Contribution à l'analyse et à l'interprétation du mouvement humain: application à la reconnaissance de postures

Vincent Girondel

▶ **To cite this version:**

Vincent Girondel. Contribution à l'analyse et à l'interprétation du mouvement humain: application à la reconnaissance de postures. Interface homme-machine [cs.HC]. Institut National Polytechnique de Grenoble - INPG, 2006. Français. NNT : . tel-00156572

HAL Id: tel-00156572
https://theses.hal.science/tel-00156572

Submitted on 21 Jun 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

$N°$ attribué par la bibliothèque

_____

THÈSE

pour obtenir le grade de

DOCTEUR DE L'INPG

Spécialité : «Signal, Image, Parole, Télécommunications»

préparée au Laboratoire des Images et des Signaux de Grenoble

dans le cadre de **l'École Doctorale EEATS**
«Électronique, Électrotechnique, Automatique, Télécommunications, Signal»

présentée et soutenue publiquement

par

Vincent GIRONDEL

le 19 Juin 2006

Titre :

CONTRIBUTION À L'ANALYSE ET À L'INTERPRÉTATION DU MOUVEMENT HUMAIN : APPLICATION À LA RECONNAISSANCE DE POSTURES

Directeurs de thèse : J.-M. CHASSERY, A. CAPLIER et L. BONNAUD

JURY

Monsieur	Pierre-Yves COULON,	Président
Monsieur	Liming CHEN,	Rapporteur
Monsieur	Olivier COLOT,	Rapporteur
Monsieur	Jean-Marc CHASSERY,	Directeur de thèse
Madame	Alice CAPLIER,	Co-encadrante de thèse
Monsieur	Laurent BONNAUD,	Co-encadrant de thèse
Madame	Catherine ACHARD,	Examinatrice

Il existe peut-être un autre moyen de savoir. C'est de renoncer à connaître, et de chercher à comprendre.
"Mon pauvre enfant", me dit un jour un vénérable vieillard curé ému par mon angoisse et qui avait lui-même trouvé depuis longtemps la paix dans les automatismes d'une foi enfantine, "mon pauvre enfant, ce sont des mystères, ne cherchez pas à comprendre..."
Si. Justement si. Je n'y parviendrai peut-être jamais, mais jusqu'à mon dernier souffle, je chercherai à comprendre. Comprendre où je suis et ce que je suis et ce que j'y fais, et à quoi ça rime.
(...)
Quelque application qu'on y mette, il est difficile de croire que le monde n'est qu'un tas confus, un ramassis de matière assemblé fortuitement et battu comme blanc d'œuf par le fouet des énergies de hasard. De l'infiniment grand à l'infiniment petit, l'examen des ensembles et des détails nous montre au contraire que *tout est en ordre*. Non seulement en ordre, mais organisé. Face aux singuliers problèmes que posent cet ordre et cette organisation, la solution rationaliste est de se satisfaire du "comment". Sachant *comment* se déroule une suite de phénomènes, ayant fait la lumière sur le fonctionnement d'un mécanisme cosmique ou biologique, on s'estime satisfait. *Le monde est, et il est ainsi. Il n'y a pas lieu de chercher à en savoir plus long.*
Cela ressemble singulièrement au ne-cherchez-pas-à-comprendre du vénérable vieillard curé.
Eh bien, qu'on m'excuse, tant que j'aurai un souffle de vie, je chercherai à en savoir plus long, même si tous mes désirs et tous mes efforts ne me font pas grimper d'un échelon.
Le loup captif qui sans cesse va et vient derrière ses barreaux me paraît plus raisonnable, sinon plus rationnel, que celui qui résigné se couche en rond dans la paille. En essayant, mille fois l'heure, de franchir l'infranchissable, qui sait ? Peut-être un jour il passera...

René BARJAVEL, "*La Faim du Tigre*".

Remerciements

Je tiens tout d'abord à remercier mon directeur de thèse, M. Jean-Marc CHASSERY, pour m'avoir permis d'effectuer ma thèse et m'avoir accueilli au LIS.

Ensuite, je tiens beaucoup à remercier mes deux co-encadrants de thèse, Mme Alice CAPLIER et M. Laurent BONNAUD, qui m'ont dirigé depuis mon arrivée au LIS et ont grandement contribué à faire de ce travail ce qu'il est. Plus particulièrement, je remercie :

- Alice, pour ses encouragements et ses recommandations. Grâce à elle et à ses nombreux conseils, j'ai pu réaliser à quel point un encadrement bénéfique est profitable.
- Laurent, pour son soutien et ses connaissances avancées en informatique, qui a su rendre passionnant et m'a facilité l'ensemble du travail de développement.

Je les remercie aussi pour leur disponibilité, leur amabilité et pour les nombreux conseils scientifiques avisés qu'ils m'ont prodigué tout au long de ces années, années qui m'ont permis de mieux appréhender le domaine de la vision par ordinateur.

Je remercie les membres de mon jury de m'avoir fait l'honneur d'évaluer mon travail :

- M. Pierre-Yves COULON, pour avoir accepté de présider mon jury. PYC a pu ainsi observer la motivation qu'ont suscité en moi ses cours de traitement d'images ;
- M. Liming CHEN, pour avoir rapporté mon mémoire et pour ses critiques constructives ;
- M. Olivier COLOT, qui a également rapporté mon mémoire et m'a permis par ses questions et remarques d'améliorer substantiellement celui-ci ;
- Mme Catherine ACHARD, pour l'intérêt qu'elle a porté à mon travail.

Pour citer M. Patrick BAS, *"Une thèse est un long voyage en solitaire, où les meilleurs compagnons sont souvent les publis"*. Même si ce n'est pas faux, un(e) doctorant(e) est amené(e) à côtoyer un grand nombre de gens au cours de sa thèse. Après les remerciements officiels, il est grand temps de passer maintenant aux remerciements officieux.

Pour commencer, je remercie les administrateurs systèmes (Hervé, Denis, Jean-Marc et plus anciennement Matthieu). On a tendance à les oublier un peu trop souvent mais ce sont eux qui volent à notre secours dès que les bécanes commencent à planter... Un grand merci donc, et en particulier à Hervé qui a sauvé ma thèse (entre autres) lors du crash magistral du 23-24 Mars 2005.

Ensuite, je voudrais remercier les permanents du labo, les chercheurs, les enseignants-chercheurs, les ingénieurs et les secrétaires. Merci pour leurs conseils, leur aide et leur soutien. Voici quelques spéciales dédicaces pour la bande formée par Bidou, Céd, Jim, Jocelyn, Nico, Pierre et à laquelle s'est ajouté (ou incrusté) plus récemment Fwansswa. Pourquoi ce favoritisme ? (En effet, il n'y en a pas, c'est par ordre alphabétique). Ces personnes m'ont fait comprendre que l'on pouvait à la fois faire de la recherche, de l'enseignement, et rester cool et ouvert. Ces petits malins ont aussi passé la majorité du temps pendant lequel je rédigeais et préparais ma soutenance à me seriner quelques phrases, avec quelquefois des variantes :

- "Alors, ça avance cette thèse ?" (rédaction) ;

– “Tu stresses, là ?” (soutenance) ;
– “Tiens, t’es là aujourd’hui ?” (rédaction et soutenance) ;
– “Je prendrais bien une graine, moi. . .” (rédaction et soutenance).

Je ne suis pas complètement sûr que ces formulations symbolisaient d’une quelconque façon des encouragements, mais je les ai considérées comme tels. Merci en particulier à Nico pour les graines, les discussions plurilangues, la “bonne” musique et surtout pour m’avoir supporté comme co-bureau. Merci à Bidou (celui dont l’accent américain est unique), Céd (celui qui arrive à fusionner art, poésie et culture gé’), Jim (celui qui porte des chemises), Jocelyn (celui qui ne m’a pas réveillé), Pierre (celui à qui on peut commander une pizza) et Fwansswa (celui qui a un certain humour, si ce n’est un humour certain). Merci à eux pour leur bonne humeur, leur humour, leur soutien inconditionnel et leurs discussions enflammées.

Après les permanents, voici venu le tour des post-docs / ATER / visiteurs, des doctorants et des stagiaires. Merci à tous pour les moments de détente et pour m’avoir soutenu quand la galère commençait à couler afin de la remettre à flots. Une liste exhaustive étant impossible, je remercie, entre autres, Alex, Amine, Anthony, Barb (qui m’a permis de faire un sauvetage *in extremis*), Ben (surtout pour sa passion du foot), Brice, Caro (alias Kro), Chouchou (le karatéka), Corentin, Franchouzze (pour sa coupe de cheveux), Frédo (Salut Fred !), Jérémy, Ju, Manu, Matthieu, Marion, Max, Mickaël, Pierre B., Sébastien, Séby (pour sa *palinca*), Sushi (pour son retard), Thomas et Zakia.

Parmi mon entourage, je voudrais aussi remercier tous mes potes de DEA, d’école d’ingé, de prépa, de lycée, de collège et d’avant, qu’ils soient de Caen, de Grenoble ou de Paris, de Bretagne, d’Isère ou de Normandie, ou même de France ou d’ailleurs, avec qui je garde et garderai encore longtemps de très bons contacts. Ces personnes ne font pas forcément partie du monde de la recherche mais m’ont fortement soutenu (les fois où l’on a réussi à se croiser). Parmi eux, je tiens d’abord à remercier Cyrille, David, Julien et Sébastien (les “vieux de la vieille” comme on dit, alias Furcoat, Morgaliel, le Maître et Seagore). Puis, par ordre alphabétique, avec peut-être des oublis (forcément involontaires), je remercie grandement aussi Alban et Lilou, Alexandra, Amélie, Anne-Lise, Anthony, Aurélie, Barb, Ben et Flo, Box et Cécile, Cécile, Céd, Chloé, Christo et Sandrine, Chouchou et Gwen, Clémence, Dimi et Isa, Élodie, Fab, Fabos, Franchouzze, Fredboowl, Furcoat et Cécilou, Fwansswa, Géraldine, Gwéna, Hervé, Isabelle, Jack, Jejen, Julien et Émilie, Kro, Ludo et Fifi, Marc et... Sophie (évidemment), Maïlys, Marion, Mickey et Véro, Manu, Matthieu T., Matthieu C., Maurane, Nico C. (alias malaka), Nico P., Nico, Pierre, Sébastien, Séb et Flore, Seby et Gina, Sly, Sophie, Stéphanie, Sushi, Tania (alias Pititouzbek), Titi, Titim’s, Tom-Tom, Valtouf, Véro et Ludo, et Vivi. Toutes ces personnes mais aussi celles que j’ai malheureusement oubliées comptent beaucoup pour moi et je les remercie du fond du cœur de faire partie de ma vie.

On dit généralement qu’on peut choisir ses amis mais pas sa famille. Je tiens quand même à remercier en premier lieu mes parents (bien sûr) et ensuite mes trois frères (Olivier, Guillaume et Étienne) pour leur soutien psychologique et affectif.

Pour terminer ces remerciements, je voudrais remercier Aurélia qui m’a supporté, soutenu et m’a témoigné sa confiance pendant toute la fin de ma thèse.

Un grand merci à toutes celles et à tout ceux qui ont traversé ma vie pendant cette grande aventure qu’a été ma thèse. Que tous soient persuadés qu’ils / elles sont et resteront dans mon cœur. Merci !

“Un proverbe chinois dit que, lorsque l’on n’a plus rien à dire, on cite généralement un proverbe chinois”.
Proverbe chinois

Table des matières

Liste des Figures

Liste des Tableaux

Glossaire

Voici le glossaire des principaux sigles, des abréviations, des définitions et des notations mathématiques utilisés dans ce manuscrit de thèse. Pour plus de clarté, nous avons séparé les sigles et abréviations des définitions et notations mathématiques.

Sigles et abréviations

- 1D : 1 Dimension
- 2D : 2 Dimensions
- 3D : 3 Dimensions
- AIM : plate-forme interactive Analyse Interprétation Multimodalités
- *Art.live (ARchitecture and authoring Tools prototype for Living Images and new Video Experiments)* : projet *IST (Information Society Technology)* n°10942
- ARTUS : projet d'Animation Réaliste par Tatouage audiovisuel à l'Usage des Sourds
- BAP ou BAPS : Boîte par Axes Principaux issue de la Segmentation
- BER ou BERS : Boîte Englobante Rectangulaire issue de la Segmentation
- BERV : Boîte Englobante Rectangulaire du Visage
- BEREP : Boîte Englobante Rectangulaire Estimée *a posteriori* de la personne
- BEREV : Boîte Englobante Rectangulaire Estimée *a posteriori* du Visage
- BERPP : Boîte Englobante Rectangulaire Prédite (estimée *a priori*) de la Personne
- BERPV : Boîte Englobante Rectangulaire Prédite (estimée *a priori*) du Visage
- BO : Boîte Octogonale
- BQ : Boîte Quadrangulaire ou quadrangle
- *DARPA (Defense Advanced Research Projects Agency)* : agence américaine pour les projets de recherche avancée de la défense
- *DTW (Dynamic Time Warping)* : transformation dynamique temporelle
- GP : Groupe de Personnes
- GPPar : mode de filtrage de Kalman (Groupe de Personnes Partiel)
- GPPre : mode de filtrage de Kalman (Groupe de Personnes Prédictif)
- *H-ANIM (Humanoid Animation Working Group)* : groupe de travail sur l'animation de personnages virtuels humanoïdes
- *HMM (Hidden Markov Models)* : chaînes de Markov cachées
- *HSI (Hue, Saturation, Intensity)* : espace couleur (Teinte, Saturation, Intensité)
- *ID (Identification Data)* : numéro d'identification
- IHM : Interfaces Homme-Machine
- LPC : Langage Parlé Complété
- *MAP (Maximum A Posteriori)* : (probabilité) maximum *a posteriori*

- *MPEG-4 (Motion Picture Expert Group)* : format informatique normalisé permettant de stocker, d'intégrer et de diffuser des objets audiovisuels naturels et synthétiques (créés sur un ordinateur). Cela inclut : la vidéo, l'audio, les graphismes en 2D, et les mondes virtuels en 3D.
- *NN (Neural Network)* : réseau de neurones
- *Pfinder (Person Finder)* : littéralement "Trouveur de Personne", système d'analyse et d'interprétation du mouvement humain
- PC : Partie du Corps
- PS : Personne Seule
- PSComp : mode de filtrage de Kalman (Personne Seule Complet)
- PSPar : mode de filtrage de Kalman (Personne Seule Partiel)
- *RGB* ou *rgb (Red, Green, Blue)* : espace couleur (Rouge, Vert, Bleu)
- *ROI (Regions Of Interest)* : régions d'intérêt
- *TBM (Transferable Belief Model)* : modèle de croyance transférable
- *VSAM (Video Surveillance And Monitoring)* : vidéo surveillance et contrôle, système d'analyse et d'interprétation du mouvement humain
- W^4 *(Who ? When ? Where ? What ?)* : littéralement "Qui ? Quand ? Où ? Quoi ?", système d'analyse et d'interprétation du mouvement humain
- $YCbCr$: espace couleur

Définitions et notations mathématiques

- A_t : matrice d'évolution du modèle, à l'instant t
- B_t : matrice de commande, à l'instant t
- $BetP$: probabilité pignistique
- C : ensemble des cliques $c = (s, r)$
- C_i : classifieurs
- C_t : matrice de mesure, d'observation, à l'instant t
- Cb : composante de chrominance de l'espace couleur $YCbCr$ (décalage bleu)
- Cr : composante de chrominance de l'espace couleur $YCbCr$ (décalage rouge)
- $Crit$: critère de décision
- D_i : distances
- D_i^{ref} : distances de référence
- E : champ d'étiquettes d'une image
- E_f : champ d'étiquettes finales d'une image
- $E[\cdot]$: espérance mathématique
- F_{Cb} : fonction cumulative de fréquence pour la composante Cb
- F_{Cr} : fonction cumulative de fréquence pour la composante Cr
- G_{Cb} : fonction de répartition pour la gaussienne théorique de la composante Cb
- G_{Cr} : fonction de répartition pour la gaussienne théorique de la composante Cr
- G_t : matrice de gain de Kalman, à l'instant t
- H *(Hue)* : teinte, composante de couleur de l'espace HSI
- H : entropie
- H_i : hypothèses et postures
- I : image
- I_{ref} : image de référence

- I_v : indicateur d'activité du visage
- I_{md} : indicateur d'activité de la main droite
- I_{mg} : indicateur d'activité de la main gauche
- $I(s,t)$: luminance du site s pour l'image I à l'instant t
- $I_{ref}(s,t)$: luminance du site s de l'image de référence I_{ref} à l'instant t
- Id : matrice identité
- L_d : Liste triée des taches de peau les plus à droite
- L_g : Liste triée des taches de peau les plus à gauche
- L_h : Liste triée des taches de peau les plus hautes
- L_s : Liste triée des taches de peau les plus grandes (surface)
- L_v : Liste triée des taches de peau les plus proches de la dernière localisation du visage
- L_{dq} : Liste triée des taches de peau les plus proches du coin droit du quadrangle
- L_{gq} : Liste triée des taches de peau les plus proches du coin gauche du quadrangle
- L_{hq} : Liste triée des taches de peau les plus proches du coin haut du quadrangle
- L_{md} : Liste triée des taches de peau les plus proches de la dernière localisation de la main droite
- L_{mg} : Liste triée des taches de peau les plus proches de la dernière localisation de la main gauche
- M_i : information fusionnée sur la décision d_i
- M_i^j : information fournie par la source S_j sur la décision d_i
- N : fonction de nécessité
- O_{mvt} : champ d'observation de la différence entre deux images consécutives
- O_{ref} : champ d'observation de la différence entre l'image courante et l'image de référence
- $Pr[\cdot]$: probabilité
- P_0 : valeur initiale de la matrice de covariance, aussi notée $P_{0/-1}$
- $P_{t/t}$: matrice de covariance estimée *a posteriori*, à l'instant t
- $P_{t/t-1}$: matrice de covariance prédite (estimée *a priori*), à l'instant t
- Q : matrice de bruit du modèle
- R : matrice de bruit des mesures
- R : redondance
- S_j : ensemble de sources ou de capteurs
- S_{min} : surface minimale pour la boîte englobante rectangulaire d'une tache de peau
- $U(e, o_{mvt}, o_{ref})$: fonction d'énergie
- $U_m(e)$: terme d'énergie
- $U_a(o_{mvt}, e)$: terme d'énergie d'adéquation lié à la réalisation o_{mvt}
- $U_a(o_{ref}, e)$: terme d'énergie d'adéquation lié à la réalisation o_{ref}
- V_c : fonction de potentiel associée à une clique particulière
- Y : composante de luminance de l'espace couleur $YCbCr$
- *bel (belief)* : crédibilité ou croyance
- $c = (s,r)$: clique, paire des sites s et r, voisin de s
- d_{bqmax} : distance maximale entre un coin du quadrangle et une tache de peau
- d_i : ensemble de décisions
- d_{tmax} : distance maximale entre une dernière position connue et une tache de peau
- $e = e(s,t), s \in I$: réalisation particulière du champ d'étiquettes E
- $e_f = e_f(s,t), s \in I$: réalisation particulière du champ d'étiquettes finales E_f
- $e_s = e(s,t) = e(x,y,t)$: étiquette du site s
- $f(x|x_1)$: densité de probabilité gaussienne conditionnelle de x sachant x_1

- $f(x|x_2)$: densité de probabilité gaussienne conditionnelle de x sachant x_2
- $f(x|x_1, x_2)$: densité de probabilité gaussienne conditionnelle de x sachant x_1 et x_2
- g_{Cb} : gaussienne théorique pour la composante Cb
- g_{Cr} : gaussienne théorique pour la composante Cr
- h_{Cb} : histogramme normalisé pour la composante Cb
- h_{Cr} : histogramme normalisé pour la composante Cr
- $height$: hauteur d'une boîte
- m_{r_i} : distributions de masses
- $n(s)$: bruit blanc gaussien associé au site s
- o_{mvt} : réalisation particulière du champ d'observation O_{mvt}
- o_{ref} : réalisation particulière du champ d'observation O_{ref}
- pl *(plausibility)* : plausibilité
- r : site ou pixel voisin du site s
- r_i : mesures pour la reconnaissance de postures
- $s = (x, y)$: site ou pixel de coordonnées (x, y) dans l'image
- s_t : vecteur de mesures, d'observations, à l'instant t
- t : temps
- u_t : vecteur de commande
- v_t : bruit de mesure, d'observation, à l'instant t
- w_t : bruit de commande à l'instant t
- $width$: largeur d'une boîte
- $\underline{x}_t$: vecteur d'état à l'instant t
- $\underline{x}_0$: valeur initiale du vecteur d'état, aussi notée $\underline{\tilde{x}}_{0/-1}$
- $\underline{\tilde{x}}_{t/t}$: vecteur d'état estimé *a posteriori*, à l'instant t
- $\underline{\hat{x}}_{t/t-1}$: vecteur d'état prédit (estimé *a priori*), à l'instant t
- z^{-1} : opérateur retard
- β_r : paramètre pour la fonction de potentiel V_c
- ϵ : ensemble des réalisations possibles du champ d'étiquettes E
- $\gamma(s, t)$: paramètre pour la mise à jour de I_{ref}
- λ_{mvt} : coefficient de pondération lié au terme d'énergie $U_a(o_{mvt}, e)$
- λ_{ref} : coefficient de pondération lié au terme d'énergie $U_a(o_{ref}, e)$
- μ_x, μ_y : moments du premier ordre
- Ω : cadre de discernement
- 2^Ω : espace de définition
- Π : fonction de possibilité
- π : distributions de possibilités
- $\psi(e(s, t))$: fonction de lien entre observation et étiquette
- σ_x : écart-type de x
- σ_x^2 : variance de x
- $trace$: opérateur de la trace d'une matrice
- $|\cdot|$: nombre d'éléments

Chapitre 1

Introduction

Le début du $21^{ème}$ siècle montre actuellement l'avènement du multimédia et des technologies de traitement et de transmission de l'information. Le développement d'Internet et les progrès scientifiques majeurs en technologie et en informatique permettent de communiquer en temps-réel par de nombreux moyens avec le monde entier. L'image et la parole, liées à l'homme par les deux sens essentiels que sont la vue et l'ouïe, sont prépondérantes : presse, radio, télévision, téléphonie fixe ou mobile, *SMS (Short Message Service)*, e-mails, forums de discussion sur Internet, vidéoconférence etc. en sont la preuve. La téléphonie est en train d'intégrer l'image conjointement à la parole dans les systèmes de visiophonie, *MMS (Multimedia Message Service)* etc. Il y a donc, d'un côté, une envie d'essayer de fusionner les types d'information pour obtenir une communication distante quasi identique à une communication normale entre deux personnes proches physiquement, où une grande quantité d'information est contenue non seulement dans la discussion elle-même, mais aussi dans l'attitude corporelle, les gestes, le regard, et les expressions du visage [Wang03a, Wang03b]. D'un autre côté, de nombreuses recherches sont actuellement menées pour l'amélioration des interfaces homme-machine (IHM). Le but de ces recherches est, entre autres, de rendre la machine suffisamment "intelligente" pour qu'elle puisse analyser et comprendre des comportements humains, et qu'elle réagisse ensuite de manière cohérente selon les différentes situations auxquelles elle est confrontée.

L'analyse du comportement humain en vision par ordinateur est donc un secteur de recherche très actif. Il consiste à **détecter**, à **suivre** au cours du temps, à **reconnaître** les actions et / ou les activités et à **réagir** aux comportements de **personnes**, et de façon plus générale, à **analyser et à interpréter le mouvement humain**.

Ce mémoire de thèse : **contribution à l'analyse et à l'interprétation du mouvement humain : application à la reconnaissance de postures**, présente les différentes étapes de traitement d'un système conçu pour analyser et interpréter le mouvement humain avec les approches utilisées, leurs avantages, leurs limitations et les résultats obtenus.

Une description générale de notre système ainsi que le plan du mémoire sont donnés dans la partie 1.3. Étant donnée la difficulté de réaliser un état de l'art général et / ou exhaustif sur l'analyse et l'interprétation du mouvement humain, car les domaines de recherche concernés sont très vastes (segmentation, suivi, reconnaissance de forme, fusion de données etc.), nous commencerons par un état de l'art sur les domaines d'applications relatifs à ce champ de la vision par ordinateur, puis nous décrirons quelques systèmes existants, et finalement nous ferons une description générale de notre système ainsi que du contenu de ce mémoire.

1.1 État de l'art sur les domaines d'applications

Le domaine de la recherche concernant la vision par ordinateur d'êtres humains a reçu depuis quelque temps un intérêt croissant [Cedras95, Gavrila99, Aggarwal99, Pentland00, Moeslund01, Wang03a, Wang03b]. Le but général de l'analyse et de l'interprétation du mouvement humain en vision par ordinateur est la conception d'une machine capable d'interagir (détecter, suivre, reconnaître et réagir) de façon intelligente et rapide dans un environnement habité par des êtres humains. Les systèmes développés dans ce but sont aussi motivés par un grand nombre d'applications prometteuses, que l'on peut classer en quatre domaines (cf. table 1.1) :

Tab. 1.1 – Applications de l'analyse du mouvement humain en vision par ordinateur.

Domaine d'applications	Application spécifique
Vidéosurveillance :	- Contrôle d'accès - Parking, magasins, distributeurs de billets etc. - Personnes âgées - Manifestations - Transport en commun
Interfaces homme-machine :	- Interfaces sociales - Reconnaissance de gestes / Langage des signes - Contrôle guidé par les gestes
Réalité mixte :	- Animation d'avatar - Jeux vidéo interactifs - Vidéoconférence
Analyse du mouvement :	- Indexation et recherche basées sur le contenu - Entraînement personnalisé - Chorégraphie - Études cliniques - Compression vidéo très bas débit

1.1.1 Vidéosurveillance

De nombreux projets de recherche ont été menés à travers le monde. L'agence américaine pour les projets de recherche avancée de la défense, *DARPA (Defense Advanced Research Projects Agency)*, par exemple, a fondé un projet multi-institutionnel sur la vidéosurveillance et le contrôle, *VSAM (Video Surveillance And Monitoring)* [Collins00] (cf. partie 1.2.3). Le domaine de la vidéosurveillance concerne l'ensemble des applications où l'on cherche à suivre

et à contrôler au cours du temps les activités d'une ou de plusieurs personnes. Les systèmes concernés sont dits "intelligents" parce qu'ils cherchent à détecter, suivre et surveiller (au sens de l'activité effectuée) un ou plusieurs êtres humains en tant que tels [Haritaoglu98, Nair02, Siebel04]. Le besoin important de systèmes de vidéosurveillance avancés provient de l'existence même d'endroits nécessitant une sécurité par rapport aux biens ou aux personnes comme les banques, les magasins, les parkings, les frontières etc. Les sorties vidéo des caméras de surveillance dans la plupart de ces endroits sont souvent enregistrées sur bandes vidéo et archivées puis utilisées, si besoin est, "après coup", principalement comme outil d'identification. Le fait que les caméras soient des moyens de traitement temps-réel est donc en général inutilisé. Afin d'avertir aussi vite que possible les personnes ou services concernés (police, pompiers etc.) et faciliter / guider le travail du personnel de surveillance, il est donc nécessaire de développer des systèmes de traitement temps-réel de vidéosurveillance.

Dans le cas d'un contrôle d'accès à un site sensible, il est nécessaire, dans un premier temps, de pouvoir détecter la présence de quelqu'un. Puis, des données biométriques, comme les caractéristiques du visage ou la démarche, voire l'analyse rétinienne ou les empreintes digitales, peuvent être utilisées afin de confirmer l'accès. De nombreux systèmes de détection de personnes, de suivi temporel et de reconnaissance de visage ou de démarche ont été développés [Yang96, Moghaddam98, Cunado99, Peng05a]. Dans d'autres applications, c'est plus la description de l'activité de la personne que la description de la personne elle-même qui est source d'intérêt [Nair02, Siebel04]. Les exemples classiques étant la surveillance d'un parking, d'un supermarché ou d'un distributeur de billets où l'on cherche à évaluer la possibilité qu'un délit ou un crime (vol, agression, etc.) soit commis. Les bénéfices de telles applications de vidéosurveillance ont parfois besoin d'être équilibrés en regard des inconvénients possibles, le respect de la vie privée par exemple.

Certaines applications visent le bien-être de la personne. C'est le cas de la vidéosurveillance de personnes âgées dans les hôpitaux ou à domicile [Noury03]. Des personnes âgées peuvent se retrouver seules chez elles et avoir besoin d'une certaine sécurité pour pouvoir vivre dans de bonnes conditions. Que ce soit dans un hôpital ou dans une maison particulière, quelqu'un peut être averti si une personne âgée est restée trop longtemps dans une position fixe, ou si elle a chuté etc. [McKenna03, Nait-Charif04, Noury04, Barralon05].

D'autres applications concernent la vidéosurveillance lors de manifestations publiques, pour des dénombrements, ou lors d'événements sportifs, pour détecter les dégradations dues à l'emportement de certains supporters [Boghossian99a, Boghossian99b]. De même, certaines applications concernent la vidéosurveillance dans ou pour les transports en commun (avions, tramways, métro, bus etc.), c'est-à-dire des systèmes embarqués ou non [Thirde05, Harasse06].

1.1.2 Interfaces homme-machine avancées

Les interfaces homme-machine, IHM ou *HCI (Human-Computer Interfaces)*, visent à faciliter la communication entre deux utilisateurs par le biais d'une machine ou entre un utilisateur et la machine elle-même. Dans ce domaine, la vision par ordinateur, conjointement à la reconnaissance de parole et au langage naturel, permet d'accéder à des interfaces multimodales évoluées grâce à l'analyse et à la reconnaissance de visage, de postures, d'activités etc. [Li98, Aggarwal99, Wang03b].

Pour les interfaces sociales, il s'agit d'interagir avec un personnage généré par ordinateur qui tente d'avoir une attitude et un comportement humain. L'analyse du regard de la personne filmée, par exemple, donne accès à des zones d'intérêt comme une personne ou un objet présent

dans la scène. Une autre application est la reconnaissance de gestes avec, par exemple, la traduction du langage des signes pour les sourds et les mal-entendants [Starner95b, Cui96, Gianni03, Braffort04, Ong05]. Ces applications focalisent alors sur la partie du torse, des bras et / ou du visage qui sont les plus pertinentes et expressives.

Il existe aussi les applications qui permettent de contrôler par les gestes des objets graphiques [Viallet98, Tsekeridou01, Enterface06].

1.1.3 Réalité mixte

La réalité mixte est un concept qui concerne le mélange des mondes réel et virtuel. Le domaine englobe l'ensemble des applications où le mouvement humain est analysé puis reproduit ou symbolisé de façon à créer une réalité mixte ou même uniquement virtuelle [Freeman96, Crowley97, Bobick99].

L'animation d'avatar est l'action de reproduire un mouvement humain réel sur un personnage imaginaire. Dans le $7^{ème}$ art, les récents grands succès du box-office (La saga *Star Wars*, les trilogies *The Lord of the Rings*, *Matrix* etc.) dans les genres science-fiction ou médiéval-fantastique font une utilisation presque abusive d'effets spéciaux de ce type. Pour obtenir ce résultat, il faut utiliser la capture de mouvement (*Motion Capture*) en ayant recours à des caméras numériques et / ou à des capteurs [Moeslund01, Horain02]. Pour l'utilisation de caméras numériques, les acteurs ou cascadeurs sont d'abord habillés en tenue de couleur uniforme et très proche du corps. Ils sont ensuite filmés sur un fond de couleur unie, généralement bleu ou vert fluo, car ces couleurs sont loin des couleurs de peau humaine. La combinaison de la tenue, du fond et de l'utilisation de plusieurs caméras permet, d'une part, l'obtention d'une segmentation quasi parfaite du corps humain et, d'autre part, une très bonne reconstruction 3D du corps humain filmé. L'autre moyen d'obtenir ces résultats est l'utilisation de capteurs. Certains capteurs sont actifs (capteurs de position 3D appelés gravitomètres, capteurs magnétiques, capteurs à infrarouge etc.) et la plupart ne nécessitent pas l'utilisation de caméras numériques car les informations auxquelles ils donnent accès sont enregistrées directement de façon numérique. D'autres sont passifs (marqueurs de formes ou de couleurs variées) et permettent de faciliter le traitement de séquences vidéos acquises avec des caméras notamment lors de l'étape de reconnaissance. Les capteurs sont quelquefois considérés comme invasifs car ils peuvent gêner les mouvements de la personne qui les porte. Néanmoins ils permettent d'avoir accès de façon robuste (surtout avec des capteurs actifs) aux mouvements capturés en obtenant des positions et des trajectoires précises. Une fois les mouvements acquis, il est possible d'animer l'avatar en calquant ces mouvements sur le modèle du personnage. Le personnage Gollum, par exemple, dans *The Lord of the Rings*, est un avatar très réaliste. Les mouvements capturés et reproduits sont non seulement des déplacements du personnage dans sa globalité mais aussi des attitudes et des expressions faciales. La contrainte majeure dans ces applications est la qualité des résultats pour un meilleur effet visuel. Les moyens utilisés sont donc généralement très précis et coûteux.

En ce qui concerne les jeux vidéo interactifs, plusieurs systèmes ont été développés dans le but de pouvoir interagir avec un environnement virtuel [Darrell95, Maes97, Wren97a]. Un exemple d'application de réalité mixte où une partie du travail présenté dans ce mémoire a été utilisée est le projet *Art.live* [Art.live02] présenté en annexe A.1. Dans le même genre d'application, la société *Alterface* propose aux utilisateurs de s'immerger en temps-réel dans des atmosphères virtuelles [Alterface06]. L'espace d'interaction est entouré de grands écrans. Toute personne pénétrant dans la zone expérimentale est filmée par une ou plusieurs caméras

et son image est projetée sur de grands écrans dans des environnements virtuels, créés à partir d'images 3D, de photographies et de narrations graphiques. Dans un contexte de communication, *Alterface* offre ainsi une immersion en temps-réel dans des mondes imaginaires, unique en son genre. L'attention des spectateurs est fortement attirée par la présence de leur alter ego dans la scène.

Dans les jeux vidéos actuels plus classiques, nous voyons apparaître une nouvelle catégorie de jeux basés sur l'analyse du mouvement humain. Le système *Eye Toy*, utilisé avec la console *Playstation 2* de *Sony*, permet, par le biais d'une simple webcam, de jouer en déplaçant son corps et en bougeant son visage et ses mains, et non plus avec une manette de jeu classique.

D'autres applications ont pour but de réaliser des vidéoconférences virtuelles, de façon à humaniser les rapports d'entreprise et à rapprocher spatialement de façon virtuelle des personnes sur des sites distants. Ici, la contrainte principale est plus le temps-réel que la précision des résultats, même si cette dernière n'est pas négligée [Wingbermühle98, Weik00].

1.1.4 Analyse du mouvement

Cette partie traite des applications qui sont plus orientées vers la réalisation effective des gestes [Cedras95, Gavrila99, Wang03b].

L'indexation et la recherche de vidéos de sport basées sur le contenu est une application possible. Dans un contexte de football, quelqu'un pourrait par exemple interroger une base de données de matches pour obtenir tous les passages où un joueur marque en lobant le gardien. Cela permettrait d'éviter à un opérateur humain de parcourir l'ensemble de la base de données.

D'autres applications concernent les systèmes d'entraînement sportif personnalisé pour la pratique de différents sports. Ces systèmes peuvent observer la réalisation technique d'un saut en hauteur lors de vues acquises avec des caméras mobiles afin de reconstituer des panoramas [Dalal02, Bartoli02, Bartoli04, Rodriguez05], ou analyser des lancers de ballon de basket-ball ou des swings de golf afin de suggérer des améliorations techniques [Lepetit03, Gehrig03]. La vision par ordinateur de l'humain est aussi utilisée dans les chorégraphies de danse et de ballet [Kojima01].

Un grand ensemble d'applications concerne l'analyse de la démarche, la reconnaissance de problèmes orthopédiques et l'analyse de parties du corps (comme le cerveau par exemple) vues selon une approche médicale et permettant l'interprétation clinique et la formulation d'un diagnostic [Köhle97, Meyer97, Capelle04].

Une autre application possible est la compression vidéo, qui considère les êtres humains dans les séquences vidéo comme des zones d'intérêt ayant certaines propriétés [Aizawa95]. C'est le cas de la norme *MPEG-4* qui se sert de connaissances *a priori* sur le corps humain pour coder plus efficacement les séquences vidéos où sont majoritairement présents des êtres humains. La norme *MPEG-4* contient des modes de codage dits *SNHC (Synthetic / Natural Hybrid Coding)*. Cette norme d'animation du visage et du corps utilise une approche appelée *FBA (Face and Body Animation)* et standardise deux types de flux pour animer un avatar. Le premier est le flux de définition qui est constitué de paramètres de définition du corps, *BDP (Body Definition Parameters)*, et de paramètres de définition du visage, *FDP (Face Definition Parameters)*. Ce flux respecte les spécifications de la norme *H-ANIM* (plus précisément, le nœud *BDP* de *MPEG-4* contient le nœud *Humanoid* de *H-ANIM*). Le second flux est spécifique à la norme *MPEG-4* et traite des paramètres de mouvement. Ainsi, afin d'animer un modèle de visage, il est nécessaire d'utiliser des paramètres d'animation du visage, *FAP (Face*

Animation Parameters), et pour animer un modèle de corps, des paramètres d'animation du corps, *BAP (Body Animation Parameters)*. Par utilisation de cette norme, l'animation d'avatar et le codage très bas-débit de séquences vidéo sont possibles [Capin00a, Capin00b].

1.2 Exemple de systèmes et de leur applications

Nous allons maintenant présenter trois systèmes qui ont des applications principalement en vidéosurveillance et en réalité mixte et qui sont très fréquemment cités dans la littérature. Ces trois systèmes sont *Pfinder*, W^4, et le système issu du projet *DARPA VSAM*.

1.2.1 *Pfinder*

1.2.1.1 Présentation du système et applications

Le système *Pfinder (Person finder)*, développé par Wren, Azarbayejani, Darrell et Pentland, est un système temps-réel pour effectuer le suivi temporel et l'interprétation du mouvement d'une personne [Wren97b]. Il permet une interaction performante grâce à du matériel informatique classique, a été testé sur des milliers de personnes dans des environnements et des conditions diverses dans le monde entier et a obtenu des résultats satisfaisants. *Pfinder* a été utilisé pour explorer plusieurs applications différentes d'interface homme-machine, principalement comme un outil d'interaction temps-réel dans des espaces de réalité mixte, avec notamment les applications découlant du système *Artificial Live IVE* (*ALIVE*) qui illustre les interactions avec une forme de vie artificielle dans un environnement virtuel 3D qui peut être contrôlé et à travers lequel la navigation est possible grâce aux gestes et à la position de l'utilisateur [Darrell95, Maes97]. L'animation d'avatar est aussi possible grâce aux positions du visage, des mains, des pieds et du torse [Darrell95]. Un autre exemple est un jeu vidéo interactif avec navigation dans un environnement 3D, *Simulated Urban Recreational Violence IVE* (*SURVIVE*) [Wren97a]. *Pfinder* a aussi été utilisé par Starner et Pentland, comme système préprocesseur pour la reconnaissance de gestes, notamment celle d'un sous-ensemble de quarante mots du langage des signes américain avec une précision quasi parfaite de 99% [Starner95b].

1.2.1.2 Étapes de traitement du système

Ce système approche les problèmes de la détection, du suivi temporel et de l'interprétation du mouvement humain avec les deux hypothèses que sont un environnement de faible dynamique (plutôt d'intérieur) et une seule personne filmée par une caméra fixe. Sa caractéristique principale est l'utilisation de *blobs* 2D basés sur de multiples classes statistiques de couleur et de forme. Les caractéristiques traitées sont les composantes spatiales, la position (x, y), et les composantes de couleur (Y, U, V). À cause de leurs différences sémantiques, ces deux types de composantes sont supposées être indépendantes.

Après avoir modélisé de façon statistique le fond fixe de la scène suivant une approche de surface texturée de couleurs, une personne seule est détectée comme un ensemble de *blobs* ayant des caractéristiques homogènes de couleur et de forme. Ces *blobs* permettent de détecter les parties du corps de la personne, notamment le visage et les mains, après une étape d'initialisation en "étoile" (debout, face à la caméra et bras étendus horizontalement). Une étape de prédiction basée sur un filtrage de Kalman et des mesures de similarité suivant un algorithme

MAP (Maximum A Posteriori probability), basé sur un critère de log-vraisemblance, permettent le suivi des *blobs* au cours du temps. De façon à obtenir une meilleure segmentation, les problèmes d'ombres projetées de la personne sont traitées en normalisant les composantes de chrominance U et V par la composante de luminance Y ($U^* = U/Y$ et $V^* = V/Y$). Lorsque surviennent des problèmes d'occultation, les *blobs* correspondants sont effacés jusqu'à leur réapparition où ils sont à nouveau détectés.

1.2.1.3 Performances / limitations du système

Ce système atteint une cadence de traitement d'environ 10 images/s sur une station de travail de 200 MHz, pour une résolution d'image de 160×120. La limitation principale de ce système est qu'une seule personne doit être présente dans le champ de la caméra, sous peine de dégrader les résultats d'interprétation de comportement et / ou de reconnaissance de gestes. De plus, en cas d'occultation du visage ou des mains, aucune estimation de position n'est disponible puisque les *blobs* correspondants ont été effacés.

1.2.2 W^4

1.2.2.1 Présentation du système

W^4 : *Who ? When ? Where ? What ?* est un système temps-réel de vidéosurveillance, proche de la cadence vidéo, et développé par Haritaoglu, Harwood et Davis pour la détection, le suivi temporel de personnes et la surveillance de leurs activités dans un environnement extérieur [Haritaoglu98]. Ce système permet la détection et le suivi temporel de plusieurs personnes, de même que des parties de leur corps (visage, mains, pieds et torse). Il a été développé pour la reconnaissance d'actions telles que les interactions entre personnes, la prise ou le dépôt d'objets dans une scène.

1.2.2.2 Étapes de traitement du système

W^4 considère des séquences obtenues en environnement extérieur, filmées par une caméra fixe et où peuvent interagir plusieurs personnes. Il ne traite que des séquences vidéo en niveaux de gris ou issues d'une caméra infrarouge. Contrairement à beaucoup de systèmes en analyse de l'humain, ce système n'utilise donc pas d'indices de couleur pour parvenir à ses fins. En lieu et place des informations de couleur, W^4 utilise une combinaison d'analyse de forme et de suivi temporel afin de localiser des personnes et les parties de leurs corps (visage, mains, pieds et torse). Ce système utilise des modèles dynamiques de mouvement et des modèles d'apparence de telle façon que les personnes puissent être suivies même en cas d'occultation.

Chaque pixel du fond de la scène est d'abord représenté au bout d'une période d'entraînement par trois valeurs, les valeurs minimale et maximale d'intensité pendant la période et la plus grande différence d'intensité entre deux images consécutives de la période. Les personnes en mouvement sont ensuite extraites de la scène par un algorithme de soustraction du fond. Puis un suivi temporel de personnes est effectué grâce à des tests de superposition entre les prédictions des positions des silhouettes des personnes suivies et celles détectées, avec un raffinement des associations suivant une estimation récursive par moindres carrés entre les contours des silhouettes traitées. Les prédictions des positions sont obtenues avec l'utilisation d'un modèle de mouvement du second ordre. Le suivi temporel des parties du corps est réalisé grâce à un modèle 2D de personne en position debout, de type *Cardboard Model*, constitué

d'un ensemble de rectangles, et d'un modèle dynamique appelé gabarit temporel de texture (*temporal texture template*). Le système prédit les positions des parties du corps, associe les parties du corps d'une image à l'image suivante et met à jour les différents modèles utilisés. Il a été testé lors d'interactions telles qu'une rencontre entre deux personnes, une personne qui s'assied etc.

1.2.2.3 Performances / limitations du système

Ce système permet une cadence de traitement d'environ 20 images/s sur une station de travail de 200 MHz, pour une résolution d'image de 320×240. L'une de ses limitations est qu'il n'exploite pas, par choix, les informations de couleur. Une autre limitation est que le modèle 2D de personne utilisé restreint l'utilisation du système à des interactions entre personnes proches de la position verticale.

1.2.3 Système issu du projet *DARPA VSAM*

1.2.3.1 Présentation du système

L'agence américaine *DARPA (Defense Advanced Research Projects Agency)* et son projet *VSAM (Video Surveillance And Monitoring)* a donné lieu à de nombreux travaux. Le but de ce projet est le développement d'une technologie de compréhension automatique de séquences vidéo qui permet à un unique opérateur humain de contrôler un ensemble d'activités dans des endroits plus ou moins complexes comme des champs de bataille ou des scènes civiles. Collins, Lipton et Kanade ont développé un système de vidéosurveillance dans le cadre de ce projet [Collins00]. Utilisant de multiples caméras, le système est capable de détecter, de classer et de suivre au cours du temps des personnes et / ou des véhicules. Il permet en particulier de déterminer la démarche et la posture d'une personne, en classant son mouvement entre la marche et la course. Ce système gère aussi les caméras multiples et leur coopération afin d'améliorer les résultats de suivi temporel et d'interprétation. Il peut déterminer la localisation 3D d'un objet dans la scène par coopération entre plusieurs vues de caméras calibrées et un modèle 3D du terrain filmé. Finalement, l'opérateur humain peut interagir avec le système, par exemple en spécifiant des régions d'intérêt et en déclenchant des alarmes lors d'évènements spécifiques.

1.2.3.2 Étapes de traitement du système

Ce système traite des séquences acquises en environnement extérieur, principalement urbain, où se déplacent des véhicules et des personnes filmés par un ensemble de caméras, fixes ou mobiles.

La première étape consiste à détecter les objets en mouvement dans la scène filmée. Cette étape d'extraction des objets en mouvement est effectuée par un algorithme de soustraction du fond de la scène, robuste aux changements dynamiques des conditions d'acquisition. Les objets extraits sont assimilés à des *blobs* dont les caractéristiques principales sont : la position et la vitesse du centre de gravité, l'apparence en tant que gabarit, la taille et l'histogramme de couleur.

La deuxième étape consiste à suivre les objets au cours du temps. Le suivi temporel des objets est réalisé entre deux images consécutives en combinant une approche d'estimation de mouvement et de gabarits. L'estimation de mouvement prédit la position d'un *blob* en fonction

des dernières position et vitesse connues. L'algorithme de suivi temporel complet combine cette estimation avec une fonction de coût d'association entre gabarits, qui détermine un coût d'association entre le gabarit d'un objet cible et celui d'un *blob* prédit [Lipton98].

La troisième étape consiste à classer les objets selon leur type. Une fois correctement suivis, ils sont ensuite classés en catégories grâce à un réseau de neurones, *NN (Neural Network)*, préalablement entraîné. Plusieurs catégories d'objets sont définies suivant des critères de forme et de couleur : personne seule, groupe de personnes, véhicule. Les caractéristiques utilisées par le réseau de neurones pour cette classification sont la disparité, la taille de l'image et le grossissement de la caméra (zoom).

Pour une personne seule, une fois détectée, suivie au cours du temps et reconnue en tant que personne unique, la dernière étape effectuée par le système est l'analyse de la démarche de cette personne. Dans ce système, les personnes filmées ne sont pas prédominantes dans l'image. Un squelette de personne obtenu par squelettisation en "étoile" ainsi que deux angles sont utilisés pour reconnaître la démarche d'une personne et la classer entre la marche et la course. Le squelette en "étoile" consiste en un ensemble de segments partant du centre de gravité de la personne et allant jusqu'aux points extrêmes de la silhouette. Les angles utilisés sont l'angle entre la verticale et l'axe du torse et l'angle entre la verticale et l'axe de la jambe du squelette la plus à gauche dans l'image. Comme une personne qui court est normalement plus penchée en avant qu'une personne qui marche, et que la fréquence du mouvement cyclique de ses jambes est plus rapide, le système peut ainsi déterminer la démarche d'un individu.

1.2.3.3 Performances / limitations du système

Ce système permet une cadence de traitement d'environ 10 images/s sur une station de travail *Pentium II* de 450 MHz, pour une résolution d'image de 320×240. La principale limitation du système, concernant l'analyse de l'humain, est due à la nature même du système qui ne se focalise pas uniquement sur les personnes mais aussi sur les véhicules et utilise des vues issues de caméras multiples assez distantes des régions d'intérêt, *ROI (Regions Of Interest)*. La reconnaissance d'activités sur les personnes est donc limitée, puisque les personnes ne sont pas les seules *ROI* du système. Elle comprend les entrées / sorties de bâtiments ou de véhicules et la reconnaissance de démarche, avec la distinction marche / course.

1.3 Description générale

Dans cette partie, nous commencerons par décrire comment le travail décrit dans ce mémoire s'est inséré dans le contexte scientifique. Puis nous ferons une présentation générale du système avec les différentes étapes de traitement et les hypothèses de développement. Dans une dernière partie, nous donnerons le plan du mémoire.

