le filtre médian Ψ. Les composantes *Teinte* et *Saturation* sont utilisées pour sélectionner les régions de l'image qui correspondent à la coloration de la peau, et sont respectivement la direction et la norme du vecteur (*Rg*, *By*). Elles sont calculées comme suit :

$$Texture = \Psi(|I - \Psi(I)|)$$
$$Teinte = \arctan^2(Rg, By)$$
$$Saturation = \sqrt{(Rg^2, By^2)}$$

La **Figure 2.13** représente l'image RGB d'entrée, et les images *Texture, Teinte* et *Saturation* (représentation en fausses couleurs).



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

Figure 2.13 : (a) Image RGB originale et les images calculées (b) *Texture,* (c) *Teinte* et (d) *Saturation.*

Comme nous venons de le présenter, divers espaces couleur ont été utilisés pour la détection de la peau, chacun ayant des avantages et des inconvénients. L'espace YCbCr et l'espace HSV sont les plus communément utilisés. L'espace THS, plus récent et spécialement conçu pour la détection de la peau, a également prouvé son efficacité [FleFB96] [ForF99] [HamBBL09] [HamBL10]. Aussi, dans l'approche que nous avons adoptée pour la détection de la peau du visage, nous utiliserons chacun de ces espaces ainsi que l'espace RGB et nous comparerons leurs résultats de détection.
Après avoir choisi, pour la tâche de détection des pixels de la peau, un espace de représentation doté d'un pouvoir discriminant suffisant, se pose encore la question de savoir sur quelle méthode de classification la sélection effective de ces derniers peut reposer. Le paragraphe suivant se propose d'apporter les éléments de réponse nécessaires.

2.6 La classification des pixels-peau

Un classifieur de « pixels-peau » peut être conçu comme un classifieur monoclasse ou deux-classes. Dans le second cas et comme le nom l'indique, les pixels de l'image seront divisés en deux classes, à savoir une classe « pixels-peau » et une classe « pixels-non-peau ». Aussi, l'ensemble de la base d'apprentissage devra contenir des données (échantillons) des deux classes. Le classifieur sera alors conçu de manière à définir une frontière de décision séparant ces deux classes. D'un autre côté, s'il est conçu comme un classifieur monoclasse, il définira alors une frontière de décision délimitant uniquement les limites de la classe peau. Les pixels situés à l'intérieur de ces limites seront classés comme pixels-peau, les autres (situés à l'extérieur) seront classés comme pixels-non-peau. L'ensemble de la base d'apprentissage n'aura à contenir ici que des échantillons de la classe peau (uniquement des exemples dits

« positifs »). Les limites (frontière de décision) de la classe peau seront alors déterminées par le classifieur lors de la phase d'apprentissage. Si ces limites ne varient pas pendant toute la phase de classification, on parlera alors de classification simple avec frontière de décision fixe. Cependant, il est toujours possible que l'emplacement de la classe peau change quelque peu et évolue dans l'espace des caractéristiques lors de la classification, et cela pour diverses raisons, dont les variations d'éclairage qui constituent un fort facteur d'influence, comme nous l'avons déjà précisé. Dans ce cas, le classifieur doit être conçu de manière à suivre ces variations en mettant à jour la frontière de décision. Dans ce cas, on parlera de données non-stationnaires et de classification adaptative avec frontière de décision dynamique. Dans ce qui suit, nous décrivons ces deux alternatives (frontières de décision fixe et frontières de décision dynamique).

2.6.1 Classification avec frontière de décision fixe

Une des méthodes de classification des pixels-peau la plus couramment utilisée et la plus simple consiste à définir explicitement les limites (frontières de décision) de la classe peau sur chaque canal de l'espace couleur utilisé [ChaN98] [KakMB07]. Dans sa forme la plus élémentaire, cette frontière peut se définir comme une sous-région dans l'espace de représentation des couleurs. En pratique, cela se traduit par l'utilisation d'intervalles (des seuils maximum et minimum) sur chaque axe de cet espace. Ainsi, considérons que C1, C2 et C3 sont les trois canaux d'un quelconque espace couleur C1C2C3. Avec des seuils empiriquement choisis $[C1_{\min}, C1_{\max}]$, $[C2_{\min}, C2_{\max}]$ et $[C3_{\min}, C3_{\max}]$, un pixel x de valeur $[x_{C1}, x_{C2}, x_{C3},]^T$ est classifié comme étant de la peau si ses valeurs x_{C1}, x_{C2} et x_{C3} sont comprises dans ces intervalles, c'est-à-dire

$$C1_{\min} < x_{C1} < C1_{\max} \quad et$$
$$C2_{\min} < x_{C2} < C2_{\max} \quad et$$
$$C3_{\min} < x_{C3} < C3_{\max}$$

La **Figure 2.14** (a) illustre un tel exemple (en 3 dimensions). On peut y voir l'emplacement des pixels-peau (les points bleus) dans l'espace caractéristiques représenté pas l'espace couleur C1C2C3. Les seuils [$C1_{min}$, $C1_{max}$], [$C2_{min}$, $C2_{max}$] et [$C3_{min}$, $C3_{max}$] qui constituent les frontières de la classe peau permettent de définir le cube. Un pixel de l'image est classé comme étant de la peau s'il est situé à l'intérieur de ce cube. Après l'étape de classification des pixels de l'image, le résultat est une image binaire où les pixels classés comme pixels-peau sont marqués. Cette image binaire sera ensuite traitée par des opérations de morphologie mathématique, afin d'éliminer les éventuels pixels isolés (assimilés à du « bruit ») et de constituer des régions de peau bien homogènes. Nous présentons dans ce qui suit l'évaluation de la détection de la peau du visage par une classification avec frontière de décision fixe.

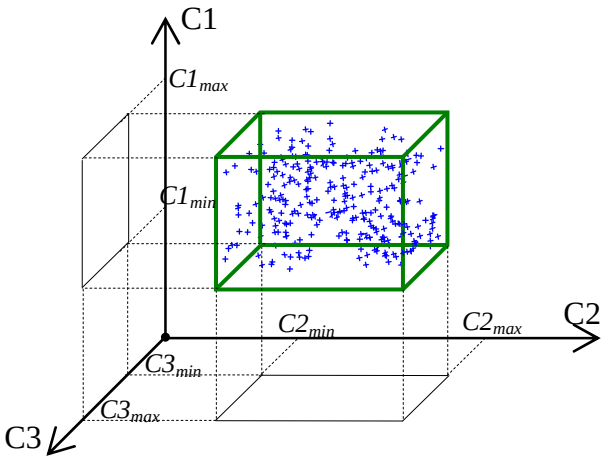


Figure 2.14 : Illustration des frontières de la classe peau dans l'espace C1C2C3.

2.6.2 Evaluation de la détection de la peau par la classification avec frontière de décision fixe

L'évaluation de notre méthode de détection de visage par détection de la peau se fera en plusieurs étapes. Dans un premier temps, nous déterminerons une zone réduite de l'image dans laquelle s'effectuera la détection de la peau. En effet, le fait qu'une détection d'éléments d'intérêt devant correspondre à des personnes en position debout ait été préalablement appliquée nous permet de poser l'hypothèse que le visage se situe dans la partie supérieure du blob détecté. Aussi, la détection de la peau s'effectuera uniquement dans une zone de l'image autour de cette partie. Puis, nous déterminerons expérimentalement les seuils fixant les frontières de la classe peau sur chacun des espaces couleur que nous avons choisi d'utiliser (RGB, YCbCr, HSV et THS). Ces seuils seront extraits sur un échantillon de notre base d'images.

2.6.3 Zone de détection de la peau

Comme indiqué précédemment, certains des pixels classés comme étant de la peau peuvent correspondre parfois à des objets ayant des couleurs proches ou identiques à la « couleur peau » (le cuir, le bois, le sable, certains vêtements, *etc.*). Aussi, afin de ne détecter que le visage, des méthodes d'extraction de caractéristiques de visage peuvent être couplées au détecteur précédent. Cette approche consiste à extraire sur les régions peau détectées des caractéristiques de visage telles que les yeux, le nez et la bouche [HsuAJ02] [SinCVS03]. Pour cela, la méthode la plus commune est de chercher dans ces régions, après une équalisation d'histogramme de l'image en niveaux de gris, des « trous » (petites zones sombres) qui correspondraient aux caractéristiques citées et ensuite leurs positions relatives [HsuAJ02]. Cette phase est utilisée afin de détecter un visage parmi les différentes zones de peau trouvées dans l'image sans aucune information *a priori*.

En ce qui nous concerne, la recherche de visage n'est effectuée que pour « certifier » les blobs candidats issus de la phase de détection de peau. Par conséquent, il n'est pas nécessaire d'appliquer celle-ci à l'intégralité de l'image mais plutôt à ces différents blobs. Qui plus est, les conditions opératoires associées à notre application permettent de dégager quelques hypothèses réalistes qui autorisent de restreindre encore davantage les zones de recherche au sein des blobs eux-mêmes. Ainsi, il nous suffira de chercher un visage uniquement dans la

partie supérieur de chacun d'entre eux, c'est-à-dire là où se trouve le visage s'il s'agit bien d'une personne. Si une région peau est détectée dans cette zone, on supposera directement qu'il s'agit d'un visage, sans avoir recours à une autre procédure de vérification. Par mesure de prudence et afin d'anticiper d'éventuelles erreurs de segmentation de personnes, nous n'effectuerons pas la détection de visage sur un blob uniquement dans sa partie supérieure, mais dans une zone autour de cette partie. Il nous faut donc définir celle-ci. La description utilisée prend la forme d'un rectangle localisé dans l'image, défini par conséquent par sa hauteur, sa largeur et sa position (précisée par les coordonnées du pixel central), et cela de manière à ce que ce rectangle de recherche soit centré sur la zone candidate du blob examiné.

Dans le détail, nous postulons, d'après les modalités de prise de vue de nos séquences d'images, que la tête d'une personne a une hauteur d'environ 1/6 de sa taille (hauteur) entière qu'on notera H. Cela signifierait que sur une personne debout, la tête se trouve entre la ligne horizontale la plus haute d'ordonnée $L1$ de son blob et la ligne horizontale d'ordonnée $L1+H/6$. Précisons à ce propos qu'on considère ici que le sens positif de l'axe des ordonnées est vers le bas avec le point d'origine de l'image est le pixel en haut à gauche, comme c'est souvent le cas lorsqu'on manipule une image numérique. Aussi, la sous-région située entre les lignes d'ordonnées respectives $L1$ et $L1+H/6$ de l'image de segmentation correspond à la tête, et aura un pixel central d'ordonnée $L1+H/12$ et d'abscisse $C1$, correspondant à la verticale partageant la sous-région en deux. La taille définitive de cette zone de recherche correspond au cinquième de la hauteur H, (ce qui correspond à un léger agrandissement par rapport à l'hypothèse préalable du sixième), ceci afin de tenir compte des éventuelles erreurs commises lors de la phase de segmentation. La **Figure 2.15** illustre le processus de détermination de la zone de détection de visage. La **Figure 2.16** illustre le processus de validation de détection de personnes.

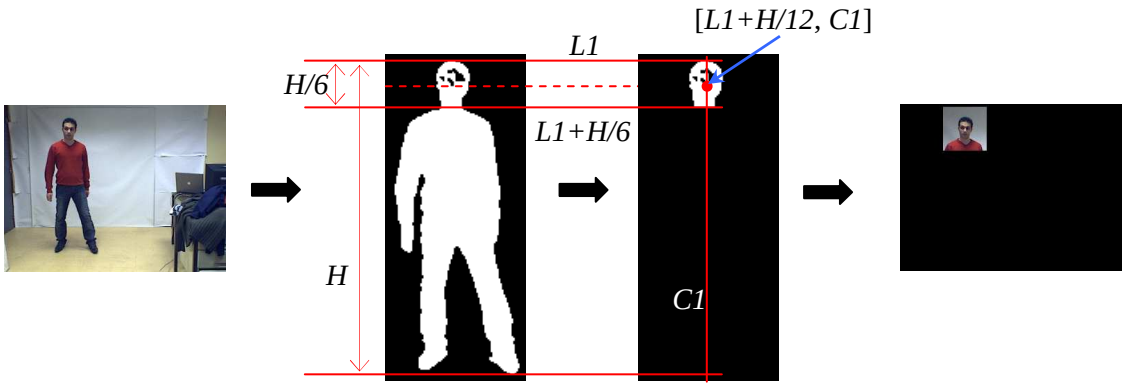
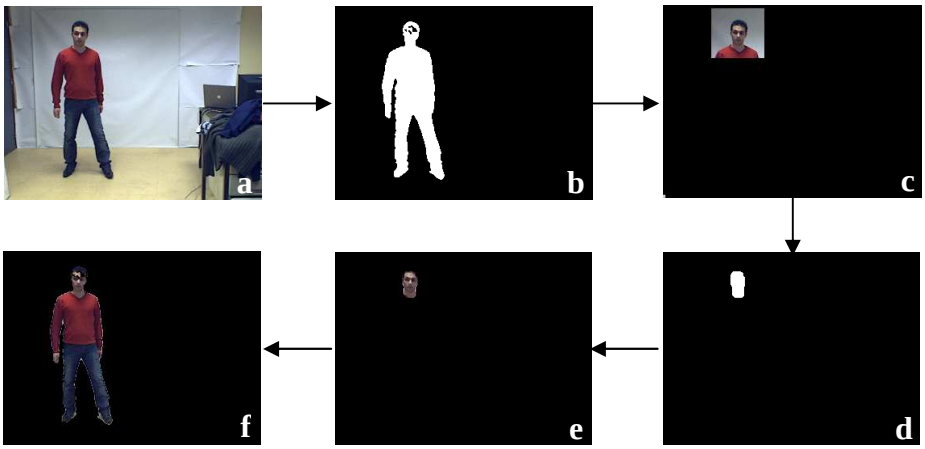


Figure 2.15 : Processus de détermination de la zone de détection de la peau du visage.



La **Figure 2.16 :** Processus de validation de détection de personnes. (a) Image d'entrée. (b) Image de détection. (c) Zone de détection de visage. (d) Détection de la peau + morphologie. (e) Visage détecté. (f) Blob validé comme étant une personne.

2.6.4 Evaluation des espaces couleur pour la détection de la peau

La base de données que nous avons construite et sur laquelle seront appliqués nos algorithmes est constituée de séquences d'images de 11 personnes, chacune d'elles portant différents vêtements dans différentes séquences, formant ainsi 54 combinaisons différentes qui seront assimilées à 54 « individus ». Chaque séquence contient 400 images (soit 8 secondes d'acquisition). Pour les besoins de cette évaluation, nous n'utiliserons ici qu'une séquence par personne pour faire l'apprentissage, ce qui fait 11 séquences au total d'où seront extraits les « patchs » de peau. Chacune de ces séquences est convertie dans les espaces RGB, YCbCr, HSV et THS. Il nous faut maintenant prélever des patches de peau de chaque personne dans chacun des quatre espaces couleur. Cette opération est réalisée manuellement, à raison de 10 images par séquence. Ces 10 images sont sélectionnées parmi celles où le visage apparait de face et est correctement éclairé (ni trop sombre, ni trop clair). Nous récupérons ainsi des échantillons de pixels de peau du visage significatifs de chaque personne dans tous les espaces couleur. On peut voir dans la **Figure 2.17** des exemples d'images (en RGB) des 11 personnes ainsi que les zones de peau extraites.

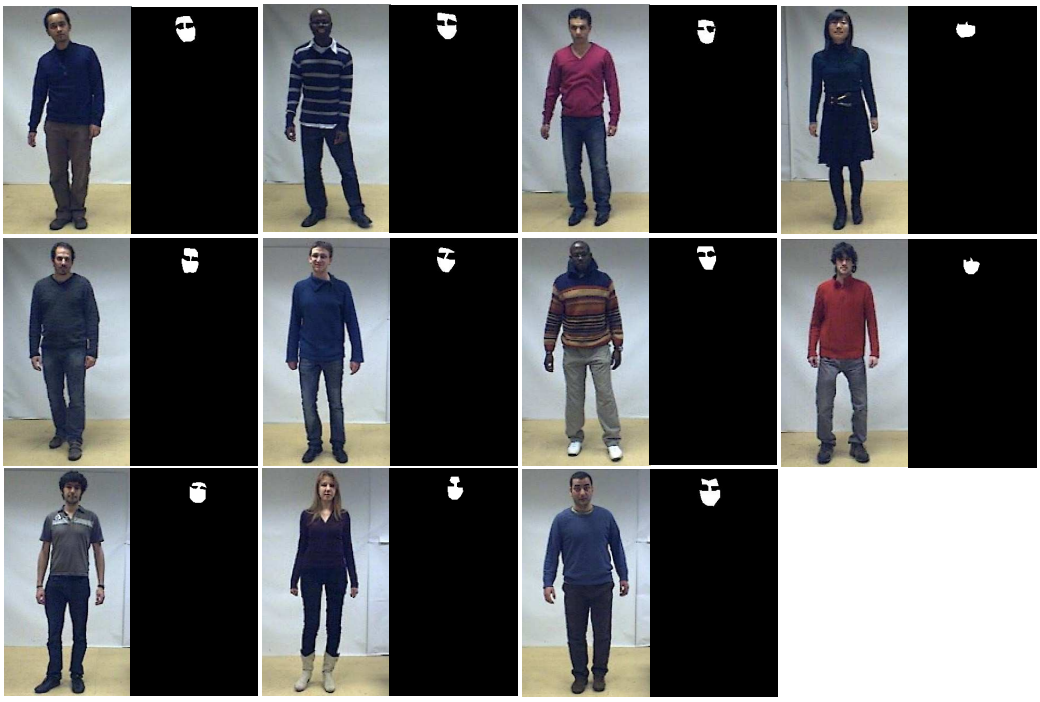


Figure 2.17 : Exemples d'images en RGB des 11 personnes que les patches de peau extraits.

A partir de ces patches, nous déterminons les seuils de valeurs des pixels-peau dans chaque espace couleur, fixant ainsi les limites de la classe peau, qui sont alors les suivantes :

- Dans l'espace RGB, un pixel $x[x_R, x_G, x_B]$ est classé comme peau si :

$$\begin{aligned} 20 < x_R < 245 \quad & et \\ 20 < x_G < 200 \quad & et \\ 20 < x_B < 180 & \end{aligned}$$

- Dans l'espace YCbCr, un pixel $x[x_Y, x_{Cb}, x_{Cr}]$ est classé comme peau si :

$$\begin{aligned} 20 < x_Y < 200 \quad & et \\ 100 < x_{Cb} < 135 \quad & et \\ 123 < x_{Cr} < 147 & \end{aligned}$$

- Dans l'espace HSV, un pixel $x[x_H, x_S, x_V]$ est classé comme peau si :

$$\begin{aligned} 0,01 < x_H < 0.97 \quad & et \\ 0,01 < x_S < 0,7 \quad & et \\ 0,04 < x_V < 200 & \end{aligned}$$

- Dans l'espace THS, un pixel $x[x_T, x_H, x_S]$ est classé comme peau si :

$$\begin{aligned} 4 < x_T < 18 \quad & et \\ 80 < x_H < 170 \quad & et \\ 6 < x_S < 20 & \end{aligned}$$

Après avoir extrait ces seuils, nous appliquons la classification en utilisant les séquences des 54 individus. Ainsi, l'évaluation se fera sur 54 séquences de 400 images chacune (ce qui fait un total de 21600 images). Avant toute chose, il est cependant indispensable de définir ce qui nous appelons une « bonne détection », à savoir :

- Plus de la moitié des pixels du visage ont été classés comme peau.
- Sur tous les pixels classés comme peau, plus de la moitié correspond aux pixels du visage.

La première condition signifie que plus de la moitié du visage doit être détectée. La seconde signifie qu'on admet des erreurs sur la classification des pixels-peau uniquement si le nombre de pixels incorrectement classés comme peau ne dépasse pas le nombre de pixels correctement classés.

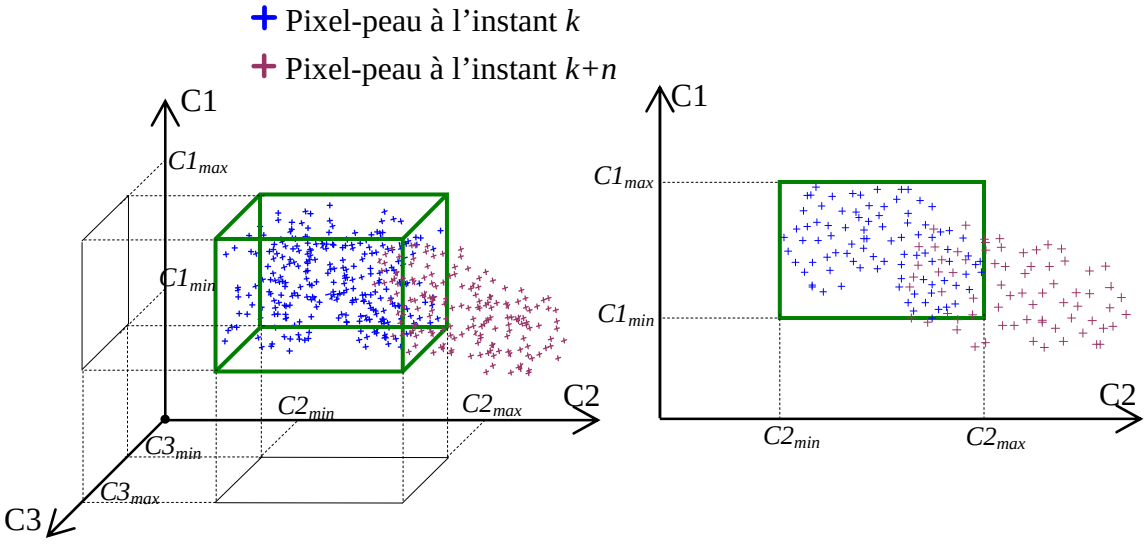
La prochaine étape maintenant est de définir quelles sont les images sur lesquelles sera appliquée la détection du visage. En effet, sur chaque séquence, la personne se déplace librement en marchant ou tournant sur elle-même. Elle peut donc se retrouver sur les images de face, de profil ou de dos. Aussi, nous prendrons en considération uniquement les images sur lesquelles le visage apparait de face ou de profil ou dans une position intermédiaire. Car bien sûr, on ne saurait évaluer une technique de détection de visage sur des images où il n'apparaît pas (en particulier lorsque les personnes apparaissent de dos). En moyenne, ces dernières représentent 20% des images de chaque séquence. Aussi, 80% des images de chacune des 54 séquences de notre base de données (sur une caméra), soit un total de 17280 images, restent donc potentiellement disponibles. Les taux (en pourcentage) de bonne détection obtenus avec les différents espaces couleurs sont présentés dans la **Tableau 2.1**. On voit clairement que le meilleur est obtenu avec l'espace THS (79%), suivi par l'espace HSV (74%) puis YCbCr (69%). Enfin, l'espace RGB obtient le taux le plus faible avec 62% de bonne détection. Les images de la première ligne de la **Figure 2.18** sont des exemples sur lesquels on obtient une bonne détection du visage. La seconde ligne fournit des exemples sur lesquelles celle-ci a échoué.

	Espaces couleur			
	RGB	YCbCr	HSV	THS
Taux de détection	62%	69%	74%	79%

Tableau 2.1 : Taux de bonnes détections du visage.

Figure 2.18 : Première ligne : images sur lesquelles il y a eu une bonne détection du visage. Seconde ligne : images sur lesquelles la détection du visage a échoué.

Nous avons constaté que la détection du visage échouait principalement lorsque les individus baissaient complètement la tête. Dans cette position, et du fait que la pièce est éclairée par des néons situés au plafond, le visage est éclairé différemment (*a priori* de façon moins intense) que dans les images où il était de position de face et à partir desquels les seuils de la classe peau ont été déterminés. Dans le cadre de nos expérimentations, il s'agit de la principale cause des non-détections. Ainsi, nous pouvons dire que la méthode de classification par seuils (limites de classe fixes) fonctionne bien (79% de bonne détection sur l'espace THS) lorsque le visage reste éclairé de la même manière tout au long de la séquence d'images, mais peut s'avérer moins efficace autrement. Mais malgré tout, comme le mettent en évidence les tests que nous avons conduits, même dans un environnement contrôlé avec un éclairage globalement invariant, des changements d'éclairage locaux issus des mouvements de la tête se manifesteront toujours de façon intermittente sur le visage. Dans de telles situations, la distribution des pixels-peau évoluera dans l'espace de représentation au fil des mouvements de la tête et donc au fil de la séquence d'images. La **Figure 2.19** illustre une telle situation. On peut voir que les pixels-peau (point bleus) à l'instant k ont changé d'emplacements à l'instant $k+n$ (points violets). Du coup, la frontière de décision définie par les seuils initialement fixés ne sera plus valable pour la suite de la séquence. Par conséquent, le modèle de décision initial de la classe peau doit être adapté. Une solution à un tel problème serait d'utiliser un classifieur capable de définir les frontières de décision et de les mettre à jour automatiquement, et ainsi d'adapter le modèle de décision dynamiquement en fonction de la non-stationnarité qui caractérise alors la classe peau. Nous nous retrouvons alors dans un contexte d'application de méthodes de classification avec frontière de décision dynamique, qui est l'objet de la section suivante.



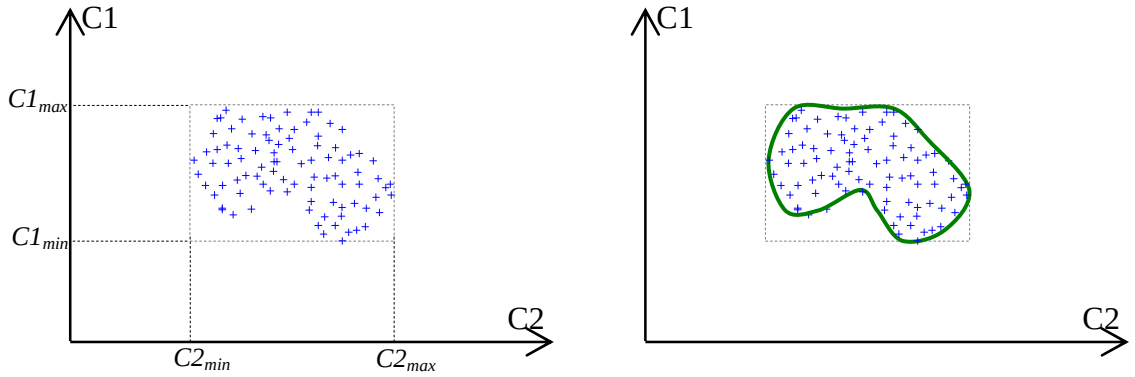
(a) (b)

Figure 2.19 : Illustration de la variation de la classe peau dans l'espace des caractéristiques. Les points bleus sont les pixels-peau à l'instant *k*, et les points violets sont les pixels-peau à l'instant *k+n*. (a) illustration en 3 dimensions et (b) illustration en 2 dimensions (à des fins de mise en évidence de la modification subie par la classe).

2.6.5 Classification avec frontière de décision ajustée et dynamique

L'objectif est ici de classifier les pixels de l'image en pixels-peau et pixels-non-peau en utilisant un modèle de décision adaptatif pouvant s'ajuster selon l'évolution de la classe peau (monoclasse). Comme évoqué précédemment, cette évolution est principalement issue des changements d'éclairage affectant le visage. Nous proposons alors d'initialiser le modèle de la classe peau en utilisant la méthode par seuil, puis de l'adapter au cours du temps. Ceci consiste à mettre à jour le modèle de classification en y incorporant de façon séquentielle les données (pixels-peau) acquises en ligne, et en y retirant les données obsolètes. Le modèle de classification est ainsi redéfini à chaque itération. Deux stratégies existent pour effectuer cette adaptation. La première, très basique, propose de recalculer complètement le modèle de classification à chaque acquisition à partir de l'ensemble des données déjà connues, auxquelles on ajoute la nouvelle. Cette technique est très gourmande en temps de calcul, même dans une version traitement par lot et ne tire aucun avantage des connaissances extraites au cours des apprentissages précédents. La seconde solution, plus en phase avec une application de suivi vidéo temps réel, consiste à utiliser les techniques d'apprentissage dotées de règles de mise à jour récursives. Grâce à ces méthodes, les informations portées par les nouvelles données sont incorporées séquentiellement dans le modèle de classification. Nous avons ainsi développé un classifieur basé sur des « séparateurs à vaste marge », plus communément appelé « machines à vecteurs supports », « Support Vector Machines » ou « SVM » [Vap95], auxquels ont été adjointes des règles d'apprentissage incrémental. Celles-ci sont capables d'adapter en ligne la frontière de décision de la classe peau au cours de son évolution dans l'espace des caractéristiques [CauP00][HamBBL09]. La raison pour laquelle nous avons choisi d'utiliser cette approche de classification est que, outre son caractère adaptatif, elle nous permet de construire le modèle monoclasse « classe peau » en exploitant uniquement des exemples positifs (pixels « peau ») lors de la phase d'apprentissage initiale. Cette démarche nous permet également d'obtenir un modèle plus fin de la classe peau. En effet, comme on peut le constater sur la **Figure 2.20** la définition de la classe peau par des seuils fixes donne un contour de la classe qui ne sera probablement qu'approximatif (en

dehors d'hypothétiques jeux de données triviaux). A l'inverse, une classification « one-class SVM » nous permet d'obtenir un contour plus fin plus adaptée à la classe car plus « resserré » autour de ses éléments dans l'espace de représentation. Les détails de cette classification SVM ainsi que des règles d'apprentissage incrémental sont présentés dans ce qui suit.



(a) (b)

Figure 2.20 : Illustration d'un contour de classe. a. modèle grossier. b. modèle plus fin.

2.6.5.1 Support Vector Machines

Pour s'en tenir au principe de base de cette technique, les SVM effectuent la classification de données entre deux classes en déterminant un hyperplan séparateur maximisant la distance (« marge ») entre ce plan et les points de l'ensemble d'apprentissage, tout en minimisant les erreurs. La **Figure 2.21** illustre des exemples d'hyperplans séparant les données de deux classes. On voit dans la **Figure 2.21** (a) trois exemples d'hyperplans séparant les deux classes parmi une infinité de solutions possibles. La **Figure 2.21** (b) illustre le cas de l'hyperplan optimal au sens de la maximisation de la marge entre les données. Comme on le verra, la définition de cet hyperplan optimal fait intervenir certains des points (les « points supports » ou « support vectors ») de l'ensemble d'apprentissage. Sur l'illustration ci-dessous, ceux-ci apparaissent en rouge.
Afin de décrire le processus de construction de l'hyperplan optimal séparant les données de deux classes différentes, on suppose l'ensemble des données linéairement séparables suivant :

$$\{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\} \text{ avec } i \in \{1,2,...,l\} \text{ et } x_i \in R^N \tag{2.1}$$

Où x_i est un point des données d'apprentissage et $y_i \in \{-1,1\}$ l'étiquette de sa classe. l est le nombre des points de données d'apprentissage tandis que N est la dimension de ces derniers.

Soit H l'hyperplan défini par $x \in H \Leftrightarrow w^T \cdot x + b = 0$, avec $x \in R^N$. Celui-ci sépare les deux classes s'il satisfait aux conditions suivantes :

$$\begin{cases} w^T \cdot x_i + b \geq +1 \Leftrightarrow y_i = +1 \\ w^T \cdot x_i + b \leq -1 \Leftrightarrow y_i = -1 \end{cases} \tag{2.2}$$

Ce qui implique :

$$y_i \cdot (w^T \cdot x_i + b) \geq 1 \tag{2.3}$$

Etant donné un point $x \in R^N$, sa distance à l'hyperplan s'exprime par $d(x) = \frac{|w^T \cdot x + b|}{\|w\|}$.

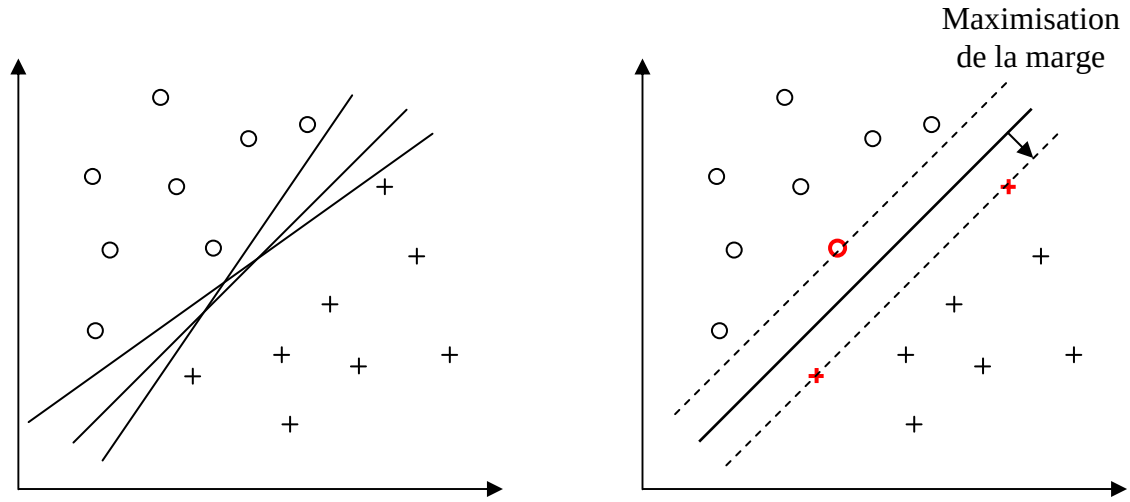


Figure 2.21 : Exemples d'hyperplans séparant les données de deux classes. *(a)* exemples d'hyperplans parmi une infinité de solutions possibles. *(b)* l'hyperplan optimal au sens de la maximisation de la marge entre des données. Les points rouges sont les points supports.

Comme indiqué en introduction, on souhaite que la distance de l'hyperplan aux points de l'ensemble d'apprentissage (« marge ») soit maximale, tout en faisant en sorte que cet hyperplan demeure séparateur. Ceci correspond alors au problème d'optimisation convexe ci-dessous :

$$\begin{cases} \min\left(\frac{1}{2} \cdot \|w\|^2\right) = \min\left(\frac{1}{2} \cdot w^T \cdot w\right) \\ \forall i \in \{1,2,...,l\},\ y_i \cdot \left(w^T \cdot x_i + b\right) \geq 1 \end{cases} \quad (2.4)$$

Pour résoudre celui-ci, la technique du Lagrangien est utilisée. Le Lagrangien $L(w,b,\alpha)$ s'exprime alors :

$$L(w,b,\alpha) = \frac{1}{2} \cdot w^T \cdot w - \sum_{i=1}^{l} \alpha_i \cdot \left[y_i \cdot \left(w^T \cdot x_i + b\right) - 1\right] \quad (2.5)$$

avec $\alpha = (\alpha_1 \quad \alpha_2 \quad ... \quad \alpha_l)^T$ les multiplicateurs de Lagrange.

Ce Lagrangien doit être minimal par rapport à w et b et maximal par rapport à α. Le point optimal est un point de selle qui vérifie :

$$\begin{matrix} \frac{\partial L(w,b,\alpha)}{\partial w} = 0 \\ \frac{\partial L(w,b,\alpha)}{\partial b} = 0 \end{matrix} \Rightarrow \begin{cases} w = \sum_{i=1}^{l} \alpha_i y_i x_i \\ \sum_{i=1}^{l} \alpha_i y_i = 0 \end{cases} \quad (2.6)$$

L'expression de w indique clairement que la solution recherchée s'exprime à partir des données de l'ensemble de l'apprentissage. Afin de préciser la valeur des $\{\alpha_i\}_{1 \leq i \leq l}$, nous

introduisons la condition complémentaire de Karush-Kuhn-Tucker $\forall i \in \{1,2,...,l\}, \alpha_i \cdot \left[y_i \cdot \left(w^T \cdot x_i + b \right) - 1 \right] = 0$. De façon explicite, un élément vérifiant $y_i \cdot \left(w^T \cdot x_i + b \right) - 1 = 0$ correspond à un individu « sur la marge » associée à l'hyperplan. En définitive, seuls les α_i associés à de tels points « sur la marge » pourront être non-nuls.
En substituant w par son expression (2.6) dans (2.5) nous aboutissons au problème dual d'optimisation, sous la forme :

$$\begin{cases} \max\left(\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \cdot \sum_{i=1}^{l} \left(\sum_{j=1}^{l} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \left(x_j^T \cdot x_i \right) \right) \right) \\ \forall i \in \{1,2,..,l\}, \alpha_i \geq 0 \\ \sum_{i=1}^{l} \alpha_i \cdot y_i = 0 \end{cases} \quad (2.7)$$

Par résolution de ce problème dual d'optimisation (2.7) on obtient les poids α_i nécessaires pour l'expression du vecteur w (2.6), et on peut donc établir la fonction de décision suivante

$$f(x) = \left(\sum_{i=1}^{l} \alpha_i \cdot y_i \cdot \left(x_i^T \cdot x \right) \right) + b \quad (2.8)$$

Comme indiqué précédemment, pour chaque classe, les points supports sont les points de la classe les plus proches de l'hyperplan séparateur, et sont les seuls pour qui les poids α_i peuvent être différents de zéro ($\alpha_i > 0$).

Nous avons insisté au début de ce développement sur le fait que l'exemple donné dans la **Figure 2.21** correspond à un cas où les données sont linéairement séparables. Cependant, ceci n'est pas le cas dans de nombreux problèmes de classification. Si cette non séparabilité correspond à la présence éventuelle de quelques individus, alors cela implique de devoir relâcher la contrainte imposant que la valeur de projection d'un vecteur x_i sur l'hyperplan défini par w soit strictement supérieure à b. Ceci est réalisé en pénalisant les valeurs de projection ne respectant pas cette condition. A cet effet, des variables de relaxation ξ_i ($\forall i,\ \xi_i \geq 0$) sont introduites pour rendre compte de la pénalisation de ces petites valeurs de projection. Il sera alors possible de relâcher la contrainte correspondante en écrivant :

$$y_i \cdot \left(w^T \cdot x_i + b \right) \geq 1 - \xi_i, \quad \text{avec } \xi_i \geq 0. \quad \forall i \in \{1,...,l\}$$

Ce qui conduit alors à minimiser

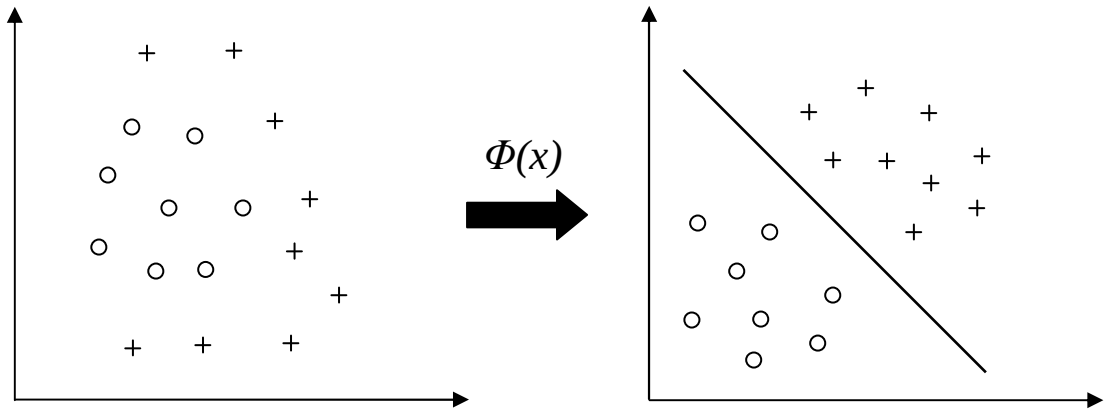
$$\frac{1}{2} \cdot w^T \cdot w + P \cdot \sum_{i=1}^{l} \xi_i$$

Ce problème se résout avec les mêmes outils que précédemment, moyennant la contrainte :
$0 \leq \alpha_i \leq P$, avec P, qui doit être défini, représente la fraction des données autorisées en dehors de la frontière afin « d'assouplir » la contrainte, et qui permet donc de faire un

compromis entre la marge maximum et les observations aberrantes [Gir97] [SmoS98] [BlaSBB96].
Dans le cas de données non linéairement séparables, la résolution peut s'effectuer en projetant les données dans un espace de grande dimension (potentiellement infinie) appelé espace de Hilbert à noyau reproduisant (*Reproducing Kernel Hilbert Space*) ou encore « espace RKHS » dans lequel les données ainsi modifiées deviennent linéairement séparables. Cette projection est effectuée à l'aide d'une fonction de projection Φ définie par :

$$\begin{cases} \Phi : R^l \mapsto R^m, \text{ où } m > l \\ x \mapsto X = \Phi(x) \end{cases}$$

La **Figure 2.22** illustre un problème de classification non linéaire. On voit dans la **Figure 2.22** (a) que les deux classes ne peuvent pas être linéairement séparées dans l'espace des caractéristiques initial. On voit dans la **Figure 2.22** (b) que la projection dans l'espace RKHS permet d'y définir un hyperplan linéaire séparant les deux classes.



(a) Espace des caractéristiques initial (b) Espace RKHS

Figure 2.22 : (a) Deux classes qui ne peuvent pas être séparées linéairement. (b) Projection des données dans l'espace RKHS dans lequel un hyperplan linéaire peut être défini.

Etant donné un couple d'individus de l'espace initial, $(x_i, x_j) \in R^l \times R^l$, « l'astuce » consiste à remarquer que le produit scalaire de leurs images par Φ peut se mettre sous la forme suivante : $\Phi(x_i)^T \cdot \Phi(x_j) = K(x_i, x_j)$, où K et une fonction symétrique vérifiant la condition de Mercer (pour toute fonction f de carré sommable, $\int K(X,Z) \cdot f(x) \cdot f(z) \cdot dx \cdot dz \geq 0$).

On pourra écrire le Lagrangien sous la forme :

$$L(w,b,\alpha,\xi,\gamma) = \frac{1}{2} \cdot w^T \cdot w - \sum_{i=1}^{l} \alpha_i \cdot \left[y_i \cdot \left(w^T \cdot x_i + b \right) - 1 + \xi_i \right] - \sum_{i=1}^{l} \gamma_i \xi_i + P \sum_{i=1}^{l} \xi_i$$

Sous la contrainte :

$$y_i \cdot \left(w^T \cdot x_i + b \right) \geq 1 - \xi_i,$$

$$\text{et} \quad \xi_i \geq 0.$$

Dans l'espace RKHS, les données étant séparables, le problème dual revêt la même forme que précédemment (on précise qu'on tolère ici aussi quelques « outliers », d'où la présence de la constante *P*). Nous pouvons cependant le réécrire en exprimant le produit scalaire à l'aide de la fonction K :

$$\begin{cases} \max\left(\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\cdot\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\cdot\alpha_j\cdot y_i\cdot y_j\cdot K\left(x_i,x_j\right)\right) \\ \forall i\in\{1,2,...,l\}, 0\le\alpha_i\le P \\ \sum_{i=1}^{l}\alpha_i\cdot y_i = 0 \end{cases} \quad (2.9)$$

La solution de (2.9) peut être apportée par les méthodes utilisées dans le cas linéairement séparable. Le point remarquable est que le problème fait intervenir K (appelé « noyau ») et pas directement la fonction de projection Φ. La difficulté se reporte alors sur la sélection d'un noyau K adéquat. Parmi ceux possibles, les noyaux gaussiens sont très populaires et sont définis par :

$K\left(x_i,x_j\right) = e^{-\frac{\left\|x_i-x_j\right\|^2}{2\cdot\sigma^2}}$, où σ est l'écart type de la distribution gaussienne.

On trouve par ailleurs les « noyaux polynomiaux », dont une formulation est :

$K\left(x_i,x_j\right) = \left(1+x_i\cdot x_j\right)^n$, avec n le degré du polynôme.

Une fois obtenue la solution de (2.9) permettant de déterminer $\{\alpha_i\}_{1\le i\le l}$ et b, la fonction de décision est donnée par :

$$f(x) = \left(\sum_{i=1}^{l}\alpha_i\cdot y_i\cdot K(x_i,x)\right) + b \quad (2.10)$$

Bien que les SVM aient été initialement développés pour traiter des problèmes binaires (séparation de deux classes), ils peuvent être étendus notamment pour des problèmes multi-classes ou monoclasse (plutôt appelés « *one-class* »).

2.6.5.2 Extensions de la formulation SVM

2.6.5.2.1 Stratégie multi-classe

Lorsque plusieurs classes sont définies dans l'ensemble d'apprentissage, une stratégie multi-classe peut être exploitée en utilisant une combinaison de classifieurs binaires. Deux approches sont alors principalement rencontrées [HsuL01] :

Un contre un : Pour *n* classes, $\frac{n(n-1)}{2}$ classifieurs sont entrainés, chacun opposant deux classes. Pour classifier une nouvelle donnée, celle-ci doit être testée par chacun des classifieur, et la classe finale est attribuée par vote.

Un contre tous : Pour *n* classes, *n* classifieurs sont entrainés, chacun opposant les données d'une classe aux données de toutes les autres classes. Lors de l'apprentissage de chaque classifieur, les ensembles d'apprentissage sont constitués par une seule classe pour les exemples positifs de la classe en cours d'apprentissage, tous les autres sont définis comme exemples négatifs. Là encore, pour classifier une nouvelle donnée, celle-ci est testée par chacun des classifieurs, et est affectée à la classe correspondant à la prédiction la plus forte parmi tous ceux-ci.

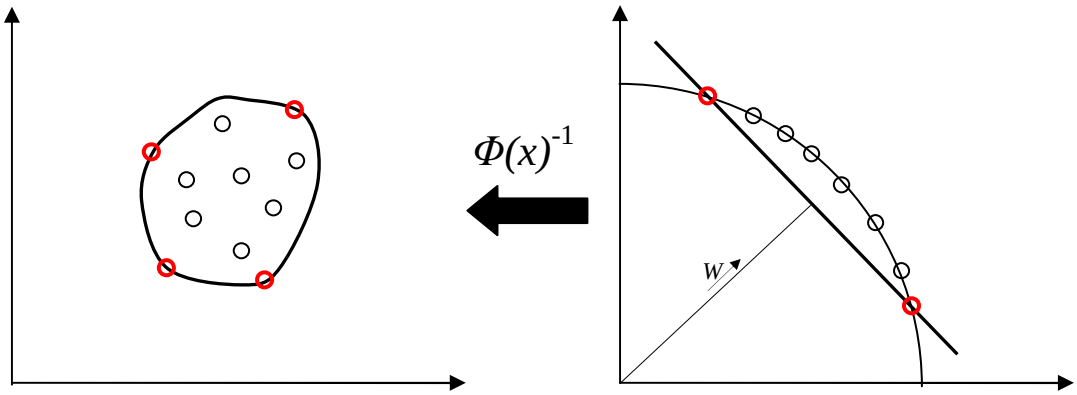
2.6.5.2.2 One–class SVM

Une variante de la formulation SVM propose de définir un classifieur dont l'hyperplan définit la frontière d'une seule classe [HaySTA00] [UnnRJ3]. La **Figure 2.23** donne un exemple de frontière obtenue. On peut voir dans la **Figure 2.23 (**b), sur une découpe 2 dimensions (raison pour laquelle seuls deux des quatre supports apparaissent sur la figure), que la projection des données dans l'espace RKHS en utilisant un noyau gaussien permet de les positionner sur le quart d'une hypersphère (dans l'espace RKHS). Là encore, la frontière de décision de la classe peut alors être définie par un hyperplan linéaire. Les points supports de la classe sont positionnés à l'intersection de l'hypersphère et de l'hyperplan. Dans l'espace des caractéristiques initial, cet hyperplan a pour représentation la plus petite hypersphère englobant les données contenues dans la classe (voir la **Figure 2.23** (a) pour un exemple en deux dimensions). Les points supports de la classe sont positionnés sur le contour de cette hypersphère. On rappelle que puisqu'ici la base d'apprentissage contient uniquement des exemples positifs, l'étiquette y_i de chaque point x_i (équation 4.10) aura la valeur 1. Si on injecte cette valeur dans la fonction de décision, nous obtenons :

$$f(x) = \left(\sum_{i=1}^{l} \alpha_i \cdot K(x_i, x) \right) + b \qquad (2.11)$$

L'un des intérêts majeurs de cette méthode est de pouvoir définir la frontière de décision d'une classe en utilisant uniquement les données de celle-ci : l'ensemble d'apprentissage contient uniquement des données de cette dernière (« exemples positifs »). Ceci permet de nous éviter de devoir incorporer à la base d'apprentissage des exemples négatifs, ce qui serait une tâche laborieuse, car nous aurions à ressembler des exemples d'image de tout ce qui n'est pas de la peau. Ainsi, cette technique peut être exploitée pour faire une classification « un-contre-tous », c'est-à-dire qu'un pixel donné serait classé soit appartenant à la classe peau soit non. L'idée originale a notamment permis d'appliquer les SVM à l'apprentissage en ligne et à la détection de nouveautés en ligne [AngC03] [CauP00]. Nous utiliserons notamment ces techniques pour la classification de personnes (chapitre 4).