1.3.1 Insertion de ce travail dans le contexte scientifique

Ce travail de thèse, réalisé au Laboratoire des Images et des Signaux (LIS), s'est inscrit à différents niveaux dans la communauté de chercheurs travaillant sur l'analyse et l'interprétation du mouvement humain en vision par ordinateur.

Au niveau national, ce thème de recherche est à la croisée de trois Réseaux Thématiques Pluridisciplinaires (RTP) du département STIC (Sciences et Technologies de l'Information et de la Communication) du CNRS :

– le RTP 15 : Interfaces Médiatisées et Réalité Virtuelle ;
– le RTP 16 : Méthodes et Outils pour l'Interaction Homme-Machine ;
– le RTP 25 : Imagerie, Vision et Analyse de Scènes.

Une partie de ce travail a été présentée lors de l'Action Spécifique GESTE du RTP 25 : "Perception, Modélisation et Interprétation du Geste Humain", animée par M. Michel Dhome les 27 et 28 Mars 2003 à Grenoble.

Ces travaux se sont aussi inscrits dans le cadre de la Réunion Française sur les Fonctions de Croyance, animée par M. Thierry Denœux et M. Philippe Smets, les 23 et 24 Mars 2005 à Compiègne.

Au niveau européen, ce travail a commencé avec le projet *Art.live*, présenté en annexe A.1. Ce projet s'est terminé en 2002 [Art.live02]. Depuis, le LIS participe activement à la recherche sur les interfaces homme-machine multimodales à travers le réseau d'excellence européen Similar [Similar05]. Le but de ce réseau est de créer des IHM qui réagissent de façon similaire aux communications homme-homme, donc basées sur des techniques de communication telles que la parole, l'attitude corporelle, le regard, les expressions du visage etc. Les IHM avancées sont composées de plus en plus de modalités différentes (gestes, parole, expressions, posture etc.).

Ces travaux s'inscrivent dans la thématique Gestes du groupe GOTA du LIS et du futur groupe GPIG du laboratoire GIPSA. En effet, au LIS sont menées des activités de recherche portant sur l'analyse et l'interprétation du mouvement humain à plusieurs niveaux :

– Analyse du mouvement et reconnaissance de postures de personnes d'où des méthodes de détection et de suivi temporel de personnes afin de faire l'analyse de gestes ou de comportement précis (postures etc.) [Girondel06].
– Analyse des expressions faciales par extraction et analyse de l'évolution temporelle des contours des traits principaux du visage (bouche, yeux, sourcils) dans le but de mettre en place un système automatique de reconnaissance d'émotions sur un visage [Hammal05].
– Analyse des mouvements de la tête et des mains avec comme objectif l'interprétation de gestes de communication non verbale (direction du regard, hochements de tête etc.) [Benoit05].
– Reconnaissance de gestes (lecture labiale et gestes de la main) pour le langage parlé complété (LPC) destiné aux sourds et aux mal-entendants [Burger06].

Le LIS dispose d'une plate-forme interactive AIM (Analyse Interprétation Multimodalités) pour mener à bien ces activités de recherche. La plate-forme AIM est décrite en annexe C.

1.3.2 Présentation du système

Notre système a été développé afin de réaliser une analyse et une interprétation du mouvement humain dans les séquences vidéo. Par rapport aux systèmes présentés précédemment, les buts poursuivis sont multiples. D'un côté, le système doit être capable de réaliser une **analyse** du mouvement humain. Le terme "analyse" concerne ici l'extraction d'informations qui seront appelées par la suite données bas-niveau. Ces données bas-niveau peuvent être, par exemple, la silhouette de la personne, la localisation de son visage ou le fait de réussir à la suivre au cours du temps. D'un autre côté, le système doit aussi permettre de réaliser une **interprétation** du mouvement humain. Quand on parle d'interprétation du mouvement ou du comportement humain, le champ de recherches est très vaste, il peut s'agir de la reconnaissance de démarche (marche, course etc.), de postures (debout, accroupi, etc.), d'interactions avec des objets (poser, prendre etc.) ou entre des personnes (gestes, attitudes etc.). Dans

notre cas, nous nous intéressons à la **reconnaissance de postures**. Cette interprétation haut-niveau se base sur la fusion des données bas-niveau. Ces dernières doivent donc être obtenues de manière **précise et robuste**. De plus, par rapport aux domaines d'applications visés, qui sont la réalité mixte et la vidéosurveillance, le système doit aussi être relativement **rapide**, assez proche de la cadence vidéo (minimum de 10 images/s).

Pour résumer, notre système peut être divisé en deux parties :

1. l'**analyse** : extraction de données (bas-niveau) ;
2. l'**interprétation** : fusion des données (haut-niveau).

1.3.2.1 Étapes de traitement du système

Notre système comporte cinq étapes de traitement :

1. segmentation 2D spatio-temporelle ;
2. première étape du suivi temporel ;
3. localisation et suivi temporel du visage et des mains ;
4. seconde étape du suivi temporel ;
5. reconnaissance de postures statiques.

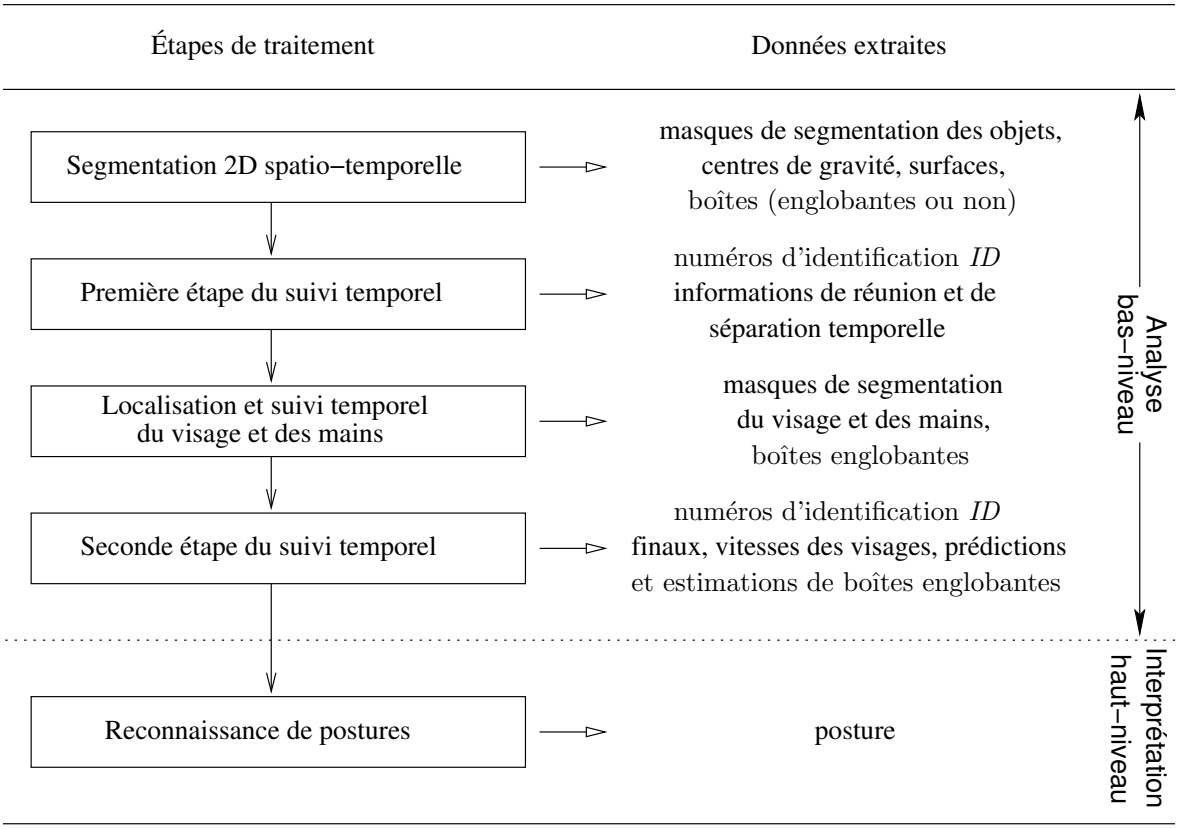
Les quatre premières étapes de traitement concernent la phase d'**analyse**. L'étape de segmentation 2D spatio-temporelle permet d'extraire les personnes de la scène filmée et des informations concernant ces personnes "segmentées" vues en tant qu'objets caractérisés par une forme spécifique. Ensuite, la première étape du suivi temporel crée des liens temporels entre deux images consécutives pour les personnes segmentées. Puis, l'étape de localisation et de suivi temporel du visage et des mains donne accès à des zones d'intérêt plus précises du corps humain. Finalement, la seconde étape du suivi temporel autorise un suivi plus complexe qui gère les occultations partielles ou complètes entre les personnes. La dernière étape de traitement concerne la phase d'**interprétation**. Grâce à la fusion de certaines données bas-niveau, l'étape d'interprétation haut-niveau présentée dans ce mémoire est la **reconnaissance de postures statiques** ("debout", "assis", "accroupi" ou "couché").

La table 1.2 présente cette description générale en grandes étapes du système avec la séparation des phases d'analyse et d'interprétation. À droite sont visibles les différentes étapes de traitement et à gauche les données bas-niveau extraites lors de ces étapes. La figure 1.1 illustre chacune des différentes étapes de traitement par des images issues d'une séquence vidéo particulière.

Les sigles utilisés dans la figure 1.1 sont les suivants :
- BAPS : Boîte par Axes Principaux issue de la Segmentation.
- BERS : Boîte Englobante Rectangulaire issue de la Segmentation.
- BERV : Boîte Englobante Rectangulaire du Visage.
- BEREP : Boîte Englobante Rectangulaire Estimée de la Personne.
- BEREV : Boîte Englobante Rectangulaire Estimée du Visage.
- BERPP : Boîte Englobante Rectangulaire Prédite de la Personne.
- BERPV : Boîte Englobante Rectangulaire Prédite du Visage.

Des précisions quant à chacune de ces boîtes seront apportées dans les chapitres suivants.

TAB. 1.2 – Description générale du système.



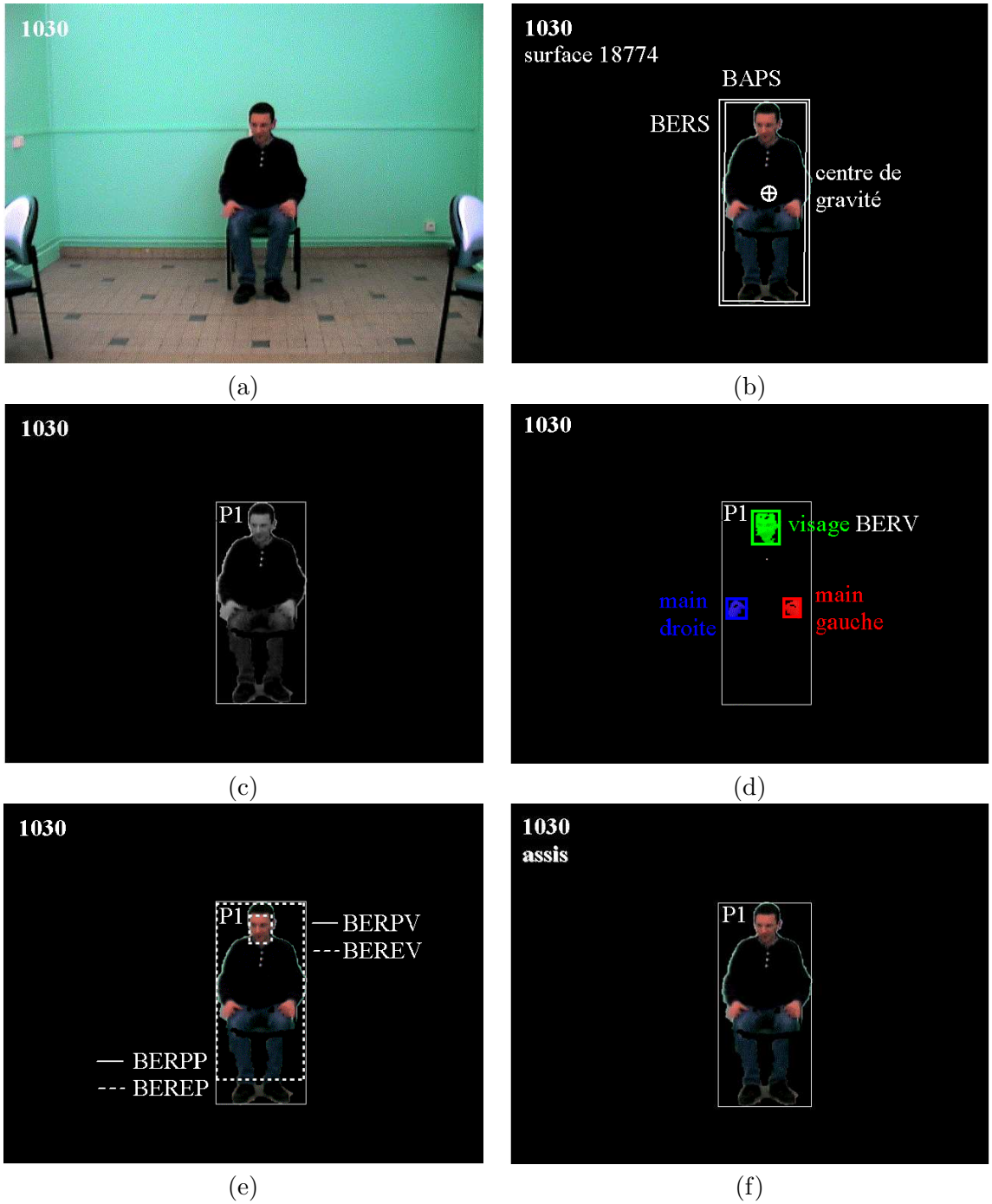


FIG. 1.1 – Exemple d'étapes de traitement. (a) image originale, (b) segmentation 2D spatio-temporelle de personnes, (c) suivi temporel (1/2), (d) localisation et suivi temporel du visage et des mains, (e) suivi temporel (2/2) et (f) reconnaissance de postures statiques.

1.3.2.2 Hypothèses du système

Comme une très grande majorité de systèmes visant l'analyse et l'interprétation du mouvement humain pour diverses applications, notre système suppose un ensemble d'hypothèses qui proviennent des applications visées et / ou de choix permettant de simplifier l'approche d'un problème donné qui peut être mal posé. Comme le traitement doit être le plus proche possible du temps-réel (25 images/s), les algorithmes développés doivent être rapides, simples mais précis et robustes. Les hypothèses du système peuvent être classées suivant différents critères : indispensables *vs* optionnelles et fortes *vs* faibles, c'est-à-dire contraignantes *vs* peu contraignantes.

Au niveau des conditions d'acquisition, les hypothèses considérées comme indispensables sont :

1 Environnement filmé par une **caméra fixe**.

2 Chaque personne **entre seule** dans la scène.

Principalement afin de faciliter l'étape de segmentation 2D spatio-temporelle, nous avons ajouté des hypothèses optionnelles, peu contraignantes par rapport aux applications visées (réalité mixte, vidéosurveillance en intérieur), qui sont :

3 L'environnement est **intérieur**.

4 Chaque séquence vidéo commence par une **scène vide**.

Nous avons aussi ajouté des hypothèses qui sont optionnelles pour les quatre premières étapes de traitement, mais indispensables pour la dernière étape de reconnaissance de postures statiques. Dans la conclusion de ce mémoire, nous proposons des solutions qui permettraient de lever ces hypothèses. Ces hypothèses sont, de la moins contraignante à la plus contraignante :

5 Chaque personne doit être au moins une fois dans une **posture de référence**, debout avec les bras étendus horizontalement. Cette posture est celle effectuée par l'Homme de Vitruve dans le dessin de Léonard De Vinci présenté figure 1.2, c'est-à-dire debout, les bras écartés.

6 Chaque personne est supposée être filmée **entièrement**, c'est-à-dire qu'elle doit rester dans le champ de la caméra et ne pas être occultée par des objets fixes. Elle peut cependant être occultée partiellement ou complètement par une autre personne.

7 Chaque personne est supposée rester à une **distance à peu près constante** de la caméra.

1.3.3 Plan du mémoire

Le corps de ce mémoire va détailler chaque étape de la chaîne de traitement de notre système. Les quatre premiers chapitres concernent l'analyse et l'extraction de données (bas-niveau) :

- chapitre 2 : segmentation 2D spatio-temporelle ;
- chapitre 3 : première étape du suivi temporel ;
- chapitre 4 : localisation et suivi temporel du visage et des mains ;
- chapitre 5 : seconde étape du suivi temporel.

Le dernier chapitre concerne l'interprétation et la fusion de données (haut-niveau) :

- chapitre 6 : fusion de données et reconnaissance de postures statiques.

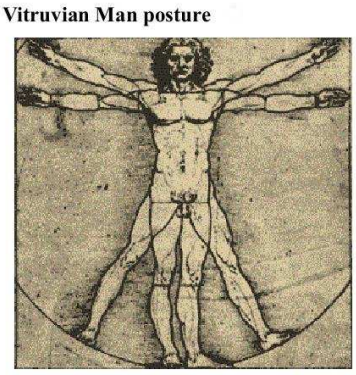


FIG. 1.2 – L'Homme de Vitruve de Léonard de Vinci.

Comme tous les chapitres décrivent chacun une étape de traitement de notre système, ils sont présentés selon une structure commune. Après une introduction, chaque chapitre comporte un état de l'art décrivant les principales approches utilisées pour réaliser l'étape de traitement correspondante. Après cet état de l'art, chaque chapitre présente la méthode ou l'approche choisie pour atteindre le but de l'étape de traitement. Puis nous détaillons, s'il y a lieu de le faire, les données bas-niveau extraites. Nous illustrons ensuite les résultats obtenus grâce à des images issues de séquences vidéo diverses. Finalement, nous donnons les avantages, les limitations et les cadences de traitement atteintes pour l'étape considérée avant de conclure avec les perspectives d'amélioration ou de développement possibles.

Concernant les états de l'art des différents chapitres, comme le domaine de recherche est très vaste quelle que soit l'étape de traitement considérée, les états de l'art sont généralement guidés par les contraintes des applications choisies et par les résultats de l'étape de traitement précédente. Les articles suivants présentent des états de l'art génériques pour les différentes étapes d'analyse et d'interprétation du mouvement humain, classés selon différents critères [Cedras95, Gavrila99, Aggarwal99, Pentland00, Moeslund01, Wang03a, Wang03b]. Des sites Internet donnent aussi accès à beaucoup plus de références bibliographiques, en particulier le site de M. Keith E. Price, régulièrement actualisé et très bien conçu pour faire des recherches selon mots-clés, auteurs, conférences ou journaux [Price06].

En ce qui concerne les différentes figures et illustrations présentant les résultats obtenus, il est conseillé de regarder ces images en couleur, sous peine de perdre une partie des informations qu'elles contiennent. Une version électronique couleur est actuellement téléchargeable à l'adresse suivante :

`http ://www.lis.inpg.fr/pages_perso/vgironde/upload/These.pdf`.

Chapitre 2

Segmentation 2D spatio-temporelle

Sommaire

La segmentation spatio-temporelle concerne l'extraction d'objets ou régions d'intérêt *ROI (Regions Of Interest)* en mouvement dans les séquences vidéo. La segmentation d'objets en mouvement dans les séquences vidéo est une étape de traitement très importante pour les étapes ultérieures. En effet, elle permet, d'une part, de réduire la quantité d'information aux *ROI*. D'autre part, quand on ne s'intéresse qu'aux objets segmentés, plus les résultats de la segmentation sont précis, plus les étapes de traitement ultérieures sont facilitées. Ce problème important en traitement d'images a été beaucoup étudié par la communauté [Mitiche96, Koprinska01, Zhang01, Lefèvre03]. Même si de très nombreuses méthodes ont été développées, les outils permettant une segmentation complète et automatique de façon générale ne sont pas encore disponibles. En fonction du système et des conditions d'acquisition, certaines approches sont préférées. De plus, des hypothèses sont généralement posées pour simplifier les approches. Les connaissances *a priori* sur les séquences vidéo traitées et sur les applications visées conditionnent donc la manière dont est réalisée la segmentation. Dans notre cas, nous nous intéressons à des personnes en mouvement dans une scène filmée par une caméra fixe, nous utiliserons donc une **segmentation 2D spatio-temporelle**.

Nous commencerons par présenter dans ce chapitre un état de l'art sur les approches utilisées pour effectuer une segmentation spatio-temporelle. Vu le grand nombre de méthodes et d'approches développées, cet état de l'art sera restreint et guidé par les applications de notre système. Nous détaillerons ensuite l'une des deux méthodes disponibles dans notre système pour réaliser une segmentation 2D spatio-temporelle d'objets en mouvement, présents dans une scène filmée par une caméra fixe. Puis nous exposerons les données bas-niveau extraites lors de cette première étape de traitement. Après cela nous illustrerons les résultats obtenus avec ces méthodes par des images issues du traitement de séquences vidéo variées. Dans une dernière partie, nous préciserons les avantages et les inconvénients de ces méthodes ainsi que les cadences de traitement atteintes avant de conclure sur cette étape de traitement en présentant les perspectives d'amélioration possibles.

2.1 État de l'art sur la segmentation spatio-temporelle

De très nombreux travaux existent sur ce sujet qui est un problème difficile car mal posé [Mitiche96, Koprinska01, Zhang01, Lefèvre03]. La segmentation spatio-temporelle combine deux sortes d'informations, qui sont différentes et complémentaires : des informations **temporelles** et des informations **spatiales**.

2.1.1 Informations temporelles

L'utilisation d'informations temporelles se justifie par l'hypothèse d'un fond fixe (c'est-à-dire un fond filmé par une caméra fixe). Ainsi, tout objet présent dans la scène et absent du fond est vu comme un objet en mouvement. Ceci se traduit par des variations temporelles de la fonction de luminance qu'il faut détecter.

Les approches basées sur la soustraction d'images pour détecter des objets en mouvement sont très utilisées [Nagel78, Yalamanchili82]. Elles consistent à calculer pixel à pixel les différences entre l'image courante et une autre image. Cette autre image peut être l'image **précédente** ou une image **de référence**.

Avec l'utilisation de l'image précédente, les différences inter-images sont calculées selon l'intensité [Polana94] ou les gradients [Amat99, Sangi04] des pixels. Une version améliorée consiste à utiliser trois images consécutives au lieu de deux [Kameda96]. Le résultat reflète

alors les mouvements (et, dans une moindre mesure, le bruit) entre les images à moins que l'objet ait la même intensité ou la même couleur que le fond. Un défaut majeur de ce choix est qu'aucun changement temporel significatif n'intervient dans la zone de glissement d'un objet sur lui-même si celui-ci n'est pas assez texturé. Un autre inconvénient est qu'un objet cesse d'être détecté s'il s'arrête.

L'autre approche utilise la différence entre l'image courante et une image de référence du fond fixe [Long90, Cavallaro01, Seki03, Lee05]. L'image de référence est une image du fond non bruitée et vide de tout objet ou personne en mouvement [Nakazawa98]. Ce choix permet de détecter l'objet complet, même s'il est peu texturé ou devient immobile. Néanmoins, la limitation principale de cette approche est le fait que, si les conditions d'acquisition (illumination de la scène etc.) varient grandement, tout ou une partie du fond peut être détecté si l'image de référence n'est pas actualisée. Une version plus avancée de cette approche consiste donc à mettre à jour l'image de référence au cours du temps [Haritaoglu98]. Le problème qui demeure est alors la construction de l'image de référence qui nécessite de nombreuses images et introduit un délai important avant l'obtention d'une segmentation correcte.

Les informations temporelles sont, dans bien des cas, une alternative robuste aux informations spatiales, car elles sont relativement aisées à extraire et se concentrent directement sur les *ROI* en mouvement.

2.1.2 Informations spatiales

L'utilisation d'informations spatiales se justifie par le fait que l'apparence des objets ou des personnes en mouvement diffère de la scène filmée (environnement). L'apparence peut par exemple différer de l'environnement par une uniformité de la couleur des vêtements ou de celle du fond. Elle peut aussi différer parce que l'objet comporte des capteurs actifs ou passifs (marqueurs). Mais la simple présence d'un objet dans la scène permet de le détecter grâce au fait qu'il cache une partie du fond par sa silhouette. La frontière visuelle entre l'objet et le fond peut être détectée par des méthodes adaptées. Les approches peuvent donc être séparées en deux types :

- les approches basées sur le **seuillage** ;
- les approches basées sur le **contour** ou la **silhouette**.

2.1.2.1 Approches basées sur le seuillage

Ces approches réalisent soit un seuillage direct de caractéristiques (couleur, température corporelle, position etc.) de l'objet, soit un seuillage basé sur leurs statistiques. Les secondes sont généralement considérées comme plus robustes que les premières.

Pour les approches basées sur un seuillage direct, un bon exemple est le cas où une personne apparaît devant un fond ou un écran de couleur uniforme, généralement bleu ou vert-fluo, et porte des vêtements de couleur différente. Un simple seuillage sur la couleur permet de segmenter facilement la personne du fond [Darrell94, Iwai99]. Cette technique est utilisée couramment pour les informations météo où le présentateur est segmenté puis incrusté devant des cartes météorologiques. L'approche inverse, où c'est la personne, cette fois, qui porte des vêtements de couleur uniforme et apparaît devant un fond de couleurs plus variées, est aussi possible [Bharatkumar94]. Une approche similaire est l'utilisation de caméra(s) infrarouge *IR (Infra Red)*. Les images thermiques obtenues permettent de segmenter une personne facilement par seuillage, comme étant le seul objet chaud dans la scène [Iwasawa97]. Outre la couleur de

ses vêtements ou sa température corporelle, l'apparence d'un sujet peut aussi être différente de l'environnement parce que sa tenue comporte des marqueurs passifs (lumineux ou colorés) qui seront aisément segmentés par seuillage, ou des capteurs actifs (gravitomètres) [Campbell95b, Goncalves98]. L'un des premiers systèmes d'analyse du mouvement humain est d'ailleurs basé sur un ensemble de marqueurs lumineux en mouvement, ce sont les travaux de Johansson en 1976 sur les *MLD (Moving Light Display)* [Johansson76].

Les approches avancées basées sur le seuillage utilisent les caractéristiques statistiques de pixels ou de groupes de pixels pour extraire la *ROI* du fond [Aach93]. Par exemple, une séquence d'images du fond est acquise et la moyenne et la variance de l'intensité et / ou de la couleur de chaque pixel sont calculées. Dans l'image courante, chaque pixel est comparé à ces statistiques et classé comme appartenant au fond ou non [Yamada98]. Une version améliorée est utilisée dans les approches avec *blobs* (groupe de pixels possédant certaines caractéristiques homogènes), où le sujet est modélisé par un ensemble de *blobs* avec des statistiques individuelles de position spatiale et de couleur. Chaque pixel de l'image courante est alors classé comme appartenant à l'un des *blobs* selon ses caractéristiques statistiques [Wren97b]. McKenna, Jubri, Duric et Wechsler combinent une approche par seuillage avec les statistiques des gradients des pixels pour éliminer les ombres projetées des sujets [McKenna00a].

La principale difficulté de ces approches basées sur le seuillage est le choix du seuil. Selon les valeurs utilisées, il peut amener des résultats de plus ou moins bonne qualité. Certaines conditions d'acquisition ou hypothèses peuvent guider sa détermination mais, de manière générale, cette dernière est souvent difficile.

2.1.2.2 Approches basées sur le contour ou la silhouette

La distinction entre le contour et la silhouette est la suivante : le **contour** est l'ensemble des pixels formant la frontière entre l'objet et le fond, la **silhouette** est l'ensemble des pixels de l'objet à l'intérieur du contour.

Les trois approches utilisées pour extraire les contours sont les méthodes différentielles (gradient, laplacien etc.), les méthodes par *template* (Roberts, Prewitt, Sobel, Kirsch etc.) et les méthodes par optimisation basées sur des modèles de contour, de bruit, de mesure de qualité de détection etc. (Marr-Hildreth, Canny etc.).

Les contours peuvent être **statiques** ou **dynamiques**. Les contours statiques reposent sur des structures rigides prédéfinies, segments, rectangles, quadrilatères, ellipses etc., qui vont schématiser en les approchant au mieux les frontières visuelles de l'objet dans l'image. Long et Yang [Long91] utilisent des quadrilatères (nommés *Logs*) pour les différentes parties du corps humain. Les *Logs* extraits de l'image sont comparés à un modèle du corps humain, lui aussi en *Logs*. Les contours dynamiques, appelés *snakes*, sont des contours actifs non rigides qui vont s'ajuster, plus précisément que les contours statiques, à la frontière entre l'objet et le fond selon des déformations qui sont contrôlées par des fonctions d'énergie interne et externe. La première permet d'estimer spatialement les frontières entre l'objet et le fond, la deuxième contrôle la régularité du contour (lissage etc.). Les contours dynamiques peuvent être utilisés pour extraire le corps du sujet en entier [Baumberg94] ou des parties de ce corps [Kakadiaris98]. Une autre approche est l'extraction de la silhouette au lieu du contour. Rigoll, Eickeler et Müller utilisent des pseudo-2D chaînes de Markov cachées *HMM (Hidden Markov Models)* pour extraire la silhouette d'une personne dans une représentation qui est la transformée en cosinus discret de l'image [Rigoll00].

2.2 Segmentation basée sur les champs aléatoires de Markov

Après avoir présenté un état de l'art sur les différentes approches pour réaliser une segmentation spatio-temporelle, nous allons maintenant présenter l'une des deux méthodes de segmentation 2D spatio-temporelle disponibles dans notre système. Cette méthode est basée sur les champs aléatoires de Markov.

2.2.1 Introduction

Cette partie décrit un algorithme de segmentation 2D spatio-temporelle d'objets en mouvement développé dans le contexte du projet européen *Art.live* [Art.live02], cf. annexe A.1. Cet algorithme est basé sur les champs aléatoires de Markov et a été développé dans la thèse d'Alice Caplier [Caplier95]. Il a ensuite été amélioré [Caplier01].

Comme notre système est composé d'une scène filmée par une caméra fixe, nous utilisons une approche basée sur la soustraction et le seuillage, en combinant deux aspects complémentaires qui sont la différence d'images successives et l'utilisation d'une image de référence réactualisée.

Afin d'obtenir une segmentation de qualité avec une certaine stabilité temporelle, ce processus de segmentation utilise une modélisation par champs de Markov qui prend en compte à la fois les différences d'images consécutives et une image de référence d'une façon unifiée. Le cadre markovien est une façon efficace de prendre en compte différentes sources d'information en vue de prendre une décision. Les différences d'images sont prédominantes lorsque l'image de référence n'est pas encore (ou pas complètement) disponible, alors que l'image de référence prédomine pour les objets mobiles faiblement texturés ou pour les objets dont le mouvement cesse.

2.2.2 Extraction d'objets en mouvement

2.2.2.1 Étiquettes et observations

La segmentation basée mouvement est vue ici comme un problème d'étiquetage binaire dont le but est d'attribuer à chaque pixel ou site $s = (x, y)$ de l'image I à l'instant t l'une des deux étiquettes suivantes :

$$e(x, y, t) = e(s, t) = \begin{cases} obj \text{ si } s \text{ appartient à un objet,} \\ bg \text{ si } s \text{ appartient au fond } (bg : background). \end{cases}$$

$e = \{e(s,t), s \in I\}$ représente une réalisation particulière (à l'instant t) du champ d'étiquettes E. De plus, nous définissons $\epsilon = \{e\}$ comme l'ensemble des réalisations possibles du champ E.

Grâce à l'hypothèse n°1 d'un environnement filmé par une **caméra fixe**, l'information de mouvement est reliée directement aux changements temporels de la fonction intensité $I(s,t)$ et aux changements entre l'image courante $I(s,t)$ et l'image de référence $I_{ref}(s,t)$. Par conséquent, nous définissons deux observations :

1. l'observation O_{mvt} définie comme la différence de deux images consécutives :

$$O_{mvt}(s,t) = |I(s,t) - I(s,t-1)|;$$

2. l'observation O_{ref} définie comme la différence entre l'image courante et l'image de référence :

$$O_{ref}(s,t) = |I(s,t) - I_{ref}(s,t)|.$$

$o_{mvt} = \{O_{mvt}(s,t), s \in I\}$ et $o_{ref} = \{O_{ref}(s,t), s \in I\}$ représentent une réalisation particulière (à l'instant t) des champs d'observation respectifs O_{mvt} et O_{ref}.

Pour trouver la configuration la plus probable du champ E étant donnés les champs o_{mvt} et o_{ref}, nous utilisons le critère *MAP (Maximum A Posteriori)* et cherchons $e \in \epsilon$ tel que ($Pr[\cdot]$ étant la probabilité) :

$$Pr[E = e/O_{mvt} = o_{mvt}, O_{ref} = o_{ref}] \text{ maximum},$$

ce qui, en utilisant la règle de Bayes, est équivalent à chercher $e \in \epsilon$ tel que :

$$Pr[E = e]\ Pr[O_{mvt} = o_{mvt}, O_{ref} = o_{ref}/E = e] \text{ maximum}.$$

2.2.2.2 Fonction d'énergie

La maximisation de cette probabilité est équivalente à la minimisation d'une fonction d'énergie U qui est la somme pondérée de plusieurs termes [Geman84] :

$$U(e, o_{mvt}, o_{ref}) \quad = \quad U_m(e) + \lambda_{mvt} U_a(o_{mvt}, e) + \lambda_{ref} U_a(o_{ref}, e). \tag{2.1}$$

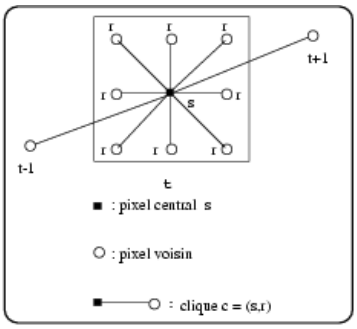


FIG. 2.1 – Voisinage spatio-temporel et cliques.

Le terme $U_m(e)$ de la fonction d'énergie, équation (2.1), peut être interprété comme un terme de régularisation qui assure l'homogénéité spatiale et temporelle du masque des objets mobiles et élimine les pixels isolés dus au bruit. Son expression, qui résulte de l'équivalence entre les champs de Markov et la distribution de Gibbs, est :

$$U_m(e) = \sum_{c \in C} V_c(e_s, e_r),$$

où c est une clique quelconque définie sur le voisinage spatio-temporel de la figure 2.1. Une clique $c = (s,r)$ est une paire quelconque de sites distincts telle que s est le pixel courant et r

un quelconque de ses voisins. C est l'ensemble de toutes les cliques. $V_c(e_s, e_r)$ est une fonction de potentiel élémentaire associée à chaque clique $c = (s, r)$. Elle prend les valeurs suivantes :

$$V_c(e_s, e_r) = \begin{cases} -\beta_r & \text{si } e_s = e_r, \\ +\beta_r & \text{si } e_s \neq e_r, \end{cases}$$

où le paramètre positif β_r dépend de la nature de la clique : $\beta_r = 20, \beta_r = 5, \beta_r = 50$ respectivement pour une clique spatiale, pour une clique temporelle vers le passé et pour une clique temporelle vers le futur. Ces valeurs ont été fixées expérimentalement une fois pour toutes en favorisant le futur par rapport au passé afin d'éliminer l'écho.

Le lien entre les étiquettes et chaque observation (notée de façon générique o) est défini par la relation suivante :

$$o(s,t) = \psi(e(s,t)) + n(s),$$
$$\text{où } \psi(e(s,t)) = \begin{cases} 0 & \text{si } e(s,t) = bg, \\ \alpha > 0 & \text{si } e(s,t) = obj. \end{cases}$$

où $n(s)$ est un bruit blanc gaussien de moyenne nulle et de variance σ^2. σ^2 est estimée de façon empirique comme la variance de chaque champ d'observation [Caplier95].

$\psi(e(s,t))$ modélise le lien entre les observations et les étiquettes de telle sorte que n représente le bruit d'adéquation :

- si le pixel s appartient au fond fixe, aucun changement temporel (autre que le bruit) n'intervient ni dans I, ni dans sa différence avec l'image de référence, si bien que chaque observation est quasiment nulle ;
- si le pixel s appartient à un objet mobile, un changement intervient dans les deux observations et chaque observation est supposée proche d'une valeur positive α_{mvt} et α_{ref} qui représente la valeur moyenne prise par chaque observation.

Les énergies d'adéquation $U_a(o_{mvt}, e)$ et $U_a(o_{ref}, e)$ sont calculées selon les relations suivantes :

$$U_a(o_{mvt}, e) = \frac{1}{2\sigma^2_{mvt}} \sum_{s \in I} [o_{mvt}(s,t) - \psi(e(s,t))]^2,$$
$$U_a(o_{ref}, e) = \frac{1}{2\sigma^2_{ref}} \sum_{s \in I} [o_{ref}(s,t) - \psi(e(s,t))]^2.$$

Dans la définition de l'énergie totale U, équation (2.1), deux coefficients de pondération λ_{mvt} et λ_{ref} sont introduits car le bon comportement de l'algorithme résulte d'un compromis entre tous les termes d'énergie.

La valeur $\lambda_{mvt} = 1$ est fixée une fois pour toutes et ne dépend pas de la séquence traitée. La valeur de λ_{ref} est fixée selon la règle suivante :

- $\lambda_{ref} = 0$ si $I_{ref}(s,t)$ n'existe pas : quand l'image de référence n'est pas encore disponible au pixel s, $o_{ref}(s,t)$ n'influence pas le processus de relaxation ;
- $\lambda_{ref} = 25$ si $I_{ref}(s,t)$ existe. Cette valeur est élevée car une grande confiance peut être accordée à l'image de référence quand elle existe. Elle permet d'obtenir des termes d'énergie d'adéquation du même ordre de grandeur.

2.2.2.3 Relaxation

L'algorithme de relaxation déterministe *ICM (Iterated Conditional Modes)* [Besag86] est utilisé pour trouver un minimum local de la fonction d'énergie totale U. Ayant constaté que la diminution la plus importante de la fonction d'énergie se produit dans les premières itérations, nous décidons de n'effectuer que quatre itérations *ICM*. De plus, une itération *ICM* sur deux est remplacée par une fermeture et une ouverture morphologique, qui sont effectuées sur le champ d'étiquettes E. Il en résulte une augmentation de la cadence de traitement sans perte de qualité puisque les itérations markoviennes restantes continuent à fonctionner directement sur les observations (différences entre images) et non pas sur les champs d'observation binarisés.

2.2.2.4 Initialisation

Cet algorithme étant non seulement itératif, mais aussi sous-optimal (au sens où il converge vers le premier minimum local), une initialisation du champ d'étiquettes E est nécessaire. Elle résulte d'un OU logique entre les deux champs d'observations O_{mvt} et O_{ref} binarisés. Cela nécessite deux seuils de binarisation qui sont choisis en fonction du type de séquence et du système d'acquisition vidéo [Caplier95].

2.2.2.5 Étiquetage en composantes connexes

Une fois obtenu le champ d'étiquettes E de l'image I à l'instant t, nous disposons donc d'une étiquette $e(x, y, t)$, pour le pixel de coordonnées (x, y) qui fixe son appartenance soit au fond ($e(x, y, t) = bg$) soit à un objet ($e(x, y, t) = obj$).

Un objet est défini comme un ensemble connexe de pixels, c'est-à-dire qu'à partir de n'importe quel pixel de cet objet, il est possible d'atteindre n'importe quel autre pixel (de cet objet) par un chemin qui n'est composé que de pixels appartenant à cet objet.

Grâce à un étiquetage en composantes connexes, nous obtenons le champ d'étiquettes finales E_f des objets vidéo, champ dont une réalisation particulière est $e_f = \{e_f(x, y, t) = e_f(s, t), s \in I\}$. Les étiquettes finales étant des entiers **consécutifs** à partir de 1 pour les objets vidéo et, par définition, 0 pour l'étiquette du fond. Chaque pixel d'un objet particulier a donc la même étiquette finale. Chaque pixel du fond a l'étiquette 0, cf. figure 3.1 page 59.

2.2.3 Construction et mise à jour de l'image de référence

Suivant le type d'environnement, la construction de l'image de référence est plus ou moins facile. Dans le cas d'un environnement intérieur, il est possible de créer l'image de référence en utilisant des images de la scène filmée lorsque personne ne se trouve dans le champ de la caméra. Dans le cas d'un environnement extérieur, par contre, où l'on contrôle généralement peu les conditions d'éclairage et où il peut être difficile d'avoir accès à une scène vide d'objets en mouvement (lieu très fréquenté), il devient nécessaire de construire l'image de référence au fur et à mesure.

Dans le cas général, l'image de référence est donc construite image après image, en réutilisant le résultat de la détection décrite précédemment, selon l'équation (2.2) où s désigne un pixel de l'image et t le temps :

$$I_{ref}(s, t+1) \quad = \quad \gamma(s, t) I_{ref}(s, t) + (1 - \gamma(s, t))\, I(s, t+1), \tag{2.2}$$

$$\text{où } \gamma(s,t) = \begin{cases} 0 & \text{si le pixel est statique et } I_{ref}(s,t) \text{ n'existe pas,} \\ 0.5 & \text{si le pixel est statique et } I_{ref}(s,t) \text{ existe,} \\ 1 & \text{si le pixel est mobile.} \end{cases}$$

Ce processus d'intégration temporelle tient compte des trois éléments suivants :
- La construction de I_{ref} au pixel s n'est possible que si s est détecté comme statique (c'est-à-dire appartient au fond fixe de la scène).
- La mise à jour de I_{ref} est nécessaire pour prendre en compte les variations d'illumination ou les changements de contenu du fond. Un délai de quelques images (~ 15) est observé avant la mise à jour de $I_{ref}(s,t+1)$ dans le cas de pixels "incohérents" : ceux qui sont statiques, mais pour lesquels la différence entre $I_{ref}(s,t)$ et $I(s,t)$ est trop élevée. Ainsi, nous évitons l'intégration dans l'image de référence de bruit temporel fort et d'objets qui deviennent brièvement immobiles.
- I_{ref} est maintenue identique pour les pixels détectés comme mobiles.

Grâce aux hypothèses optionnelles n°3 et n°4, qui supposent respectivement que l'environnement est **intérieur** et que la scène filmée est **vide au début de la séquence vidéo**, les conditions de construction de l'image de référence sont facilitées. L'hypothèse optionnelle n°3 permet d'avoir des conditions d'acquisition où l'illumination est relativement constante, l'hypothèse n°4 permet d'obtenir une image de référence robuste puisque l'on utilise des images où la scène ne comporte aucune *ROI* (personne). Ces deux hypothèses permettent d'obtenir de très bons résultats de segmentation.

2.3 Segmentation optimisée en vitesse

Dans notre système, une deuxième méthode de segmentation 2D spatio-temporelle d'objets vidéo en mouvement est disponible. C'est une segmentation développée par Umeda, Hernandez, Marques et Marichal, dans le cadre du projet *Art.live* [Art.live02], cf. annexe A.1. Elle donne accès aux objets segmentés et étiquetés en composantes connexes avec les mêmes conventions que la segmentation basée sur les champs aléatoires de Markov.

Nous ne détaillerons pas cette segmentation car elle a été optimisée en assembleur et il n'a pas été possible de déterminer comment elle a été mise en œuvre. Néanmoins, nous pensons que cette segmentation a été améliorée dans [Umeda01, Marichal03, Umeda04], principalement à cause des cadences de traitement présentées dans ces travaux.

2.4 Données bas-niveau extraites

2.4.1 Masques de segmentation

La première donnée extraite lors de cette étape de segmentation est l'ensemble des masques de segmentation des *ROI* présentes dans la scène. Ces masques ont été étiquetés en composantes connexes à la fin de l'étape de segmentation. Un masque de segmentation en tant que tel est défini comme l'ensemble des pixels connexes d'un objet ayant été extrait de la scène par rapport au fond. Chaque masque possède une étiquette individuelle. Les étiquettes sont des valeurs numériques entières consécutives, qui vont de 1 au nombre d'objets détectés. L'étiquette 0 désigne les pixels du fond. À partir d'un masque de segmentation, nous pouvons calculer d'autres données (**descripteurs**) qui décrivent l'objet.

Les descripteurs calculés sont :
– la surface, c'est-à-dire le nombre de pixels qui composent l'objet ;
– le centre de l'objet ;
– le centre de gravité de l'objet ;
– des boîtes englobant ou non l'objet.

2.4.2 Boîtes englobant ou non les objets

À partir des masques de segmentation, nous pouvons calculer des boîtes qui peuvent contenir l'objet dans son entier, nous parlons alors de boîtes englobantes, ou non. Quatre boîtes sont calculées, la boîte englobante rectangulaire, la boîte quadrangulaire, la boîte octogonale et une boîte définie selon les axes principaux de l'objet. Les trois premières boîtes, dans l'ordre, approchent de plus en plus la forme segmentée de l'objet. La quatrième apporte de l'information sur la forme de l'objet. Chaque boîte possède deux définitions, une définition **intrinsèque** qui correspond aux coordonnées des points qui appartiennent à cette boîte et une définition **paramétrique** qui peuvent permettre de calculer les trajectoires des centres et les moyennes et les variances des dimensions des boîtes après l'étape de suivi temporel.

2.4.2.1 Boîte englobante rectangulaire (BER)

La boîte englobante rectangulaire, ou rectangle englobant, est définie comme la boîte rectangulaire de dimensions minimales contenant entièrement l'objet. Elle peut être définie de façon intrinsèque ou paramétrique (cf. figure 2.2). La définition intrinsèque est composée de quatre entiers définissant les coins de la boîte :

$$(x_{min}, y_{min}, x_{max}, y_{max})$$

La définition paramétrique est composée des coordonnées du centre de l'objet (ou centre de la boîte) et des dimensions de la boîte, qui sont aussi des entiers :

$$(x_{center}, y_{center}, width, height)$$

(a) (b)

FIG. 2.2 – Définitions intrinsèque (a) et paramétrique (b) de la BER.

L'indice $_{center}$, *width* et *height* correspondent respectivement au centre, à la largeur et à la hauteur de la boîte englobante rectangulaire. Il est possible de passer d'une définition à l'autre simplement, par les formules suivantes :

$$\begin{cases} x_{center} = (x_{min} + x_{max})/2 \\ y_{center} = (y_{min} + y_{max})/2 \\ width = x_{max} - x_{min} \\ height = y_{max} - y_{min} \end{cases} \quad \begin{cases} x_{min} = x_{center} - width/2 \\ y_{min} = y_{center} - height/2 \\ x_{max} = x_{center} + width/2 \\ y_{max} = y_{center} + height/2 \end{cases} \tag{2.3}$$

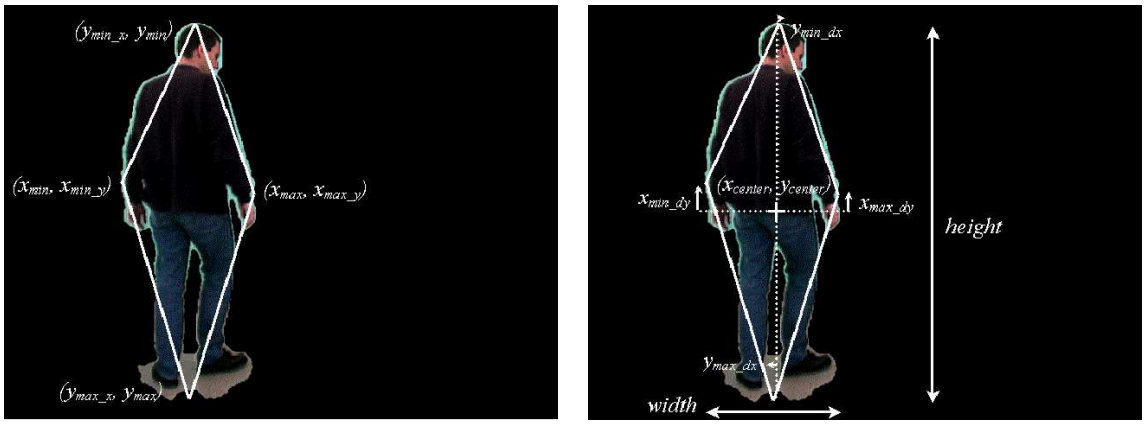
2.4.2.2 Boîte quadrangulaire (BQ)

La boîte quadrangulaire, ou quadrangle, est obtenue à partir de la boîte englobante rectangulaire. Elle ne contient généralement pas l'objet dans son entier. Les deux définitions, intrinsèque et paramétrique, sont chacune composées de huit entiers (cf. figure 2.3). La définition intrinsèque est composée des huit entiers définissant les coins de la boîte :

$$(x_{min}, x_{min_y}, y_{min_x}, y_{min}, x_{max}, x_{max_y}, y_{max_x}, y_{max})$$

La définition paramétrique est composée des coordonnées du centre de l'objet (ou centre de la boîte), des dimensions de la boîte et des décalages nécessaires pour obtenir les coordonnées des coins :

$$(x_{center}, y_{center}, width, height, x_{min_{dy}}, x_{max_{dy}}, y_{min_{dx}}, y_{max_{dx}})$$



(a) (b)

FIG. 2.3 – Définitions intrinsèque (a) et paramétrique (b) de la BQ.

En parcourant un par un les bords de la boîte englobante rectangulaire à partir des coins, les coins de la boîte quadrangulaire sont situés au milieu du segment formé par les deux premiers pixels du bord appartenant à l'objet. Il est possible de passer d'une définition à l'autre en rajoutant aux formules de (2.3) les formules suivantes :

$$\begin{cases} x_{min_{dy}} = x_{min_y} - (y_{min} + y_{max})/2 \\ x_{max_{dy}} = x_{max_y} - (y_{min} + y_{max})/2 \\ y_{min_{dx}} = y_{min_x} - (x_{min} + x_{max})/2 \\ y_{max_{dx}} = y_{max_x} - (x_{min} + x_{max})/2 \end{cases} \quad \begin{cases} x_{min_y} = y_{center} + x_{min_{dy}} \\ y_{min_x} = x_{center} + y_{min_{dx}} \\ x_{max_y} = y_{center} + x_{max_{dy}} \\ y_{max_x} = x_{center} + y_{max_{dx}} \end{cases} \tag{2.4}$$

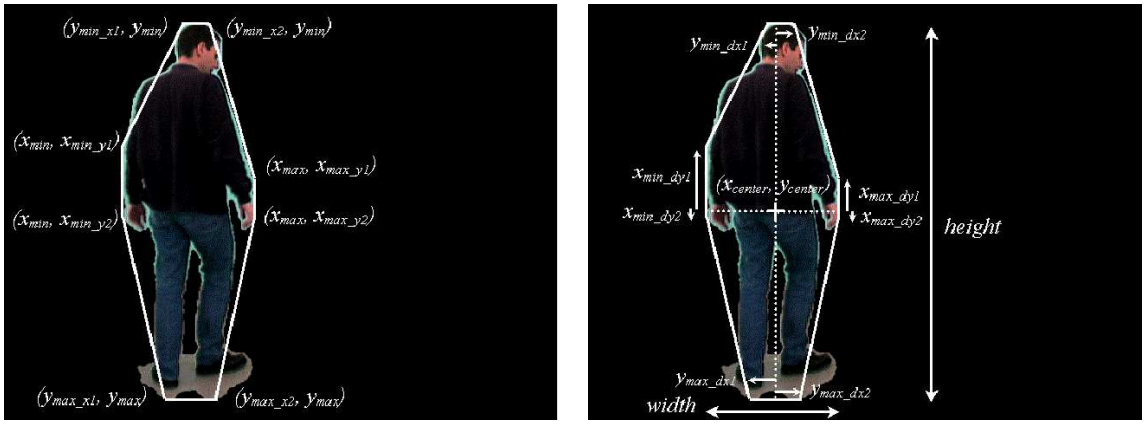
2.4.2.3 Boîte octogonale (BO)

La boîte octogonale est aussi obtenue à partir de la boîte englobante rectangulaire. Elle ne contient généralement pas non plus l'objet dans son entier. Les deux définitions, intrinsèque et paramétrique, sont chacune composées de douze entiers (cf. figure 2.4). La définition intrinsèque est composée des douze entiers définissant les coins de la boîte :

$$(x_{min}, x_{min_{y1}}, x_{min_{y2}}, y_{min_{x1}}, y_{min_{x2}}, y_{min}, x_{max}, x_{max_{y1}}, x_{max_{y2}}, y_{max_{x1}}, y_{max_{x2}}, y_{max})$$

La définition paramétrique est composée des coordonnées du centre de l'objet (ou centre de la boîte), des dimensions de la boîte et des décalages nécessaires pour obtenir les coordonnées des coins :

$$(x_{center}, y_{center}, width, height, x_{min_{dy1}}, x_{min_{dy2}}, x_{max_{dy1}}, x_{max_{dy2}}, y_{min_{dx1}}, y_{min_{dx2}}, y_{max_{dx1}}, y_{max_{dx2}})$$



(a) (b)

FIG. 2.4 – Définitions intrinsèque (a) et paramétrique (b) de la BO.

En parcourant un par un les bords de la boîte englobante rectangulaire à partir des coins, les coins de la boîte octogonale sont les deux premiers pixels du bord appartenant à l'objet. Dans le cas où un seul pixel de l'objet est sur le bord, la boîte octogonale a une forme dégénérée pouvant aller jusqu'à un quadrilatère, qui est alors identique à la boîte quadrangulaire. Il est possible de passer d'une définition à l'autre en rajoutant aux formules de (2.3) les formules suivantes :

$$\begin{cases} x_{min_{dy1}} = x_{min_{y1}} - (y_{min} + y_{max})/2 \\ x_{min_{dy2}} = x_{min_{y2}} - (y_{min} + y_{max})/2 \\ x_{max_{dy1}} = x_{max_{y1}} - (y_{min} + y_{max})/2 \\ x_{max_{dy2}} = x_{max_{y2}} - (y_{min} + y_{max})/2 \\ y_{min_{dx1}} = y_{min_{x1}} - (x_{min} + x_{max})/2 \\ y_{min_{dx2}} = y_{min_{x2}} - (x_{min} + x_{max})/2 \\ y_{max_{dx1}} = y_{max_{x1}} - (x_{min} + x_{max})/2 \\ y_{max_{dx2}} = y_{max_{x2}} - (x_{min} + x_{max})/2 \end{cases} \quad \begin{cases} x_{min_{y1}} = y_{center} + x_{min_{dy1}} \\ x_{min_{y2}} = y_{center} + x_{min_{dy2}} \\ y_{min_{x1}} = x_{center} + y_{min_{dx1}} \\ y_{min_{x2}} = x_{center} + y_{min_{dx2}} \\ x_{max_{y1}} = y_{center} + x_{max_{dy1}} \\ x_{max_{y2}} = y_{center} + x_{max_{dy2}} \\ y_{max_{x1}} = x_{center} + y_{max_{dx1}} \\ y_{max_{x2}} = x_{center} + y_{max_{dx2}} \end{cases} \tag{2.5}$$

2.4.2.4 Boîte par axes principaux (BAP)

La boîte par axes principaux est calculée à part. Elle est issue de l'approximation de l'objet par une ellipse 2D. C'est, par définition, le rectangle englobant cette ellipse 2D d'approximation. La figure 2.5 illustre sur un exemple une ellipse 2D d'approximation, ses paramètres et la boîte par axes principaux qui en découle.

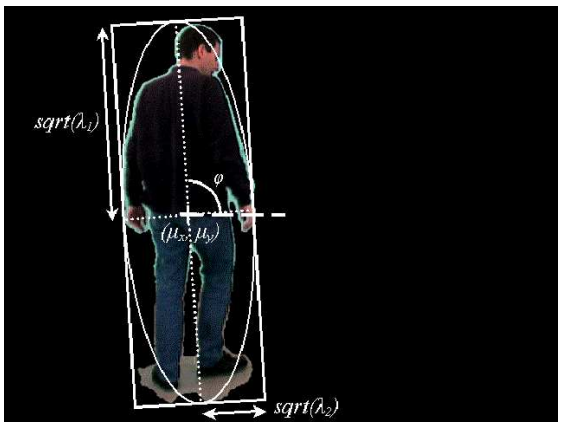


FIG. 2.5 – Ellipse 2D d'approximation et boîte par axes principaux.

Pour obtenir la boîte par axes principaux, nous calculons les moments du premier ordre et les moments centrés du second ordre de l'ensemble des pixels définissant l'objet. Les moments du premier ordre sont notés $x_{gravity} = \mu_x$ et $y_{gravity} = \mu_y$ (cf. figure 2.6). L'indice $_{gravity}$ correspond au centre de gravité. En calculant les valeurs propres λ_1 et λ_2 de la matrice d'inertie définie selon les moments centrés du second ordre, on peut en déduire les dimensions de la boîte par axes principaux, c'est-à-dire la largeur *width* et la hauteur *height*, et l'angle *angle* formé entre l'axe des abscisses dans l'image (horizontale) et l'axe vertical de la boîte par axes principaux.

Nous avons alors les cinq paramètres utiles au tracé d'une ellipse 2D, et donc de la boîte par axes principaux d'inertie :

– les coordonnées de son centre, qui correspond au centre de l'ellipse 2D et au centre de gravité de l'objet : $x_{gravity}$ et $y_{gravity}$;

– les dimensions de la boîte, qui sont les longueurs des axes d'inertie de l'ellipse 2D : *width* et *height* ;
– l'angle de la boîte par rapport à l'axe des abscisses, qui est celui de l'ellipse : *angle*.

La boîte par axes principaux ne contient généralement pas non plus l'objet dans son entier. Elle peut être définie de façon intrinsèque ou paramétrique (cf. figure 2.6). La définition intrinsèque de la boîte par axes principaux comporte huit entiers définissant les coordonnées des coins de la boîte :

$$(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$$

La définition paramétrique de la boîte par axes principaux est composée de cinq valeurs numériques, quatre entières et une réelle, qui sont : les coordonnées du centre de gravité de l'objet, les dimensions de la boîte et l'angle de la boîte par axes principaux :

$$(x_{gravity}, y_{gravity}, width, height, angle)$$

(a) (b)

FIG. 2.6 – Définitions intrinsèque (a) et paramétrique (b) de la BAP.

2.4.3 Résumé

En résumé, lors de l'étape de segmentation 2D spatio-temporelle, pour chaque objet segmenté, nous disposons des données bas-niveau suivantes :

– le masque de segmentation ;
– l'étiquette ;
– la surface ;
– le centre de l'objet ;
– le centre de gravité de l'objet ;
– la boîte englobante rectangulaire ;
– la boîte quadrangulaire ;
– la boîte octogonale ;
– la boîte par axes principaux.

2.5 Résultats

Les pages suivantes présentent quelques résultats de segmentation obtenus pour différentes séquences vidéo acquises dans des conditions variées. Chaque figure présente des séries de trois images, à gauche les images originales des séquences, au milieu les masques de segmentation obtenus et à droite le produit des masques de segmentation avec les images originales.

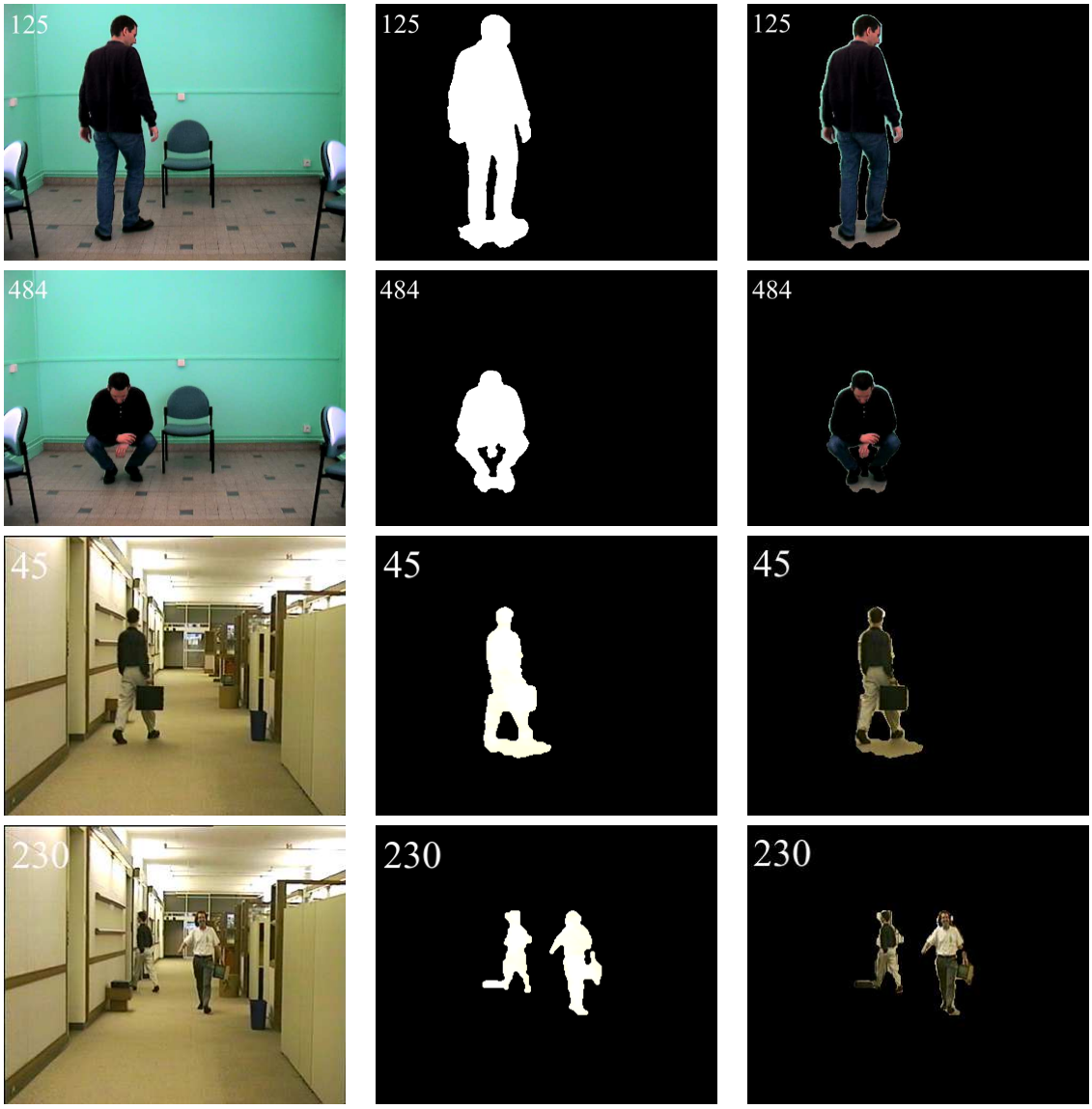


FIG. 2.7 – Environnement intérieur : segmentation basée sur les champs aléatoires de Markov.

La figure 2.7 illustre les résultats obtenus en environnement intérieur avec la segmentation 2D spatio-temporelle basée sur les champs aléatoires de Markov. Tout d'abord, nous pouvons remarquer que les masques de segmentation sont lisses et réguliers et approchent très bien les silhouettes des personnes. Ensuite, sur les trois premières séries d'images, les masques sont troués et ne comprennent pas les pixels du fond entre les jambes de la personne. Néanmoins,

nous pouvons aussi observer que les masques “bavent” un peu autour de la silhouette et particulièrement au niveau des pieds des personnes (principalement à cause des ombres portées des personnes en mouvement). Enfin, sur la dernière série d'images, la personne de droite a perdu une partie de son avant-bras gauche, qui n'a pas été segmenté. Sur cette même série, la personne de gauche, qui a entre-temps posé son attaché-case sur des boîtes en carton, et qui est en train de sortir du couloir, a un masque de segmentation commun avec l'attaché-case, alors que ce sont deux objets différents. Nous reparlerons de cet inconvénient lors des chapitres sur le suivi temporel (cf. chapitre 3 et chapitre 5).

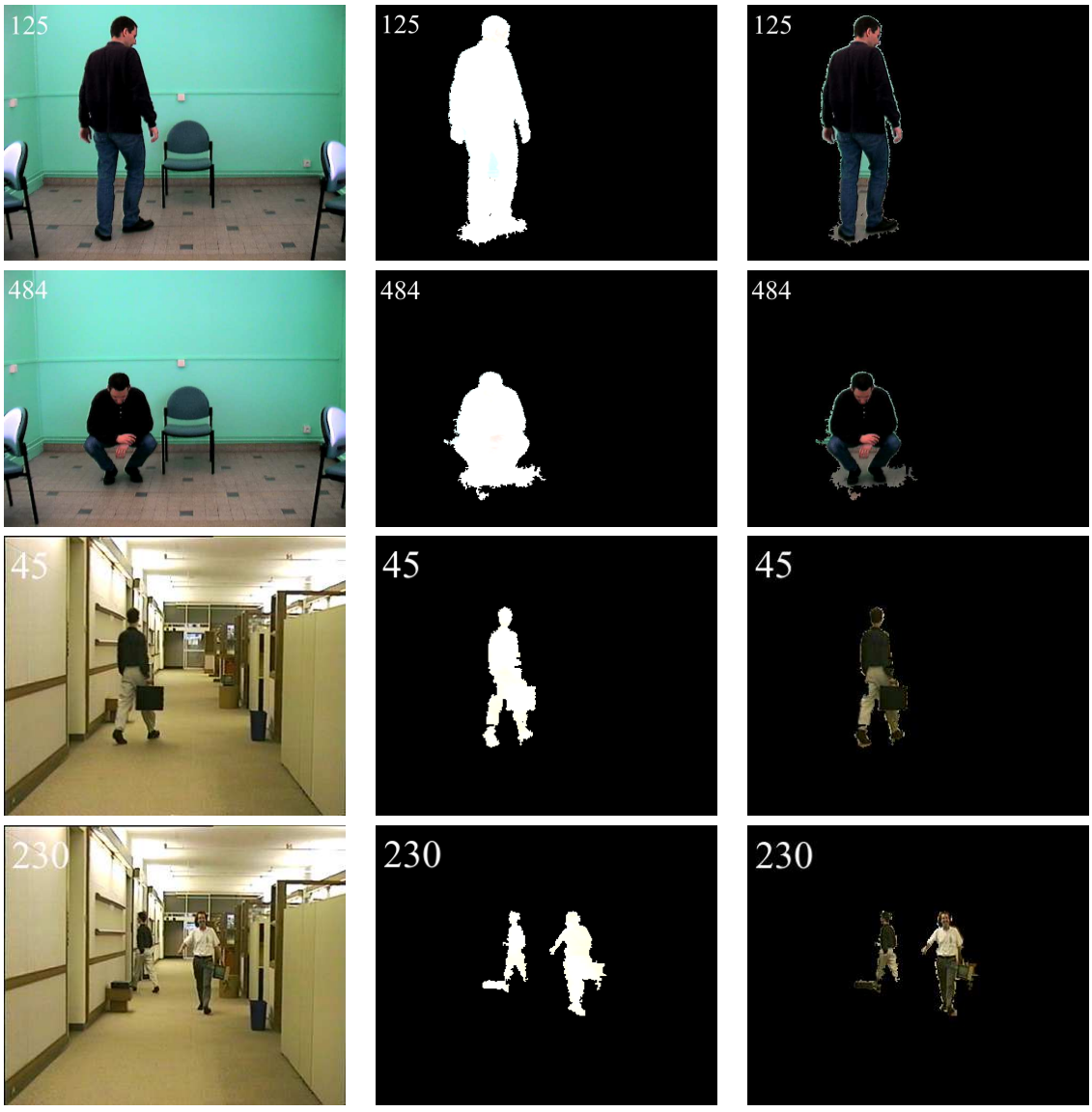


Fig. 2.8 – Environnement intérieur : segmentation optimisée en vitesse.

La figure 2.8 illustre les résultats obtenus en environnement intérieur pour les mêmes images mais avec la segmentation optimisée en vitesse. Tout d'abord, en comparant avec la figure 2.7, nous pouvons observer que les masques sont plus précis, mais moins lisses et réguliers

que ceux obtenus avec la première segmentation. Les masques obtenus avec la segmentation optimisée approchent encore plus les silhouettes des personnes, mais comprennent les pixels entre les jambes. De plus, sur les deux dernières série d'images, nous pouvons voir des masques "rognés". Dans la troisième série, une partie de la cuisse gauche de la personne filmée n'a pas été segmentée. Dans la quatrième série, ce sont ses pieds qui n'ont pas été segmentés.

Bien que, dans ce travail, nous nous focalisions sur le cas d'analyse de scènes intérieures, nous présentons pour chacune des méthodes, des résultats obtenus pour des scènes extérieures, ceci afin de montrer que ces segmentations 2D spatio-temporelles sont génériques et ne sont pas dédiées uniquement au traitement de scènes intérieures.

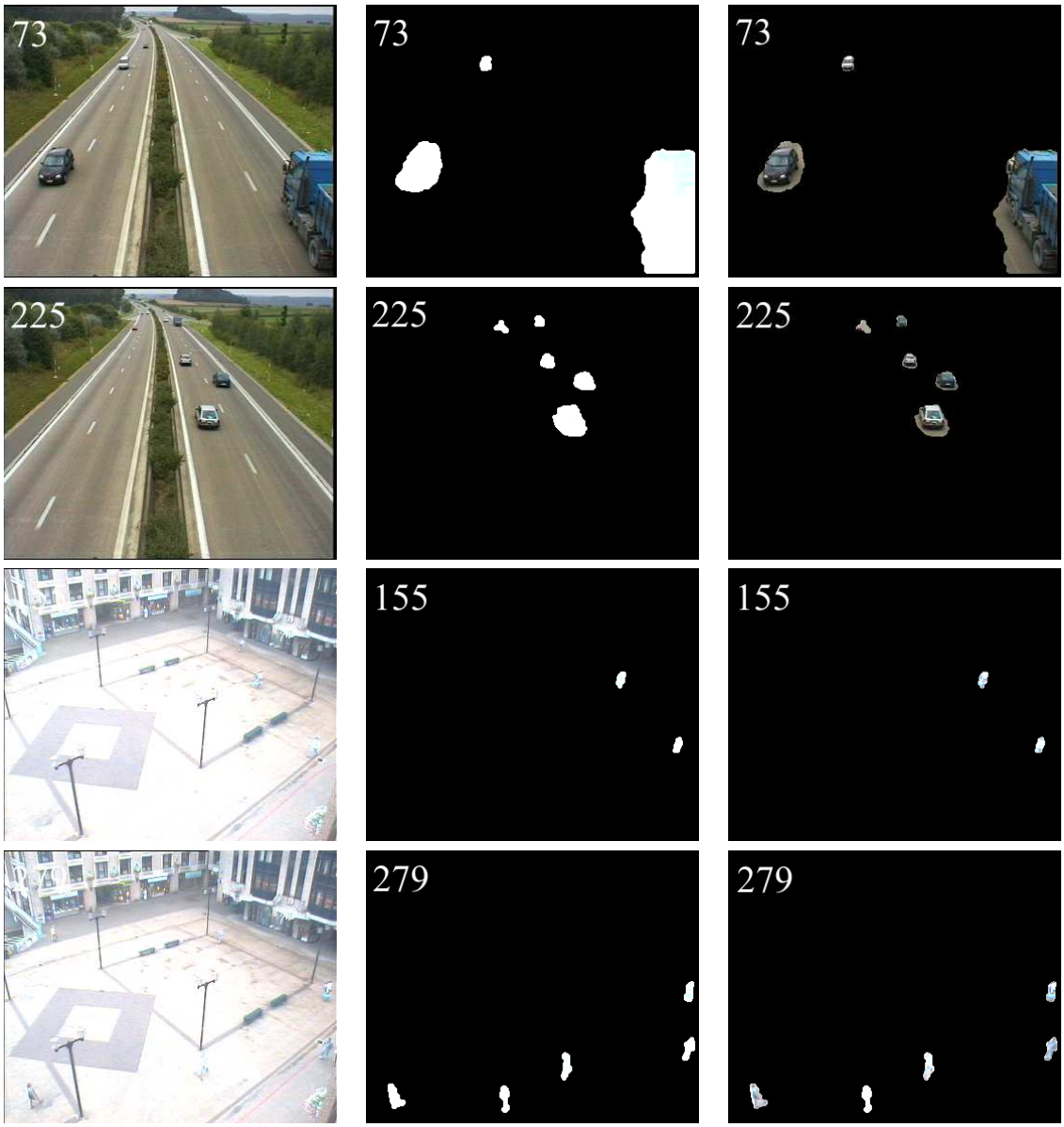


FIG. 2.9 – Environnement extérieur : segmentation basée sur les champs aléatoires de Markov.

La figure 2.9 illustre les résultats obtenus en environnement extérieur avec la segmentation 2D spatio-temporelle basée sur les champs aléatoires de Markov. Ici encore, nous pouvons voir que les masques “bavent” légèrement sur les côtés des objets segmentés, pour le camion en bas à droite de la première série d’images, par exemple.

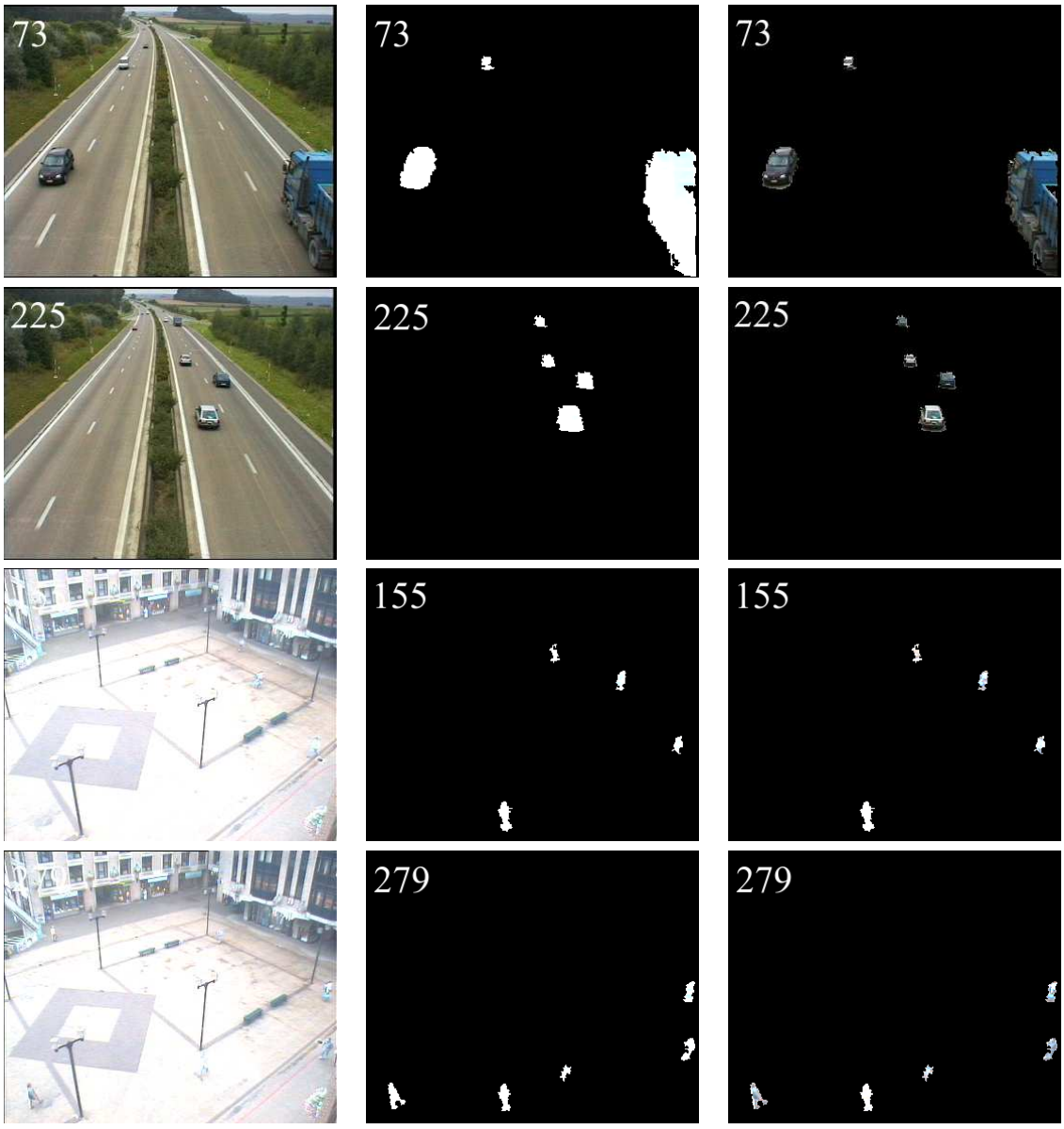


Fig. 2.10 – Environnement extérieur : segmentation optimisée en vitesse.

La figure 2.10 illustre les résultats obtenus en environnement extérieur pour les mêmes images mais avec la segmentation 2D spatio-temporelle optimisée. En comparant avec la figure 2.9, nous retrouvons les mêmes défauts qu’en environnement intérieur, les masques sont plus précis mais “rognent” aussi les objets segmentés (cf. quatrième série d’images). Toujours en comparant les figures 2.9 et 2.10, pour la troisième série d’images, nous pouvons voir que la

segmentation optimisée en vitesse permet de détecter plus d'objets (quatre personnes contre deux avec la segmentation basée sur les champs aléatoires de Markov) mais que les masques sont fortement "rognés". Sur la quatrième série d'images, la troisième personne en partant de la gauche est bien segmentée avec la segmentation basée sur les champs aléatoires de Markov, mais s'apparente à du bruit avec la segmentation optimisée en vitesse.

La figure 2.11 illustre quelques exemples de mauvaise segmentation lorsque les masques "bavent" sur de grandes zones de l'image, ou lorsque les masques sont assez "rognés", avec les mêmes images pour les deux types de segmentation. Pour cette figure, les deux séries de cinq images sont, de gauche à droite, l'image originale, le masque de segmentation obtenu avec la segmentation basée sur les champs aléatoires de Markov, le produit de ce masque avec l'image originale, le masque de segmentation obtenu avec la segmentation optimisée en vitesse et le produit de ce masque avec l'image originale.

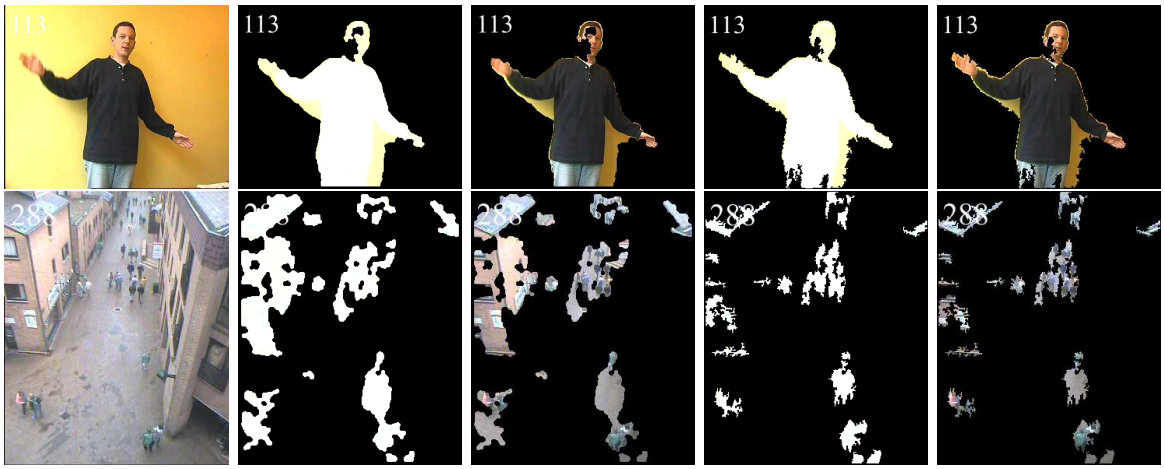


FIG. 2.11 – Mauvaise segmentation en environnement intérieur ou extérieur.

Sur la première série d'images, en environnement intérieur, nous pouvons voir qu'une partie du visage de la personne et une partie de ses jambes ont été "rognées" et que le masque "bave" sur une assez grande zone du fond, sous le bras droit et sur le côté gauche de la personne. Sur la deuxième série d'images, en environnement extérieur, il y a encore plus de masques qui sont "rognés" ou "bavent" et nous détectons même des parties du bâtiment en haut à gauche comme des objets en mouvement. Nous venons de voir que ces mauvais résultats peuvent survenir avec les deux segmentations, mais de manière générale, ils surviennent plus fréquemment en environnement extérieur qu'en environnement intérieur, principalement parce que les conditions d'acquisition sont très peu contrôlées en extérieur.

2.6 Avantages, limitations et cadences de traitement

Les segmentations 2D spatio-temporelles disponibles dans notre système présentent chacune des avantages et des limitations. Les résultats présentés dans la partie 2.5 illustrent ce que nous allons maintenant décrire. Les avantages et limitations présentées découlent du compromis qualité / cadence de traitement. Nous définissons la qualité des résultats en fonction de la précision (plus ou moins proche de la silhouette des objets) et de la régularité des masques

de segmentation obtenus. La cadence de traitement représente la vitesse des méthodes.

Concernant la qualité de la segmentation, les deux segmentations sont robustes et performantes car les masques de segmentation obtenus sont réguliers et approchent bien la forme des objets. Les masques obtenus par la segmentation basée sur les champs aléatoires de Markov sont lisses et réguliers, mais ont légèrement tendance à "baver", c'est-à-dire qu'ils comportent certains pixels du fond. Mais nous obtenons l'ensemble des pixels des objets en mouvement. Le fait d'avoir deux seuils de binarisation à régler, lorsque les conditions d'acquisition changent, n'est pas véritablement une limitation car il est possible de modifier ces seuils directement pendant le traitement de la séquence, en même temps que l'acquisition. Les masques obtenus par la segmentation optimisée sont, en un sens, meilleurs que ceux obtenus avec la segmentation basée sur les champs aléatoires de Markov, car ils sont plus proches de la silhouette des objets, mais sont aussi moins lisses et moins réguliers. De plus il arrive que la segmentation se dégrade et donne des masques "rognés", dont certaines parties sont manquantes.

Concernant les cadences de traitement de la segmentation, les résultats sont très différents selon la segmentation utilisée. La table 2.1 résume les pourcentages de temps de calcul et les cadences de traitement atteintes, selon la segmentation utilisée (champs aléatoires de Markov ou optimisée en vitesse) et selon la résolution d'image (320×240 ou 640×480). Les pourcentages de temps de calcul par rapport au temps de calcul total sont donnés pour l'ensemble des étapes de traitement du système jusqu'à l'étape de segmentation 2D spatio-temporelle. Comme c'est la première étape de traitement, il n'y a donc que l'acquisition et la segmentation. Les pourcentages obtenus ont été calculés en réalisant une moyenne des temps de traitement pour une demi-douzaine de séquences vidéo dans des conditions d'acquisition variées. Ils représentent donc une estimation du pourcentage réel de temps de calcul nécessaire à chaque étape de traitement.

Tab. 2.1 – Pourcentages de temps de calcul et cadences de traitement pour l'étape de segmentation 2D spatio-temporelle.

Segmentation	Champs aléatoires de Markov		Optimisée en vitesse	
Résolution d'image	320×240	640×480	320×240	640×480
Acquisition	0.3%	0.4%	6.6%	11.5 %
Segmentation	99.7%	99.6%	93.4%	88.5%
Cadences de traitement	**9.65** images/s	**2.2** images/s	**300** images/s	**65** images/s

Pour la segmentation basée sur les champs aléatoires de Markov, la complexité de l'algorithme l'empêche d'être très rapide. C'est la principale limitation de cette segmentation. Les estimations sont néanmoins des estimations basses des cadences de traitement, car la segmentation n'est pas optimisée. L'algorithme proposé est aussi suffisamment générique pour être utilisé dans d'autres applications telles que la vidéosurveillance par exemple. Pour la segmentation optimisée en vitesse, les cadences de traitement atteintes sont très élevées. Ces cadences sont beaucoup plus élevées que les cadences vidéo classiques (25 images/s) et respectent donc parfaitement la contrainte de temps-réel. Les ratios entre les cadences de traitement selon la résolution ne sont pas exactement les mêmes car ces mesures sont des estimations réalisées sur une demi-douzaine de séquences vidéo enregistrées.

2.7 Conclusion

Ce premier chapitre a présenté une méthode de segmentation 2D spatio-temporelle d'objets en mouvement, présents dans une scène filmée par une caméra fixe. Cette méthode est basée sur des champs aléatoires de Markov. Nous avons ensuite présenté les données bas-niveau extraites lors de cette étape de traitement, notamment les masques de segmentation et les boîtes englobant ou non les objets. La boîte englobante rectangulaire en particulier sera beaucoup utilisée par la suite puisqu'elle permet, avec quatre entiers seulement, et conjointement avec le masque de segmentation, de ne traiter que les pixels appartenant à un objet segmenté particulier. Nous y ferons alors référence sous le nom de BERS (Boîte Englobante Rectangulaire issue de la Segmentation). Puis nous avons illustré différents résultats pour les deux méthodes de segmentation disponibles dans notre système et nous avons ensuite donné les avantages, les limitations et les cadences de traitement pour chacune d'elles, la première donnant des masques de segmentation plus réguliers et la seconde étant beaucoup plus rapide.

Nous allons maintenant présenter quelques perspectives concernant cette première étape de traitement. Tout d'abord, il serait possible, à partir du masque de segmentation, d'extraire d'autres données bas-niveau. Nous avons la silhouette de l'objet, nous pourrions en déduire facilement le contour et tenter de réaliser les étapes de traitement ultérieures avec cette information, par exemple réaliser une approche de suivi temporel basé contour. Nous pourrions ensuite caractériser l'objet sous forme de *blobs* de couleur avec des caractéristiques statistiques (position et couleur) et réaliser un suivi temporel sur les différentes parties extraites.

Puis, il faudrait améliorer les résultats de segmentation selon la méthode choisie. Pour la segmentation basée sur les champs aléatoires de Markov, la perspective principale serait l'amélioration du code pour accélérer la cadence de traitement. Une perspective secondaire concernant cette méthode pourrait être de rajouter un traitement spécifique sur les ombres, par exemple en utilisant des informations sur le contour et sur la couleur. Des modèles d'ombre basés sur la couleur avec des techniques d'invariance pourraient aussi être utilisés pour résoudre cette difficulté [Salvador01]. Pour la segmentation optimisée en vitesse, une perspective intéressante serait l'amélioration de la régularité des masques obtenus.

Chapitre 3

Première étape du suivi temporel

Sommaire

Le suivi temporel d'objets en mouvement dans les séquences vidéo est un thème de recherche très étudié en vision par ordinateur. En effet, dans de nombreux systèmes, il est nécessaire de détecter et de suivre au cours du temps des objets ou des personnes en mouvement passant dans le champ d'une caméra. Le but du suivi temporel est de réaliser des liens temporels entre les objets détectés à l'instant $t-1$ et les objets détectés à l'instant t. Par rapport à l'étape de segmentation, c'est une étape de plus haut niveau. Cependant, les algorithmes de suivi temporel ont souvent beaucoup de points communs avec ceux de segmentation. Le suivi au cours du temps inclut typiquement l'association d'objets sur des images consécutives en utilisant des caractéristiques telles que les points, les lignes, les *blobs*, voire des modèles plus complexes tels que des squelettes 3D, des volumes 3D etc. En d'autres termes, le suivi temporel correspond à l'établissement de relations temporelles cohérentes entre les *ROI* considérées selon leurs caractéristiques et / ou des critères comme la position, la vitesse, la forme, la texture, la couleur etc. De même que pour l'étape de segmentation, il existe donc de très nombreuses façons de réaliser un suivi temporel :

- **régions** ou silhouettes *vs* **contours** ;
- **avec modèle** *vs* **sans modèle** ;
- **2D** *vs* **3D** ;
- **mono**-caméra *vs* **multi**-caméras ;
- etc.

Le suivi temporel comporte une difficulté inhérente, c'est le phénomène d'**occultation**. L'occultation est le fait qu'un objet peut se voir partiellement ou complètement caché du point vue de la caméra. Par exemple, avec une caméra monoculaire, des objets, même distants, peuvent ne constituer qu'un seul objet du point de vue de la caméra à cause de la projection 2D. Des objets qui rentrent en contact physiquement vont aussi amener le même résultat. L'occultation peut être **directe**, si l'objet est occulté par une autre *ROI*, ou **indirecte**, si l'objet est occulté par autre chose qu'une *ROI*. Le phénomène d'occultation directe est lié aux **réunions** et aux **séparations temporelles** d'objets. Une occultation directe commence par une réunion temporelle et se termine par une séparation temporelle d'objets. La gestion de cette difficulté selon l'approche choisie est ce qui fait de l'étape de suivi temporel une tâche complexe.

Dans les systèmes qui considèrent surtout les êtres humains, le suivi temporel peut être réalisé pour une personne seule, vue comme un ensemble de parties du corps humain ou comme un tout ; ou pour des groupes d'individus, vus comme des objets formés de plusieurs personnes ou comme un tout. Concernant les êtres humains, le suivi temporel est une tâche difficile, car le corps humain n'est pas une structure rigide, mais articulée et de mouvement complexe. Les possibilités de déformations du corps humain en terme de mouvement sont en effet considérables. Le suivi temporel est une étape cruciale en analyse du comportement humain car il permet de faire un lien temporel entre les caractéristiques choisies pour interpréter le comportement humain. Au niveau des occultations, des personnes peuvent se regrouper visuellement en passant l'une devant l'autre. Elles peuvent aussi entrer en contact physiquement, par exemple en se serrant la main.

Comme la segmentation, cette première étape du suivi temporel est générique et peut être utilisée pour des objets (*ROI*) autres que des personnes dans des séquences vidéo. Les étapes de traitement ultérieures sont dédiées au traitement de séquences vidéo dont les *ROI* sont des êtres humains.

Nous commencerons par présenter dans ce chapitre un état de l'art sur les techniques utilisées pour le suivi temporel en général mais néanmoins orientées selon les résultats de notre

segmentation. Nous détaillerons ensuite la méthode utilisée pour réaliser le suivi temporel d'objets présents dans une scène filmée par une caméra fixe. Après cela, nous exposerons les données bas-niveau extraites lors de cette seconde étape de traitement. Puis nous illustrerons les résultats obtenus avec cette méthode par des images issues du traitement de séquences vidéo variées. Nous préciserons ensuite les avantages, les limitations et les cadences de traitement atteintes avant de conclure sur cette étape de traitement.

3.1 État de l'art sur le suivi temporel

Un des problèmes les plus difficiles pour un système de vision par ordinateur dynamique est le suivi temporel d'objets articulés, tels que des corps humains dans un environnement complexe. Comme les personnes peuvent porter des vêtements de formes, de couleurs et de textures variées, il est difficile d'identifier les contours du corps et les limites entre les parties du corps. Cette partie présente un état de l'art sur le suivi temporel en général guidé par les résultats fournis par notre étape de segmentation.

Le suivi temporel peut être divisé en diverses catégories selon différents critères. Selon le type d'objets suivis, par exemple, si l'on considère des êtres humains, il y a alors le suivi temporel de parties du corps humain (main, visage, bras, jambe etc.), d'individu seul (corps considéré dans son entier), d'individus multiples et séparés et de groupes de personnes. Si le nombre de vues est considéré, il y a le suivi temporel vue simple, vues multiples et vue omnidirectionnelle. Le suivi temporel peut être classé selon d'autres critères comme la dimension de l'espace (2D *vs* 3D), l'environnement (intérieur *vs* extérieur), le nombre d'objets ou de personnes suivies (seul(e), multiples, en groupe), l'état de la caméra (fixe *vs* mobile), la multiplicité des caméras (mono *vs* stéréo) etc.

Nous classerons le suivi temporel d'objets en mouvement en quatre types d'approche :
- suivi temporel basé **région** ;
- suivi temporel basé **contour** ;
- suivi temporel basé **caractéristique** ;
- suivi temporel basé **modèle**.

Nous ne présenterons dans ce chapitre qu'un état de l'art restreint sur les trois premiers types d'approche. En effet, comme pour l'étape de segmentation, il existe un très grand nombre de méthodes et il serait impensable de vouloir réaliser un état de l'art exhaustif. De plus, comme le chapitre 5 de ce mémoire est dédié plus spécifiquement au suivi temporel d'êtres humains, l'état de l'art sur le suivi temporel basé sur des modèles de corps humain est donc présenté au chapitre 5.

3.1.1 Suivi temporel basé région

L'idée ici est d'identifier une région connexe associée à chaque objet en mouvement dans une image, et de la suivre au cours du temps en utilisant une mesure de corrélation. L'approche de suivi temporel basé région a été largement étudiée [Wren00].

Par exemple, Wren *et al.* ont exploré l'utilisation de caractéristiques de *blobs* 2D pour suivre au cours du temps un humain seul dans un environnement intérieur [Wren97b]. Dans leur système *Pfinder*, un corps humain est considéré comme un ensemble de quelques *blobs*, qui représentent des parties du corps humain telles que le visage, le torse et les quatre membres. Pendant ce temps, à la fois le fond et le corps humain sont modélisés sous forme de distributions Gaussiennes. Finalement, les pixels appartenant au corps humain sont assignés aux

différents *blobs* des parties du corps selon un critère de log-vraisemblance. Par conséquent, en suivant temporellement chaque *blob*, la personne en mouvement est suivie avec succès.

Dans les travaux de Haritaoglu *et al.*, le système W^4 de vidéosurveillance utilise une échelle en niveaux de gris ou infrarouge sur des séquences vidéo [Haritaoglu98]. W^4 ne se sert pas des informations de couleur, au lieu de cela, il emploie une combinaison d'analyse de forme, avec des tests de superposition d'occupation spatiale, et du suivi temporel de région et un appariement de bords de silhouettes avec une estimation récursive par moindres carrés.

McKenna *et al.* proposent une méthode de soustraction du fond qui combine des informations de couleur et de gradient pour gérer efficacement les ombres et les informations invalides de couleur dans une segmentation de mouvement [McKenna00b]. Le processus de suivi temporel est ensuite réalisé à un triple niveau d'abstraction : régions, personnes et groupes d'individus. Chaque région pouvant se séparer ou fusionner possède une boîte englobante. Une personne est composée d'une ou plusieurs régions groupées ensemble à la condition de respecter certaines contraintes liées à la structure géométrique d'un corps humain, et un groupe d'individus consiste en plusieurs personnes groupées ensemble. Par conséquent, en utilisant le suivi temporel de régions et un modèle d'apparence individuel basé sur la couleur, ils réussissent à suivre temporellement plusieurs individus, même pendant les occultations.

3.1.2 Suivi temporel basé contour

Le suivi temporel basé sur les modèles de contour, ou *snakes*, vise à extraire directement la forme de l'objet. L'idée est d'obtenir une représentation du contour englobant l'objet et de continuer à le mettre à jour dynamiquement au cours du temps. Le suivi temporel basé contour a été intensivement étudié durant les quelques dernières années.

Isard et Blake, par exemple, ont adopté l'équation différentielle stochastique pour décrire un modèle de mouvement complexe, et combinent cette approche avec des modèles de référence déformables pour le suivi temporel [Isard96].

Le travail de Paragios et Deriche présente un cadre d'études variationnel pour détecter et suivre au cours du temps des objets en mouvement dans les séquences d'images [Paragios00]. La détection de mouvement est effectuée dans un cadre statistique, où la fonction de densité de différences inter-images observée est approchée en utilisant un modèle de mélange. Ce modèle comporte deux composantes, celle du fond et celle des objets mobiles. Les problèmes de détection et de suivi temporel sont alors définis dans un cadre commun qui emploie une fonction objective de contour géodésique actif. Cette fonction est minimisée selon une méthode de descente du gradient, dans laquelle un flot déforme le contour initial vers le minimum de la fonction, selon des forces internes et externes définies à partir de l'image. En utilisant le schéma de formulation d'ensemble de niveau, des courbes complexes peuvent être détectées et suivies temporellement.

Peterfreund a exploré un modèle de contour actif basé sur le filtrage de Kalman pour suivre au cours du temps des objets non rigides en mouvement comme des personnes dans un espace 2D de position et de vitesse [Peterfreund99]. Le modèle emploie des mesures de potentiel d'une image basée gradient et de flot optique le long du contour comme mesures du système. En parallèle, afin d'améliorer la robustesse au bruit et aux occultations, un mécanisme de détection basé sur un flot optique est proposé.

Contrairement au suivi temporel basé région, l'avantage de posséder une représentation basée sur les contours actifs est une réduction relative de la complexité calculatoire. Cependant cette approche requiert une bonne initialisation. S'il était possible d'obtenir d'une

quelconque façon une bonne initialisation de multiples contours actifs pour chaque objet en mouvement, alors un suivi temporel serait possible même en présence d'occultations partielles. Mais l'initialisation est une étape assez difficile, spécialement dans le cas d'objets complexes articulés.

3.1.3 Suivi temporel basé caractéristique

Si l'on abandonne l'idée de suivre au cours du temps les objets dans leur entier, cette approche de suivi temporel utilise des caractéristiques comme des points ou des lignes de l'objet pour réaliser le suivi. Son bénéfice est que, même en présence d'occultation partielle, quelques unes des caractéristiques restent visibles. Le suivi temporel basé caractéristique comprend l'extraction et l'association de caractéristiques. Les caractéristiques bas-niveau comme les points sont plus faciles à extraire que des caractéristiques de plus haut niveau comme les lignes ou les *blobs*. Il y a donc ici un compromis à faire entre la complexité des caractéristiques à extraire, la précision et la robustesse du suivi temporel.

Un exemple de suivi temporel basé sur des points caractéristiques est décrit dans les travaux de Polana et Nelson [Polana94]. Dans ces travaux, une personne est entourée par une boîte englobante rectangulaire, dont le centre est sélectionné comme point caractéristique pour le suivi temporel. Même quand une occultation partielle survient entre deux individus pendant le suivi temporel, du moment que la vitesse des centres peut être effectivement déterminée, le suivi temporel est efficace.

Le système de suivi temporel développé par Segen et Pingali utilise en plus les coins des silhouettes en mouvement comme points caractéristiques à suivre, et ces points sont associés en utilisant une mesure de distance basée sur les positions et les courbures des points entre images consécutives [Segen96].