Comme nous l'avons précisé, nous avons choisi de développer un classifieur basé sur l'approche des one-class SVM pour la définition de la classe « peau » et d'une métrique de similarité permettant de définir la proximité du modèle courant et son éventuelle mise à jour. Cette technique de classification nous permet d'obtenir un modèle plus fin, c'est-à-dire plus ajusté de la classe peau. Par ailleurs, elle nous permet de construire le modèle « classe peau » en utilisant, lors de l'apprentissage initial, uniquement des pixels-peau (exemples positifs). La méthode que nous avons développée à cette fin est équipée de règles d'apprentissage incrémental de sorte à adapter en ligne la frontière de décision de la classe peau au cours de son évolution dans l'espace des caractéristiques [CauP00] [HamBBL09], évolution due au phénomènes perturbateurs évoqués précédemment. Les règles d'apprentissage et de désapprentissage incrémentaux sont présentées ci-après.



(a) Espace des caractéristiques initial (b) Espace RKHS

Figure 2.23 : Définition de la frontière de décision d'une classe. Seuls les points appartenant à la classe elle-même (points positifs) sont contenus dans l'ensemble d'apprentissage. La frontière de décision déterminée dans l'espace RKHS (b) est représentée dans l'espace des caractéristiques initial (a). Les points rouges sont les points supports de la classe.

2.6.5.3 Classification incrémentale des pixels-peau

Après avoir initialisé la frontière de décision du modèle de la classe peau au début de la classification, l'objectif est d'adapter de manière séquentielle cette frontière en fonction de l'évolution du modèle de cette classe au fil de la séquence. La **Figure 2.24** illustre ce procédé. Ici, les *croix* représentent les pixels qui appartiennent à la classe peau à l'instant k. A l'inverse, les *cercles* représentent les pixels qui appartiennent à la classe peau à l'instant $k+1$. De fait, la frontière de décision à l'instant k doit englober les *croix* tandis qu'à l'instant $k+1$ elle doit englober les *cercles*. Aussi, l'adaptation de la frontière de décision entre les instants k et $k+1$doit se faire en ajoutant à la base d'apprentissage les *cercles* (caractérisant les données les plus récentes), tout en y supprimant les *croix* (données les plus anciennes et potentiellement obsolètes). Puisque la mise à jour de la frontière de décision se fait de manière progressive, on parlera alors « d'apprentissage en ligne incrémental » pour l'ajout des nouvelles données, et « d'apprentissage en ligne décrémental » pour la suppression des anciennes. Ce procédé permet de passer de la frontière de décision à l'instant k (contour continu) à ce qu'elle est à l'instant $k+1$ (contour en pointillés).

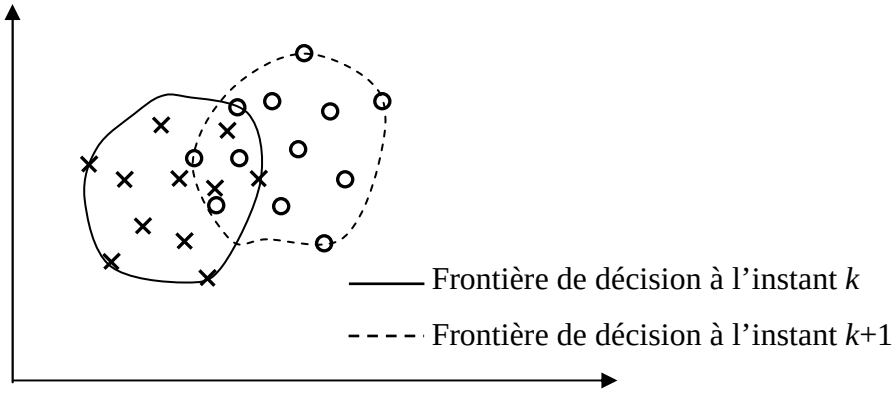


Figure 2.24 : Adaptation de la frontière de décision suivant l'évolution des pixels du modèle de la classe peau dans l'espace des caractéristiques.

Afin de formaliser la méthode mise en œuvre, posons C l'ensemble des pixels-peau et f_k sa fonction de décision temporelle à l'instant k, correspondant en pratique à la $k^{\text{ième}}$ image de la séquence qui est supposée en comporter un total de N. Dans l'image de la séquence traitée à cet instant k, étant donné un pixel p dont la couleur est définie par $x = \begin{pmatrix} x_{C1} & x_{C2} & x_{C3} \end{pmatrix}^T$, celui-ci sera classé comme étant un « pixel-peau » si $f_k(x) \geq 0$, c'est-à-dire :

$$\begin{aligned} &\text{si} \quad f_k(x) \geq 0, \quad \text{alors} \quad x \in C \\ &\text{si} \quad f_k(x) < 0, \quad \text{alors} \quad x \notin C \end{aligned}$$

f_k est définie dans l'espace des caractéristiques par :

$$f_k(x) = \left(\sum_{i=1}^{d} \alpha_i \cdot K(x_i, x) \right) + b \tag{2.12}$$

Où *b* est le biais (*offset*) de la fonction, *K(•,•)* est un noyau gaussien et *d* est le nombre total de pixels-peau à l'instant *k*. Les poids α_i (multiplicateurs de Lagrange) sont obtenus en minimisant une fonction objective quadratique convexe [CauP00]:

$$\min_{0 \leq \alpha \leq C} : W = \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} \alpha_i \alpha_j \cdot K(x_i, x_j) - \sum_{i=1}^{d} \alpha_i + b \sum_{i=1}^{d} \alpha_i \tag{2.13}$$

A chaque nouvelle image à traiter (qu'on appellera image courante), la fonction de décision f_k est adaptée en ajoutant à la base d'apprentissage les pixels-peau de cette image courante puis en y retirant ceux des plus anciennes images (suivant un critère d'ancienneté qui est présenté dans ce qui suit). Par exemple, à l'instant *k*+1 la fonction de décision sera recalculée de manière itérative en ajoutant à l'ensemble d'apprentissage les pixels de l'image *k*+1 classés comme pixels-peau (apprentissage en ligne incrémental), puis en retirant de manière itérative (apprentissage en ligne décrémental) les pixels-peau de toutes les images antérieures à l'image *k-m*. (avec *m* une constante à fixer selon la cadence d'acquisition des images). L'ajout des nouvelles données puis la suppression des anciennes permettent d'ajuster en ligne la frontière de décision de la classe peau. La clé est d'ajouter les pixels-peau de l'image courante à la solution puis d'en retirer ceux des images trop anciennes tout en conservant les conditions de Karush Kuhn Tucker (KKT) [CauP00] satisfaites sur tous les pixels-peau classés. Les conditions du premier ordre sur le gradient de *W* mènent aux conditions de KKT [CauP00] :

$$\begin{aligned} g_i &= \frac{\partial W}{\partial \alpha_i} = \sum_{j=1}^{s} \alpha_j K(x_i, x_j) + b = f(x_i) \begin{cases} > 0; \alpha_i = 0 \\ = 0; 0 < \alpha_i < a, a = cste \\ < 0; \alpha_i = a \end{cases} \\ \frac{\partial W}{\partial b} &= \sum_{j=1}^{s} \alpha_j - 1 = 0 \end{aligned} \tag{2.14}$$

Où *s* est le nombre de points supports de la classe peau.

Cette opération permet de classer tous les pixels de l'image courante en 3 ensembles :

- L'ensemble *D*, qui contient les pixels dits « points intérieurs », situés à l'intérieur de la frontière de décision, donc avec $\forall x_i \in D,\ g_i = f(x_i) > 0$.
- L'ensemble *S*, qui contient les points supports, situés sur frontière de décision, donc avec $\forall x_j \in S,\ g_j = f(x_j) = 0$.
- L'ensemble *U*, qui contient les pixels dits « points extérieurs », situés à l'extérieur de la frontière de décision, donc avec $\forall x_u \in U,\ g_u = f(x_u) < 0$.

Puisque *C* est l'ensemble des pixels-peau, alors lorsqu'un un pixel candidat (pixel en cours de classification) x_c de l'image courante *k* est classé dans *D* (renvoyant une valeur $g_c > 0$) ou *S* (renvoyant une valeur $g_c = 0$), il est immédiatement classé comme étant un pixel-peau (donc ajouté à *C*). En outre, à tout pixel-peau $x_c \in (C = D \cup S)$, sera associée sa valeur g_c. Avec $g_c = 0$ pour tout point support, et $g_c > 0$ pour tout point intérieur. Nous enregistrons l'ensemble de ces valeurs *g* dans un tableau *G*. Le tableau *G* aura donc *d* éléments, avec *d* le nombre total de pixels-peau à l'instant *k*.

Il est important de noter que lorsque x_c est ajouté à *C*, la frontière de décision doit être adaptée et recalculée jusqu'à ce que ce pixel se retrouve positionné sur celle-ci et devienne un point support (donc jusqu'à ce que $g_c = 0$). Les détails de cette procédure seront présentés ultérieurement.

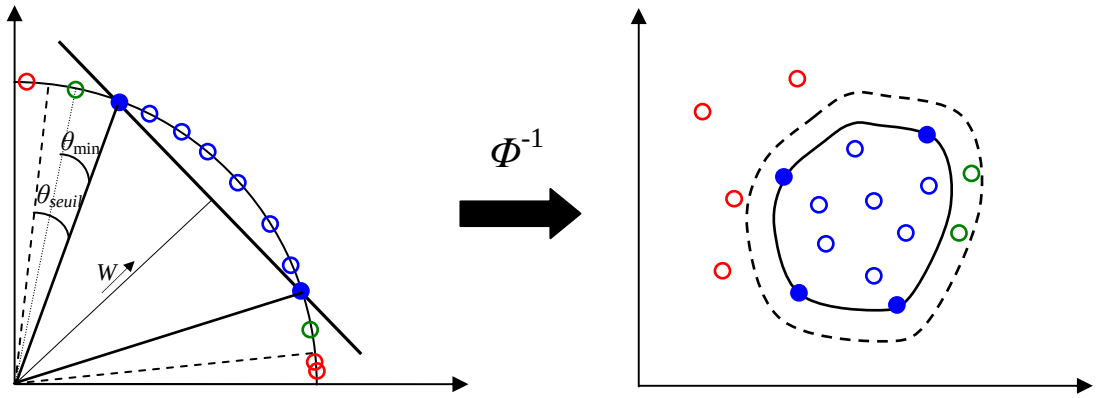
Lorsqu'un pixel candidat x_c de l'image courante *k* renvoie une valeur $g_c < 0$, cela signifie que ce pixel est situé à l'extérieur de la frontière de décision. Ici, on considérera deux cas possibles. Le premier est que ce pixel corresponde simplement à un objet autre que de la peau, et doit donc naturellement être rejeté et classé comme pixel-non-peau. L'autre cas est qu'il puisse s'agir d'un pixel qui corresponde à de la peau dans l'image mais dont les attributs auraient subi des changements dus à des variations d'éclairage par exemple, et se retrouve ainsi à l'extérieur de la frontière de décision du modèle peau. Dans ce cas, il faut ajouter ce pixel au modèle peau et mettre à jour la frontière de décision et l'adapter, ici aussi, jusqu'à ce que ce pixel se retrouve positionné sur celle-ci et devienne un point support (donc jusqu'à ce que $g_c = 0$).

Pour décider entre les deux cas possibles, on doit vérifier la proximité de ce pixel avec le modèle de la classe peau. On définit cette proximité en calculant la distance entre ce pixel et la frontière de décision de la classe peau, ce qui revient à faire une approximation de la distance entre ce pixel et le point support le plus proche. Si cette distance est grande, on considérera alors qu'on est dans le premier cas. Mais si cette distance est petite, c'est-à-dire que le pixel est assez proche de la classe peau, on se positionnera alors dans le deuxième cas. Bien évidemment, nous devons définir ce qu'est une « petite distance ». Nous mettons alors au point une mesure permettant d'évaluer la distance entre un pixel extérieur et la frontière de décision du modèle peau.

2.6.5.4 Mesure de distance

Nous définissons une mesure de distance entre un pixel extérieur et la frontière de décision de la classe peau dans l'espace RKHS. Comme il a été précisé précédemment, lorsque le noyau de la fonction de projection est un noyau gaussien, toutes les données sont positionnées sur un quart d'hypersphère, et la frontière de décision de la classe est définie par un hyperplan

linéaire (**Figure 2.23** (b)). Dans cet espace, les points supports de la classe peau sont positionnés à l'intersection de l'hyperplan et de l'hypersphère (points bleus pleins dans la **Figure 2.25** (a), et les autres données de la classe sont à l'intérieur, c'est-à-dire entre l'hyperplan et l'hypersphère (points bleus creux dans la **Figure 2.25** (a). Toute autre donnée n'appartenant pas à la classe peau (pixels-non-peau) est positionnée à l'extérieur (point verts et points rouges dans la **Figure 2.25**. Dans l'espace des caractéristiques initial, l'hyperplan de la classe peau se résume à la plus petite sphère (en 3 dimensions selon les trois canaux de l'espace couleur utilisé) englobant les données contenues dans la classe (contour en ligne continue dans la **Figure 2.25** (b), représenté en 2 dimensions).



(a) Espace RKHS (b) Espace des caractéristiques initial

Figure 2.25 : Positionnement des données dans : (a) l'espace RKHS, (b) l'espace des caractéristiques initial.

Dans l'espace RKHS, afin de définir la distance entre un point x_c (pixel candidat x_c situé à l'extérieur de l'hyperplan ($g = f_k(x_c) < 0$)) et cet hyperplan (ce qui revient à calculer la distance entre x_c et le point support le plus proche), on calcule l'angle entre ce point et chaque point support (**Figure 2.25** (a). Puis, on retient l'angle le plus petit, correspondant au point support le plus proche. Cet angle minimum correspond alors à la distance (ou plutôt pseudo-distance) entre ce point et l'hyperplan. Dans cet espace RKHS, le produit entre le point x_c et un point support x_j est donné par [BouL08]:

$$\left\langle \phi(x_c), \phi(x_j) \right\rangle = \left\| \phi(x_c) \right\| . \left\| \phi(x_j) \right\| . \cos(\phi(x_c), \phi(x_j)) \tag{2.15}$$

Aussi, en utilisant le noyau gaussien, le plus petit angle $\theta_{\min}$ est exprimé par :

$$\theta_{\min} = \min\left\{ \cos^{-1}\left(K(x_c, x_j) \right) \right\}, \quad j = 1...s \tag{2.16}$$

On définit alors un angle seuil θ_{seuil} avec lequel on va comparer l'angle minimum $\theta_{\min}$.

- Si $\theta_{\min} > \theta_{seuil}$, alors le point x_c sera rejeté et classé comme pixel-non-peau.

- Si $\theta_{\min} < \theta_{seuil}$, alors le point x_c sera ajouté au modèle de la classe peau, et la frontière de décision sera adaptée et ajustée jusqu'à ce que ce pixel se retrouve positionné sur celle-ci et devienne un point support (donc jusqu'à ce que $g = f_k(x) = 0$).

Comme on vient de le voir, lorsque $\theta_{\min} < \theta_{seuil}$, on estimera que le point x_c est « assez proche » de la frontière de décision (les ajustements expliqués précédemment seront effectués). L'angle seuil θ_{seuil} représente alors la limite de ce qu'on admettra comme distance « assez petite » et jusqu'à laquelle on estimera qu'un point est « assez proche ». Dans l'espace des caractéristiques initial, cela correspond à une région située entre le contour de la classe peau (contour en ligne continue dans la **Figure 2.25** (b)) défini donc par l'hyperplan de la classe peau, et un contour enveloppant (contour en ligne discontinue dans la **Figure 2.25** (b)) défini par l'angle seuil θ_{seuil}. Dans l'exemple de la **Figure 2.25** (b), les points verts sont situés dans cette région.

Pour résumer, chaque pixel x_c candidat de l'image à classer peut être traité de quatre façons. La première est qu'il soit classé comme pixel-peau car $f_k(x_c) > 0$ et devra donc avoir un poids $\alpha_c = 0$ dans la solution finale. La frontière de décision n'aura pas à être adaptée (puisque $\alpha_c = 0$). La deuxième est qu'il soit classé comme pixel-peau car $f_k(x_c) = 0$ et sera donc un point support et donc pourra avoir un poids $\alpha_c > 0$ dans la solution finale. La frontière de décision devra alors être adaptée. La troisième est qu'il soit classé comme pixel-peau car $f_k(x_c) < 0$ et $\theta_{\min} \leq \theta_{seuil}$, et sera donc un point support et pourra avoir un poids $\alpha_c > 0$ dans la solution finale. La frontière de décision devra alors être adaptée. Et la dernière, est qu'il soit classé comme pixel-non-peau et rejeté car $f_k(x_c) < 0$ et $\theta_{\min} > \theta_{seuil}$, et ne sera évidemment pas ajouté à la solution finale.

Aussi, dans le deuxième et le troisième cas, x_c sera ajouté au modèle peau en tant que point support et son poids α_c initialement égal à zéro devra être recalculé de manière itérative de sorte qu'il soit supérieur à zéro ($\alpha_c > 0$) tout en gardant les conditions de Karush Kuhn Tucker (KKT) satisfaites (équation (2.14)). Aussi, on dira que le modèle de la classe peau va apprendre en ligne le point x_c en l'ajoutant à la solution finale. Et comme cet apprentissage s'effectuera de manière incrémentale, on parlera alors d'apprentissage incrémental.

2.6.5.5 Apprentissage incrémental

Lorsque le pixel x_c est ajouté à l'ensemble des points supports *S*, la fonction de décision de la classe peau est adaptée et mise à jour itérativement, c'est-à-dire que ses paramètres α_j, b sont mis à jour et recalculés de manière itérative et incrémentale. A chaque itération, $g_c = f_k(x_c)$ est recalculée jusqu'à ce que $g_c = f_k(x_c) = 0$ tout en gardant les conditions KKT satisfaites. On doit alors définir les pas d'incrémentation Δg_i (avec i = 1,...,d), $\Delta\alpha_j$ (avec j = 1,..., s) et Δb [BouL08]. On pose z comme l'ensemble des paramètres $\{b, \alpha_j\}$. Ces paramètres varient de manière à garder leurs conditions KKT satisfaites. Pour cela, ces conditions sont exprimées de manière différentielle par :

$$\begin{aligned} \Delta g_i &= \Delta\alpha \cdot K(x_i, x_c) + \sum_{j=1}^{s} \Delta\alpha_j \cdot K(x_i, x_j) + \Delta b \\ 0 &= \Delta\alpha + \sum_{j=1}^{s} \Delta\alpha_j \end{aligned} \tag{2.17}$$

Puisque $g_i = 0$ pour chaque point support, les changements des poids doivent satisfaire :

$$\underbrace{\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & K(x_1,x_1) & \cdots & K(x_1,x_s) \\ \vdots & & & \\ 1 & K(x_s,x_1) & \cdots & K(x_s,x_s) \end{bmatrix}}_{Jacobienne\,Q} \cdot \underbrace{\begin{bmatrix} \Delta b \\ \Delta\alpha_1 \\ \vdots \\ \Delta\alpha_s \end{bmatrix}}_{Delta} = -\underbrace{\begin{bmatrix} 1 \\ K(x_1,x_c) \\ \vdots \\ K(x_s,x_c) \end{bmatrix}}_{h} .\Delta\alpha$$

Aussi, $Delta = -\underbrace{Q^{-1}}_{R} \times h \times \Delta\alpha$

Ainsi, seront satisfaits

$$\begin{aligned} \Delta b &= \beta_0 .\Delta\alpha \\ \Delta\alpha_j &= \beta_j .\Delta\alpha, \quad \forall x_j \in S \end{aligned} \tag{2.18}$$

Avec les poids donnés β par

$$[\beta_0 \beta_1 \ldots \beta_s] = -R \cdot h \tag{2.19}$$

Où $R = Q^{-1}$, et $\beta = 0$ pour tout pixel x qui ne soit pas point support. De cette façon, les valeurs des pas d'incrémentation ($\Delta\alpha_j$, Δb, Δg_i) de tous les paramètres sont calculées.

La valeur associée g_c et le poids α_c du pixel x_c ajouté à S seront respectivement ajoutés au tableau G et à un ensemble A comme suit : $G^{s+1} \leftarrow G^s \bigcup \{g_c\}$ et $A^{s+1} \leftarrow A^s \bigcup \{\alpha_c\}$. La matrice R (Q) sera mise à jour en ajoutant une ligne et une colonne correspondant au nouveau pixel x_c.

Quand x_c est ajouté à S, selon la mise à jour de la fonction de décision, tous les éléments de G doivent être modifiés comme suit :

$$\forall x_i \in D, \qquad \Delta g_i = \gamma_i \cdot \Delta\alpha, \qquad i = 1 \ldots d \tag{2.20}$$

Où γ_i est défini par:

$$\gamma_i = K(x_i, x_c) + \sum_{j=1}^{s} K(x_i, x_j) \cdot \beta_j + \beta_0, \qquad i = 1 \ldots d \tag{2.21}$$

Il est à noter que durant l'ajustement de la fonction de décision, un point support x_j qui se trouvait sur l'hyperplan peut se retrouver à l'intérieur, c'est-à-dire que par la procédure d'incrémentation, pour garder les conditions KKT satisfaites, α_c peut se retrouver égal à 0. Dans ce cas x_j sera éliminé de S et mis dans D, et les paramètres b, α_j et R seront mis à jour.

On rappelle enfin que lorsqu'un pixel x_c est classé comme point intérieur ($g_c = f_k(x_c) > 0$) avec donc $\alpha_c = 0$, seul G est mis à jour : $G^{s+1} \leftarrow G^s \bigcup \{g_c\}$.

Algorithme d'apprentissage incrémental : On peut résumer la procédure d'apprentissage incrémental d'un pixel candidat x_c par le pseudo-code suivant :

Initialiser α_c à zéro
Si $g_c > 0$, ajouter x_c à *D*, mettre à jour *G* ($G^{s+1} \leftarrow G^s \cup \{g_c\}$), sortir.
Si $g_c = 0$, ajouter x_c à *S*, mettre à jour les paramètres α_j, *b*, *R* et *G*, sortir.
Si $g_c < 0$, ajouter x_c à *U* calculer l'angle $\theta_{\min}$,
    Si $\theta_{\min} < \theta_{seuil}$,
        Ajouter x_c à *S*
        Tant que $g_c < 0$ faire
            $\alpha_c = \alpha_c + \Delta\alpha$
            Calculer β
            Calculer Δb, puis $b = b + \Delta b$
            Pour chaque $x_j \in S$,
                Calculer $\Delta\alpha_j$, puis $\alpha_j = \alpha_j + \Delta\alpha_j$
            Pour chaque $x_i \in D$,
                Calculer Δg_i, puis $g_i = g_i + \Delta g_i$
        Vérifier si un point support se retrouve à l'intérieur de l'hyperplan ($\alpha_j \leq 0$). Si oui, l'enlever de *S* et l'ajouter à *D*, and et mettre à jour tous les paramètres.
        Répéter jusqu'à ce que $g_c = 0$.

2.6.5.6 Apprentissage décrémental

La procédure de suppression des plus anciennes données, complémentaire avec la procédure d'ajout de nouvelles données nous permet de suivre l'évolution de la classe peau. Quand le système traite la $k^{\text{ème}}$ image, les pixels-peau appris sur la $(k\text{-}m)^{\text{ième}}$ image correspondent à des informations potentiellement obsolètes qu'ils seraient nécessaire de supprimer. Aussi, quand un pixel x_j est retiré de *S*, g_j sera retiré de *G*, et *z* (ensemble des paramètres $\{b, \alpha_j\}$) sera mis à jour de façon décrémentale et la fonction de décision f_k sera ajustée jusqu'à ce que x_j soit à l'extérieur ($\alpha_i \leq 0$). La matrice *R* est mise à jour en éliminant de la matrice *Q* la colonne *j*+1 et la ligne *j*+1 (correspondant à x_j qui a été retiré). Quand un pixel x_i est retiré de *D*, seul *G* est mis à jour en lui retirant g_i.

Algorithme d'apprentissage décrémental : En retirant le pixel x_r de *C*, les paramètres $\{\alpha_j^{s-1}, b^{s-1}\}$ sont exprimés en fonctions des paramètres $\{\alpha_j^s, b^s\}$, la matrice *R*, et x_r par :

Si $g_r > 0$, retirer x_r de C, et retirer g_r de G ($G \leftarrow G - \{g_r\}$), terminer.
Si $g_r = 0$, retirer x_r de S (et donc de C),
Tant que $\alpha_r > 0$, faire
$\alpha_r = \alpha_r - \Delta\alpha$
Calculer β
Calculer Δb, $b = b - \Delta b$
Pour chaque $x_j \in S$,
Calculer $\Delta\alpha_j$, puis, $\alpha_j = \alpha_j - \Delta\alpha_j$
Pour chaque $x_i \in C$,
Calculer Δg_i, puis $g_i = g_i - \Delta g_i$
Vérifier si un point intérieur $x_i \in D$ se retrouve à l'extérieur de la frontière de décision ($g_i \leq 0$). Si oui, interrompre la procédure de suppression, et appliquer la procédure d'ajout sur x_i, puis reprendre la procédure d'apprentissage d'oubli jusqu'à ce que $\alpha_r = 0$.

2.6.6 Evaluation de la détection de la peau par la classification avec frontière de décision dynamique

Nous appliquons ici notre technique de classification dynamique pour la détection de la peau afin d'évaluer l'amélioration des taux de détection par rapport à la classification avec seuils fixes. Nous avons donc utilisé les mêmes images avec les mêmes prétraitements (Zone de détection de la peau). Les taux (en pourcentage) de bonne détection obtenue avec les différents espaces couleurs sont présentés dans la **Tableau 2.2**. Les taux de bonne détection obtenue avec la classification avec seuils fixes y sont également rappelés. On voit clairement ici l'augmentation des taux de bonnes détection obtenus. Le meilleur score est toujours obtenu avec l'espace THS avec 90% de bonne détection et qui augmente de 11%, suivi par l'espace HSV avec 83% et qui augmente de 9%. Les espaces YCbCr et RGB obtiennent tous deux un score de 81% avec une augmentation de 12% pour le premier et de 19% pour le deuxième. On constate ici que la plus forte amélioration est obtenue avec l'espace RGB. Ceci peut s'expliquer par le fait que cet espace est, tel que nous l'avons évoqué précédemment, le plus sensible au changement d'éclairage. Ainsi, l'utilisation de la classification dynamique permettant de parer des changements d'éclairage du visage a permis de rattraper la sensibilité de l'espace RGB. Cela a évidement également permis d'améliorer les taux de détection sur l'ensemble des espaces couleur utilisés.

	Espaces couleur			
	RGB	YCbCr	HSV	THS
Taux de détection avec classification dynamique	81%	81%	83%	90%
Taux de détection avec classification par seuils	62%	69%	74%	79%

Tableau 2.2 : Taux de bonnes détections de pixels-peau du visage.

2.7 Conclusion

Nous avons présenté dans ce chapitre les différentes étapes de traitement afin de détecter et d'isoler chaque individu présent dans la scène afin de tenter, par la suite, de le reconnaitre. Notre approche de détection utilise la soustraction de fond. Nous avons ainsi les processus de génération et de mise à jour du modèle du fond ainsi que les post-traitements effectués afin de segmenter les blobs correspondant aux éléments au premier-plan. Nous avons également détaillé la méthode que nous avons développée afin de nous assurer que les blobs détectés étaient bien des humains. Cette méthode se base sur la détection de la peau du visage. Nous avons ainsi mis en œuvre un algorithme de classification dynamique permettant d'améliorer la robustesse aux changements d'éclairage et ainsi d'améliorer le taux de bonne détection. A ce stade, le résultat des différents traitements appliqués aux images est la détection des individus présents dans la scène. Ainsi, chaque individu pourra être reconnu et identifié. La procédure de reconnaissance de personnes sera présentée par la suite. Mais au préalable, nous présentons dans le chapitre 3 les approches de reconnaissance de personnes qu'on trouve dans la littérature.

CHAPITRE 3

3. METHODES DE RECONNAISSANCE DE PERSONNES

3.1 Introduction

Assurer une sécurité totale dans une installation surveillée exigerait que les systèmes puissent en permanence analyser les scènes filmées, et cela en suivant la position et reconnaissant, en continu, l'identité et l'activité des personnes dans l'espace surveillé. Traditionnellement, les systèmes de surveillance ont été axés sur le suivi, l'analyse de trajectoire et la reconnaissance de l'activité [IvaDHE09], alors que les systèmes biométriques ont, quant à eux, été axés sur l'identification des individus [JaiRP04]. Aujourd'hui, alors que les technologies de surveillance intelligente atteignent un certain degré de maturité, il devient possible d'intégrer les techniques de reconnaissance et celles basées sur l'analyse de l'activité dans un même système, permettant ainsi une compréhension de la scène plus complète.

La reconnaissance des personnes peut s'effectuer soit par des méthodes biométriques ou par des méthodes basées sur l'apparence. La reconnaissance par méthodes biométriques fait référence à l'utilisation de différentes caractéristiques physiologiques (empreintes digitales, visage, rétine, iris, voix) et comportementales (manière de marcher, dynamique de la signature, *etc.*), appelées « caractéristiques biométriques », ou simplement « biométrie », pour reconnaître automatiquement les individus. Toutes les caractéristiques biométriques d'une personne sont, en fait, une combinaison de caractéristiques physiologiques et comportementales uniques et propres à chaque personne. A l'inverse, les méthodes basées sur l'apparence font référence à l'apparence extérieure des personnes, et utilisent généralement des caractéristiques décrivant leur forme et leur taille ou les caractéristiques de leurs vêtements tels que la couleur et la texture. Ainsi, les éléments pris en compte dans les modèles d'apparence peuvent être communs à plusieurs individus (la taille ou la couleur des vêtements par exemple), chacun d'entre eux se distinguant par la combinaison qu'il le caractérise.

A l'origine, les systèmes de reconnaissance biométrique ont été développés pour des applications autres que la vidéosurveillance proprement dite, telles que le contrôle d'accès par exemple. Dans ce genre d'application, la personne à reconnaitre est consentante ou, quoiqu'il en soit, prend part d'une manière active à l'extraction de ces attributs biométriques afin de recueillir les données utiles à la reconnaissance. Il peut s'agir, suivant les cas, de présenter son visage devant une caméra (reconnaissance de visage) ou son œil (reconnaissance d'iris) voire de poser son doigt sur un appareil dans le cas de la reconnaissance d'empreintes. Plus récemment, quand les systèmes de vidéosurveillance se sont fortement développés et déployés, les chercheurs ont essayé d'intégrer la reconnaissance biométrique à ces systèmes pour permettre la reconnaissance de personnes. En pratique, bien que certaines techniques telles que la reconnaissance de visage aient été adoptées, les méthodes biométriques sont soumises à des contraintes très difficiles à satisfaire sur le terrain, notamment si l'on considère les conditions de prise de vue et la résolution des caméras. Dans des applications de vidéosurveillance, où les individus n'ont pas à participer de manière active au processus d'extraction de caractéristiques, obtenir celles-ci est particulièrement difficile. Aussi, les seules méthodes de reconnaissance biométrique développées pour la vidéosurveillance portent sur la reconnaissance de visage et la reconnaissance de la démarche, du fait de leur caractère non intrusif dans l'activité réalisée par le sujet, au sens où elle ne requiert pas la coopération de ce dernier.

Contrairement aux méthodes biométriques, les méthodes de reconnaissance basées sur l'apparence manquent du caractère univoque des caractéristiques utilisées. En effet, si le visage ou la démarche sont propres à chaque personne, l'apparence d'un individu dépend en grande partie de sa tenue vestimentaire. Sachant que la plupart des gens changent de vêtements tous les jours, les méthodes de reconnaissance basées sur l'apparence ne peuvent être utilisées que pour des applications à court terme, et deviennent inutilisables pour des applications de reconnaissance qui portent sur une longue durée (plusieurs jours), en l'absence de caractère obligatoire (port de l'uniforme, par définition peu discriminant !) ou cyclique (assez rare) des tenues portées. Cependant, dans des applications de vidéosurveillance où, pour la plupart, le système ne conserve les données vidéo que sur une période limitée (rarement plus de quelques jours), ces méthodes sont très appropriées. En effet, si la majorité des gens changent de vêtements tous les jours, il est beaucoup plus rare qu'on en change durant une même journée. Les avantages des méthodes basées sur l'apparence sont multiples. Par exemple, les images utilisées dans ces méthodes sont des images de personnes dont le corps entier est visible, et c'est précisément le genre d'images acquises par les caméras de surveillance. De plus, dans leur application, ces méthodes ne posent pas autant de contraintes et de restrictions que les méthodes biométriques.

Afin de mieux cerner ce qu'il est possible d'attendre de celles-ci, nous décrivons dans ce chapitre les trois grandes approches de reconnaissance de personne utilisées dans des applications de vidéosurveillance, à savoir la reconnaissance de visage, la reconnaissance de la démarche et la reconnaissance basée sur l'apparence. Nous parlerons des principales méthodes développées et mettrons en évidence les faiblesses majeures et les difficultés inhérentes à chacune des trois approches présentées. Compte-tenu de leur adéquation avec les applications de type vidéosurveillance, nous nous focaliserons particulièrement sur les méthodes de l'approche basée sur l'apparence.

3.2 Reconnaissance du visage

3.2.1 Introduction

La reconnaissance du visage est la technique la plus commune et la plus populaire parmi les techniques de reconnaissance biométrique. Elle reste la plus intuitive puisqu'elle correspond à ce que les humains utilisent pour se reconnaitre entre eux. Elle a reçu une attention accrue dans le monde de la recherche du fait de son caractère non invasif, au sens où elle ne requiert pas la coopération du sujet (dans le cadre de la vidéosurveillance). En effet, le visage d'un individu peut être enregistré par une caméra à distance. Le développement d'outils performants de suivi a permis son déploiement à large échelle. Ainsi, la reconnaissance du visage est utilisée dans un grand nombre d'applications incluant la sécurité, le contrôle d'accès, le fichage, la communication et le loisir informatique.

On précise ici qu'il faut distinguer la « reconnaissance de visage » et la « détection de visage », que nous avons évoquée dans le chapitre 2. En effet, la détection de visage est l'opération qui consiste à « localiser » l'emplacement du ou des visages présents dans l'image. Quant à la reconnaissance de visage, elle consiste à « identifier » une personne par son visage parmi un groupe de personnes. Naturellement, ces deux opérations sont liées du fait qu'il faille d'abord localiser un visage dans l'image avant de tenter de l'identifier.

De nombreuses méthodes de reconnaissance de visage ont été proposées durant les trois dernières décennies. De ce fait, la littérature sur cette thématique est très vaste et en aborde

des aspects très variés [Zha03] [DelGB08]. Les méthodes de reconnaissance de visage peuvent être regroupées suivant qu'elles opèrent une caractérisation globale, locale ou hybride des visages [Zha03], ce que nous définissons dans ce qui suit :

1. **Méthodes globales** : ces méthodes utilisent la totalité du visage comme entrée du système de reconnaissance. Une des représentations les plus utilisées pour coder le visage est basée sur les *Eigenfaces* [KirS90], qui s'appuie sur l'analyse en composantes principales.

2. **Méthodes locales** : typiquement, dans ces méthodes, les caractéristiques locales telles que les yeux, le nez et la bouche sont extraites et leurs positions ainsi que leurs statistiques locales (issues d'informations géométriques, colorimétriques ou texturales) sont fournies à un classifieur structurel.

3. **Méthodes hybrides** : ici le système de reconnaissance utilise les caractéristiques locales (voir point précédent) et la région globale du visage pour le reconnaître, comme le fait le système de perception humaine. Ces méthodes peuvent potentiellement obtenir de meilleurs taux de reconnaissance que les deux types de méthodes précédentes.

Les informations globales et les caractéristiques locales sont cruciales pour la perception et l'identification des visages [WecPBS98]. Les recherches suggèrent une première phase de descriptions globales suivie d'une phase de perception plus fine basée sur les caractéristiques locales.

3.2.2 Les difficultés inhérentes à la reconnaissance de visage

Pour que les systèmes de reconnaissance de visages soient fiables, ils devraient idéalement rester invariants à tout facteur indépendant de l'identité, même si ce facteur engendre des changements d'apparence du visage. De manière pratique, en plus des difficultés liées à la détection de visage (chapitre 2, section 2.3.3.2), de nombreux facteurs extérieurs au visage ou en lien avec sa nature intrinsèque peuvent influer sur la qualité de la reconnaissance. Les conditions de prise de vue, notamment l'angle sous lequel le visage est observé et les conditions d'éclairage, modifient considérablement l'apparence d'un visage [DelGB08]. Dans [GroSC01], Gross *et al* fournissent une étude systématique de l'impact de différents paramètres sur les performances d'un système de reconnaissance de visage. Les six facteurs considérés sont : la pose de la tête, les changements d'éclairage, l'expression faciale, les occultations, l'intervalle de temps entre deux prises de vue et le sexe (genre). Leurs influences relatives sont étudiées en faisant varier isolément ou conjointement ces paramètres. Nous reprenons ci-dessous les principales conclusions issues de cette étude :

3.2.2.1 Influence des variations de la pose

Les changements d'orientation et les changements de l'angle d'inclinaison du visage engendrent de nombreuses modifications d'apparence dans les images collectées. Une phase préliminaire de normalisation de l'image du visage permet de corriger d'éventuelles rotations dans le plan de celle-ci. Les rotations en profondeur engendrent l'occultation de certaines parties du visage comme pour les vues de trois-quarts. D'autre part, elles amènent des différences de profondeur qui, projetées sur le plan 2D de l'image, provoquent des déformations qui font varier la forme globale du visage. Ces déformations qui correspondent à

l'étirement de certaines parties du visage et la compression d'autres régions font varier aussi les distances entre les caractéristiques faciales.

3.2.2.2 Influence des changements d'éclairage

L'intensité et la direction d'éclairage lors de la prise de vue influent énormément sur l'apparence du visage dans l'image. En effet, dans la plupart des applications courantes, des changements dans les conditions d'éclairage sont inévitables, notamment lorsque les vues sont collectées à des heures différentes, en intérieur ou en extérieur. Etant donnée la forme spécifique d'un visage humain, ces variations d'éclairage peuvent y faire apparaître des ombres accentuant ou masquant certaines caractéristiques faciales.

3.2.2.3 Influence des expressions faciales

Les visages sont des éléments non rigides. Les expressions faciales véhiculant des émotions, combinées avec les déformations induites par la parole, peuvent produire des changements d'apparence importants, et le nombre de configurations possibles est trop important pour que celles-ci soient décrites *in extenso* de façon réaliste. L'influence de l'expression faciale sur la reconnaissance est donc difficile à évaluer avec précision. Cependant, du fait que ce facteur affecte la forme géométrique et les positions des caractéristiques faciales, les techniques globales ou hybrides y sont généralement plus robustes que la plupart des techniques géométriques.

3.2.2.4 Influence des occultations partielles

Le visage peut être partiellement masqué par des objets ou par le port d'accessoires tels que des lunettes, un chapeau, une écharpe, *etc*. Les occultations peuvent être intentionnelles ou non. Dans le contexte de la vidéosurveillance, il peut s'agir d'une volonté délibérée d'empêcher la reconnaissance. Il est clair que la reconnaissance sera d'autant plus difficile que peu d'éléments discriminants seront simultanément visibles.

3.2.3 Conclusion

Bien qu'il s'agisse d'une démarche dont l'utilisation semble s'imposer de façon intuitive, l'utilisation de la reconnaissance faciale à des fins de vidéosurveillance n'est pas sans poser certains défis pratiques. En effet, si les images issues de systèmes à vocation biométrique proposent une qualité adaptée à la tâche de reconnaissance, il en va tout autrement pour les informations provenant de systèmes de vidéosurveillance « généralistes ». De fait, le visage ne représente alors qu'une faible proportion d'une image d'une résolution parfois basse. De plus, l'influence des facteurs que nous avons cités (pose, changements d'éclairage, expression faciale, occultations) se trouvent dans ce contexte augmentée. Quand bien même la région comportant le visage proposerait une quantité d'information en théorie suffisante pour l'opération de classification, encore faut-il être capable de l'extraire du reste du signal image disponible, et ceci en respectant les contraintes de temps de réponse associées à ce type de dispositif. Au final, malgré le fait que la reconnaissance de visage soit développée et largement utilisée depuis trois décennies dans différentes applications, son intégration dans des systèmes de vidéosurveillance reste une tâche assez délicate.

3.3 Reconnaissance de la démarche

3.3.1 Introduction

Parmi les méthodes de reconnaissance biométrique, la reconnaissance de la démarche est une technique relativement nouvelle et vise à reconnaître les personnes par leur façon de marcher. Aussi, l'utilisation de la démarche pour l'identification des personnes dans les applications de surveillance a attiré récemment les chercheurs du domaine de la vision par ordinateur. L'adéquation de la reconnaissance de la démarche pour les systèmes de surveillance vient du fait que celle-ci peut être perçue à distance, d'une manière non invasive. En effet, les séquences d'images de personnes enregistrées par les caméras de surveillance représentent des individus se déplaçant (pour la plupart d'entre eux en marchant) dans l'espace surveillé. De plus, comme nous le verrons, les méthodes de reconnaissance de la démarche sont naturellement assez robustes aux perturbations affectant les méthodes de reconnaissance de visage. Elles n'ont pas non plus spécialement besoin d'images de haute résolution, et peuvent donc être utilisées pour la reconnaissance à distance ou à basse résolution quand le sujet humain occupe peu de pixels dans l'image [NixC04] [WagN04].

Des études médicales ont montré que la démarche de chaque individu était unique, variant de personne à personne et était difficile à déguiser [MurDK64]. D'autres études [Mur67] montrent également que la démarche humaine dispose de 24 composantes différentes et est propre à chaque individu [Gaf07] [LiuJZ09]. Du point de vue de la biomécanique, la structure musculo-squelettique varie d'une personne à une autre et il est possible d'identifier un individu par sa démarche, bien que tous les humains suivent le même schéma de marche de base [Mur67]. En outre, il a été démontré que les démarches sont si caractéristiques que l'on reconnaît les amis de par leur démarche [CutK77] et qu'une démarche permet même de connaître le sexe d'un individu [BarCK78].

Dans des applications de vision par ordinateur, la démarche est capturée à distance par une caméra vidéo. Des techniques de traitement d'images sont utilisées pour extraire les caractéristiques de la démarche pour la reconnaissance. La plupart des algorithmes de reconnaissance de la démarche est basé sur la silhouette des personnes. Aussi, le fond de l'image est supprimée et la silhouette de la personne est extraite et analysée pour la reconnaissance. Dans ce qui suit, nous allons présenter les principales techniques rencontrées.

3.3.2 Les méthodes de reconnaissance de la démarche

On peut diviser les approches de reconnaissance de la démarche en deux catégories, à savoir, les approches avec modèle et les approches sans modèle.

3.3.2.1 Les approches avec modèle

Les approches avec modèle visent à modéliser de façon explicite le corps humain ou les mouvements humains en fonction de connaissances *a priori*. Habituellement, un modèle d'appariement est appliqué dans chaque trame d'une séquence de marche afin de mesurer les paramètres physiques de la démarche tels que les trajectoires d'éléments de référence, la longueur des membres et les vitesses angulaires [YamNC04] [LeeE06]. Certaines approches tirent directement les paramètres structurels de la séquence de la démarche [WanNTH04].

Les approches avec modèle peuvent utiliser des caractéristiques statiques, dynamiques ou les deux. Les caractéristiques statiques reflètent les mesures basées sur la géométrie de la

structure anatomique du corps humain telles que la taille de la personne et la longueur ou la largeur des différents segments du corps. Les caractéristiques statiques peuvent aussi être dérivées de l'allure observée telles que la longueur de la foulée [JohB01] [LeeG02]. Les caractéristiques dynamiques sont les indices qui décrivent la cinématique du processus de locomotion, tels que le mouvement angulaire des membres inférieurs extraits des informations des trajectoires des articulations [BouN07]. Il existe également des méthodes qui utilisent conjointement des caractéristiques statiques et dynamiques afin d'augmenter les taux de reconnaissance [WagN04] [LeeE06].

Les caractéristiques statiques sont moins difficiles à calculer, mais malgré que des taux de reconnaissance élevés aient été obtenus sous certaines conditions, il a été clairement indiqué que les paramètres liés au corps ne sont pas robustes car ils sont fortement influencés par les vêtements portés. En général, ces méthodes tendent à être complexes et impliquent des coûts de calculs élevés. En outre, les approches avec modèle sont sensibles au bruit de l'image, aux auto-occultations, aux ombres et aux changements de vue, ce qui conduit à des performances inférieures sur les ensembles de bases de données publiques de démarche.

3.3.2.2 Les approches sans modèle

La plupart des recherches sur la reconnaissance de la démarche utilise la reconnaissance sans modèle (dite holistique). Les approches sans modèle n'ont, par nature, pas besoin de connaissance préalable d'un modèle de la démarche, mais utilisent directement les informations de mouvement. Ces approches caractérisent généralement la distribution spatio-temporelle générée par le mouvement de la marche dans son continuum. Les signatures spatio-temporelles de démarche utilisent des données de la séquence d'images de la marche telles que les images en niveaux de gris, les images binaires de silhouettes ou des images de flux optique pour caractériser le mouvement de marche dans un espace 3D (XYT). Il existe un nombre important d'approches sans modèle proposées dans la littérature [TurP91] [CheLZ09] [YuTH09].

Parmi celles-ci, les approches de représentation de la démarche telles que *Gait Energy Image* (*GEI*) et *Motion Silhouettes Image* (*MSI*) capturent uniquement les informations de mouvement et négligent les informations d'égales importances, mais moins fiables, que sont les directions des mouvements relatifs. Ils sacrifient ainsi un pouvoir discriminant pour gagner en robustesse. A l'inverse, dans [BasXG09], les auteurs proposent une représentation de la démarche basée sur les champs de flux optique calculés à partir d'images de personnes normalisées centrées sur un cycle de marche complet. Dans leur représentation, l'intensité de mouvement et l'information de direction de mouvement sont capturées dans un ensemble de descripteurs de mouvement. Pour obtenir une robustesse au bruit, plutôt que de s'appuyer sur la valeur exacte des vecteurs de flux, la direction du flux est quantifiée et une représentation de direction basée sur des histogrammes est utilisée. Par rapport aux représentations de démarche sans modèle existantes, leur représentation est moins sensible aux modifications de certains paramètres tels que les vêtements, les chaussures, les objets portés ainsi que la vitesse de déplacement.
Quelle que soit l'approche retenue, la reconnaissance de la démarche est sous l'influence de facteurs qui peuvent en limiter l'efficacité. Les difficultés qui en résultent sont évoquées ci-dessous.