Le suivi temporel de points et de lignes caractéristiques basé sur le filtrage de Kalman a été beaucoup étudié. Dans le travail de Jang et Choi [Jang00], un modèle de référence actif qui résume les caractéristiques de structure et de région d'un objet, est construit dynamiquement sur les informations de forme, de texture, de couleur et de frontières de la région. En utilisant une estimation de mouvement basée sur un filtrage de Kalman, le suivi temporel d'un objet en mouvement non rigide peut être réalisé par minimisation d'une fonction d'énergie caractéristique pendant le processus d'association entre les observations et les estimations.

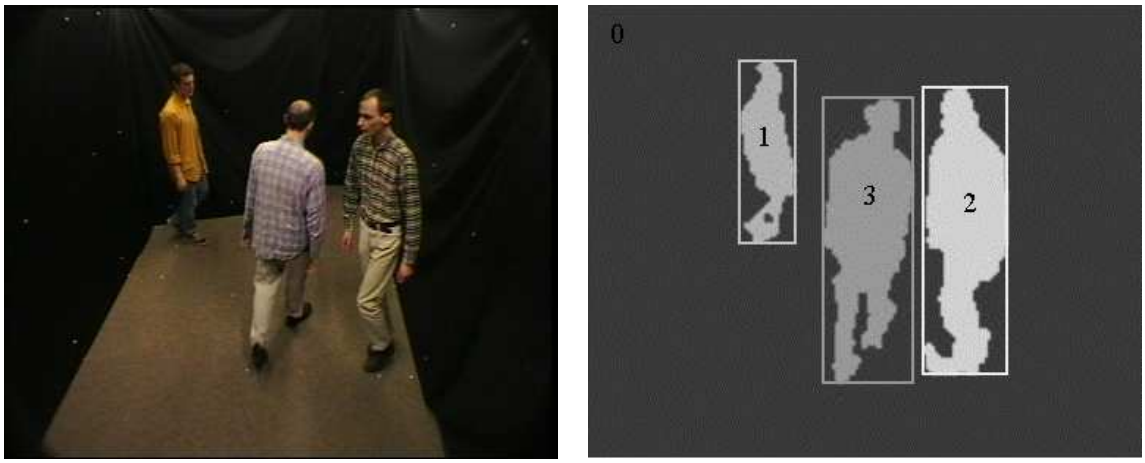
3.2 Suivi temporel basé sur l'intersection de boîtes

Après avoir présenté un état de l'art sur différentes méthodes destinées à réaliser un suivi temporel, nous allons maintenant présenter une méthode de suivi temporel basée sur l'intersection de boîtes. Cette méthode a été développée pour être très rapide et simple et est basée sur l'hypothèse que les **mouvements** des objets sur deux images consécutives sont **de faible amplitude**, notamment grâce à une cadence d'acquisition de 30 images/s.

3.2.1 Introduction

Cette partie décrit un algorithme de suivi temporel d'objets en mouvement présents dans une scène filmée par une caméra fixe. De même que l'étape de segmentation 2D spatio-temporelle décrite au chapitre 2, cette première étape de suivi temporel est générique et peut-être utilisée pour des *ROI* autres que des personnes. Une deuxième étape de suivi temporel,

plus complexe et plus précise, développée spécifiquement pour des personnes, sera présentée dans un chapitre ultérieur (cf. chapitre 5).



(a) (b)

FIG. 3.1 – Résultats de segmentation. (a) image originale, (b) masques de segmentation avec étiquettes et boîtes englobantes rectangulaires.

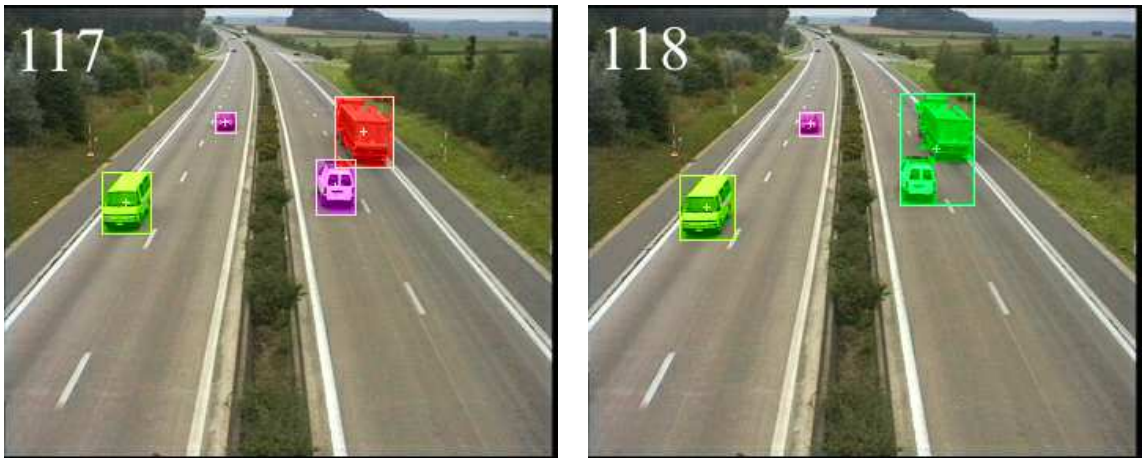
Pour cette seconde étape de traitement le point de départ est la fin de l'étape de segmentation 2D spatio-temporelle. La figure 3.1 illustre des résultats de segmentation dans le cas d'une séquence en environnement intérieur où les *ROI* sont des personnes avec en (a), l'image originale et en (b), les masques de segmentation obtenus avec leurs étiquettes et leurs boîtes englobantes rectangulaires.

Avant de décrire la première étape de notre suivi temporel, nous allons illustrer les phénomènes de réunion et de séparation temporelle qui représentent la difficulté inhérente du suivi au cours du temps et sont liés aux occultations. La figure 3.2 illustre une réunion temporelle d'objets et la figure 3.3 une séparation temporelle d'objets. Les deux figures sont extraites d'une séquence d'autoroute, en environnement extérieur, et présentent deux séries d'images consécutives. Sur chaque image, nous pouvons voir le produit des masques de segmentation en couleur des objets avec l'image originale, leurs centres et leurs boîtes englobantes rectangulaires.

Sur la figure 3.2, en (a) nous avons quatre objets segmentés, en (b) nous en avons seulement trois. Le camion et la camionnette à droite se sont réunis en un seul objet. Cela est illustré par la grande boîte englobante rectangulaire.

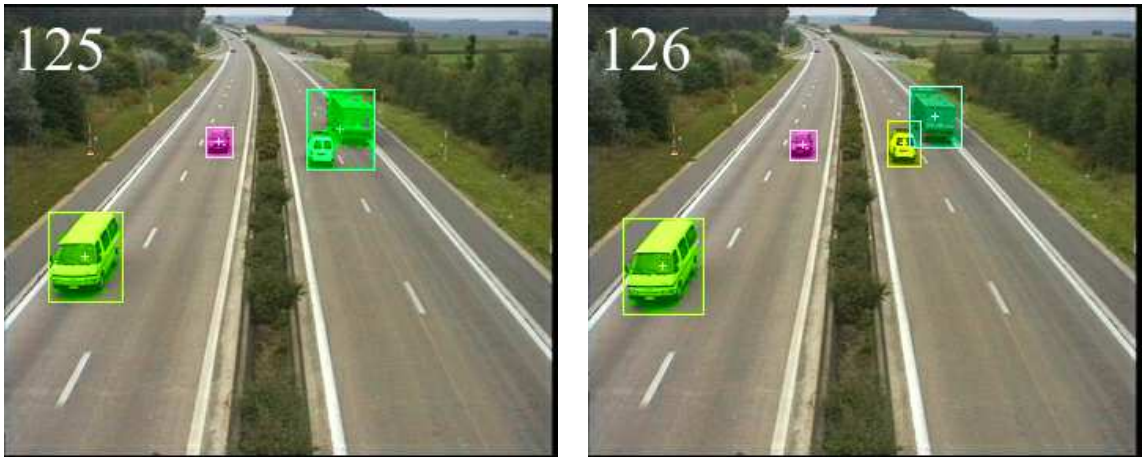
Sur la figure 3.3, une dizaine d'images après celles de la figure 3.2, en (a) nous avons trois objets segmentés, en (b) nous en avons de nouveau quatre. Le camion et la camionnette se sont séparés en deux objets. Ces objets ont chacun leur boîte englobante rectangulaire.

La première étape du suivi temporel présentée ici a été développée pour être très rapide et simple. Elle est basée sur le **calcul d'intersection de boîtes**. Cette méthode ne gère pas les problèmes d'occultation entre objets mais permet la détection de séparation ou de réunion temporelle (cf. figures 3.2 et 3.3). Le suivi temporel a pour but d'établir une liaison temporelle entre les objets de l'image courante (instant t) et ceux de l'image précédente (instant $t-1$). Ce lien temporel est appelé numéro d'identification *ID (Identification Data)*. Un objet est correctement suivi d'une image à la suivante si son *ID* est le même dans les deux images.



(a) (b)

FIG. 3.2 – Réunion temporelle d'objets. Images (a) 117 et (b) 118.



(a) (b)

FIG. 3.3 – Séparation temporelle d'objets. Images (a) 125 et (b) 126.

Dans le cas de réunions temporelles d'objets, l'objet résultant est suivi comme un tout de la même manière que si l'objet était seul. Au niveau des *ID*, chaque nouvel objet détecté se voit attribuer un *ID* unique. Tant que cet objet est correctement suivi au cours du temps, son *ID* est inchangé. Dans les cas où cet objet participe à une séparation ou à une réunion temporelle, les *ID* peuvent être gérés de différentes façons. Nous reviendrons sur la gestion des *ID* plus loin dans ce chapitre (cf. partie 3.2.4).

3.2.2 Méthode pour la première étape du suivi temporel

Le principe du suivi temporel d'objets dans cette partie s'appuie sur l'hypothèse suivante : comme la cadence d'acquisition de la caméra est de 30 images/s nous pouvons supposer que les objets ont un **mouvement de faible amplitude** sur des images consécutives, c'est-à-dire qu'il y a toujours une zone de superposition non nulle entre un objet au temps t et ce même objet au temps $t-1$. *A fortiori*, il y a toujours une zone de superposition non nulle entre les boîtes de cet objet au temps t et celles au temps $t-1$.

Par conséquent, un suivi temporel est possible en ne considérant que les intersections entre les boîtes détectées au temps $t-1$ et les boîtes détectées au temps t. Nous n'utilisons pas de compensation de mouvement pour les boîtes parce que cela nécessiterait une estimation de mouvement sur les pixels des objets qui serait coûteuse en temps de calcul. L'intersection de deux boîtes est définie comme la surface de superposition entre les boîtes. Nous ne calculons pas la surface de superposition au niveau des masques des objets mais bien au niveau des boîtes. Pour les calculs d'intersection entre les boîtes des objets, il est possible de choisir n'importe quel type de boîtes, englobantes rectangulaires, quadrangulaires, octogonales ou par axes principaux. Néanmoins, pour des raisons de temps de calcul, nous avons choisi de faire ces calculs sur les boîtes englobantes rectangulaires.

La première étape du suivi temporel, qui est donc basée sur les intersections entre les boîtes englobantes rectangulaires des objets détectés sur deux images consécutives (temps $t-1$ et t), résulte de la combinaison d'une phase de suivi temporel vers l'avant et d'une phase de suivi temporel vers l'arrière. Ces deux phases sont effectuées afin d'obtenir une association cohérente entre les objets issus d'images consécutives.

3.2.2.1 Suivi temporel vers l'avant (*forward tracking*)

Pour la phase de suivi temporel vers l'avant, nous considérons les successeurs potentiels de chaque objet détecté au temps $t-1$ en calculant les intersections entre la boîte de cet objet et l'ensemble des boîtes détectées au temps t. Dans le cas de successeurs multiples, c'est-à-dire s'il existe plusieurs boîtes au temps t qui ont une intersection non nulle avec celle au temps $t-1$, les objets correspondants à ces boîtes sont triés par ordre décroissant de surface d'intersection dans une **liste de successeurs**. Nous dénommons alors "successeur le plus probable" le premier élément de cette liste, qui correspond à l'objet dont la boîte a la plus grande intersection en suivi temporel vers l'avant. En effet, selon notre hypothèse de mouvement de faible amplitude, par rapport à la boîte englobante rectangulaire d'un objet au temps $t-1$, l'objet au temps t dont la boîte englobante rectangulaire a la plus grande zone de superposition a de fortes chances d'être le même objet.

Si un objet a une liste vide de successeurs, c'est-à-dire aucun successeur, c'est que cet objet à disparu à l'instant t. En faisant le postulat que la segmentation est satisfaisante, les objets disparaissent en quittant la scène.

3.2.2.2 Suivi temporel vers l'arrière (*backward tracking*)

Pour la phase de suivi temporel vers l'arrière, la procédure est semblable : nous considérons les prédécesseurs potentiels de chaque objet détecté au temps t. Le ou les prédécesseurs sont triés par ordre décroissant de surface d'intersection dans une **liste de prédécesseurs**. Le premier élément de cette liste est appelé "prédécesseur le plus probable".

Si un objet a une liste vide de prédécesseurs, c'est-à-dire aucun prédécesseur, c'est qu'il s'agit d'un nouvel objet apparu à l'instant t. Si la segmentation a échoué et qu'un objet présent dans la scène n'a pas été segmenté, il sera considéré comme un nouvel objet s'il réapparaît.

3.2.2.3 Combinaison des suivis temporels vers l'avant et vers l'arrière

Comme chaque prédécesseur d'un objet donné peut avoir plusieurs successeurs, il est nécessaire de définir une méthode de parcours dans ces listes multiples de successeurs et de prédécesseurs. Cette méthode va permettre de trouver un résultat cohérent pour le suivi temporel.

Cette méthode de parcours des listes triées de successeurs et de prédécesseurs est basée sur la notion de **rang** dans une liste. Le tableau présenté table 3.1 montre l'ordre de parcours des listes triées de successeurs et de prédécesseurs, les lignes représentant le rang du successeur et les colonnes le rang du prédécesseur :

Tab. 3.1 – Méthode de parcours des listes triées de successeurs et de prédécesseurs.

Rang du prédécesseur →

Rang du successeur ↓	1	2	3	4	...
1	1	2_a	4_a	7_a	
2	2_b	3	5_a	8_a	
3	4_b	5_b	6	...	
4	7_b	8_b	...	...	
...					

La lecture de ce tableau se fait de la manière suivante : tant que nous n'obtenons pas une combinaison positive, c'est-à-dire une cohérence des résultats, nous regardons successivement, dans l'ordre :

1. le 1^{er} successeur du 1^{er} prédécesseur (1) ;
2. à la fois :
   - le $2^{ème}$ successeur du 1^{er} prédécesseur (2_a) et
   - le 1^{er} successeur du $2^{ème}$ prédécesseur (2_b) ;
3. le $2^{ème}$ successeur du $2^{ème}$ prédécesseur (3) ;
4. à la fois :
   - le $3^{ème}$ successeur du 1^{er} prédécesseur (4_a) et
   - le 1^{er} successeur du $3^{ème}$ prédécesseur (4_b) ;
5. ...

Quand un couple prédécesseur-successeur cohérent est trouvé, l'*ID* du prédécesseur est transmis à son successeur. Dans le cas où l'on regarde deux possibilités en même temps et que les deux donnent des résultats cohérents, le prédécesseur choisi est celui qui a la plus grande intersection.

Il est toujours possible, de trouver un couple prédécesseur-successeur cohérent grâce à cette méthode, sauf si la boîte englobante rectangulaire de l'objet considéré au temps t n'a aucune intersection avec l'ensemble des boîtes englobantes rectangulaires des objets détectés au temps $t-1$, mais dans ce cas c'est un nouvel objet, il se voit donc attribuer un nouvel *ID*. En effet, à partir du moment où un objet détecté au temps t a une intersection non nulle avec un objet détecté au temps $t-1$, ces deux objets seront forcément, par définition, dans les listes respectives de successeurs et de prédécesseurs. En regardant suffisamment loin dans les rangs, un résultat cohérent sera trouvé.

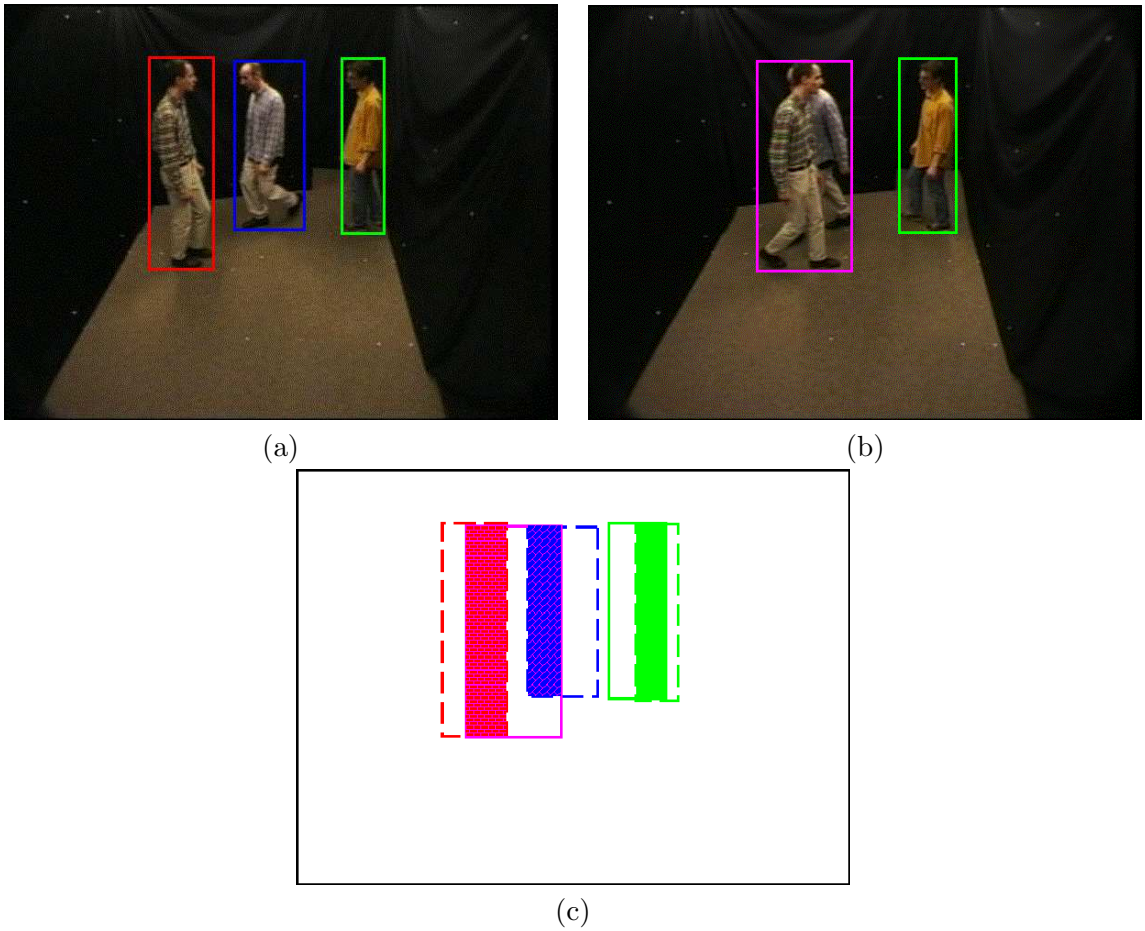


FIG. 3.4 – Illustration du suivi temporel pour des personnes. (a) image à l'instant $t-1$, (b) image à l'instant t, (c) image fictive d'intersection.

La figure 3.4 illustre les phases de suivi temporel vers l'avant et vers l'arrière pour une séquence en environnement intérieur comportant trois personnes. En (a), trois objets sont segmentés, et en (b), seulement deux objets sont segmentés. L'image (c) illustre l'image fictive d'intersection des boîtes englobantes rectangulaires. La combinaison des suivis amène un

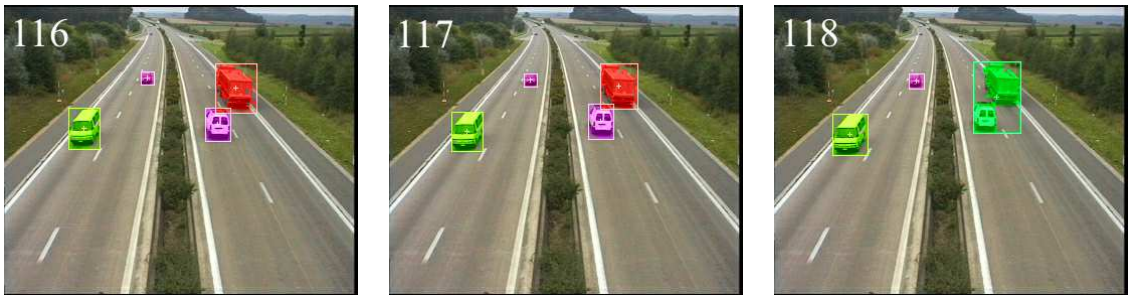
suivi temporel correct pour l'individu sur la droite (l'objet a un unique successeur, ce successeur ayant lui-même un unique prédécesseur). Pour les personnes à gauche la phase de suivi temporel vers l'arrière apporte deux prédécesseurs et la phase de suivi temporel vers l'avant un unique successeur. L'objet résultant, un groupe de deux individus dans le cas présent, sera suivi temporellement comme un tout jusqu'à ce qu'il se sépare ou qu'il disparaisse.

3.2.3 Détection de séparation et de réunion temporelle

La première étape du suivi temporel permet de façon très simple de détecter les séparations et les réunions temporelles de *ROI*, grâce aux listes triées de successeurs et de prédécesseurs.

3.2.3.1 Réunion temporelle

Si un objet, à l'instant t, a une liste de prédécesseurs qui contient plusieurs éléments, c'est un objet qui **peut** résulter d'une réunion temporelle. En effet, ce n'est pas parce qu'un objet a plusieurs prédécesseurs qu'il résulte forcément d'une réunion temporelle. La figure 3.5 illustre cette affirmation avec des images consécutives extraites de la séquence d'autoroute en environnement extérieur. Sur les deux premières images de la figure 3.5, nous pouvons voir que le camion et la camionnette à droite ont chacun deux successeurs (suivi temporel vers l'avant) et deux prédécesseurs (suivi temporel vers l'arrière). Or ces deux objets ne se sont pas réunis, il n'y a donc pas de réunion temporelle. En revanche, les deux dernières images de la figure 3.5 illustrent une véritable réunion temporelle. Le camion et la camionnette se sont réunis en un seul objet.



(a) (b) (c)

FIG. 3.5 – Images (a) 116, (b) 117 et (c) 118.

Nous pouvons déduire de ces observations une définition pour la réunion temporelle. Pour qu'une **réunion temporelle** soit détectée, il faut qu'au moins deux objets à l'instant $t-1$ aient un même et unique successeur au temps t.

La figure 3.4 illustre, pour les personnes à gauche, la détection d'une réunion temporelle. Le groupe a deux prédécesseurs qui ont chacun le même et unique successeur.

3.2.3.2 Séparation temporelle

De façon similaire à la détection de réunion temporelle, un objet, à l'instant $t-1$, ayant une liste de successeurs qui contient plusieurs éléments, est un objet qui **peut** donner lieu à une séparation temporelle. En suivant le même raisonnement qu'au paragraphe précédent, on

peut définir une séparation temporelle. Pour qu'une **séparation temporelle** soit détectée, il faut qu'au moins deux objets à l'instant t aient un même et unique prédécesseur au temps $t-1$.

3.2.4 Gestion des numéros d'identification *ID*

Dans notre système, les *ID* sont les liens temporels du suivi. Quand un nouvel objet est détecté, il se voit attribuer un *ID* unique. Un nouvel objet est détecté quand sa liste de prédécesseurs est vide. Tant que cet objet est correctement suivi, son *ID* est inchangé. Quand des séparations et / ou des réunions temporelles surviennent, il faut alors décider comment gérer les *ID*, c'est-à-dire comment attribuer les numéros d'identification.

Dans notre système, deux choix sont possibles :

1. n'attribuer que des nouveaux *ID* :
   - à tous les objets résultant d'une séparation temporelle ;
   - à l'objet résultant d'une réunion temporelle.
2. conserver l'un des *ID* et, éventuellement, attribuer des nouveaux *ID* :
   - en cas de séparation temporelle, l'*ID* du prédécesseur est transmis à son successeur le plus probable, le ou les autres ayant des nouveaux *ID* ;
   - en cas de réunion temporelle, l'*ID* du prédécesseur le probable est transmis à son successeur.

Le premier choix ne tolère aucune séparation ou réunion temporelle. Le second permet de suivre éventuellement au cours du temps le plus gros objet (du point de vue de la caméra) dans la scène. La partie 3.5 illustrera les deux choix possibles.

3.3 Suivi temporel de personnes

Nous allons maintenant analyser les performances de la première étape du suivi temporel lorsque les *ROI* sont des êtres humains.

Après l'étape de segmentation, nous pourrions supposer que chaque boîte englobante rectangulaire contient une personne seule ou un groupe, dans le cas d'une réunion temporelle. Mais il peut aussi arriver, plus rarement, dans le cas d'une mauvaise segmentation, que le corps d'un individu soit segmenté en plusieurs parties (main(s), bras, visage, pied(s), jambe(s)). Une boîte englobante rectangulaire peut donc contenir finalement une partie du corps, une personne seule ou un groupe de personnes. Pour faciliter les explications, notons ces trois types d'objets :

- PC : Partie du Corps ;
- PS : Personne Seule ;
- GP : Groupe de Personnes.

L'hypothèse n°2 de notre système assure que **chaque personne entre seule dans la scène filmée**. Donc quand un nouvel objet apparaît, nous pouvons supposer que cet objet est une PS. Quand surviennent des réunions ou des séparations temporelles, nous pouvons regarder quels sont les types d'objets possibles. Les tableaux suivants (tables 3.2 et 3.3) illustrent respectivement le type d'objet pouvant résulter d'une réunion temporelle entre deux objets et les types d'objet possibles pour les objets résultant d'une séparation temporelle.

Pour une réunion temporelle, à part quand deux PC se réunissent, il n'y a pas de doute, l'objet résultant d'une réunion temporelle a un type clairement défini. Pour deux PC se

Tab. 3.2 – Types d'objet pour une réunion temporelle.

Réunion temporelle	PC	PS	GP
PC	PC / PS	PS	GP
PS	PS	GP	GP
GP	GP	GP	GP

réunissant, nous pouvons avoir soit encore une PC, soit une PS, dans le cas où cette personne s'est retrouvé séparée en deux PC qui se réunissent à nouveau.

Tab. 3.3 – Types d'objet pour une séparation temporelle.

Séparation temporelle	PC	PS	GP
	PC	PC / PS	PC / PS / GP

Pour une séparation temporelle, il y a toujours un doute sur les types possibles pour les objets résultant de cette séparation, à part quand une PC se sépare et donne naissance à d'autres PC.

Même en faisant l'hypothèse que les mauvaises segmentations ne surviennent que rarement, et donc en ne considérant que les objets de type PS ou GP, pour trois individus présents dans la scène, nous pouvons avoir les cas suivants :

- trois PS ;
- une PS et un GP de 2 PS ;
- un GP de 3 PS.

Le cas difficile est la détermination des types d'objet dans le second cas. Avec des tests simples sur la surface des objets segmentés, il serait possible de différencier le groupe et la personne seule.

Néanmoins, pour plus de trois personnes présentes dans la pièce, par exemple quatre individus, il devient très difficile de déterminer si un GP de 4 PS s'est séparé en deux GP de 2 PS chacun ou s'il s'est séparé en un GP de 3 PS et une PS.

Il faudrait alors que le suivi temporel soit plus complexe pour pouvoir distinguer les types d'objet résultant.

3.4 Données bas-niveau extraites

3.4.1 Numéros d'identification *ID*

Les numéros d'identification *ID* sont la principale information extraite lors de cette étape. Ce sont eux qui définissent les liens temporels entre les objets de l'image $t-1$ et ceux de l'image t. À chaque objet on associe un unique *ID*, au moment de sa détection. L'*ID* d'un objet suivi avec succès est inchangé de son apparition jusqu'à sa disparition. Si cet objet participe à une réunion temporelle ou donne lieu à une séparation temporelle, son *ID* peut changer ou non suivant le choix fait pour la gestion des *ID* (cf. partie 3.2.4).

Dans le cas où une seule personne est présente dans la scène, cette personne devrait avoir un unique *ID* pendant toute la séquence vidéo, puisqu'il ne peut y avoir de réunion ou de séparation temporelle avec une autre personne.

3.4.2 Trajectoires

Les trajectoires des objets vidéo suivis au cours du temps sont une information qui découle de l'étape de suivi temporel. Les trajectoires sont les positions successives des centres des boîtes englobantes rectangulaires. Il est possible aussi d'obtenir les trajectoires des centres des autres boîtes ou des centres de gravité. La structure de données utilisée lors du suivi temporel est définie sur un certain nombre d'images, une dizaine par défaut, c'est une sorte de fenêtre glissante qui possède une mémoire des résultats de suivi successifs sur la durée de la fenêtre. Les numéros d'identification *ID* font le lien temporel entre les objets détectés sur deux images consécutives. En remontant la structure de données jusqu'au début de la fenêtre, nous pouvons donc obtenir les trajectoires d'un objet d'*ID* particulier.

3.4.3 Autres données bas-niveau

Les autres données bas-niveau extraites lors de cette première étape du suivi temporel de base sont les informations de séparation et de réunion temporelle et le nombre d'occurrences.

Les informations de séparation et de réunion temporelle sont directement accessibles en regardant les listes triées de successeurs et de prédécesseurs. Nous rappelons les définitions d'une séparation et d'une réunion temporelle :

- pour qu'une **séparation temporelle** soit détectée, il faut qu'au moins deux objets à l'instant t aient un même et unique prédécesseur au temps $t-1$.
- pour qu'une **réunion temporelle** soit détectée, il faut qu'au moins deux objets à l'instant $t-1$ aient un même et unique successeur au temps t.

La dernière donnée calculée est le nombre d'occurrences, c'est un compteur qui donne le nombre d'images consécutives pour lesquelles un objet a été correctement suivi.

3.4.4 Résumé

En résumé, lors de l'étape de suivi temporel de base, nous disposons des données suivantes pour chaque objet :

- le numéro d'identification *ID* ;
- la trajectoire ;
- les informations de séparation et de réunion temporelle ;
- le nombre d'occurrences.

3.5 Résultats

Nous allons présenter des images illustrant les résultats obtenus pour la première étape du suivi temporel. Les images présentées montrent des points importants des résultats de suivi temporel et se regardent dans le sens de lecture classique. Elles sont composées du produit des masques de segmentation en couleur des objets avec les images originales, de leurs boîtes englobantes rectangulaires, de leurs centres et de leurs trajectoires. Pour des raisons de visualisation, les *ID* définissent chacun une unique couleur utilisée pour le dessin

des données. Ces couleurs sont déterminées de façon aléatoire. Les trajectoires illustrent le fait que les objets ont été correctement suivis (même couleur).

La figure 3.6 illustre les résultats de suivi temporel pour la séquence d'autoroute en environnement extérieur. Pour la gestion des *ID*, le choix a été fait d'attribuer un nouvel *ID* en cas de réunion ou de séparation temporelle. Le camion et la camionnette sont donc mal suivis dès qu'ils sont trop proches. Les autres objets sont tous correctement suivis.

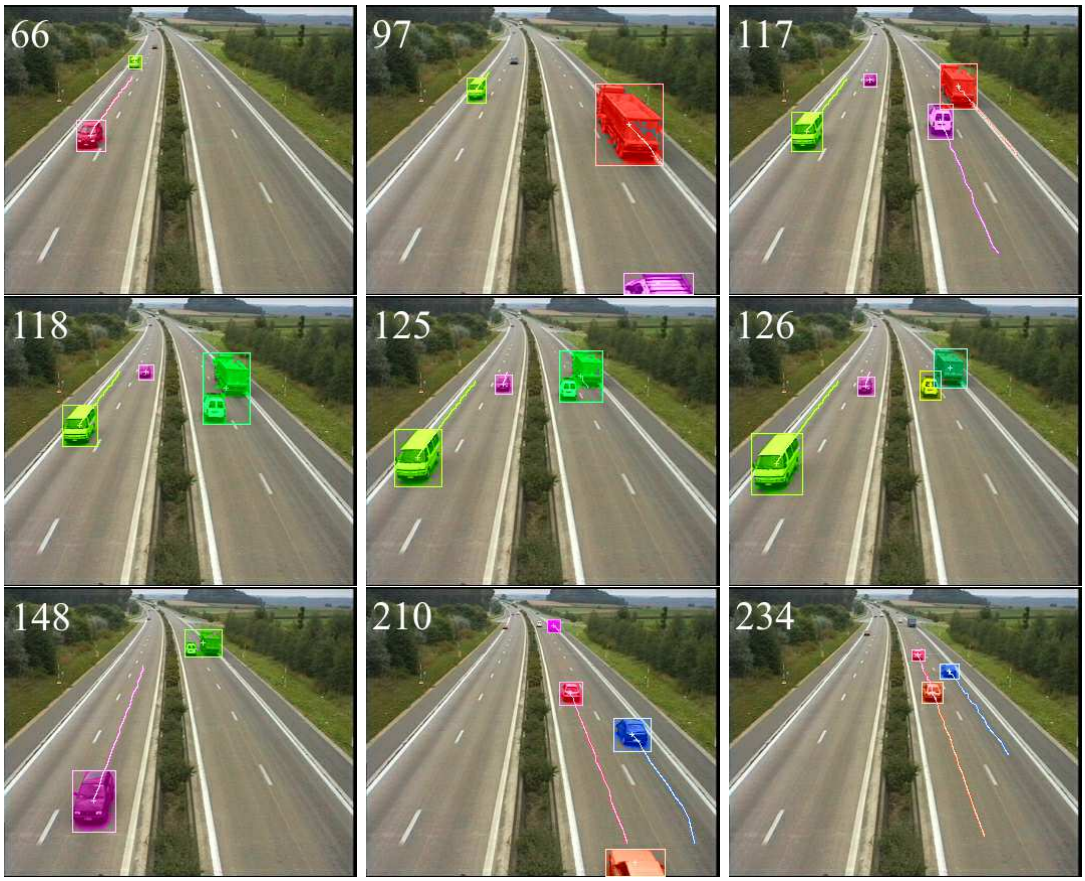


FIG. 3.6 – Environnement extérieur : suivi temporel avec nouvel *ID*.

La figure 3.7 illustre, sur les mêmes images, l'autre choix pour la gestion des *ID*, où l'on conserve l'*ID* de l'objet le plus probable en cas de réunion ou de séparation temporelle. La différence principale avec la figure 3.6 est le fait que sur cette séquence, le camion à droite est correctement suivi malgré les nombreuses réunions et séparations temporelles avec la camionnette. Ceci est dû au fait que le camion est l'objet le plus probable puisqu'il est plus gros que la camionnette. Sa trajectoire comporte des discontinuités liés au fait que les centres sont distants (cf. images 118, 125, 126 et 148). Les autres objets, à part la camionnette, sont tous correctement suivis.

La figure 3.8 illustre les résultats obtenus pour une séquence vidéo en environnement intérieur où deux personnes se croisent, quand elles se réunissent en un groupe et quand ce groupe se sépare. Pour la gestion des *ID*, le choix a été fait d'attribuer un nouvel *ID* en cas de réunion ou de séparation temporelle. Les images illustrent successivement l'apparition

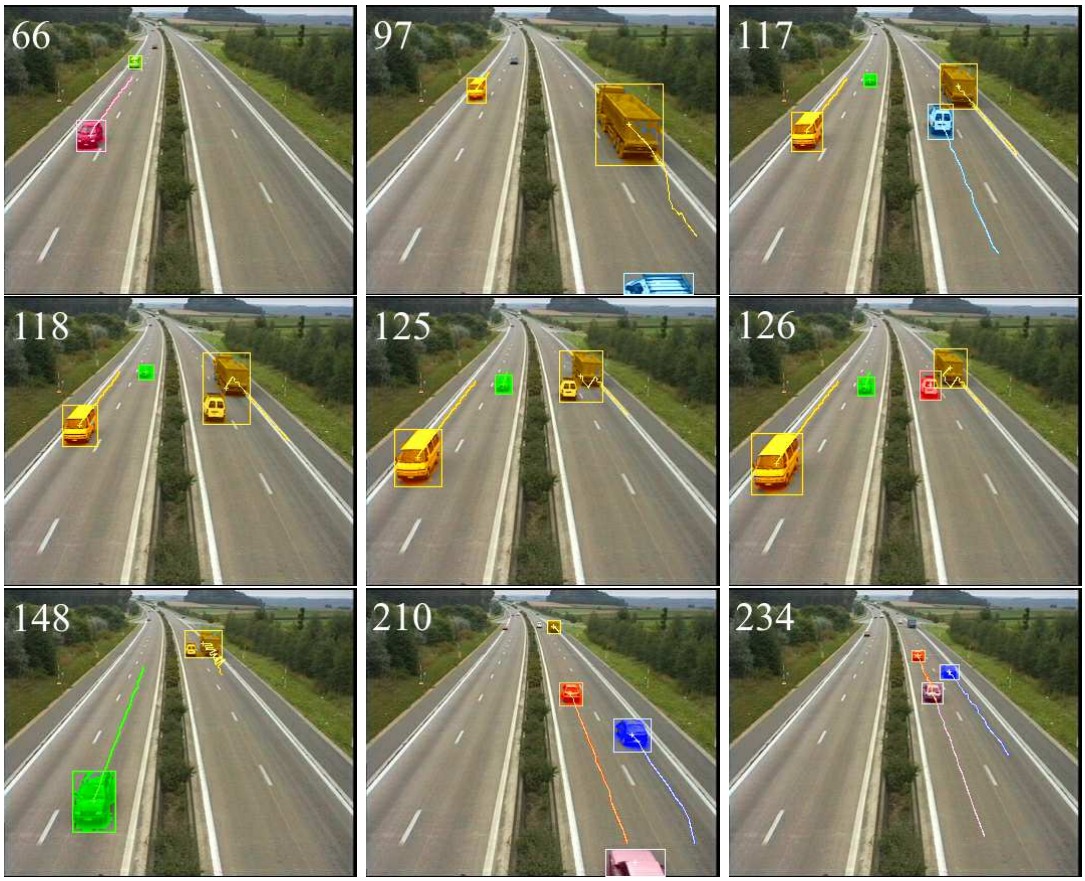


FIG. 3.7 – Environnement extérieur : suivi temporel avec conservation d'un *ID*.

des deux individus, un suivi temporel correct pour ces deux individus jusqu'à une réunion temporelle, donc un nouvel objet. Puis un suivi temporel correct pour le groupe de personnes jusqu'à une séparation temporelle en deux nouveaux objets, suivis de façon correcte jusqu'à leur disparition.

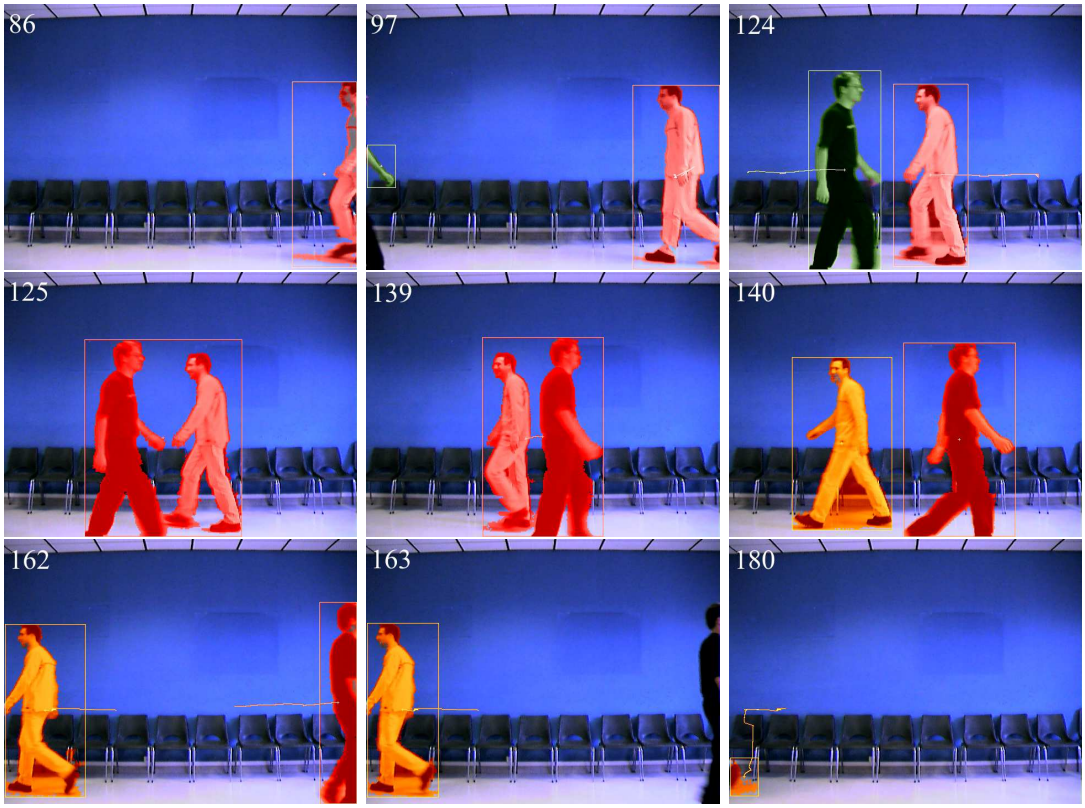


FIG. 3.8 – Environnement intérieur : suivi temporel avec nouvel *ID*.

La figure 3.9 illustre, sur les mêmes images, l'autre choix pour la gestion des *ID*, où l'on conserve l'*ID* de l'objet le plus probable en cas de réunion ou de séparation temporelle. La différence avec la figure 3.8 est le fait que l'individu de gauche qui est le plus proche de la caméra est suivi de façon correcte tout au long de la séquence. L'autre personne est correctement suivie avant la réunion temporelle et après la séparation temporelle jusqu'à sa disparition.

3.6 Avantages, limitations et cadences de traitement

La première étape du suivi temporel, basée sur l'intersection de boîtes englobantes rectangulaires, présente des avantages et une limitation. Ses principaux avantages sont sa rapidité et sa simplicité. La méthode présentée permet en outre la détection des séparations et des réunions temporelles.

Néanmoins, cette méthode de suivi temporel possède aussi une limitation. Même si elle détecte les séparations et les réunions temporelles, qui sont les conséquences de phénomènes d'occultation, elle ne les corrige pas. En effet, elle ne peut garder séparés des objets qui

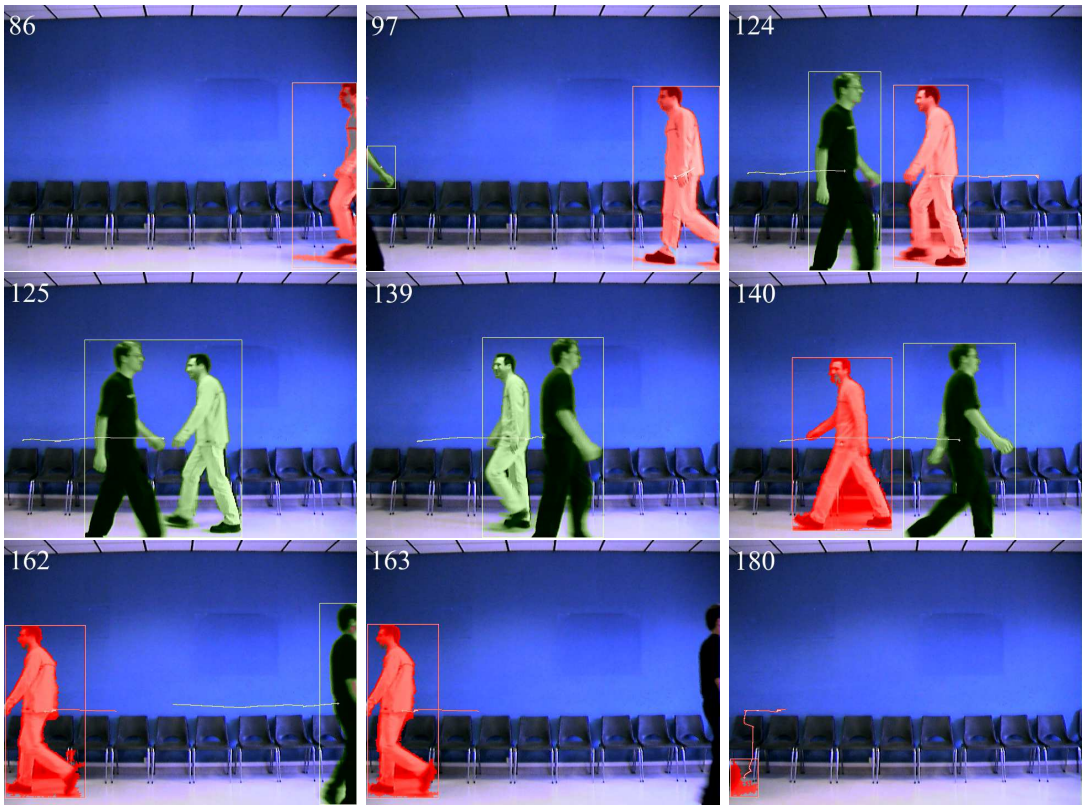


FIG. 3.9 – Environnement intérieur : suivi temporel avec conservation d'un *ID*.

ont participé à une réunion temporelle de même qu'elle ne peut réunir les différentes parties d'un objet qui a subi une séparation temporelle. Elle ne gère donc pas en particulier les difficultés propres aux phénomènes d'occultation. Ceci est lié au fait que cette méthode est extrêmement simple. Nous avons vu que, si nous voulons développer une méthode de suivi temporel générique, nous ne pouvons pas, à moins d'utiliser un suivi plus complexe, déterminer les types d'objet résultant principalement de séparations temporelles (cf. partie 3.3).

Dans ce mémoire, les résultats présentés en environnement intérieur ont été obtenus sur des séquences vidéo comportant, en grande majorité, une ou deux personnes, et sur quelques séquences comportant trois ou quatre personnes. Nous considérons que c'est suffisant pour les applications visées (IHM, vidéosurveillance en intérieur et applications de réalité mixte). Dans le cas où plus de trois personnes sont présentes dans le champ de la caméra, quelques algorithmes nécessiteraient d'être revus et améliorés.

Au niveau des pourcentage de temps de calcul et des cadences de traitement atteintes, la table 3.4 résume les résultats obtenus. Dans cette table, la première étape du suivi temporel est notée (1/2).

Tab. 3.4 – Pourcentages de temps de calcul et cadences de traitement pour la première étape du suivi temporel.

Segmentation	Champs aléatoires de Markov		Optimisée en vitesse	
Résolution d'image	320×240	640×480	320×240	640×480
Acquisition	0.2%	0.4%	5.9%	11.2%
Segmentation	89.4%	97.9%	83.6%	86.3%
Suivi temporel (1/2)	10.4%	1.7%	10.5%	2.5%
Cadences de traitement	**8.65** images/s	**2.07** images/s	**266** images/s	**61** images/s

Nous pouvons constater que la méthode pour la première étape du suivi temporel ne réduit que très peu les cadences de traitement par rapport à celles de l'étape de segmentation (cf. table 2.1, page 51). En particulier, les cadences de traitement atteintes avec la segmentation optimisée en vitesse sont encore bien au delà de la cadence vidéo (30 images/s). Cette première étape du suivi temporel est donc très rapide.

3.7 Conclusion

Ce deuxième chapitre a détaillé une méthode pour la première étape du suivi temporel d'objets présents dans une scène filmée par une caméra fixe. Cette méthode est extrêmement simple et il est évident qu'il existe des méthodes probablement plus performantes mais aussi beaucoup plus compliquées et plus lourdes. L'objectif de cette étape de traitement était de réaliser, à partir des informations issues de la segmentation 2D spatio-temporelle, un suivi temporel le plus rapide possible.

Nous avons décrit notre méthode de suivi temporel qui est basée sur un calcul d'intersection de boîtes englobantes rectangulaires. Puis les données bas-niveau extraites lors de cette première étape du suivi temporel ont été exposées, notamment les numéros d'identification *ID* et les trajectoires des objets. Ensuite nous avons donné quelques images de résultats de suivi temporel issus du traitement de séquences vidéo variées. Après cela, nous avons présenté

les avantages, les limitations et les cadences de traitement atteintes pour cette première étape du suivi temporel. Ses avantages principaux sont sa rapidité et le fait qu'elle permette de détecter les réunions et les séparations temporelles. La limitation principale est le fait qu'elle ne gère pas les occultations. Comme l'étape de segmentation 2D spatio-temporelle, elle peut être utilisée en environnement intérieur ou extérieur. Elle a été testée pour le suivi temporel de véhicules sur une autoroute, sans tenir compte des résultats de séparation et de réunion temporelle. Cet aspect générique fait de cette étape de traitement le lien entre diverses applications possibles et les applications visées où nous nous focalisons plus sur les personnes.