3.3.3 Les difficultés liées à la reconnaissance de la démarche

Bien que les résultats des approches de reconnaissance de personnes par leur démarche soient encourageants, il y a plusieurs facteurs qui influent négativement sur la performance de telles approches et qui sont :

- **Les conditions de prise de vue :** les conditions d'éclairage telles que l'intensité et la direction de la lumière, l'heure de la journée (jour/nuit) et les environnements intérieurs ou extérieurs influencent l'aspect des images acquises et donc l'apparence des personnes.
- **La nature du sol :** l'inclinaison du sol (on ne marche pas de la même manière dans une montée, une descente ou sur un sol plat), et la nature du sol (dur, mou, gazon, béton, escaliers, *etc.*) influencent considérablement la démarche.
- **Les vêtements :** ces derniers peuvent ou non constituer une gêne au mouvement et par conséquent influencer la manière de marcher. Qui plus est, les vêtements peuvent contrarier la détection des primitives utilisées pour la reconnaissance de la démarche.
- **Le type de chaussures portées :** en fonction du confort qu'elles procurent (ou non !), elles peuvent faire changer la manière dont on marche.
- **Les objets portés :** porter un cartable, un sac à dos, une valise, *etc.*, influence non seulement la démarche, mais peut également affecter les primitives visuelles utilisées.
- **La vitesse de marche :** la démarche n'est pas la même lorsqu'on marche doucement et lorsqu'on marche vite.
- **La présence de handicaps temporaires ou permanents :** des blessures au pied, des troubles des membres inférieurs, une maladie de Parkinson, *etc.*, peuvent également changer la démarche.
- **Changements physiologiques :** la vieillesse, la grossesse, un gain ou une perte de poids changent également la manière de marcher.

3.3.4 Conclusion

Malgré des résultats prometteurs en milieux maitrisés, la reconnaissance de la démarche n'est pas encore assez mature pour être déployée dans des systèmes de vidéosurveillance. Néanmoins, le fait que les séquences vidéo puissent être acquises à distance ainsi que sa nature non invasive présentent de grands avantages.

3.4 Reconnaissance basée sur l'apparence

3.4.1 Introduction

Comparativement à la reconnaissance biométrique, notamment à la reconnaissance de visage, la reconnaissance basée sur l'apparence a été peu développée. En effet, on ne trouve dans la littérature qu'un nombre limité de travaux dédiés à ce sujet. La plupart des méthodes de reconnaissance basée sur l'apparence suivent globalement le même schéma. Elles abordent le problème en s'appuyant sur la reconnaissance du corps entier des individus, basée sur l'usage de caractéristiques de forme, de couleur ou de texture déduites de leur apparence extérieure. On peut classer les méthodes existantes en deux catégories :

- **Les méthodes avec apprentissage hors ligne** : ici, une base d'apprentissage est construite en apprenant au préalable l'apparence de tous les individus devant être reconnus plus tard. La base d'apprentissage contiendra donc un modèle d'apparence de

chaque individu. Puis, pour reconnaitre un individu lors de la phase de reconnaissance, son modèle d'apparence est construit et comparé par le classifieur à chaque modèle de la base d'apprentissage.

- **Les méthodes avec apprentissage en ligne** : ici, aucune base d'apprentissage n'est construite au préalable. Une utilisation typique est celle correspondant au problème du suivi de personnes dans des séquences vidéo. Dans ce type d'application, un vecteur caractéristique (modèle) par personne est constitué à la volée sur chaque trame de la séquence, puis une distance est calculée entre ce modèle et celui de chaque personne présente dans la trame précédente. On apparie entre elles les personnes dont les vecteurs de caractéristiques sont les plus proches au sens de la métrique utilisée.

Nous détaillons ci-dessous des méthodes de reconnaissance basée sur l'apparence avec apprentissage (respectivement hors ligne et en ligne) qui nous semblent les plus significatives dans le contexte de la vidéosurveillance. Il est à noter, cependant, que toutes les évaluations rapportées se sont faites sur des séquences d'images récoltées en environnements très maîtrisés et qu'il convient d'être prudent quant à la facilité d'application de ces méthodes en utilisation extérieur et en environnement non maîtrisé.

3.4.2 Méthodes de reconnaissance avec apprentissage hors ligne

Les méthodes de reconnaissance avec apprentissage hors ligne opèrent en deux temps : durant une phase d'apprentissage, un modèle de chaque personne à reconnaitre est créé. Puis, en se basant sur ces modèles, la reconnaissance de ces mêmes personnes est effectuée durant une phase de classification. Concernant cette dernière, toutes ces méthodes suivent les mêmes étapes :

1) une étape de « filtrage », consistant à isoler le signal utile du reste. En pratique, ce filtrage prend très souvent la forme d'une opération de soustraction de fond qui permet de détecter et d'isoler les personnes du reste de l'image,
2) une étape d'extraction de caractéristiques de forme, de couleur, de texture ou une combinaison de celles-ci,
3) une étape de classification utilisant les modèles appris lors de la phase d'apprentissage.

Aussi, la différence entre ces méthodes réside principalement dans les caractéristiques extraites et les algorithmes de classification utilisés. Au niveau des différentes approches proposées, on notera également qu'il n'y a qu'une seule personne à la fois présente dans l'image lors des phases d'apprentissage et de classification. La **Figure 3.1** illustre ce processus de traitement associé aux étapes de reconnaissance avec apprentissage hors ligne :

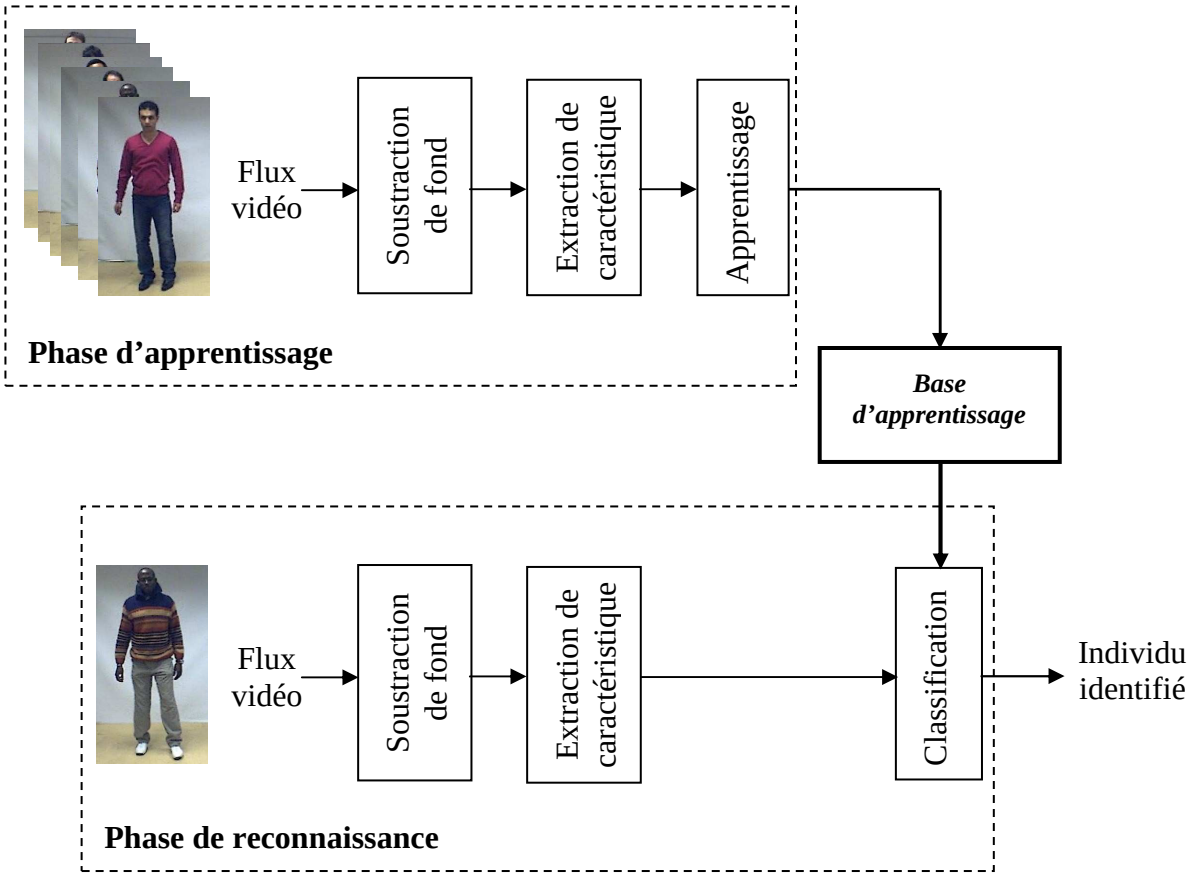


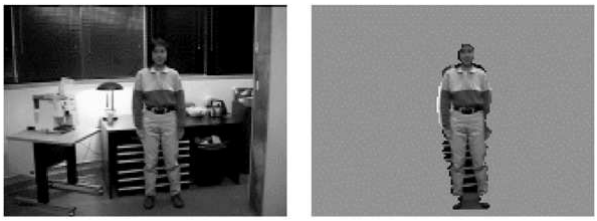
Figure 3.1 : Processus de traitement des étapes de reconnaissance de personnes avec apprentissage hors ligne.

A des fins de référence, nous détaillons ci-dessous trois des algorithmes de reconnaissance basées sur l'apparence avec apprentissage hors ligne les plus adaptées à notre cas d'application.

3.4.2.1 Algorithme de Nakajima et al.

Nakajima et al. [NakPHP03] ont proposé un système qui apprend, à partir d'exemples, à reconnaitre des personnes ainsi que quatre de leurs postures (de face, profil droit, profil gauche, de dos) dans des séquences d'images acquises en intérieur. L'apparence entière des individus est décrite par des caractéristiques de forme et de couleur, tandis que la reconnaissance est effectuée par des combinaisons de classifieurs SVM. Différentes stratégies multi-classes SVM sont alors utilisées et comparées à un classifieur du *k*-plus-proches-voisins (ou kNN pour *k*-Nearest-Neighbor).

Conformément au schéma général de la figure 3.1, la première opération effectuée est la détection de « formes candidates » en utilisant la soustraction de fond. Puis, pour s'assurer que tout ou partie de ces formes correspondent effectivement à des personnes, le système extrait la silhouette de la personne potentielle en utilisant une détection de contour. Présumant qu'une personne se déplace lentement entre deux images successives, le système réalise une détection de contours sur l'image obtenue par la soustraction de deux images successives. Si le nombre des pixels du contour est supérieur à un seuil, l'image est retenue et la personne est segmentée. La **Figure 3.2** illustre un exemple d'image et le résultat de la détection.



(a) (b)

Figure 3.2 : Exemple de détection de personne de l'algorithme de Nakajima et al. [NakPHP03].

Une fois la personne détectée et isolée du reste de l'image, différentes caractéristiques sont extraites :

- **Caractéristiques couleur :**

  - ***Histogramme couleur RGB*** : un histogramme permet de représenter la distribution des intensités (ou des couleurs) des pixels d'une image ou d'une région de l'image. Il est défini comme une fonction discrète qui associe à chaque valeur d'intensité le nombre de pixels prenant cette valeur. La détermination de l'histogramme est réalisée en comptant le nombre de pixels pour chaque intensité. Ainsi, l'histogramme d'une image en 256 niveaux de gris pourrait être représenté par un vecteur possédant 256 valeurs. Cependant, on effectue parfois une quantification qui regroupe plusieurs valeurs d'intensité en une seule classe (appelée bin). Les histogrammes sont intrinsèquement invariants en rotation et en translation. Ils peuvent également être normalisés, en divisant la valeur de chaque classe (encore appelée « bin ») par le nombre total de pixels. La valeur d'une classe varie alors entre 0 et 1, et peut s'interpréter comme une estimation de la probabilité d'occurrence de la classe.

    Ici, le système extrait un histogramme de 32 bins de l'image du corps supposé (blob) sur chaque canal couleur (*R, G, B*). La taille du vecteur est alors de $3 \times 32 = 96$ caractéristiques sur chaque forme candidate de l'image.

  - ***Histogramme chrominance rg*** : le système extrait un histogramme sur chaque canal (r, g) dans l'espace rg (espace RGB normalisé), avec : r = R/(R+G+B) et g = G/(R+G+B). Cet histogramme extrait uniquement des informations couleur (ou chrominance) sans prendre en compte des informations d'intensité. Ici encore, chaque canal est échantillonné sur 32 bins, ce qui ajoute 64 composantes au vecteur de caractéristiques.

- **Caractéristiques de forme :**

  - ***Histogramme (Pseudo-histogramme) de forme*** : ici, le système extrait de simples caractéristiques de forme des personnes en calculant les pixels sur les lignes et les colonnes du blob. Les auteurs ont choisi une résolution de 10 bins pour les histogrammes colonnes et de 30 bins pour les histogrammes lignes.

  - ***Caractéristiques de forme locale*** : elles sont obtenues en convoluant l'image avec les signatures de forme de la **Figure 3.3**. Ces signatures ont été introduites dans

[KurHM98] pour la détection de personnes avec posture invariante. Deux différents types d'opérations de convolution sont considérés ici, une convolution linéaire et une convolution non-linéaire. Les caractéristiques de forme sont extraites ici pour chacun des plans couleurs suivant : (*R*+*G*-*B*), (*R*-*G*), et (*R*+*G*). Le système extrait en tout 75 (25×3) caractéristiques de ces trois canaux couleur.

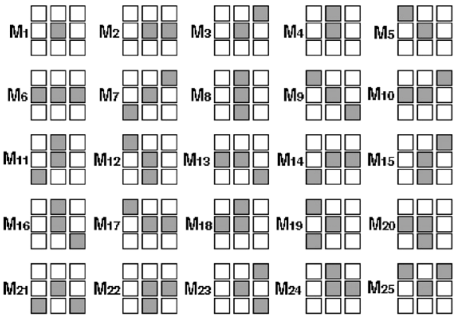


Figure 3.3 : Exemples de signatures de forme [NakPHP03].

• **<u>La classification :</u>**

Comme évoqué lors de la présentation de cette méthode, les auteurs utilisent différentes stratégies multi-classes SVM ainsi qu'un classifieur *k* plus-proches-voisins (*k*NN) et comparent les résultats de ces classifieurs.

- ***Séparateurs à vaste marge*** (***SVM***) : ont été présentés dans le chapitre 2 (section 2.5.2). Bien qu'initialement les SVM aient été développés pour traiter des problèmes binaires (séparation de deux classes), ils peuvent être étendus à des problèmes multi-classes en exploitant des stratégies utilisant des combinaisons de classifieurs binaires (chapitre 2, section 1.5.2.1). Il existe principalement deux types d'approches de classification multi-classes utilisant des combinaisons de classifieurs binaires : le premier est appelé *un contre un* et le second est appelé *un contre tous*. Dans le premier type, afin de séparer *q* classes, *q*(*q* - 1)/2 SVM sont appris, chacun séparant deux classes [PonV98] [ClaM99]. Dans le deuxième type, *q* SVM sont appris, chacun séparant une classe de toutes les autres [CorV95][SchBV95].

  Ici, les auteurs ont utilisé trois stratégies de classification multi-classes. La première est de type *un contre un*, le deuxième de type *arbre de décision bottom-up* (*un contre un*), et le troisième de type *graphe de décision top-down* (*un contre un*).

- ***k plus-proches-voisins*** (***k*NN**) : fait partie des méthodes de reconnaissance de forme non-paramétriques. C'est une méthode assez intuitive qui classifie des données non-étiquetées en se basant sur leur similarité aux données d'apprentissage. Le seul outil à définir est une distance entre les éléments que l'on veut classifier, partant de la simple distance usuelle (euclidienne), allant jusqu'à des mesures plus élaborées. Disposant d'une base d'apprentissage où chaque élément connu est affecté à une classe, dès qu'il y a un nouvel élément à classer, on calcule sa distance à tous les éléments de la base. Si cette base comporte *N* éléments, alors on calcule *N* distances et on obtient donc *N* nombres réels. On cherche alors les *k* plus petits nombres parmi ces *N* nombres. Ces *k* nombres correspondent donc aux *k* éléments de la base qui sont les plus proches de

l'élément que l'on souhaite classifier. On décide d'attribuer à l'élément à classifier la classe majoritaire parmi ces *k* éléments.

- **Evaluation :**

Les séquences vidéo utilisées pour évaluer le système de Nakajima et al. [NakPHP03] ont été prises par une caméra placée dans une pièce en face d'une machine à café à une distance de 4,5 mètres. La caméra filmait des membres du laboratoire venant prendre un café à la machine. L'éclairage et le fond de l'image sont invariants, et les personnes filmées portaient les mêmes vêtements durant une même journée. La **Figure 3.4** présente des exemples d'images de la base d'apprentissage.

Figure 3.4 : Exemples d'images de la base d'apprentissage [NakPHP03].

La première série de tests a été effectué sur 4 individus. 160 images ont été utilisées (40 par individu) pour faire l'apprentissage des différents classifieurs (3 classifieurs à base de SVM : un-contre-tous, top-down et bottom-up, et 3 classifieurs à base de kNN : pour *k*=1, *k*=3 et *k*=5) et sur 418 images de test. Le **Tableau 3.1** présente les taux de reconnaissance (en %) obtenus.

Vecteur caractéristiques	**SVM**			***k*NN**		
	Top-down	*Bottom-up*	*Un-contre-tous*	$k = 1$	$k = 3$	$k = 5$
Hist. RGB	99.5	99.2	99.5	99.0	98.7	98.5
Hist. rg	100	100	100	100	100	100
Hist. Forme	91.4	91.6	96.2	94.7	94.4	94.1
Cars. Forme locale	99.5	99.5	97.5	88.3	85.0	84.4

Tableau 3.1 : Taux de reconnaissance (%) sur 4 individus [NakPHP03].

On voit ici que les meilleurs taux de reconnaissance ont été obtenus avec les caractéristiques couleurs RGB normalisées (dimension 1024). Les résultats obtenus par les trois classifieurs SVM sont très similaires et sont légèrement supérieurs à ceux obtenus par les classifieurs *k*NN. On note que les taux de reconnaissance baissent de manière significative lorsqu'on ajoute des caractéristiques de forme. Ceci est dû au fait que les mouvements des personnes et de leurs bras et la variation de la distance entre eux et la caméra conduisent à des changements considérables de la forme et de la taille des personnes dans l'image.

Une deuxième évaluation a été menée en filmant 8 personnes dans les mêmes conditions mais sur une période de 16 jours. 1127 images ont été utilisées, et seules les caractéristiques couleurs normalisées ont été exploitées du fait qu'elles ont obtenu les meilleurs résultats précédemment. Les auteurs ont effectué ici cinq séries d'expérimentations où le système a été

entrainé différemment à reconnaitre ces 8 individus. Les quatre premières séries d'expérimentation ont été menées en utilisant 90%, 80%, 50% et 20% des 1127 images pour faire l'apprentissage du système (et donc 10%, 20, 50% et 80% respectivement des 1127 images pour les tests). La cinquième série a été effectuée en utilisant les images acquises lors des 15 premiers jours pour faire l'apprentissage du système, et les images acquise lors du 16ème jour comme images de test (ces personnes changeaient de vêtements tous les jours aléatoirement). Les taux de reconnaissance obtenus sont présentés dans le **Tableau 3.2**.

	Images Apprentissage % : Test%	90 : 10	80 : 20	50 : 50	20 : 80	Image test du 16ème jour
SVM	***Top-down***	92.3	91.2	90.5	73.3	45.9
	Bottom-up	90.6	91.7	90.6	66.1	45.9
	Un-contre-tous	87.2	90.6	85.9	84.6	49.2
	Un-contre-tous (Pol.)	98.3	96.4	94.7	88.1	52.9
***k*NN**	***k* = 1**	92.9	92.0	92.7	85.1	53.3
	***k* = 3**	92.9	92.0	92.2	81.3	50.0
	***k* = 5**	94.7	91.0	90.1	76.0	50.8

Tableau 3.2 : Taux de reconnaissance (%) sur 8 individus dans une période de 16 jours.

On peut voir ici que les taux de reconnaissance obtenus sont élevés lorsqu'ils utilisent des images provenant des 16 jours pour faire à la fois l'apprentissage et les tests. Puis les taux baissent dans certains cas au dessous des 50% quand le système est testé sur les images provenant du 16ème jour. Ceci s'explique par le fait que les personnes portent différents vêtements tous les jours, et le système ne peut donc les reconnaitre dans ces conditions.

Cette étude montre deux choses : la première est que des caractéristiques « couleur » sont efficaces pour la reconnaissance et donnent de meilleurs résultats que les caractéristiques de forme et de taille qui ont été testées ici. En effet, la forme et la taille d'une personne dans des images changent considérablement en fonction des mouvements du corps et/ou par rapport à sa position et sa distance par rapport à la caméra. La deuxième chose est qu'en se basant sur des caractéristiques couleurs pour reconnaitre une personne, on reconnait en réalité ses vêtements, ce qui semble intuitivement évident. Par conséquent, un système de reconnaissance basé sur la couleur ne peut reconnaitre une personne que si elle ne change pas de vêtements.
Une des possibilités que n'ont pas exploitées les auteurs de cette méthode est l'utilisation de la texture, qui est un outil adapté à la description des éléments tels que les rayures ou les « carreaux » qui peuvent décorer un vêtement. Un tel outil peut mettre en évidence des différences qui échapperaient à une simple analyse des distributions colorimétriques. L'algorithme présenté ci-après introduit l'utilisation des attributs texturaux.

3.4.2.2 Algorithme de Hahnel et al.

Hahnel et al. [HahKK04] ont proposé une méthode de reconnaissance de personnes avec apprentissage hors ligne. Leur méthode commence par segmenter le blob correspondant au corps-entier de l'individu du reste de l'image, puis à extraire ses caractéristiques couleur et texture. La classification est ensuite effectuée avec un classifieur à réseau de neurones RBF (classifieur à base de fonctions radiales) ainsi qu'avec un classifieur du *k* plus-proches-voisins (*k*NN).

- **Caractéristiques couleur :**

  - ***Histogramme RGBL*** : ici, le système extrait un histogramme sur chaque canal couleur (*R*, *G*, *B*), et sur le canal de luminance *L* qui s'est exprimé par : $L = (\min(R,G,B) + \max(R,G,B))/2$. Comme indiqué précédemment, les histogrammes sont invariants en rotation et en translation. Tous les histogrammes sont normalisés pour être invariants à l'échelle, puis passent par un filtre gaussien de taille 3×3 pour diminuer le bruit de l'image.

  - ***Histogramme chrominance rg*** : le système extrait un histogramme de 12 bins sur chaque canal (r, g) dans l'espace rg (espace RGB normalisé). Cet histogramme est normalisé puis filtré par un filtre passe-bas.

  - ***Descripteur couleur-structure*** (***Color Structure Descriptor*** ou ***CSD***) : Il fait partie des descripteurs visuels définis par le standard MPEG-7 (officiellement appelé *Multimedia Content Description Interface*) afin de caractériser la couleur, la texture, la forme et le mouvement. Ce descripteur est couramment utilisé dans le domaine de l'indexation de flux vidéo par le contenu [MesVE01]. Le CSD est un histogramme couleur enrichi par l'introduction d'informations décrivant la distribution spatiale de la couleur. Il code la structure locale de celle-ci en utilisant un élément structurant, représenté par un masque binaire, de dimension (8x8) comme le montre la **Figure 3.5**. Le CSD est obtenu en translatant l'élément structurant à chaque pixel de l'image. Ensuite, la fréquence d'occurrence des couleurs dans chaque élément structurant est représentée par quatre possibilités de quantification dans l'espace de couleur HMMD (décrit ci-après): 256, 128, 64 et 32 bins [MesVE01].

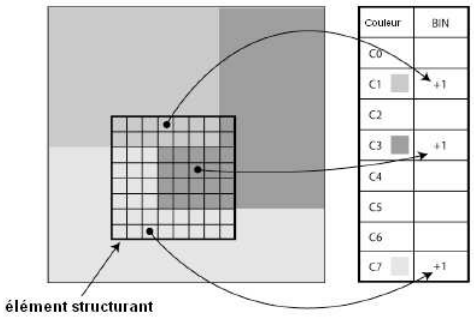


Figure 3.5 : Représentation de l'élément structurant pour le calcul du descripteur CSD [MesVE01]

Bien que le CSD puisse être appliqué sur n'importe quel espace couleur, le standard MPEG-7 suggère cependant l'utilisation de l'espace non-linéaire HMMD, dérivé de l'espace HSV. L'espace HMMD regroupe la teinte (Hue), le maximum (Max) et le minimum (Min) des valeurs RGB ainsi que la différence (Différence) entre les valeurs Max et Min. Ici, les auteurs ont utilisé les valeurs de différence et de somme, soit (Min+Max)/2, à la place du maximum et du minimum.

- **Caractéristiques texture :**

Les caractéristiques de texture utilisées ici sont extraites sur l'image en niveaux de gris et sont :

- ***Descripteur de texture homogène*** (***Homogeneous Texture Descriptor*** ou ***HTD***) : Il décrit la texture d'une région en utilisant des statistiques de fréquences locales. Ce descripteur est fondé sur des mesures énergétiques dans des sous-bandes du domaine fréquentiel correspondant à un banc de filtres de Gabor [ManWNS00] [Chu92]. Dans la spécification MPEG-7, cinq échelles et six orientations sont considérées, ce qui correspond à un total de 30 sous-bandes spectrales [OlmCKK02].

- ***Descripteur d'histogramme de contour*** (***Edge Histogram Descriptor*** ou ***EHD***) : il exprime la distribution locale des contours dans l'image. Un histogramme de contour représente la fréquence et l'orientation des changements de luminosité dans l'image, essentiellement sur 5 types de bords (horizontaux, verticaux, diagonaux, orientés à 45°, à 135° et non-directionnels).

- ***Filtres miroir quadratiques*** (***Quadrature Mirror Filters*** (***QMF***)) : Ceux-ci consistent en un filtre passe-bas et un filtre passe-haut complémentaires. Le résultat des filtrages successifs et des post-traitements est une image d'énergie sur laquelle un histogramme est construit et utilisé comme vecteur de caractéristiques.

- ***Dérivées gaussiennes orientées*** (***Oriented Gaussian Derivatives*** (***OGD***)) : pour obtenir des OGD, l'image passe par une succession de filtres et de fonctions d'interpolation. Le résultat est un vecteur de caractéristiques invariant en rotation.

• **La classification :**

Comme nous l'avons précisé, les auteurs utilisent ici un classifieur du type réseau de neurones (Neural Network RBF) ainsi qu'un classifieur du *k* plus-proches-voisins. Les résultats obtenus par l'un et l'autre sont comparés.

- ***Réseau de Neurones RBF*** : le réseau RBF (Radial Basis Functions) fait partie des réseaux de neurones supervisés [Bro88] [Vog93]. Il est constitué de trois couches : une couche d'entrée qui retransmet les entrées sans distorsion, une seule couche cachée qui contient les neurones RBF qui sont généralement des gaussiennes et une couche de sortie dont les neurones sont généralement animés par une fonction d'activation linéaire. Chaque couche est complètement connectée à la suivante et il n'y a pas de connexions à l'intérieur d'une même couche. Cet algorithme repose sur le fait que toute fonction peut être approchée d'aussi près que l'on veut par une combinaison linéaire de fonctions gaussiennes judicieusement choisies.

• **Evaluation :**

Pour l'évaluation de leur méthode de reconnaissance, les auteurs ont enregistré une base de données d'individus se déplaçant sur un fond simple en face d'une caméra fixe. 24 individus ont été utilisés dont certains portant différents vêtements. Ils supposent ainsi avoir 53 individus d'apparence différente dans la base de données. Chaque individu a été appris avec 10 images et testé sur 200 images, aléatoirement choisies dans la base de données. La **Figure 3.6** donne des exemples d'images de la base d'apprentissage.

Figure 3.6 : Exemples d'images de la base d'apprentissage [HahKK04].

La première série de tests a été menée pour comparer la pertinence des différents vecteurs caractéristiques utilisés pour la reconnaissance. Les auteurs ont exploité ici uniquement le classifieur RBF. Le **Tableau 3.3** présente les taux de reconnaissance (en %) obtenus sur 10, 20, 30, 40 et 53 individus.

		Nombre d'individus				
		10	20	30	40	53
Caractéristiques Couleur	**Hist. RGBL**	96.0	94.6	94.4	94.1	91.3
	Hist. Chrominance	96.7	93.4	93.6	90.1	86.2
	CSD	99.6	98.5	98.3	97.8	95.8
Caractéristiques Texture	**QMF**	47.3	40.2	31.3	28.6	24.5
	OGD	88.7	87.2	81.3	79.1	76.6
	EHD	53.4	41.8	39.1	36.0	31.3
	HTD	80.9	80.6	76.6	73.3	66.3

Tableau 3.3 : Taux de reconnaissance(%) utilisant un classifieur **RBF**.

On voit ici que, prises seules, les caractéristiques « couleur » ont de meilleures performances que les caractéristiques « texture ». Les taux de reconnaissance les plus élevés sont obtenu avec le CSD, suivi de près par les histogrammes RGBL et les histogrammes de chrominance (95.8%, 91.3% et 86.2% respectivement, sur 53 individus). Les taux obtenus avec les caractéristiques texture OGD et HTD sont assez satisfaisants sur 10 individus (88.7% et 80.9% respectivement), et baissent jusqu'à 76.6% et 66.3% respectivement, sur 53 individus.

Les auteurs ont tenté d'améliorer les taux de reconnaissance en combinant des informations couleur et texture. Le **Tableau 3.4** présente les taux de reconnaissance obtenus avec le CSD (couleur), la HTD (texture) et par la combinaison des deux (couleur + texture). On peut y voir que l'utilisation conjointe des informations couleur et texture améliore quelque peu les taux de reconnaissance.

	Nombre d'individus				
Caractéristiques	10	20	30	40	53
CSD	99.6	98.5	98.3	97.8	95.8
HTD	80.9	80.6	76.6	73.3	66.3
CSD + HTD	99.7	99.1	99.3	99.5	97.4

Tableau 3.4 : Taux de reconnaissance avec la couleur, et la texture et la combinaison des deux.

Enfin, les auteurs ont comparé les performances du classifieur RBF avec celles du classifieur *k*NN. Le **Tableau 3.5** présente les taux de reconnaissance obtenus par ces deux classifieurs en

utilisant respectivement le CSD, l'histogramme de chrominance et la HTD. On peut voir que les performances du classifieur RBF sont légèrement supérieures à celles basées sur le classifieur kNN.

	Nombre d'individus				
Caractéristiques	10	20	30	40	53
CSD (RBF)	99.6	98.5	98.3	97.8	95.8
CSD (kNN)	99.6	91.3	93.2	95.9	92.8
Hist. Chrominance (RBF)	96.7	93.4	93.6	90.1	86.2
Hist. chrominance (kNN)	94.1	90.6	90.3	86.9	81.0
HTD (RBF)	80.9	80.6	76.6	73.3	66.3
HTD (kNN)	71.3	67.1	63.5	60.3	54.1

Tableau 3.5 : Taux de reconnaissance obtenus avec les classifieurs RBF et kNN avec les caractéristiques sélectionnées.

Cette étude a permis de mettre en évidence plusieurs choses. La première est que, prises isolément ; les caractéristiques « couleur » sont plus performantes pour la reconnaissance des personnes que les caractéristiques texture, même si ces dernières obtiennent des résultats assez satisfaisants. Ceci peut s'expliquer notamment en raison de la faible résolution des images ainsi que du manque de motifs de texture sur les vêtements des individus. Rien ne garantit que les bases d'apprentissage et de test utilisées permettent de tirer des conclusions d'ordre général sur la pertinence comparée de la couleur et de la texture à des fins de reconnaissance.

La deuxième chose est que le descripteur CSD est légèrement plus performant que les histogrammes couleur. Néanmoins, il faut noter que le calcul des caractéristiques définies par le standard MPEG-7 (dont le descripteur CSD) est plus couteux sur le plan des calculs que la détermination de caractéristiques plus simples comme les histogrammes. Donc, si on prend en considération cette complexité algorithmique, les histogrammes couleurs sont plus appropriés pour être utilisés dans des applications de reconnaissance nécessitant le respect de contraintes temps réel. Par ailleurs, cette étude a également pu montrer que l'utilisation conjointe de la couleur et de la texture améliore les performances du système de reconnaissance. Enfin, il a été constaté que les taux de reconnaissance diminuaient à mesure que le nombre d'individus à reconnaitre dans la base de données augmentait.

Dans le cadre de l'approche proposée, on constate que si le vecteur de caractéristiques utilisé est assez « complet » (au sens où il intègre des attributs à la fois géométriques, colorimétriques et texturaux), les méthodes de classification mises en œuvre sont, somme toute, relativement simples. La question se pose alors de savoir si, combinées à des vecteurs de caractéristiques reprenant une quantité d'information comparable à celle utilisée ici, des classifieurs plus élaborés ne donneraient pas de meilleurs résultats. L'exemple de l'algorithme de Goldmann (ci-dessous) permet d'apporter certains éléments de réponse.

3.4.2.3 Algorithme Goldmann et al.

Goldmann et al. [GolKMS06] ont également proposé une méthode de reconnaissance de personnes avec apprentissage hors ligne. Leur système détecte et segmente les personnes dans les séquences d'images acquises par une caméra fixe et sur un fond invariant puis, comme dans les méthodes précédemment décrites, en extrait des caractéristiques « couleur » et

« texture ». Pour la reconnaissance, les auteurs utilisent ici encore un classifieur kNN (à des fins de référence), mais aussi un classifieur SVM et un classifieur à base de Modèle de Mélange de Gaussiennes (Gaussian mixture model ou GMM). Là aussi, les performances obtenues permettent de dresser un comparatif.

Pour l'extraction des formes candidates au sein des séquences d'image, les auteurs utilisent une méthode de soustraction de fond s'appuyant sur les travaux de Horprasert [HorH99]. Cette méthode utilise un modèle de fond basé-pixel incluant des informations de luminance et de chrominance et qui est appris à l'avance à partir de plusieurs images de fond. Chaque pixel est classifié comme faisant partie d'un nouvel élément en fonction de changements de luminance et de chrominance. La classification entre les pixels du fond et les autres donne lieu à une image binaire qui est ensuite traitée par des opérations de morphologie pour constituer des blobs. L'application à ces blobs de règles heuristiques basées sur des critères de taille et de forme permet d'isoler parmi ceux d'entre eux qui peuvent être « raisonnablement » associés à un individu. Une fois un de ces « individus » détecté, différentes caractéristiques couleur et texture sont extraites, à savoir :

- **Caractéristiques « couleur » :**

  - ***Descripteur de structure de couleur*** (***CSD***) : strictement identique à celui utilisé par Hahnel *et al* (décrit dans ce qui précède).

  - ***Valeur moyenne RGB*** : consiste à calculer la valeur moyenne des pixels qui composent le blob sur chaque canal couleur R, G et B. La moyenne RGB est intrinsèquement invariante en rotation, translation et en échelle.

  - ***Histogramme couleur HMMD*** : identique à celui utilisé par Hahnel *et al*. (décrit dans ce qui précède, section 3.4.2.2). Ici l'image RGB est transformée vers l'espace HMMD (voir algorithme de Hahnel *et al*. [HahKK04]) puis un histogramme en 32 bins est calculé sur chacun des 4 canaux de cet espace. Cet histogramme est finalement normalisé.

- **Caractéristiques « texture » :**

  - ***Caractéristiques basées sur les histogrammes d'intensité*** (***Intensity Histogram-based features*** ou ***IH***) : l'histogramme d'intensité est un histogramme calculé sur l'image en niveaux de gris. Puisque les pixels sont considérés comme indépendants, cet histogramme contient les informations statistiques du premier ordre du blob. Aussi, différentes caractéristiques (statistiques) peuvent être calculées par cet histogramme pour décrire la texture. Ici, l'algorithme calcule la moyenne, la variance et l'énergie de l'histogramme.

  - ***Caractéristiques basées sur la matrice de cooccurence*** (***Cooccurence Matrix-based features*** ou ***CM***)) : ces caractéristiques se calculent à partir des indices du second ordre dérivés de la matrice de cooccurrence. Pour une image en niveaux de gris, une matrice de cooccurrence est une matrice carrée dont la dimension est égale au nombre des niveaux d'intensité (niveaux de gris) pris en considération et ses coefficients représentent les valeurs de probabilité jointe $p_{d,\theta}(i,j)$ d'une paire de pixels avec les valeurs d'intensité respectives I_i et I_j pour différents angles θ et distances d et une

partition donnée $\bigcup_{k=1}^{N} I_k$ de l'ensemble des niveaux d'intensité possibles. Haralick [Har79] a définit 14 coefficients (paramètres ou descripteurs) qui peuvent être calculés à partir de la matrice de cooccurrence afin de décrire la texture d'une image. Dépendant de la nature de l'image, certains de ces coefficients peuvent avoir des valeurs indéfinies ou alors contenir très peu d'informations. Dans cet algorithme, parmi les caractéristiques extraites, seuls l'énergie, la valeur absolue et le contraste sont utilisés.

- **La classification :**

Pour la classification, les auteurs ont utilisé et comparé trois différents classifieurs : *k*NN, SVM (identiques à ceux déjà décrits) ainsi que des modèles de mélanges de gaussiennes (Gaussian Mixture Models (GMM)).

- **Modèles de mélanges de gaussiennes (Gaussian Mixture Models (GMM))** : L'algorithme GMM fait partie des méthodes de classification paramétriques [McLP00]. Il modélise la fonction de densité de probabilité d'une classe en utilisant une combinaison linéaire de plusieurs distributions gaussiennes. Les paramètres des distributions gaussiennes et les probabilités antérieures utilisées pour la combinaison linéaire sont calculés au cours de la phase d'apprentissage en se basant sur l'algorithme de maximisation de l'espérance (Expectation Maximisation algorithm ou EM). Pour un problème multi-classe, un GMM est appris pour chaque classe c_i. Pour une donnée x à étiqueter, les probabilités $p(x|c_i)$ sont calculées pour chaque classe connue et la décision est prise en utilisant le critère du maximum *a posteriori*. Ici, un GMM est appris pour chaque individu.

- **Evaluation :**

Afin de tester leur méthode et évaluer les performances des caractéristiques et classifieurs choisis, les auteurs ont commencé par construire leur propre base de données en enregistrant des séquences d'image couleur contenant 10 personnes. La première expérimentation a été effectuée en utilisant chaque classifieur avec chaque type de caractéristique prise isolément. Les taux de reconnaissance obtenus (en %) sont présentés dans le **Tableau 3.6** :

		Classifieur		
Caractéristiques		**GMM**	**kNN**	**SVM**
Caractéristiques Couleur	**Moyenne RGB**	90.1	86.5	90.2
	Hist. HMMD	90.3	94.7	89.7
	CSD	90.6	94.5	90.3
Caractéristiques Texture	**Hist. d'intensité**	59.1	55.2	39.4
	Mat. Coocurrence	24.8	23.7	27.1

Tableau 3.6 : Taux de reconnaissance.

Les résultats obtenus par les trois classifieurs sont très similaires, et dépendent des caractéristiques utilisées, mais le GMM semble être le plus « constant » avec toutes les caractéristiques (couleur en particulier). On voit ici clairement que, sur les séquences vidéos utilisées, les taux de reconnaissance obtenus par les caractéristiques « couleur » sont nettement supérieurs à ceux obtenus par les caractéristiques texture. De manière générale, les

résultats obtenus par les trois caractéristiques « couleur » sont très similaires, avec un léger avantage pour la CSD. Mais si la complexité de calcul est prise en compte, la valeur moyenne RGB serait le meilleur compromis.

Afin d'améliorer les taux de reconnaissance, les auteurs ont combiné les caractéristiques couleur (moyenne RGB) et texture (histogramme d'intensité). Pour cela, ils ont fusionné la sortie (probabilité) du classifieur GMM construit sur l'utilisation des caractéristiques couleur et la sortie (probabilité) du classifieur GMM basé sur les caractéristiques texture. Les opérations utilisées pour la fusion sont le produit, la somme, la moyenne et le maximum. Le **Tableau 3.7** présente les taux de reconnaissance obtenus par la caractéristique couleur (moyenne RGB), la caractéristique texture (histogramme d'intensité) ainsi que par la fusion des deux. On peut voir que la fusion des caractéristiques couleur et texture améliore le taux de reconnaissance en utilisant l'opération Produit pour la fusion des sorties du classifieur (91.8% contre 90.1% pour la couleur et 59.1% pour la texture).

		GMM
Caractéristiques Couleur	**Moyenne RGB**	90.1
Caractéristiques Texture	**Hist. D'Intensité**	59.1
Fusion	**Produit**	91.8
	Somme	74.2
	Moyenne	74.2
	Max	73.9

Tableau 3.7 : Taux de reconnaissance avec la couleur, la texture et la combinaison des deux.

Cette étude a permis de montrer trois choses. La première est que, sur l'ensemble de données utilisé, l'utilisation de caractéristiques couleur simple telles que les histogrammes ou la moyenne permet d'obtenir des performances aussi élevées que celles obtenues par des caractéristiques couleur plus complexes et plus couteuses à calculer telles que celles définies par le standard MPEG-7, dont le descripteur CSD. La deuxième est que les caractéristiques « couleur » sont plus pertinentes pour la reconnaissance des personnes que les caractéristiques « texture », en tout cas sur les bases de données vidéos utilisées par les auteurs. En effet, il est normal de constater que sur des vêtements peu texturées des caractéristiques texture soient moins discriminantes. L'écart se creuse d'autant plus si la base contient des tenues vestimentaires aux couleurs très variées, à l'inverse des textures. Enfin, comme précédemment, il a été montré que la fusion des caractéristiques « couleur » et « texture » permet d'améliorer les performances de la reconnaissance.

Comme nous venons de le voir, les méthodes de reconnaissance avec apprentissage hors ligne commencent par construire une base d'apprentissage contenant les modèles d'apparence de tous les individus concernés. Le système tentera alors de reconnaitre chacun d'eux lors de la phase de classification. Quant aux méthodes de reconnaissance avec apprentissage en ligne, elles ne construisent aucune base d'apprentissage au préalable. Les détails de ce type d'approche sont présentés dans ce qui suit.

3.4.3 Méthodes de reconnaissance avec apprentissage en ligne

Comme précisé en introduction, dans les méthodes avec apprentissage en ligne, aucun apprentissage préalable des individus à reconnaître n'est effectué. L'utilisation des techniques associées à cette approche est fréquente au sein de systèmes plus globaux, dédiés au suivi de personnes. Dans ce cadre, la reconnaissance est utilisée lorsqu'il y a intersection entre des individus afin de retrouver la trajectoire spécifique à chaque individu après une occultation. La reconnaissance ici mise en œuvre ne s'apparente pas à une décision basée sur des classes portant des labels relatifs à l'identification des individus (nom, n° de matricule, n° dans la base de connaissance, *etc.*). En fait, les labels sont définis en termes d'index caractérisant les individus exprimés selon une date, un numéro d'apparence, *etc.* Dans le cadre d'une application de suivi d'individus, un vecteur de caractéristiques (modèle) par personne est constitué à la volée sur chaque trame de la séquence. Une approche simple pour la classification, servant alors de base au suivi, consiste alors à calculer une distance entre ce modèle et celui de chaque personne présente dans la trame précédente. Cette personne est alors appariée avec celle ayant obtenue la distance la plus faible au sens d'une métrique particulière. Lorsqu'un individu quitte la scène, son vecteur de caractéristiques peut-être stocké à des fins de ré-appariement ultérieur. Nous détaillons dans ce qui suit trois des algorithmes de reconnaissance basées sur l'apparence avec apprentissage en ligne les plus proches de notre cadre applicatif.

3.4.3.1 Algorithme de Tao et al.

Tao et al. [TaoT03] ont développé un système de suivi qui intègre un module de reconnaissance de personnes en modélisant chacune d'entre elles par son histogramme couleur. Leur système se caractérise également par le fait qu'il comprend une étape de suppression d'ombres dans la partie traitement bas-niveau. La première phase mise en œuvre dans ce système vise à détecter et isoler chaque individu présent dans la scène dans une boite englobante (*bounding box*). Ils utilisent pour cela une méthode de soustraction de fond pour segmenter l'individu du reste de l'image. Les séquences d'images sont acquises par une caméra fixe avec un fond invariant. Chaque pixel de l'image est modélisé dans l'espace couleur HSV, et les pixels classés comme ayant varié sont marqués. L'image est donc convertie en image binaire où apparaissent uniquement les pixels classés comme n'appartenant pas au modèle du fond préenregistré. Les auteurs proposent d'utiliser ensuite une analyse de moments pour regrouper les différentes sous-régions détectées en régions compactes entières représentant des individus. Chaque individu est alors encadré par sa boite englobante. Quand il n'y pas d'occultation parmi les individus suivis dans la scène, l'algorithme CAMSHIFT (*Continuously Adaptive Mean Shift*) est utilisé pour permettre un suivi continu de ces individus. L'algorithme CAMSHIFT est un algorithme de segmentation d'images couleur basé sur l'algorithme du Mean Shift initialement introduit par Gary Bradski [Bra98] pour le suivi d'un visage.

Lors du suivi des individus, un histogramme couleur normalisé H_p dans l'espace HSV faisant office de vecteur caractéristique (ou modèle) est construit pour chaque individu p suivi. Pour cet histogramme couleur, les trois canaux de l'espace couleur HSV sont échantillonnés en 12 bins (composante teinte « Hue »), 4 (composante Saturation) et 4 (composante Value) respectivement. Afin de s'adapter aux changements d'apparence des individus, ce modèle d'histogramme est continuellement mis à jour comme suit :

$$H_p^k = (1-\alpha)H_p^{k-1} + \alpha H_p^{new},$$

Avec α est le facteur d'apprentissage et H_p^{new} l'histogramme couleur calculé sur l'image courante à « l'instant *k* », H_p^k est le modèle d'histogramme à l'instant *k* et H_p^{k-1} est le modèle d'histogramme à l'instant *k*-1.

Lorsque des individus se croisent (intersection), leurs boites englobantes se chevauchent. Dans ce cas, il y a occultation entre ces individus. Les opérations d'extraction et de mise à jours des histogrammes sont alors suspendues jusqu'à ce qu'il n'y ait plus d'occultation (réapparition de plusieurs boites englobantes). Ainsi, après un tel évènement, chaque individu est détecté séparément et son histogramme est extrait. Cependant, à ce stade, le suivi de chacun des individus a été rompu et leurs « identités » éventuellement confondues (on ne sait plus qui est qui). Par conséquent, afin de réaffecter à chaque individu la séquence de déplacements correcte, leurs histogrammes couleur respectifs sont comparés aux histogrammes calculés juste avant l'occurrence de l'occultation. On note ainsi *c* et *s* un individu segmenté avant et après l'occultation respectivement. Les histogrammes respectifs sont désignés par H_c et H_s. Afin de mesurer la ressemblance entre un individu *c* et un individu *s*, l'intersection entre leurs histogrammes H_s et H_c est calculée par :

$$P(s|c) = \sum_{i=1}^{d} \min(H_s(i), H_c(i))$$

Avec *d* le nombre de classes de l'histogramme, et $H(i)$ représente l'effectif de la classe i de l'histogramme H (pour H_c et H_s). Ayant N individus impliqués dans l'occultation et M individus segmentés après celle-ci, une fonction de reconnaissance est définie $w(s) = (w(s_1), w(s_2), \ldots, w(s_M))$, où $w(s_m) = n$ et $1 \leq n \leq N$. Les auteurs considèrent des probabilités *a priori* égales pour tous les N individus. Par conséquent, chaque personne segmentée peut être identifiée comme étant une des personnes impliquées dans l'occultation en utilisant une règle de classification de maximum de probabilité, donnée par :

$$w(s_m) = \arg\max_{n} P(s_m|c_n)$$

Pour poser la contrainte que chaque personne *c* ne peut physiquement correspondre qu'à une seule personne *s*, on obtient la reconnaissance optimale comme celle maximisant les probabilités conjointes $P(s_1,\ldots,s_M|c_1,\ldots,c_N)$ pour toutes les personnes *s*, qui peut être calculée par la fonction log-probabilité suivante :

$$w_{opt} = \arg\max_{w} \sum_{i=1}^{M} log\, P(s_i|w(s_i))$$

Afin d'évaluer leur méthode de reconnaissance, les auteurs l'ont d'abord testée sur des séquences d'images où des individus viennent de différents côtés, se croisent puis se séparent et partent dans différentes directions. Malheureusement, aucun taux de reconnaissance n'a été rapporté. La seconde évaluation a été effectuée sur deux séquences d'images acquises par une caméra filmant l'unique entrée d'une pièce. La première séquence contient onze individus et la deuxième en contient huit. Dans les deux, on y voit les individus (un à la fois) rentrer et sortir dans/de la pièce (une ou plusieurs fois). Initialement, la pièce est vide et tous les individus sont marqués comme « out ». Lorsqu'un individu rentre dans la pièce il est marqué comme « in » et son histogramme couleur est extrait, stocké et comparé à ceux des individus de la base de données et qui sont marqués comme « out ». Si le maximum de probabilité est

obtenu pour un certain individu et que cette valeur est supérieure à un seuil (fixé à 0,7) alors l'individu est considéré comme reconnu. Si le maximum de probabilité est inférieur au seuil, alors cet individu est considéré comme nouveau et est ajouté à la base de données. Le tableau suivant représente les taux de reconnaissance obtenus sur les deux séquences :

Séquence de test	Individus détectés	Individus reconnus	Taux de reconnaissance
Séquence I	27	23	88,9%
Séquence II	48	38	79,2%

Figure 3.8 : taux de reconnaissance.

On voit ici que les taux de reconnaissance obtenus sont assez hauts, mais diminuent néanmoins lorsque le nombre de personnes à reconnaitre augmente.

3.4.3.2 Algorithme de Cappellades et al.

Cappellades *et al* [CapDDC03] ont développé un système de suivi de personnes et de détection d'interactions personne-objet dans des environnements en intérieur. Leur objectif est de détecter, suivre et reconnaitre l'identité des personnes tout au long de la séquence vidéo, et ce même lorsqu'il y interaction entre les personnes ou avec des objets, ou lorsque ces personnes quittent la scène puis y reviennent. Toutes les séquences sont prises avec une caméra fixe avec un fond invariant. La première opération est une soustraction de fond pour segmenter les personnes et les objets de l'image. Puis, l'apparence des personnes est modélisée sur chaque image par une combinaison d'histogrammes et de corrélogrammes. Le corrélogramme couleur d'une image *I* est défini comme un tableau indexé par des couples de couleurs, dont la *d*-ième entrée pour le couple de couleurs $(\vec{c}_\alpha, \vec{c}_\beta)$ spécifie la probabilité de trouver un pixel de couleur $\vec{c}_\beta$ à la distance *d* d'un pixel de couleur $\vec{c}_\alpha$. Le corrélogramme couleur exprime comment la corrélation spatiale de couples de couleurs évolue en fonction de la distance dans l'image *I*. Pour une distance *d* et un couple de couleurs $(\vec{c}_\alpha, \vec{c}_\beta)$, le corrélogramme couleur $\gamma^{(d)}(\vec{c}_\alpha, \vec{c}_\beta)$ est défini par :

$$\gamma^{(d)}(\vec{c}_\alpha, \vec{c}_\beta) = card\{(P_1, P_2) \in I \times I \mid \vec{c}(P_1) = \vec{c}_\alpha,\ \vec{c}(P_2) = \vec{c}_\beta,\ \|P_1 - P_2\| = d\}$$

Où $\vec{c}(P_1)$ et $\vec{c}(P_2)$ correspondent respectivement aux couleurs des pixels P_1 et P_2.
Le correlogramme couleur $\gamma^{(d)}(\vec{c}_\alpha, \vec{c}_\beta)$ contient le nombre de couples de pixels (P_1, P_2) d'une image *I*, séparés par la distance *d* tels que la couleur du pixel P_1 soit $\vec{c}_\alpha$ et celle du pixel P_2 soit $\vec{c}_\beta$. Le corrélogramme couleur prend simultanément en compte la distribution globale des couleurs des pixels ainsi que la corrélation des couleurs entre pixels voisins. Il est notamment utilisé comme signature dans le cadre de l'indexation d'images couleur [HuaKMZ97].

Dans ce système, la distance normalisée L_1 est utilisée comme mesure de similarité entre deux corrélogrammes. Soient deux corrélogrammes $\gamma_1^{(d)}$ et $\gamma_2^{(d)}$ (de deux individus respectivement), la distance entre les deux est calculée par :

$$D(\gamma_1^{(d)}, \gamma_2^{(d)}) = \frac{\sum_{\forall i,j,k} \left| \gamma_1(c_i, c_j, k) - \gamma_2(c_i, c_j, k) \right|}{\sum_{\forall i,j,k} \gamma_1(c_i, c_j, k) + \sum_{\forall i,j,k} \gamma_2(c_i, c_j, k)}$$

Sur chaque trame de la séquence d'images, chaque individu présent est segmenté. Puis le système extrait son modèle (corrélogramme) et calcule la distance (L_1) entre ce modèle et le modèle de chacun des individus présents dans la base de données. Le système identifie alors l'individu à reconnaitre comme étant l'individu de la base de données qui est le plus similaire (distance minimale, et uniquement si cette distance minimale est inférieure à un seuil prédéfini). Si cette distance est supérieure au seuil, l'individu est marqué comme « nouveau » et son modèle est ajouté à la base de données. Les expérimentations ont été effectuées sur 5 personnes mais aucun taux de reconnaissance n'a été rapporté.

3.4.3.3 Algorithme de Seitner et al.

Seitner et al. [SeiH06] ont développé un système pour la détection et le suivi de piétons. Après avoir utilisé un algorithme de soustraction de fond (dans l'espace couleur HSV) pour segmenter les piétons potentiels, ils utilisent une cascade de classifieurs optimisés par boosting pour s'assurer qu'il s'agit bien de personnes. Sur chaque trame de la séquence, chaque individu détecté est divisé en trois blobs (tête, haut du corps et bas du corps) en utilisant des proportions fixes $[1/4, 3/8, 3/8]^T$. Un vecteur caractéristique est construit pour chaque blob. Ce vecteur contient la position P (pixel au centre de gravité) du blob et sa taille T, ainsi que son histogramme couleur $H_{\{h,s,v\}}$ (échantillonné en 10 bins), les valeurs de moyenne couleur $\mu_{\{h,s,v\}}$ et les valeurs de variance couleur $\sigma_{\{h,s,v\}}$ calculées sur chaque canal de l'espace HSV. Puis, à la prochaine trame, ces opérations sont répétées et une mesure de distance est calculée entre le vecteur caractéristique de chacun des trois blobs de chaque individu avec le vecteur caractéristiques correspondant (tête-tête, haut-haut et bas-bas) de tous les individus présents dans l'image précédente. Différentes mesures de distance sont appliquées selon les caractéristiques utilisées.

La distance Euclidienne est utilisée pour mesurer la distance spatiale $D_{position}$ entre deux positions (pixels) P_1 et P_2. La distance D_{hist} entre deux histogrammes couleur (échantillonné sur 10 bins) H_1 et H_2 est calculée par la distance de Bhattacharyya :

$$D_{hist} = (H_1, H_2) = 1 - \sum_{k=1}^{10} \sqrt{H_1(k).H_2(k)}\,.$$

La distance D_{dist} entre deux distributions couleur $C(\mu_1, \sigma_1)$ et $C(\mu_2, \sigma_2)$ est calculée par une distance de Mahalanobis modifiée :

$$D_{dist}(C(\mu_1,\sigma_1), C(\mu_2,\sigma_2)) = \frac{|\mu_1 - \mu_2|}{2\min(\mu_1, \mu_2)}\,.$$

Une fois toutes les distances calculées, elles sont pondérées et additionnées pour obtenir une distance finale D_{finale} (on précise que les auteurs ne précisent pas comment sont fusionnés les distances des trois blobs). L'individu est identifié comme étant celui qui a obtenu la distance la plus petite et inférieure à un seuil fixé. Si la distance minimale est supérieure au seuil, l'identification n'est pas validée, et une autre tentative sera effectuée à la prochaine image. Aucun taux de reconnaissance n'a été rapporté.

Comme précisé précédemment, les méthodes de reconnaissance de personnes avec approche en ligne sont principalement développées pour des problèmes de suivi. La classification est effectuée ici par le calcul de simples métriques. Lorsqu'un individu sort de la scène, son vecteur de caractéristiques le plus récent est stocké. Dans ce contexte, on travaille sur une fenêtre temporelle assez réduite (quelques trames) et on peut faire certaines hypothèses de

« continuité » ou de relative invariance de l'apparence qui permettent de simplifier cette mise en correspondance. C'est principalement pour ces raisons que les approches à base de distance se révèlent efficaces.

3.4.4 Conclusion

Nous avons présenté dans ce chapitre les approches de reconnaissance de personne utilisées en vidéosurveillance. Nous avons décrit deux approches biométriques qui sont la reconnaissance de visage et la reconnaissance de la démarche, et les limitations et les difficultés qui se posent à leur implémentation dans un système de vidéosurveillance ont été montrées. Nous avons également présenté les méthodes de reconnaissance basées sur l'apparence avec apprentissage hors ligne et apprentissage en ligne et nous avons détaillé les méthodes les plus pertinentes dans ces catégories d'approches ainsi que leurs performances.

Dans le cadre de cette thèse, nous avons développé une nouvelle méthode de reconnaissance basées sur l'apparence que nous présenterons au chapitre 4. Dans un premier temps, notre approche est positionnée parmi les méthodes de reconnaissance avec apprentissage hors ligne. En effet, la méthode se déroule en deux parties. Dans la première, les modèles d'apparence de tous les individus à reconnaitre sont construits et stockés (ce qui constitue un apprentissage hors ligne). Cependant, à l'inverse des méthodes que nous venons de voir [NakPHP03] [HahKK04] [GolKMS06], où la reconnaissance est basée sur la caractérisation de chaque individu à partir de son apparence global, le modèle d'apparence que nous utilisons pour chaque individu est issu de la modélisation séparée de ses vêtements du haut et du bas. Ainsi, chaque individu est identifié par la combinaison des vêtements qu'il porte. Un autre point important par lequel notre approche de reconnaissance se démarque des méthodes existantes est la procédure de « fusion des classes ». Cette procédure consiste à rassembler en amont les classes de vêtements similaires, et ce afin d'anticiper et d'éviter les confusions et les erreurs qui pourraient se produire lors de la phase de reconnaissance.

Il est important de rappeler que les expérimentations rapportées dans les approches de reconnaissance présentées dans [NakPHP03] [HahKK04] [GolKMS06] sont basées sur un « scénario fermé », c'est-à-dire que seuls les individus appris sont soumis à la phase de classification ultérieure. Ces systèmes ne sont donc pas conçus *a priori* pour pouvoir détecter l'émergence d'un « nouvel individu », c'est à dire non appris initialement. D'autre part, nous avons pu voir quelques méthodes de reconnaissance pourvu d'un mécanisme simple d'apprentissage en ligne [TaoT03] [CapDDC03] [SeiH06]. Cependant, et comme nous l'avons évoqué, ces méthodes ont été principalement développées afin de s'incorporer dans des systèmes plus globaux dédiés au suivi de personnes au sein d'une même séquence. Dans ce cadre, la reconnaissance est utilisée lorsqu'il y a intersection entre des individus afin de retrouver la trajectoire spécifique à chacun après occultation. En pratique, de par la manière dont elles sont conçues, ces méthodes ne construisent aucune base de connaissance proprement dite permettant de gérer un nombre relativement important de modèles d'apparence.

Dans l'approche que nous avons développée, nous proposons d'adjoindre à notre méthode de reconnaissance avec apprentissage hors ligne, des mécanismes de reconnaissance de « nouveauté » permettant d'effectuer un apprentissage en ligne et accroitre la base d'apprentissage. Cet apprentissage en ligne peut également être utilisé directement et sans apprentissage hors ligne, ce qui permet de constituer progressivement base de connaissance. Les détails de notre méthode de reconnaissance sont présentés dans le chapitre suivant.

CHAPITRE 4

4. CLASSIFICATION D'INDIVIDUS PAR LE BIAIS D'UN MODELE D'APPARENCE

4.1 Introduction

Comme évoqué précédemment, les travaux effectués dans cette thèse s'inscrivent dans le cadre du projet CAnADA, dont l'objectif est l'analyse de scènes et la génération d'alertes basées sur l'analyse comportementale dans un lieu accueillant du public, par le biais d'une installation multi caméras. Nous avons pu voir que, dans un tel contexte, l'une des contraintes majeures consiste à pouvoir observer et suivre les activités des individus filmés sur une fenêtre temporelle suffisamment large, en dépit des périodes où ces individus se retrouvent hors du champ de la caméra. Dans cette optique, la reconnaissance de personnes apparaît comme une solution efficace au problème « d'émergence » et « réémergence » des individus concernés. Ainsi, après avoir présenté dans le chapitre qui précède les approches issues de la littérature, nous présentons dans ce chapitre la méthode de reconnaissance de personnes que nous avons développée.

A l'inverse des méthodes de reconnaissance biométrique telles que la reconnaissance de visage, notre objectif est ici de distinguer des personnes entre elles, par le biais de leur « modèle d'apparence », plutôt que d'associer un identifiant unique à chaque individu. Comme nous le verrons, ce modèle d'apparence fournit des éléments pertinents pour réaliser ce processus de « caractérisation » (ultérieurement appelé « reconnaissance »). Contrairement aux méthodes présentées dans le chapitre 3, celui que nous utilisons est issu de la modélisation séparée des parties supérieure et inférieure (appelées respectivement éléments « haut » et « bas ») des corps des individus à reconnaitre. Bien sûr, ces éléments constitutifs sont avant tout conditionnés par les vêtements portés. Par conséquent, chaque individu observé est identifié par son « modèle d'apparence » qui est en pratique défini par une combinaison de vêtements (*haut* + *bas*) plutôt que par son apparence globale correspondant au corps tout entier (qu'on appellera ultérieurement la partie « *global* »). Une telle démarche se rapproche de la façon dont les gens en général, ou les policiers en particulier, donnent le signalement d'un individu lorsqu'ils doivent le retrouver parmi une multitude de personnes. A titre d'exemple, si on a rendez-vous avec une personne qu'on n'a jamais rencontrée (et qui ne connait pas notre visage) dans un espace public fréquenté, on lui décrira prioritairement notre tenue vestimentaire afin de lui permettre de nous reconnaitre au milieu de la foule : « *je porte un pull rouge et un jean bleu.* »

La méthode de reconnaissance que nous avons développée peut être utilisée dans les trois scénarios d'apprentissage intitulés « ensemble fermé », « ensemble ouvert » et « ensemble vierge ». Ces derniers sont décrits ci-dessous :

« Ensemble fermé ».
Principe : On exploite ici un apprentissage hors ligne, au sens où tous les M individus à reconnaitre sont appris au préalable. De plus, lors de la phase de reconnaissance, les personnes présentées au système sont celles déjà contenues dans la base d'apprentissage. Le système devra alors assigner chacune d'entre elles à la classe qui lui a été attribuée lors de la phase d'apprentissage.

- *Méthode* : On commence par enregistrer des courtes séquences d'images d'apprentissage des M individus (c'est-à-dire une séquence par personne) afin d'effectuer l'apprentissage hors ligne. De façon complémentaire, on enregistre des séquences

d'images de test des mêmes M individus (au moins une séquence par personne), qui serviront à effectuer la reconnaissance proprement dite. Lors des phases d'apprentissage et de reconnaissance, le système commence par effectuer des prétraitements sur chaque séquence. Ces derniers consistent en les procédures de détection d'individus par soustraction de fond et de détection de visage que nous avons présentées dans le chapitre 2. Le résultat de ces phases de segmentation est un « blob », relatif à la partie globale de l'individu. Celle-ci est ensuite scindée afin d'en extraire les parties « *haut* » et « *bas* ». Un vecteur de caractéristiques est ensuite extrait de chacune de ces entités. L'ensemble des N vecteurs de caractéristiques extraits sur une fenêtre temporelle de durée $\Delta T = N / f_{fps}$ (avec f_{fps} la fréquence d'acquisition des images) constitue ce qu'on va appeler sa « signature » et représente une description de son apparence. Cette signature permet ensuite la construction d'une classe en appliquant ici la technique one-class SVM. A la fin de ce procédé, nous disposons d'autant de classifieurs one-class qu'il y a de classes. Aussi, après que l'ensemble de M individus ait été présenté au système lors de la phase d'apprentissage, nous aurons initialement à disposition M classes pour les parties supérieures des corps (classes constituant l'ensemble $IH = \bigcup_{i=1}^{M} \{IH_i\}$ pour « initial haut ») ainsi que M classes (formant l'ensemble noté $IB = \bigcup_{i=1}^{M} \{IB_i\}$) pour les parties inférieures.

Arrivé à ce point, nous introduisons un point clé de notre approche de reconnaissance. Celle-ci réside en une procédure que nous avons appelée la « fusion des classes ». Compte-tenu de la nature du problème traité, un tel mécanisme est *a priori* indispensable. En effet, l'apparence des personnes telle qu'elle est considérée ici est avant tout conditionnée par la tenue vestimentaire. Il est assez fréquent que deux personnes différentes portent des vêtements presque identiques. Par conséquent, les classes correspondantes peuvent être extrêmement proches (un critère de similarité étant à définir). Plutôt que de chercher à mettre en place un mécanisme destiné à exacerber les subtiles différences entre deux classes réputées proches, nous avons au contraire pris l'option de prendre acte de cette similarité en fusionnant les deux classes concernées. En pratique, cette procédure consiste, lors de la phase d'apprentissage, à rassembler ou « fusionner » en une seule classe plusieurs classes de vêtements (*haut* et *bas* étant traités séparément) qui sont « similaires ». A titre d'exemple, si parmi les individus appris deux d'entre eux portent des pulls identiques (par exemple, des pulls rouges), ces deux pulls constitueraient deux classes initiales distinctes « pull rouge ». Si rien n'était fait, lors de la phase de reconnaissance, lorsqu'une de ces deux versions du « pull rouge » réapparait dans la scène, l'assignation à la classe correcte peut mener à une confusion, ce qui diminuerait le pouvoir discriminant du classifieur. Ainsi, la fusion en amont des deux classes correspondantes permet une pré-structuration de la connaissance au sein de la base. Dans l'exemple précédent, cela conduit à mettre en évidence le fait que deux personnes différentes portent en définitive un pull affecté à une même classe (la classe « pull rouge »). Cette procédure de fusion de classe génère un ensemble de « classes finales » à partir de celui des classes initiales.

Pour effectuer la fusion, le système commence par détecter de manière automatique les classes similaires. Il effectue pour cela une « classification croisée » parmi les classes initiales. En effet, dans la mesure où celles-ci sont construites grâce à une approche « one-class » n'utilisant que des exemples positifs de la classe en cours de construction (c'est-à-dire sans considérer les données des autres classes), il est possible *a posteriori* que des

éléments puissent être affectés simultanément à plusieurs d'entre elles. Le degré de similitude entre classes peut donc être estimé en comptabilisant la proportion d'éléments d'une classe donnée satisfaisant la fonction de décision d'une autre (avant opération de fusion). Avec cette approche, on peut alors définir une matrice de confusion dont chaque coefficient correspond à la valeur du « degré d'enchevêtrement » entre deux classes. Plus cette valeur est grande, plus ces deux classes sont réputées « proches » ou « similaires ».

En appliquant ce procédé, nous calculons une matrice de confusion pour les classes initiales (une matrice de confusion pour les classes *haut* et une autre pour les classes *bas*). En se basant sur le contenu de cette matrice, les classes initiales similaires sont repérées. Le système peut alors les fusionner afin de constituer des classes finales. On obtient ainsi un ensemble de classes finales *haut* noté $H = \bigcup_{i=1}^{N_{haut}} \{H_i\}$ et un ensemble de classes finales *bas* noté $B = \bigcup_{i=1}^{N_{bas}} \{B_i\}$, avec $N_{haut} \leq M$ et $N_{bas} \leq M$. Cette procédure de fusion permet ainsi de réduire les erreurs liées à la confusion (voire de les éliminer). La dernière opération effectuée lors de cette phase d'apprentissage hors ligne est d'associer aux individus observés les combinaisons correspondantes entre « classes finales *haut* » et « classes finales *bas* ». Nous vérifions l'éventuelle présence de « doublons » de combinaisons, qui sont regroupés le cas échéant. Ainsi, chaque combinaison spécifique constitue un « modèle d'apparence » décrivant un ou plusieurs individus.

A la fin de cette phase d'apprentissage hors ligne, on disposera d'une base d'apprentissage constituée par l'ensemble de classes *haut H*, l'ensemble de classes *bas B* et l'ensemble des « modèles d'apparence ». A l'aide de celle-ci, lors de la phase de reconnaissance, chaque individu observé dans les séquences de tests se voit attribué une classe « *haut* » et une classe « *bas* ». Celui-ci pourra alors être identifié par son « modèle d'apparence », c'est-à-dire par la combinaison (H_i + B_j) qui lui correspond dans les prises de vue.

« ensemble ouvert »

Principe : Dans ce scénario, nous partons toujours d'une base d'individus connus. La différence réside dans le fait que lors de phase de classification, les personnes à reconnaitre pourront différer de celles présentes dans la base initiale.

Méthode : On combine ici un apprentissage hors ligne et un apprentissage en ligne. L'apprentissage hors ligne est exactement le même que celui que nous avons expliqué pour dans le scénario précédent (« ensemble fermé »). C'est la phase de reconnaissance qui se distingue. En effet, lors de celle-ci, les individus présentés au système peuvent soit avoir été appris (et donc leurs « modèles d'apparence » sont contenus dans la base d'apprentissage) ou non. Le système devra alors identifier le « modèle d'apparence » de chaque individu observé soit comme étant un de ceux appris, soit comme étant « nouveau ». Si le « modèle d'apparence » d'un individu est identifié comme « nouveau », il sera alors ajouté à la base d'apprentissage. Cette procédure d'identification de « nouveaux modèles d'apparence » et leur ajout à la base d'apprentissage, lors de la phase de reconnaissance, correspond à un « apprentissage en ligne ». La détection de la « nouveauté » est effectuée en utilisant une des propriétés de la classification one-class SVM (qui sera précisée par la suite) qui permet de détecter lorsque des vecteurs de caractéristiques à classer n'appartiennent à aucune des classes apprises.

Dans l'un et l'autre de ces deux premiers scénarios, la classification proprement dite s'effectue en deux phases. La première consiste à reconnaitre le *haut* et le *bas* de l'individu observé, et la seconde à l'identifier par son « modèle d'apparence » défini par sa combinaison (*haut* + *bas*). Ainsi, lors de la reconnaissance dans le deuxième scénario (« ensemble ouvert » avec apprentissage en ligne), cinq cas sont possibles :

1 – la partie *haut* et la partie *bas* sont identifiées comme connues (existent dans la base d'apprentissage), ainsi que la combinaison (*haut* + *bas*). Dans ce cas, le « modèle d'apparence » de l'individu est bien identifié.
2 – la partie *haut* et la partie *bas* sont identifiées comme connues, mais la combinaison de ces deux vêtements n'a été observée sur aucun des individus appris (n'est pas contenue dans l'ensemble des « modèles d'apparence »). Aussi, cette combinaison est considérée comme « nouvelle » (inédite) et est ajoutée à l'ensemble des « modèles d'apparence ». Evidemment, l'individu observé ici est détecté comme « nouveau ».
3 – la partie *haut* est identifiée comme connue mais la partie *bas* est détectée comme « nouvelle » (n'existe pas dans la base d'apprentissage). Elle y est alors ajoutée. La combinaison (*haut* + *bas*) est donc forcement « nouvelle ». Cette nouvelle combinaison est alors ajoutée à la base des « modèles d'apparence ». Le « modèle d'apparence » de l'individu observé est donc, là encore, détecté comme « nouveau ».
4 – De façon complémentaire au cas précédent, la partie *bas* est ici bien identifiée (connue) mais la partie *haut* est détectée comme « nouvelle » (et sera donc ajoutée). Il en est de même pour la combinaison (*haut* + *bas*), le modèle d'apparence et l'individu observé.
5 – le dernier cas est celui où les deux parties *haut* et *bas* sont détectées comme « nouvelles ». Elles sont alors ajoutées à la base d'apprentissage, tout comme leur combinaison qui est forcement « nouvelle » aussi. Naturellement, le « modèle d'apparence » de l'individu observé ici est également détecté comme « nouveau ».

« **ensemble vierge** »
principe : Ici, aucun apprentissage hors ligne n'est effectué. Le système commence la phase de reconnaissance avec une base d'apprentissage vide (c'est-à-dire l'ensemble des classes *haut*, l'ensemble des classes *bas*, et l'ensemble des « modèles d'apparence » sont tous initialement vides). Le « modèle d'apparence » du premier individu observé ainsi que ses éléments constitutifs (parties « *haut* » et « *bas* ») seront automatiquement considérés comme « nouveaux » et ajoutés à la base d'apprentissage. Par la suite, comme dans le scénario précédent, le système identifiera le « modèle d'apparence » de chaque individu observé soit comme étant un de ceux contenus dans cette base, soit comme étant « nouveau » (et l'ajoutera, le cas échéant) construisant ainsi progressivement la base d'apprentissage. Le système effectue donc un apprentissage entièrement en ligne.

Dans la suite de ce chapitre, nous présentons dans le détail les étapes de prétraitement associées à l'ensemble des scénarii évoqués, à savoir :

- la détection des personnes et la scission du corps,
- l'extraction du vecteur de caractéristiques (contenant des valeurs correspondant à différentes informations de couleur et de texture, pour chacune des parties « *haut* » et « *bas* ».

Concernant ce dernier élément, nous justifierons la pertinence des composantes exploitées pour la description de l'apparence des vêtements (utilisées, comme indiqué, pour la construction des signatures) par le biais d'une évaluation de leur « pouvoir discriminant ». Nous utilisons pour cela deux techniques de classification « traditionnelles » qui sont les kppv

(k plus proches voisins, avec $k=1, k=3$ puis $k=5$) et les GMM (modèles de mélanges de gaussiennes), qui furent toutes deux présentées dans le chapitre 3.

Après avoir justifié le vecteur de caractéristiques employé, nous détaillons les différentes procédures d'apprentissage et de reconnaissance exploitées dans les différents scénarios. Ainsi, nous commençons par présenter les procédures d'apprentissage hors ligne et de reconnaissance de personnes pour le scénario « ensemble fermé ». Nous évaluerons alors notre approche sur une base de données constituée de 54 individus. Nous pourrons alors démontrer l'intérêt de reconnaitre chaque individu en définissant son « modèle d'apparence » par sa combinaison de vêtement (*haut* + *bas*) au lieu de le reconnaitre en définissant celui-ci au travers d'une approche de type « corps entier » (partie *global*). Par la suite, nous exposons les mécanismes qui permettent dans notre méthode d'effectuer la détection de la nouveauté et l'apprentissage en ligne. Nous présentons alors les résultats des évaluations effectuées dans le scénario « ensemble ouvert », puis dans le scénario « ensemble vierge ». En guise de conclusion, nous finissons ce chapitre en discutant des performances exhibées par le procédé utilisé.

4.2 Détection de personnes et scission du corps

Les premières opérations qu'effectue le système sont la détection et la segmentation du reste de l'image de chaque individu présent dans la scène. Ces opérations correspondent aux procédures de soustraction du fond et de soustraction du contour du fond que nous avons présentées dans le chapitre 2. Le résultat sera un ensemble de formes détectées (« blobs ») qui correspondent aux éléments au premier plan. Nous vérifions ensuite que chaque blob correspond bien à une personne à l'aide des procédures de sélection des blobs et de détection de visage présentées dans le chapitre 2. Le résultat de ces opérations est illustré dans la **Figures 4.1** (c). A partir du blob entier de l'individu, on extrait deux parties correspondant respectivement au « vêtement du haut » et au « vêtement du bas ». Ces deux parties sont obtenues après un procédé de division du blob (ou « scission ») basé sur un rapport de hauteur et détaillé ci-dessous.

On commence par diviser le blob entier en trois zones en utilisant des rapports de hauteur relatifs à sa taille totale. Dans nos conditions opératoires, déterminées notamment par la distance « moyenne » à laquelle se situent les individus filmés ainsi que par la hauteur à laquelle est placée la caméra et son inclinaison par rapport à l'horizontale (voir chapitre 2, section 2.3.2), nous avons fixé ces rapports à [1/5, 3/10, 1/2] de la taille totale du blob (la première valeur permet d'isoler la zone correspondant à la tête). Les rapports exacts sont sujets à une certaine variabilité en fonction des individus, en plus de l'influence du positionnement de la caméra. Quoiqu'il en soit, d'un point de vue pragmatique, les essais que nous avons pu conduire ont mis en évidence une pertinence certaine des ratios utilisés.
La **Figure 4.1** (d) illustre cette procédure de scission. Les trois zones obtenues correspondent respectivement à la tête, au haut du corps (le buste) et au bas de celui-ci (les jambes). Comme nous l'avons expliqué, seules les deux dernières nous intéressent. Ces deux parties subissent ensuite une opération d'érosion pour n'en garder que la zone centrale et afin de supprimer les bords qui pourraient être bruités en raison d'une segmentation du fond imparfaite ou d'un chevauchement entre les vêtements du haut et du bas. On obtient ainsi les parties *haut* et *bas* de l'individu observé (ou plutôt de ses vêtements). Une fois ce processus achevé, un vecteur de caractéristiques est extrait pour chaque partie.

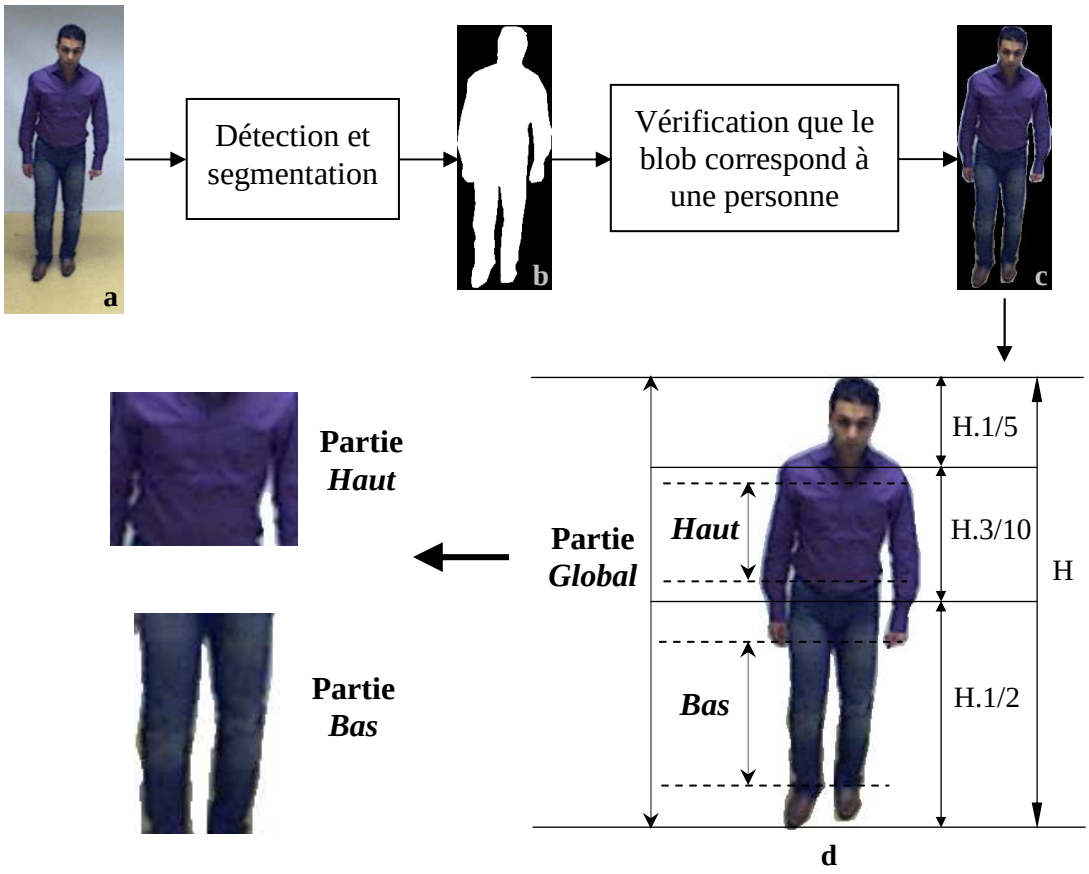


Figure 4.1 : Procédure de détection et de scission du corps. (a) image d'entrée. (b) blob détecté et segmenté du reste de l'image. (c) individu détecté. (d) scission du blob et extraction des parties *haut* et *bas*.

4.3 Extraction de signature

Comme indiqué dans l'introduction, la signature d'un blob, ou en l'occurrence d'un sous-blob, consiste en un ensemble de vecteurs de caractéristiques extraits sur plusieurs images d'une séquence. Aussi, puisque l'apparence extérieure d'une personne dépend principalement des vêtements qu'elle porte avec une grande variété de couleurs et de motifs, nous pouvons présumer que des caractéristiques décrivant la distribution des couleurs et des textures sont pertinentes pour réaliser la tâche de reconnaissance. Par ailleurs, nous avons montré dans le chapitre 3 que les méthodes de reconnaissance de personnes se basent principalement sur des descripteurs bâtis autour de ces éléments. Il est à noter néanmoins que de précédentes approches ont tenté d'aborder le problème de reconnaissance en se basant notamment sur des caractéristiques de forme [NakPHP03]. Cependant, obtenir la forme (notamment pour une analyse de silhouette ultérieure) requiert une extrême précision de la segmentation de fond, qui n'est pas toujours aisée à réaliser. Par ailleurs, la détermination de la silhouette est généralement coûteuse en calcul à cause de la nature non-rigide du corps humain.

Nous utilisons ainsi des caractéristiques couleur et texture pour construire les vecteurs constituant les signatures des parties *haut* et *bas*. Afin de mettre en évidence l'intérêt de reconnaitre chaque individu en définissant son « modèle d'apparence » par sa combinaison (*haut* + *bas*) plutôt que d'en construire un global (partie *global*), nous extrayons également la signature du corps entier de chaque individu. Nous pourrons alors comparer par la suite les

résultats de la reconnaissance par combinaison avec ceux de la reconnaissance basée sur l'apparence globale. Dans ce qui suit, nous décrivons les caractéristiques couleur et texture que nous avons choisies d'utiliser *a priori*. A des fins de justification de l'espace de représentation des couleurs utilisé (en complément des éléments déjà fournis dans le chapitre 3), celles-ci seront soumises une batterie de tests comparatifs mettant en œuvre des métriques simples et des méthodes de classification très « classiques ».
Les caractéristiques exploitées ici sont les suivantes :

- **Caractéristiques couleur :**

  - ***Valeurs moyennes des couleurs*** : modélisent la couleur d'un blob en calculant la valeur moyenne des pixels qui le composent sur chaque canal couleur (noté C1, C2, et C3). Les moyennes des couleurs sont intrinsèquement invariantes en rotation, translation et par changement d'échelle. Trois valeurs sont donc obtenues (valeur moyenne sur C1, valeur moyenne sur C2 et valeur moyenne sur C3).

  - ***Valeurs des variances couleurs*** : donnent une mesure de la distribution de la couleur du blob autour de sa moyenne. Les valeurs des variances sont également calculées pour chacun des canaux couleur (C1, C2, C3) de l'image, sur tous les pixels qui composent le blob. Les valeurs de variances des couleurs sont également invariantes en rotation, translation et par changement d'échelle.

  - ***Histogrammes couleurs normalisés*** : comme nous l'avons vu précédemment, l'histogramme d'un blob au sein de l'image couleur permet d'obtenir une représentation de la distribution des intensités (niveaux de gris) des pixels constituant ce blob, et ceci sur chaque canal couleur C1, C2 et C3. Il est donc défini comme une fonction discrète qui associe à chaque valeur d'intensité le nombre de pixels de ce blob prenant cette valeur. La détermination de l'histogramme est réalisée en comptant le nombre de pixels par plage d'intensité. Ici, on calcule un histogramme comportant 16 plages ou bins pour chaque canal (C1, C2, C3), de même « largeur ». La taille du vecteur se trouve alors augmentée de 48 composantes sur chaque image. Les histogrammes sont intrinsèquement invariants en rotation et en translation. Nous les normalisons ici en divisant la valeur de l'effectif de chaque bin par le nombre total de pixels du blob afin de les rendre invariants aux changements d'échelle.

- **Caractéristiques texture :**

  - ***Caractéristiques texture basées sur la matrice de cooccurrence*** : comme nous l'avons précédemment défini dans le chapitre 3, ces caractéristiques se calculent à partir des indices du second ordre dérivés de la matrice de cooccurrence. Pour une image en niveaux de gris, une matrice de cooccurrence est une matrice carrée dont la dimension est égale au nombre des niveaux d'intensité (niveaux de gris) pris en considération et ses coefficients représentent les valeurs de probabilité jointe $p_{d,\theta}(i,j)$ d'une paire de pixels avec les valeurs d'intensité respectives I_i et I_j pour différents angles θ et distances d et une partition donnée $\bigcup_{k=1}^{N} I_k$ de l'ensemble des niveaux d'intensité possibles. Haralick [Har79] a défini 14 coefficients (paramètres ou descripteurs) qui peuvent être calculés à partir de la matrice de cooccurrence afin de décrire la texture d'une image. Dépendant de la nature de

l'image, certains de ces coefficients peuvent avoir des valeurs indéfinies ou alors contenir très peu d'informations. Beaucoup de tests doivent être faits sur des données relatives aux applications envisagées afin de sélectionner les coefficients d'Haralick les plus pertinents dans un contexte précis. Ainsi, dans notre cas, après avoir mené des expérimentations, nous avons choisi d'utiliser les cinq paramètres suivants : le contraste, la corrélation, l'uniformité, l'énergie et l'entropie [Har79].

Dans la section suivante, nous allons évaluer les différentes caractéristiques que nous venons de décrire et que nous extrayons sur les espaces couleur que nous avons évoqués. Comme précisé au début de cette section, nous utilisons pour cela deux techniques de classification « traditionnelles » qui sont les kppv et les GMM. Cette évaluation s'effectuera sur les parties *haut* et *bas* mais également sur la partie « *global* », qui correspond à l'apparence construite à l'aide du corps entier de l'individu. Notre objectif de mettre en évidence l'intérêt d'identifier chaque individu par sa combinaison (*haut* + *bas*) plutôt que par son apparence globale. Avant de passer à l'évaluation des caractéristiques proprement dite, nous décrivons la structure de la base associée à ce test.

Détails de la base vidéo utilisée **:**
Afin de construire notre base de données, nous avons enregistré des séquences vidéo de 54 individus. Ces séquences ont été enregistrées par une webcam Philips SPC900NC/00 montée sur un trépied à 1,5 mètres de haut. La cadence d'acquisition a été fixée à 25 images par seconde, et la taille des images à 240x320 pixels, au format RGB. La caméra a été placée de manière à filmer l'intérieur d'une pièce avec un mur en arrière plan à une distance de 5 mètres, devant lequel des individus viennent marcher, se tenir debout (de face, de profil ou de dos) ou tourner sur eux-mêmes. L'éclairage de la scène par des néons au plafond est resté inchangé durant toutes les acquisitions. Le fond des images est constitué d'un mur gris sans motif. Onze membres de notre laboratoire ont été sollicités pour participer aux acquisitions. Chacun d'eux a fait une ou plusieurs apparitions en mettant différents vêtements (pull, chemise, pantalon, *etc.*), constituant ainsi une base de 54 « prototypes » ou « individus » (par abus de langage, car il s'agit de 54 combinaisons *a priori* différentes de vêtements portés).
Dans le détail, à ces onze personnes, nous avons associé 36 vêtements du haut (pull, T-shirt, chemise, *etc.*) et 13 vêtements du bas (pantalons). Ces vêtements ont été portés au hasard par une ou plusieurs des onze personnes impliquées dans nos séquences d'images. On peut voir par exemple dans la **Figure 4.2** qu'un même pull est porté par les individus #5, #25 et #30, ou qu'un même pantalon est porté par les individus #9, #10, #11 et #12 (qui sont en fait la même personne, en l'occurrence *Moussa*, qui a mis quatre pulls différents). Le **Tableau 4.1** présente les 36 vêtements du haut utilisés et le numéro des individus qui les portent. De même, le **Tableau 4.2** présente les 13 pantalons utilisés et le numéro des individus qui en furent vêtus. Aussi, nous disposons en définitive de 36 classes de vêtements du haut et de 13 vêtements du bas. Chacun des 54 « individus » a effectué deux passages devant la caméra, le premier durant 8 secondes et le second 16 secondes. Durant le premier passage on enregistre les 200 images (8x25) d'un individu portant une combinaison de vêtements particulière (*haut* + *bas*). Toutes les images d'une même séquence représentent donc des exemples positifs de « l'étiquette » caractérisant cette combinaison de vêtements spécifique. Un vecteur de caractéristiques est extrait de chaque partie (*haut*, *bas* et *global*) sur chaque image. Aussi, nous obtenons 200 vecteurs de caractéristiques pour chaque partie, constituant alors la signature exploitée lors de la phase d'apprentissage. Durant le deuxième passage on enregistre les 400 (16x25 pour chaque individu) images d'où seront extraites les signatures de chaque partie et qui serviront pour faire la classification. La caméra a ainsi enregistré 54 séquences d'apprentissage de 200 images chacune et 54 séquences de classification de 400 images à des fins de test (réalisées

pour chaque partie). La **Figure 4.2** présente un exemple d'images de chacun des 54 individus enregistrés et le numéro qui leur a été attribué.

Figure 4.2 : Exemple d'images de chacun des 54 individus enregistrés.

4.4 Evaluation des caractéristiques

Afin d'effectuer l'apprentissage des classifieurs kppv et GMM, nous utilisons uniquement les données d'apprentissage d'un seul individu par vêtement. Par exemple, pour faire l'apprentissage de la classe *haut* 5, nous utiliserons uniquement la signature d'apprentissage du vêtement du haut de l'individu #5. Par contre, pour effectuer la classification, nous utilisons l'ensemble des signatures de test relatives à ce même vêtement. Ainsi, lors de classification, l'association à la classe vêtement du haut 5, par exemple, sera réalisée à l'aide des données de test relatives à la partie « haut » des individus #5, #25 et #30. Le même procédé est bien sûr utilisé pour les classes des vêtements du bas. Par exemple, pour faire l'apprentissage de la classe *bas* 3, nous utiliserons uniquement les données d'apprentissage du vêtement du bas de l'individu #9. Par contre, lors de la classification, l'appariement à la classe *bas* 3 sera évalué à l'aide des données de test des parties « bas » des individus #9, #10, #11 et #12. En ce qui concerne les classes *global*, nous présumons que nous avons à faire à autant de classes qu'il y a d'individus, c'est-à-dire 54. Ainsi, les 54 signatures d'apprentissage des parties *global* serviront à faire donc l'apprentissage, et les 54 signatures de tests des parties *global* serviront à faire la classification.

	Vêtement du haut (pull, T-shirt, chemise)											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
												
Individus	#1	#2 #27	#3 #37	#4 #49	#5 #25 #30	#6	#7 #40	#8 #38	#9	#10 #47 #51	#11 #42	#12 #39
	Vêtement du haut (pull, T-shirt, chemise)											
	13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
												
Individus	#13	#14	#15	#16	#17	#18	#19 #48 #50	#20 #43	#21 #44	#22	#23	#24 #35
	Vêtement du haut (pull, T-shirt, chemise)											
	25	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
												
Individus	#26	#28	#29	#31	#32	#33	#34	#36 #53	#41	#45	#46	#52 #54

Tableau 4.1 : Les 36 vêtements du haut utilisés et les numéros des individus qui les portent.

	Vêtement du bas (pantalon)												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
													
Individus	#1 #2 #3 #4 #5 #53 #54	#6 #7 #8	#9 #10 #11 #12	#13 #14	#15	#16 #17	#18 #19 #20 #21	#22 #23 #24 #25	#26 #27 #28 #29 #30	#31 #32	#33 #34 #35 #36 #37 #38 #39	#40 #41 #42 #43 #44 #49 #50 #51 #52	#45 #46 #47 #48

Tableau 4.2 : Les 13 vêtements du bas utilisés et les numéros des individus qui les portent.

Les tableaux **Tableau 4.3**, **Tableau 4.4** et **Tableau 4.5** présentent les taux de classification (en pourcentage) obtenus par les différentes caractéristiques et classifieurs utilisés pour les parties *haut*, *bas* et *global* respectivement. Le **Tableau 4.3** présente les taux de classification des classes de vêtements du haut. On peut voir que les valeurs des moyennes couleur RGB obtiennent des scores entre 94% et 95% (94% avec le kppv, avec k=1, 3 et 5, et 95% avec le GMM). Les valeurs des moyennes RGB normalisées obtiennent des scores plus faibles entre 79% et 83%. Les valeurs des moyennes YCbCr obtiennent des scores entre 93% et 95%, et les valeurs des moyennes HSV obtiennent des scores entre 91% et 93 %. Les valeurs des variances couleur obtiennent des scores nettement plus faibles sur l'ensemble des espaces couleurs (entre 66% et 69% avec le classifieur GMM). Les scores obtenus par l'histogramme RGB sont élevés avec un maximum de 95% obtenus par la classification kppv avec k=3 et k=5. Quant aux valeurs de texture, les scores obtenus sont entre 65% et 69% et nettement inférieurs à ceux obtenus par des caractéristiques couleur (ces résultats sont cependant à nuancer). On peut déduire des résultats présentés dans ce tableau que les valeurs moyennes des couleurs ainsi que l'histogramme RGB sont très efficaces pour la classification des vêtements du haut avec des scores de reconnaissance tournant autour de 94%, avec un très léger avantage pour les espaces RGB et YCbCr. Les valeurs de variances le sont moins avec un taux de reconnaissance maximum de 69% obtenu par l'espace YCbCr. Enfin, il apparait clair que la texture est moins efficace pour effectuer cette reconnaissance avec un taux maximum de 69%. Cependant, il est important de préciser que les vêtements utilisés dans cette étude sont très peu texturés (pour la majorité d'entre eux), ce qui peut expliquer la relative non-efficacité des caractéristiques de texture à les différencier. En outre, nous avons refait ces procédures d'apprentissage et de classification en utilisant uniquement les classes de vêtement haut 1, 4, 5, 10, 15, 19, et 34 qui sont relativement texturés, et les caractéristiques correspondantes ont obtenu des taux de reconnaissance de 83% avec les classification GMM ainsi qu'avec kppv (k=5). Ceci démontre que la texture serait plus efficace pour la reconnaissance sur des vêtements suffisamment texturés.

Caractéristiques	Classifieurs			
	kppv k=1	kppv k=3	kppv k=5	GMM
Moyenne RGB	**94** %	**94** %	**94** %	**95** %
Moyenne rgbN	79 %	81 %	82 %	83 %
Moyenne YCbCr	**93** %	**94** %	**94** %	**95** %
Moyenne HSV	**91** %	**92** %	**92** %	**93** %
Variance RGB	59 %	60 %	61 %	69 %
Variance rgbN	58 %	60 %	62 %	66 %
Variance YCbCr	60 %	60 %	62 %	69 %
Variance HSV	62 %	63 %	65 %	67 %
Histogramme RGB	**94** %	**95** %	**95** %	**94** %
Texture	65 %	67 %	69 %	65 %

Tableau 4.3 : Taux de classification des classes de vêtements du haut.

Analysons maintenant les taux de reconnaissance obtenus sur les classes de vêtements du bas présentés dans le **Tableau 4.4**. On peut voir que les scores obtenus par les différentes valeurs moyennes des couleurs sont entre 87% et 92% sur les espaces RGB, YCbCr et HSV avec un très léger avantage pour les valeurs en RGB. On peut noter les faibles taux de reconnaissance obtenus par les valeurs de variances des couleurs, avec un avantage sur l'espace HSV (66%). L'histogramme couleur, comme les valeurs moyennes, obtient des scores élevés, entre 91% et 92%. Les scores des caractéristiques de texture sont, là encore, faibles, entre 48% et 52%. Ici aussi, on peut voir que les pantalons portés par les individus ne sont pas texturés. On peut déduire également ici que les caractéristiques couleur sont très efficaces pour la reconnaissance des vêtements du bas avec des scores tournant autour de 90% (à l'exception des valeurs RGB normalisées), avec un léger avantage pour les valeurs des moyennes RGB et l'histogramme couleur (92%). Cependant, on note que ces taux de reconnaissance sont légèrement inférieurs à ceux obtenus sur les vêtements du haut. Ceci peut s'expliquer par le fait que les pantalons utilisés ici se ressemblent presque tous (jeans bleus ou pantalons foncés pour la plupart), alors que les pulls sont plus dissemblables et contiennent un peu plus de couleurs.