Nous allons maintenant présenter quelques perspectives pour cette seconde étape de traitement. Tout d'abord il serait utile de réaliser cette première étape du suivi en calculant les intersections entre les masques de segmentation des objets et non entre leurs boîtes englobantes rectangulaires. Les résultats seraient sans doute un peu plus précis et le suivi plus robuste. Ensuite, avec une approche basée contour, nous pourrions améliorer le suivi temporel au niveau des occultations en tentant de détecter et de conserver les frontières entre les objets même en cas de réunion temporelle.

Les autres perspectives pour cette étape de traitement sont similaires à celles pour l'étape de segmentation 2D spatio-temporelle. Outre l'approche basée contour dont nous avons parlé, une autre approche, basée sur les *blobs* serait aussi une perspective intéressante, ceci afin de voir si nous pouvons réaliser un suivi temporel, par exemple pour une personne, grâce au suivi temporel de l'ensemble des parties de son corps. Ceci permettrait éventuellement de comparer les résultats, dans un premier lieu, avec le système *Pfinder* dont c'est l'approche et ensuite avec ceux du chapitre suivant, qui consiste à localiser et à suivre au cours du temps le visage et les mains d'une personne.

Chapitre 4

Localisation et suivi temporel du visage et des mains

Sommaire

De nombreux articles sur l'analyse et l'interprétation du comportement humain se focalisent sur des *ROI* plus précises du corps humain, telles que le visage, les mains, les bras ou les jambes. De fait, quand on regarde une personne et que l'on interagit avec elle, notre regard est attiré d'abord par le visage, car c'est notre principal moyen de communication, puis par les mains (notre gestuelle) et ensuite par notre attitude corporelle globale. Nos expressions faciales et notre regard peuvent être très significatifs et représentatifs de nos émotions et de notre état d'esprit à l'instant présent (joie, tristesse, peur, surprise etc.).

Dans les applications multimédia qui nécessitent une interface homme-machine avancée, la détection et le suivi temporel du visage et / ou des mains sont souvent considérées comme des étapes préliminaires indispensables pour l'analyse et l'interprétation des gestes et des activités d'êtres humains. Concernant ces *ROI*, trois domaines de recherche se dégagent, selon le type d'application visée :

– l'étude du visage (seul) ;
– l'étude des mains (une ou les deux), à l'exclusion du visage ;
– l'étude conjointe du visage et des mains (une ou les deux).

Pour certaines applications, comme la visiophonie par exemple, la seule *ROI* est le visage, car c'est le point focal de la conversation. Lorsque l'on cherche à analyser les expressions faciales, un modèle simplifié, ou squelette d'expression (yeux, sourcils, bouche etc.) permet d'atteindre ce but [Hammal05]. En vidéosurveillance, le visage, et les yeux en particulier, peuvent servir pour vérifier l'identité de personnes [Peng05a]. Les applications qui ne considèrent que la main comme *ROI* peuvent être médicales ou d'IHM comme le contrôle d'objets graphiques par les gestes. D'autres applications encore combinent le visage et les mains, pour le langage des signes pour les sourds et les mal-entendants [Ong05], ou pour le LPC (Langage Parlé Complété) qui utilise, en plus de la lecture labiale, les positions de la main [Burger06] (cf. le projet ARTUS, présenté en annexe A.2).

Ces *ROI*, comme les individus dans leur entier, nécessitent d'être analysées (segmentées, puis éventuellement suivies au cours du temps), avant d'être interprétées. La majorité des approches possibles pour des *ROI* génériques peut être utilisée pour le visage et les mains, ainsi que d'autres approches, comme les modèles de corps humain, développées de façon plus spécifique pour les personnes (cf. chapitre 5). Toutes les difficultés vues précédemment pour les étapes de segmentation et de suivi temporel, notamment les occultations et les réunions temporelles, vont aussi survenir lors de cette étape de localisation et de suivi temporel du visage et des mains. De plus, par rapport aux approches proposées pour les étapes de traitement précédentes, certaines ne sont plus possibles pour le visage ou les mains. Par exemple, les mains peuvent se déplacer trop vite pour que l'on puisse envisager un suivi temporel par intersection de boîtes.

À partir de ce chapitre, les solutions proposées ont été développées de façon plus spécifique dans le cas de **personnes en mouvement dans un environnement intérieur**. Dans ce chapitre, nous cherchons à localiser et à suivre le visage et les mains, avec la distinction main droite / main gauche, de plusieurs individus dans la scène qui sont segmentées individuellement. Dans le cas d'un groupe de personnes, nous obtenons au mieux un visage, une main droite et une main gauche, sans être certains que ces *ROI* appartiennent au même individu. Nous commencerons par présenter un état de l'art sur les approches utilisées pour détecter le visage et les mains, et un état de l'art sur quelques espaces couleur utilisés pour réaliser une détection de peau par une approche couleur. Ensuite, nous détaillerons nos méthodes pour extraire les pixels de peau, localiser les positions initiales du visage et des mains et les suivre au cours du temps. Après cela, nous présenterons les données bas-niveau extraites

lors de cette étape de traitement puis nous illustrerons les résultats obtenus par des images issues de diverses séquences vidéo. Dans une dernière partie, nous décrirons les avantages, les limitations et les cadences de traitement pour cette étape de traitement avant de conclure en donnant quelques perspectives de développement ou d'amélioration possibles.

4.1 État de l'art sur la détection du visage et / ou des mains

Dans les applications qui nécessitent une interface homme-machine par le biais du traitement d'une séquence d'images, nous cherchons à obtenir des informations sur les personnes présentes dans la scène. Ces informations peuvent décrire l'ensemble de chaque individu : silhouette, trajectoire, posture, etc. mais peuvent aussi décrire des parties plus précises du corps : localisations et trajectoires du visage et des mains, estimation de la direction du regard etc. Détecter et suivre le visage et les mains d'un individu placé devant une caméra sont des fonctionnalités indispensables des interfaces homme-machine avancées. C'est en effet une première étape pour l'analyse et l'interprétation des gestes et des actions d'un être humain.

Il existe de nombreuses méthodes pour détecter et suivre au cours du temps le visage et les mains d'une personne [Chellappa95, Pavlovic97, Hjelmäs01, Yang02, Canton-Ferrer05]. Deux grandes approches se détachent principalement :

1. l'extraction de traits caractéristiques de ces *ROI* ;
2. la détection de peau par une approche couleur.

Nous allons présenter quelques travaux développés pour chacune de ces approches. Nous parlerons brièvement ensuite des autres approches utilisées qui sont plus liées à la manière de réaliser le suivi temporel de ces *ROI*. Dans une dernière partie, nous présenterons un état de l'art sur quelques espaces utilisés pour réaliser une détection de peau par une approche couleur.

4.1.1 Extraction de traits caractéristiques

Les méthodes basées sur l'extraction de traits caractéristiques du visage et / ou des mains nécessitent la définition de modèles pour ces traits, et peuvent aussi nécessiter un modèle global du corps humain plus ou moins détaillé. L'utilisation de traits caractéristiques est guidée par la *ROI* considérée.

Dans le cas du visage, outre sa symétrie, il est composé de plusieurs traits caractéristiques comme les yeux, le nez, la bouche, les sourcils, les oreilles etc. Jacquin et Eleftheriadis ont utilisé une région rectangulaire basée sur les yeux, le nez et la bouche pour suivre un visage au cours du temps [Jacquin95]. Une approche similaire a été proposée par Jebara et Pentland qui sélectionnent une région candidate pour le visage selon un critère de maximum de vraisemblance [Jebara97]. D'autres traits caractéristiques comme les sourcils, les iris, les paupières ont aussi été utilisés et testés avec une précision de suivi de 98% [Tian99]. Même le clignement des yeux est une particularité qui a retenu l'attention [Crowley97]. Les yeux sont souvent des régions d'intérêt prisées du visage. Plus récemment, Peng, Chen, Ruan et Kukharev ont proposé une méthode de détection des yeux dans des images de visage en niveaux de gris [Peng05b]. Ces travaux ont menés à une carte d'identité "intelligente" pouvant réaliser l'identification du visage de son porteur grâce à des modèles d'apparence active (*Active Appearance Models*) [Peng05a]. Concernant le visage, l'une des difficultés pour le choix des traits caractéristiques est la diversité des cas possibles, principalement suivant la chevelure de la

personne, le port de lunettes, le maquillage et le port de bijoux (boucles d'oreille, colliers etc.).

La main a comme principal trait caractéristique sa structure anatomique. C'est un objet non rigide articulé complexe, composé de cinq doigts ayant la même structure, formés de trois phalanges reliées entre elles et à la main par des articulations. Des modèles 2D et 3D sont donc utilisés, parfois conjointement avec des gants, des marqueurs ou des capteurs, pour détecter et suivre la main au cours du temps. Le système *DigitEyes* est un système de suivi temporel de main qui la modélise avec vingt-sept degrés de liberté [Rehg93, Rehg94]. Dans [Nirei96], la main est constituée d'un modèle 3D de vingt et un segments reliés par vingt articulations, d'après les connaissances anatomiques humaines. Le pouce, chacun des quatre autres doigts et une partie du poignet sont respectivement décrits par trois, quatre et deux segments. Chaque segment est modélisé par un cône elliptique tronqué. Cette approche ne considère toutefois pas les variations de taille et les problèmes d'occultations. Les processus d'ajustement des doigts (positions, orientations) basés sur la cinématique inverse mènent à des erreurs plus fréquentes en cas d'occultation. Cependant, grâce à des modèles de référence classés selon la visibilité et mis à jour, et à des fonctions de fenêtrage, il est possible de réduire le nombre d'erreurs dans une certaine mesure [Rehg95]. Dans le cas des mains, le choix d'un modèle ou non dépend souvent de l'application (et donc de la prise de vue) et est lié au nombre de pixels dans l'image pour les différentes *ROI* considérées.

4.1.2 Détection de peau par une approche couleur

Les méthodes basées sur la détection de peau par une approche couleur travaillent sur les composantes de couleur de la peau. De nombreuses études montrent que la couleur de la peau forme une distribution compacte dans certains espaces couleur [Terrillon99] et que c'est une information **discriminante**, c'est-à-dire que les informations de couleur (chrominance) de la peau ne dépendent pas de l'information de luminosité (luminance). Néanmoins, contrairement à ce que l'on pourrait croire, il n'existe pas d'espace couleur prépondérant pour résoudre ce problème. Le choix de l'espace dans lequel la détection est réalisée dépend du type d'application visée et du système d'acquisition (bruit d'acquisition et balance des blancs de la caméra, éclairage, etc.). L'utilisation ou non d'un modèle pour faire la détection, à partir d'un apprentissage grâce à une base de données, est envisageable selon les applications. Les méthodes de détection de peau par une approche couleur sont principalement des méthodes statistiques.

Les méthodes statistiques pour réaliser une détection de peau par une approche couleur sont celles où l'on utilise une représentation de la distribution de probabilités pour la ou les composante(s) de couleur. La représentation peut-être une Gaussienne [Darrell96], des mélanges de Gaussiennes [McKenna98, McKenna99] ou des histogrammes [Birchfield98, Yoo99]. Dans la majorité des cas, on se place dans des espaces de couleur à luminance et chrominance séparées ou dans des espaces normalisés. Les *blobs* sont aussi utilisés, par exemple dans [Isard98] où le suivi temporel combine des *blobs* de couleur avec un modèle de contour. Après avoir extrait le visage et les mains grâce à la couleur, les *blobs* sont calculés et suivis grâce à un filtrage de Kalman. Cependant, le suivi temporel peut échouer à cause de la difficulté à distinguer les multiples objets que sont la main droite, la main gauche et le visage.

4.1.3 Autres approches

De nombreuses autres approches pour le suivi temporel du visage et / ou des mains existent, mais elles sont principalement liées à la façon de réaliser le suivi temporel. Nous ne les détaillerons pas, car elles utilisent peu, de façon explicite, les détails que sont la couleur de la peau ou les traits caractéristiques du visage et / ou des mains. Parmi ces autres approches, il y a celles qui découlent du suivi d'objets articulés, présentées de façon générale pour le suivi temporel d'objets, mais appliquées ici aux parties du corps que sont le visage et les mains. Nous retrouverons donc par exemple les méthodes basées sur l'utilisation de modèles 2D ou 3D avec des modèles de référence déformables [Essa94, Zhong00], les méthodes basées contour, avec les contours actifs ou *snakes* [Heap96, Sobottka98, Kim01], les méthodes basées sur le flot optique [Black95] etc.

Dans une méthode basée modèle, la combinaison d'un modèle stochastique avec un modèle de mouvement et un modèle de caméra est utilisée [Yang95, Yang96]. Ces trois modèles compensent les différents problèmes pouvant survenir. Le modèle stochastique s'adapte aux différents individus et aux conditions d'illumination de la scène en temps-réel ; le modèle de mouvement estime le mouvement de l'image et prédit la fenêtre de recherche ; finalement le modèle de caméra prédit et compense le mouvement de la caméra.

4.1.4 Espaces couleur RGB, $YCbCr$ et HSI

Avec les contraintes de notre système, nous cherchons à détecter et à suivre au cours du temps le visage et les mains, avec la distinction main droite / main gauche. Dans nos conditions de prise de vue, étant donné la complexité de la définition de modèles différents pour extraire les traits caractéristiques du visage et des mains et par rapport à des contraintes de temps de calcul, une détection de peau par une approche couleur est bien plus adaptée. Lorsqu'on réalise une détection de ce genre, il faut choisir l'espace couleur le plus adapté pour la réaliser.

Cette partie présente donc un état de l'art sur quelques espaces utilisés pour réaliser une détection de peau. Il existe une multitude d'espaces couleur, entre autres :

- *RGB (Red, Green, Blue)*, $CMY(K)$ *(Cyan, Magenta, Yellow (Black))* ;
- XYZ, *Lab*, *Luv*, *Lhs*, *Lhc* etc. définis par la CIE (Commission Internationale de l'Éclairage) ;
- *HSL (Hue, Saturation, Darkness)*, HSI *(Intensity)*, HSV *(Value)*, HCI *(Colourfulness)*, TSL (Teinte, Saturation, Luminosité) etc. ;
- $YCbCr$, YUV, YIQ etc.

Il existe en effet plusieurs manières de représenter les couleurs, soit selon une approche purement physique (RGB, CIE XYZ etc.) soit selon une approche physique corrigée par les données de la perception visuelle (CIE *Lab*, CIE *Luv* etc.). Pour une description plus complète des espaces et les formules de conversion pour passer d'un espace à un autre, se reporter à [Ford98, Fernandez01, Couleur04].

Dans cette partie, nous décrirons les trois espaces couleur suivants : RGB, $YCbCr$ et HSI. Ces trois espaces ont été retenus parce qu'il est préférable d'utiliser des espaces à luminance et chrominance séparées. Dans le cas de l'espace couleur RGB, même si cet espace n'est pas à luminance et chrominance séparées, il est possible de s'y ramener (cf. partie 4.1.4.1), de plus, cet espace est très utilisé par les systèmes informatiques, ce qui justifie qu'on le retienne. Nous présenterons leurs définitions habituelles et quelques travaux de détection de peau par une

approche couleur dans chacun de ces espaces.

4.1.4.1 Espace couleur *RGB*

L'espace *RGB (Red, Green, Blue)* est l'espace couleur le plus utilisé dans la majorité des systèmes informatiques et sur Internet. Il intervient dans la plupart des appareils de prise d'images couleurs, ainsi que dans les moniteurs couleurs (écrans de télévision, d'ordinateur etc.). La plupart des formats de codage d'images (GIF, BMP, PNG, PPM etc.) utilisent cet espace couleur.

L'espace RGB utilise trois composantes numériques pour représenter une couleur. C'est un espace additif de couleur. Les capteurs fournissent en sortie trois informations issues des trois canaux R, G et B et la couleur d'un pixel quelconque est reconstituée par synthèse additive de ces trois informations. Cet espace couleur peut donc être vu comme un système de coordonnées cartésiennes en 3D où les axes Ox, Oy et Oz représentent chacun l'intensité des trois couleurs R, G et B. Sa représentation est donc un cube. Chaque composante a une gamme de valeurs comprise entre 0 et 255, quand l'information est quantifiée sur 8 bits. $R = G = B$ définit une droite appelée axe achromatique ou axe des niveaux de gris. En particulier, lorsque $R = G = B = 0$, le pixel correspondant est noir, lorsque $R = G = B = 255$, il est blanc.

En normalisant l'amplitude de chaque composante par la somme $R + G + B$, l'espace couleur RGB se transforme en l'espace couleur rgb, qui est l'espace RGB **normalisé**. Dans cet espace couleur, le noir correspond à $r = g = b = 0$ et le blanc à $r = g = b = 1$.

La représentation 3D de l'espace couleur rgb est visible sur la figure 4.1.

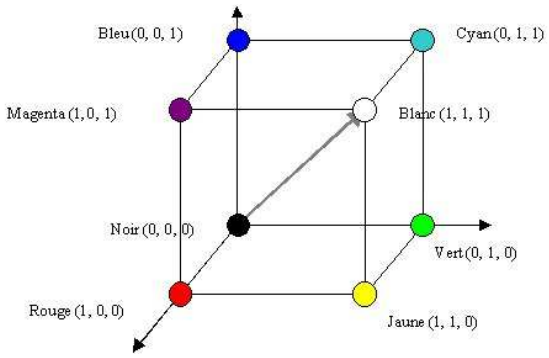


FIG. 4.1 – Représentation 3D de l'espace couleur rgb.

Dans l'espace rgb, il y a une redondance d'information. L'une des trois composantes n'est plus nécessaire puisque la somme des trois composantes vaut 1. Cet espace couleur est donc aussi souvent représenté uniquement comme son sous-espace couleur rg.

De nombreux articles conseillent une détection de peau dans les espaces RGB, rgb ou dans le sous-espace couleur rg [Yang95, Sobottka96, Yang98, Schwerdt00]. Ces articles utilisent souvent une modélisation gaussienne pour les composantes de couleur. Voici, par exemple, les moyennes et les écart-types pour les trois composantes R, G et B selon une modélisation gaussienne. Ces valeurs sont issues de tests portant sur une base de données d'un millier de visages [Yang98] :

$$\begin{aligned}\mu_R &= 234.2947 \quad \sigma_R = 26.7735,\\ \mu_G &= 185.7177 \quad \sigma_G = 30.4088,\\ \mu_B &= 151.1090 \quad \sigma_B = 25.6779.\end{aligned}$$

[Yang98] donne aussi (pour l'ensemble des visages de la base de données) ces moyennes pour le sous-espace rg, ce qui montre l'utilité de la normalisation, car les écart-types sont réduits d'un facteur compris entre 5 et 10 :

$$\begin{aligned}\mu_r &= 104.2225 \quad \sigma_r = 4.9317,\\ \mu_g &= \ 81.5879 \quad \sigma_g = 3.8858.\end{aligned}$$

L'utilisation de l'espace couleur RGB (ou rgb) pour une détection de peau ne semble pas naturelle puisque cet espace ne tient pas compte de la séparation de l'information en luminance et chrominance.

4.1.4.2 Espace couleur $YCbCr$

L'espace couleur $YCbCr$ a été développé en tant que partie de la norme *ITU-RBT.601*[1], d'ancien nom CCIR 601[2], mondialement définie pour les standards de vidéo numérique et utilisée dans les transmissions télévisuelles. D'abord, notons qu'il existe aussi les espaces couleur YIQ et YUV et que la confusion est fréquente parmi ces trois espaces[3]. Les espaces couleur $YCbCr$, YIQ et YUV sont des espaces couleur semblables. YUV est utilisé pour le codage des couleurs dans le système de télévision *PAL*[4], YIQ pour le système *NTSC*[5]. Ces deux systèmes sont dépendants des appareils, quant à l'espace $YCbCr$, standard en vidéo numérique, c'est une version mise à l'échelle et décalée de l'espace couleur YUV. De ce point de vue, $YCbCr$ est le plus intéressant. Y correspond à la composante de luminance, Cb et Cr aux deux composantes de chrominance (respectivement compléments bleu et rouge). La luminance Y et les chrominances Cb et Cr peuvent s'obtenir à partir des composantes R, G et B par la conversion suivante :

$$\begin{bmatrix} Y-16 \\ Cb-128 \\ Cr-128 \end{bmatrix} = \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ -0.148 & -0.291 & 0.439 \\ 0.439 & -0.368 & -0.071 \end{bmatrix} \times \begin{bmatrix} R \\ G \\ B \end{bmatrix}.$$

Dans notre système, les caméras fournissent des images dans l'espace couleur $YCbCr$. La conversion inverse est donnée par :

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1.164 & 0 & 1.596 \\ 1.164 & -0.392 & -0.813 \\ 1.164 & 2.017 & 0 \end{bmatrix} \times \begin{bmatrix} Y-16 \\ Cb-128 \\ Cr-128 \end{bmatrix}.$$

[1] *ITU-RBT : International Telecommunications Union - Remote Batch Terminal.*
[2] CCIR : Comité Consultatif International sur la Radio.
[3] Par exemple dans le code même de notre système, où Cb et Cr sont respectivement notés U et V.
[4] *PAL : Phase Alternating Line.*
[5] *NTSC : National Television Standards Committee.*

Il est alors aisé de passer éventuellement de R, G et B à r, g et b.

Cet espace couleur est fréquemment cité dans la littérature pour effectuer une détection de peau [Chai99, Ahlberg99, Bruce00, Marcel00a, Marcel00b]. La table 4.1 donne les seuils utilisés par [Chai99] pour réaliser une détection de visage en (a) et ceux utilisés par [Ahlberg99] afin de réaliser une extraction des paramètres de couleur du visage en (b).

TAB. 4.1 – Seuils de détection de peau dans le sous-espace couleur $CbCr$. (a) [Chai99], (b) [Ahlberg99].

$Cb \in [110; 123]$
$Cr \in [136; 156]$

(a)

$Cb \in [77; 127]$
$Cr \in [133; 173]$

(b)

Nous pouvons remarquer que les seuils de [Chai99] sont inclus dans ceux de [Ahlberg99].

4.1.4.3 Espace couleur HSI

L'espace RGB est utile pour représenter les couleurs sur des appareils, mais est peu intuitif pour l'observateur humain. L'espace couleur HSI *(Hue, Saturation, Intensity)* utilise un système de coordonnées cylindriques (z, r, θ) pour représenter les couleurs. H est la teinte, S la saturation et I l'intensité.

L'intensité I est définie selon l'axe partant du point d'intensité la plus faible ($I = 0$, noir) au point d'intensité la plus forte ($I = 1$, blanc), et représente la quantité de lumière contenue dans une couleur. C'est l'axe des niveaux de gris et de coordonnée z. La saturation S est la distance d'une couleur à l'axe d'intensité, et donne la quantité de coloration. La saturation peut être représentée par des cylindres coaxiaux ayant comme axe l'axe d'intensité. C'est la coordonnée r. La teinte H définit la couleur le long de l'arc-en-ciel. C'est la coordonnée θ, et son origine ($\theta = 0$) correspond généralement au rouge. Les composantes de l'espace couleur HSI sont décorrélées. L'espace HSI est relié à l'espace couleur RGB par des formules de conversion non linéaires :

$$
\begin{aligned}
H &= \begin{cases} \arccos\left(\frac{2R-G-B}{2\sqrt{(R-G)^2+(R-B)(G-B)}}\right) & \text{si } G > B, \\ 2\pi - \arccos\left(\frac{2R-G-B}{2\sqrt{(R-G)^2+(R-B)(G-B)}}\right) & \text{si } B > G, \end{cases} \\
S &= 1 - 3\frac{\min(R,G,B)}{R+G+B}, \\
I &= \frac{R+G+B}{3}.
\end{aligned}
$$

La saturation S n'est pas définie pour $I = 0$ et la teinte H n'est pas définie si $S = 0$.

La représentation 3D de l'espace couleur HSI est visible sur la figure 4.2.

[Mottin00] propose une détection de peau par seuillage dans l'espace couleur HSI, avec ou sans utilisation d'une modélisation gaussienne pour la teinte H. Les seuils pour une détection de peau sans modélisation gaussienne sont $H \in [0°; 60°]$. Voici la moyenne et l'écart-type pour H selon une modélisation gaussienne :

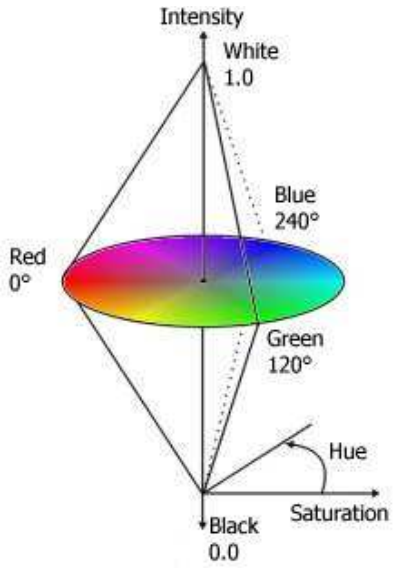


FIG. 4.2 – Représentation 3D de l'espace couleur HSI.

$$\begin{aligned} \mu_H &= 22.08^\circ, \\ \sigma_H &= 1.375^\circ. \end{aligned}$$

Et voici les seuils obtenus avec $\mu_H \pm 3\sigma_H$: $H \in [17.95^\circ; 26.21^\circ]$. Les meilleurs résultats sont obtenus avec une modélisation gaussienne.

[Sobottka96] utilise, en plus de la teinte H, la saturation S. Les seuils proposés par [Sobottka96] sont les suivants :

$$\begin{aligned} H &\in [0^\circ; 50^\circ], \\ S &\in [0.23; 0.68]. \end{aligned}$$

Nous pouvons constater que les seuils pour H de [Sobottka96] ne sont pas très différents de ceux de [Mottin00] sans modélisation gaussienne.

4.2 Détection de peau par une approche couleur

4.2.1 Introduction

Dans le cadre de notre système, les conditions de prise de vue peuvent amener des zones de petite taille (faible nombre de pixels pour les mains ou le visage). De plus, la définition et l'utilisation de traits caractéristiques précis et différents pour le visage et les mains seraient trop contraignantes au niveau de la complexité et du temps de calcul. Par conséquent, une approche couleur pour la détection de peau pouvant être utilisée de la même manière pour détecter le visage et les mains a été préférée.

Avant de réaliser la détection de peau, il est nécessaire de déterminer l'espace couleur qui sera utilisé. Dans notre système, c'est l'espace $YCbCr$ qui a été retenu. Nous commencerons donc par présenter comment et pourquoi cet espace couleur a été retenu. Nous détaillerons ensuite la méthode utilisée pour la détection de peau qui est basée sur un seuillage des composantes de couleur (chrominances) Cb et Cr. Quatre seuils sont nécessaires, deux pour la chrominance Cb (Cb_{min} et Cb_{max}) et deux pour la chrominance Cr (Cr_{min} et Cr_{max}). Puis nous préciserons la méthode d'adaptation automatique des seuils de détection.

4.2.2 Détermination de l'espace couleur utilisé ($YCbCr$)

Comme nous l'avons vu dans l'état de l'art sur les espaces, il existe de nombreux espaces couleur et il nous faut déterminer lequel est le plus adapté à notre système. Les deux critères principaux pour déterminer l'espace optimal sont les suivants :

1. la qualité des résultats ;
2. la cadence de traitement.

Bien sûr, il nous faut un espace dans lequel les résultats de détection de peau sont satisfaisants. Mais nous voulons aussi que cette détection de peau soit rapide. C'est, une fois encore, le compromis qualité / temps de calcul. Les caméras que nous utilisons fournissent des images dans l'espace couleur $YCbCr$. Pour une détection dans un autre espace que celui-ci, il est nécessaire de faire des conversions entre espaces. De plus, certains espaces couleur présentent des composantes qui nécessitent des fonctions mathématiques avancées comme, par exemple, la composante H de l'espace HSI qui nécessite les fonctions racine carrée et arccos .

4.2.2.1 Base de données de pixels de peau

La base de données de pixels de peau que nous utilisons provient de deux sources distinctes. La première est l'image de la base de peaux de Von Luschan présentée figure 4.3.

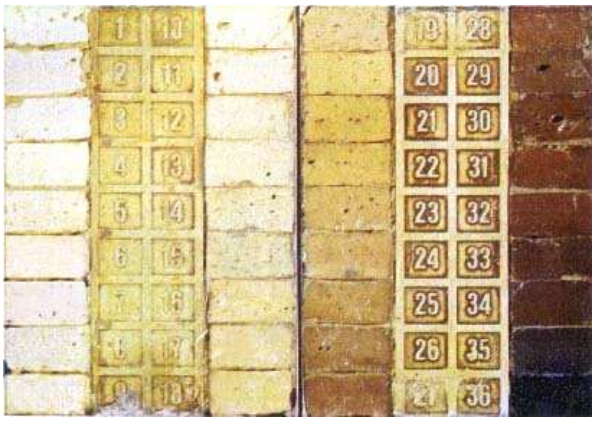


FIG. 4.3 – Image de la base de peaux de Von Luschan [Von Luschan27].

L'image de la base de peaux de Von Luschan, de dimensions 409×286 pixels, contient 36 imagettes de peau depuis les peaux les plus claires jusqu'aux peaux les plus mates. Deux remarques peuvent être faites à propos de cette image. La première est qu'elle ne contient

pas uniquement des pixels de peau, principalement à cause des délimitations entre imagettes mais aussi à cause des numéros et de la présence de dégradations dans l'image. Par exemple, en bas à gauche entre les deux premières séries d'imagettes, l'image est nettement dégradée. La seconde remarque est que les couleurs de peau présentées sont plutôt dans les tons jaunes et marrons, alors qu'il existe aussi des peaux de couleur rosée. Ceci s'explique principalement par le fait que **les couleurs de peaux dépendent du système d'acquisition.**

Pour cette raison, nous avons décidé de créer un ensemble d'images de peau propre à notre système d'acquisition. C'est la seconde source de notre base de données de pixels de peau.

Ces images ont été acquises avec une caméra *Sony DFW-VL500*, ceci afin de tenir compte des conditions d'acquisition et des caractéristiques de la caméra (balance des blancs, bruit d'acquisition, gain etc.). Elles ont été prises en environnement intérieur, dans les mêmes conditions d'acquisition (illumination etc.). Afin d'avoir une grande quantité de pixels de peau et une variabilité assez représentative, les images sont assez grandes (résolution d'image 640×480) et contiennent uniquement des pixels de peau pour des personnes de couleur de peau assez variées. Nous avons demandé à une dizaine d'individus de contribuer à cette base de données en nous permettant de filmer leurs mains et / ou leurs bras. Ainsi nous disposons d'une vingtaine d'images de peaux à majorité caucasiennes, mais aussi asiatiques, noires et indiennes.

La figure 4.4 montre les images de peaux de cette base de données.

FIG. 4.4 – Images de peaux acquises avec notre système.

4.2.2.2 Espace couleur retenu : $YCbCr$ (comparaison avec HSI)

Plusieurs espaces ont été testés pour déterminer lequel est le plus adapté à la détection des pixels de peau. Nous ne présentons ici que la comparaison entre les espaces couleur $YCbCr$ et HSI qui ont donné tous deux des résultats de bonne qualité [Girondel02].

Notre système fournit des images dans l'espace couleur $YCbCr$ donc la luminance Y et

les chrominances Cb et Cr sont des informations accessibles immédiatement, à l'acquisition. Pour changer d'espace, il faut soit faire une ou plusieurs conversions entre espaces, soit utiliser une table de conversion, ou *LUT (Look Up Table)*. Dans [Marcel00a], une table de conversion est utilisée dans l'espace couleur YUV pour déterminer si un triplet (Y, U, V) correspond à un pixel de peau. Pour passer dans l'espace HSI, par exemple, deux solutions sont possibles :

1. faire deux conversions successives entre espaces couleur :
   - de $YCbCr$ vers RGB ;
   - de RGB vers HSI.
2. utiliser une table de conversion.

La seconde solution amène une contrainte d'espace mémoire disponible de 16 Mo (256^3) pour l'utilisation d'une table de conversion. Si on n'utilise pas de table, le temps de calcul est plus grand que dans l'espace couleur $YCbCr$ à cause des conversions et des fonctions mathématiques racine carrée et arccos .

4.2.2.3 Analyse des bases de données

Nous présentons ici les projections d'une partie ou de l'ensemble de notre base de données de pixels de peau dans les espaces $YCbCr$ et HSI. Dans ces espaces couleur, les composantes de couleur pertinentes sont Cb et Cr pour l'espace $YCbCr$; H pour l'espace couleur HSI.

Base de peaux de Von Luschan : Tout d'abord, nous présentons les projections 2D des pixels de la base de peaux de Von Luschan sur les sous-espaces $CbCr$, CbY et YCr. La figure 4.5 montre les résultats obtenus.

Ces projections 2D illustrent quelques points importants :

- La luminance Y prend quasiment toutes les valeurs possibles entre 0 et 255, ce qui montre bien que la luminance n'est pas une information discriminante pour détecter la peau.
- La distribution dans le sous-espace $CbCr$ est assez compacte, ce qui prouve que cet espace couleur peut servir à réaliser une détection de peau.

Les seuils délimitant le rectangle blanc de la projection 2D de la base de peaux de Von Luschan sur le sous-espace couleur $CbCr$ sont :

$$\boxed{\begin{array}{c} Cb \in [\,64; 148] \\ Cr \in [111; 167] \end{array}}$$

Par rapport aux seuils de l'état de l'art, donnés par [Chai99] et [Ahlberg99], ces seuils couvrent une zone plus large, sauf par rapport à Cr_{max}.

La figure 4.6 illustre l'histogramme 1D de la teinte H pour les pixels de la base de peaux de Von Luschan. Les valeurs extrêmes de la teinte H sont les suivantes :

$$H \in [0\ rad; 1.047\ rad] \equiv H \in [0°; 60°]\,.$$

Ensemble de la base de données : La figure 4.7 illustre les histogrammes des chrominances Cb et Cr pour les pixels de peaux de l'ensemble de notre base de données.

La figure 4.8 présente la comparaison entre les projections 2D sur le sous-espace $CbCr$ de la base de peaux de Von Luschan en (a), de la base de peaux acquise avec notre système en (b) et de l'ensemble des deux en (c).

La table 4.2 donne les seuils délimitant les rectangles noirs englobant les projections 2D de la figure 4.8. Ces derniers englobent intégralement ceux de l'état de l'art.

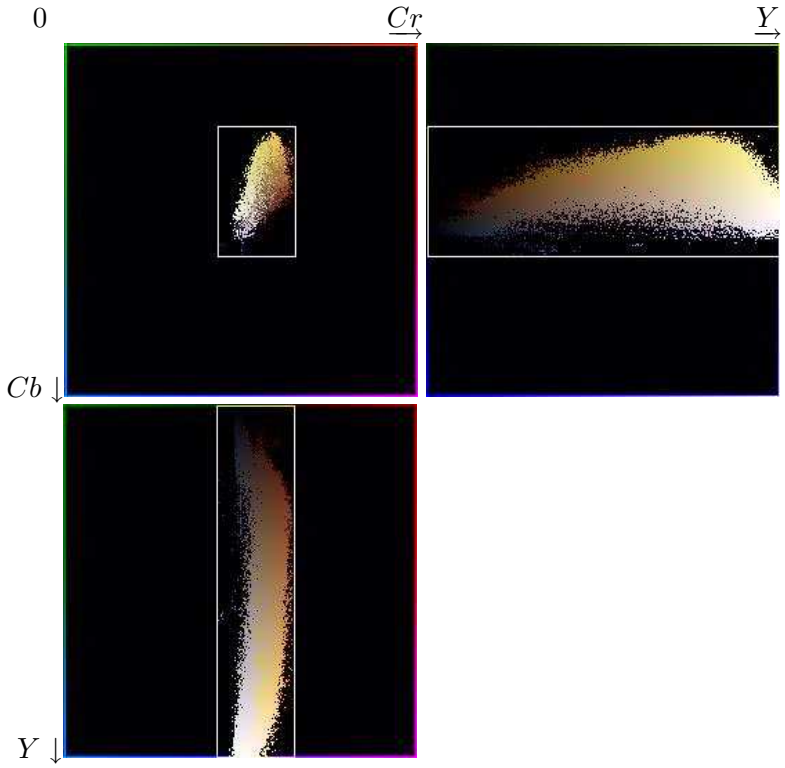


FIG. 4.5 – Projections 2D de la base de peaux de Von Luschan dans l'espace couleur $YCbCr$.

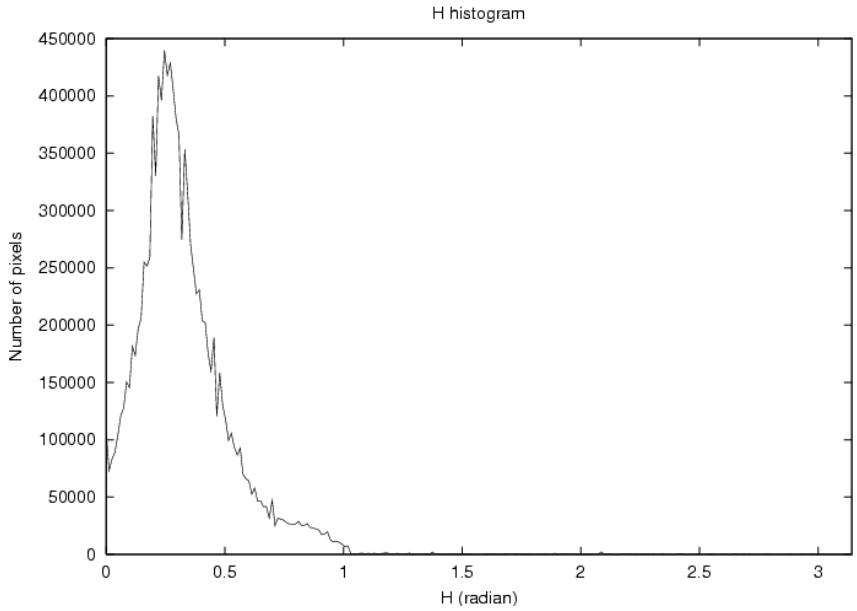


FIG. 4.6 – Histogramme 1D de la teinte H pour la base de peaux de Von Luschan.

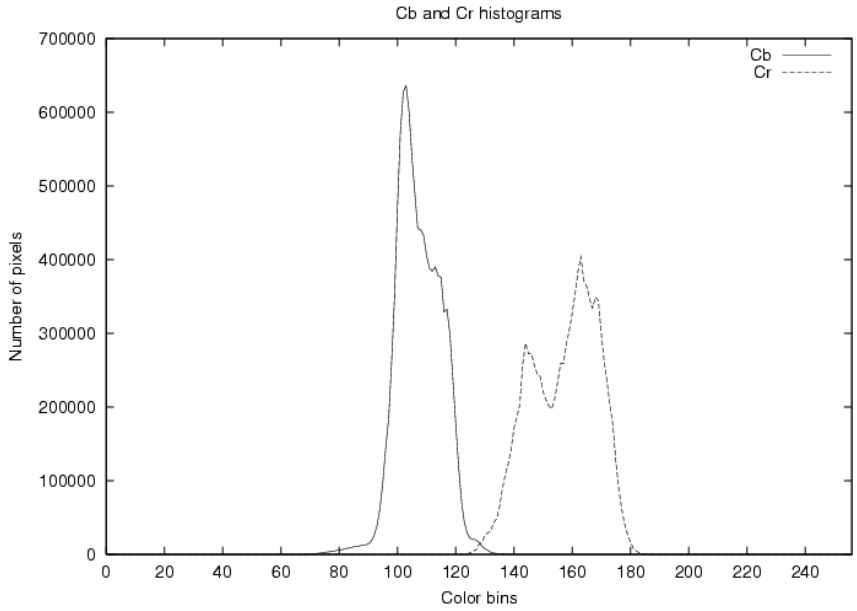
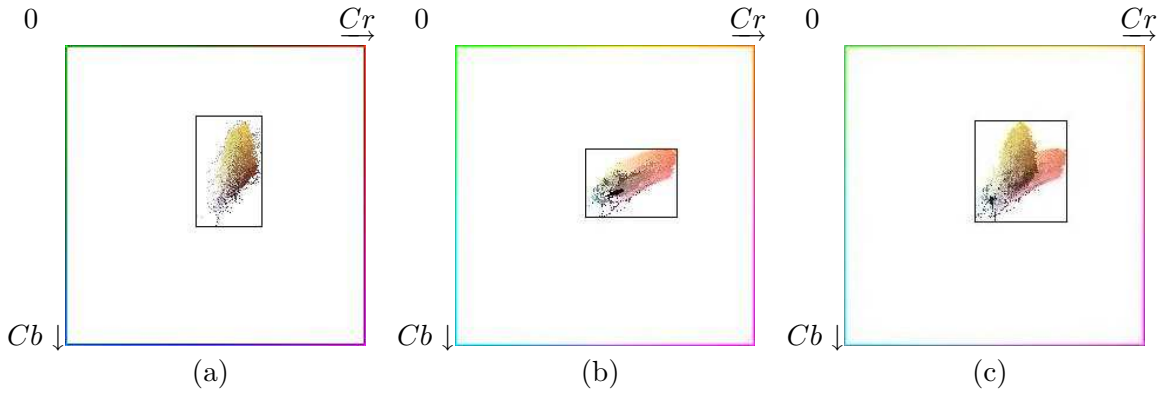


FIG. 4.7 – Histogrammes 1D des chrominances Cb et Cr pour l'ensemble de la base de données.



(a) (b) (c)

FIG. 4.8 – Projections 2D sur le sous-espace couleur $CbCr$. (a) base de peaux de Von Luschan, (b) base de peaux acquise avec notre système, (c) ensemble des deux.

TAB. 4.2 – Seuils des rectangles de détection des projections 2D de la figure 4.8. (a) base de peaux de Von Luschan, (b) base de peaux acquise avec notre système, (c) ensemble des deux.

(a)	(b)	(c)
$Cb \in [64; 148]$ $Cr \in [111; 167]$	$Cb \in [88; 140]$ $Cr \in [105; 189]$	$Cb \in [64; 148]$ $Cr \in [105; 189]$

4.2.3 Méthode de seuillage dans le sous-espace couleur $CbCr$

Pour chaque pixel, en fonction de l'espace choisi pour réaliser la détection de peau, nous prenons la décision soit selon ses chrominances Cb et Cr, pour une détection dans l'espace couleur $YCbCr$, soit selon la composante H, pour une détection dans l'espace HSI. La détection se fait par seuillage sur la ou les composante(s) considérée(s).

Dans l'espace couleur $YCbCr$, quatre seuils sont nécessaires, deux pour la chrominance Cb (Cb_{min} et Cb_{max}) et deux pour la chrominance Cr (Cr_{min} et Cr_{max}). Les caméras que nous utilisons fournissent des images dans l'espace $YCbCr$ sous-échantillonnées en chrominances (format $4:2:0$). La luminance Y est définie pour chaque pixel et les chrominances Cb et Cr sont sous-échantillonnées d'un facteur 2 selon les lignes et selon les colonnes, soit d'un facteur 4 au total. Afin de rendre la détection de peau plus robuste, un moyennage des chrominances Cb et Cr est effectué sur 16 pixels (moyennage 4×4) et la décision est prise pour les 4 pixels centraux.

Après un certain nombre de tests, dont les détails sont présentés dans [Girondel02], nous avons choisi de restreindre l'influence de la base de peaux de Von Luschan en accentuant celle de la base de peaux acquise avec notre système. En effet, il s'avère que la prise en compte des spécificités du système d'acquisition est indispensable pour une bonne détection. En d'autres termes, les seuils de détection sont tributaires du système d'acquisition. Les seuils initiaux ont été déterminés par un opérateur humain, après une étude qualitative des résultats de détection sur une dizaine de séquences vidéo. Une étude de la variabilité des seuils en fonction de l'opérateur aurait été pertinente.

Les seuils initiaux, pour une détection de peau dans l'espace couleur $YCbCr$, sont :

$$\boxed{\begin{array}{c} Cb \in [\,86; 140] \\ Cr \in [139; 175] \end{array}}$$

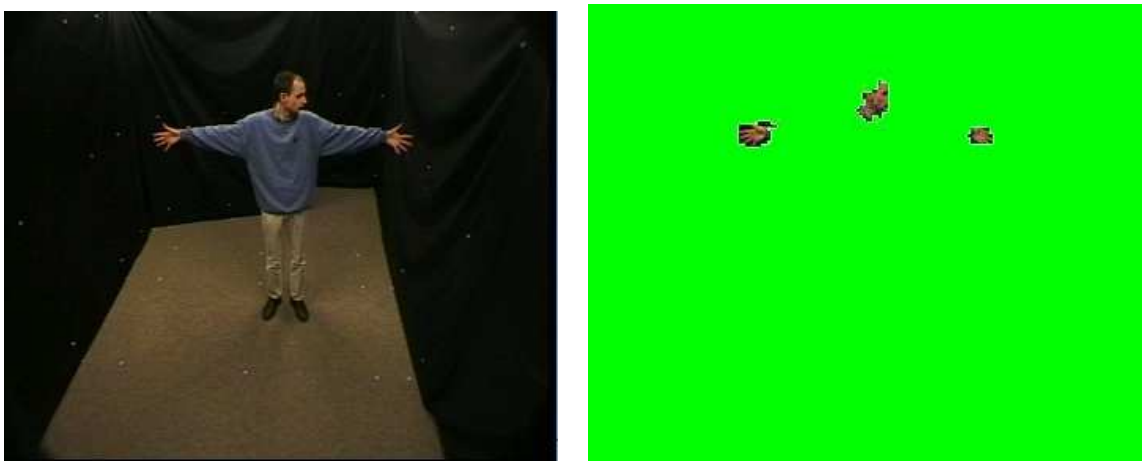
Un exemple de détection de peau dans l'espace $YCbCr$ est donné figure 4.9 avec l'image originale en (a), et les pixels de peau segmentés en (b). On notera que les pixels de peau sont très bien détectés et que ni le pantalon, ni le sol de la scène, de couleurs pourtant proches de couleurs de peau, n'ont été détectés.

Dans l'espace HSI, deux seuils sont nécessaires pour la composante H (H_{min}, H_{max}). Les seuils utilisés sont ceux de [Mottin00], avec modélisation gaussienne de la teinte H :

$$H \in [17.95°; 26.21°] \equiv H \in [0.3134\ rad; 0.4574\ rad]\,.$$

La figure 4.10, page 92, donne un exemple de détection de peau dans l'espace HSI et permet aussi de comparer avec la détection de peau dans l'espace $YCbCr$. En (a), nous avons l'image originale, en (b) l'image segmentée, en (c) les pixels de peau segmentés dans l'espace couleur HSI et en (d) les pixels de peau segmentés dans l'espace couleur $YCbCr$. Les résultats sont assez proches, même si ici, il y a un peu moins de fausses détections dans l'espace $YCbCr$ et les régions de peau sont plus précises (cf. le bras gauche et le visage sur la figure 4.10).

Afin de détecter les pixels de peau rapidement dans une image, nous parcourons l'ensemble des pixels segmentés situés à l'intérieur des boîtes englobantes rectangulaires issues de la segmentation. Ceci permet aussi de ne pas traiter les pixels du fond de la scène et rend la détection plus robuste, même si le fond a une couleur proche des couleurs de peau.



(a) (b)

FIG. 4.9 – Exemple de détection de peau dans l'espace couleur $YCbCr$. (a) image originale, (b) pixels de peau.