Caractéristiques	Classifieurs			
	kppv k=1	kppv k=3	kppv k=5	GMM
Moyenne RGB	**89** %	**91** %	**92** %	**92** %
Moyenne rgbN	80 %	**82** %	84 %	84 %
Moyenne YCbCr	**87** %	88 %	89 %	**90** %
Moyenne HSV	**88** %	**89** %	**90** %	**91** %
Variance RGB	46 %	48 %	49 %	56 %
Variance rgbN	54 %	55 %	57 %	62 %
Variance YCbCr	48 %	51 %	53 %	57 %
Variance HSV	63 %	65 %	66 %	67 %
Histogramme RGB	**92** %	**92** %	**91** %	**92** %
Texture	48 %	48 %	50 %	52 %

Tableau 4.4 : Taux de classification des classes de vêtements du bas.

Passons maintenant au **Tableau 4.5** qui contient les taux de reconnaissance obtenus sur les classes corps global. On peut y voir que les scores obtenus par les moyennes des couleurs restent ici relativement élevés sur tous les espaces couleur (excepté encore une fois pour RGB normalisé), situées entre 81% et 84%. Les espaces RGB et YCbCr sont les plus « performants » d'après ce critère. Les scores des valeurs des variances restent ici aussi assez bas, entre 51% et 72%, avec un net avantage pour l'espace HSV. Cependant, les scores des histogrammes couleur sont ici bien supérieurs à tous les autres (88% avec presque tous les classifieurs). A l'instar des cas précédents, les caractéristiques de texture, obtiennent ici aussi des scores bas, entre 52% et 54% (valeurs explicables par le faible effectif de vêtements texturés au sein de la base).

Caractéristiques	Classifieurs			
	kppv k=1	kppv k=3	kppv k=5	Gmm
Moyenne RGB	**82** %	**83** %	**84** %	**84** %
Moyenne rgbN	67 %	70 %	71 %	71 %
Moyenne YCbCr	**82** %	**83** %	**84** %	**84** %
Moyenne HSV	**81** %	**83** %	**83** %	**82** %
Variance RGB	51 %	53 %	54 %	58 %
Variance rgbN	53 %	55 %	57 %	60 %
Variance YCbCr	60 %	62 %	63 %	68 %
Variance HSV	68 %	70 %	72 %	71 %
Histogramme RGB	**87** %	**88** %	**88** %	**88** %
Texture	54 %	52 %	53 %	54 %

Tableau 4.5 : Taux de classification des classes global.

Après avoir analysé les résultats présentés dans les tableaux **Tableau 4.3, Tableau 4.4** et **Tableau 4.5**, on peut clairement déduire que les caractéristiques couleurs, principalement les valeurs des moyennes des couleurs et l'histogramme sont très efficaces et adéquates pour la représentation et la discrimination de l'apparence extérieure des groupes de personnes présents dans notre base (supposée représentatives des cas d'application réels), que ce soit pour l'apparence de leurs vêtements du haut ou du bas pris isolément ou pour leur apparence globale. Bien que les scores obtenus avec les valeurs moyennes sur les espaces RGB, YCbCr et HSV soient très similaires, ceux associés aux valeurs RGB sont en moyenne légèrement supérieurs et presque aussi élevés que ceux obtenus par les histogrammes RGB. Il aurait été possible ici de tenter d'améliorer les taux de reconnaissance en combinant les caractéristiques couleur et texture, tel que ça a été fait avec certaines méthodes de reconnaissance présentées dans le chapitre 3 [HahKK04] [GolKMS06]. On rappelle par exemple que dans [GolKMS06], les auteurs arrivent à améliorer légèrement les taux de reconnaissance en fusionnant les caractéristiques couleur et texture (91.8% avec la fusion contre 90.1% pour la couleur et 59.1% pour la texture). Il aurait été également possible d'utiliser une analyse en composantes principales (ACP) sur l'ensemble des caractéristiques pour en tirer une meilleure discrimination. Cependant, pour des raisons tenant à la capacité à effectuer les différents traitements en temps-réel, nous avons souhaité réaliser un bon compromis entre « pouvoir discriminant » et complexité des traitements à réaliser. Il découle des études précédentes que les valeurs moyennes sur RGB sont suffisantes pour constituer nos vecteurs de caractéristiques. Il est clair que l'histogramme peut apporter une information propre à améliorer encore le taux de reconnaissance. Toutefois, l'intégration de cet élément apporterait 48 composantes supplémentaires (à raison de 16 classes par canal), alors qu'il apparaît qu'un

vecteur à 3 composantes nous apporte déjà des performances compatibles avec celles recherchées.

Un fait remarquable apparaissant dans les tableaux précédents est que les taux de reconnaissance obtenus sur les classes vêtements du haut et les classes vêtements du bas sont supérieurs à ceux obtenus sur le corps entier. On pourrait alors penser que discriminer les pulls ou discriminer les pantalons d'un groupe d'individus est plus « efficace » que de discriminer leurs apparences entières. Cependant, cette comparaison ne serait pas ici tout à fait justifiée, car si les classes correspondant aux vêtements du haut et aux vêtements du bas ont bénéficié d'un regroupement préalable entre classes « visiblement identiques » lors de la phase d'apprentissage, ce n'est pas le cas des classes associées à l'apparence globale. Si une telle opération n'avait pas été effectuée, il est très probable que les taux de reconnaissance auraient été bien plus faibles à cause des confusions engendrées par les classes identiques. Ce fait a été vérifié en réalisant l'apprentissage sans regroupement et les résultats de la classification ont été considérablement dégradés car nous avons trouvé des scores de 74% sur *haut* et 34% sur *bas* avec les moyennes RGB et la classification GMM.
Ces premiers résultats mettent en lumière que rassembler en une seule des classes identiques ou même similaires (et cela en amont du processus, c'est-à-dire lors de la phase d'apprentissage) permet d'améliorer considérablement les taux de reconnaissance. Si ce regroupement s'est effectué de manière manuelle durant ces premières manipulations, il va de soi que dans une application de reconnaissance automatique intégrée à un système de vidéosurveillance celle-ci doit s'opérer sans l'intervention d'expert humain. Par conséquent, une approche permettant dans un premier temps d'identifier automatiquement en amont les classes similaires, et qui permette par la suite d'effectuer le rassemblement (« fusion ») de ces classes, doit être utilisée.

Pour cela, nous présentons une stratégie d'apprentissage et classification se basant sur la technique one-class SVM. Nous rappelons que la classification one-class SVM permet de définir une fonction de décision dont l'hyperplan définit la frontière d'une seule classe [HaySTA00] [UnnRJ3]. Notre choix s'est porté sur cette technique à cause notamment de deux propriétés qu'elle présente, qui sont la « fusion des classes » permettant de gérer les confusions entre classes, et la « détection de la nouveauté » permettant d'effectuer un apprentissage en ligne. La première propriété est illustrée dans la **Figure 4.3**. On peut voir sur celle-ci deux classes, à savoir la classe *rond* et la classe *triangle*. On tente ici de classer une nouvelle donnée symbolisée par une *croix*. Si on essaye de résoudre le problème par une approche de classification simple, telle que les kppv ou les GMM, la nouvelle donnée serait attribuée à une des deux classes existantes. Cette manière d'aborder le problème de classification revient à poser au classifieur utilisé la question : *à laquelle de ces classes appartient cette donnée ?* En posant le problème de cette manière, on suppose d'emblée que la nouvelle donnée doit forcement appartenir à une des deux classes existantes. Ceci exclut alors deux choses. La première est que cette nouvelle donnée puisse appartenir aux deux classes simultanément, et la deuxième est qu'elle puisse n'appartenir à aucune d'entre elles.

Pour éventuellement autoriser ces deux dernières alternatives, nous abordons le problème différemment. Au lieu de définir un hyperplan pour séparer les deux classes existantes, nous allons définir un hyperplan pour chaque classe qui définit uniquement sa frontière. Ainsi, nous définirons autant de classifieurs qu'il y a initialement de classes (dans l'exemple de la **Figure 4.3**, deux). Chacun de ces classifieurs devra alors répondre à la question : *est ce que, oui ou non, cette nouvelle donnée appartient à cette classe ?* Très classiquement, si cette donnée est située à l'intérieur du contour défini par la frontière de décision, la réponse sera *oui*, sinon, la

réponse sera *non*. Par hypothèses, les classes pouvant se recouvrir, toutes les combinaisons sont possibles. Ainsi, il sera possible de déterminer si la nouvelle donnée appartient à une des deux classes, ou aux deux, ou à aucune. Dans l'exemple de la **Figure 4.3** (b), après avoir défini une frontière de décision pour chaque classe, on peut voir que la donnée à classer se situe à l'intérieur du contour de décision des deux classes. Aussi, cette donnée ne serait pas attribuée exclusivement à une des deux classes mais bien aux deux. Ceci implique, évidemment, que les contours de ces deux classes soient « enchevêtrées ». Si on souhaite maintenant classifier les données de la classe *rond*, on se rendrait compte qu'un certain nombre d'entre elles seraient non seulement classifiées dans la classe *rond* mais également dans la classe *triangle*, puisqu'elles se situent également à l'intérieur du contour de la celle-ci. La même chose peut être constatée sur un certain nombre de données de la classe *triangle*.

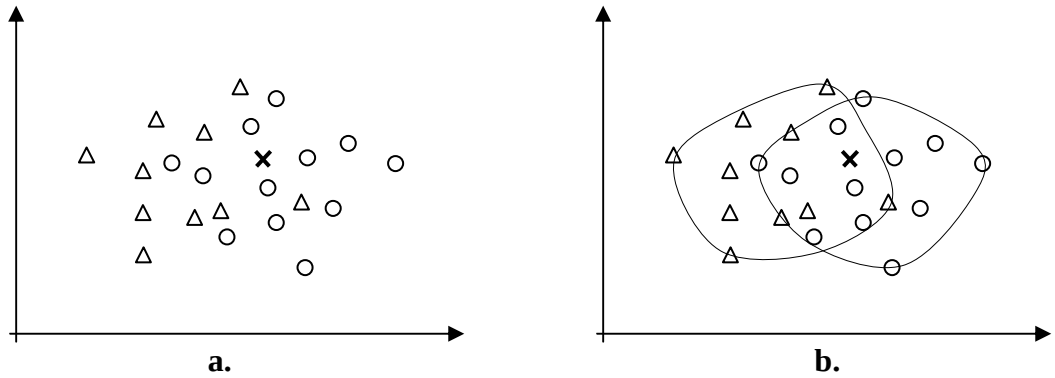


Figure 4.3 : Exemple de deux classes (classe *rond* et classe *triangle*) et une nouvelle donnée (*croix*) à classer. (a) la nouvelle donnée serait attribuée à une des deux classes. (b) approche one-class SVM : la nouvelle donnée serait attribuée aux deux classes.

Prenons maintenant le cas où toutes les données d'une classe se situent à l'intérieur du contour de l'autre classe. Dans ce cas, il est clair que ces deux classes doivent en pratique n'en former qu'une seule, c'est à dire être « fusionnées ». La frontière de la classe « englobante » pourrait alors être assimilée à celle de la classe résultante de la fusion. Dans les cas intermédiaires, c'est-à-dire où une proportion significative (à préciser à l'aide d'un seuil) d'une classe donnée se situe à l'intérieur du contour d'une autre, alors on décide que ces deux classes ne doivent en former qu'une seule, et nous appliquons ce processus de fusion. C'est ce qui est pratiqué dans notre application lors de la phase d'apprentissage.

Le second avantage que présente l'approche one-class SVM est illustrée dans la **Figure 4.4** (a). Ici aussi, on y voit deux classes (*rond* et *triangle*) et une nouvelle donnée *croix* à classer. Si on essaye de résoudre le problème par une approche simple, telle que l'application d'une métrique, il faut adjoindre à celle-ci des mécanismes permettant de gérer d'une part les rejets d'ambigüité (la donnée présentée se trouve à des distances très semblables de plusieurs classes, d'où difficulté à l'affecter) et d'autre part les rejets de distance : la donnée se trouve « trop loin » (notion qu'il faut alors quantifier) de l'ensemble des classes connues. Dans l'exemple ci-dessous, en appliquant une simple métrique, la donnée *croix* serait attribuée à une des deux classes existantes, en dépit d'une distance apparemment trop grande. Mais si on résout ce problème de classification par une approche one-class, illustrée dans la **Figure 4.4** (b), on pourrait voir que la donnée *croix* ne serait attribuée à aucune des deux classes, du fait que cette donnée se situe à l'extérieur de leurs contours de décision respectifs. Dans ce cas, on estimera que la nouvelle donnée représente une observation « inédite » qui n'appartient à

aucune classe connue, et pourra être considérée comme une « nouveauté ». De la sorte, nous pourrons créer une « nouvelle » classe à laquelle appartient la donnée *croix*. En utilisant ce principe dans notre étude, nous pouvons déterminer lors de la phase de reconnaissance, qu'un ensemble d'observations (une signature) issues d'un vêtement donné représente une « nouveauté » parmi les classes existantes et ainsi créer une nouvelle classe qui sera ajouter à notre base de connaissance. Ceci permettra alors d'effectuer un apprentissage en ligne.

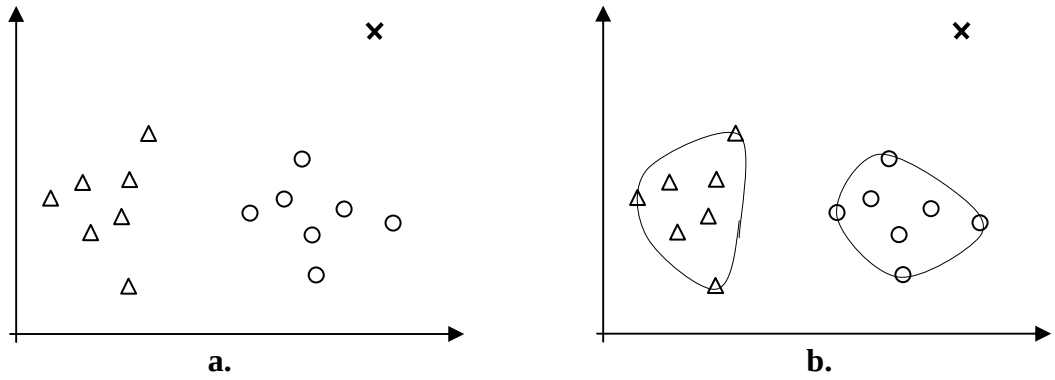


Figure 4.4 : Exemple de deux classes (classe *rond* et classe *triangle*) et une nouvelle donnée (*croix*) à classer. **(a)** la nouvelle donnée serait attribuée à une des deux classes. **(b)** approche one-class SVM : la nouvelle donnée ne serait attribuée à aucune des deux classes.

Dans le cadre de nos expérimentations, nous rappelons que nous disposons d'une base de test de 54 individus (nous présumons alors ne pas savoir comment ils sont vêtus). Dans ce qui suit, nous commençons par présenter notre approche de reconnaissance d'individus et les procédures d'apprentissage et de classification dans un scénario « ensemble fermé ». Nous présentons à cette occasion la procédure d'apprentissage hors ligne durant laquelle le système est capable en amont de détecter automatiquement lorsqu'il y a des vêtements similaires et de fusionner les classes correspondantes. Nous présentons alors les résultats de reconnaissance de notre approche par identification du « modèle d'apparence » d'individu (dont on rappelle qu'elle est définie par la combinaison de vêtements du haut et bas), résultats que nous comparons à ceux de l'identification menée grâce à l'apparence globale. Enfin, nous présentons les processus de traitement permettant d'effectuer la reconnaissance d'individus dans le scénario « ensemble ouvert», puis dans le scénario « ensemble vierge ».

4.5 Reconnaissance d'individus dans un scénario « ensemble fermé »

La **Figure 4.5** illustre la procédure d'apprentissage hors ligne des individus. Celle-ci s'effectue en deux étapes. La première étape, mise en œuvre après avoir appliqué les prétraitements définis précédemment (détection et segmentation d'individu et de scission de corps) et extrait de chaque individu ses signatures *haut* et *bas*, consiste à calculer la fonction contour de chaque signature qui délimite ses frontières et en modélise ainsi la classe. Ainsi, après avoir traité les 54 séquences d'apprentissage, on obtient 54 classes *haut* et 54 classes *bas*. Cependant, comme nous l'avons déjà indiqué, nous avons volontairement fait porter à certains individus les mêmes vêtements (même pull, même pantalon ou les deux), ceci afin d'évaluer nos algorithmes de reconnaissance dans des conditions réalistes. En effet, dans les cas pratiques de vidéosurveillance, aucune hypothèse ne peut être posée quant au fait que les personnes filmées porteront toutes des vêtements différents. A l'inverse, la variabilité entre

individus peut être assez faible de ce point de vue. On peut par exemple voir sur la **Figure 4.2** que dans notre base de données les individus #5, #25 et #30 portent le même pull (pull col v bleu foncé avec une bande belge au niveau du torse), de même que les individus #7 et #40 (pull rouge, col à fermeture éclair), vêtement lui-même similaire (sans être parfaitement identique) à celui porté par les individus #12 et #39 (pull rouge col v). Dans le même ordre d'idée, on peut également y constater que les individus #9, #10, #11 et #12 portent le même pantalon. Dans ce dernier exemple, il apparait clair que faire l'apprentissage des signatures de la partie *bas* en supposant avoir 54 différentes classes serait mal à propos. En effet, il est clair que les vêtements du bas portés par ces quatre individus ne devraient pas constituer quatre différentes classes mais une seule, puisqu'il s'agit du même pantalon. Il s'agit donc d'un cas sur lequel notre procédure de fusion devrait s'appliquer (de même que sur tout ou partie des similitudes évoquées auparavant).

Dans notre base de données, ayant initialement extrait 54 signatures de vêtements du bas, on disposera donc de 54 classes initiales *bas* (notées $\{IB_i\}$ avec ici $1 \le i \le 54$). A l'issue du processus de fusion, nous disposerons au final d'un nombre $N_{bas} \le 54$ de classes *bas* (notées $\{B_j\}$ avec $1 \le j \le N_{bas}$). Par exemple, les classes initiales *bas* $\{IB_9\}, \{IB_{10}\}, \{IB_{11}\}$ et $\{IB_{12}\}$ seront *a priori* fusionnées pour ne former qu'une seule classe finale *bas* $\{B_k\}$, $1 \le k \le N_{bas}$. Lors de la phase de reconnaissance, lorsqu'un de ces quatre individus apparaitra dans la scène, il sera identifié comme celui portant le pantalon représenté par cette classe finale $\{B_k\}$. Bien entendu, la même procédure de fusion sera appliquée aux classes initiales *haut* (notées $\{IH_i\}$ avec $1 \le i \le 54$) pour obtenir un nombre de classes finales après opération $N_{haut} \le 54$ (classes notées $\{H_j\}$ avec $1 \le j \le N_{haut}$). Ainsi, les classes initiales *haut* $\{IH_5\}, \{IH_{25}\}$ et $\{IH_{30}\}$seront *a priori* fusionnées pour ne former qu'une seule classe finale *haut* $\{H_m\}$ ($1 \le m \le N_{haut}$). À la fin de la procédure de fusion, nous obtenons donc un ensemble de N_{haut} classes finales *haut* et un ensemble de N_{bas} finales *bas*. La dernière procédure appliquée lors de la phase d'apprentissage consiste ensuite à reconstituer les combinaisons (*haut* + *bas*) de tous les individus observés et à vérifier s'il y a des « doubles ». Le cas échéant, ceux-ci sont regroupés. Finalement, chaque combinaison spécifique constitue un « modèle d'apparence » décrivant un individu (ou plusieurs « similaires ») et qui servira à le caractériser. Les grandes lignes ayant été tracées, nous détaillons dans ce qui suit les différentes procédures effectuées lors de l'apprentissage hors ligne.

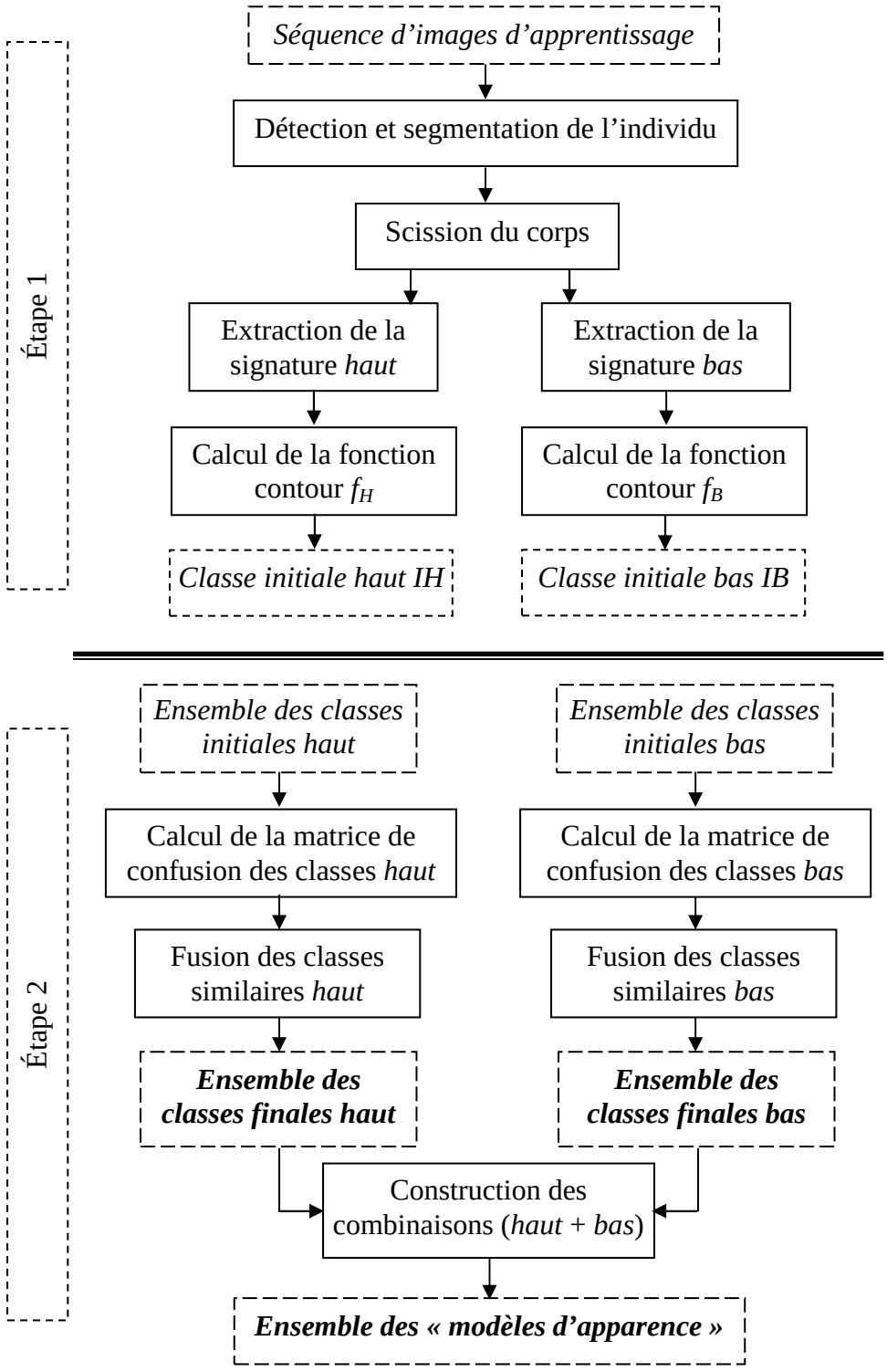


Figure 4.5 : Procédure d'apprentissage hors ligne.

4.5.1 Construction des classes initiales

L'objectif est de définir une fonction de décision pour chaque signature d'apprentissage, qui constituera alors une classe initiale. Après que toutes les signatures d'apprentissage aient été enregistrées, une fonction de décision est calculée pour chacune d'entre elles (procédure appliquée pour la partie *haut* et pour la partie *bas* de chacun des 54 individus enregistrés dans la base). On obtient ainsi 54 fonctions contour f_{IH_i} des classes initiales *haut*, et 54 fonctions contour f_{IB_i} des classes initiales *bas*, avec $1 \le i \le 54$. La **Figure 4.6** illustre l'exemple de quatre classes initiales *bas* et de leurs fonctions contour respectives. La définition de ces fonctions de décision est précisée dans ce qui suit.

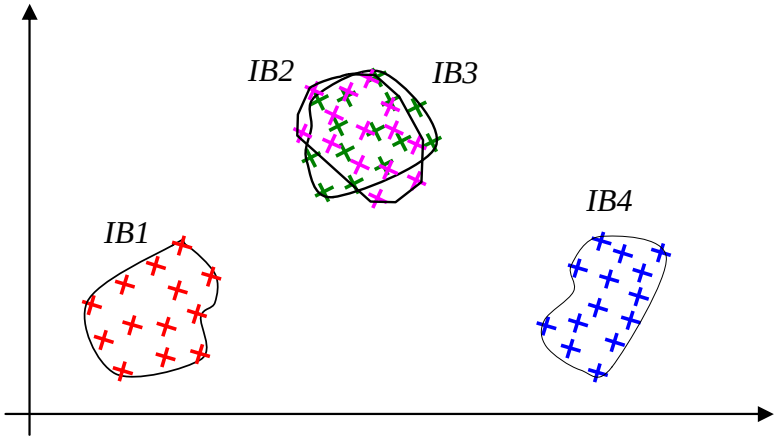


Figure 4.6 : Exemple de quatre classes initiales *bas* et de leurs fonctions contour respectives.

4.5.2 Classification one-class SVM

Comme nous l'avons présenté dans le chapitre 2, la classification one-class SVM est une méthode basée noyau [CorV95] [HaySTA00] [UnnRJ3], qui construit un hyperplan définissant la frontière d'une classe en utilisant uniquement un ensemble de données d'apprentissage positives. Ainsi, afin de définir la frontière de la signature d'apprentissage d'un vêtement donné, on utilisera uniquement l'ensemble des vecteurs de caractéristiques qui la constituent. Pour 54 signatures *haut* on utilisera 54 classifieurs one-class. De même, pour 54 signatures *bas* on utilisera également 54 classifieurs one-class. Pour une classe donnée, si l'ensemble des données ne contient pas des valeurs atypiques (outliers), le one-class SVM calcule un hyperplan (dans l'espace de Hilbert à noyau reproduisant RKHS) qui possède une marge maximum $(\mathrm{w} \cdot \phi(\mathrm{x}_i) = b)$ pour séparer l'ensemble des données des origines. On rappelle ici que ϕ est la fonction de projection non linéaire entre l'espace d'entrée (espace des caractéristiques qui est dans notre cas constitué de trois valeurs [*moyenne R, moyenne G, moyenne B*] de l'espace colorimétrique RGB) et l'espace RKHS, w est le vecteur de pondération et b est la marge associée à l'hyperplan (pour rappel, voir **Figure 2.22**, au chapitre 2).

On rappelle que sur chaque image d'une séquence d'apprentissage d'un individu donné, on extrait un vecteur de caractéristiques de chacune de ses parties *haut* et *bas*, correspondant respectivement à son pull et à son pantalon. Cependant, il est possible qu'une de ces parties (ou les deux) soit mal segmentée, c'est-à-dire qu'elle ne corresponde pas avec précision à la région de l'image contenant le vêtement concerné. Dans ce cas, le vecteur de caractéristiques

extrait ne correspondrait pas à l'élément désiré et serait alors une observation « aberrante » ou « outlier ». Cette éventualité implique de devoir relâcher la contrainte imposant que la valeur de projection d'un vecteur x_i sur l'hyperplan défini par w soit strictement supérieure à b (équation 2.8, 2.9 et 2.10, au chapitre 2).

Après les développements présentés dans le chapitre 2, la fonction de décision peut être écrite comme suit :

$$f(x) = w \cdot \phi(x) - b = \left(\sum_{j=1}^{n} \alpha_j \cdot K(x, x_j) \right) - b$$

où x_j est un vecteur support de la classe.

De cette façon, pour une classe donnée C dont le contour est défini par la fonction de décision f_C, un vecteur de caractéristiques x est classé comme appartenant à cette classe si $f_C(x) \geq 0$, et donc classé comme n'appartenant pas à C si $f_C(x) \leq 0$.

Nous utilisons ici une fonction à noyau gaussien, c'est-à-dire :

$$K(\mathrm{x}_i, \mathrm{x}_j) = \exp(-\left\| \mathrm{x}_i - \mathrm{x}_j \right\|^2 / 2\sigma^2)$$

C'est sur ce modèle que sont construites les fonctions contour associées aux classes *haut* et *bas* lors de la phase d'apprentissage.

4.5.3 Fusion des classes

Afin de produire des classes qui soient suffisamment « distinctes » pour éviter les problèmes associés à la confusion lors de l'étape de classification, il importe d'être capable de mesurer le degré « d'enchevêtrement » de deux classes données, et de disposer d'un mécanisme de fusion efficace. Le processus mis en œuvre repose donc sur deux étapes.

La première consiste à calculer le degré de « similarité » entre deux classes. Afin d'illustrer la démarche, notons IB_1 et IB_2 deux classes initiales *bas* dont les fonctions de décision sont respectivement f_{IB_1} et f_{IB_2}. Afin de calculer leur degré de similarité, on calcule la valeur v_1 qui représente le pourcentage des vecteurs de caractéristiques constituant la classe IB_1 qui se situent à l'intérieur du contour de la classe IB_2, puis la valeur v_2 qui représente le pourcentage des vecteurs de caractéristiques constituant la classe IB_2 qui se situent à l'intérieur du contour de la classe IB_1. Ces valeurs constituent la mesure recherchée de « l'enchevêtrement » entre ces deux classes. Il est clair que si l'une de ces valeurs était égale à 100% (ce qui signifie que l'une des classes est incluse dans l'autre, on ne saurait aborder la classification (reconnaissance) en considérant deux classes distinctes. Mais il faudrait les fusionner afin de constituer une seule et même classe. Dans les cas où cette inclusion n'est pas totale, mais « significative » et matérialisée par le dépassement d'un seuil S par v_1 ou v_2, on estimera que les classes IB_1 et IB_2 sont suffisamment « proches » et donc suffisamment « similaires » pour être fusionnées. Pour fusionner deux (ou plusieurs) classes, on calcule un contour de décision englobant l'ensemble des vecteurs de caractéristiques de ces classes. La **Figure 4.7** (a) illustre le cas de quatre classes initiales *bas* IB_1, IB_2, IB_3 et IB_4. Comme on peut le voir, les classes IB_2 et IB_3 sont très largement enchevêtrées. Par conséquent, elles vont fusionner pour constituer

une seule classe finale *bas*. Dans cet exemple, les classes initiales IB_1 et IB_4 ne sont enchevêtrées avec aucune autre classe, et formeront chacune une classe finale.

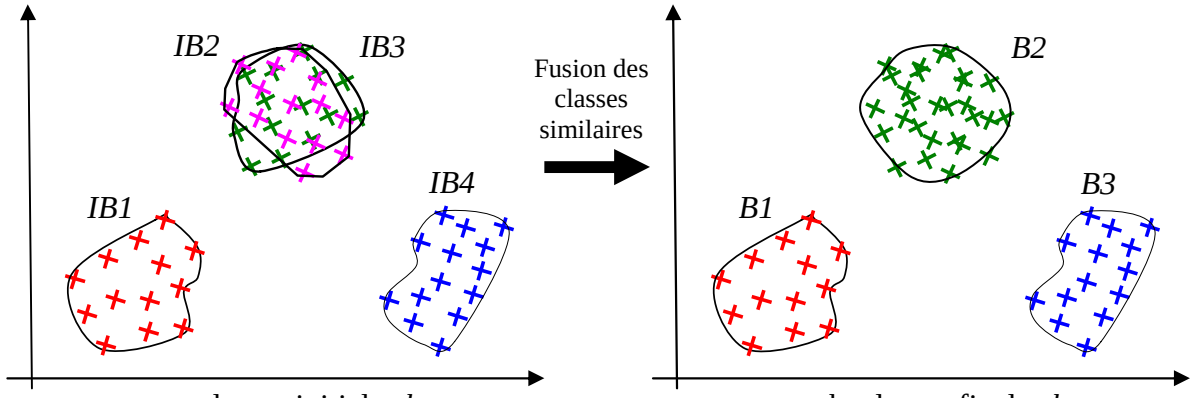


Figure 4.7 : Exemple de fusion des classes initiales *bas* et formation des classes finales *bas*.

Dans ce qui suit, nous parlerons uniquement des classes *haut*, sachant que le même procédé est appliqué simultanément aux classes *bas*. On notera alors IH_i (classe initiale *haut* i) le vêtement du haut porté par l'individu #i, et IB_i (classe initiale *bas* i) le vêtement du bas porté par l'individu #i.

Afin d'aborder la fusion de nos N_{ini} ($N_{ini} = 54$) classes initiales *haut*, on commence par effectuer une classification croisée, c'est-à-dire que les vecteurs de caractéristiques de chaque classe seront testés par chacun des N_{ini} classifieurs one-class afin de déterminer si ces vecteurs se situent à l'intérieur du contour d'une autre classe que la leur. On calcule ainsi une matrice de confusion notée M_H de taille $N_{ini} \times N_{ini}$. Le coefficient $M_H(i, j)$ de cette matrice contient le pourcentage des vecteurs de caractéristiques de la classe IH_i situés à l'intérieur de la fonction contour de la classe IH_j, avec $1 \le i \le N_{ini}$ et $1 \le j \le N_{ini}$. Pour calculer ce pourcentage, chaque vecteur caractéristique x_{m,IH_i} de la classe IH_i, avec $1 \le m \le n_{IH_i}$ (n_{IH_i} étant le nombre de vecteurs de caractéristiques de la classe considérée) est testé par la fonction contour f_{IH_j}. Si la condition $f_{IH_j}\left(x_{m,IH_i}\right) \ge 0$ est vraie, cela signifie que x_{m,IH_i} se situe à l'intérieur de la classe IH_j. Si c'est le cas inverse qui se vérifie ($f_{IH_j}\left(x_{m,IH_i}\right) < 0$), cela signifie x_{m,IH_i} se situe à l'extérieur de la classe IH_j. Ainsi, plus le pourcentage est élevé, plus la classe IH_i n'apparaît que comme un « sous-ensemble » ou une presque réplique (si « l'enchevêtrement » est mutuel) de IH_j : elle est « enchevêtrée » dans IH_j. A la lumière de l'interprétation précédente, il est clair que la matrice M_H n'est pas forcément symétrique, loin s'en faut.
Une fois toutes les valeurs de toutes les cellules de la matrice de confusion M_H calculées, chaque cellule de chaque colonne de la matrice est comparée à une valeur seuil S_{haut}. Ainsi, si la valeur du coefficient $M_H(i, j)$ est supérieure à S_{haut}, alors les classes IH_i et IH_j seront considérées comme « similaires » et seront fusionnées après que toutes les colonnes aient été

testées. En d'autres termes, si $M_H(i,j) \geq S_{haut}$ ou $M_H(j,i) \geq S_{haut}$ alors les classes IH_i et IH_j seront fusionnées. Il est à noter que ce procédé est appliqué à l'ensemble des classes satisfaisant la condition mentionnée. En toute rigueur, cette opération est réalisée en deux passes : la première sert à repérer les classes qui devront être assemblées en évaluant la condition évoquée plus haut. A la fin de celle-ci, nous disposons d'une liste de classes initiales *haut* à fusionner. Etant donné F_{IH} l'ensemble des indices des classes à fusionner, une fonction contour f_H est alors calculée en utilisant tous les vecteurs de caractéristiques appartenant à $\bigcup_{i \in F_{IH}} \{IH_i\}$, définissant ainsi une classe finale *haut* notée *H*. Bien entendu, une classe initiale qui ne serait similaire à aucune autre constituera elle-même une classe finale. A la fin de cette procédure de fusion, nous obtiendrons un nombre de N_{haut} classes finales *haut*, avec $N_{haut} \leq N_{ini}$. Comme précisé plus tôt, les mêmes procédés sont simultanément appliqués aux classes initiales *bas* IB pour obtenir un nombre de N_{bas} ($N_{bas} \leq N_{ini}$) de classes finales *bas B*. Dans l'exemple de la **Figure 4.7**, on peut voir qu'on obtient trois classes finales *bas* à partir de quatre classes initiales.

Le **Tableau 4.13** illustre la matrice de confusion M_H obtenue sur nos 54 classes initiales *haut*. Afin de pouvoir fusionner les classes similaires, il nous faut maintenant fixer le seuil S_{haut} qui définit à partir de quel niveau de similarité (ou de confusion) des classes vont être fusionnées. Nous déterminons cette valeur de façon empirique (en observant les coefficients de la matrice de confusion) et nous la fixons à 80%. Les classes finales *haut* obtenues en appliquant la fusion en utilisant cette valeur sont présentées dans la **Tableau 4.6**. On voit qu'on obtient 31 classes finales *haut*. A titre d'exemple, on peut voir que le pourcentage des vecteurs de caractéristiques de la classe initiale IH_2 situés à l'intérieur de la fonction contour de la classe IH_{27} est de 81% (qui est donc supérieur au seuil de 80%), ces deux classes initiales sont alors fusionnées pour donnée naissance à une classe finale notée H_2. On compare maintenant les résultats de cette procédure de fusion automatique avec ce que nous avions fait de façon manuelle présentés dans le **Tableau 4.1**. On peut voir que le système a pris la même décision que « l'expert humain », ce qui démontre que le système détecte et reconnait automatiquement les vêtements du haut identiques au sein du jeu de données présenté. Au-delà des fusions réalisées sur les classes effectivement identiques, d'autres ont été faites à bon escient sur des classes d'aspects très similaires, ainsi :

- la procédure automatique a fusionné le vêtement du haut porté par les individus #7 et #40 (pull rouge col à fermeture éclair) avec celui porté par les individus #12 et #39 (pull rouge col v) ainsi qu'avec celui porté par l'individu #34 (chemise rouge).
- On peut voir également que le vêtement du haut porté par l'individu #17 (pull noir col roulé), celui porté par l'individu #18 (pull noir col rond) et celui porté par les individus #52 et #54 (pull noir col v) ont été fusionnés en une seule classe finale.
- Enfin, on constate qu'ont été fusionnés le vêtement du haut porté par les individus #24 et #35 (pull marron foncé col haut à boutons) et le vêtement du haut porté par l'individu #41 (pull marron foncé col v).

On peut estimer que les fusions automatiques effectuées ici avec un seuil S_{haut} = 80% sont tout à fait adéquates. En effet, en plus d'avoir fusionné les vêtements connus pour être identiques, le système a fusionné des vêtements du haut extrêmement similaires (la distinction étant difficile à faire même pour « l'expert humain »).

On rappelle ici que pour fusionner deux classes, il suffit que le pourcentage des données de l'une d'entre elle soit supérieur eu seuil fixé. A titre d'exemple, on peut constater sur le **Tableau 4.13** (matrice de confusion des classes initiales *haut*) que les classes initiales *haut* IH_{12} et IH_{34} seront fusionnées, car le pourcentage des données de IH_{34} situées dans IH_{12} est de 93% (supérieur au seuil fixé S_{haut} = 80%). On remarque ici que le pourcentage des données de IH_{12} situées dans IH_{34} n'est que de 79% (inférieur au seuil).

Un autre essai a été réalisé en prenant $S_{haut} = 70\%$. De cette façon, d'autres classes ont alors été fusionnées. Les classes finales obtenues avec cette valeur sont réduites au nombre 28 et sont présentées dans le **Tableau 4.7**. Parmi les nouvelles fusions effectuées par rapport à la valeur de seuil précédente, on notera :

- le vêtement du haut porté par #1 (T-shirt de couleur entre beige et gis) et celui porté par #36 et #53 (gilet de couleur similaire) on été fusionnés.
- On constate également la fusion du vêtement du haut porté par #6 (pull bleu foncé col v) avec celui porté par #8 et #38 (pull col rond bleu foncé avec des bandes blanches sur les côtés) et celui porté par #21 et #44 (pull bleu foncé col à boutons).

Même si ces fusions restent tout à fait acceptables, nous estimons cependant que les résultats obtenus avec S_{haut} = 80% représentent un meilleur compromis pour obtenir le rassemblement des classes similaires tout en gardant un haut niveau de discrimination. Ainsi, nous utilisons ici $S_{haut} = 80\%$ pour la suite du processus d'apprentissage, débouchant ainsi sur 31 classes finales *haut*.

	Classes finales *haut*											
	H1	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**	**H11**	**H12**
Individus	#1	#2 #27	#3 #37	#4 #49	#5 #25 #30	#6	#7 #12 #34 #39 #40	#8 #38	#9	#10 #47 #51	#11 #42	#13

	Classes finales *haut*											
	H13	**H14**	**H15**	**H16**	**H17**	**H18**	**H19**	**H20**	**H21**	**H22**	**H23**	**H24**
Individus	#14	#15	#16	#17 #18 #52 #54	#19 #48 #50	#20 #43	#21 #44	#22	#23	#24 #35 #41	#26	#28

	Classes finales *haut*						
	H25	**H26**	**H27**	**H28**	**H29**	**H30**	**H31**
Individus	#29	#31	#32	#33	#36 #53	#45	#46

Tableau 4.6 : Classes finales *haut*. S_{haut} = 80%.

	Classes intiales *haut*																																																				
Classes intiales *haut*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**
1	**95**	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0	0	0	0	59	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	67	0
2	0	**95**	0	0	1	0	0	0	0	0	0	0	36	0	0	0	0	2	0	59	0	0	0	0	1	0	**81**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	30	0	22	0	0	0	0	0	0	0	0	0
3	0	0	**94**	0	0	0	8	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	**86**	0	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	**92**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**91**	0	0	0	0	0
5	0	1	0	0	**94**	0	0	2	0	0	69	0	15	0	1	0	0	0	2	2	0	0	20	0	**96**	1	1	0	0	79	0	0	0	0	0	0	0	5	0	0	0	53	1	0	1	0	0	3	0	9	0	0	0	0
6	0	0	0	0	0	**96**	0	60	0	0	0	0	0	17	30	56	51	0	17	0	64	0	0	4	0	2	0	0	0	0	28	0	0	0	0	0	0	58	0	0	4	0	0	50	0	0	0	21	0	8	0	0	0	0
7	0	0	0	0	0	0	**94**	0	0	0	0	**86**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**82**	0	0	0	0	77	**87**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	1	0	0	2	77	0	**94**	0	0	0	0	0	0	56	5	0	1	59	1	70	0	2	0	2	17	0	0	0	0	13	0	0	0	0	0	0	**89**	0	0	0	0	3	72	16	0	0	63	0	57	0	0	0	0
9	0	0	0	0	0	0	0	0	**94**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	**94**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**88**	0	0	0	**83**	0	2	0
11	3	0	0	0	61	0	0	0	0	0	**94**	0	18	0	0	0	0	0	0	0	0	0	0	0	47	0	0	0	0	53	0	0	0	0	0	0	0	0	0	0	0	**88**	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	1	0	0	0	**91**	0	0	0	0	**94**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	79	0	0	0	0	**87**	**81**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	59	0	0	1	0	0	0	0	0	4	0	**93**	0	0	0	0	0	0	40	0	0	0	0	6	0	69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	26	0	16	0	0	0	0	0	0	0	3	0
14	0	0	0	0	0	19	0	0	0	0	0	0	0	**97**	0	20	31	34	0	0	0	0	0	69	0	0	0	0	0	0	0	0	0	0	41	0	0	0	0	0	69	0	0	0	2	0	0	0	0	0	0	9	0	5
15	0	20	0	0	1	50	0	56	0	0	0	0	1	0	**96**	1	0	1	69	60	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0	0	0	62	0	68	0	0	67	0	68	0	0	0	0
16	0	0	0	0	0	66	0	4	0	0	0	0	0	47	0	**99**	77	0	0	0	3	0	0	62	0	0	0	0	0	0	0	0	0	0	51	0	0	0	0	0	23	0	0	10	1	0	0	0	0	0	0	58	0	60
17	0	0	0	0	0	57	0	0	0	0	0	0	0	44	0	79	**100**	**87**	0	0	0	0	0	66	0	0	0	0	0	0	0	0	0	0	64	0	0	0	0	0	36	0	0	8	0	0	0	0	0	0	0	78	0	**81**
18	0	1	0	0	0	49	0	0	0	0	0	0	0	58	0	71	**82**	**99**	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0	41	0	0	0	0	0	39	0	0	0	16	0	0	0	0	0	0	**81**	0	**83**
19	0	0	0	0	0	50	0	63	0	0	0	0	0	0	66	0	0	0	**98**	6	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	61	0	0	0	0	6	0	68	0	0	**100**	0	**90**	0	0	0	0
20	0	67	0	0	0	0	0	28	0	0	0	0	59	0	63	0	0	1	39	**97**	0	0	0	0	0	0	67	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	**90**	0	55	0	0	61	0	53	0	0	0	0
21	0	0	0	0	0	56	0	50	0	0	0	0	0	0	0	40	31	0	0	0	**100**	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	67	0	0	0	0	0	**93**	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	35	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	**93**	0	26	28	0	0	0	2	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	4	0	11	0	0	0	0
24	0	0	0	0	0	1	0	0	0	0	0	0	0	62	0	22	32	17	0	0	0	0	0	**97**	0	0	0	0	0	0	0	0	0	0	**82**	0	0	0	0	0	**90**	0	0	0	1	0	0	0	0	0	0	31	0	20
25	3	0	0	0	**89**	0	0	1	0	0	57	0	11	0	0	0	0	0	1	0	0	0	14	0	**92**	1	1	0	0	**89**	0	0	0	0	0	0	0	4	0	0	0	62	0	0	0	0	0	2	0	7	0	0	0	0
26	0	0	0	0	0	3	0	16	0	0	0	0	0	0	2	0	0	0	9	0	0	0	38	0	0	**97**	0	0	0	0	11	0	0	0	0	0	0	46	0	0	0	0	0	0	0	0	0	23	0	38	0	0	0	0
27	0	**91**	0	0	0	0	0	0	0	0	0	0	43	0	0	0	0	0	0	48	0	0	0	0	0	0	**96**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	0	33	0	0	0	0	0	0	0	0	0
28	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	**96**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**94**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	2	0	0	0	**80**	0	0	0	0	0	64	0	4	0	0	0	0	0	0	0	0	0	13	0	**96**	0	0	0	0	**94**	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	5	0	0	0	0
31	0	0	0	0	0	43	0	24	0	0	0	0	0	0	0	0	0	0	0	0	19	0	1	0	0	5	0	0	0	0	**96**	0	0	0	0	0	0	29	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**92**	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	**94**	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	**100**	0	0	0	0	**93**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**96**	0	0	0	0	**87**	**97**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0	0	0	0	44	0	40	59	37	0	0	0	0	0	**93**	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	75	0	0	0	0	0	0	0	0	0	0	62	0	48
36	63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**95**	0
37	0	0	**89**	0	0	0	0	0	0	0	0	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	8	57	0	**80**	0	0	0	0	0	0	47	0	0	0	57	0	67	0	23	0	4	44	0	0	0	1	22	0	0	0	0	0	0	**97**	0	0	0	0	0	71	15	0	0	66	0	68	0	0	0	0
39	0	0	0	0	0	0	**99**	0	0	0	0	**98**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**90**	0	0	0	0	**94**	**90**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	**95**	0	0	0	0	71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	72	0	0	0	0	73	**94**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	2	0	0	0	0	0	0	0	60	0	36	35	44	0	0	0	0	0	**96**	0	0	0	0	0	0	0	0	0	0	**82**	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	13	0	5
42	1	0	0	0	67	0	0	0	0	0	**85**	0	17	0	0	0	0	0	0	0	0	0	0	0	59	0	0	0	0	56	0	0	0	0	0	0	0	0	0	0	0	**94**	0	0	0	0	0	0	0	0	0	0	0	0
43	0	67	0	0	0	0	0	28	0	0	0	0	38	0	63	0	0	2	20	**89**	0	0	0	0	0	0	65	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	**97**	0	68	0	0	53	0	44	0	0	0	0
44	0	0	0	0	0	76	0	63	0	0	0	0	0	2	0	38	28	10	0	0	**88**	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	70	0	0	0	0	0	**98**	0	0	0	0	0	0	0	1	0	1
45	0	35	0	0	0	40	0	37	0	0	0	0	11	2	65	6	13	20	40	58	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	58	0	**95**	0	0	55	0	39	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	**93**	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**93**	0	0	0	**90**	0	0	0
48	0	2	0	0	2	43	0	68	0	0	0	0	0	1	61	1	0	1	**82**	22	0	0	2	0	1	30	0	0	0	0	0	0	0	0	0	0	0	68	0	0	0	0	21	0	66	0	0	**96**	0	**85**	0	0	0	0
49	2	0	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**93**	0	0	0	0	0
50	0	1	0	0	14	41	0	67	0	0	1	0	0	0	47	1	0	1	**83**	20	0	0	17	0	11	25	0	0	0	1	0	0	0	0	0	0	0	64	0	0	0	1	11	0	54	0	0	**84**	0	**94**	0	0	0	0
51	0	0	0	0	0	0	0	0	0	**90**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**96**	0	0	0	**96**	0	0	0
52	0	0	0	0	0	14	0	0	0	0	0	0	0	33	0	73	**83**	78	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	55	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	**98**	0	**97**
53	72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**86**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**93**	0
54	0	0	0	0	0	18	0	0	0	0	0	0	0	31	0	65	**91**	**90**	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	51	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	**99**	0	**99**

Tableau 4.13 : Matrice de confusion des classes initiales *haut*.

	Classes finales *haut*											
	H1	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**	**H11**	**H12**
Individus	#1 #36 #53	#2 #27	#3 #37	#4 #49	#5 #25 #30	#6 #8 #21 #38 #44	#7 #12 #34 #39 #40	#9	#10 #47 #51	#11 #42	#13	#14

	Classes finales *haut*											
	H13	**H14**	**H15**	**H16**	**H17**	**H18**	**H19**	**H20**	**H21**	**H22**	**H23**	**H24**
Individus	#15	#16	#17 #18 #52 #54	#19 #48 #50	#20 #43	#22	#23	#24 #35 #41	#26	#28	#29	#31

	Classes finales *haut*			
	H25	**H26**	**H27**	**H28**
Individus	#32	#33	#45	#46

Tableau 4.7 : Classes finales *haut*. S_{haut} = 70%.

Passons maintenant aux vêtements du bas. Le **Tableau 4.14** représente le contenu de la matrice de confusion M_B. Afin de pouvoir fusionner les classes similaires, nous devons ici aussi fixer la valeur d'un seuil noté S_{bas}. Comme nous l'avons évoqué, les vêtements du bas portés par nos 54 individus sont moins dissemblables et contiennent moins de couleurs que les vêtements du haut, et sont pour la plupart des jeans bleus ou des pantalons de couleur plus ou moins foncée. Par conséquent, on peut constater que les taux de confusion sont plus élevés pour les classes *bas* (M_B) que pour les classes *haut* (M_H). Aussi, nous fixons la valeur seuil $S_{bas} = 90\%$, obtenant ainsi après fusion 11 classes finales *bas* (**Tableau 4.8**).

La comparaison avec la fusion manuelle des vêtements du bas (**Tableau 4.2**) montre que les pantalons identiques ont été ici correctement fusionnés de manière automatique par le système. En plus de ceux-ci, des fusions ont été réalisées sur des éléments similaires. Par exemple :

- le pantalon porté par les individus #1, #2, #3, #4, #5, #53 et #54 et celui porté par les individus #45, #46, #47 et #48 ont été ici détectés comme similaires et ont été fusionnés, (il s'agit de deux pantalons noirs).
- le pantalon porté par les individus #31 et #32 et celui porté par les individus #33, #34, #35, #36, #3,7 #38 et #39 ont également été fusionnés : Il s'agit de deux jeans de couleur bleu foncé légèrement délavé au niveau des cuisses et des tibias, qui sont effectivement d'aspect très similaires.

Classes intiales *bas* (colonnes) / **Classes intiales *bas*** (lignes)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
1	99	98	96	96	97	0	0	0	0	0	0	0	21	18	0	1	3	0	0	0	0	2	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47	36	32	36	24	65	92	84	66	48	55	38	35	96	98
2	99	99	99	97	98	0	0	0	0	0	0	0	10	12	0	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	33	18	29	22	69	98	94	78	43	38	35	33	99	99
3	100	100	100	97	98	0	0	0	0	0	0	0	29	8	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	35	29	15	27	16	65	97	92	75	36	43	34	31	100	100
4	98	92	90	99	99	0	0	0	0	0	0	0	35	30	0	0	1	0	0	0	0	2	1	3	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	52	44	24	35	26	55	78	65	58	53	53	34	31	89	95
5	96	90	90	97	100	0	0	0	0	0	0	0	36	30	0	0	0	0	0	0	0	2	0	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	52	41	20	20	20	50	79	66	52	48	43	23	34	89	94
6	0	0	0	0	0	96	97	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	98	99	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	98	98	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	94	95	95	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	92	95	92	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	92	94	93	94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	92	94	93	94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	9	7	7	9	11	7	8	7	0	0	0	0	96	72	26	0	1	0	0	0	0	19	14	17	22	0	0	0	2	0	6	2	23	32	25	16	9	8	4	15	15	14	14	14	2	6	5	3	15	15	11	11	7	8
14	11	14	11	11	13	0	0	6	0	0	0	0	99	92	20	0	0	0	0	0	0	6	5	5	12	0	0	0	1	0	4	0	16	34	34	9	3	4	1	3	3	3	2	3	0	0	0	0	3	3	1	1	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	10	17	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	10	15	17	21	5	6	9	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	11	8	17	0	0	0	0	0	0
17	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	12	17	24	0	0	0	0	2	1
18	0	0	0	0	0	0	0	0	0	0	0	0	10	20	0	0	0	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	12	19	0	0	0	97	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	20	28	0	0	0	98	100	99	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	11	16	0	0	0	99	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	2	0	0	4	4	0	0	0	0	0	0	0	12	5	0	0	0	0	0	0	0	99	99	99	98	0	0	0	0	0	0	0	0	0	0	1	0	0	0	71	73	77	75	68	0	0	0	0	80	82	79	76	0	1
23	0	0	0	0	1	0	0	0	0	0	0	0	9	4	0	0	0	0	0	0	0	97	97	97	94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	78	71	80	82	71	0	0	0	0	68	79	62	63	0	0
24	1	0	0	2	2	0	0	0	0	0	0	0	8	4	0	0	0	0	0	0	0	95	96	98	88	0	0	0	0	0	0	0	0	0	0	1	0	0	0	82	73	71	70	77	0	0	0	0	71	73	68	63	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	7	21	0	0	0	0	0	0	0	93	91	90	98	0	0	0	0	0	0	0	0	2	1	0	0	0	0	56	63	56	60	68	0	0	0	0	56	60	66	57	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	20	29	0	0	0	0	0	0	0	0	0	0	0	99	96	99	100	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	21	30	0	0	0	0	0	0	0	0	0	0	0	99	98	99	99	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	19	26	0	0	0	0	0	0	0	0	0	0	0	94	91	95	96	93	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	23	35	0	0	0	0	0	0	0	0	0	0	0	89	81	89	96	87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	11	12	0	0	0	0	0	0	0	0	0	0	0	96	94	96	98	97	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	9	8	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98	90	99	100	100	66	72	63	67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	6	11	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	85	98	100	100	88	64	70	71	62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	10	12	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	81	92	98	100	90	78	80	84	84	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	7	3	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	75	85	92	95	82	78	79	79	79	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0	0	0	9	2	30	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	86	72	90	100	98	76	81	80	83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	1	0	0	3	3	0	0	0	0	0	0	0	9	5	23	0	0	0	0	0	0	1	0	1	2	0	0	0	0	0	65	60	78	88	85	98	94	88	88	1	1	0	0	0	0	0	0	0	1	1	0	0	0	1
37	0	0	0	0	0	0	0	0	0	0	0	0	8	7	30	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	73	78	85	93	90	96	99	93	92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	0	0	0	0	0	8	2	47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86	83	89	98	97	91	94	95	93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	1	0	0	0	0	0	0	0	8	4	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	88	82	91	98	98	97	98	99	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	40	25	20	40	44	0	0	0	0	0	0	0	9	2	0	0	0	0	0	0	0	69	74	81	73	0	0	0	0	0	0	0	0	0	0	2	0	0	0	97	99	96	95	97	1	6	6	1	99	99	87	87	35	43
41	31	20	17	27	31	0	0	0	0	0	0	0	9	3	0	0	0	0	0	0	0	69	74	80	72	0	0	0	0	0	0	0	0	0	0	1	0	0	0	96	98	96	94	98	0	3	4	1	98	99	90	91	32	39
42	35	20	18	30	32	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	76	71	88	67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	97	99	98	96	98	0	2	3	1	99	99	93	94	30	48
43	32	26	21	35	36	0	0	0	0	0	0	0	9	6	0	0	0	0	0	0	0	63	72	81	64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	100	100	99	99	0	4	6	1	100	100	97	97	48	46
44	34	20	15	27	29	0	0	0	0	0	0	0	4	1	0	0	0	0	0	0	0	69	78	85	63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96	97	97	96	97	0	1	3	0	97	97	97	97	30	35
45	73	68	70	53	55	0	0	0	0	0	0	0	1	0	0	18	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	78	97	98	0	0	1	1	77	66
46	99	99	99	97	97	0	0	0	0	0	0	0	9	2	0	10	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	4	1	5	1	84	99	98	93	9	7	6	6	89	88
47	90	88	89	75	77	0	0	0	0	0	0	0	11	4	0	11	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	3	2	3	2	93	91	100	99	7	4	5	5	81	88
48	91	87	89	73	76	0	0	0	0	0	0	0	4	0	0	11	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	98	93	98	97	2	1	1	1	83	87
49	38	28	21	37	40	0	0	0	0	0	0	0	11	1	0	0	0	0	0	0	0	71	76	83	67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	95	95	95	95	2	6	8	3	97	97	94	94	28	34
50	33	21	26	29	31	0	0	0	0	0	0	0	12	4	0	0	0	0	0	0	0	75	82	85	65	0	0	0	0	0	0	0	0	0	0	1	0	0	0	99	99	98	95	96	1	4	4	1	99	99	93	93	22	30
51	51	27	21	32	36	0	0	0	0	0	0	0	8	7	0	0	0	0	0	0	0	65	70	79	59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	92	95	94	97	97	0	3	11	2	97	97	99	99	61	41
52	44	27	17	24	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	63	80	78	56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	90	96	93	97	97	2	2	10	3	97	96	99	99	30	38
53	99	98	98	96	97	0	0	0	0	0	0	0	5	1	0	2	3	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	44	36	23	37	24	70	86	95	69	49	44	44	42	98	98
54	99	98	99	98	98	0	0	0	0	0	0	0	6	2	0	0	1	0	0	0	0	1	0	3	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	54	58	38	57	44	52	85	78	69	61	53	52	59	99	98

Tableau 4.14 : Matrice de confusion des classes initiales *bas*.

Là encore, les fusions effectuées sont tout à fait comparables à celles qu'effectuerait « l'expert humain ». Un essai complémentaire a été fait en fixant $S_{bas} = 80\%$. Les classes finales *bas* obtenues avec cette valeur se réduisent au nombre 10 et sont présentées dans la **Tableau 4.9**. Au titre des fusions supplémentaires réalisées, on note en particulier que le pantalon porté par les individus #22, #23, #24, et #25 et celui porté par les individus #40, #41, #42, #43, #44, #49, #50, #51, et #52 ont été fusionnées : il s'agit de deux pantalons de couleur marron foncé assez similaire. Nous gardons ici pour les classes *bas* $S_{bas} = 90\%$, qui apparaît comme un bon compromis « pratique », disposant ainsi de 11 classes finales *bas* pour la suite du processus d'apprentissage.

	Classes finales *bas*										
	B1	**B2**	**B3**	**B4**	**B5**	**B6**	**B7**	**B8**	**B9**	**B10**	**B11**
Individus	#1	#6	#9	#13	#15	#16	#18	#22	#26	#31	#40
	#2	#7	#10	#14		#17	#19	#23	#27	#32	#41
	#3	#8	#11				#20	#24	#28	#33	#42
	#4		#12				#21	#25	#29	#34	#43
	#5								#30	#35	#44
	#45									#36	#49
	#46									#37	#50
	#47									#38	#51
	#48									#39	#52
	#53										
	#54										

Tableau 4.8 : Classes finales *bas*. S_{bas} = 90%.

	Classes finales *bas*									
	B1	**B2**	**B3**	**B4**	**B5**	**B6**	**B7**	**B8**	**B9**	**B10**
Individus	#1	#6	#9	#13	#15	#16	#18	#22	#26	#31
	#2	#7	#10	#14		#17	#19	#23	#27	#32
	#3	#8	#11				#20	#24	#28	#33
	#4		#12				#21	#25	#29	#34
	#5							#40	#30	#35
	#45							#41		#36
	#46							#42		#37
	#47							#43		#38
	#48							#44		#39
	#53							#49		
	#54							#50		
								#51		
								#52		

Tableau 4.9 : Classes finales *bas*. S_{bas} = 80%.

Arrivé à ce stade de nos expérimentations, nous disposons de 31 classes finales *haut* et de 11 classes finales *bas*. On peut alors entamer la deuxième étape de l'apprentissage hors ligne, qui est de reconstituer les combinaisons de classes finales de vêtement, sous la forme de couples $\left(H_p + B_q\right)$ avec $1 \le p \le 31$ et $1 \le q \le 11$ pour chaque individu observé, ce qui permettra alors de disposer au final de la base des « modèles d'apparence » de tous les individus appris.

4.5.4 Construction de l'ensemble des modèles d'apparence

Durant cette phase de construction de l'ensemble des « modèles d'apparence », chaque individu est caractérisé par sa combinaison de vêtements *haut* et *bas*. Le mécanisme simple mis en place fait appel à une indirection. Plus précisément, lors de la phase de fusion des classes (qu'il s'agisse des classes *haut* ou *bas*), nous mettons en correspondance les indices des classes initiales avec les indices des classes finales générées. Si nous notons R_H et R_B les applications faisant correspondre les indices entre classe initiales et finales respectivement pour les classes *haut* et *bas*, un individu représenté initialement par le couple $\left(IH_i, IB_j\right)$ sera finalement représenté par $\left(H_{R_H(i)}, B_{R_B(j)}\right) = \left(H_p, B_q\right)$.

La **Figure 4.8** illustre l'exemple d'une procédure de construction des combinaisons pour quatre individus. On peut voir que les quatre individus portent des vêtements du haut différents, et que les individus #2 et #3 portent le même vêtement du bas (représenté par la classe finale *bas B2*) qui est différent de celui porté par l'individu #1 et celui porté par l'individu #4. Aussi, quatre combinaisons sont obtenues constituant ainsi quatre « modèles d'apparence », chacun décrivant un des quatre individus.

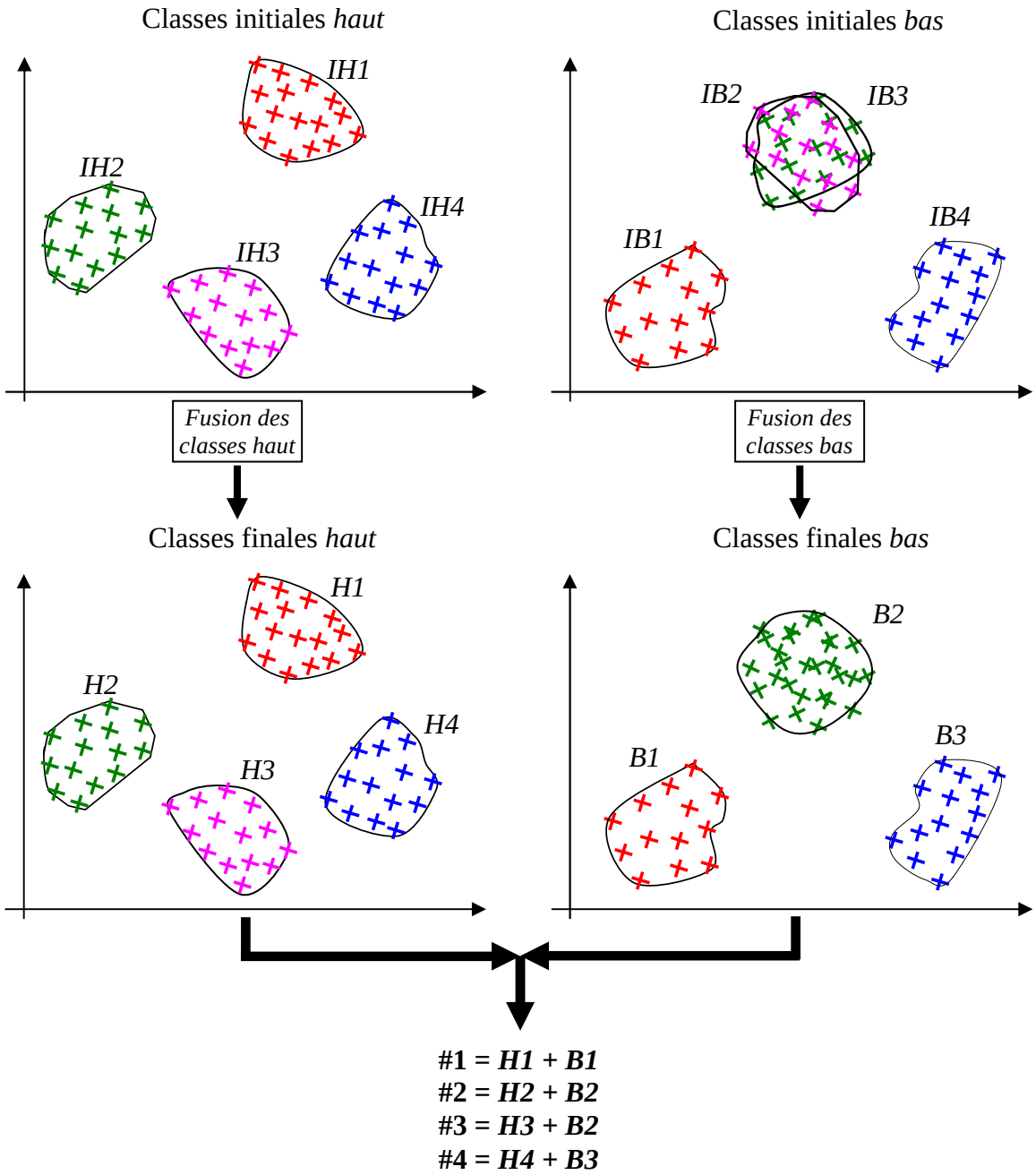


Figure 4.8 : Exemple de construction des combinaisons pour quatre individus.

Le même procédé est appliqué à chacun de nos 54 individus. Nous pouvons alors construire 54 combinaisons à partir des tableaux des classes finales *haut* et *bas* **Tableau 4.6** et **Tableau 4.8**. Le résultat est l'ensemble des « modèles d'apparence » et est présenté dans le **Tableau 4.10**. On rappelle que les classes finales ont été obtenues avec un seuil de fusion de 80% pour les classes *haut* et de 90% pour les classes *bas*.

Individu	Combinaison	Individu	Combinaison	Individu	Combinaison
#1	$H_1 + B_1$	#19	$H_{17} + B_7$	#37	$H_3 + B_{10}$
#2	$H_2 + B_1$	#20	$H_{18} + B_7$	#38	$H_8 + B_{10}$
#3	$H_3 + B_1$	#21	$H_{19} + B_7$	#39	$H_7 + B_{10}$
#4	$H_4 + B_1$	#22	$H_{20} + B_8$	#40	$H_7 + B_{11}$
#5	$H_5 + B_1$	#23	$H_{21} + B_8$	#41	$H_{22} + B_{11}$
#6	$H_6 + B_2$	#24	$H_{22} + B_8$	#42	$H_{11} + B_{11}$
#7	$H_7 + B_2$	#25	$H_5 + B_8$	#43	$H_{18} + B_{11}$
#8	$H_8 + B_2$	#26	$H_{23} + B_9$	#44	$H_{19} + B_{11}$
#9	$H_9 + B_3$	#27	$H_2 + B_9$	#45	$H_{30} + B_1$
#10	$H_{10} + B_3$	#28	$H_{24} + B_9$	#46	$H_{31} + B_1$
#11	$H_{11} + B_3$	#29	$H_{25} + B_9$	#47	$H_{10} + B_1$
#12	$H_7 + B_3$	#30	$H_5 + B_9$	#48	$H_{17} + B_1$
#13	$H_{12} + B_4$	#31	$H_{26} + B_{10}$	#49	$H_4 + B_{11}$
#14	$H_{13} + B_4$	#32	$H_{27} + B_{10}$	#50	$H_{17} + B_{11}$
#15	$H_{14} + B_5$	#33	$H_{28} + B_{10}$	#51	$H_{10} + B_{11}$
#16	$H_{15} + B_6$	#34	$H_7 + B_{10}$	#52	$H_{16} + B_{11}$
#17	$H_{16} + B_6$	#35	$H_{22} + B_{10}$	#53	$H_{29} + B_1$
#18	$H_{16} + B_7$	#36	$H_{29} + B_{10}$	#54	$H_{16} + B_1$

Tableau 4.10 : Ensemble des « modèles d'apparence » des 54 individus.

La dernière opération dans la phase d'apprentissage consiste à vérifier s'il y a des « doublons », c'est-à-dire des combinaisons identiques. Si c'est le cas, cela signifie que deux (ou plusieurs) individus portent la même combinaison (c'est-à-dire le même vêtement du haut et le même vêtement du bas). Ici, dans le **Tableau 4.10**, on constate qu'on a effectivement le cas de deux individus portant la même combinaison, et qui sont les individus #34 et #39. Ces derniers portent effectivement le même jean (représenté par la classe H_7), un pull rouge pour l'un et une chemise de la même couleur pour l'autre (représentés par la classe B_{10}). La **Figure 4.9** illustre cet état de fait. Dans ce cas, ces deux individus seront « regroupés », c'est-à-dire que le système estime ici qu'il s'agit du même « individu », puisqu'ils ont tous les deux le même « modèle d'apparence » défini par la combinaison (H_7, B_{10}).

Individu #34 = (H_7, B_{10}) Individu #39 = (H_7, B_{10})

Figure 4.9: Illustration de la combinaison identique détectée parmi les 54 individus.

Ainsi s'achève la phase d'apprentissage hors ligne. Nous disposons ainsi de notre base d'apprentissage, constituée d'un ensemble de 31 classes finales *haut*, d'un ensemble de 11 classes finales *bas* et d'un ensemble de 53 « modèles d'apparence » (car il y avait un « doublon » parmi les 54 individus appris).

4.6 Phase de reconnaissance

Afin d'effecteur la reconnaissance des individus, nous utilisons les 54 séquences de test (contenant 400 images chacune) sur lesquelles apparaissent les individus appris dans un ordre aléatoire (un individu par séquence). La **Figure 4.10** illustre le processus de reconnaissance appliqué sur chaque séquence.

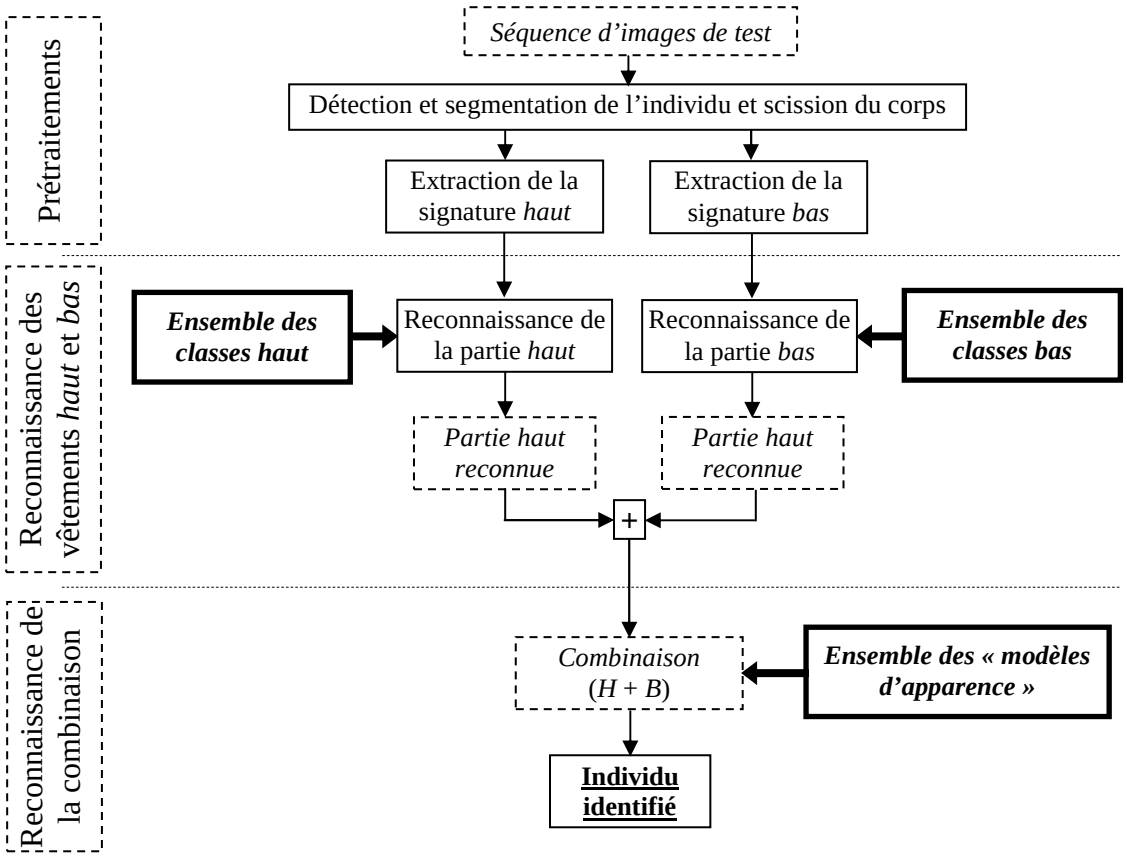


Figure 4.10 : Processus de reconnaissance de chaque individu par sa combinaison spécifique.

Sur chacune d'entre elle, le système commence par appliquer les procédures de prétraitement précédemment décrites (dans le chapitre 2) afin de détecter l'individu présent dans la scène et d'en extraire les signatures de classification *haut* et *bas*. Par la suite, tous les vecteurs de caractéristiques de la signature *haut* extraite sont testés par toutes les fonctions contour des classes *haut* f_{H_i}, $1 \le i \le 31$. L'affectation d'un vecteur x à la classe H_i est effective si $f_{H_i}(x) \ge 0$. Dans une séquence de classification, chacun des 400 vecteurs de caractéristiques *haut* est testé par chacune des 31 fonctions de décision classes *haut*. Pour une classe donnée H_i, si le pourcentage des vecteurs de caractéristiques classés comme appartenant à

cette classe est supérieur au seuil S_{haut} (fixé à 80%), alors le vêtement *haut* de cet individu est reconnu comme étant le vêtement appris représenté par la classe H_i. Le même processus, au seuil près qui vaut alors $S_{bas} = 90\%$ est appliqué à la partie *bas* : le vêtement *bas* d'un individu donné est reconnu comme étant le vêtement appris représenté par la classe B_j si le pourcentage de vecteurs de caractéristiques qui lui ont été affectés le long de la séquence est supérieur à S_{bas}. Au bout de cette phase, le vêtement du haut *H* et le vêtement du bas *B* de l'individu observé sont tous les deux reconnus et forment une combinaison $(H + B)$. La dernière phase de reconnaissance consiste alors à rechercher cette combinaison parmi celles contenues dans l'ensemble des « modèles d'apparence ». Ainsi, l'individu observé est identifié comme étant celui possédant le « modèle d'apparence » $(H + B)$ parmi ceux appris.

Le **Tableau 4.15** présente la matrice issue de la classification des parties *haut*, où chaque ligne représente un des 54 vêtements du haut à classer, et chaque colonne représente une des 31 classes *haut* apprises. La valeur du coefficient (i, j) de cette matrice (avec *i*=1...54 et *j*=1...31) représente la pourcentage de vecteurs de caractéristiques du vêtement du haut de l'individu #i classés comme appartenant à la classe *haut* H_j. Comme indiqué précédemment, S_{haut} étant fixé à 80%, le vêtement du haut d'un individu #i est reconnu comme étant le vêtements *haut* appris H_j si la valeur du coefficient (i, j) est supérieure à 80%. Le **Tableau 4.15** montre que tous les vêtements du haut ont été correctement classés, c'est-à-dire qu'on obtient un taux de bonne reconnaissance de 100%. En effet, non seulement tous les vêtements *haut* ont été reconnus, mais aucune confusion n'a été commise, c'est-à-dire qu'à chaque vêtement correspond à une et une seule classe *haut* apprise. Ceci démontre que la procédure de fusion effectuée lors de l'apprentissage des classes *haut* a parfaitement rempli son rôle, qui était de rassembler les classes similaires afin qu'il ne se produise pas de confusion lors de la reconnaissance.

En s'appuyant sur le même principe, le **Tableau 4.16** présente la matrice produite suite à la classification des parties *bas*, où chaque ligne représente un des 54 vêtements du bas à classer, et chaque colonne représente une des 11 classes *bas* apprises. Ici, S_{bas} étant fixé à 90%, le vêtement du bas d'un individu #i est reconnu comme étant le vêtement *bas* appris B_j (avec *i*=1...54 et *j*=1...11) si la valeur du coefficient (i, j) de cette matrice est supérieure à 90%. Le **Tableau 4.16** montre qu'ici aussi, non seulement tous les vêtements *bas* ont été reconnus, mais aucune confusion ne s'est produite, c'est-à-dire qu'à chaque vêtement correspond une et une seule classe *bas* apprise. Ce qui montre là encore que la procédure de fusion a parfaitement rempli son rôle sur les classes *bas*.

De cette façon, sur chaque séquence de test, les vêtements du haut et du bas de l'individu observé étant correctement reconnus, son « modèle d'apparence » défini par sa combinaison $(H + B)$ est donc reconstitué. Il suffit alors de rechercher ce « modèle d'apparence » parmi ceux appris. Cet individu est alors identifié comme étant l'individu appris représenté par ce dernier. Au final, nous avons pu constater que l'ensemble des 54 individus à reconnaitre a été parfaitement identifié. On rappelle que les individus #34 et #39 sont tous les deux identifiés par le « modèle d'apparence » défini par la combinaison $(H_7 + B_{10})$.

Tableau 4.15 : Matrice de classification *haut*.

Vêtements du haut à classer \ Classes apprises *haut*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	**95**	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	69	0	0
2	0	**94**	0	0	0	0	0	0	0	0	0	34	0	0	0	1	1	54	0	0	0	0	0	0	0	0	0	0	0	21	0
3	0	0	**94**	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	**93**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	2	0	0	**93**	0	0	6	0	0	61	14	0	0	0	0	7	1	0	0	17	0	1	0	0	0	0	0	0	0	0
6	0	0	0	0	0	**94**	0	59	0	0	0	0	16	28	56	1	18	0	68	0	0	4	3	0	0	28	0	0	0	0	0
7	0	0	0	0	0	0	**93**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	1	77	0	**93**	0	0	0	0	0	62	6	1	69	2	70	0	2	0	18	0	0	10	0	0	0	23	0
9	0	0	0	0	0	0	0	0	**97**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0
10	0	0	0	0	0	0	0	0	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
11	3	0	0	0	77	0	0	0	0	0	**93**	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
12	0	0	0	0	0	0	**92**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	71	0	0	4	0	0	0	0	0	5	**94**	0	0	0	0	0	43	0	0	0	0	0	0	0	0	0	0	1	16	0
14	0	0	0	0	0	18	0	0	0	0	0	0	**97**	0	10	46	0	0	0	0	0	78	0	0	0	0	0	0	0	3	0
15	0	29	0	0	0	52	0	56	0	0	0	1	0	**97**	1	0	69	70	0	0	0	0	0	0	0	0	0	0	0	67	0
16	0	0	0	0	0	73	0	0	0	0	0	0	28	0	**96**	73	0	0	13	0	0	52	0	0	0	0	0	0	0	1	0
17	0	0	0	0	0	30	0	0	0	0	0	0	47	0	72	**100**	0	0	7	0	0	76	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	55	0	0	0	0	0	0	62	0	41	**95**	0	0	0	0	0	38	0	0	0	0	0	0	0	12	0
19	0	0	0	0	0	51	0	69	0	0	0	0	0	78	0	0	**97**	19	0	0	0	0	19	0	0	0	0	0	0	66	0
20	0	69	0	0	0	0	0	34	0	0	0	59	0	64	0	0	68	**97**	0	0	0	0	0	0	0	0	0	0	0	50	0
21	0	0	0	0	0	79	0	75	0	0	0	0	0	0	45	15	0	0	**100**	0	0	0	0	0	0	20	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**95**	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	40	0	0	15	0	0	0	0	0	0	0	0	7	0	0	0	**94**	0	27	0	0	0	0	0	0	0	0
24	0	0	0	0	0	1	0	0	0	0	0	0	74	0	22	37	0	0	0	0	0	**97**	0	0	0	0	0	0	0	2	0
25	3	0	0	0	**94**	0	0	4	0	0	60	10	0	0	0	0	5	0	0	0	15	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	3	0	58	0	0	0	0	0	2	0	0	32	0	0	0	48	0	**96**	0	0	13	0	0	0	0	0
27	0	**97**	0	0	0	0	0	0	0	0	0	39	0	0	0	0	1	41	0	0	0	0	0	0	0	0	0	0	0	39	0
28	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**96**	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**96**	0	0	0	0	0	0
30	1	0	0	0	**95**	0	0	0	0	0	59	4	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	52	0	20	0	0	0	0	0	0	0	0	0	0	13	0	0	0	4	0	0	**96**	0	0	0	0	0
32	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**92**	1	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	**94**	0	0	1
34	0	0	1	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0	0	0	44	0	19	69	0	0	0	0	0	**98**	0	0	0	0	0	0	0	0	0
36	74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**96**	0	0
37	0	0	**97**	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	13	75	0	**95**	0	0	0	0	0	47	0	0	72	0	74	0	23	0	33	0	0	22	0	0	0	15	0
39	0	0	0	0	0	0	**98**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	**93**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	4	0	0	0	0	0	0	60	0	32	27	0	0	0	0	0	**98**	0	0	0	0	0	0	0	0	0
42	0	0	0	0	72	0	0	0	0	0	**95**	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	0	77	0	0	0	0	0	32	0	0	0	37	0	73	0	0	60	**97**	0	0	0	0	0	0	0	0	0	0	0	67	0
44	0	0	0	0	0	74	0	76	0	0	0	0	3	0	30	24	0	0	**98**	0	0	0	0	0	0	22	0	0	0	0	0
45	0	45	0	0	0	40	0	51	0	0	0	8	2	62	0	8	68	67	0	0	0	0	0	0	0	0	0	0	0	**95**	0
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	**94**
47	0	0	0	0	0	0	0	0	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	0	7	0	0	1	40	0	73	0	0	0	0	0	72	1	0	**94**	32	0	0	3	0	26	0	0	0	0	0	0	69	0
49	2	0	0	**94**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50	0	7	0	0	6	33	0	68	0	0	1	0	0	40	1	0	**92**	20	0	0	16	0	36	0	0	0	0	0	0	47	0
51	0	0	0	0	0	0	0	0	0	**96**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
52	0	0	0	0	0	10	0	0	0	0	0	0	27	12	63	**96**	0	0	0	0	0	41	0	0	0	0	0	0	0	0	0
53	78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**92**	0	0
54	0	0	0	0	0	12	0	0	0	0	0	0	33	9	69	**97**	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0

Vêtements du bas à classer	Classes apprises *bas*										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	**99**	0	0	5	0	2	0	2	0	0	36
2	**100**	0	0	3	0	0	0	1	0	0	20
3	**100**	0	0	2	0	1	0	0	0	0	22
4	**97**	0	0	15	0	0	0	1	0	0	33
5	**93**	0	0	20	0	1	0	1	0	0	36
6	0	**96**	0	0	0	0	0	0	0	0	0
7	0	**97**	0	0	0	0	0	0	0	0	0
8	0	**97**	0	0	0	0	0	0	0	0	0
9	0	0	**94**	0	0	0	0	0	0	0	0
10	0	0	**95**	0	0	0	0	0	0	0	0
11	0	0	**93**	0	0	0	0	0	0	0	0
12	0	0	**93**	0	0	0	0	0	0	0	0
13	13	7	1	**100**	15	0	0	13	0	36	14
14	7	8	0	**97**	11	0	0	8	0	31	2
15	0	0	0	22	**99**	0	0	0	0	46	0
16	5	0	0	0	0	**100**	0	0	0	0	0
17	6	0	0	0	0	**100**	0	0	0	0	0
18	0	0	0	20	0	0	**100**	0	0	0	0
19	0	0	0	25	0	0	**100**	0	0	0	0
20	0	0	0	24	0	0	**100**	0	0	0	0
21	0	0	0	8	0	0	**99**	0	0	0	0
22	1	0	0	11	0	0	0	**100**	0	0	71
23	1	0	0	12	0	0	0	**97**	0	0	84
24	1	0	0	16	0	0	0	**98**	0	1	86
25	0	0	0	8	0	0	0	**94**	0	2	72
26	0	0	0	15	0	0	0	0	**98**	0	0
27	0	0	0	21	0	0	0	0	**99**	0	0
28	0	0	0	10	0	0	0	0	**96**	0	0
29	0	1	0	24	0	0	0	0	**92**	0	0
30	0	0	0	9	0	0	0	0	**97**	0	0
31	0	0	0	18	31	0	0	0	0	**100**	0
32	0	0	0	21	38	0	0	0	0	**96**	0
33	0	0	0	13	34	0	0	0	0	**95**	0
34	0	0	0	19	29	0	0	0	0	**92**	0
35	0	0	0	23	39	0	0	0	0	**99**	0
36	0	0	0	12	40	0	0	0	0	**96**	0
37	0	0	0	19	38	0	0	0	0	**94**	0
38	0	0	0	22	41	0	0	0	0	**98**	0
39	0	0	0	21	29	0	0	0	0	**98**	0
40	27	0	0	9	0	0	0	81	0	0	**97**
41	19	0	0	16	0	0	0	85	0	0	**97**
42	18	0	0	9	0	0	0	83	0	0	**98**
43	34	0	0	12	0	0	0	69	0	0	**99**
44	26	0	0	16	0	0	0	74	0	0	**99**
45	**95**	0	0	0	0	23	0	0	0	0	3
46	**100**	0	0	1	0	22	0	0	0	0	12
47	**98**	0	0	0	0	13	0	0	0	0	8
48	**96**	0	0	0	0	14	0	0	0	0	5
49	28	0	0	4	0	0	0	79	0	0	**97**
50	22	0	0	8	0	0	0	80	0	0	**98**
51	42	0	0	12	0	0	0	75	0	0	**96**
52	36	0	0	3	0	0	0	71	0	0	**98**
53	**100**	0	0	1	0	2	0	0	0	0	47
54	**99**	0	0	0	0	0	0	1	0	0	51

Tableau 4.16 : Matrice de classification *bas*.

Arrivé à ce stade, nous devons comparer notre approche de reconnaissance avec celles utilisant un apprentissage hors ligne proposées dans la littérature, dont celles présentées dans le chapitre 3 [NakPHP03] [HahKK04] [GolKMS06]. Comme nous l'avons expliqué dans le chapitre 3, ces méthodes construisent une base d'apprentissage contenant les modèles d'apparence des individus appris. Le modèle de chaque individu correspondant à une description de l'apparence de son corps entier, c'est-à-dire de son apparence globale. Ainsi, afin de pouvoir comparer notre approche de reconnaissance par combinaison de vêtements avec l'approche de reconnaissance utilisant l'apparence globale, nous appliquons sur les parties *global* de nos 54 individus les mêmes procédures d'apprentissage hors ligne et de classification que celles effectuées sur les parties *haut* et *bas*. Les résultats obtenus sont présentés ci-après.

4.7 Comparaison avec la reconnaissance par l'apparence globale

La **Figure 4.11** illustre la procédure d'apprentissage hors ligne effectuée sur les parties *global*. Il s'agit de la même procédure que celle effectuée sur les parties *haut* et *bas*, hormis évidemment la phase de construction des combinaisons qui n'a ici plus lieu d'être. Dans ces conditions, le « modèle d'apparence » de chaque individu n'est plus représenté par sa combinaison de vêtements mais par son corps tout entier. Le système commence donc par extraire les 54 signatures d'apprentissage *global*, puis calcule la fonction de décision de chacune d'elles. On obtient de cette manière un ensemble de 54 classes initiales *global* sur lequel on applique la procédure de fusion.

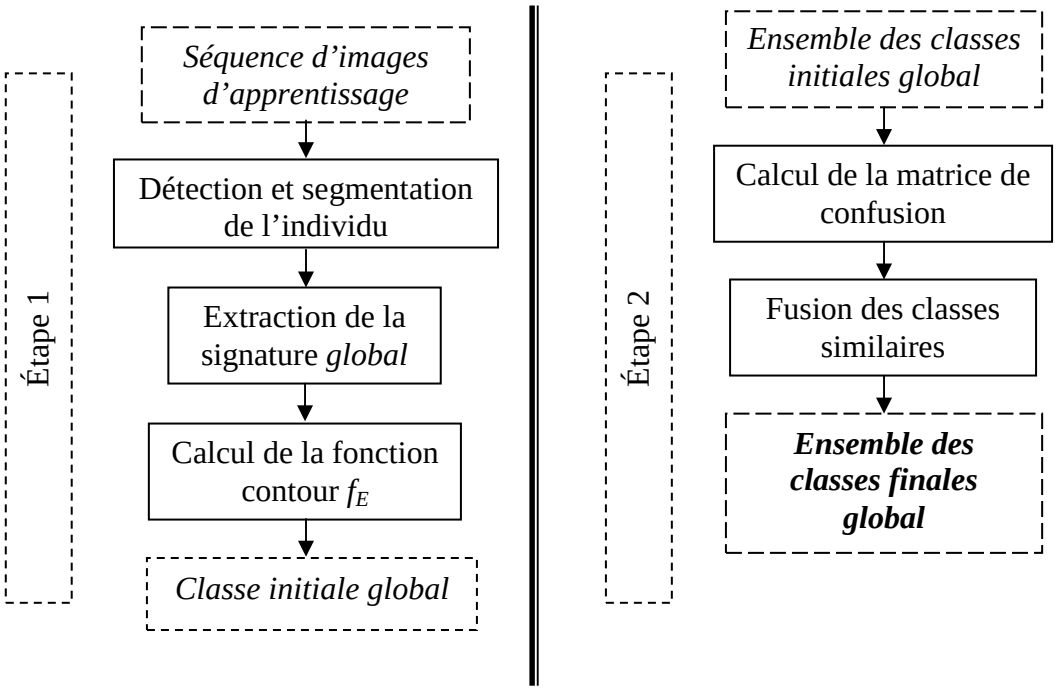


Figure 4.11 : Procédure d'apprentissage hors ligne des classes *global*.

Le **Tableau 4.17** présente la matrice de confusion M_G obtenue sur les classes initiales *global*. La fusion des classes similaires est effectuée ici en fixant le seuil $S_{global} = 90\%$, afin d'avoir un niveau suffisant de discrimination. Les fusions effectuées entre les classes initiales et les classes finales obtenues, qui sont au nombre de 48, sont présentées dans le **Tableau 4.11**. En

fixant le seuil $S_{global} = 80\%$, le nombre des classes finales obtenues serait de 47 (voir le **Tableau 4.12**). On garde pour la phase de reconnaissance S_{global} = 90%, construisant ainsi un ensemble de 48 classes finales *global*, et qui constituent ici la base d'apprentissage.

	Classes finales *global*												
	G1	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**	**G11**	**G12**	**G13**
Individus	#1	#2	#3	#4 #49	#5	#6 #8	#7	#9	#10	#11	#12	#13	#14

	Classes finales *global*												
	G14	**G15**	**G16**	**G17**	**G18**	**G19**	**G20**	**G21**	**G22**	**G23**	**G24**	**G25**	**G26**
Individus	#15	#16 #17	#18	#19	#20	#21	#22	#23	#24 #41	#25	#26	#27	#28

	Classes finales *global*												
	G27	**G28**	**G29**	**G30**	**G31**	**G32**	**G33**	**G34**	**G35**	**G36**	**G37**	**G38**	**G39**
Individus	#29	#30	#31	#32	#33	#34 #39	#35	#36	#37	#38	#40	#42	#43

	Classes finales *global*								
	G40	**G41**	**G42**	**G43**	**G44**	**G45**	**G46**	**G47**	**G48**
Individus	#44	#45	#46	#47	#48	#50	#51	#52 #54	#53

Tableau 4.11 : Classes finales *global*. S_{global} = 90%.