4.2.4 Adaptation automatique des seuils dans le sous-espace couleur $CbCr$

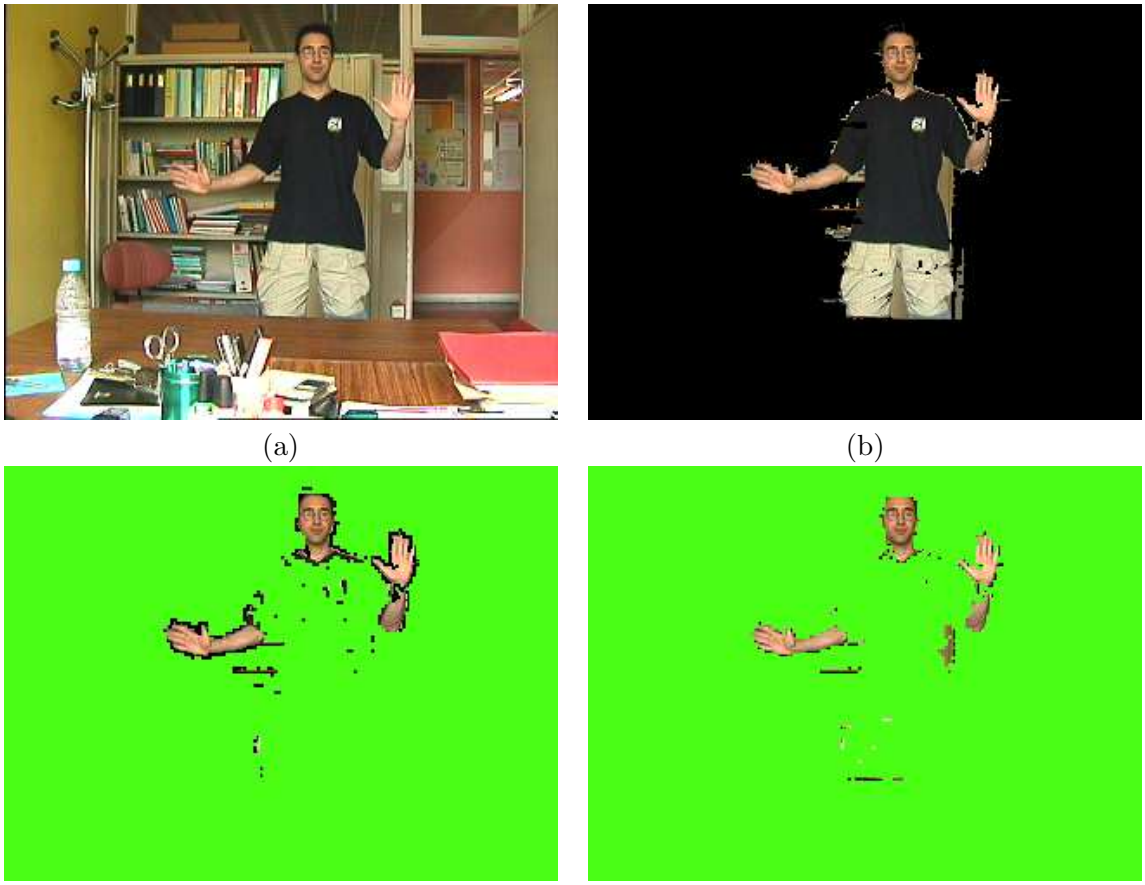
L'adaptation automatique des seuils pour la détection de peau est nécessaire principalement pour les deux raisons suivantes :

1. tenir compte des variations des conditions d'acquisition (illumination de la scène, paramètres de la caméra, etc.) ;
2. tenir compte de l'homogénéité de la couleur de peau pour un individu donné (la couleur de la peau d'une personne représente un sous-ensemble des couleurs de peaux possibles).

Pour réaliser cette adaptation automatique, nous pouvons essayer de nous appuyer sur les caractéristiques statistiques des chrominances des pixels de peau détectés. En effet, de nombreux travaux affirment que les composantes de couleur de la peau présentent la propriété de gaussianité, c'est-à-dire que la distribution de probabilités de ces composantes est une gaussienne [Darrell96, Yang98, McKenna98, McKenna99, Mottin00]. Cette propriété de gaussianité a été utilisée dans l'espace RGB par [Yang98] et dans l'espace couleur HSI par [Mottin00]. Nous avons donc réalisé une étude statistique sur la gaussianité des chrominances Cb et Cr des pixels de peau, ceci pour voir si, dans l'espace $YCbCr$ retenu, la peau présente aussi cette propriété de gaussianité et si, le cas échéant, nous pouvions utiliser cette propriété pour adapter les seuils de façon automatique.

4.2.4.1 Étude statistique sur la gaussianité de la peau dans l'espace couleur $YCbCr$

Il existe plusieurs façons de déterminer si une distribution est gaussienne, citons, entre autres, les tests du χ^2, de Kolmogorov-Smirnov, de Cramer Von Mises etc. [Saporta90]. Afin d'étudier et de tester la propriété de gaussianité des chrominances Cb et Cr des pixels de peau, nous avons utilisé le test de Kolmogorov-Smirnov. Ce test a été choisi pour sa simplicité de mise en œuvre.



(a) (b)

(c) (d)

FIG. 4.10 – Comparaison HSI *vs* $YCbCR$. (a) image originale, (b) image segmentée, (c) pixels de peau dans l'espace couleur HSI et (d) pixels de peau dans l'espace couleur $YCbCr$.

Test de Kolmogorov-Smirnov : C'est un test d'ajustement non paramétrique. Les tests d'ajustement ont pour but de vérifier si un échantillon provient ou non d'une variable aléatoire de fonction de répartition connue $G(x)$ (cf. [Saporta90]). Soit $F(x)$ la fonction de répartition de la variable échantillonnée, il s'agit donc de tester l'hypothèse $H^0 : \ F(x) = G(x)$ contre l'hypothèse $H^1 : \ F(x) \neq G(x)$. Si F_n représente la fonction de répartition empirique d'un n-échantillon d'une variable aléatoire de distribution connue $G(x)$, alors $D_n = \sup_x |F_n(x) - G(x)|$ est asymptotiquement distribuée comme suit :

$$\lim_{n \to +\infty} P(\sqrt{n} D_n < y) = \sum_{-\infty}^{+\infty} (-1)^k e^{-2k^2 y^2} = K(y).$$

Le test de Kolmogorov-Smirnov s'appuie sur la table de Kolmogorov-Smirnov, présentée table 4.3, qui fournit un test de :

$$\begin{cases} H^0 : F(x) = G(x), \\ H^1 : F(x) \neq G(x). \end{cases}$$

Dans la table de Kolmogorov-Smirnov, la région critique est définie par $D_n > d(n)$ (cf. table 4.3). $d(n)$ représente la valeur du seuil critique et $\alpha = 1 - P$ correspond au risque de première espèce, c'est-à-dire au fait de rejeter H^0 alors qu'elle est vraie. Par exemple, pour $P = 0.95$, soit au seuil $\alpha = 0.05$, si $n > 100$, la région critique est $D_n > \frac{1.358}{\sqrt{n}}$; pour $P = 0.99$, $\alpha = 0.01$ et $D_n > \frac{1.629}{\sqrt{n}}$.

Si $n < 100$, ces valeurs sont trop grandes et correspondent à un seuil plus petit, se reporter à la table de Kolmogorov-Smirnov.

En résumé, le test de Kolmogorov-Smirnov est basé sur la comparaison entre la fonction cumulative de fréquence F pour l'échantillon et la fonction de répartition G pour la distribution.

En pratique, nous calculons, pour l'ensemble des pixels de peau considérés, deux histogrammes normalisés, h_{Cb} et h_{Cr}, qui donnent respectivement les fréquences d'observation des chrominances Cb et Cr. Nous calculons ensuite les moyennes et les écart-types de ces histogrammes normalisés. À partir de ces valeurs statistiques, nous créons deux distributions gaussiennes théoriques g_{Cb} et g_{Cr}. La figure 4.11, page 95, illustre un exemple d'histogramme normalisé en Cb, h_{Cb}, et la gaussienne théorique associée, g_{Cb}.

Puis nous calculons deux fonctions cumulatives de fréquence, F_{Cb} et F_{Cr}, définies par rapport aux histogrammes normalisés correspondants selon les formules suivantes :

$$\begin{aligned} F_{Cb}(c) &= \sum_{c_i} h_{Cb}(c_i), \\ F_{Cr}(c) &= \sum_{c_j} h_{Cr}(c_j). \end{aligned}$$

Nous calculons aussi deux fonctions de répartition, G_{Cb} et G_{Cr}, définies par rapport aux gaussiennes théoriques associées selon les formules suivantes :

TAB. 4.3 – Table de Kolmogorov-Smirnov.

n	$P=.80$	$P=.90$	$P=.95$	$P=.98$	$P=.99$
1	.90000	.95000	.97500	.99000	.99500
2	.68377	.77639	.84189	.90000	.92929
3	.56481	.63604	.70760	.78456	.82900
4	.49265	.56522	.62394	.68887	.73424
5	.44698	.50945	.56328	.62718	.66853
6	.41037	.46799	.51926	.57741	.61661
7	.38148	.43607	.48342	.53844	.57581
8	.35831	.40962	.45427	.50654	.54179
9	.33910	.38746	.43001	.47960	.51332
10	.32260	.36866	.40925	.45662	.48893
11	.30829	.35242	.39122	.43670	.46770
12	.29577	.33815	.37543	.41918	.44905
13	.28470	.32549	.36143	.40362	.43247
14	.27481	.31417	.34890	.38970	.41762
15	.26588	.30397	.33760	.37713	.40420
16	.25778	.29472	.32733	.36571	.39201
17	.25039	.28627	.31796	.35528	.38086
18	.24360	.27851	.30936	.34569	.37062
19	.23735	.27136	.30143	.33685	.36117
20	.23156	.26473	.29408	.32866	.35241
21	.22617	.25858	.28724	.32104	.34427
22	.22115	.25283	.28087	.31394	.33666
23	.21645	.24746	.27490	.30728	.32954
24	.21205	.24242	.26931	.30104	.32286
25	.20790	.23768	.26404	.29516	.31657
26	.20399	.23320	.25907	.28962	.31064
27	.20030	.22898	.25438	.28438	.30502
28	.19680	.22497	.24993	.27942	.29971
29	.19348	.22117	.24571	.27471	.29466
30	.19032	.21756	.24170	.27023	.28987
31	.18732	.21412	.23788	.26596	.28530
32	.18445	.21085	.23424	.26189	.28094
33	.18171	.20771	.23076	.25801	.27677
34	.17909	.20472	.22743	.25429	.27279
35	.17659	.20185	.22425	.25073	.26897
36	.17418	.19910	.22119	.24732	.26532
37	.17188	.19646	.21826	.24404	.26180
38	.16966	.19392	.21544	.24089	.25843
39	.16753	.19148	.21273	.23786	.25518
40	.16547	.18913	.21012	.23494	.25205
41	.16349	.18687	.20760	.23213	.24904
42	.16158	.18468	.20517	.22941	.24613
43	.15974	.18257	.20283	.22679	.24332
44	.15796	.18053	.20056	.22426	.24060
45	.15623	.17856	.19837	.22181	.23798
46	.15457	.17665	.19625	.21944	.23544
47	.15295	.17481	.19420	.21715	.23298
48	.15139	.17302	.19221	.21493	.23059
49	.14987	.17128	.19028	.21277	.22828
50	.14840	.16959	.18841	.21068	.22604

n	$P=.80$	$P=.90$	$P=.95$	$P=.98$	$P=.99$
51	.14697	.16796	.18659	.20864	.22386
52	.14558	.16637	.18482	.20667	.22174
53	.14423	.16483	.18311	.20475	.21968
54	.14292	.16332	.18144	.20289	.21768
55	.14164	.16186	.17981	.20107	.21574
56	.14040	.16044	.17823	.19930	.21384
57	.13919	.15906	.17669	.19758	.21199
58	.13801	.15771	.17519	.19590	.21019
59	.13686	.15639	.17373	.19427	.20844
60	.13573	.15511	.17231	.19267	.20673
61	.13464	.15385	.17091	.19112	.20506
62	.13357	.15263	.16956	.18960	.20343
63	.13253	.15144	.16823	.18812	.20184
64	.13151	.15027	.16693	.18667	.20029
65	.13052	.14913	.16567	.18525	.19877
66	.12954	.14802	.16443	.18387	.19729
67	.12859	.14693	.16322	.18252	.19584
68	.12766	.14587	.16204	.18119	.19442
69	.12675	.14483	.16088	.17990	.19303
70	.12586	.14381	.15975	.17863	.19167
71	.12499	.14281	.15864	.17739	.19034
72	.12413	.14183	.15755	.17618	.18903
73	.12329	.14087	.15649	.17498	.18776
74	.12247	.13993	.15544	.17382	.18650
75	.12167	.13901	.15442	.17268	.18528
76	.12088	.13811	.15342	.17155	.18408
77	.12011	.13723	.15244	.17045	.18290
78	.11935	.13636	.15147	.16938	.18174
79	.11860	.13551	.15052	.16832	.18060
80	.11787	.13467	.14960	.16728	.17949
81	.11716	.13385	.14868	.16626	.17840
82	.11645	.13305	.14779	.16526	.17732
83	.11576	.13226	.14691	.16428	.17627
84	.11508	.13148	.14605	.16331	.17523
85	.11442	.13072	.14520	.16236	.17421
86	.11376	.12997	.14437	.16143	.17321
87	.11311	.12923	.14355	.16051	.17223
88	.11248	.12850	.14274	.15961	.17126
89	.11186	.12779	.14195	.15873	.17031
90	.11125	.12709	.14117	.15786	.16938
91	.11064	.12640	.14040	.15700	.16846
92	.11005	.12572	.13965	.15616	.16755
93	.10947	.12506	.13891	.15533	.16666
94	.10889	.12440	.13818	.15451	.16579
95	.10833	.12375	.13746	.15371	.16493
96	.10777	.12312	.13675	.15291	.16408
97	.10722	.12249	.13606	.15214	.16324
98	.10668	.12187	.13537	.15137	.16242
99	.10615	.12126	.13469	.15061	.16161
100	.10563	.12067	.13403	.14987	.16081
$n > 100$	$1.073/\sqrt{n}$	$1.223/\sqrt{n}$	$1.358/\sqrt{n}$	$1.518/\sqrt{n}$	$1.629/\sqrt{n}$

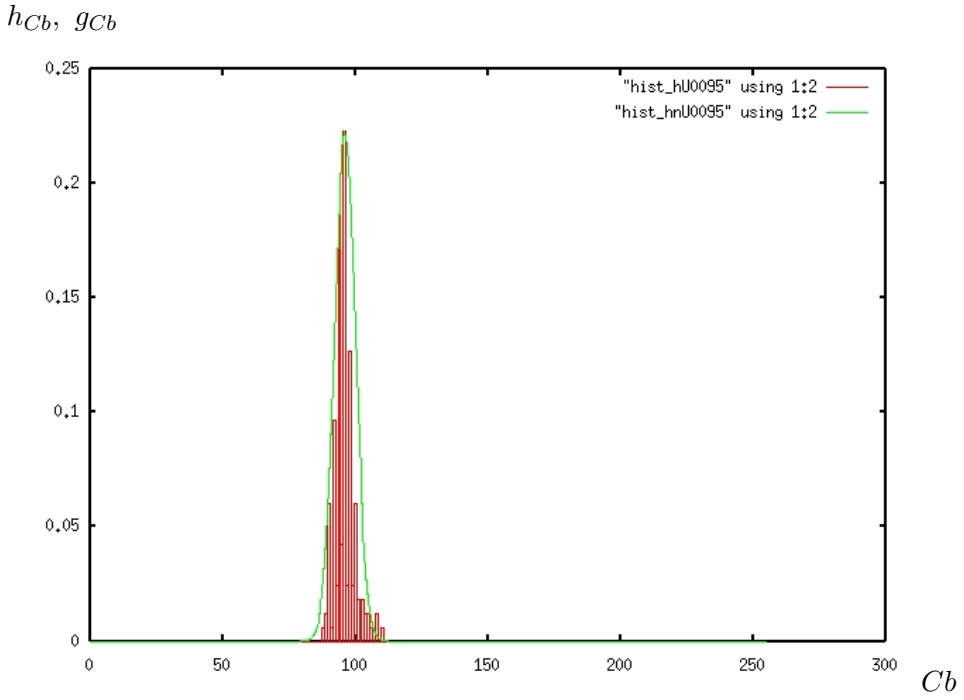


FIG. 4.11 – Exemple d'histogramme normalisé h_{Cb} et de gaussienne théorique g_{Cb} pour la composante Cb.

$$G_{Cb}(c) = \sum_{c_i} g_{Cb}(c_i) \equiv \int_{c_i} g_{Cb}(c_i),$$
$$G_{Cr}(c) = \sum_{c_j} g_{Cr}(c_j) \equiv \int_{c_i} g_{Cr}(c_i).$$

La figure 4.12 présente les fonctions F_{Cb} et G_{Cb} correspondant respectivement à l'histogramme normalisé et à la gaussienne théorique associée de la figure 4.11.

Ensuite nous cherchons l'écart maximum entre la fonction cumulative de fréquence et la fonction de répartition et finalement nous comparons cet écart aux seuils critiques proposés dans la table de Kolmogorov-Smirnov et obtenons le résultat souhaité sur la gaussianité de la distribution. De façon plus précise, par exemple pour une distribution de chrominance Cb, nous cherchons l'écart maximum entre F_{Cb} et G_{Cb} : $D_n = \sup_{c_i} |F_{Cb}(c_i) - G_{Cb}(c_i)|$ et nous comparons cet écart avec les valeurs critiques tabulées, nous rejetons l'hypothèse H^0 aux niveaux $\alpha = 5\%$ et $\alpha = 1\%$ lorsque l'écart maximum observé est respectivement supérieur à $D_n > \frac{1.358}{\sqrt{n}}$ et $D_n > \frac{1.629}{\sqrt{n}}$. Dans le cas où l'écart est supérieur à la valeur critique, la distribution de chrominance Cb considérée n'est alors pas gaussienne.

Résultats : Les tests de Kolmogorov-Smirnov ont été effectués sur les distributions en Cb et Cr formées, d'une part, par l'ensemble des pixels de peau de notre base de données et, d'autre part, par des pixels de peau détectés pour différents individus dans plusieurs séquences vidéo. Les seuils pour cette détection sont les seuils initiaux déterminés par un

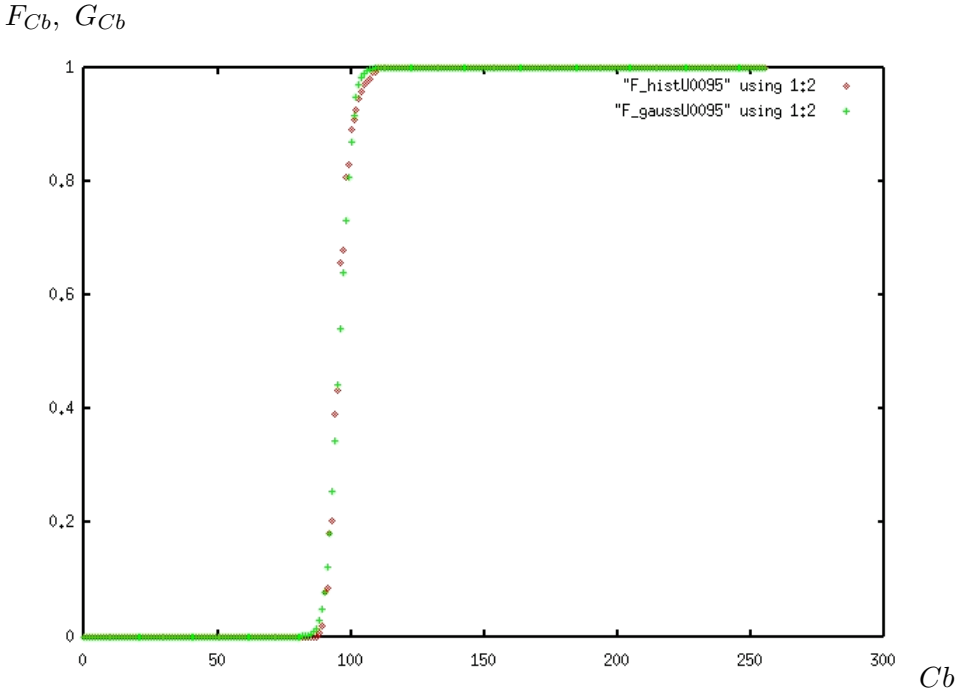


FIG. 4.12 – Exemple de fonction cumulative de fréquence F_{Cb} et de fonction de répartition G_{Cb} pour la composante Cb.

opérateur humain (cf. partie 4.2.3). En prenant en compte les pixels de peau du visage et ceux des mains, nous obtenons des distributions dont les effectifs sont assez importants (quelques centaines de pixels) pour que les tests soient pertinents.

Concernant l'ensemble de notre base de données de pixels de peau, les histogrammes présentés sur la figure 4.7 ne semblent pas gaussiens. Sur l'ensemble des images testées, ce qui représente environ 9600 images, quand nous utilisons les seuils initiaux pré-cités déterminés manuellement pour la détection de peau, au seuil $\alpha = 5\%$, environ 70% des distributions en Cb sont gaussiennes et environ 75% des distributions en Cr aussi. Néanmoins les résultats varient grandement en fonction des séquences vidéo. Quelques séquences vidéo présentent des distributions quasiment tout le temps gaussiennes en Cb et Cr, alors que d'autres présentent des distributions très rarement gaussiennes.

Nous pouvons donc affirmer, au terme de cette étude statistique, que la peau a des composantes de chrominance à majorité gaussiennes dans l'espace $YCbCr$. Néanmoins, la variabilité des résultats que nous avons observés montre que ce n'est pas une raison suffisante pour réaliser une adaptation automatique des seuils en se basant sur cette propriété de gaussianité. Nous avons donc réalisé une adaptation automatique des seuils sans tenir compte d'un modèle gaussien pour les chrominances Cb et Cr.

4.2.4.2 Adaptation automatique des seuils Cb_{min}, Cb_{max}, Cr_{min} et Cr_{max}

Dans notre système, la détection de peau dans l'espace couleur $YCbCr$ est basée sur quatre seuils qui sont notés Cb_{min}, Cb_{max}, Cr_{min} et Cr_{max}. Ces seuils définissent un rectangle dans le sous-espace couleur $CbCr$ que nous appellerons rectangle de détection. Tous les pixels à

l'intérieur sont considérés comme étant des pixels de peau.

Lorsqu'un nouvel objet est détecté par les étapes de segmentation et de suivi temporel, cet objet se voit attribuer les quatre seuils initiaux de détection de peau (cf. partie 4.2.3). Ces seuils initiaux, qui définissent le rectangle de détection initial, sont : $Cb \in [86, 140]$, $Cr \in [139, 175]$. Le rectangle initial est le grand rectangle noir de la figure 4.13.

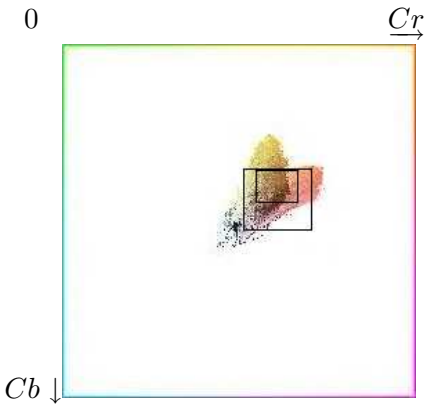


FIG. 4.13 – Rectangle initial de détection en $CbCr$ et rectangle adapté.

La figure 4.13 donne en plus un rectangle de détection pour une personne donnée, c'est le petit rectangle noir inclus dans le rectangle initial. Ce rectangle adapté a été obtenu en ne prenant que les pixels de peau de cette personne sur plusieurs images.

L'idée de l'adaptation automatique est de passer du rectangle initial au rectangle adapté de façon progressive par des transformations élémentaires sur le rectangle de détection, ou, plus précisément, par des opérations élémentaires sur les intervalles de détection.

Sans utiliser de propriété de gaussianité pour les chrominances Cb et Cr, nous adaptons automatiquement les seuils Cb_{min}, Cb_{max}, Cr_{min} et Cr_{max}, c'est-à-dire les intervalles de détection et, par conséquent, le rectangle de détection, en fonction des moyennes statistiques μ_{Cb} et μ_{Cr} calculés sur les pixels de peau détectés.

Trois transformations élémentaires sont envisagées pour chaque intervalle de détection :

1. réduction ;
2. translation ;
3. réinitialisation.

Les distributions de pixels considérés sont constituées de l'ensemble des pixels formés par le visage et les mains. Les transformations élémentaires sont appliquées aux deux intervalles de détection. Nous allons les détailler pour un intervalle de détection. Le rectangle de détection courant est contraint de rester dans le rectangle de détection initial afin que les chrominances ne s'éloignent pas de la gamme de couleurs de peau autorisée au départ.

Réduction : L'intervalle est réduit d'une unité de couleur[6] afin que le milieu de l'intervalle soit plus proche de la moyenne de chrominance correspondante. Soit le seuil haut est décrémenté, soit le seuil bas est incrémenté. Quand l'intervalle atteint une taille de 15 unités de couleur (valeur choisie de manière empirique), la réduction stoppe afin de garder

[6]Une unité de couleur correspond à la valeur du pas de quantification.

une gamme étroite de couleurs différentes. Lors de la recherche des seuils initiaux, les rectangles étaient, en taille, dans ces ordres de grandeur. Le rectangle adapté minimal auquel nous pouvons aboutir a donc les dimensions 15×15 en $Cb \times Cr$. Nous ne réduisons le rectangle que d'une ligne et d'une colonne pour obtenir une adaptation progressive.

Translation : Si le milieu de l'intervalle est distant de moins de 5 unités de couleur (valeur choisie de manière empirique) de la moyenne de la chrominance correspondante, l'intervalle est translaté de 2 unités vers cette moyenne. Les seuils haut et bas sont tous deux incrémentés ou décrémentés. Le fait de contrôler la distance maximum du milieu de l'intervalle à la moyenne permet de ne pas se focaliser sur une mauvaise gamme de couleurs. Le fait de ne modifier les seuils que d'une unité permet encore une fois une adaptation progressive. À cause de la contrainte pour le rectangle courant de rester dans le rectangle initial, il faut d'abord qu'il y ait eu au préalable quelques réductions du rectangle courant pour que ces translations puissent avoir lieu.

Réinitialisation : Cette réinitialisation n'a lieu que si, dans l'image courante, aucune tache de peau n'a été détectée. Sinon, nous considérons que les seuils sont toujours valides. Dans le cas où aucune zone de peau n'a été détectée, les seuils sont réinitialisés aux seuils du rectangle initial, afin de recommencer le processus.

Ces trois transformations élémentaires sont appliquées lors du traitement de chaque nouvelle image de la séquence vidéo. La détection de peau s'améliore, en s'adaptant progressivement à la couleur de peau de chaque individu détecté et en réduisant par conséquent les mauvaises détections. La stabilisation des seuils lors de l'adaptation prend généralement une trentaine d'images, ce qui correspond à environ une seconde de temps d'acquisition.

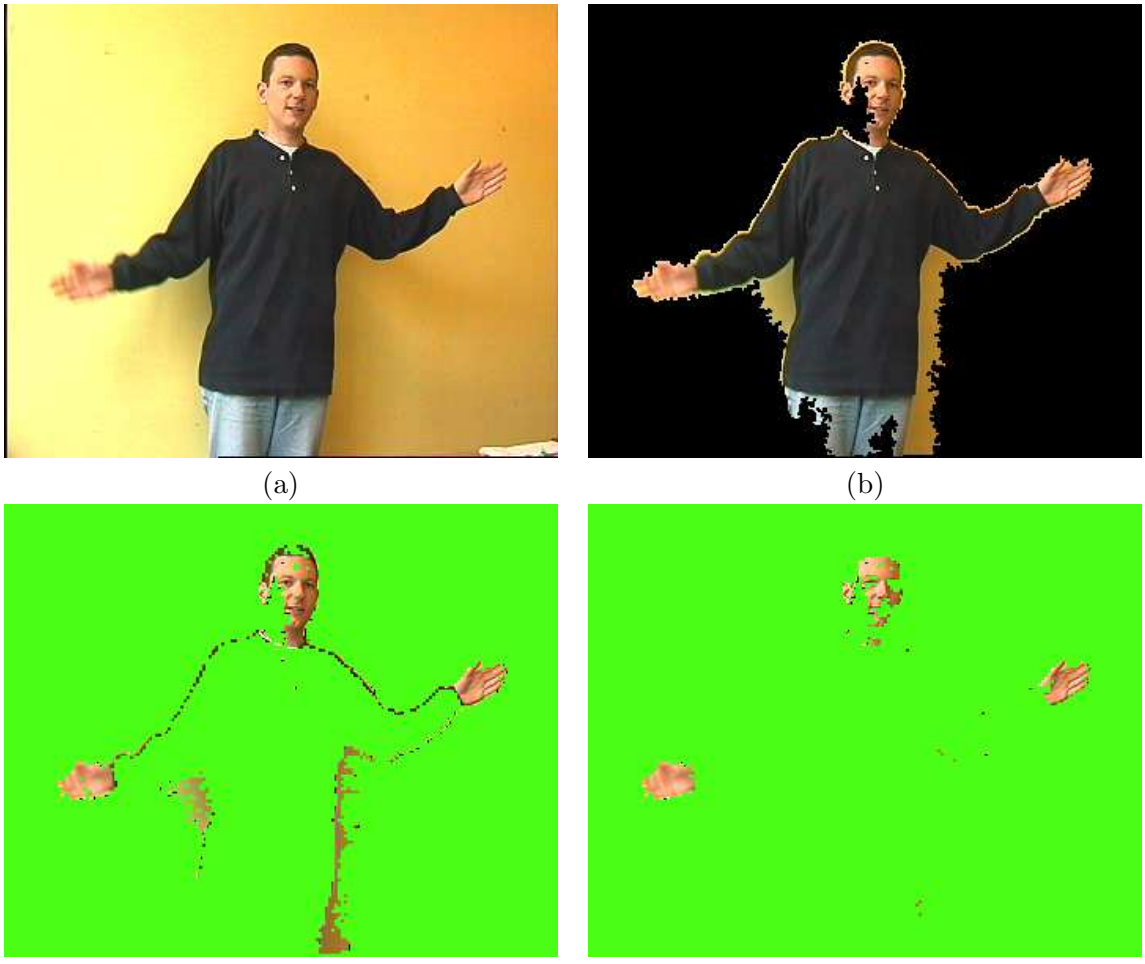
La figure 4.14 illustre un exemple d'adaptation. L'image originale est en (a), l'image segmentée en (b), (c) est l'image de pixels de peau détectés sans adaptation et (d) l'image des pixels de peau détectés avec adaptation. Nous voyons que l'adaptation permet de ne détecter que les zones de peau et réduit les fausses détections, dues ici au fond de couleur jaune qui, avec l'ombre de la personne, se retrouve dans les tons bruns.

4.3 Localisation et suivi temporel

4.3.1 Introduction

Dans notre système, nous voulons détecter et suivre au cours du temps le visage et les mains, avec la distinction main droite / main gauche. Le suivi temporel est rendu difficile par les occultations et les réunions temporelles entre *ROI*. Nous considérons, pour cette étape de localisation et de suivi temporel du visage et des mains, qu'une seule personne est présente dans chaque boîte englobante rectangulaire issue de la segmentation. Si plusieurs individus se sont réunis et ne forment qu'un seul objet du point de vue de la caméra (réunion temporelle), notre algorithme tentera de trouver et de suivre au cours du temps un visage et deux mains mais il y aura beaucoup plus d'erreurs de localisation et de suivi temporel étant donné la complexité de la tâche.

Dans cette partie, nous commencerons par détailler le traitement préliminaire réalisé à l'issue de l'étape de détection de peau. Ce traitement permet de passer d'un ensemble de pixels de peau à un ensemble de régions de peau affectées à chaque personne selon leur position dans l'image. Puis nous présenterons les données et les méthodes utilisées pour réaliser la localisation initiale du visage et des mains et la méthode choisie pour éventuellement



(a) (b)

(c) (d)

FIG. 4.14 – Exemple d'adaptation. (a) image originale, (b) image segmentée, (c) pixels de peau détectés sans adaptation et (d) pixels de peau détectés avec adaptation.

réinitialiser leurs localisations. Ensuite nous détaillerons l'approche choisie pour suivre au cours du temps le visage et les mains.

4.3.2 Traitement préliminaire

4.3.2.1 Étiquetage en composantes connexes

Après la détection de peau, nous disposons d'une image qui contient tous les pixels ayant été reconnus comme étant des pixels de peau. Ces pixels peuvent être éparpillés dans l'ensemble formé par l'union des boîtes englobantes rectangulaires des individus segmentées et ne forment pas, au début, de zones de peau à proprement parler. La première étape après la détection de peau est l'étiquetage en composantes connexes des pixels de peau détectés. Elle permet d'obtenir des taches de peau étiquetées suivant leur place dans l'image.

4.3.2.2 Calcul des boîtes englobantes rectangulaires

Pour pouvoir traiter efficacement les régions de peau nous devons avoir plus d'informations sur chaque région, comme sa surface, ses dimensions, les coordonnées de son centre, etc. Une boîte englobante rectangulaire est donc calculée pour chaque zone de peau par la même méthode que le calcul des boîtes englobantes rectangulaires autour des objets lors de l'étape de segmentation 2D spatio-temporelle (cf. partie 2.4.2). Certains descripteurs, comme la surface par exemple, sont aussi calculés.

4.3.2.3 Filtrage sur la surface

La détection de peau pouvant amener un grand nombre de taches parasites, il est utile de réaliser un filtrage pour éliminer les taches de peau trop petites. Ces dernières peuvent être du bruit ou des fausses détections. Nous considérons que les zones de peau intéressantes ont au moins une surface de quelques dizaines de pixels en résolution 320×240.

Le filtrage est effectué sur la surface des boîtes englobantes rectangulaires des régions de peau. Lors de la localisation initiale du visage et des mains, aucun filtrage n'est fait afin de ne pas éliminer une tache pertinente. Après localisation initiale, quand nous avons trouvé les positions initiales du visage et des deux mains, le filtrage sur la surface est effectué. Nous notons S_{min} la surface de la zone de peau la plus petite entre celles retenues comme étant le visage ou les mains, en supposant que la détection et la localisation sont satisfaisantes. Toutes les taches de peau qui ont alors une boîte englobante rectangulaire dont la surface est inférieure au tiers de S_{min} sont éliminées. Le facteur $1/3$ a été choisi de manière empirique en fonction de tests menés sur l'évolution de la surface des boîtes englobantes rectangulaires des taches de peau du visage et des mains sur une douzaine de séquences vidéo.

4.3.2.4 Affectation des taches de peau à la personne correspondante

Nous possédons les boîtes englobantes rectangulaires des objets segmentés lors de l'étape de segmentation 2D spatio-temporelle et celles des zones de peau.

Comme précisé dans l'introduction de cette partie, nous considérons qu'un seul individu est présent dans chaque boîte englobante rectangulaire issue de la segmentation.

En fonction des positions des centres des boîtes englobantes rectangulaires des régions de peau par rapport aux coordonnées des boîtes englobantes rectangulaires issues de la segmentation, les taches de peau sont affectées à l'objet dont la boîte englobante rectangulaire

contient le centre de ces taches. Les taches de peau sont donc normalement affectées à la personne correspondante. Cela nous permet de ne rechercher le visage et les mains d'un individu que parmi les zones de peau qui sont susceptibles de l'être.

4.3.3 Données et méthodes utilisées

Après avoir réalisé le traitement préliminaire pour l'ensemble des pixels de peau, nous disposons de régions de peau de taille suffisante affectées aux personnes selon leur position. Nous allons maintenant présenter les méthodes utilisées pour réaliser la localisation initiale et le suivi temporel du visage et des mains.

Nous voulons utiliser une méthode semblable pour localiser et suivre au cours du temps le visage et les mains. L'idée est d'utiliser certains critères spatiaux comme la taille, la position et la distance des taches de peau à des points précis, pour trouver la zone la plus probable d'être l'une des *ROI* que nous recherchons. Ces critères spatiaux et certains paramètres vont permettre le calcul de listes triées de taches de peau. Ensuite, grâce à des heuristiques liées à la morphologie humaine et à une méthode de somme de rangs minimale dans des listes choisies, nous pourrons localiser et suivre au cours du temps le visage et les mains.

4.3.3.1 Présentation des listes et des paramètres

La localisation et le suivi temporel du visage et des mains se font grâce à l'utilisation de listes triées. Les listes utilisées sont basées sur des critères de taille, de position et de distance. Selon la liste, l'ensemble des régions de peau affectées à l'individu sont présentes ou non.

Les listes basées sur les critères de taille et de position sont :

- L_s : Liste des taches les plus grandes (surface) ;
- L_h : Liste des taches les plus hautes ;
- L_d : Liste des taches les plus à droite ;
- L_g : Liste des taches les plus à gauche.

L'ensemble des taches de peau affectées à chaque personne sont présentes dans les listes L_s, L_h, L_d et L_g. Ces listes sont triées par **ordre décroissant**. Pour les listes L_h, L_d et L_g, la position est définie par rapport à la boîte englobante rectangulaire de l'individu correspondant. Les indices $_d$ et $_g$, respectivement mis pour droite et gauche, sont à interpréter du point de vue d'une personne filmée et non du point de vue de la caméra.

Les listes basées sur le critère de distance sont de deux types selon la distance (spatiale ou spatio-temporelle). Pour les listes du premier type, nous utilisons la boîte quadrangulaire, ou quadrangle, présentée dans la partie 2.4.2.

Les listes du premier type, basées sur un critère de distance spatiale, sont :

- L_{hq} : Liste des taches les plus proches du coin haut du quadrangle ;
- L_{dq} : Liste des taches les plus proches du coin droit du quadrangle ;
- L_{gq} : Liste des taches les plus proches du coin gauche du quadrangle.

Les listes L_{hq}, L_{dq} et L_{gq} sont construites en fonction d'un paramètre d_{bqmax} qui définit la distance maximale à un coin du quadrangle où une zone de la liste peut se trouver. Ces listes peuvent donc ne pas contenir l'ensemble des taches de peau affectées à une personne. Le tri se fait par **ordre croissant** en fonction de la distance, les zones en tête de liste sont donc les plus proches des coins. Le paramètre d_{bqmax} est adaptatif en fonction des dimensions *width* et *height* de la boîte englobante rectangulaire de l'individu correspondant. Le facteur

1/4 a été choisi de manière empirique en fonction de la distance respective entre les mains et le visage lorsqu'une personne est debout les bras écartés.

$$d_{bqmax} = \frac{\max(width, height)}{4}$$

Le second type de liste est celui des listes qui utilisent le suivi temporel, c'est-à-dire les dernières localisations connues du visage et des mains. Au vu de la cadence d'acquisition, les localisations consécutives du visage et des mains ne sont pas très éloignées. Il est donc pertinent de calculer les listes triées des régions de peau les plus proches des dernières localisations.

Voici les listes du second type, basées sur un critère de distance spatio-temporelle :
- L_v : Liste des taches les plus proches de la dernière localisation du visage ;
- L_{md} : Liste des taches les plus proches de la dernière localisation de la main droite ;
- L_{mg} : Liste des taches les plus proches de la dernière localisation de la main gauche.

Les listes L_v, L_{md} et L_{mg} sont construites en fonction d'un paramètre d_{tmax} qui définit la distance maximale à une dernière localisation connue où une zone de la liste peut se trouver. Ces listes peuvent donc aussi ne pas contenir toutes les régions de peau. Le tri se fait par **ordre croissant** en fonction de la distance, les taches en tête de liste sont donc les plus proches des dernières localisations connues du visage et des mains. Le paramètre d_{tmax} est adaptatif en fonction des dimensions $width$ et $height$ de la boîte englobante rectangulaire. Le facteur 1/2 a aussi été choisi de façon empirique en fonction de la distance respective entre les mains et le visage lorsqu'une personne est debout les bras écartés.

$$d_{tmax} = \frac{\max(width, height)}{2}$$

Le paramètre d_{bqmax} sert essentiellement à la localisation initiale du visage et des mains. Le paramètre d_{tmax} sert essentiellement au suivi temporel du visage et des mains. d_{tmax} est deux fois plus grand que d_{bqmax} afin d'une part, que la localisation initiale soit précise (prise en compte de peu de zones de peau) et, d'autre part, que le suivi temporel soit robuste en prenant en compte plus de zones de peau.

Une fois ces listes calculées, nous utilisons des heuristiques liées à la morphologie humaine et une méthode de somme de rangs minimale. Les heuristiques seront présentées dans les parties correspondant à la localisation initiale et au suivi temporel du visage et des mains (cf. parties 4.3.4 et 4.3.5).

4.3.3.2 Méthode de somme de rangs minimale

Pour déterminer les taches de peau correspondant aux *ROI* que sont le visage et les mains, que ce soit lors de la localisation initiale ou lors du suivi temporel, nous utilisons une méthode de somme de rangs minimale dans des listes triées.

Cette méthode, comme celle présentée lors du suivi temporel, est aussi basée sur la notion de **rang** dans une liste. Contrairement à la méthode de la première étape du suivi temporel, qui fixe plutôt le rang maximal pour déterminer un couple prédécesseur-successeur cohérent, la méthode utilisée pour la localisation et le suivi temporel du visage et des mains effectue le parcours des zones de peau présentes dans une liste et calcule la somme des rangs de ces taches de peau dans un certain nombre de listes triées. La région de la liste parcourue qui a la somme de rangs minimale dans les listes triées considérées est alors la tache de peau retenue.

4.3.4 Localisation initiale du visage et des mains

Après avoir présenté les listes triées et la méthode de somme de rangs minimale, il faut maintenant déterminer la localisation initiale du visage et des mains. Nous commençons par localiser le visage, d'une part parce que le visage est plus gros que les mains et, d'autre part, parce qu'il a un mouvement plus lent et plus stable que les mains.

4.3.4.1 Localisation initiale du visage

Les heuristiques choisies pour la localisation initiale du visage sont les suivantes : étant en haut du corps, de surface plus grande et situé généralement plus haut que les mains, nous déterminons la localisation initiale du visage en utilisant la liste triée basée sur le critère de taille, L_s et celle basée sur le critère de position en hauteur, L_h.

Nous parcourons la liste L_s et calculons, grâce à la méthode de somme de rangs minimale, la somme des rangs des zones de cette liste pour cette liste et la liste L_h. Comme ces listes sont triées par ordre décroissant, la tache de somme de rangs minimale qui est la plus grande et la plus haute est retenue pour être la localisation initiale du visage.

Ces listes ne sont vides que dans le cas où aucune région de peau n'a été détectée, ou s'il n'y a personne dans l'image. Dans ces deux cas, il n'est pas possible de déterminer une localisation initiale du visage. Dans le cas contraire, il y a forcément une zone de peau qui sera choisie pour être la localisation initiale du visage.

La partie encadrée en pointillés du schéma figure 4.16, page 106, résume la localisation initiale du visage.

4.3.4.2 Localisation initiale des mains

Les mains sont des taches de peau bien plus difficiles à localiser que le visage. Elles sont très ressemblantes entre elles, de taille variable selon qu'elles sont ouvertes ou fermées, mais de plus un être humain peut croiser les bras, cacher ses mains derrière son dos, etc. Ce sont donc des *ROI* difficiles à localiser parmi les éventuelles taches de peau. La vitesse de déplacement des mains est aussi plus élevée que celle du visage.

Les heuristiques choisies pour la localisation initiale des mains exploitent le fait que dans les applications visées, une personne se trouve souvent de face par rapport à la caméra. Pour un individu de face, les mains se trouvent de façon générale de chaque côté de son corps. Nous déterminons donc les positions initiales des mains en utilisant les listes triées basées sur les critères de distance spatiale par rapport au quadrangle : L_{hq}, L_{dq} et L_{gq}.

Si les listes L_{dq} et L_{gq} ne contiennent qu'une seule zone de peau, que ces régions de peau sont différentes entre elles, et différentes de la tache retenue comme localisation du visage, ces régions sont retenues pour être les localisations initiales de la main droite et de la main gauche.

Grâce à cette méthode, les localisations initiales des mains sont déterminées uniquement dans le cas où nous avons peu de chances de commettre une erreur de localisation.

4.3.4.3 Réinitialisation des localisations du visage et des mains

La méthode utilisée pour la localisation initiale des mains est aussi utilisée pour réinitialiser éventuellement les positions du visage et des mains. Cette réinitialisation a lieu à trois conditions :

– Les listes L_{hq}, L_{dq} et L_{gq} contiennent chacune une unique tache.
– Ces trois taches sont distinctes deux à deux.
– Les mains ne se sont pas croisées (le suivi temporel n'a pas localisé la main droite du côté gauche et la main gauche du côté droit).

Dans ces conditions, les nouvelles localisations du visage, de la main droite et de la main gauche sont respectivement la tache de la liste L_{hq}, celle de la liste L_{dq} et celle de la liste L_{gq}.

La position parfaite pour les localisations initiales du visage et des mains est donc de se tenir debout face à la caméra, les bras écartés du corps au niveau des épaules. Si nous considérons la contrainte n°4 de notre système, qui est que la personne doit au moins se trouver une fois dans une **posture de référence**, debout avec les bras écartés, nous sommes sûrs que les localisations initiales des mains seront correctes.

Le schéma figure 4.15 résume d'une part la localisation initiale des mains et d'autre part l'éventuelle réinitialisation de la localisation du visage.

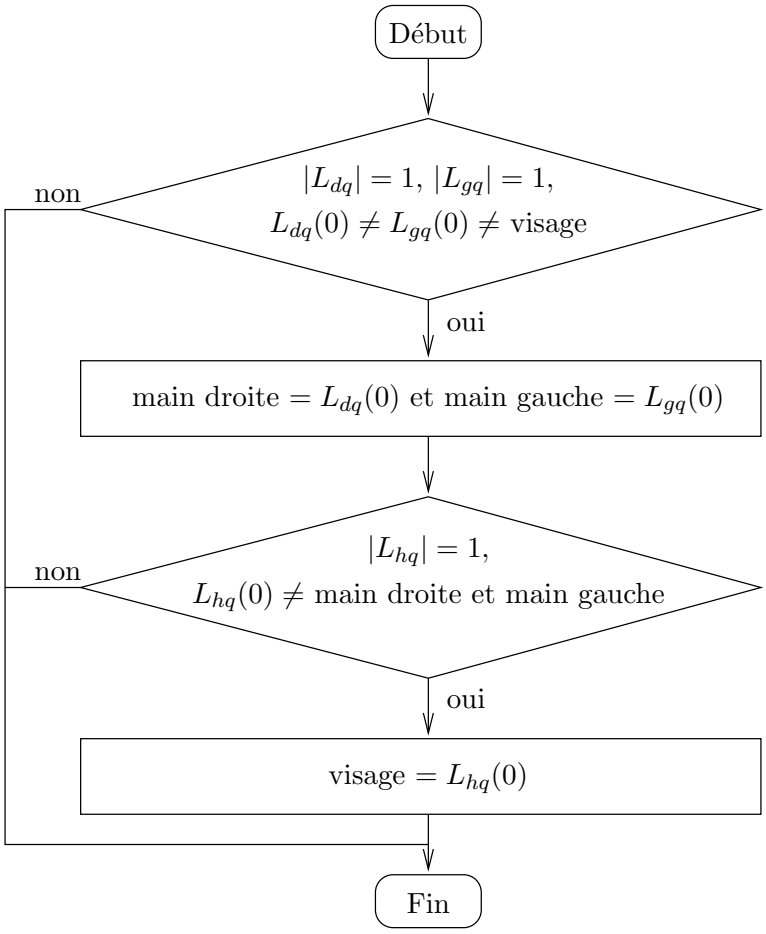


Fig. 4.15 – Schéma explicatif pour la localisation initiale des mains.

4.3.5 Suivi temporel du visage et des mains

Grâce à la localisation initiale du visage et des mains, nous disposons des localisations dans l'image précédente. Nous allons donc utiliser ces informations pour déterminer plus facilement les localisations du visage et des mains dans l'image courante. Nous réalisons d'abord le suivi temporel du visage, parce qu'il a un mouvement plus lent et plus stable que les mains.

4.3.5.1 Suivi temporel du visage

Le suivi temporel du visage utilise les listes triées suivantes : L_s, L_h et L_v, c'est-à-dire les listes qui contiennent respectivement les zones les plus grandes, les plus hautes et les plus proches de la dernière localisation connue du visage.

Pour suivre le visage au cours du temps, nous utilisons la méthode de somme de rangs minimale sur les listes L_s, L_h et L_v, et nous utilisons aussi un calcul d'intersection de boîtes englobantes rectangulaires.

Nous parcourons les régions de la liste L_v si la localisation initiale du visage a eu lieu et que cette liste n'est pas vide, sinon nous parcourons les taches de la liste L_s. Dans le premier cas, la tache de peau qui a une somme de rangs minimale dans les listes L_s, L_h et L_v et qui possède une intersection non nulle avec la boîte englobante rectangulaire du visage dans l'image précédente est retenue pour être la nouvelle localisation du visage dans l'image courante.

Le calcul d'intersection de boîtes englobantes rectangulaires pour le suivi temporel du visage permet de gérer les réunions temporelles entre une (ou les deux) main(s) et le visage. Quand une main passe devant le visage, elle se retrouve réunie avec lui pendant un instant puis se sépare et poursuit sa trajectoire. Grâce à l'intersection de boîtes englobantes rectangulaires, nous sommes sûrs que le visage sera correctement suivi pendant la réunion temporelle et surtout après la séparation temporelle. Une illustration de ce phénomène est donnée dans la partie 4.5.

Dans le second cas, où il n'y a aucune zone de peau dans la liste L_v, si la localisation du visage avait été initialisée, alors aucune région de peau n'est désignée comme nouvelle localisation du visage et la localisation reste la même. Cela arrive dans certains cas de mauvaise segmentation (visage "rogné"). Sinon, dans le cas où la localisation du visage n'avait pas été initialisée, c'est la méthode de localisation initiale du visage, présentée précédemment, qui est utilisée, avec le parcours des taches de la liste L_s et le calcul de la somme de rangs minimale dans les listes L_s et L_h uniquement.

Le schéma figure 4.16 résume la localisation initiale et le suivi temporel du visage, la localisation initiale (encadrée en pointillés) ne sert qu'une fois pour chaque individu.

4.3.5.2 Suivi temporel des mains

Le suivi temporel des mains est très délicat pour plusieurs raisons :

- Les mains sont de petites taches de peau qui peuvent être confondues avec de fausses détections.
- Les mains ont un mouvement souvent beaucoup plus rapide que celui du visage, donc moins stable. Il n'est pas possible, contrairement au visage, d'envisager une intersection de boîtes englobantes rectangulaires pour les mains.
- Les mains peuvent être occultées, donc disparaître de l'image, et il ne faut pas détecter une autre zone dans ce cas, mais garder la dernière localisation connue, ceci afin de

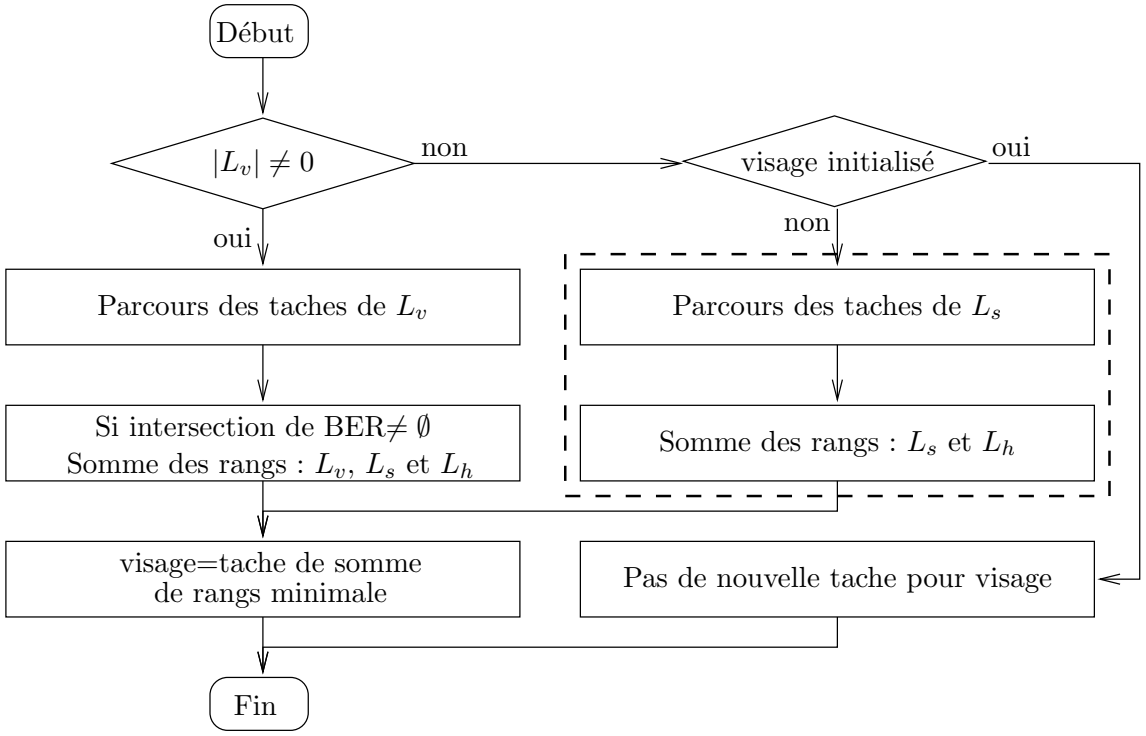


FIG. 4.16 – Schéma explicatif pour la localisation initiale et le suivi temporel du visage.

s'aider du suivi temporel quand elles réapparaîtront.

– Les mains peuvent se réunir au cours du temps entre elles et / ou avec le visage pour ne plus former qu'une seule région et, ici aussi, il ne faut pas détecter une autre tache mais garder la dernière localisation connue.

Un traitement symétrique des mains gauche et droite est proposé, de façon à ne favoriser le suivi temporel d'aucune d'entre elles en particulier.

Nous commençons par déterminer quelle main doit être suivie en premier, puis nous réalisons le suivi temporel d'abord pour cette main et ensuite pour l'autre.

Méthode pour déterminer quelle main doit être suivie en premier : Pour ce faire, nous utilisons des indicateurs pour savoir depuis combien de temps le suivi temporel du visage et des mains est satisfaisant. Ces indicateurs d'activité, notés I_v, I_{md} et I_{mg}, respectivement pour visage, main droite et main gauche, sont à 0 si tout se passe bien, sinon ils indiquent depuis combien d'images le suivi temporel a échoué pour la *ROI* considérée. Nous utilisons aussi la localisation courante du visage, déjà déterminée à cette étape du suivi temporel, et les listes L_{md} et L_{mg}, qui contiennent respectivement les taches de peau les plus proches de la dernière localisation de la main droite et celles les plus proches de la dernière localisation de la main gauche.

Certains cas permettent de faire le choix très rapidement, en fonction de la taille des listes L_{md} et L_{mg}, et de la localisation courante du visage. En effet, si l'une de ces deux listes est vide ou si la première zone de l'une de ces deux listes est la région de peau retenue comme nouvelle localisation du visage, il faut commencer par suivre la main correspondant à l'autre liste car il y possibilité de réunion temporelle entre une main et le visage.

Exemple : si la tache en tête de la liste L_{mg} est la tache retenue comme nouvelle localisation du visage pour l'image courante, il y a un risque que la main gauche se soit réunie avec le visage, il faut donc commencer par suivre la main droite, nous tenterons de suivre la main gauche, pour laquelle le suivi temporel est plus difficile, après.

Si nous ne sommes pas dans le cas ci-dessus, ce qui arrive assez souvent quand les mains sont visibles et assez loin du visage, le choix se fait en fonction de la distance entre les dernières localisations des mains droite et gauche, des zones respectivement en tête des listes L_{md} et L_{mg} et des indicateurs d'activité des mains I_{md} et I_{mg}. Comme la nouvelle localisation d'une main est généralement la première région de sa liste de proximité, ces tests s'appliquent dans les cas restants. Nous suivons la main dont la tache en tête de sa liste de proximité est la plus proche de la dernière localisation connue si son indicateur d'activité est plus petit que celui de l'autre main.

Exemple : si la distance entre la dernière localisation de la main droite et la tache en tête de la liste L_{md} est plus petite que la distance entre la dernière localisation de la main gauche et la zone en tête de la liste L_{mg} et si $I_{md} < I_{mg}$, nous allons suivre d'abord la main droite puis la main gauche. Sinon nous suivrons d'abord la main gauche puis la main droite.

Le schéma figure 4.17 résume cette méthode pour déterminer quelle main doit être suivie en premier.

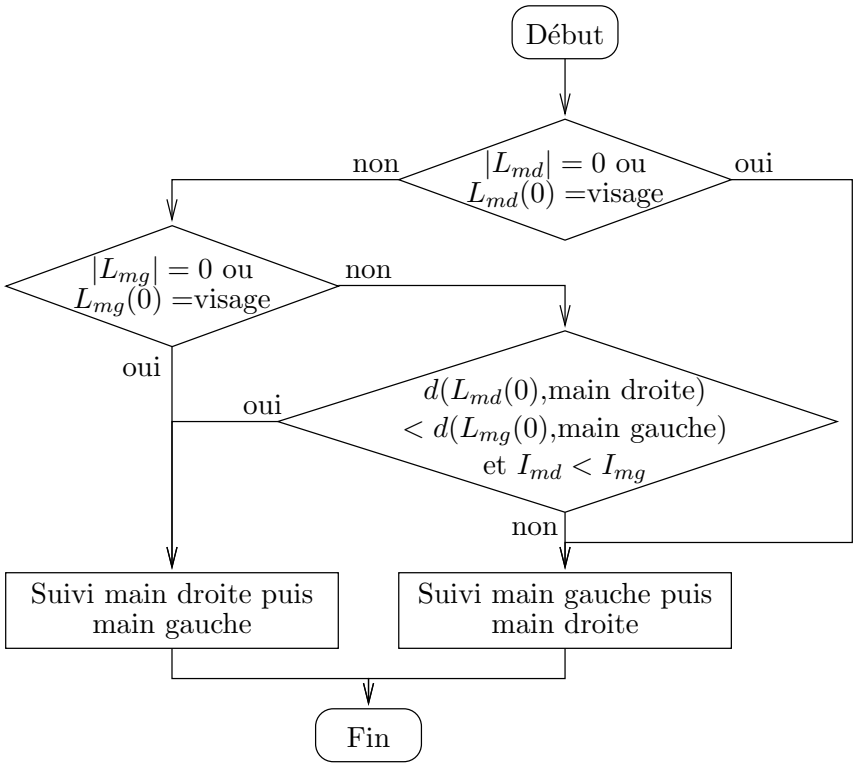


FIG. 4.17 – Schéma explicatif pour déterminer la main à suivre en premier.

Méthode (commune aux deux mains) pour le suivi temporel d'une main : Cette méthode ressemble beaucoup à celle utilisée pour le suivi temporel du visage, à ceci près qu'il n'y a pas de calcul d'intersection de boîtes englobantes rectangulaires et qu'un traitement

particulier est ajouté si la main est proche du visage, pour gérer les réunions temporelles entre une main (ou les deux) et le visage.

Le suivi temporel d'une main utilise les listes suivantes : L_{md} ou L_{mg}, L_s, L_h, L_d ou L_g et L_v, c'est-à-dire les listes qui contiennent respectivement les régions les plus proches de la dernière localisation de la main considérée, les plus grandes, les plus hautes, les plus à droite ou à gauche du côté de la main considérée et les plus proches de la dernière localisation du visage.

Nous parcourons les taches de la liste L_{md} ou L_{mg} et la méthode de somme de rangs minimale sur les listes L_{md} ou L_{mg}, L_s, L_h et L_d ou L_g est appliquée. La zone de peau qui a une somme de rangs minimale est retenue pour être la nouvelle localisation de la main considérée dans l'image courante.

Si cela n'a pas permis de retenir une tache de peau comme nouvelle localisation de la main considérée et que nous sommes dans le cas où la main est proche du visage, nous appliquons de nouveau la méthode de somme de rangs minimale mais en remplaçant la liste L_{md} ou L_{mg} par la liste L_v. Cela permet de gérer les réunions temporelles entre la main considérée et le visage. Dans ce cas (main proche du visage), s'il n'y a toujours aucune zone retenue comme nouvelle localisation de la main, nous considérons que la main et le visage se sont réunis, ne formant qu'une seule région de peau, identifiée comme le visage. La nouvelle localisation de la main considérée est alors la même que sa dernière localisation connue, ceci afin de conserver une distance non nulle entre la localisation de la main et la localisation du visage.

Si, finalement, la main n'est pas proche du visage mais qu'aucune zone de peau n'a été retenue comme nouvelle localisation de la main considérée (en parcourant la liste L_{md} ou L_{mg}), nous considérons que la main est occultée et la nouvelle localisation est identique à la dernière localisation connue, ceci pour s'aider du suivi temporel quand cette main réapparaîtra.

Quand une réunion temporelle survient entre les deux mains, la tache résultante est choisie comme nouvelle localisation pour l'une des deux mains, la première suivie, et lors de la séparation temporelle suivante, la zone de peau la plus à gauche devient la main gauche et la tache de peau la plus à droite, la main droite. Il n'est pas possible, en effet, de faire un choix toujours correct avec les informations dont nous disposons. C'est le même problème quand il survient une réunion temporelle entre les deux mains et le visage, nous ne pouvons savoir si, après la séparation temporelle, les mains se sont croisées ou non. En cas de réunion temporelle entre les deux mains, la première suivie est alors suivie correctement, sa nouvelle localisation est celle de la zone de peau résultante, l'autre main est considérée comme occultée et sa nouvelle localisation est identique à la dernière localisation connue, ceci pour conserver une distance non nulle entre les localisations retenues pour les mains.

Le schéma figure 4.18 résume la méthode utilisée pour le suivi temporel d'une main (droite ou gauche). La main droite est considérée dans la figure 4.18.

4.4 Données bas-niveau extraites

Plusieurs données bas-niveau sont extraites lors de cette étape de traitement. Pour chaque boîte englobante rectangulaire issue de la segmentation 2D spatio-temporelle, nous disposons de :

- la localisation du visage ;
- la localisation de la main droite ;
- la localisation de la main gauche ;

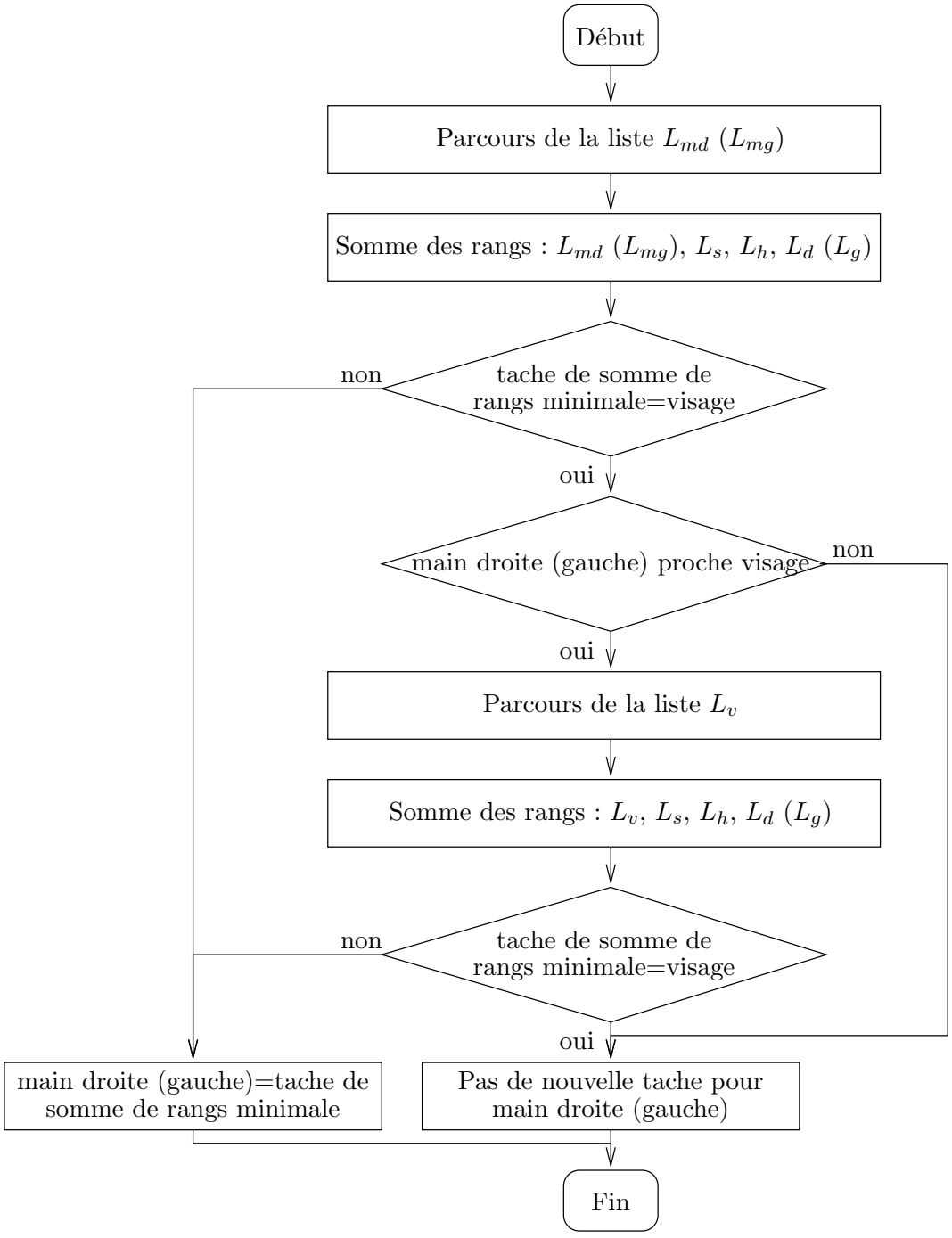


FIG. 4.18 – Schéma explicatif de la méthode de suivi temporel pour une main.

– la dernière localisation du visage ;
– la dernière localisation de la main droite ;
– la dernière localisation de la main gauche.

Pour chacune des taches de peau spécifiques que sont le visage, la main droite et la main gauche, nous avons aussi :

– le masque de segmentation ;
– le centre de la tache de peau ;
– la boîte englobante rectangulaire ;
– l'indicateur d'activité ;
– la surface ;
– la trajectoire.

Parmi toutes ces données bas-niveau extraites, les données concernant le visage seront utilisées lors de la seconde étape du suivi temporel et lors de l'étape d'interprétation de haut niveau sur la reconnaissance de postures.

4.5 Résultats

Les méthodes de localisation et de suivi temporel du visage et des mains se sont révélées précises et robustes. Les résultats sont satisfaisants et la majorité des difficultés liés aux occultations, aux réunions et séparations temporelles qui peuvent survenir est résolue de façon correcte. Afin de voir si l'algorithme est robuste, nous avons testés des *scenarii* présentant ces difficultés. Les voici, des cas très faciles aux cas très difficiles :

– mains basses, moyennes ou hautes ;
– croisement des mains en bas et en haut sans réunion temporelle ;
– réunion temporelle entre les mains sans réunion temporelle avec le visage, dans le sens vertical et dans le sens horizontal ;
– occultation d'une main, voire des deux, et réapparition(s) pas forcément au(x) même(s) endroit(s) ;
– occultation du visage par une main ou un bras ;
– réunion temporelle entre une main et le visage, avec ou sans occultation et passage ou non de l'autre côté ;
– réunion temporelle entre les deux mains et le visage.

L'algorithme présenté pour la localisation et le suivi temporel du visage et des mains prend en compte la majorité des cas présentés et les gère correctement. Il est rapide et travaille conjointement avec l'étape de détection de peau puisque l'adaptation des seuils de détection est réalisée en fonction des résultats de localisation. Les résultats de localisation et de suivi temporel sont donc de bonne qualité.

Les pages suivantes présentent quelques résultats de localisation et de suivi temporel obtenus pour différentes séquences vidéo. La figure 4.19, page 112, illustre les résultats obtenus pour une séquence vidéo particulière. Nous avons choisi, pour la présentation des résultats, d'encadrer les régions de peau localisées et suivies au cours du temps par leurs boîtes englobantes rectangulaires, selon un code de couleur particulier :

– vert pour la localisation courante du visage ;
– bleu pour la localisation courante de la main droite ;
– rouge pour la localisation courante de la main gauche ;
– mauve pour les dernières localisations.

La figure 4.19 montre quelques résultats de localisation et de suivi temporel du visage et des mains avec les conventions présentées ci-dessus. Les images de gauche présentent les images originales de la séquence, les images du milieu les résultats de détection de peau (à part la dernière série d'images) et les images de droite les résultats de localisation et de suivi temporel (à part la dernière série d'images), avec la boîte englobante rectangulaire de la personne, son quadrangle et les boîtes englobantes rectangulaires des zones de peau localisées.

Nous pouvons tout d'abord observer que la détection de peau est quasi parfaite (colonne du milieu). La première série d'images illustre une localisation initiale correcte pour le visage et les mains. Nous pouvons voir sur les deuxième et troisième séries d'images une occultation du visage par un bras et le fait qu'à sa réapparition le suivi temporel localise le visage correctement. Sur la quatrième série d'images, nous illustrons le phénomène cité plus haut de croisement des mains sans réinitialisation des localisations. La cinquième série d'images illustre ce qui se passe en cas de mauvaise segmentation, l'individu a été coupé en deux au niveau du torse et son visage n'est pas détecté, l'image du milieu illustre ce qui se passe quand on réalise une fusion des boîtes englobantes rectangulaires lors de la première étape du suivi temporel, l'individu est "raccommodé" et ses mains sont correctement suivies, de même que son visage, qui, ne pouvant être détecté, est localisé au même endroit que précédemment. L'image de droite illustre ce qui se passe lorsqu'on n'effectue pas cette fusion de boîtes. Les deux parties sont considérées comme des personnes seules et l'algorithme tente de détecter un visage, qu'il localise à la place des mains, pour chaque partie du corps. La fusion de boîtes englobantes rectangulaires est un algorithme simple non présenté dans ce mémoire.

La figure 4.20, page 113, présente d'autres résultats avec d'autres personnes dans d'autres conditions d'acquisition (cf. première et deuxième séries d'images). Nous avons les individus segmentés avec leur boîte englobante rectangulaire et celles des résultats de localisation et de suivi temporel du visage et des mains. Même si la segmentation est assez mauvaise, les résultats sont corrects. Les trois dernières séries d'images illustrent l'adaptation automatique des seuils, à gauche nous avons les images originales, au milieu, les images de détection de peau avec les résultats de localisation et de suivi temporel obtenus sans adaptation et à droite les images de détection de peau et les résultats avec adaptation. Nous pouvons voir que la couleur du fond est incluse dans la gamme de couleurs de peau acceptées, néanmoins grâce à l'adaptation automatique des seuils, les résultats de localisation et de suivi temporel sont grandement améliorés par rapport à ceux obtenus sans adaptation.

4.6 Avantages, limitations et cadences de traitement

Grâce à un seuillage adaptatif dans l'espace $YCbCr$ et grâce à l'utilisation de listes triées, la méthode de localisation et de suivi temporel du visage et des mains arrive à détecter, localiser et suivre au cours du temps le(s) visage(s) et les mains de plusieurs individus, avec la distinction main droite / main gauche. De plus, les seuils de détection de peau sont automatiquement adaptés en fonction de la couleur de peau de la personne.

Les avantages principaux des méthodes proposées sont leur rapidité et leur robustesse. Une première limitation survient quand des réunions temporelles entre personnes ont lieu lors de l'étape de segmentation. Ces méthodes sont efficaces si les masques de segmentation ne se regroupent pas. Dans le cas où les masques de segmentation de plusieurs êtres humains se regroupent, la localisation et le suivi temporel du visage et des mains ne seront pas aussi fiables que lorsqu'une seule personne est présente à l'intérieur d'une boîte englobante rectangulaire

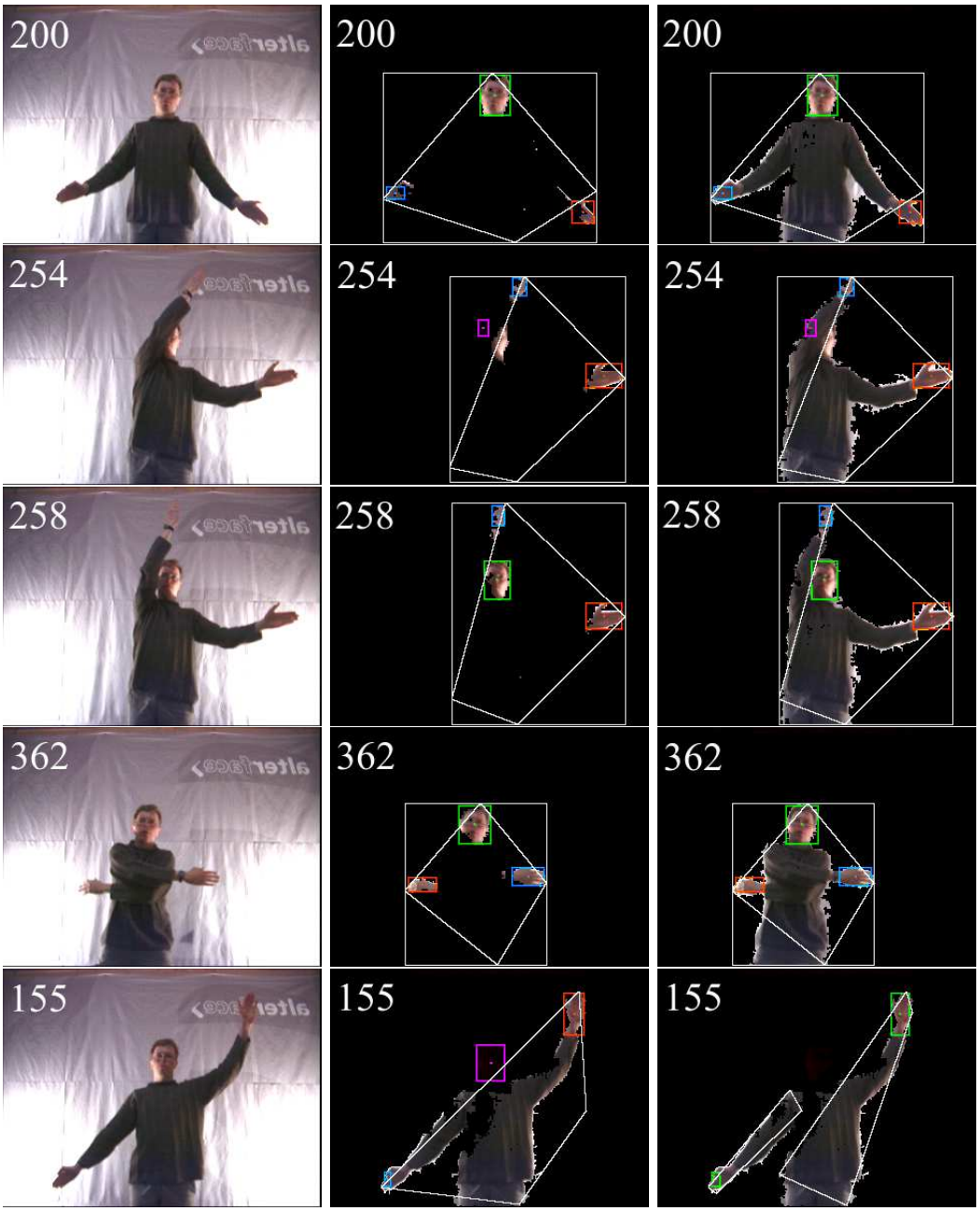


FIG. 4.19 – Exemple 1 de localisation et de suivi temporel du visage et des mains.

FIG. 4.20 – Exemple 2 de localisation et de suivi temporel du visage et des mains.

issue de la segmentation. En effet, nos méthodes tenteront de localiser et de suivre au cours du temps un visage et deux mains uniquement. Une autre limitation est la couleur des vêtements portés. Si elle est trop proche des couleurs de peau, alors le visage ou les mains peuvent inclure une partie de ces vêtements et par conséquent cela dégradera les performances et la précision des résultats. Une dernière limitation est le fait qu'en cas de réunions temporelles entre les deux mains, nous ne pouvons être sûrs des localisations des mains après la séparation temporelle. Par défaut, la main droite sera localisée à droite et la main gauche à gauche. Sur l'ensemble des séquences vidéo testées en environnement intérieur, le taux moyen de réussite pour la localisation et le suivi temporel se situe aux environs de 90%. Les 10% d'échec correspondent principalement à des personnes portant des vêtements de couleur proche de celle de la peau. Lorsque les limitations ci-dessus sont évitées, les méthodes donnent des résultats très satisfaisants et n'échouent pratiquement jamais.

De façon générale, les modèles de couleur de peau sont sensibles au système d'acquisition utilisé (caméra, format d'espace couleur, balance des blancs, bruit d'acquisition etc.) et aux conditions d'éclairement (intérieur, extérieur, éclairage uniforme etc.). Les seuils présentés ont été testés dans différents environnements intérieurs et sont assez robustes pour réaliser une détection de peau fiable dans des conditions d'acquisition variées. Du moment que les couleurs des pixels de peau sont visuellement proches des couleurs de peau perçues par l'œil humain, il ne devrait pas y avoir besoin de régler les seuils initiaux de détection si la même caméra est utilisée. Si la caméra est différente ou si les conditions d'acquisition varient fortement, il peut être utile de régler ou au moins de tester ces seuils initiaux. Dans notre cas, il n'a jamais été nécessaire de les régler, même si l'environnement intérieur était différent. Cela peut être dû à deux raisons : premièrement, cette étape de traitement n'utilise que les pixels des masques de segmentation qui ne devraient pas inclure les pixels du fond et deuxièmement, les conditions d'acquisition dans les environnements intérieur utilisés sont relativement bien contrôlées.

Au niveau des pourcentages de temps de calcul et des cadences de traitement atteintes, la table 4.4 résume les résultats obtenus pour l'ensemble des étapes de traitement jusqu'à la localisation et le suivi temporel du visage et des mains :

Tab. 4.4 – Pourcentages de temps de calcul et cadences de traitement pour la localisation et le suivi temporel du visage et des mains.

Segmentation	Champs aléatoires de Markov		Optimisée en vitesse	
Résolution d'image	320 × 240	640 × 480	320 × 240	640 × 480
Acquisition	0.2%	0.4%	2.4%	5.7%
Segmentation	81.7%	95.4%	33.7%	43.5%
Suivi temporel (1/2)	9.4%	1.7%	4.2%	1.3%
Localisation et suivi du visage et des mains	8.7%	2.5%	59.7%	49.5%
Cadences de traitement	**7.91** images/s	**2.02** images/s	**107** images/s	**31** images/s

Avec la méthode de segmentation basée sur les champs aléatoires de Markov, cette étape de traitement ne ralentit pas beaucoup les cadences de traitement. Avec la segmentation optimisée en vitesse, l'étape de localisation et de suivi temporel du visage et des mains ralentit les cadences de traitement, principalement à cause de l'étiquetage en composantes connexes

du traitement préliminaire. Les autres méthodes présentées sont en effet relativement rapides, que ce soit la détection de peau, l'adaptation des seuils ou la méthode de somme de rangs minimale utilisée pour effectuer le suivi temporel. Au regard des cadences de traitement atteintes avec la segmentation optimisée en vitesse, la cadence vidéo est encore parfaitement respectée.

4.7 Conclusion

Ce chapitre a explicité une méthode de localisation et de suivi temporel du visage et des mains, et les informations bas-niveau extraites lors de cette étape.

Dans une première partie, nous avons détaillé notre méthode pour extraire les pixels de peau grâce à un seuillage sur les composantes de couleur Cb et Cr dans l'espace couleur $YCbCr$. Nous avons ensuite décrit l'adaptation automatique des seuils de la détection de peau par rapport à des caractéristiques statistiques des pixels de peau correctement identifiés. Dans une seconde partie, nous avons présenté les méthodes utilisées pour localiser et suivre au cours du temps le visage, la main droite et la main gauche. Ces méthodes sont principalement basées sur l'utilisation de listes triées suivant des critères de taille, de position et de distance et sur une méthode de somme de rangs minimale dans ces listes. Puis nous avons présenté les données bas-niveau extraites lors de cette étape de traitement. Après cela, nous avons illustré les résultats de nos méthodes par quelques images de localisation et de suivi temporel issus du traitement de séquences vidéo variées. Dans une dernière partie, nous avons donné les principaux avantages et limitations de nos méthodes, ainsi que les cadences de traitement atteintes. Les méthodes proposées ont l'avantage d'être relativement rapides, que ce soit pour la détection de peau par seuillage, l'adaptation des seuils, la localisation ou le suivi temporel du visage et des mains. De plus, grâce à l'adaptation automatique des seuils de détection, les résultats sont de bonne qualité vis-à-vis des applications visées. Les limitations principales sont les réunions temporelles entre individus et les vêtements de couleur proche de la couleur de peau.

Nous allons maintenant présenter quelques perspectives concernant cette étape de traitement. Pour la méthode de détection de peau par seuillage, il serait possible de modéliser plus précisément la répartition spatiale des couleurs de la base de données de peau acquises avec notre système. Cette répartition, cf. figure 4.8(b), page 89, a une forme elliptique et cela pourrait améliorer les résultats. Par rapport à la localisation du visage et des mains, une méthode basée sur des distances géodésiques par rapport à la forme de la personne pourrait donner une autre estimation des localisations. Néanmoins, cela rendrait plus difficile le suivi temporel quand les mains sont devant le corps. Une autre perspective intéressante, concernant le suivi temporel, serait d'utiliser des modèles de vitesse pour gérer les cas de réunions temporelles entre les mains où, pour l'instant, nous ne pouvons prendre une décision sûre par rapport aux informations dont nous disposons. Il serait intéressant aussi de regarder si le suivi temporel pourrait être réalisé grâce à des informations de forme et de contour. Une dernière perspective serait de réaliser un suivi temporel 3D, basé sur la forme, la position, la couleur et la vitesse afin de gérer différemment et de comparer les résultats sur les occultations et les réunions / séparations temporelles.

Chapitre 5

Seconde étape du suivi temporel

Sommaire

L'idée de suivre au cours du temps et d'analyser les mouvements d'une ou de plusieurs personnes par une approche basée sur un modèle de corps humain n'est pas récente [Johansson76]. Le corps humain est une structure rigide articulée présentant de nombreux degrés de liberté. Par conséquent, l'analyse du mouvement du corps humain présente une difficulté intrinsèque liée à la nature de ce dernier. Le mouvement d'un corps humain est, en première approximation, grandement lié à celui de son squelette. Le corps humain peut être représenté par des modèles de plus ou moins grande complexité. Cette complexité est liée au nombre de dimensions utilisées pour définir le modèle (1D, 2D ou 3D). En général, plus le nombre de dimensions est élevé, plus le modèle est proche d'un véritable corps humain. En fonction des paramètres du modèle, il faut trouver un compromis entre la difficulté de les extraire et celle de les associer. Les paramètres peuvent être extraits avec ou sans utilisation de capteurs ou de marqueurs. L'utilisation de capteurs ou de marqueurs (lumineux, colorés, etc.) facilite l'extraction des paramètres mais ces approches sont souvent considérées comme invasives car l'appareillage est susceptible de gêner les mouvements des individus filmés. Les méthodes de capture de mouvement (*Motion Capture*) utilisent par exemple des capteurs de position dans l'espace qui donnent accès aux positions 3D des articulations principales du corps humain. Les méthodes qui permettent l'analyse du corps humain grâce à des capteurs peuvent servir de vérité terrain pour la comparaison avec d'autres modèles. En effet, comme l'utilisation de capteurs n'est pas toujours possible selon le type d'application visée, un grand nombre d'autres méthodes qui ne nécessitent pas de capteurs ont été développées parallèlement.

La définition d'un modèle de corps humain dépend très souvent du type d'application visée. Certaines applications ne nécessitent qu'un modèle très approximatif (vidéosurveillance dans une foule par exemple) alors que d'autres ont besoin d'un modèle plus précis et détaillé (applications médicales où il faut par exemple diagnostiquer des problèmes de démarche).

Nous nous sommes intéressés aux modèles de corps humain afin de pouvoir améliorer le suivi temporel de personnes, notamment en cas d'occultations. En effet, la première étape du suivi temporel peut détecter les réunions et les séparations temporelles entre individus mais elle ne les corrige pas. Cette première étape réalise le suivi d'un groupe d'êtres humains de la même façon que celui d'une personne seule.

Nous commencerons par présenter dans ce chapitre un état de l'art sur les modèles de corps humain. Puis nous présenterons une méthode pour réaliser un suivi temporel de personnes dans un groupe basé sur un filtrage de Kalman partiel et une poursuite du visage. Cette méthode constitue la seconde étape du suivi temporel et permet de gérer les problèmes d'occultations entre individus. Ensuite nous exposerons les données bas-niveau extraites lors de cette étape de traitement, et nous illustrerons les résultats obtenus. Nous préciserons alors les avantages, les limitations et les cadences de traitement atteintes pour cette étape avant de conclure. Ce chapitre présente la dernière étape de notre système concernant l'extraction de données bas-niveau.

5.1 État de l'art sur les modèles de corps humain

Les approches de suivi temporel basées sur des modèles du corps humain utilisent des connaissances *a priori* sur la structure du corps humain pour le suivre au cours du temps. Le corps humain est considéré comme un ensemble de parties (tête, avant-bras, bras, mains, torse, cuisses, jambes, pieds) qui peuvent être définies de façon plus ou moins précise. De façon courante, la structure du corps humain est associée aux mouvements du squelette,

qui constitue un ensemble de segments reliés par des articulations. Les segments peuvent être estimés en tant que tels, au sens géométrique (1D), en utilisant un modèle simplifié de squelette humain. C'est une estimation linéaire, à une dimension, des segments, où ils sont approchés par des lignes. Il est aussi possible de les estimer par des paramètres 2D, en se basant sur des informations de contour ou de silhouette. C'est alors une estimation surfacique, à deux dimensions, des segments. Dans une dernière approche, il est aussi possible de considérer ces segments au sens volumique, à trois dimensions, et en les approchant par des parallélépipèdes, des cylindres, des *blobs* 3D ou un autre modèle volumique [Aggarwal98]. Ainsi, les segments du corps humain sont approchés respectivement en tant que lignes, surfaces et volumes qui seront respectivement appelées approches 1D, 2D, ou 3D. Nous allons maintenant détailler un peu plus l'état de l'art pour chaque approche.

5.1.1 Approches 1D

L'essence du mouvement humain est contenue dans les mouvements de la tête, du torse et des quatre membres. La représentation la plus simple d'un corps humain consiste en un squelette de personne formé de segments de ligne (*stick figure*) reliés par des articulations. Le mouvement des articulations et des segments permet d'estimer le mouvement d'un corps humain dans son entier. Différentes méthodes donnent accès à un squelette de personne, par exemple une transformée d'axe médian [Bharatkumar94] ou une transformée de distance [Iwasawa97].

Les approches 1D, qui conduisent à des modèles de corps humain en squelette, peuvent être classées en deux catégories, suivant le fait qu'elles utilisent ou non des connaissances *a priori* sur la forme.

5.1.1.1 Approches 1D sans connaissances *a priori* sur la forme

Le squelette de personne, dont un exemple est donné figure 5.1, a été initialement proposé par Johansson qui a montré que le regard humain est capable d'interpréter une structure de corps humain en mouvement formée de points lumineux se déplaçant visuellement (*MLD : Moving Light Display*) [Johansson76]. Les *MLD* sont des points lumineux brillants placés sur les articulations d'un individu habillé en noir se déplaçant devant un fond sombre. Il est ainsi possible de trouver la structure du corps humain en faisant l'hypothèse que les articulations appartenant à un même objet ont de plus grandes corrélations de positions et de vitesses projetées en 2D [Rashid80]. En 3D, la reconstruction du squelette est aussi possible si l'on suppose que le mouvement de chaque partie rigide de l'objet est contraint de telle façon que son axe de rotation reste fixe [Webb81, Webb82]. D'autres études se focalisent sur les trajectoires des articulations des *MLD*, [Bobick95] où le mouvement humain est représenté grâce à des courbes dans des sous-espaces de l'espace des phases [Campbell95a].

5.1.1.2 Approches 1D avec connaissances *a priori* sur la forme

Chen et Lee utilisent un squelette de personne pour représenter les caractéristiques de la tête, du torse, des bras et des jambes avec un squelette formé de dix-sept segments et quatorze articulations [Chen92] (cf. figure 5.1). Ils retrouvent la configuration 3D d'un individu en mouvement grâce à sa projection 2D dans l'image. Différentes contraintes sont imposées afin de réaliser ensuite une analyse de la démarche (marche, course, etc.). La méthode est très coûteuse en temps de calcul, étant donné qu'elle recherche la solution parmi l'ensemble des

configurations 3D possibles à partir de la projection 2D obtenue. De plus, elle nécessite une grande précision pour l'extraction du squelette.

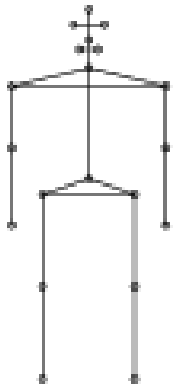


FIG. 5.1 – Modèle de corps humain par une approche 1D [Chen92].

Guo, Xu et Tsuji représentent la structure du corps humain en silhouette comme un modèle de squelette de personne composé de dix segments et de six articulations [Guo94a]. Le problème est alors la recherche d'un squelette d'énergie minimale dans un champ de potentiel. De plus, des contraintes de prédiction et d'angles entre les articulations sont introduites de façon à réduire la complexité du processus d'association.

D'autres modèles en squelette ne traitent que certaines parties du corps, par exemple, le modèle de Bharatkumar *et al.* [Bharatkumar94], qui ne concerne que les jambes avec les articulations de la hanche, du genou et de la cheville. Ces travaux visent à définir un modèle cinématique général pour l'analyse de la démarche d'une personne. Des transformées d'axe médian sont utilisées pour extraire les segments 2D du squelette au niveau des membres inférieurs. L'angle des segments et le déplacement des articulations sont mesurés et lissés à partir de séquences d'images réelles. Un schéma cinématique commun est alors détecté lors de chaque cycle de marche. Une forte corrélation ($> 95\%$) a été trouvée entre les images réelles et le modèle. L'inconvénient de ce modèle est qu'il est dépendant du point de vue et donc sensible aux changements d'angle de vue de la caméra. De plus, il est restreint à la partie inférieure du corps humain.

Un modèle en squelette amélioré a été développé par Huber [Huber96] où les articulations connectant les segments dépendent de contraintes qui découlent de ressorts virtuels. Des mesures de mouvement des articulations basées sur une méthode stéréoscopique sont réalisées dans un espace à trois dimensions appelé espace de proximité (*PS : Proximity Space*). Pour suivre au cours du temps l'ensemble des articulations dans le *PS*, le point de départ est la tête.

Les travaux d'Iwasawa se sont focalisés sur l'extraction de squelettes de personne à partir de séquences d'images thermiques monoculaires [Iwasawa97]. La taille du sujet et la distance à la caméra ayant été pré-calibrées, l'orientation du haut du corps est calculée comme étant l'axe d'inertie principal de la silhouette. Alors, les points significatifs comme le haut de la tête, les extrémités des mains et des pieds sont extraits comme étant, de façon heuristique, ceux qui sont les plus éloignés du centre de la silhouette. Finalement, les articulations principales que sont les coudes et les genoux sont estimées selon les positions des points détectés grâce à un apprentissage. L'inconvénient de cette méthode est qu'elle aussi est dépendante de la vue.

De plus, la méthode d'obtention des points significatifs limite les gestes reconnus. En effet, pour quelqu'un ayant les bras devant lui, il n'y a aucun moyen d'extraire les extrémités des mains et, par conséquent, la méthode choisie échouera à estimer les articulations des coudes.

Fujiyoshi et Lipton utilisent un genre d'approche similaire, avec un squelette en "étoile" (*"star" skeleton*). Grâce à une modélisation simple de la forme du corps humain, ils réalisent une analyse du mouvement en extrayant uniquement des caractéristiques intérieures grossières de l'objet cible [Fujiyoshi98]. Les indices sur le mouvement obtenus à partir du squelette en étoile sont le mouvement cyclique des segments représentant les jambes et la pose du segment du torse. Ces indices, utilisés conjointement, permettent d'analyser le mouvement d'un être humain et de le classer en marche ou course.

Karaulova, Hall et Marshall ont aussi utilisé ce type de représentation du corps humain en modèle de squelette de personne pour construire un modèle hiérarchique de la dynamique du corps humain en utilisant des chaînes de Markov cachées. Ils réalisent ainsi un suivi temporel du corps humain indépendant de la vue dans des séquences vidéo monoculaires [Karaulova00].

5.1.2 Approches 2D

Ce type de représentation du corps humain est directement lié à la projection du corps dans le plan de l'image. Dans une telle représentation, les différentes parties du corps peuvent être représentées par des régions définies par des rectangles ou des parallélogrammes (aussi appelés rubans) 2D [Jain79, Kurakake92, Leung94, Leung95, Ju96], par des *blobs* 2D [Shio91], par des contours 2D [Kakadiaris94, Niyogi94a, Niyogi94b] ou par des gabarits (*templates*) [Bobick96, Rosales98]. L'avantage de ces approches par rapport aux approches 1D est que l'utilisation de surfaces par rapport à des lignes peut réduire la probabilité de mauvaise association. Il est aussi possible généralement d'extraire un squelette de personne d'un modèle par une approche 2D.

Les approches 2D peuvent aussi être classées en deux catégories, suivant le fait qu'elles utilisent ou non des connaissances *a priori* sur la forme.

5.1.2.1 Approches 2D sans connaissances *a priori* sur la forme

L'utilisation de *blobs* 2D est un exemple d'approche 2D sans connaissances *a priori* sur la forme. Les pixels sont groupés en *blobs* selon des caractéristiques qui peuvent être la position, la couleur, la vitesse etc. Par exemple, les *blobs* peuvent être groupés selon l'amplitude et la direction de la vitesse 2D des pixels, qui est obtenue par des méthodes basées sur le flot optique [Shio91]. La vitesse de chaque partie du corps est supposée converger vers une valeur globale moyenne après plusieurs images consécutives. Cette vitesse moyenne correspond au mouvement de l'ensemble du corps humain et conduit à l'identification du corps dans son entier par le regroupement de régions ayant des vitesses moyennes similaires.

Les articulations des objets et leurs mouvements grossiers peuvent être estimés en utilisant des rubans 2D [Kurakake92]. Avec l'hypothèse de mouvements de faible amplitude sur deux images consécutives, la correspondance entre rubans retenus après avoir filtré les rubans mal associés est réalisée sous différentes contraintes géométriques. Les articulations sont localisées aux alentours des rubans connectés ou proches, souvent comme le centre de la surface de superposition des rubans.

La segmentation, la forme et l'estimation de mouvement peuvent être combinées pour construire des modèles déformables [Kakadiaris94]. L'approche de Kakadiaris, Metaxas et

Bacjsy utilisent des contours 2D selon plusieurs vues pour estimer la position du corps humain en 3D [Kakadiaris95]. Les contours 2D sont utilisés pour la segmentation et l'estimation de mouvement, où le lieu d'une articulation est détecté comme étant le centre de la surface de superposition de deux contours connectés, à partir du moment où ces contours se mettent en mouvement. Au début, la personne est supposée être un modèle déformable. Au fur et à mesure qu'elle se déplace et que de nouvelles parties du corps apparaissent, de nouveaux contours 2D sont créés pour remplacer les anciens, chacun d'eux représentant une sous-partie du corps en mouvement. Les articulations sont alors déterminées en fonction du mouvement relatif et de la forme de ces sous-parties.

Rowley et Rehg ont exploré la segmentation par flot optique d'objets articulés [Rowley97]. Leur travail est une extension de l'analyse de mouvement d'objets rigides aux objets articulés en utilisant l'algorithme *EM (Expectation Maximization)*. Ils ajoutent des contraintes cinématiques de mouvement à chaque pixel de données. La force de cette approche est la combinaison performante de la segmentation de mouvement et de l'estimation dans le calcul *EM*. La segmentation est réalisée pendant l'étape *E*, et l'analyse du mouvement, l'estimation, est réalisée pendant l'étape *M*. Ces deux étapes sont calculées de façon itérative d'une manière avant-arrière afin de minimiser la fonction d'énergie globale de l'image. Les mouvements considérés sont restreints à des transformations 2D affines.

Le mouvement humain peut aussi être décrit comme un ensemble de gabarits (*templates*) où la composante temporelle est comprise dans le modèle sans analyse temporelle explicite ou association de séquence. Bobick *et al.* proposent une approche basée sur la vue pour la représentation et la reconnaissance d'action en se servant de gabarits temporels [Bobick96, Bobick97]. Ils utilisent une image d'énergie du mouvement (*MEI : Motion Energy Image*) et une image d'histoire du mouvement (*MHI : Motion History Image*) pour analyser le mouvement humain dans une séquence vidéo. Dans un premier temps, des images de mouvement sont extraites d'une séquence vidéo par différenciation et seuillage, et sont accumulées temporellement pour former les *MEI*. Les *MEI* sont des images binaires qui décrivent le mouvement dans sa globalité. Elles sont ensuite transformées en *MHI* dont l'intensité des pixels est fonction de l'histoire du mouvement. Les *MHI* sont des images de valeurs scalaires. Utilisées conjointement, les *MEI* et *MHI* peuvent être considérées comme la version d'un gabarit temporel à deux composantes. Finalement, ces gabarits de vue spécifique sont associés avec les modèles enregistrés d'actions connues pendant le processus de reconnaissance.

5.1.2.2 Approches 2D avec connaissances *a priori* sur la forme

Certaines études utilisent deux ensembles de rubans 2D (un pour chaque côté du bord en mouvement, soit une partie du corps, soit une partie du fond) pour l'identification de parties du corps humain selon leurs changements de forme au cours du temps. Ces parties du corps sont alors étiquetées selon un modèle de corps humain. Sur ce principe, une description des parties du corps et des articulations est obtenue. Il est possible de retrouver ainsi les différents membres du corps humain [Jain79].

Leung et Yang appliquent un modèle de corps humain avec des rubans 2D pour reconnaître les postures d'une personne en train de réaliser des mouvements de gymnastique [Leung94, Leung95]. Ils estiment le mouvement uniquement à partir de la silhouette de la personne. Leur système comprend deux étapes, l'extraction des rubans 2D à partir de la silhouette et l'interprétation du mouvement. Leur modèle de corps humain en rubans se compose du tronc du corps, de cinq rubans en forme de U avec leurs axes principaux, de sept points

d'articulation, et de plusieurs points milieux des segments. Dans leur travail, l'extérieur de la silhouette du sujet est donc estimée comme des régions frontières représentées par des rubans 2D qui sont des segments de bord en forme de U (cf. figure 5.2). Un processus de relaxation spatio-temporelle est proposé pour déterminer si un ruban 2D appartient bien à l'individu ou au fond. Le modèle de ruban 2D est utilisé pour décrire les relations de structure, de forme et de mouvement entre les différentes parties et ainsi guider l'étiquetage des données de l'image en parties du corps. Ils obtiennent ainsi une description des parties du corps humain et les articulations entre ces parties.