Classes intiales *global* (colonnes) / **Classes intiales *global*** (lignes)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
1	**95**	0	0	26	14	18	0	47	0	0	0	0	69	19	16	0	0	2	14	14	2	0	17	0	46	0	32	1	0	36	0	0	0	0	0	26	0	1	0	0	0	34	1	0	0	0	2	0	12	10	0	0	42	0
2	0	**97**	0	0	31	0	0	0	0	0	0	0	2	67	0	0	0	18	4	1	48	0	0	68	0	0	0	0	0	0	0	0	0	0	71	0	0	0	0	0	63	0	71	16	65	0	3	8	0	20	0	6	0	0
3	0	0	**98**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	17	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	25	0	0	**94**	0	1	0	5	0	0	8	0	5	0	0	0	0	0	0	0	0	0	1	0	9	10	9	46	0	35	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	**91**	0	0	0	0	0
5	28	14	0	0	**97**	61	0	27	0	0	0	0	36	56	53	0	0	0	41	14	47	2	54	4	30	0	3	0	0	0	4	0	0	0	8	0	0	73	0	0	2	49	22	2	28	0	0	66	0	78	0	0	20	0
6	14	0	0	0	71	**96**	0	79	0	0	0	0	19	24	72	0	0	0	46	11	8	0	85	0	76	0	30	0	0	5	4	0	0	0	0	9	0	75	0	0	0	70	2	0	2	0	0	31	0	55	0	0	18	0
7	0	0	0	0	0	0	**95**	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	6	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	44	0	0	3	33	**92**	0	**96**	0	0	0	0	35	23	45	0	0	0	19	0	0	0	61	0	73	10	57	0	0	19	0	0	0	0	0	28	0	32	0	0	0	68	0	0	0	0	0	0	0	10	0	0	44	0
9	0	0	0	0	0	0	0	0	**93**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	**94**	1	0	0	0	0	0	0	0	0	0	0	67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	11	0	0	0	0	0	7	**94**	0	0	0	0	0	0	0	0	0	0	73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	0	0	0	0	0
12	0	0	0	0	0	0	1	0	0	0	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	66	3	0	0	13	4	0	22	0	0	0	0	**94**	66	20	0	0	25	37	50	13	0	0	1	54	0	13	0	0	2	0	0	0	0	1	22	0	3	0	0	1	41	10	0	1	0	26	0	0	16	10	0	63	0
14	18	54	0	0	34	36	0	12	0	0	0	0	54	**93**	26	1	0	30	21	33	41	0	2	36	23	0	7	0	0	0	0	0	0	0	29	4	0	0	0	0	25	27	38	9	36	0	32	8	0	32	1	4	18	0
15	21	0	0	0	73	75	0	52	0	0	0	0	15	31	**100**	0	0	0	61	16	7	0	78	0	64	0	14	0	0	0	0	0	0	0	0	1	0	59	0	0	0	73	1	0	0	0	0	25	0	56	0	0	51	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**98**	74	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	13	0	0	44	8	0	0	0	0	0	0	67	0	67
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**94**	**98**	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	16	0	0	0	0	0	0	0	62	0	75
18	0	25	0	0	13	0	0	0	0	0	0	0	31	41	0	0	0	**99**	58	54	57	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	5	66	0	0	0	49	0	0	28	21	0	22	0
19	31	3	0	0	47	57	0	19	0	0	0	0	28	32	77	0	0	71	**100**	75	34	0	28	0	54	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	68	21	0	2	0	38	1	0	66	0	0	53	0
20	28	2	0	0	11	12	0	1	0	0	0	0	38	38	20	0	0	65	71	**96**	9	0	0	0	48	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	64	37	0	0	0	47	0	0	24	11	0	62	0
21	0	39	0	0	65	7	0	0	0	0	0	0	18	50	6	0	0	46	22	9	**100**	0	3	35	1	0	0	0	0	0	0	0	0	0	41	0	0	4	0	0	16	4	63	3	71	0	10	56	0	61	0	0	2	0
22	0	0	0	0	0	0	0	0	0	63	63	0	0	0	0	0	0	0	0	0	0	**94**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	32	0	0	1	57	86	0	44	0	0	0	0	0	11	50	0	0	0	2	0	6	0	**97**	0	48	17	20	0	0	1	23	0	0	0	0	0	0	61	0	0	0	20	0	0	1	0	0	31	0	43	0	0	4	0
24	0	63	0	0	3	0	0	0	0	0	0	0	0	43	0	17	3	4	1	0	16	0	0	**96**	0	0	0	0	0	0	0	0	0	0	69	0	0	0	0	0	87	0	28	51	44	0	1	0	0	5	0	59	0	11
25	53	0	0	5	49	74	0	62	0	0	0	0	49	26	34	0	0	3	28	28	3	0	48	0	**95**	0	44	1	0	19	0	0	0	0	0	53	0	7	0	0	0	56	2	0	0	0	1	1	1	17	0	0	62	0
26	0	0	0	22	0	7	0	12	0	0	0	0	0	0	1	0	0	0	0	0	0	0	14	0	14	**97**	2	1	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
27	35	0	0	10	3	34	0	63	0	0	0	0	20	8	12	0	0	0	2	1	0	0	17	0	60	0	**99**	0	0	27	0	0	0	0	0	60	0	0	0	0	0	32	0	0	0	0	0	0	3	1	0	0	44	0
28	13	0	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	**97**	0	63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**94**	0	0	9	32	0	0	0	0	0	0	0	0	0	0	0	0	43	0	0	0	0	0	0	0	0
30	19	0	0	47	0	5	0	22	0	0	0	0	15	0	0	0	0	0	0	0	0	0	1	0	27	22	18	48	0	**97**	0	0	0	0	0	25	0	0	0	0	0	1	0	0	0	0	0	0	35	0	0	0	1	0
31	0	0	0	0	18	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	18	0	0	1	0	0	0	0	**98**	0	0	0	0	0	0	30	0	0	0	0	0	0	1	0	0	16	0	1	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**93**	76	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	0	77	**94**	0	0	0	0	0	0	0	0	0	0	0	0	65	0	0	0	0	0	0	0	0
34	0	0	7	0	0	0	36	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**95**	0	0	42	0	**91**	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	0	73	0	0	19	0	0	0	0	0	0	0	0	21	0	9	1	0	0	0	48	0	0	77	0	0	0	0	0	0	0	0	0	0	**99**	0	0	0	0	0	72	0	52	56	76	0	0	16	0	23	0	12	0	2
36	28	0	0	3	1	4	0	27	0	0	0	0	54	4	1	0	0	0	2	3	0	0	0	0	59	0	40	0	0	31	0	0	0	0	0	**96**	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	39	0
37	0	0	18	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	63	0	0	**98**	0	35	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	11	0	0	0	78	78	0	33	0	0	0	0	9	0	50	0	0	0	1	0	5	0	71	0	18	1	2	0	0	0	37	0	0	0	0	0	0	**98**	0	0	0	10	0	0	8	0	0	66	0	78	0	0	1	0
39	0	0	5	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**96**	0	0	36	0	**99**	73	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	59	0	0	12	0	24	**96**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	0	61	0	0	0	0	0	0	0	0	0	0	0	10	0	22	8	0	0	0	13	0	0	**93**	0	0	0	0	0	0	0	0	0	0	79	0	0	0	0	0	100	0	26	74	60	0	0	0	0	0	0	56	0	16
42	23	1	0	0	30	68	0	66	0	0	0	0	30	24	62	0	0	11	49	57	5	0	57	0	72	0	50	0	0	0	0	0	0	0	0	28	0	2	0	0	0	**97**	4	0	1	0	2	1	0	24	0	0	73	0
43	0	70	0	0	39	0	0	0	0	0	0	0	9	51	0	0	0	51	17	6	73	0	0	53	0	0	0	0	0	0	0	0	0	0	47	0	0	0	0	0	27	0	**99**	2	60	0	14	13	0	68	0	1	0	0
44	0	12	0	0	1	0	0	0	0	0	0	0	0	10	0	55	18	0	0	0	1	0	0	69	0	0	0	0	0	0	0	0	0	0	60	0	0	0	0	0	66	0	1	**100**	53	0	0	1	0	1	0	48	0	24
45	0	72	0	0	19	2	0	0	0	0	0	0	0	43	2	6	0	0	1	0	50	0	2	61	0	0	0	0	0	0	1	0	0	0	61	0	0	8	0	0	42	0	57	43	**98**	0	0	51	0	49	0	5	0	0
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	14	20	0	0	0	0	0	0	0	0	0	0	0	0	**93**	0	0	0	0	0	0	0	0
47	1	7	0	0	0	0	0	0	0	0	0	0	25	45	0	0	0	63	21	52	2	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	9	0	0	0	**98**	0	0	0	63	0	23	0
48	1	9	0	0	77	25	0	11	0	0	0	0	1	45	21	0	0	0	0	0	38	0	38	2	1	0	0	0	0	0	34	0	0	0	11	0	0	72	0	0	0	0	8	1	59	0	0	**98**	0	78	0	0	0	0
49	16	0	0	86	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	21	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	**95**	0	0	0	0	0
50	11	25	0	0	77	49	0	28	0	0	0	0	18	40	49	0	0	5	65	6	48	0	55	7	19	0	1	0	0	0	0	0	0	0	11	0	0	72	0	0	3	34	46	2	34	0	0	69	0	**96**	0	0	13	0
51	0	0	0	0	0	0	0	0	0	0	0	0	23	19	0	0	0	31	1	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	76	0	0	0	**97**	0	28	0
52	0	6	0	0	0	0	0	0	0	0	0	0	0	2	0	75	51	0	0	0	0	0	0	73	0	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	68	0	1	57	3	0	0	0	0	0	0	**98**	0	87
53	58	0	0	0	12	14	0	30	0	0	0	0	58	25	16	0	0	23	25	43	2	0	2	0	63	0	32	0	0	0	0	0	0	0	0	49	0	0	0	0	0	73	2	0	0	0	20	0	0	7	0	0	**97**	0
54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	63	75	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	6	0	0	27	0	0	0	0	0	0	0	**93**	0	**99**

Tableau 4.17 : Matrice de confusion des classes initiales *global*.

	Classes finales *global*												
	G1	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**	**G11**	**G12**	**G13**
Individus	#1	#2	#3	#4 #49	#5	#6 #8 #23	#7	#9	#10	#11	#12	#13	#14

	Classes finales *global*												
	G14	**G15**	**G16**	**G17**	**G18**	**G19**	**G20**	**G21**	**G22**	**G23**	**G24**	**G25**	**G26**
Individus	#15	#16 #17	#18	#19	#20	#21	#22	#24 #41	#25	#26	#27	#28	#29

	Classes finales *global*												
	G27	**G28**	**G29**	**G30**	**G31**	**G32**	**G33**	**G34**	**G35**	**G36**	**G37**	**G38**	**G39**
Individus	#30	#31	#32	#33	#34 #39	#35	#36	#37	#38	#40	#42	#43	#44

	Classes finales *global*							
	G40	**G41**	**G42**	**G43**	**G44**	**G45**	**G46**	**G47**
Individus	#45	#46	#47	#48	#50	#51	#52 #54	#53

Tableau 4.12 : Classes finales *global*. S_{global} = 80%.

La procédure de reconnaissance des individus par leur apparence globale est illustrée par la **Figure 4.12**. Le **Tableau 4.17** illustre la matrice de classification des parties *global*, où chaque ligne représente un des 54 individus à classer, et chaque colonne représente une des 48 classes *global* apprises. La valeur du coefficient (i, j) de cette matrice (avec *i*=1...54 et *j*=1...48) représente la pourcentage de vecteurs de caractéristiques du corps entier de l'individu #i classés comme appartenant à la classe *global* G_j. S_{entier} étant fixé à 90%, un individu #i est associé à la classe G_j si la valeur de la cellule (i, j) est supérieure à 90%. On peut voir ici (**Tableau 4.18**) que toutes les signatures *global* ont non seulement ont été correctement reconnus, mais qu'aucune confusion ne s'est produite, ce qui démontre encore une fois que la procédure de fusion effectuée lors de l'apprentissage a parfaitement rempli son rôle.

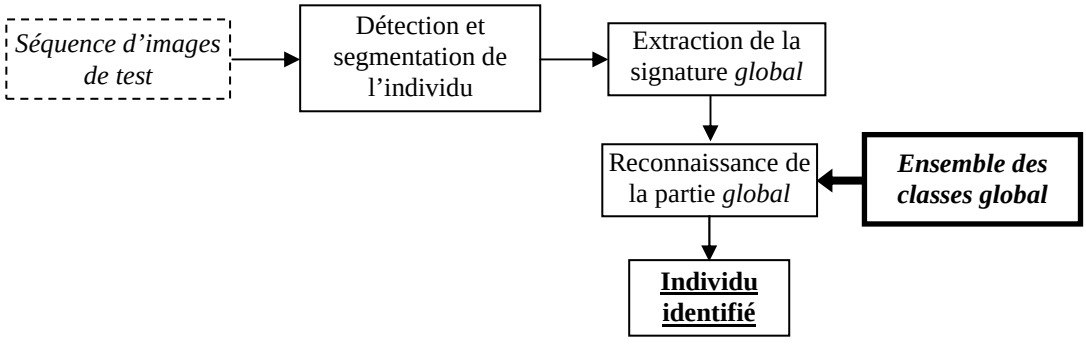


Figure 4.12 : Processus de reconnaissance de chaque individu par son apparence globale.

Tableau 4.18: Matrice de confusion des classes initiales *global*.

Apparences globalee à classer \ Classes apprises *global*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
1	**95**	0	0	25	12	28	0	0	0	0	0	65	20	16	0	2	16	9	1	0	23	0	50	0	43	1	0	42	0	0	0	0	0	42	0	1	0	31	0	0	0	0	1	0	6	0	0	40
2	0	**96**	0	0	23	0	0	0	0	0	0	1	67	0	1	9	3	0	56	0	0	75	0	0	0	0	0	0	0	0	0	0	72	0	0	0	0	0	78	18	75	0	2	3	31	0	4	0
3	0	0	**97**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	30	0	0	**94**	0	2	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	7	10	7	43	0	40	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	30	9	0	0	**94**	69	0	0	0	0	0	40	64	56	0	0	42	10	38	0	61	4	36	0	2	0	0	0	2	0	0	0	7	0	0	73	0	51	14	1	23	0	0	69	80	0	0	12
6	27	0	0	0	74	**94**	0	0	0	0	0	25	27	80	0	0	43	10	6	0	88	0	74	0	29	0	0	2	2	0	0	0	0	6	0	64	0	72	1	0	1	0	0	34	47	0	0	25
7	0	0	0	0	0	0	**93**	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0
8	42	0	0	2	25	**93**	0	0	0	0	0	39	19	36	0	0	15	0	0	0	66	0	70	15	67	0	0	14	0	0	0	0	0	27	0	28	0	63	0	0	0	0	0	0	6	0	0	56
9	0	0	0	0	0	0	0	**93**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	**91**	1	0	0	0	0	0	0	0	0	0	62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	35	0	0	0	0	8	**93**	0	0	0	0	0	0	0	0	0	78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	**92**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	71	2	0	0	9	27	0	0	1	0	0	**95**	58	15	0	22	38	50	8	0	0	1	60	0	11	0	0	0	0	0	0	0	1	40	0	0	0	37	6	0	1	0	22	0	8	4	0	63
14	19	51	0	0	38	28	0	0	0	0	0	40	**92**	19	0	37	30	34	50	0	0	35	22	0	6	0	0	0	0	0	0	0	31	6	0	0	0	31	44	4	35	0	26	0	45	0	3	26
15	35	0	0	0	70	83	0	0	0	0	0	20	32	**98**	0	0	58	13	4	0	86	0	71	0	18	0	0	0	0	0	0	0	0	1	0	60	0	81	1	0	1	0	0	28	58	0	0	63
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**95**	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	44	3	0	0	0	0	0	59	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**95**	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	69	0
18	0	16	0	0	6	0	0	0	0	0	0	40	43	0	0	**95**	58	56	46	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	66	0	1	0	63	0	15	16	0	21
19	24	2	0	0	51	60	0	0	0	0	0	25	40	67	0	67	**96**	84	37	0	25	0	51	0	1	0	0	0	0	0	0	0	0	0	0	0	0	71	13	0	1	0	35	1	76	0	0	71
20	21	1	0	0	20	13	0	0	0	0	0	38	38	15	0	66	78	**94**	6	0	0	0	58	0	0	0	0	0	0	0	0	0	0	1	0	0	0	67	35	0	0	0	59	0	10	4	0	87
21	0	40	0	0	71	2	0	0	0	0	0	15	48	3	0	36	19	4	**96**	0	1	41	0	0	0	0	0	0	0	0	0	0	38	0	0	2	0	2	82	2	72	0	6	47	70	0	0	1
22	0	0	0	0	0	0	0	0	59	60	0	0	0	0	0	0	0	0	0	**94**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	43	0	0	1	55	85	0	0	0	0	0	38	12	54	0	0	1	0	4	0	**94**	0	55	19	15	0	0	2	20	0	0	0	0	0	0	63	0	13	0	0	1	0	0	26	39	0	0	2
24	0	51	0	0	3	0	0	0	0	0	0	0	40	0	12	2	1	0	22	0	0	**96**	0	0	0	0	0	0	0	0	0	0	68	0	0	0	0	0	21	56	40	0	1	0	3	0	46	0
25	53	0	0	4	34	78	0	0	0	0	0	52	33	33	0	2	24	27	2	0	44	0	**93**	0	54	0	0	17	0	0	0	0	0	49	0	3	0	62	1	0	0	0	0	1	12	0	0	74
26	0	0	0	20	0	13	0	0	0	0	0	0	0	1	0	0	0	0	0	0	18	0	8	**95**	1	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	38	0	0	6	2	70	0	0	0	0	0	20	6	8	0	0	1	0	0	0	21	0	68	0	**95**	0	0	36	0	0	0	0	0	72	0	0	0	25	0	0	0	0	0	0	0	0	0	43
28	15	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	**95**	0	54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**93**	0	0	7	29	0	0	0	0	0	0	0	0	0	0	53	0	0	0	0	0	0
30	26	0	0	53	0	27	0	0	0	0	0	20	0	0	0	0	0	0	0	0	5	0	24	21	22	45	0	**96**	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	1
31	0	0	0	0	11	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	18	0	0	1	0	0	0	0	**95**	0	0	0	0	0	0	28	0	0	0	0	0	0	0	10	1	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**91**	83	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	65	0	0	80	**93**	0	0	0	0	0	0	0	0	0	69	0	0	0	0	0	0
34	0	0	6	0	0	0	38	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**95**	0	0	38	0	37	0	0	0	0	0	0	0	0	0	0	0
35	0	75	0	0	13	0	0	0	0	0	0	0	18	0	5	0	0	0	53	0	0	82	0	0	0	0	0	0	0	0	0	0	**97**	0	0	0	0	0	56	48	75	0	0	10	16	0	10	0
36	30	0	0	2	1	25	0	0	0	0	0	53	4	1	0	0	1	2	0	0	0	0	63	0	41	0	0	27	0	0	0	0	0	**93**	0	0	0	13	0	0	0	0	0	0	0	0	0	36
37	0	0	13	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	65	0	0	**97**	0	0	0	0	0	0	0	0	0	0	0	0	0
38	8	0	0	0	74	75	0	0	0	0	0	12	6	63	0	0	1	0	3	0	81	0	19	1	1	0	0	0	41	0	0	0	0	0	0	**94**	0	6	0	0	5	0	0	59	72	0	0	0
39	0	0	3	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**98**	0	0	33	0	69	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	63	0	0	7	0	**93**	0	0	0	0	0	0	0	0	0	0	0
41	0	60	0	0	0	0	0	0	0	0	0	0	13	0	11	0	0	0	7	0	0	**97**	0	0	0	0	0	0	0	0	0	0	72	0	0	0	0	0	24	71	56	0	0	0	1	0	55	0
42	25	1	0	0	39	72	0	0	0	0	0	29	16	57	0	6	47	52	3	0	55	0	71	0	44	0	0	0	0	0	0	0	0	23	0	1	0	**95**	2	0	1	0	2	1	17	0	0	80
43	0	73	0	0	41	0	0	0	0	0	0	6	50	1	0	49	11	3	75	0	0	58	0	0	0	0	0	0	0	0	0	0	47	0	0	0	0	1	**98**	1	69	0	7	6	56	0	0	1
44	0	8	0	0	0	0	0	0	0	0	0	0	5	0	49	0	0	0	1	0	0	72	0	0	0	0	0	0	0	0	0	0	51	0	0	0	0	0	0	**96**	51	0	0	1	0	0	30	0
45	0	58	0	0	41	0	0	0	0	0	0	0	48	1	1	0	1	0	51	0	1	71	0	0	0	0	0	0	1	0	0	0	72	0	0	13	0	0	56	40	**95**	0	0	50	42	0	2	0
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	15	19	0	0	0	0	0	0	0	0	0	0	**92**	0	0	0	0	0	0
47	0	4	0	0	0	0	0	0	0	0	0	37	40	0	0	77	21	42	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	11	0	0	0	**95**	0	0	52	0	17
48	0	5	0	0	80	7	0	0	0	0	0	0	48	12	0	0	0	0	26	0	41	0	1	0	0	0	0	0	42	0	0	0	7	0	0	63	0	0	8	1	58	0	0	**96**	77	0	0	0
49	14	0	0	**92**	0	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	28	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50	8	18	0	0	82	39	0	0	0	0	0	17	48	48	0	4	68	3	61	0	59	13	22	0	1	0	0	0	0	0	0	0	13	0	0	79	0	24	44	1	30	0	0	56	**93**	0	0	9
51	0	0	0	0	0	0	0	0	0	0	0	29	13	0	0	28	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	67	0	0	**94**	0	18
52	0	4	0	0	0	0	0	0	0	0	0	0	1	0	75	0	0	0	0	0	0	81	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	54	2	0	0	0	0	0	0	**96**	0
53	65	0	0	0	12	33	0	0	0	0	0	51	30	5	0	25	30	45	1	0	8	0	67	0	30	0	0	0	0	0	0	0	0	45	0	0	0	60	1	0	0	0	13	0	6	0	0	**93**
54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	79	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	**98**	0

Nous venons de présenter deux approches de reconnaissance de personnes basées sur l'apparence, la première en constituant le « modèle d'apparence » de chaque individu par sa combinaison de vêtements, l'autre en constituant son « modèle d'apparence » par son apparence globale, tel que ça a été fait notamment dans [NakPHP03] [HahKK04] [GolKMS06]. Afin de confronter les performances de ces deux approches, nous n'allons pas faire ici une comparaison en termes de résultats de taux de « bonnes reconnaissance » et de « taux de mauvaises reconnaissance ». En effet, la procédure de fusion automatique des classes similaires que nous avons effectuée lors de la phase d'apprentissage nous permet d'éviter les confusions (et donc les « erreurs ») lors de la phase de reconnaissance, et donc d'obtenir un taux de reconnaissance de 100%. Ainsi, en effectuant cette procédure de fusion en amont, nous anticipons les confusions qui pourraient se produire lors de la phase de reconnaissance. Ainsi, dans l'approche que nous proposons, il ne s'agit plus de reconnaitre chaque individu comme étant une personne ayant une identité « univoque », mais de le reconnaitre par son « modèle d'apparence », qui pourrait correspondre éventuellement à d'autres « individus ».

Ainsi, grâce à la procédure de fusion, nous obtenons un taux de bonne reconnaissance de 100% sur les classes *global*, contre un taux de 88% obtenu sans fusion (voir **Tableau 4.5**). Evidemment, la fusion des classes a réduit le nombre de classes. En effet, en utilisant la fusion sur les classes *global*, on n'obtient plus que 48 « modèles d'apparence » caractérisant un ensemble de 54 « individus ». Et c'est exactement sur ce point qu'on peut se rendre compte de l'intérêt et de la pertinence de caractériser chaque individu par un modèle d'apparence constitué par sa combinaison de vêtements du haut et du bas, plutôt que par son apparence entière. En effet, la modélisation d'apparence par combinaison nous a permis de constituer 53 « modèles d'apparence », tout en obtenant un taux de reconnaissance de 100%.

Au final, en utilisant l'approche de reconnaissance par combinaison le système arrive à discriminer 53 différents « modèles d'apparence » sur un ensemble contenant 54 individus (avec un taux de reconnaissance de 100%). Avec l'approche de reconnaissance par l'apparence globale, et sur le même ensemble, le système arrive à discriminer uniquement 48 « modèles d'apparence ». Ce résultat est parfaitement prévisible. En effet, dans le second cas, il y a équivalence directe entre les classes finales et le modèle d'apparence. Ainsi, le processus de fusion conduit à regrouper des classes initiales et donc, nécessairement, à diminuer le nombre de modèles d'apparence disponibles. A l'inverse, dans le premier cas traité, les modèles d'apparence résultent de la combinaison entre classes finales de vêtements de type *haut* et *bas*. Par conséquent, même si l'opération de fusion réduit forcément le nombre de combinaisons disponibles (le nombre de « classes de base » ayant considérablement diminué), la combinatoire restant réalisable reste en pratique très largement suffisante pour représenter l'ensemble des modèles d'apparence. Par ailleurs, comme nous l'avons évoqué dans l'introduction de ce chapitre, cette approche présente l'attrait d'être plus proche d'une description « humaine », qui présenterait un individu en détaillant sa tenue vestimentaire en séparant, justement, les différentes parties de celle-ci.

Dans le scénario que nous venons de décrire, tous les individus présentés au système ont préalablement été appris. Aussi, à chacun de leur vêtement du haut ou du bas ainsi qu'à leur combinaison correspondait une classe existante dans la base d'apprentissage (classe *haut*, *bas* et combinaison $(H+B)$ respectivement). Cependant, dans une application de vidéosurveillance, l'ensemble des individus qui seront impliqués dans les scènes ne peut pas être fermé. En effet, il faut toujours admettre la survenue d'un « nouvel » individu, initialement absent de la base d'apprentissage. Dans cette situation, le système devra alors être

capable de détecter automatiquement cette nouveauté et de générer le « modèle d'apparence » correspondant, afin de l'ajouter à la base d'apprentissage : en d'autres termes, il s'agit d'effectuer un apprentissage « en ligne ». Nous présentons dans ce qui suit les procédures qui permettent à notre système d'effectuer une telle opération.

4.8 Reconnaissance d'individus avec apprentissage en ligne

4.8.1 Scénario « ensemble ouvert »

Ici, on commence par faire l'apprentissage hors ligne d'un certain nombre d'individus issus d'une base de connaissance initiale. Par rapport au cas « fermé » précédent, nous partons d'un nombre de classe *haut*, *bas* et de combinaisons plus réduit. Ainsi, lors de la phase de reconnaissance, l'individu observé peut être soit déjà appris, soit non. La procédure de reconnaissance avec apprentissage en ligne est présentée dans la **Figure 4.14**.
Dans le cadre de celle-ci, le système commence par détecter et segmenter l'individu observé puis par extraire la signature de sa partie *haut* et la signature de sa partie *bas*. Chacun des 400 vecteurs de caractéristiques x qui composent la signature *haut* est testé par chacune des fonctions de décision des classes *haut* $\left\{ f_{H_i} \right\}_{1 \le i \le N_{haut}}$ contenues dans la base d'apprentissage. La règle définissant l'affectation à une classe donnée reste celle évoquée dans le cas fermé. Cependant, il est alors possible qu'aucune des classes existantes ne contienne plus de 80% (S_{haut}) de ces vecteurs. La signature correspondante est alors considérée comme « nouvelle ». Dans ce cas, une nouvelle fonction de décision, englobant l'ensemble des vecteurs de cette signature, est calculée et ajoutée à la base des classes *haut*.

La **Figure 4.13** présente l'exemple de l'apprentissage en ligne d'une nouvelle classe *haut*. On voit dans un premier temps dans cet exemple (**Figure 4.13** (a)) qu'il existe quatre classes *haut* (H_1, H_2, H_3 et H_4). On voit également un ensemble de vecteurs de caractéristiques (*croix noires*) appartenant à la signature actuellement en cours de reconnaissance. On constate dans cet exemple que les vecteurs se situent à l'extérieur des quatre contours de décision existants. Dans ce cas, cette signature correspond à une « nouveauté ». Un fonction de décision est alors calculée pour cette nouvelle classe et ajoutée à la base de connaissance (**Figure 4.13** (b)). Evidemment, l'individu portant ce haut sera considéré comme « nouveau », que son pantalon soit déjà appris ou pas, puisque sa combinaison est de toute manière inédite. Ce principe reste vrai si on inverse les classes *haut* et *bas*. En effet, en parallèle à cette procédure de classification de la signature *haut*, le système effectue la classification de la signature *bas*, reprenant les mêmes procédures que celles appliquées à la partie *haut*. Une fois les partie *haut* et *bas* classifiées, le système devra classer la combinaison $(haut + bas)$. Ainsi, pour chaque individu observé, cinq cas sont possibles :

1 – la partie *haut* et la partie *bas* sont identifiées comme connues (existent dans la base d'apprentissage), ainsi que la combinaison (*haut* + *bas*). Dans ce cas, le « modèle d'apparence » de l'individu est bien identifié.
2 – la partie *haut* et la partie *bas* sont identifiées comme connues, mais la combinaison de ces deux vêtements n'a été observée sur aucun des individus appris (n'est pas contenue dans l'ensemble des « modèles d'apparence »). Aussi, cette combinaison est considérée comme « nouvelle » et est ajoutée à l'ensemble des « modèles d'apparence ». Evidemment, l'individu observé ici est détecté comme « nouveau ».

3 – la partie *haut* est identifiée comme connue mais la partie *bas* est détectée comme « nouvelle » (n'existe pas dans la base d'apprentissage). Elle y est alors ajoutée. La combinaison (*haut* + *bas*) est donc forcement « nouvelle ». Cette nouvelle combinaison est alors ajoutée à la base des « modèles d'apparence ». Le « modèle d'apparence » de l'individu observé est donc détecté ici comme « nouveau ».
4 – à l'inverse du cas précédent, ici la partie *bas* est bien identifiée comme connue mais la partie *haut* est détectée comme « nouvelle » (et donc ajoutée). La combinaison (*haut* + *bas*) est donc détectée comme « nouvelle » et ajoutée à l'ensemble des « modèles d'apparence », et le « modèle d'apparence » de l'individu observé est ici aussi détecté comme « nouveau ».
5 – le dernier cas est celui où les deux parties *haut* et *bas* sont détectées comme « nouvelles ». Elles sont alors ajoutées à la base d'apprentissage, tout comme leur combinaison qui est forcement « nouvelle », elle aussi. Naturellement, le « modèle d'apparence » de l'individu observé ici est également détecté comme « nouveau ».

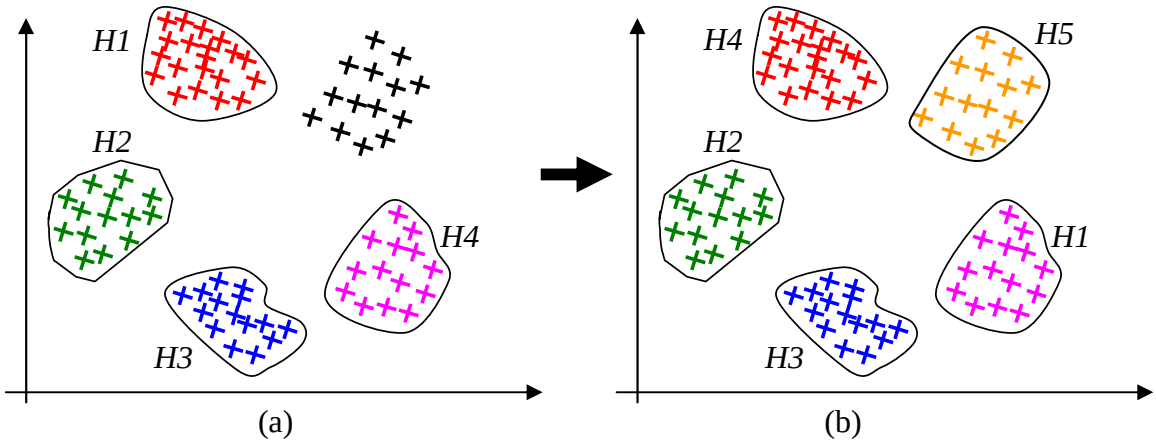


Figure 4.13 : Exemple illustrant le cas de l'apprentissage en ligne d'une nouvelle classe *haut*.

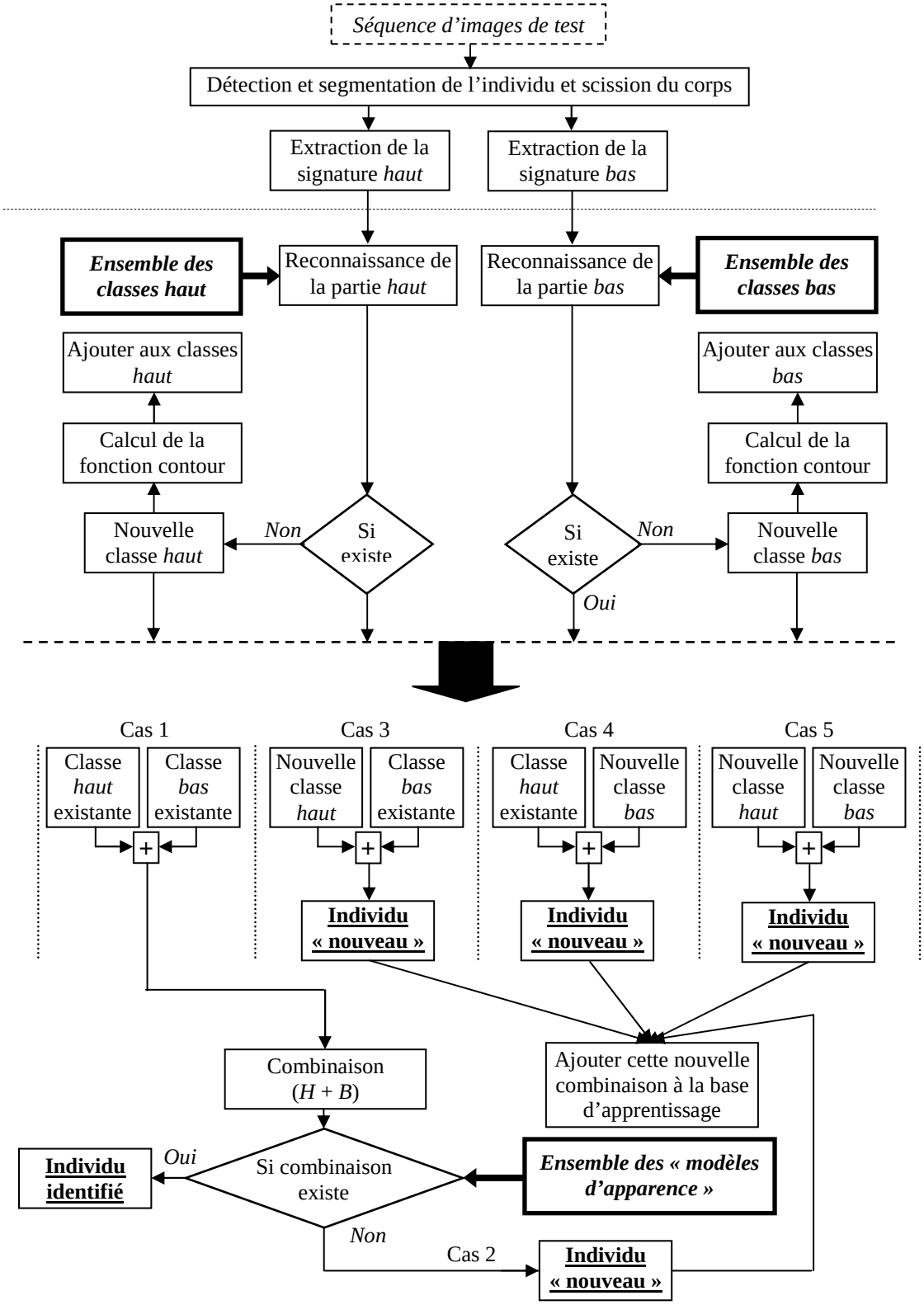


Figure 4.14 : Processus de reconnaissance avec apprentissage en ligne.

Afin d'évaluer notre algorithme avec reconnaissance de nouveautés et apprentissage en ligne sur notre base de données de 54 individus, nous effectuons une procédure de *leave-one-out cross validation*, c'est-à-dire que nous commençons par faire un apprentissage hors ligne de 53 individus, puis on lance la procédure de reconnaissance sur le 54ème. Ainsi, nous avons effectué cette procédure 54 fois (une fois pour chaque individu de notre base).

Les résultats obtenus sont que tous les individus à reconnaitre ont été, à juste titre, détectés comme « nouveaux », à l'exception de l'individu #34 qui a été reconnu comme ayant le même « modèle d'apparence » que l'individu #39 et de l'individu #39 qui a été reconnu comme ayant le même « modèle d'apparence » que l'individu #34.
Ainsi, tous les vêtements du haut ont été correctement été classifiés, c'est-à-dire que les haut des individus #2, #3, #4, #5, #7, #8, #10, #11, #12, #17, #18, #19, #20, #21, #24, #25, #27, #30, #34, #35, #36, #37, #38, #39, #40, #41, #42, #43, #44, #47, #48, #49, #50, #51, #52, #53 et #54 ont été identifiés comme connus, et ceux des individus #1, #6, #9, #13, #14, #15, #16, #22, #23, #26, #28, #29, #31, #32, #33, #45 et #46 ont été détectés comme nouveaux.
De la même manière, le vêtement du bas de l'individu observé a été à chaque fois correctement classé, c'est-à-dire que seul celui de l'individu #15 a été détecté comme nouveau, et ceux de tous les autres individus ont été identifiés comme connus.
Au final, les combinaisons de tous les individus ont été reconnues comme nouvelles, à l'exception naturellement de celles des individus #34 et #39 qui ont été identifiées comme connues. Ainsi, à la fin de la reconnaissance et donc de l'apprentissage en ligne, nous avons obtenus au final exactement la même base d'apprentissage que celle construite lors de l'apprentissage hors ligne dans le scénario « ensemble fermé » sur les 54 individus, c'est-à-dire que nous obtenons les mêmes classes *haut*, les mêmes classes *bas* et les mêmes combinaisons.

A partir de ce principe, et en utilisant la procédure d'apprentissage en ligne, on peut utiliser notre algorithme de reconnaissance dans un scénario « ensemble vierge » décrit ci-après.

4.8.2 Scénario « ensemble vierge »

Dans ce scénario, aucun apprentissage hors ligne n'est effectué. Le système commence directement par la phase de reconnaissance avec une ensemble d'apprentissage vide (c'est-à-dire l'ensemble des classes *haut*, l'ensemble des classes *bas*, et l'ensemble des combinaisons sont tous initialement vides). Les parties *haut* et *bas* ainsi que la combinaison du premier individu à reconnaitre sont automatiquement détectées comme « nouvelles » et ajoutées à leurs ensembles d'apprentissage respectifs. Par la suite, pour tous les individus suivants, le système suit la même procédure de reconnaissance que celle effectuée dans le scénario « ensemble fermé » (voir **Figure 4.14**).
Afin d'évaluer cette procédure dans le dernier scénario, nous avons présenté au système de reconnaissance les 54 individus un à un, dans l'ordre des numéros qu'ils portent, c'est-à-dire le l'individu #1 à l'individu #54. Ici, le système a reconnu, à raison, tous les individus comme étant « nouveaux », hormis l'individu #39 qui a été identifié comme ayant le même « modèle d'apparence » que l'individu #34.
Les individus dont les parties *haut* ont été détectées comme nouvelles sont #1, #2, #3, #4, #5, #6, #7, #8, #9, #10, #11, #13, #14, #15, #16, #17, #19, #20, #21, #22, #23, #24, #26, #28, #29, #31, #32, #33, #36, #45, et #46, et ceux dont les parties *haut* ont été détectées comme connues sont #12, #18, #25, #27, #30, #34, #35, #37, #38, #39, #40, #41, #42, #43, #44 #47, #48, #49, #50, #51, #52, #53 et #54.

De même, les individus dont les parties *bas* ont été détectées comme nouvelles sont #1, #6, #9, #13, #15, #16, #18, #22, #26, #31 et #40, et ceux dont les parties *bas* ont été détectées comme connues sont #2, #3, #4, #5, #7, #8, #10, #11, #12, #14, #17, #19, #20, #21, #23, #24, #25, #27, #28, #29, #30, #32, #33, #34, #35, #36, #37, #38, #39, #41, #42, #43, #44, #45, #46, #47, #48, #49, #50, #51, #52, #53 et #54.

Comme nous venons de le constater, la procédure d'apprentissage en ligne (dans les scénarios « ensemble ouvert » et « ensemble vierge ») s'est parfaitement déroulée. En effet, sur l'ensemble des classes *haut* et *bas*, toutes les classes appris ont été correctement identifiées, et toutes les classes non apprises ont été détectées comme « nouvelles ». Par ailleurs, toutes les combinaisons ont été correctement détectées comme « nouvelles », hormis, à juste titre, la combinaison des individus #34 et #39 qui a été identifiée comme connue puisque l'autre individu portant la même combinaison avait déjà été appris.

4.9 Conclusion

Nous avons présenté dans ce chapitre notre méthode de reconnaissance de personnes. Au lieu de se baser sur la description de l'apparence globale des personnes, tel que ça se fait dans les méthodes qu'on trouve dans la littérature [NakPHP03] [HahKK04] [GolKMS06], notre approche décrit de manière parallèle et séparée la partie du haut et la partie du bas de chaque individu, qui correspondent respectivement à ces vêtements du haut (pull, t-shirt, chemise, *etc.*) et du bas (pantalon). Le « modèle d'apparence » de chaque individu à reconnaitre est alors représenté par la combinaison que constituent les vêtements qu'il porte. Nous avons démontré que grâce à cette description par combinaison, on pouvait effectuer une plus grande discrimination parmi un groupe de personnes. En effet, dans notre base de données composée de 54 individus, on arrive à distinguer 53 « modèles d'apparence » avec la description par combinaison, alors qu'uniquement 48 ont pu l'être avec une description globale.

Afin de modéliser l'apparence de chaque vêtement, une « signature », composée d'un ensemble de vecteurs de caractéristiques, est extraite. Ainsi, nous avons testé et comparé différentes caractéristiques couleur et texture. Parmi ces dernières, nous avons pu constater que les valeurs moyennes RGB étaient tout à fait adéquates et efficaces pour la description d'apparence. Afin d'effectuer les phases d'apprentissage et de reconnaissance, nous avons mis au point une stratégie de classification (basée sur la technique one-class SVM) nous permettant d'effectuer notamment deux opérations. La première opération, la « fusion des classes », consiste à détecter en amont (c'est-à-dire lors de l'apprentissage) les classes de vêtement similaires et les rassembler en une seule classe. Cette opération est effectuée afin d'anticiper les confusions entre les classes et ainsi de les éviter lors de la phase reconnaissance. La deuxième opération est un « apprentissage en ligne ». Cette dernière permet au système, lors de la phase de reconnaissance, de reconnaitre que l'individu présent dans la scène correspond à une « nouveauté » (c'est-à-dire que son modèle d'apparence n'a pas été appris), puis de l'incorporer dans la base d'apprentissage. Grâce à ces processus de reconnaissance de nouveauté et d'apprentissage en ligne, le système peut fonctionner en scénario « ouvert », c'est-à-dire en faisant l'apprentissage d'un certain nombre d'individus tout en étant capable de reconnaitre un individu non appris et l'ajouter à la base d'apprentissage. En utilisant les mêmes principes, le système peut fonctionner en scénario « vierge » et construire progressivement, en partant d'une base vide, une base d'apprentissage contenant les modèles d'apparence de tous les individus vus par le système. Les expérimentations que nous avons menées ont montré une efficacité avérée de cette approche.

CONCLUSION GENERALE

Comme nous l'avons plusieurs fois évoqué, les travaux effectués dans cette thèse s'inscrivent dans le cadre du projet CAnADA (Comportements Anormaux : Analyse, Détection, Alerte). Ce projet a pour objectif l'analyse de scènes et la génération d'alertes basées sur l'analyse comportementale dans un lieu accueillant du public, et muni d'une installation multicaméra. Dans un tel cadre, nous avons pu constater qu'il y avait trois points clés pour l'analyse de scène et l'interprétation d'activités : la détection d'individus d'intérêt, le suivi de ces individus sur une fenêtre temporelle suffisamment large et l'analyse de leurs trajectoires ainsi que de leurs actions. Nous avons alors posé l'objectif de la détection comme étant de localiser les individus d'intérêt dans la scène en vue d'analyses ultérieures. Nous avons ainsi présenté dans le chapitre 2 l'approche que nous utilisons pour effectuer cette tâche, approche qui est basée sur une combinaison de soustraction du fond et de soustraction du contour du fond. Nous y avons également présenté une méthode de détection de la peau du visage dont le but était de nous permettre de nous assurer, avec suffisamment de confiance, que les blobs issus de la phase de soustraction de fond correspondaient bien à des personnes. Nous avons ainsi mis œuvre un algorithme de classification dynamique basé sur les SVM permettant d'améliorer la robustesse aux changements d'éclairage et ainsi d'améliorer le taux de bonne détection.

Nous avons également assimilé le suivi d'un individu cible à la détermination de sa position de manière continue et fiable tout au long du flux vidéo, ce qui permet de générer sa trajectoire. La représentation de la trajectoire permet alors de décrire les déplacements de cet individu, et peut être considérée comme une caractéristique déterminante qui aide à la discrimination entre des activités normales et anormales. Dans ce contexte, un des points importants consiste à mettre à jour les données de suivi des personnes et stocker l'historique de leurs déplacements afin d'effectuer par la suite une analyse d'activité. Etant donné que les personnes peuvent aléatoirement sortir ou entrer dans une pièce ou un quelconque espace surveillé, le processus de suivi instantané sera interrompu à chaque fois qu'un individu ciblé disparaitra du champ des caméras concernées. Pour arriver à reconstituer les déplacements et la trajectoire d'un individu sur une fenêtre temporelle suffisamment large, il faut alors pouvoir le caractériser et le reconnaitre à chaque fois qu'il apparaitra dans le champ de la caméra. Pour ce faire, la solution que nous proposons consiste à intégrer au processus de suivi un module de reconnaissance de personnes consistant à identifier un individu cible parmi un ensemble d'individus.

Comme présenté précédemment, les approches de reconnaissance de personnes qui paraissent être les plus utilisées aujourd'hui sont la reconnaissance de visages, la reconnaissance de la démarche (qui sont toutes deux des méthodes dites « biométriques ») et la reconnaissance basée sur l'apparence. Les caractéristiques biométriques sont basées sur un ensemble d'attributs connus pour être suffisamment spécifiques pour discriminer différentes personnes, amenant ainsi à un haut niveau de sélectivité. Cependant, les difficultés liées à ces méthodes (telles que les variations de la pose du visage, les expressions faciales, les occultations partielles, *etc.*, pour la reconnaissance du visage, et la vitesse de marche, la nature du sol, le type de chaussures portées, *etc.*, pour la reconnaissance de la démarche) et les fortes restrictions que ces caractéristiques imposent aux données (telle que la nécessité d'une résolution assez élevée des images pour la reconnaissance du visage, ou l'angle de prise de vue pour la démarche) limitent considérablement leur utilisation dans le cadre de la vidéosurveillance. D'un autre côté, les caractéristiques non-biométriques décrivant

l'apparence extérieure manquent du caractère univoque de leurs homologues biométriques et possèdent un temps de validité plus court (pas plus d'une journée). Néanmoins, elles font peser beaucoup moins de restrictions sur les données. Les méthodes basées sur l'apparence semblent alors les plus adaptées pour une intégration dans des systèmes de vidéosurveillance. Nous avons présenté dans le chapitre 3 les détails de différentes approches de reconnaissance de personnes et nous avons notamment classé les méthodes basées sur l'apparence en deux catégories : méthodes avec apprentissage hors ligne et avec apprentissage en ligne.

Les méthodes de la première catégorie opèrent en deux étapes. Durant une première étape qui constitue une phase d'apprentissage supervisé, un modèle d'apparence est créé pour chaque individu en utilisant des données « vérité de terrain » représentant les réponses correctes au problème de classification. La seconde étape correspond à la phase de reconnaissance proprement dite, qui consiste à essayer de retrouver l'identité d'un individu présenté en entrée parmi les individus que le système a appris. Nous avons décrit trois des méthodes les plus significatives dans cette catégorie [NakPHP03] [HahKK04] [GolKMS06]. Nous avons alors précisé que les méthodes décrites sont évaluées dans le contexte d'un scénario dit « d'ensemble fermé », où seuls des individus précédemment appris sont supposés être présentés au classifieur. Aussi, une limitation de ces méthodes est qu'elles ne sont pas capables de reconnaitre la « nouveauté », et ce par construction même. Or, dans des applications de vidéosurveillance avec un « scénario ouvert », c'est-à-dire où on ne connait pas à l'avance les personnes qui seront impliquées dans les scènes, la capacité de reconnaissance de l'occurrence d'un individu « nouveau » (cas non appris) est indispensable.

Pour ce qui touche aux méthodes de reconnaissance avec apprentissage en ligne, nous avons précisé qu'elles s'inséraient généralement dans un système plus global dédié au suivi de personnes, et où la reconnaissance n'est essentiellement effectuée que lorsqu'il y a intersection entre les individus. Aussi, aucun apprentissage préalable n'était effectué, mais un modèle constitué d'un seul vecteur de caractéristiques était extrait à la volée (pour chaque individu détecté) puis comparé aux modèles existants par le calcul d'une simple distance. Et si cette distance est supérieure à un seuil prédéfini, cet individu est considéré comme « nouveau ». Cependant, la faiblesse de cette catégorie de méthodes réside dans le fait que la base de connaissance est constituée d'un vecteur de caractéristiques par individu observé, ce qui la rend peu fiable à cause de la variabilité de l'apparence, du fait notamment à des variations de poses (de face, de côté, *etc.*). Aussi, il est nécessaire d'enregistrer et de classifier l'apparence d'un individu sur une période plus longue afin de saisir ces éventuelles variations. De plus, lorsqu'il s'agira de traiter et de reconnaitre un nombre assez important d'individus, un simple calcul de distance deviendra insuffisant pour en effectuer l'identification. Une méthode de classification plus élaborée sera alors nécessaire pour gérer la base de données contenant tous ces individus.

Une autre limitation importante des méthodes basées sur l'apparence que nous avons présentées est qu'elles effectuent la reconnaissance en modélisant l'apparence extérieure globale des personnes, c'est-à-dire que les blobs sur lesquels sont extraites les caractéristiques d'apparence (couleur, texture, forme) correspondent au corps entier. Ceci représente une limitation de ces méthodes. En effet, ces dernières ne tirent alors pas avantage des différences de couleur et de texture qui existent généralement entre les vêtements du haut et les vêtements du bas que les gens portent. A titre d'exemple, ces systèmes ne feraient pas la différence entre un individu portant un pull rouge et pantalon bleu et un individu portant un pull bleu et pantalon rouge.

Dans le cadre de cette thèse, nous avons alors développé une approche de reconnaissance de personnes basée sur l'apparence. L'approche que nous proposons se démarque des méthodes existantes en plusieurs points. En premier lieu, notre méthode se base sur la modélisation de manière séparée des vêtements du haut et du bas des personnes observées (au lieu de modéliser leur apparence entière). Par la suite, pour un individu donné, le système reconnait (classifie) séparément ses deux vêtements, puis la combinaison correspondante (*haut + bas*) représentera son « modèle d'apparence » par laquelle il sera identifié. Ceci s'inspire de la façon dont les gens en général, ou les policiers en particulier, donnent le signalement d'un individu lorsqu'ils doivent décrire son apparence afin de le retrouver parmi une multitude de personnes.

Par ailleurs, notre approche se positionne dans un premier temps parmi les méthodes de reconnaissance avec apprentissage hors ligne, et est alors développée en deux phases. Durant la première, qui consiste en un apprentissage hors ligne, les modèles d'apparence des individus à reconnaitre sont construits. Une base de connaissance est donc constituée à la fin de cette étape. La deuxième phase représente la reconnaissance des individus. Durant celle-ci, le système tente de reconnaitre chaque individu qui se présente dans la scène. Par la suite, nous avons incorporé une procédure de reconnaissance permettant au système de détecter lorsque le modèle de l'individu cible est « nouveau » (non appris au préalable), et de l'ajouter à la base d'apprentissage. Cette procédure d'ajout représente donc un apprentissage en ligne, au cours duquel de « nouveaux » individus sont intégrés. Nous pouvons ainsi considérer que notre méthode de reconnaissance se rapproche de celles utilisant un apprentissage hors ligne dans le sens où elle construit une base de connaissance a priori, mais s'en démarque toute fois par sa capacité à identifier l'occurrence de la « nouveauté » et à effectuer un apprentissage en ligne. A l'extrême, la base a priori peut être intégralement vide et l'apprentissage se faire entièrement en ligne.

Un autre point clé de notre approche de reconnaissance réside dans la procédure que nous avons appelée la « fusion des classes ». Cette procédure consiste, lors de la phase d'apprentissage, à rassembler en une seule classe plusieurs classes de vêtements qui sont « similaires ». La mesure de similarité dans ce cas étant définie par le degré « d'enchevêtrement » entre classes. Cette phase de fusion est appliquée afin de structurer la connaissance extraite de la phase d'apprentissage afin d'anticiper, et ainsi d'éviter, les confusions entre classes lors de la reconnaissance.

En définitive, au cours de cette thèse, nous avons eu l'occasion de présenter notre contribution au domaine de la classification d'individus observés dans des séquences vidéo. L'approche que nous avons proposée se veut une démarche efficace de résolution du problème de la distinction entre personnes sur la base d'un modèle d'apparence basé principalement sur la tenue vestimentaire. Les résultats que nous avons obtenus montrent que la méthode proposée donne des résultats tout à fait pertinents, en conformité avec les décisions prises par « l'expert humain » soumis au même problème. Qui plus est, le vecteur choisi est en lien étroit avec la « perception » que ce même expert peut avoir sur l'apparence des personnes observées. Cette proximité est suffisante pour pouvoir imaginer, à titre de perspective, qu'une signature puisse être construite non plus à partir d'une séquence d'images, mais sur la base d'un « signalement ». Il est alors envisageable d'utiliser cette dernière pour rechercher une correspondance individu / signalement dans des banques de données vidéos en appliquant les mêmes outils que ceux que nous avons développés. Au titre des aspects qui mériteraient, entre autre, d'être creusés plus en profondeur, la texture est très certainement un paramètre qui pourrait améliorer encore les performances obtenues. Comme nous l'avons évoqué dans ce

mémoire, les résultats *a priori* décevants générés avec ce type de caractéristique tiennent certainement autant (si ce n'est davantage) au manque de données « texturées » dans les bases d'apprentissage et de test (qu'il s'agisse des nôtres où de celles exploitées par la communauté) que des difficultés liées à une mise en œuvre efficace.

Enfin, nous espérons avoir pu communiquer au lecteur du présent mémoire la passion qui nous a animée tout au long de notre modeste incursion au sein de ce domaine en devenir que constitue la « vidéo analyse ».

RÉFÉRENCES BIBLIOGRAPHIQUES

[AbdE99] M. Abdel-Mottaleb, A. Elgammal, Face detection in complex environments from color lmages, Proceedings of the IEEE International Conference on Image Processing (ICIP), pages 22-626, Kobe, Japan, 1999

[Adelie] Adelie Surveillance Des Forets, Paratronic, http://www.paratronic.fr/feux/

[AhmL08] M. Ahmad, S.W. Lee, Human action recognition using shape and clg-motion flow from multi-view image sequences, Pattern Recognition, vol. 41 (7), pages 2237-2252, 2008

[AlbTD01] A. Albiol, L. Torres, E.J. Delp, Optimum color spaces for skin detection, Proceedings of the IEEE International Conference on Image Processing (ICIP), pages 122-124, Thessaloniki, Greece, 2001

[AlpOM06] Y. Alper, J. Omar and S. Mubarak, Object tracking: a survey, ACM Computing Surveys, vol. 38 (4), 2006

[AngC03] C. Angulo, A. Català, Online learning with kernels for smart adaptive systems: a review, European Network for Intelligent Technologies (ENIT03), Oulu, Finland, 2003

[AnjC07] N. Anjum, A. Cavallaro, Unsupervised fuzzy clustering for trajectory analysis, Proceedings of the IEEE International Conference on Image Processing (ICIP), vol. 3, pages 213-216, San Antonio, Texas, USA, 2007

[Auto07ID] IEEE Workshop on Automatic Identification Advanced Technologies (Auto ID), Alghero, Italy, 2007

[Avi01] S. Avidan, Support vector tracking, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 184-191, Kauai, HI, USA, 2001

[AVSS03] 1st IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), Miami, FL, USA, 21-22 Jul 2003

[AVSS05] 2nd IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), Como, Italy, 15-16 Sep 2005

[AVSS06] 3rd IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), Sydney, NSW, Australia, 22-24 Nov 2006

[AVSS07] 4th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), London, United Kingdom, 05-07 Sep 2007

[AVSS08] 5th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), Santa Fe, New Mexico, USA, 01-03 Sep 2008

[AVSS09] 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), Genova, Italy, 02-04 Sep 2009

[AVSS10] 7th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), Boston, MA, USA, 29 Aug-01 Sep 2010

[BarCK78] C. Barclay, J. Cutting, L. Kozlowski, Temporal and spatial factors in gait perception that influence gender recognition, Percept. Psychophys, vol. 23 (2) pages, 145-152, 1978

[BasXG09] K. Bashir, T. Xiang and S. Gong, Gait representation using flow fields, Proceedings of the British Machine Vision Conference (BMVC), London, UK, 2009

[BaySM09] Á. Bayona, J.C. SanMiguel, J.M. Martínez, Comparative evaluation of stationary foreground object detection algorithms based on background subtraction techniques, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 25-30, Genova, Italy, 2009

[BBCnews05] BBCnews online. CCTVvoyeurism story. 2005.
http://www.news.bbc.co.uk/1/hi/england/merseyside/4521342.stm

[BenJSR09] Y. Benezeth, P.M. Jodoin, V. Saligrama, C. Rosenberger, Abnormal events detection based on spatio-temporal co-occurences, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2458-2465, Miami, Florida, USA, 2009

[BerSR00] M. Bertalmio, G. Sapiro, G. Randall, Morphing active contours, IEEE Transactions on Pattern Analysis, vol. 22 (7), pages 733-737, Minneapolis, MN, USA, 2000

[BeyK99] D. Beymer, K. Konolige, Real-time tracking of multiple people using continuous detection, Proceedings of the IEEE International Conference on Computer Vision (ICCV) Frame-Rate Workshop, Kerkyra, Corfu, Greece, 1999

[BhaCRA07] M. Bhargava, C. Chen, M. Ryoo, and J. Aggarwal, Detection of abandoned objects in crowded environments, Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 271-276, London, United Kingdom, 2007

[BioSurv07] NSF Workshop on BioSurveillance Systems and Case Studies (BioSurveillance 2007), New Jersey, USA, 2007

[BlaGSJ09] C.R. del-Blanco, N. García, L. Salgado, F. Jaureguizar, Object tracking from unstabilized platforms by particle filtering with embedded camera ego motion, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 400-405, Genova, Italy, 2009

[BlaJ98] M. Black, A. Jepson, Eigentracking: Robust matching and tracking of articulated objects using a view-based representation, International Journal of Computer Vision, vol. 26 (1), pages 63-84, 1998

[BlaSBB96] V. Blanz, B. Schölkopf, H. Büthoff, C. Burges, V. Vapnik, and T. Vetter, Comparison of view-based object recognition algorithms using realistic 3D models, Artificial Neural Networks-ICANN, vol. 1112, pages 251 - 256, Berlin, Germany, 1996

[BouBPB09] P.L.M. Bouttefroy, A. Bouzerdoum, S.L. Phung, A. Beghdadi, Vehicle tracking using projective particle filter, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 7-12, Genova, Italy, 2009

[BouL08] K. Boukharouba and S. Lecoeuche, Online clustering of non-stationary data using incremental and decremental SVM, 18th International Conference of Artificial Neural Networks, pages 336-345, Prague, Czech Republic, 2008

[BouN07] I. Bouchrika and M. S. Nixon, Model-based feature extraction for gait analysis and recognition, Proceedings of Mirage: Computer Vision/Computer Graphics Collaboration Techniques and Applications, pages 150-160, INRIA Rocquencourt, France, 2007

[Bra98] G.R. Bradski, Computer vision face tracking for use in a perceptual user interface, Intel Technology Journal, 1998

[Bro88] D.S. Broomhead, D. Lowe, Multivariate functional interpolation and adaptive networks, Complex Systems, vol. 2, pages 321-355, 1988

[BroBM09] T. Brox, C. Bregler, J. Malik, Large displacement optical flow, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 41-48, Miami, Florida, USA, 2009

[BroC86] T. Broida, and R. Chellappa, Estimation of object motion parameters from noisy images, IEEE Transactions on Pattern Analysis Machine Intelligence, vol.8 (1), pages 90-99, 1986

[BroC06] G.J. Brostow, R. Cipolla, Unsupervised bayesian detection of independent motion in crowds, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 594-601, New York, NY, USA, 2006

[BroCL01] D. Brown, I. Craw, and J. Lewthwaite, A som based approach to skin detection with application in real time systems, Proceedings of the British Machine Vision Conference (BMVC), pages 491-500, Manchester, UK, 2001

[BSYM07] Biometrics Symposium (BSYM), Baltimore, Maryland, USA, 11-13 Sep 2007

[BSYM08] Biometrics Symposium (BSYM), Tampa, Florida, USA, 23-25 Sep 2008

[BugP09] A. Bugeau and P. Pérez, Detection and segmentation of moving objects in complex scenes, Computer Vision and Image Understanding, vol. 112, pages 459-476, 2009

[BurB08] W. Burger, M. Burge, Digital Image Processing, An algorithmic introduction using Java, Springer, 2008

[CamM99] D. Comaniciu and P. Meer, Mean shift analysis and application, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 1197-1203, Kerkyra, Corfu, Greece, 1999

[CapDDC03] M.B. Capellades, D. Doermann, D. Dementhon and R. Chellappa, An Appearance Based Approach for Human and Object Tracking, Proceedings of the IEEE

International Conference on Image Processing (ICIP), pages 85-88, Barcelona, Catalonia, Spain, 2003

[CauP00] G. Cauwenberghs and T. Poggio, Incremental and decremental support vector machine learning, Advances in Neural Information Processing Systems, M.I.T. Press, vol. 13, pages 409-415, Cambridge, MA, USA, 2000

[ChaN98] D. Chai and K.N. Ngan, Locating facial region of a head-and-shoulders color image, Proceedings 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pages 124-129, Nara, Japan, 1998

[CheLZ09] C.H. Chen, J.M. Liang, H. Zhao, et al., Factorial HMM and parallel HMM for gait recognition, IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews, vol. 39 (1), pages 114-123, 2009

[CheOO04] L. Chen, M. T. Ozsu, and V. Oria, Symbolic representation and retrieval of moving object trajectories, Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, pages 227-234, New York, NY, USA, 2004

[Chu92] C. K. Chui, Introduction to wavelets, Academic Press, Boston, MA, USA, 1992

[ClaM99] P.R. Clarkson, P.J. Moreno, On the use of support vector machines for phonetic classification, Proceedings of the IEEE International Conference on Speech and Signal Processing, vol. 2, pages 585-588, Phoenix, Arizona, USA, 1999

[Cliris] Cliris group, http://www.clirisgroup.com/?pagex=6

[ColLK00] R. Collins, A. Lipton, T. Kanada, et al., A system for video surveillance and monitoring: VSAM final report, Technical report CMU-RI-TR-00-12, Carnegie Mellon University, May 2000

[ComRM03] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, pages 564-577, 2003

[CorV95] C. Cortes, V. Vapnik, Support vector networks, Machine Learning 20, pages 273-297, 1995

[CraTB92] I. Craw, D. Tock, and A. Bennett, Finding face features, Proceedings of the 2nd European Conference Computer Vision (ECCV), Santa Margherita Ligure, Italy, pages 92-96, 1992

[CreKS002] D. Cremers, T. Kohlberger, and C. Schnorr, Non-linear shape statistics in mumford-shah based segmentation, Proceedings of the 7th European Conference on Computer Vision (ECCV), Copenhagen, Denmark, 2002

[CucGPP01] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, and S. Sirotti, Improving shadow suppression in moving object detection with HSV color information, Proceedings of the IEEE International Conference Intelligent Transportation Systems, pages 334-339, Oakland, CA, USA, 2001

[CutK77] J. Cutting, L. Kozlowski, Recognizing friends by their walk: gait perception without familiarity cues, Bulletin of the Psychonomic Society, Vol. 9, pages 353-356, 1977

[DaiN96] Y. Dai, Y. Nakano, Face-texture model based on SGLD and its application in face detection in a color scene, Pattern Recognition, vol. 29 (6), pages 1007-1017, 1996

[DavT02] J.W. Davis, S.R. Taylor, Analysis and recognition of walking movements, International Conference on Pattern Recognition (ICPR), pages 315-318, Quebec City, Canada, 2002

[DelGB08] K. Delac, M. Grgic, M.S. Bartlett, Recent advances in face recognition, Publisher: IN-TECH, 2008

[DimSG09] K. Dimitropoulos, T. Semertzidis and N. Grammalidis, Video and signal based surveillance for airport applications, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 170-175, Genova, Italy, 2009

[DioG03] J.J. de Dios, N. Garcia, Face detection based on a new color space YCgCr, Proceedings of the IEEE International Conference on Image Processing (ICIP), Barcelona, Spain, 2003

[DLPAJ07] Direction des libertés publiques et des affaires juridiques, Rapport 2007 relatif aux conditions dapplication de larticle 10 de la loi du 21 janvier 1995 modifiée dorientation et de programmation relative a la sécurité, 2007

[DocT01] S. Dockstader, and M. Tekalp, On the tracking of articulated and occluded video object motion, Real Time Image, vol. 7 (5), pages 415-432, 2001

[DeeV08] H.M. Dee, and S.A. Velastin, How close are we to solving the problem of automated visual surveillance ? A review of real-world surveillance, scientific progress and evaluative mechanisms, Machine Vision and Applications, vol. 19 (5-6), pages 329-343, 2008

[FehSLY09] D. Fehr, R. Sivalingam, O. Lotfallah and Y. Park, Counting people in groups, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 152-157, Genova, Italy, 2009

[FleFB96] M. Fleck, D. Forsyth, C. Bregler, Finding naked people, Proceedings of the European Conference on Computer Vision (ECCV), pages 593-602, Cambridge, UK, 1996

[ForF99] D. Forsyth and M. Fleck, Automatic Detection of Human Nudes, International Journal of Computer Vision, vol. 31 (1), pages 63-77, 1999

[FraM99] A. Francois, and G. Medioni, Adaptive color background modeling for real-time segmentation of video streams, Proceedings of the International Conference on Imaging Science Systems and Technology, pages 227-232, 1999

[Gaf07] D. Gafurov, A survey of biometric gait recognition: approaches, security and challenges, Annual Norwegian Computer Science Conference, Oslo, Norway, 2007

[Gir97] F. Girosi, An equivalence between sparse approximation and Support Vector Machines. A.I. Memo 1606, MIT Artificial Intelligence Laboratory, 1997

[GolKMS06] L. Goldmann, M. Karaman, J. Minquez, T. Sikora, Appearance-based person recognition for surveillance applications, Technical University of Berlin, Communication Systems Group, Berlin, Germany, 2006

[GomSS02] G. Gomez, M. Sanchez, L.E. Sucar, On selecting an appropriate colour space for skin detection, Springer-Verlag: Lecture Notes in Artificial Intelligence, vol. 2313, pages 70-79, 2002

[GorBSIB07] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29 (12), pages 2247-2253, 2007

[GreK95] L. Grewe, A. Kak, Interactive learning of a multi-attribute hash table classifier for fast object recognition, Computer Vision and Image Understanding, vol. 61 (3), pages 387-416, 1995

[GroSC01] R. Gross, J. Shi, J.F. Cohen, Quo vadis face recognition?, Proceedings of the 3rd Workshop on Empirical Evaluation Methods in Computer Vision, pages 119-132, Santa Barbara, CA, USA, 2001

[HahKK04] M. Hahnel, D. Klunder, K Kraiss, Color and texture features for person recognition, Proceedings of the IEEE International Joint Conference on Neural Networks, pages 25-29, Budapest, Hungary, 2004

[HamBBD09] A. Hampapur, R. Bobbitt, L. Brown, M. Desimone, R. Feris, Video analytics in urban environments, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 128-133, Genova, Italy, 2009

[HamBBL09] L. Hamoudi, K. Boukharouba, J. Boonaert and S. Lecoeuche, On-line face tracking under large lighting condition variations using incremental learning, The International Conference on Computer Vision Theory and Applications (VISSAPP), pages 636-643, Lisboa, Portugal, 2009

[HamBL10] L. Hamoudi, J. Boonaert, and S. Lecoeuche, People Recognition using color and texture features: application to camera based video games, The International Conference on Signal Processing Pattern Recognition and Applications (SPPRA), Innsbruck, Austria, 2010

[HamBL10+1] L. Hamoudi, J. Boonaert, and S. Lecoeuche, People recognition based on clothes characterization, The SPIE's Defense, Security and Sensing Symposium, Orlando, USA, 2010

[HamBL10+2] L. Hamoudi, J. Boonaert, and S. Lecoeuche, Appearance-based person recognition using clothes classification, The International Conference on Image Processing Computer Vision and Pattern Recognition (IPCV), pages 415-421, Las Vegas, USA, 2010

[HanYJ08] Z. Han, Q. Ye, J. Jiao, Online feature evaluation for object tracking using Kalman Filter, 19th IEEE International Conference onPattern Recognition (ICPR), pages 1-4, Tampa, Florida, USA, 2008

[Har79] R.M. Haralick, Statistical and structural approaches to texture, Proceedings of the IEEE, vol. 67 (5), pages 786-804, 1979

[HarBD05] S. Harasse, L. Bonnaud, and M. Desvignes, People counting in transport vehicles, Proceedings of the International Conference on Pattern Recognition and Computer Vision, pages 221-224, St. Augustine, FL, USA, 2005

[HarHD00] I. Haritaoglu, D. Harwood, L.S. Davis, W4: real-time surveillance of people and their activities, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22 (8), pages 809-830, 2000

[HaySTA00] P. Hayton, B. Schölkopf, L. Tarassenko, and P. Anuzis, Support vector novelty detection applied to jet engine vibration spectra, In Neural Information Processing Systems, vol. (13), pages 946-952, 2000

[HesF09] R. Hess and A. Fern, Discriminatively trained particle filters for complex multi-object tracking, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 240-247, Miami, Florida, USA, 2009

[HorH99] T. Horprasert and D. Harwood, A statistical approach for real-time robust background subtraction and shadow detection, Technical Report, University of Maryland, 1999

[HorS81] B.K.P. Horn and B.G. Schunk, Detenninig optical flow, Artificial Intelligence, vol. 17, pages 185-203, 1981

[HsuAJ02] R. Hsu, M. Abdel-Mottaleb, A. Jain, Face detection in color images, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24 (5), pages 696-706, 2002

[HsuL01] C. Hsu and C. Lin, A comparison of methods for multi-class support vector machines, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, 2001

[HuaKMZ97] J. Huang, S. Ravi Kumar, M.Mitra, W. Zhu, R. Zabih, Image indexing using color correlograms, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 762-768, San Juan, Puerto Rico, 1997

[HuTWM04] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 34 (3), pages 334-352, 2004

[ICCST09] International Carnahan Conference on Security Technology (ICCST), Zürich, Switzerland, 05-08 Oct 2009

[ICCV09] IEEE 12th International Conference on Computer Vision (ICCV), Kyoto, Japan, 29 Sep-02 Oct 2009

[ICDSC09] 3rd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), Como, Italie, 30 Aug-02 Sep 2009

[ICNS06] Integrated Communications, Navigation and Surveillance Conference (ICNS), Baltimore, MD, USA, 01 May-03 May 2006

[ICNS07] Integrated Communications, Navigation and Surveillance Conference (ICNS), Herndon, VA, USA, 30 Apr-03 May 2007

[ICNS08] Integrated Communications, Navigation and Surveillance Conference (ICNS), Bethesda, MA, USA, 05 May-07 May 2008

[ICNS09] Integrated Communications, Navigation and Surveillance Conference (ICNS), Crystal City, VA, USA, 13 May-15 May 2009

[ICNS10] Integrated Communications, Navigation and Surveillance Conference (ICNS), Herndon, USA, 10 May-13 May 2010

[IPSOS08] Sondage IPSOS, Mars 2008. http://www.ipsos.fr/CanalIpsos/articles/2509.asp

[IsaB96] M. Isard, A. Blake, Contour tracking by stochastic propagation of conditional density, Proceedings of the 4th European Conference on Computer Vision (ECCV), pages 343-356, Cambridge, UK, 1996

[IsaB98] M. Isard, A. Blake, Condensation-conditional density propagation for visual tracking, International Journal of Computer Vision, vol. 29 (1), pages 5-28, 1998

[IsaM01] M. Isard, J. Maccormick, Bramble: A bayesian multiple-blob tracker, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 34-41, Vancouver, Canada, 2001

[ISSNIP09] International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), Melbourne, Australia, 07-10 Dec 2009

[IvaDHE09] I. Ivanov, F. Dufaux, T.M. Ha, T. Ebrahimi, Towards generic detection of unusual events in video surveillance, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 61-66, Genova, Italy, 2009

[JaiRP04] A.K. Jain, A. Ross, S. Prabhakar, An introduction to biometric recognition, IEEE Transactions on Circuits and Systems for Video Technology, vol. 14 (1), pages 4-20, 2004

[JohB01] A. Johnson and A. Bobick, A multi-view method for gait recognition using static body parameters, Proceedings of the 3rd International Conference on Audio-and Video-Based Biometric Person Authentication, pages 301-31, Halmstad, Sweden, 2001

[JohH95] N. Johnson and D. Hogg, Learning the distribution of object trajectories for event recognition, Proceedings of the British Machine Vision Conference (BMVC), pages 583-592, Birmingham, UK, 1995

[JonR02] M.J. Jones, J.M. Rehg, Statistical color models with application to skin detection, International Journal of Computer Vision (IJCV), vol. 46 (1), pages 81-96, 2002

[Jun09] C.R. Jung, Efficient background subtraction and shadow removal for monochromatic video sequences, IEEE Transactions on Multimedia, vol. 11 (3), 2009

[JunJS04] I. Junejo, O. Javed, and M. Shah, Multi feature path modeling for video surveillance, International Conference on Pattern Recognition (ICPR), vol. 2, pages 716-719, Cambridge, UK, 2004

[KakMB07] P. Kakumanu, S.Makrogiannis, N. Bourbakis, A survey of skin-color modeling and detection methods, Pattern Recognition, vol. 40 (3), pages 1106-1122, 2007

[KasWT88] M. Kass, A. Witkin, D. Terzopoulos, Snakes: active contour models, International Journal of Computer Vision, vol. 1 (4), pages 321-331, 1988

[KirRIR08] N. Kiryati, T. iklin-Raviv, Y. Ivanchenko, S. Rochel, Real-time abnormal motion detection in surveillance video (ICPR), pages 1-4, Tampa, Florida, USA, 2008

[KorP97] C. Kotropoulos, I. Pitas, Rule-based face detection in frontal views, Proceedings International Conference Acoustics, Speech and Signal Processing, vol. 4, pages 2537-2540, Munich, Germany, 1997

[Kos09] K. Robert, Night-time traffic surveillance a robust framework for multi-vehicle detection, Classification and Tracking, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 1-6, Genova, Italy, 2009

[KovUSHP09] L. Kovács, Á. Utasi, Z. Szlávik, L. Havasi, I. Petrás, T. Szirányi, Digital video event detector framework for surveillance applications, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 565-570, Genova, Italy, 2009

[KraN09] L. Kratz, K. Nishino, Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1446-1453, Miami, FL, USA, 2009

[KraTYP09] N. Krahnstoever, P. Tu, T. Yu K. Patwardhan, D. Hamilton, B. Yu, C. Greco and G. Doretto, Intelligent video for protecting crowded sports venues, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 116-121, Genova, Italy, 2009

[KirS90] M. Kirby, L. Sirovich, Application of the karhunen-loeve procedure for the characterization of human faces, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12 (1), pages 103-108, 1990

[KurHM98] T. Kurita, K. Hotta, T. Mishima, Scale and rotation invariant recognition method using higher-order local autocorrelation features of log-polar image, Proceedings of the 3rd Asian Conference on Computer Vision (ACCV), vol. 2, pages 89-96, Hong Kong, 1998

[LanTC95] A. Lanitis, C.J. Taylor, and T.F. Cootes, An automatic face identification system using flexible appearance models, Image and Vision Computing, vol. 13 (5), pages 393-401, 1995

[LeeE06] C. S. Lee, A. Elgammal, Gait tracking and recognition using persondependent dynamic shape model, Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, pages 553-559, Southampton, UK, 2006

[LeeG02] L. Lee, W. Grimson, Gait analysis for recognition and classification, 5th IEEE International Conference on Automatic Face and Gesture Recognition, pages 155-162, Washington, DC, USA, 2002

[LeeK07] J. Lee, Y. Kuo, P. Chung, and E. Chen, Naked image detection based on adaptive and extensible skin color model, Pattern Recognition, vol. 40 (8), pages 2261-2270, 2007

[LeiY09] C. Lei and Y.H. Yang, Optical flow estimation on coarse-to-fine region-trees using discrete optimization, International Conference on Computer Vision (ICCV), pages 1562-1569, Kyoto, Japan, 2009

[LiaCC08] H. Liao, J Chang, L. Chen, A localized approach to abandoned luggage detection with foreground -mask sampling, Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 132-139, Santa Fe, New Mexico, USA, 2008

[Liberty09] Liberty CCTV, 2009
http://www.liberty-human-rights.org.uk/ privacy/cctv.shtml.

[LiTCW08] Z. Li, E. Tan, J. Chen, T. Wassantachat, On Traffic Density Estimation with a Boosted SVM classifier, Proceedings of the Digital Image Computing: Techniques and Applications (DICTA), pages 117-123, Washington, DC, USA, 2008

[LiuJZ09] L. Liu, W. Jia, Y. Zhu, Survey of gait recognition, Proceedings of the 5th International Conference on Intelligent Computing (ICIC), page 652-659, Ulsan, South Korea, 2009

[LlaZ08] Y. Llach, J. Zhang, A method of small object detection and tracking based on particle filters, Proceedings of the IEEE 19th International Conference on Pattern Recognition (ICPR), pages 1-4, Tampa, Florida, USA, 2008

[MagTBS09] M. Magno, F. Tombari, D. Brunelli, L. Di Stefano, L. Benini, Multimodal abandoned/removed object detection for low power video surveillance systems, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 188-193, Genova, Italy, 2009

[ManWNS00] B. S. Manjunath, P. Wu, S. Newsam, H. D. Shin, A Texture descriptor for browsing and similarity retrieval, Journal of Signal Processing: Image Communication, vol. 16, pages 33-43, 2000

[MarV00] F. Marqués, V. Vilaplana, A morphological approach for segmentation and tracking of human face, Proceedings 15th International Conference on Pattern Recognition (ICPR), pages 1064-1067, Barcelona, Spain, 2000

[MatYZ05] R. Mathew, Z. Yu, J. Zhang, Detecting new stable objects in surveillance video, Proceedings of the IEEE 7th Workshop on Multimedia Signal Processing, pages 1-4, Shanghai, 2005

[McLP00] G.J. McLachlan, D. Peel, Finite Mixture Models, New York: Wiley, 2000

[MesVE01] D. Messing, P. Van Beek, J. Errico, The MPEG-7 color structure descriptor : image description using color and local spatial information, In International Conference on Image Processing, vol 1, pages 670-673, Thessaloniki, Greece, 2001

[MigM08] J. San Miguel, J. Martnez, Robust unattended and stolen object detection by fusing simple algorithms, Proceedings IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 18-25, Santa Fe, New Mexico, USA, 2008

[MecP05] A. Mecocci, M. Pannozzo, A completely autonomous system that learns anomalous movements in advanced videosurveillance applications, Proceedings of the IEEE International Conference on Image Processing (ICIP), vol. 2, pages 586-9, Genova, Italy, 2005

[MicSF06] C. Micheloni, L. Snidaro, G.L. Foresti, Statistical event analysis for video surveillance, the IEEE International Workshop on Video Surveillance, pages 81-88, Santa Barbara, CA, USA, 2006

[MicSF09] C. Micheloni, L. Snidaro, G. Foresti, Exploiting temporal statistics for events analysis and understanding, Image and Vision Computing, vol. 27 (2), pages 1459-1469, 2009

[MN03] M. McCahill, C. Norris, CCTV systems in London: their structures and practices. On the threshold to Urban Panopticon?: Analysing the Employment of CCTV in European Cities and Assessing its Social and Political Impacts. Technical University Berlin, 2003

[MoeHK06] T.B. Moeslund, A. Hilton, V. Kruger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding, vol. 104 (2), 90-126, 2006

[MonMK09] E. Monari, J. Maerker, K. Kroschel, A Robust and efficient approach for human tracking in multi-camera systems, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 134-139, Genova, Italy, 2009

[MorT08] B. Morris and M. Trivedi, A survey of vision-based trajectory learning and analysis for surveillance, IEEE Transactions on Circuits and Systems for Video Technology, vol. 18 (8), pages 1114-1127, 2008

[Mur67] M.P. Murray, Gait as a total pattern of movement, American Journal of Physical Medicine, vol. 46 (1), pages 290-332, 1967

[MurDK64] M.P. Murray, A.B. Drought, R.C. Kory, Walking patterns of normal men, The Journal of Bone and Joint Surgery, vol. 46 (2), pages 335-360, 1964

[NakPHP03] C. Nakajima, M. Pontil, B. Heisele and T. Poggio, Full-body person recognition system, Pattern Recognition, Kernel and Subspace Methods for Computer Vision, vol. 36 (9), pages 1997-2006, 2003

[NixC04] M.S. Nixon, J.N. Carter, Advances in automatic gait recognition, 6th IEEE International Conference on Automatic Face and Gesture Recognition, pages 139-144, Seoul, Korea, 2004

[NorA99] C. Norris, G. Armstrong, CCTV and the social structuring of surveillance, Crime Prevention Studies, vol. 10, pages 157-178, 1999

[NVWB03] N. T. Nguyen, S. Venkatesh, G. West, H.H. Bui, Multiple camera coordination in a surveillance system, Acta Automatica Sinica, vol.29 (3), pages 408-421, 2003

[OliPB97] N. Oliver, A.Pentland, F. Berard, Lafter: Lips and face real time tracker, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 123-129, Washington, DC, USA, 1997

[OlmCKK02] J. R. Ohm, L. Cieplinski, H. J. Kim, S. Krishnamacha, B. S. Manjunath, D.S. Messing. and A. Yamada, Color descriptors, introduction to MPEG-7: B.S. Manjunath, P. Salembier. Th. Sikora (Eds.), John Wiley & Sons, Ltd., pages 187-212, 2002

[PapOP98] C. Papageorgiou, M. Oren, T. Poggio, A general framework for object detection, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 555-562, Bombay, India, 1998

[PASSIVE08] New Trends for Environmental Monitoring Using Passive Systems (PASSIVE 2008), 14-17 Oct 2008

[PETS03] IEEE Visual Surveillance and Performance Evaluation and Tracking Surveillance (VS PETS), Nice, France, 12-13 Oct 2003

[PETS05] IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (IWVS PETS), Beijing, China, 15-16 Oct 2005

[PETS09] Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), Boston, MA, USA, 24-25 Jun 2009

[Pic04] M. Piccardi, Background subtraction techniques: a review, Proceedings of IEEE International Conference on Systems, Man and Cybernetics (SMC), vol. 4, pages 3099-3104, The Hague, The Netherlands, 2004

[PioNC09] N. Piotto, F. De Natale, N. Conci, Hierarchical matching of 3D pedestrian trajectories for surveillance applications, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 146-151, Genova, Italy, 2009

[PonV98] M. Pontil, A. Verri, Support vector machines for 3-d object recognition, IEEE Transactions on Pattern Analysis Machine Intelligence, vol. 20 (6), pages 637-646, 1998

[Pop07] R. Poppe, Vision-based human motion analysis: an overview, Computer Vision and Image Understanding, vol. 108 (2), pages 4-18, 2007

[Por04] F. M. Porikly, Trajectory distance metric using hidden markov model based representation, European Conference on Computer Vision (ECCV), Prague, Czech Republic, 2004

[Por07] F. Porikli, Detection of temporarily static regions by processing video at different frame rates, Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 236-241, London, UK, 2007

[PorIH08] F. Porikli, Y. Ivanov, T. Haga, Robust Abandoned Object Detection Using Dual Foregrounds, EURASIP Journal on Advances in Signal Processing, vol. 2008 (3), 2008

[Poy95] Poynton, C. A. 1995. Frequently asked questions about colour. ftp://www.inforamp.net/pub/users/poynton/doc/colour/ColorFAQ.ps.gz.

[RadAAR05] R.J. Radke, S. Andra, O. Al-Kofahi, B. Roysam, Image change detection algorithms: a systematic survey, IEEE Transactions on Image Processing, vol. 14 (3), pages 294-307, 2005

[RADAR09] International Radar Conference Radar Surveillance for a Safer World (RADAR 2009), Bordeaux, France, 12-16 Oct 2009

[RemBGH97] P. Remagnino, A. Baumberg, T. Grove, D. Hogg, T. Tan, A. Worral, K. Baker, An integrated traffic and pedestrian model-based vision system, Proceedings of the British Machine Vision Conference (BMVC), pages 380-389, Essex, UK, 1997

[Ron94] R. Ronfard, Region based strategies for active contour models, International Journal of Computer Vision, vol. 13 (2), page 229-251,1994

[RowBA98] H. Rowley, S. Baluja, T. Andkanade, Neural network-based face detection, IEEE Transactions on Pattern Analysis Machine Intelligence, vol. 20 (1), pages 23-38, 1998

[SatA04] K. Sato, J.Aggarwal, Temporal spatio-velocity transform and its application to tracking and interaction, Comput Vision Image Understand, vol. 96 (2), pages 100-128, 2004

[SchBV95] B. SchUolkopf, C. Burges, V. Vapnik, Extracting support data for a given task, Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining, pages 252-257, Montreal, Canada, 1995

[SchK98] H. Schneiderman and T. Kanade, Probabilistic modeling of local appearance and spatial relationships for object recognition, Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, pages 45-51, Santa Barbara, CA, USA, 1998

[SebL00] N. Sebe, M. S. Lew, A maximum likelihood, investigation into color indexing, Proceedings Visual Interface, pages 101-106, Montreal, Canada, 2000

[SeiH06] F. Seitner, A. Hanbury, Fast pedestrian tracking based on spatial features and colour, 11th Computer Vision Winter Workshop (CVWW), pages 105-110, Telč, Czech Republic, 2006

[ShiCT02] M.C. Shin, K.I. Chang, L.V. Tsap, Does colorspace transformation make any difference on skin detection? In: WACV: Proceedings of the 6th IEEE Workshop on Applications of Computer Vision, Orlando, Flordia, USA, 2002

[ShiKKL05] J. Shin, S. Kim, S. Kang, S.-W. Lee, J. Paik, B. Abidi, M. Abidi, Optical flow-based real-time object tracking using non-prior training active feature model, Real-Time Imaging, vol. 11 (3), pages 204-218, 2005

[ShiT94] J. Shi, C. Tomasi, Good features to track, IEEE Conference on Computer Vision and Pattern Recognition, pages 593-600, Seattle, WA, USA, 1994

[SinCVS03] S. Kr. Singh, D. S. Chauhan, M. Vatsa, R. Singh, A robust skin color based face detection algorithm, Tamkang Journal of Science and Engineering, vol. 6 (4), pages 227-234, 2003

[SinSMM09] A. Singh, S. Sawan, M. Hanmandlu, V.K. Madasu, B.C. Lovell, An abandoned object detection system based on dual background segmentation, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 352-357, Genova, Italy, 2009

[Smi04] G.J.D. Smith, Behind the screens : examining constructions of deviance and informal practices among CCTV control room operators in the UK, Surveillance & Society, vol. 2 (2), pages 376-395, 2004

[SmoS98] A. J. Smola, B. Schölkopf, On a kernel-based method for pattern recognition, regression, approximation and operator inversion, Algorithmica, vol. 22 (1), pages 211-231, 1998

[SniPF06] L. Snidaro, C. Piciarelli, G.L. Foresti, Activity analysis for video security systems, Proceedings of the IEEE International Conference on Image Processing (ICIP), pages 1753-1756, Atanta, GA, USA, 2006

[SorHML00] M. Soriano, S. Huovinen, B. Martinkauppi, Skin detection in video under changing illumination conditions, Proceedings 15th International Conference on Pattern Recognition, vol. 1, pages 839-842, Barcelona, Spain, 2000

[SPJ7] Special issue on Visual Surveillance, International Journal of Computer Vision, June 2000

[SPJ8] Special issue on Visual Surveillance, IEEE Transactions on Pattern Analysis and Machine Intelligence, August 2000

[SPJ9] Special issue on third generation surveillance systems, Proceedings of IEEE, October 2001

[SPJ10] Special issue on human motion analysis, Computer vision and Image Understanding, March 2001

[Sta03] C. Stauffer, Estimating tracking sources and sinks, Proceedings of 2nd IEEE workshop on event mining, pages 259-266, Madison, WI, USA, 2003

[Sun10] H. Sun, Skin detection for single images using dynamic skin color modeling, Pattern Recognition, vol. 43 (4), pages 1413-1420, 2010

[SunP98] K. Sung and T. Poggio, Example-based learning for view-based human face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20 (1), pages 39-51, 1998

[TaoT03] J. Tao et Y. Tan, Color appearance-based approach to robust tracking and recognition of multiple people, Proceedings of the Joint Conference of the 4th International Conference on Information Communications and Signal Processing, and the 4th Pacific Rim Conference on Multimedia, vol. 1, pages 95-99, 2003

[TIWDC08] Tyrrhenian International Workshop on Digital Communications-Enhanced Surveillance of Aircraft and Vehicles (TIWDC/ESAV), 03-05 Sep 2008

[TurP91] M. Turk, A. Pentland, Face recognition using eigenfaces, Proceedings Computer Vision Pattern Recognition, pages 586-591, Maui, HI, USA, 1991

[ToyKBM99] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, Wallflower: principles and practice of background maintenance, Proceedings International Conference Computer Vision, Kerkyra, Greece, 1999

[UnnRJ3] R. Unnthorsson, T. Runarsson and M. Jonsson, Model selection in one class nu-svms using rbf kernels, In 16th conference on Condition Monitoring and Diagnostic Engineering Management, August 2003

[Vap95] V. Vapnik, The nature of statistical learning theory, New York: Springer-Verlag, 1995

[VeeRB01] C. Veenman, M. Reinders, E. Backer, Resolving motion correspondence for densely moving points, IEEE Transactions on Pattern Analysis Machine Intelligence, vol. 23 (1), pages 54-72, 2001

[VeiB05] T. Veit, F. Cao, P. Bouthemy, A maximality principle applied to a contrario motion detection, Proceedings of the IEEE International Conference Image Processing (ICIP), pages 1061-1064, Genova, Italy, 2005

[VelW05] S. Velipasalar, W. Wolf, Multiple object tracking and occultation handling by information exchange between uncalibrated cameras, Proceedings of the IEEE International Conference on Image Processing (ICIP), Genova, Italy, 2005

[VioJS03] P. Viola, M. Jones, D. Snow, Detecting pedestrians using patterns of motion and appearance, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 734-741, Nice, France, 2003

[Vog93] M. Vagt, Combination of Radial basis function neural network with optimized laming vector quantization, Proceedings of the IEEE International Conj on Neural Network, CNN-93. vol. 3, pages 1841-1846, San Francisco, CA, USA, 1993

[WagN04] D.K Wagg, M.S. Nixon, On automated model-based extraction and analysis of gait, Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition, pages 11-16, Seoul, Korea, 2004

[WanAR09] P. Wang, G. Abowd, J. Rehg, Quasi-periodic Event analysis for social game retrieval, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 112-119, Kyoto, Japan, 2009

[WanHY09] X. Wang, T. Han, S. Yan, An HOG-LBP human detector with partial occultation handling, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 32-39, Kyoto, Japan, 2009

[WanMY08] J. Wang, H. Man, Y. Yin, Tracking human body by using particle filter gaussian process markov-switching model, Proceedings of the IEEE 19th International Conference on Pattern Recognition (ICPR), pages 1-4, Tampa, Florida, USA, 2008

[WanNTH04] L. Wang, H. Ning, T. Tan, and W. Hu, Fusion of Static and dynamic body biometrics for gait recognition, IEEE Transactions on Circuits and Systems for Video Technology, vol. 14 (2), pages 149-158, 2004

[WanP03] J. R. Wang and N. Parameswaran, Survey of Sports video analysis: research issues and applications, Proceedings of the Pan-Sydney area workshop on Visual Information Processing (VIP), Darlinghurst, Australia, 2003

[WanTNH03] L. Wang, T. Tan, H. Ning, W. Hu, Silhouette analysis-based gait recognition for human identification, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25 (12), pages 1505-1518, 2003

[WasLCW09] T. Wassantachat, Z. Li, J. Chen, Y. Wang, E. Tan, Traffic density estimation with on-line SVM classifier, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 13-18, Genova, Italy, 2009

[WD88] E. Wallace, C. Diffley, CCTV control room ergonomics. Technical Report 14/98, Police Scientific Development Branch (PSDB), UK Home Office, 1988

[WelB01] G. Welch, G. Bishop, An Introduction to the Kalman Filter, the Association for Computing Machinery's Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH) Course, 2001

[WIFS09] 1st IEEE International Workshop on Information Forensics and Security (WIFS), London, UK, 06-09 Dec 2009

[WonLS03] K.W.Wong, K.M. Lam,W.C. Siu, A robust scheme for live detection of human faces in color images, Signal Processing: Image Communication, vol. 18 (2), pages 103-114, 2003

[WecPBS98] H. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulie, T.S. Huang, Face recognition : from theory to applications, Springer-Verlag, pages 51-72, 1998

[WuN07] B. Wu, R. Nevatia, Detection and tracking of multiple partially occluded humans by bayesian combination of edgelet based part detectors, International Journal of Computer Vision, vol. 75 (2), pages 247-266, 2007

[XiaG06] T. Xiang, S. Gong, Beyond tracking: modelling activity and understanding behaviour, International Journal of Computer Vision, vol. 67 (1), pages 21-51, 2006

[XiaL08] L. Xiao, P. Li, Improvement on Mean Shift based tracking using second-order information, Proceedings of the IEEE 19th International Conference on Pattern Recognition (ICPR), pages 1-4 Tampa, Florida, USA, 2008

[XioRD03] Z. Xiong, R. Radhakrishnan, A. Divakaran, Generation of sports highlights using motion activities in combination with a common audio feature extraction framework, Mitsubishi Electric Research Laboratories Inc., 2003

[XinLS04] A. Xin and X. Li, M. Shah, Object contour tracking using level sets, Asian Conference on Computer Vision (ACCV), Jeju, Korea, 2004

[YamNC04] C.Y. Yam, M.S. Nixon, J.N. Carter, Automated person recognition by walking and running via model-based approaches, Pattern Recognition, vol. 37 (5), pages 1057-1072, 2004

[YanH94] G. Yang and T. S. Huang, Human face detection in complex background, Pattern Recognition, vol. 27 (1), pages 53-63, 1994

[YanKM02] M. Yang, D. Kriegman, N. Ahuja, Detecting faces in images: a survey, IEEE Transactions on Pattern Analysis and Machine Intelligence In Pattern Analysis and Machine Intelligence, vol. 24 (1), pages 34-58, 2002

[YanLW98] J. Yang, W. Lu, A. Waibel, Skin-color modeling and adaptation, Proceedings of Asian Conference on Computer Vision (ACCV), pages 687-694, Hong Kong, China, 1998

[YilXS04] A. Yilmaz, L. Xin, M. Shah, Contour-based object tracking with occultation handling in video acquired using mobile cameras, Transactions on Pattern Analysis and Machine Intelligence, vol. 26 (11), pages 1531-1536, 2004

[YonCJ04] A. Yoneyama, Y. Chia-Hung, and C. Jay Kuo, Robust vehicle and traffic information extraction for highway surveillance, EURASIP Journal on Applied Signal Processing vol. (14), pages 2305-2321, 2004

[YowC97] K.C. Yow and R. Cipolla, Feature-based human face detection, Image and Vision Computing, vol. 15 (9), pages 713-735, 1997

[YuTH09] S.Q. Yu, T.N. Tan, K.Q. Huang, et al., A Study on gait based gender classification, IEEE Transactions on Image Processing, vol. 18 (8), pages 1905-1910, 2009

[ZarSQ99] B. D. Zarit, B. J. Super, and F. K. H. Quek, Comparison of five color models in skin pixel classification, In ICCV International Workshop on recognition, analysis and tracking of faces and gestures in Real-Time systems, pages 58-63, Kerkyra, Greece, 1999

[Zha03] W. Zhao, R. Chellapa, P. J. Phillips, A. Rosenfeld, Face recognition: a literature survey, ACM Computing Surveys, vol. 35 (4), pages 399-458, 2003

[ZhaDC09] X. Zhao, Em. Dellandrea and L. Chen, A People counting system based on face detection and tracking in a video, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 67-72, Genova, Italy, 2009

[ZhaH09] Y. Zhai, A. Hampapur, Virtual boundary crossing detection without explicit object tracking, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 518-522, Genova, Italy, 2009

[ZhaKR08] C. Zhao, A. Knight, Reid, Target tracking using mean-shift and affine structure, Proceedings of the IEEE 19th International Conference on Pattern Recognition (ICPR), pages 1-5, 2008

[ZheFZ05] J. B. Zheng, D. Feng, and R. Zhao, Trajectory matching and classification of video moving objects, Proceedings of the IEEE Multimedia Signal Processing, pages 1-4, Shanghai, 2005

[ZheLSZ09] Y. Zheng, G. Li, X. Sun, X. Zhou, A geometric active contour model without re-initialization for color images, Image and Vision Computing, vol. 27 (9), pages 1411-1417, 2009

[ZheZW04] Q.F. Zheng, M.J. Zhang, W.Q. Wang, A hybrid approach to detect adult web images, Advances in Multimedia Information Processing PCM, vol. 3332 (2005), pages 609-616, 2004

[ZhuHLL09] P. Zhu, W. Hu, X. Li, L. Li, Segment model based vehicle motion analysis, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 514-517, Genova, Italy, 2009

[ZouBT09] N. Zouba, F. Bremond, M. Thonnat, Multisensor fusion for monitoring elderly activities at home, Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 98-103, Genova, Italy, 2009