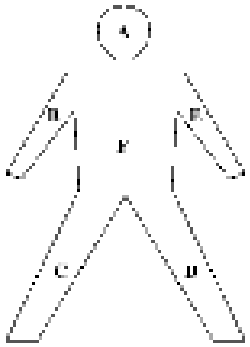


FIG. 5.2 – Modèle de corps humain par une approche 2D [Leung95].

Une silhouette ou un contour sont relativement faciles à extraire à la fois du modèle et de l'image. En se basant sur une représentation 2D du contour, Niyogi et Adelson utilisent des schémas spatio-temporels dans l'espace XYT pour suivre, analyser et reconnaître des individus qui se déplacent en marchant [Niyogi94a, Niyogi94b]. Ils examinent d'abord les schémas caractéristiques produits par les membres inférieurs, puis la projection des mouvements du visage est localisée dans le domaine spatio-temporel, suivie de l'identification des trajectoires d'autres articulations. Finalement, le contour d'un individu marchant est détecté en utilisant ces trajectoires, et une méthode de reconnaissance de démarche plus précise est utilisée avec ce contour 2D actif pour la reconnaissance d'un individu particulier.

Ju, Black et Yacoob proposent un modèle de personne, nommé *Cardboard Model*, dans lequel les membres du corps humain sont représentés par un ensemble de rubans 2D connectés afin d'analyser les mouvements de démarche [Ju96]. Un modèle paramétré de flot optique est utilisé pour traiter du mouvement articulé des membres humains. Un mouvement explicite du modèle qui est basé sur des courbes de mouvement dérivées analytiquement est utilisé pour représenter le corps humain de même que son mouvement. Ceci permet d'estimer les positions 3D et les postures des gens dans les séquences.

5.1.3 Approches 3D

Le principal inconvénient des modèles 2D est leur restriction suivant l'angle de vue de la caméra, alors de nombreux chercheurs essaient de trouver une structure géométrique du corps humain plus détaillée en utilisant des modèles 3D comme des cylindres ou des ellipsoïdes [Hogg83, Rohr94, Wachter99, Sminchisescu03], des cônes tronqués [Wachter99, Goncalves95], des sphères, aussi appelées balles, [O'Rourke80, Goncalves95] etc. Plus les

modèles 3D sont complexes, meilleurs sont les résultats mais ils requièrent plus de paramètres et conduisent souvent à des calculs beaucoup plus coûteux durant le processus d'association du suivi temporel. Il est possible d'utiliser des modèles 3D avec une seule caméra, il faut alors faire correspondre la projection 2D de la personne dans l'image à une configuration du modèle 3D. Mais différentes vues permettent d'améliorer l'analyse, problème bien connu en stéréovision.

Les modèles volumiques à base de sphères sont fréquemment utilisés. O'Rourke et Badler se servent de six cents sphères qui se superposent pour définir un modèle 3D de corps humain très élaboré comprenant vingt-quatre segments et vingt-cinq articulations [O'Rourke80]. Leur système est basé sur quatre étapes de traitement : prédiction, simulation, analyse d'image et analyse du modèle. Tout d'abord, l'étape d'analyse d'image localise précisément les parties du corps selon les précédents résultats de prédiction. Lorsque la gamme de possibilités des localisations 3D prédites pour les partie du corps est suffisamment réduite, l'étape d'analyse du modèle transforme les relations spatio-temporelles (entre les localisations et le temps) en certaines fonctions linéaires. Alors l'étape de prédiction estime les positions des parties du corps dans l'image suivante en utilisant ces dernières fonctions linéaires. Finalement l'étape de simulation, qui comprend de nombreuses connaissances *a priori* sur le corps humain, translate les données de prédiction en régions 3D correspondantes, qui seront vérifiées par l'étape d'analyse d'image de la prochaine boucle.

Hogg et Rohr utilisent chacun un ensemble de quatorze cylindres elliptiques pour modéliser le corps humain selon une approche 3D [Hogg83, Rohr94]. Le modèle de Hogg est présenté figure 5.3. L'origine du système de coordonnées est fixé au centre du torse. L'ajustement des axes des cylindres est réalisé en fonction de vecteurs propres pour modéliser l'extérieur du corps humain et ensuite les projections 2D sont ajustées pour correspondre aux modèles 3D selon une mesure de similarité basée sur la distance.

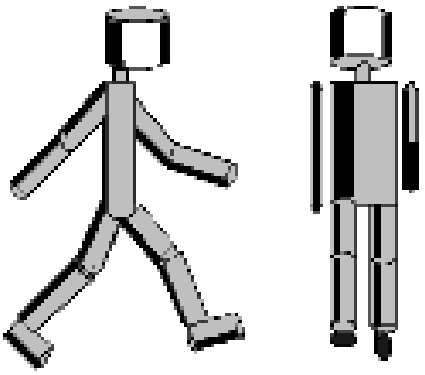


FIG. 5.3 – Modèle de corps humain par une approche 3D [Hogg83].

Les modèles sphériques peuvent être utilisés conjointement avec d'autres modèles volumiques pour définir le corps humain, comme dans [Goncalves95] où à la fois le bras et l'avant-bras sont modélisés comme des cônes tronqués de bases circulaires, et les épaules et les coudes sont supposés être des articulations sphériques. Une projection du modèle 3D du bras est utilisée pour être associée à une image indistincte d'un véritable bras. L'association est réalisée en minimisant récursivement l'erreur entre la projection du modèle et l'image réelle par une adaptation dynamique de la taille et de l'orientation du modèle.

Kakadiaris et Metaxas, d'un autre côté, ont considéré le problème d'un suivi temporel basé sur des modèles 3D de parties du corps humain à partir de trois caméras calibrées placées dans une configuration d'orthogonalité mutuelle [Kakadiaris95, Kakadiaris96]. Les modèles 3D de parties du corps sont extraits à partir des projections 2D du point de vue de chaque caméra. Les caméras se succèdent pour le suivi temporel, selon la visibilité des parties du corps en fonction de leurs mouvements prédits. La correspondance entre les points des contours occultés et les points correspondants sur le modèle 3D est établie en utilisant des concepts de géométrie perspective. Un filtrage de Kalman est utilisé pour prédire le mouvement des parties du corps.

Gavrila et Davis ont ensuite étendu l'approche de [Kakadiaris95] à quatre caméras calibrées pour effectuer une modélisation 3D du corps humain [Gavrila95, Gavrila96]. La classe des super-quadriques graduellement de plus en plus étroites (*tapered super-quadrics*), comprenant des cylindres, des sphères, des ellipsoïdes et des hyper-rectangles, a été utilisée pour l'obtention d'un modèle 3D à vingt-deux degrés de libertés. Le cadre de travail général pour la reconstruction de la posture et le suivi temporel des parties du corps est similaire à celui présenté dans [O'Rourke80]. L'association entre la vue du modèle est la scène réelle est basée sur un calcul de similarité au niveau des bords des contours selon une variante de distance de chanfrein [Barrow77].

Voulant générer une description 3D de personnes par modèle volumique, Wachter et Nagel ont tenté d'établir une correspondance entre un modèle 3D du corps humain basé sur un ensemble de cônes elliptiques et une séquence réelle d'images [Wachter97]. Grâce à un filtrage de Kalman étendu itératif, qui incorpore de l'information à la fois sur les bords et les régions afin de déterminer les degrés de libertés des articulations et leurs orientations par rapport à la caméra, ils ont obtenu une description qualitative du mouvement humain dans des séquences vidéo monoculaires.

Hunter *et al.* ont appliqué une modélisation par mélange de densités à la reconnaissance de postures de corps humains 3D dans une séquence d'images [Hunter97]. Le modèle 3D de corps humain consiste en cinq formes de composantes ellipsoïdes représentant le tronc et les bras avec quatorze degrés de liberté. Un algorithme *EM* modifié est utilisé pour résoudre l'estimation du mélange dans les images segmentées.

Un avantage important d'une modélisation 3D du corps humain est la capacité de gérer, dans une certaine mesure, les phénomènes d'occultations et d'obtenir plus de données significatives pour la reconnaissance d'actions. Cependant, elle est restreinte à des hypothèses de simplicité comparées aux possibilités de mouvement du corps humain et a aussi une complexité calculatoire élevée.

5.1.4 Approches génériques appliquées aux êtres humains

Rappelons aussi qu'il est possible de définir des modèles de corps humain qui ne sont pas explicitement basés sur la structure anatomique du corps humain. Ces modèles peuvent être utilisés pour d'autres *ROI* que des êtres humains mais, par extension, sont aussi appelés modèles de corps humain parce qu'ils ont été appliqués à des êtres humains. Dans cette catégorie, nous pouvons inclure par exemple, certains des modèles par approche 2D basés sur la silhouette ou le contour, et aussi d'autres modèles comme les modèles d'apparence basés sur la couleur [Huang99, Capellades03].

McKenna, Jabri, Duric, Rosenfeld et Wechsler [McKenna00b] proposent une segmentation spatio-temporelle basée sur une méthode de soustraction de fond qui combine une information de gradient et des caractéristiques de couleur (moyennes et variances dans le sous-espace

couleur rg) afin de traiter les ombres en segmentation de mouvement. Ils différencient trois niveaux de suivi temporel : les régions, les personnes seules et les groupes de personnes. Ils utilisent des boîtes englobantes rectangulaires pour chaque région qui peut se séparer ou se regrouper. Un groupe est formé de plusieurs individus, chacune d'elles étant formée de plusieurs régions sous certaines contraintes géométriques. Ils obtiennent ainsi de bons résultats de suivi temporel de plusieurs personnes même en cas d'occultations. Mais si deux personnes sont vêtues de la même façon, le suivi échouera quand elles se réuniront puis se retrouveront séparées à nouveau.

Dockstader et Tekalp [Dockstader01] présentent une méthode de suivi temporel quasi temps-réel de plusieurs personnes dans un système de vidéosurveillance. L'algorithme mélange des estimations de mouvement, des informations de détection de changement et des prédictions pour créer des trajectoires précises d'objets en mouvement. Ces caractéristiques bas-niveau et cette stratégie de mélange de composantes utilisent un mécanisme de filtrage de Kalman modifié. Il y a peu de contraintes dans leur système mais il ne gère pas les occultations complètes.

Dans [Capellades03], Capellades *et al.* décrivent un système pour le suivi temporel d'êtres humains et la détection des interactions personne-objet en environnement intérieur. Une combinaison d'information par histogramme et corrélogramme est utilisée pour modéliser les distributions de couleur de personne et d'objet. Cependant la particularité de la couleur de peau humaine n'est pas prise en compte. Les modèles d'apparence sont construits au fur et à mesure du traitement et sont utilisés pour suivre au cours du temps les personnes d'une image à la suivante. Le système est capable de détecter des personnes qui se regroupent et peut les segmenter durant les occultations. Les résultats de suivi temporel et de segmentation d'individus à l'intérieur d'un groupe sont très bons mais le prix à payer pour l'utilisation des corrélogrammes est un temps de calcul élevé.

Mostafaoui, Achard et Milgram ont proposé récemment une méthode de suivi temporel de personnes dans des séquences d'images couleur utilisant simultanément la cinématique, la forme et un modèle d'apparence [Mostafaoui05]. Afin de gérer correctement les problèmes d'occultations et de sur-segmentation, l'algorithme utilise des pistes élémentaires qui décrivent les trajectoires des individus ou des groupes quand ne surviennent aucune réunion ou séparation temporelle pour ces régions. Les pistes élémentaires sont ensuite regroupées en pistes associées aux personnes selon les paramètres cités précédemment. Ainsi, il est possible de suivre temporellement des personnes en tenant compte des phénomènes d'occultations.

Résumé

Toutes ces approches, 1D, 2D, 3D et génériques, décrivent le problème d'associer l'image d'un être humain à sa représentation abstraite par des modèles de corps humain de différentes complexités. Le problème en lui-même n'est pas trivial. La complexité du processus d'association est régie par le nombre de paramètres du modèle et par l'efficacité de la segmentation du corps humain. Quand peu de paramètres sont utilisés pour le modèle, il est plus facile d'associer la caractéristique extraite au modèle mais il est souvent plus difficile d'extraire cette caractéristique. Par exemple, l'approche 1D de squelette de personne est une façon simple de représenter un corps humain, par conséquent il est relativement aisé d'associer les segments de lignes extraits aux segments du squelette correspondant. Cependant, extraire un squelette d'images réelles nécessite plus de précautions que la recherche de *blobs* ou de volumes.

5.2 Suivi temporel basé sur un filtrage de Kalman et une poursuite du visage

5.2.1 Introduction

La première étape du suivi temporel, présentée au chapitre 3, détecte mais ne gère pas les réunions et les séparations temporelles d'individus ou de groupes de personnes. Quand deux individus suivis au cours du temps se réunissent, cette première étape détecte la réunion temporelle mais ne la corrige pas et suit alors le groupe résultant dans son entier jusqu'à ce que les deux personnes se séparent à nouveau. Alors elles sont de nouveau suivies individuellement mais sont considérées comme de nouveaux individus. Aucun lien temporel n'est fait entre les personnes avant la réunion temporelle et après la séparation.

Dans la figure 3.8, page 70, deux personnes, notées par exemple P_1 et P_2, se réunissent en un groupe. Quand ce groupe se sépare en deux individus, elles sont suivies en tant que P_3 et P_4, et non en tant que P_1 et P_2. Les réunions et les séparations temporelles rendent la tâche de suivi d'êtres humains dans un groupe beaucoup plus difficile [McKenna00b, Dockstader01].

Cette partie présente la seconde étape du suivi temporel, qui utilise une combinaison de filtrage de Kalman partiel et de poursuite du visage afin de réaliser le suivi temporel de plusieurs personnes en temps-réel même dans le cas d'occultations complètes.

5.2.2 Méthode pour la seconde étape du suivi temporel

Il n'est pas possible de segmenter les individus à l'intérieur d'un groupe pendant les occultations mais nous obtenons des estimations pour les boîtes englobantes rectangulaires de leur positions. Cette méthode est basée sur un filtrage de Kalman partiel et une poursuite du visage.

Le filtrage de Kalman est un algorithme bien connu optimal et récursif pour l'estimation de paramètres [Kalman60]. Une présentation de la théorie sur le filtrage de Kalman est donnée en annexe B. Grâce à un modèle d'évolution des paramètres, cet algorithme calcule des prédictions et ajoute l'information, provenant de mesures, de façon optimale pour produire des estimations *a posteriori* des paramètres.

Nous définissons un filtre de Kalman pour chaque nouvelle personne détectée[1], c'est-à-dire chaque nouvel objet. Par rapport à la contrainte de temps-réel, des choix simples doivent être faits : nous assimilons le déplacement global d'une personne dans la scène au mouvement apparent 2D de son visage. Associé à un modèle d'évolution à vitesse constante, ceci conduit à un vecteur d'état $\underline{x}$ de dix composantes pour chaque filtre de Kalman : les boîtes englobantes rectangulaires de la personne et de son visage (quatre coordonnées chacune) et deux composantes pour la vitesse apparente 2D du visage :

$$\underline{x}^T = (x_{pg}, x_{pd}, y_{ph}, y_{pb}, x_{vg}, x_{vd}, y_{vh}, y_{vb}, v_x, v_y).$$

Les indices $_p$ et $_v$ correspondent respectivement aux boîtes englobantes rectangulaires de la personne et de son visage, $_g$, $_d$, $_h$ et $_b$ correspondent respectivement aux coordonnées gauche, droite, haut et bas d'une de ces boîtes. v_x et v_y sont les deux composantes de la vitesse apparente 2D du visage.

[1] Dans les limites exposées au chapitre 3 au niveau du nombre de personnes.

Le modèle choisi pour représenter une personne est donc composé de deux boîtes englobantes rectangulaires, celle de la personne et celle de son visage, et de la vitesse apparente 2D de son visage.

Le modèle d'évolution à vitesse constante conduit donc à la matrice d'évolution suivante pour chaque filtre de Kalman :

$$A_t = A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Les équations générales de prédiction et de filtrage de Kalman sont (cf. annexe B) :
équations de prédiction :

$$\begin{aligned} \underline{\hat{x}}_{t/t-1} &= A_{t-1}\underline{\tilde{x}}_{t-1/t-1} + B_{t-1}\underline{u}_{t-1}, \\ P_{t/t-1} &= A_{t-1}P_{t-1/t-1}A_{t-1}^T + Q, \end{aligned}$$

équations de filtrage :

$$\begin{aligned} \underline{\tilde{x}}_{t/t} &= \underline{\hat{x}}_{t/t-1} + G_t(s_t - C_t\underline{\hat{x}}_{t/t-1}), \\ G_t &= P_{t/t-1}C_t^T(R + C_tP_{t/t-1}C_t^T)^{-1}, \\ P_{t/t} &= (Id - G_tC_t)P_{t/t-1}, \end{aligned}$$

conditions initiales :

$$\begin{aligned} \underline{\tilde{x}}_{0/-1} &= \underline{x}_0, \\ P_{0/-1} &= P_0 = \lambda Id. \end{aligned}$$

Avec les notations de l'annexe B, pour l'image précédente, à l'instant $t-1$:
- s_t : vecteur de mesures ;
- $\underline{u}_{t-1}$: vecteur de commande ;
- $\underline{\hat{x}}_{t/t-1}$: vecteur d'état prédit ;
- $\underline{\tilde{x}}_{t/t}$: vecteur d'état estimé *a posteriori* ;
- $\underline{x}_0$: vecteur d'état initial ;
- A_{t-1} : matrice d'évolution ;
- B_{t-1} : matrice de commande ;
- C_t : matrice d'observation des mesures ;
- G_t : matrice de gain de Kalman ;
- $P_{t/t-1}$: matrice de covariance prédite ;
- $P_{t/t}$: matrice de covariance estimée *a posteriori* ;
- P_0 : matrice de covariance initiale ;

– Q : matrice de bruit du modèle ;
– R : matrice de bruit d'observation des mesures ;
– Id : matrice identité.

Nous devons maintenant expliquer les simplifications de notations qui découlent de notre modélisation et les hypothèses concernant certaines données.

1. Il n'y a pas de vecteur de commande, soit $\forall t\ \underline{u}_t = 0$.
2. Les variables du vecteur d'état étant les mêmes que celles mesurées, la matrice d'observation $C_t = Id$. En effet, toutes les mesures sont directement observables. Par conséquent, nous pouvons noter $s_t = \underline{x}_t^m$, le ^{m} signifiant mesures.
3. Les mesures sont indépendantes et il n'y a que peu de bruit sur les mesures. Par conséquent, la matrice de bruit d'observation des mesures R est diagonale $R = Diag(\sigma_i)$. La valeur de σ_i a été fixée à quelques pixels. Étant données les performances de la segmentation et de la localisation du visage, nous pouvons estimer que la dispersion est limitée à quelques pixels.
4. La matrice de bruit du modèle a été choisie diagonale $Q = Diag(\sigma_i^{'})$. La valeur de $\sigma_i^{'}$ a été fixée à quelques pixels.
5. Concernant les conditions initiales, le vecteur d'état initial $\underline{x}_0$ est défini par les premières mesures $\underline{x}_0 = \underline{x}_0^m$ et la matrice de covariance initiale P_0 a été choisie diagonale et égale à la matrice identité $P_0 = Id$.

Les mesures étant disponibles uniquement lors du traitement de l'image courante, à l'instant t, nous commençons par estimer les variables en combinant les prédictions faites lors du traitement de l'image précédente, à l'instant $t-1$, et les mesures provenant de l'image courante, à l'instant t. Ensuite, nous prédisons les variables pour l'image future. Voici les équations de prédiction et de filtrage de Kalman appliquées à notre modélisation, et simplifiées grâce aux hypothèses et aux choix précédemment décrits. Elles sont présentées dans l'ordre où elles sont calculées, pour l'instant t :

équations de filtrage :

$$G_t = P_{t/t-1}(R + P_{t/t-1})^{-1}, \tag{5.1}$$
$$P_{t/t} = (Id - G_t)P_{t/t-1}, \tag{5.2}$$
$$\underline{\tilde{x}}_{t/t} = (Id - G_t)\underline{\hat{x}}_{t/t-1} + G_t\underline{x}_t^m, \tag{5.3}$$

équations de prédiction :

$$\underline{\hat{x}}_{t+1/t} = A\underline{\tilde{x}}_{t/t}, \tag{5.4}$$
$$P_{t+1/t} = AP_{t/t}A^T + Q. \tag{5.5}$$

5.2.2.1 Notations

L'étape de segmentation 2D spatio-temporelle fournit des BERS (Boîtes Englobantes Rectangulaires issues de la Segmentation) qui peuvent contenir une ou plusieurs personnes (dans le cas d'une réunion temporelle) alors que le vecteur d'état de Kalman, et par conséquent, la boîte englobante rectangulaire de la personne qu'il définit, est associé à une personne seule. Ainsi, trois boîtes englobantes rectangulaires différentes existent et sont associées à chaque individu :

– la BERS : Boîte Englobante Rectangulaire issue de la Segmentation ;
– la BERPP : Boîte Englobante Rectangulaire Prédite de la Personne ;
– la BEREP : Boîte Englobante Rectangulaire Estimée *a posteriori* de la Personne.

De manière similaire, trois boîtes englobantes rectangulaires différentes existent pour le visage et sont associées à chaque individu :

– la BERV : Boîte Englobante Rectangulaire du Visage issue de la localisation du visage ;
– la BERPV : Boîte Englobante Rectangulaire Prédite du Visage ;
– la BEREV : Boîte Englobante Rectangulaire Estimée *a posteriori* du Visage.

5.2.2.2 Estimation de la vitesse apparente 2D du visage

Pour chaque visage localisé lors de l'étape de traitement précédente (localisation et suivi temporel du visage et des mains) dans l'image à l'instant $t-1$, nous estimons la vitesse apparente 2D de $t-1$ à t par une méthode de *block-matching* afin d'obtenir les deux composantes de cette vitesse : v_x et v_y. Comme les visages peuvent être de taille relativement petite, tous les pixels à l'intérieur de la boîte englobante rectangulaire du visage sont utilisés comme support pour cette estimation de vitesse.

Il est à noter que la détection de peau, la localisation et le suivi temporel du visage et des mains, pour cette étape de traitement, sont maintenant réalisées sur l'ensemble des boîtes englobantes rectangulaires des personnes prédites par le filtrage de Kalman (BERPP).

5.2.3 Modes de filtrage de Kalman

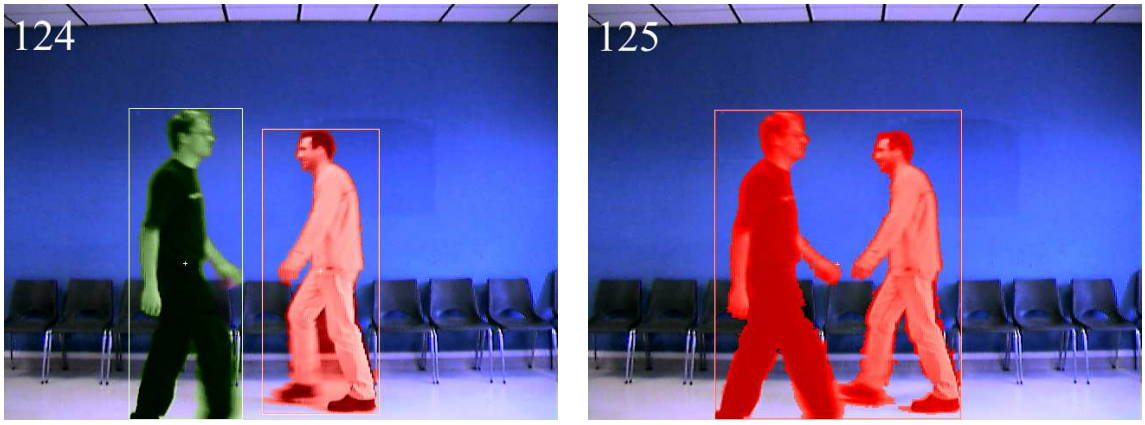
Les mesures qui sont injectées dans les filtres de Kalman proviennent des BERS, des BERV et des estimations de vitesse des visages. Toutes les mesures ne sont pas nécessairement disponibles. Par exemple, si deux individus viennent de se réunir, quelques-unes des mesures qui devraient être combinées aux BERPP de chaque personne ne sont pas disponibles (par exemple, la mesure d'un côté ne sera pas disponible). La figure 5.4 illustre un exemple de mesures indisponibles pour les BERS des individus. Dans l'image à gauche, en (a), les deux personnes sont séparées et ont chacune leur BERS. Dans l'image à droite, en (b), elles se sont réunies et il n'y a plus qu'une seule BERS pour le groupe. Nous pouvons observer que le côté haut de la BERS du groupe, en (b), est très proche de celui de la BERS de la personne à gauche en (a). Les mesures suivantes sont donc indisponibles :

– mesure pour le côté gauche de la BERS de la personne à droite dans (b) ;
– mesure pour le côté haut de la BERS de la personne à droite dans (b) ;
– mesure pour le côté droit de la BERS de la personne à gauche dans (b).

Selon le nombre et le type de mesures disponibles, il y a quatre modes de filtrage de Kalman :

1. PSComp : mode Personne Seule Complet ;
2. PSPar : mode Personne Seule Partiel ;
3. GPPar : mode Groupe de Personnes Partiel ;
4. GPPre : mode Groupe de Personnes Prédictif.

Tout d'abord, nous devons déterminer si nous sommes dans un mode personne seule ou groupe de personnes. Ceci est une information fournie par la première étape du suivi temporel, qui peut détecter les réunions (et les séparations) temporelles entre individus. Ainsi nous savons s'il y a une ou plusieurs personnes dans chaque BERS.



(a) (b)

FIG. 5.4 – Illustration de mesures indisponibles. (a) avant réunion temporelle, (b) après réunion temporelle.

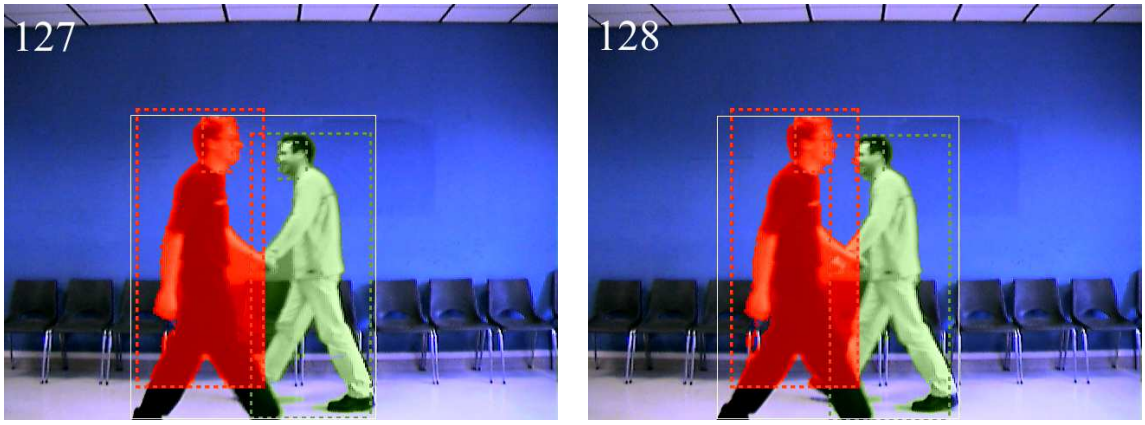
Si la BERS contient une seule personne, toutes les mesures utilisées pour estimer la BEREP sont disponibles. Dans ce cas, le visage a été correctement localisé aux instant $t-1$ et t ou non. S'il l'a été, nous sommes dans le mode PSComp comme toutes les mesures pour chaque composante du vecteur d'état sont disponibles. Sinon, nous sommes dans le mode PSPar comme quelques mesures liées au visage ne sont pas disponibles. Par rapport à la première étape du suivi temporel, les modes de filtrage PSComp et PSPar permettent d'obtenir une estimation de la position du visage même si la personne est de dos du point de vue de la caméra.

Si la BERS contient plus d'une personne, quelques mesures ne sont pas disponibles pour l'estimation de la BEREP (cf. figure 5.4). Selon le fait qu'il y a une unique BERV intersectée par la BERPP ou non, nous sommes respectivement dans le mode GPPar ou dans le mode GPPre. La figure 5.5, page 133, illustre les deux cas pour les intersections entre les BERPP des individus et les BERV. Dans l'image à gauche, en (a), les BERPP n'intersectent chacune qu'une unique BERV. Dans l'image à droite, en (b), les BERPP intersectent toutes les deux les deux BERV. Par rapport à la première étape du suivi temporel, les modes de filtrage GPPar et GPPre permettent d'obtenir les estimations des positions des individus et de celles de leurs visages.

Quand il y a des mesures indisponibles, deux choix sont possibles :

1. Réaliser un filtrage de Kalman uniquement sur les composantes du vecteur d'état dont les mesures sont disponibles.
2. Réaliser un filtrage de Kalman sur toutes les composantes du vecteur d'état en remplaçant les mesures indisponibles.

Le premier choix conduit d'une part à une difficulté théorique, puisqu'un filtrage de ce type amènerait une perte d'information (les estimations, les variances et les covariances des mesures indisponibles) et d'autre part à une mise en œuvre complexe, puisque toutes les tailles possibles de vecteurs et de matrices doivent être prévues afin de prendre en compte tous les cas possibles de mesures indisponibles. Cette solution n'a donc pas été retenue.



(a) (b)

FIG. 5.5 – Intersection entre BERV et BERPP. (a) une BERV , (b) deux BERV.

Le second choix permet de limiter la perte d'information et la complexité de la mise en œuvre. Les mesures indisponibles sont remplacées par les valeurs prédites. Cette procédure intuitive permet de réaliser un filtrage de Kalman pour toutes les composantes du vecteur d'état même quand des observations (mesures disponibles) sont manquantes. Agir ainsi ne semble pas influencer grandement les résultats parce que les variances des erreurs d'estimation sont seulement de quelques pixels, par rapport aux mesures disponibles.

5.2.3.1 Mode Personne Seule Complet (PSComp)

Ce mode est sélectionné quand il n'y a eu aucune réunion temporelle détectée et quand toutes les mesures liées au visage sont disponibles :

- La BERS contient une seule personne (toutes les mesures pour l'estimation de la BEREP sont disponibles).
- Le visage de la personne a été localisé à l'instant t (toutes les mesures pour l'estimation de la BEREV sont disponibles).
- Le visage de la personne a été localisé à l'instant $t-1$ (les mesures d'estimation de vitesse du visage sont disponibles).

La figure 5.6 illustre deux cas de mode de filtrage PSComp pour deux individus dans la scène. Les BERS sont en traits blancs continus et les BEREP et les BEREV sont en traits couleurs pointillés. Nous pouvons noter une mauvaise localisation du visage dans l'image de gauche, mais cette erreur sera corrigée avant la réunion temporelle entre individus.

Dans ce mode, le filtrage de Kalman est réalisé pour toutes les composantes du vecteur d'état. Les équations (5.1) à (5.5) sont utilisées.

5.2.3.2 Mode Personne Seule Partiel (PSPar)

Ce mode est sélectionné quand il n'y a aucune réunion temporelle mais que certaines mesures liées au visage ne sont pas disponibles. Si c'est le cas, l'étape de localisation du visage a échoué à l'instant t et / ou à l'instant $t-1$, alors les mesures indisponibles sont remplacées par les valeurs prédites par le filtrage de Kalman.

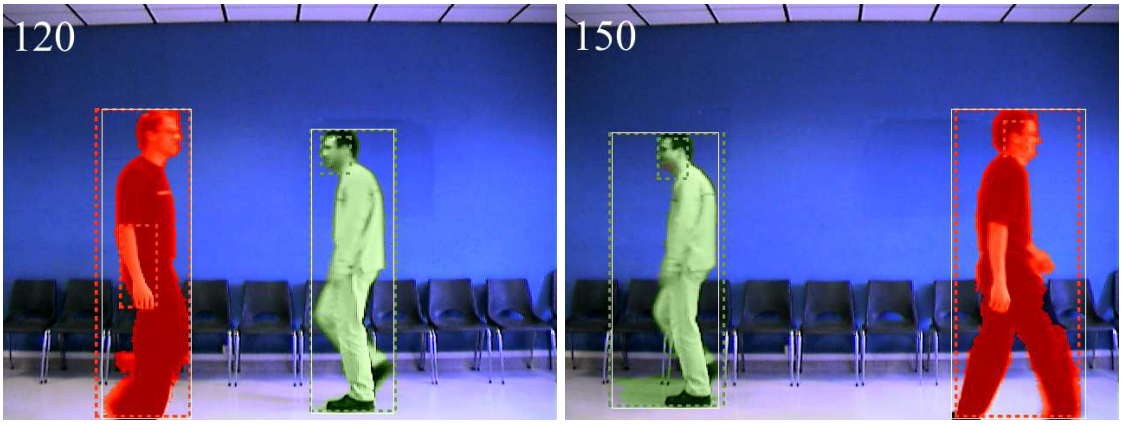


FIG. 5.6 – Exemples de mode de filtrage PSComp.

Le filtrage de Kalman est alors réalisé pour toutes les composantes du vecteur d'état, même celles qui ont été remplacées par les valeurs prédites par Kalman. Les équations (5.1) à (5.5) sont utilisées.

5.2.3.3 Mode Groupe de Personnes Partiel (GPPar)

Ce mode est sélectionné quand il y a une réunion temporelle (c'est-à-dire que certaines mesures pour l'estimation de la BEREP ne sont pas disponibles) et quand la BERPP intersecte un seul visage.

La figure 5.7 illustre deux cas de mode de filtrage GPPar pour deux individus dans la scène. Les BERS sont en traits blancs continus et les BEREP et les BEREV sont en traits couleurs pointillés.

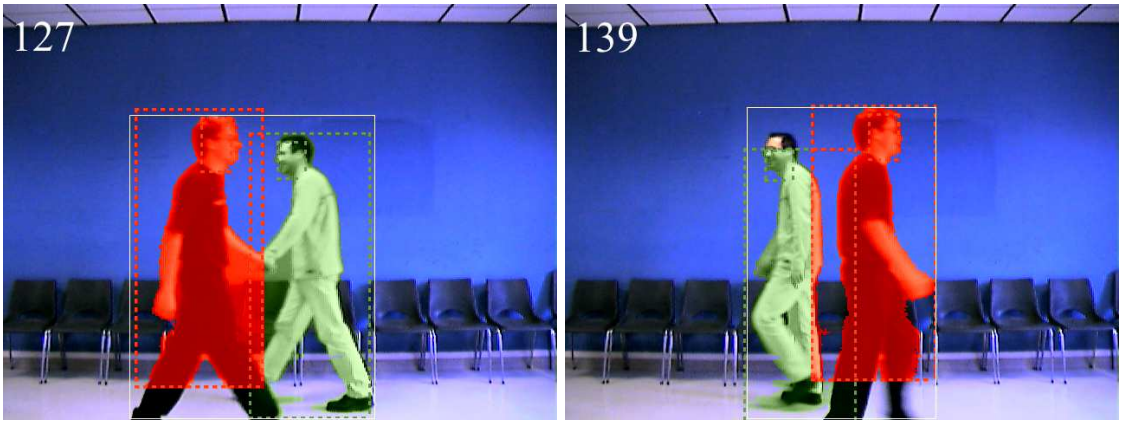


FIG. 5.7 – Exemples de mode de filtrage GPPar.

Comme la BERS contient un groupe de personnes, les mesures disponibles peuvent servir à l'estimation de plusieurs BEREP. L'attribution des mesures disponibles à une personne donnée d'un groupe est décidée en considérant les centres des différentes boîtes et les coordonnées des

côtés. La comparaison est effectuée en deux étapes entre la BERS et les différentes BERPP. La figure 5.8 illustre le principe de l'attribution des mesures.

Dans la première étape, nous comparons les coordonnées des centres des BERPP à celles du centre de la BERS. Par rapport au quart de la BERS où se situe le centre d'une BERPP donnée, les deux coordonnées des côtés les plus proches sont utilisées comme mesures pour l'estimation de la BEREP correspondante. Par exemple, sur la figure 5.8, si deux individus viennent de se réunir (les mains semblent en contact visuellement), nous avons seulement quatre mesures disponibles (au lieu de huit) qui peuvent être utilisées comme observations pour l'estimation des deux BEREP. Avec cette première étape, la personne à gauche dans l'image, notée P_1, recevra les coordonnées des côtés gauche et bas comme mesures, la personne à droite dans l'image, notée P_2, quant à elle, recevra celles des côtés droit et bas. Grâce à cette étape, nous sommes sûrs qu'au moins deux mesures sont transmises et utilisées pour l'estimation de chaque BEREP.

Dans la seconde étape, nous comparons chaque coordonnée des côtés de la BERPP à celle correspondante de la BERS. Si la distance entre les deux est inférieure à un seuil et si cette coordonnée n'a pas encore été prise en compte en tant que mesure, cette coordonnée du côté de la BERS est ajoutée aux mesures disponibles pour l'estimation de la BEREP correspondante. Avec cette étape, dans notre exemple, la personne P_1 reçoit la coordonnée du côté haut de la BERS comme mesure disponible supplémentaire. Cette étape permet généralement l'ajout d'une ou deux mesures afin de réaliser une meilleure estimation. Le seuil de distance a été fixé à une dizaine de pixels, afin de ne pas ajouter des mesures qui sont trop loin des observations réelles.

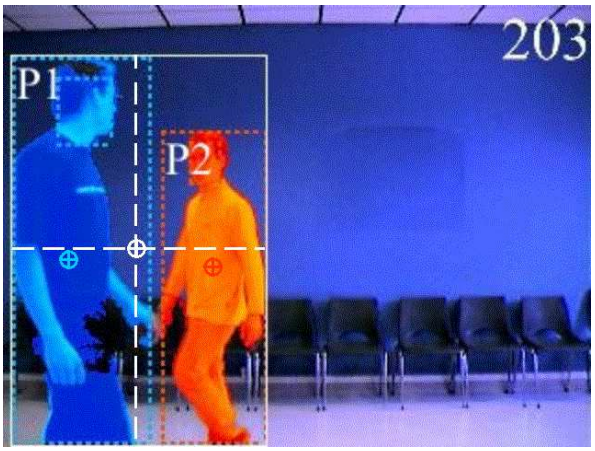


FIG. 5.8 – Principe d'attribution des mesures.

Dans notre exemple, les coordonnées des côtés gauche, haut et bas de la BERS sont utilisées pour l'estimation de la BEREP de la personne P_1 à gauche. Les coordonnées des côtés droit et bas sont utilisées pour l'estimation de la BEREP de la personne P_2 à droite. Comme nous pouvons le constater pour la coordonnée du côté bas dans notre exemple, il est possible que certaines mesures disponibles soient utilisées pour différentes personnes. Pour chaque individu, dans ce mode GPPar, nous avons généralement deux ou trois mesures disponibles (coordonnées des côtés haut et / ou bas, et la coordonnée de l'un des côtés droite ou gauche).

Si certaines mesures liées au visage sont indisponibles, les valeurs prédites par le filtrage

de Kalman remplacent les mesures manquantes. Les équations (5.1) à (5.5) sont utilisées aussi longtemps que chaque BERPP n'intersecte qu'un unique visage. S'il arrive que la BERPP intersecte un deuxième visage, alors le mode de filtrage devient le mode GPPre.

5.2.3.4 Mode Groupe de Personnes Prédictif (GPPre)

Ce mode est sélectionné quand une réunion temporelle est survenue (c'est-à-dire que certaines des mesures pour l'estimation des BEREP sont indisponibles) et quand la BERPP considérée intersecte plus d'un visage.

La figure 5.9 illustre deux cas de mode de filtrage GPPre pour deux individus dans la scène. Les BERS sont en traits blancs continus et les BEREP et les BEREV sont en traits couleurs pointillés.

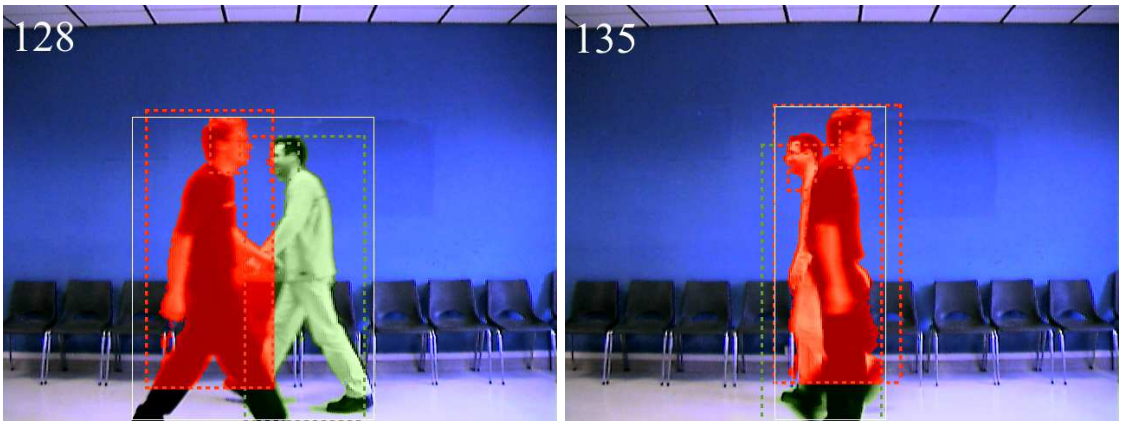


Fig. 5.9 – Exemples de mode de filtrage GPPre.

Quand cette situation survient, c'est que les BERPP s'interpénètrent fortement. Cela arrive quand une personne commence à être majoritairement occultée par une autre personne. L'attribution des mesures par la méthode présentée lors du mode GPPar est trop aléatoire.

Aucune mesure n'est alors prise en compte. Toutes les composantes du vecteur d'état sont prédites selon la dernière estimation de la vitesse du visage. Seules les équations (5.4) et (5.5) sont utilisées. Le filtre de Kalman fonctionne en mode GPPre jusqu'à ce que la BERPP n'intersecte à nouveau qu'un seul visage. Le mode de filtrage avant et après le mode GPPre est très souvent le mode de filtrage GPPar.

5.2.4 Gestion des numéros d'identification *ID*

Dans notre système, les *ID* sont les liens temporels du suivi. Lors de la première étape du suivi temporel, quand un nouvel objet est détecté, il se voit attribuer un *ID* unique. Tant que cet objet est correctement suivi, son *ID* est inchangé. Nous allons maintenant détailler la gestion des numéros d'identification *ID* qui découlent de cette seconde étape du suivi temporel.

Quand un nouvel objet est détecté, l'*ID* qui lui a été attribué par la première étape du suivi temporel est transmis à sa BERPP et à sa BEREP. Ces boîtes conserveront cet *ID* tant qu'il n'aura pas disparu de la scène.

Dans les situations qui n'amènent pas de difficultés par rapport à la première étape du suivi, les modes de filtrage de Kalman utilisés sont les modes PSComp et PSPar. Le numéro d'identification *ID* de la personne reste donc le même que celui attribué par la première étape du suivi temporel. Il n'y a en effet aucun problème de réunion ou de séparation temporelle entre individus.

Quand survient une réunion temporelle entre individus, les modes de filtrage de Kalman utilisés sont les modes GPPar et GPPre. Quel que soit le type d'attribution des *ID* choisi pour l'objet résultant de la réunion temporelle, les BERPP et les BEREP des personnes impliquées dans la réunion temporelle conservent leur *ID*, qui correspond bien à chaque personne. Nous ne considérons donc pas le nouvel *ID* qui correspond à la première étape du suivi temporel pour le groupe de personnes dans son entier. Pendant toute la réunion temporelle, les BERPP et les BEREP conservent leur *ID*. Quand survient la séparation temporelle du groupe de personnes, les modes de filtrage de Kalman utilisés sont à nouveau les modes PSComp et PSPar et les personnes ont été suivies correctement par rapport à la première étape du suivi temporel.

De cette façon, il est possible de suivre au cours du temps plusieurs individus même en cas d'occultations, qu'elles soient partielles ou complètes.

5.3 Données bas-niveau extraites

Les données bas-niveau extraites lors de la seconde étape de suivi temporel sont :
- les numéros d'identification *ID* finaux ;
- les BERPP ;
- les BEREP ;
- les BERPV ;
- les BEREV ;
- les vitesses des visages.

Les principales données extraites lors de cette étape de traitement sont bien sûr les numéros d'identification *ID* finaux. Ils permettent, si le suivi temporel est un succès, de faire les liens temporels pour une même personne sur des images consécutives et ce, même s'il survient des réunions ou des séparations temporelles, en particulier lors d'occultations partielles ou complètes.

Les données bas-niveau secondaires extraites lors de cette étape découlent de l'approche choisie, c'est-à-dire du filtrage de Kalman. Nous avons accès aux prédictions et aux estimations de toutes les composantes du vecteur d'état, donc la boîte englobante rectangulaire prédite de la personne (BERPP), la boîte englobante rectangulaire estimée de la personne (BEREP), la boîte englobante rectangulaire prédite du visage (BERPV), la boîte englobante rectangulaire estimée du visage (BEREV) et la vitesse apparente du visage.

5.4 Résultats

Les figures 5.10 et 5.11 illustrent un suivi temporel réussi dans des séquences vidéo où deux individus se croisent et où l'un d'eux est complètement occulté. Les personnes segmentées et suivies temporellement sont dessinées sur les images originales de la séquence. Les BERS des individus ou des groupes de personnes sont dessinées en lignes blanches. Les BEREP et les BEREV, boîtes englobantes rectangulaires respectivement des individus et de leurs visages, sont dessinées en traits pointillés de couleur.

Sur la figure 5.10, les images 200 et 228 montrent un suivi temporel par filtrage de Kalman en mode PSComp avec toutes les mesures disponibles pour les filtres de Kalman avant la réunion temporelle (image 203) et après la séparation (image 228). Les images 212 et 219 illustrent un suivi temporel en mode GPPre quand un visage est occulté. Les images 203 et 221 (juste avant la séparation) illustrent un suivi temporel en mode GPPar.

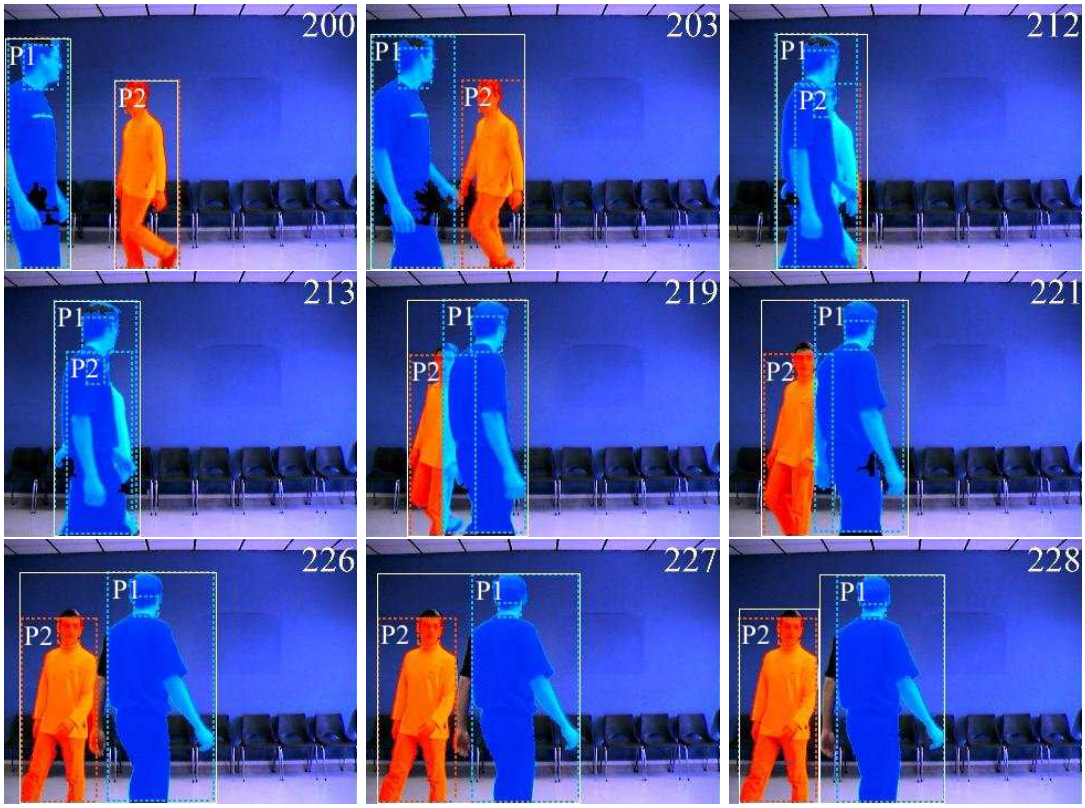


Fig. 5.10 – Exemple 1 : suivi temporel de plusieurs personnes avec occultation complète.

La seconde séquence vidéo, illustrée par la figure 5.11, montre aussi des résultats corrects de suivi temporel pour deux autres individus, bien que le fond de la scène, plus complexe et difficile, amène des problèmes de segmentation non corrigés.

5.5 Avantages, limitations et cadences de traitement

La méthode présentée pour la seconde étape du suivi temporel, qui permet le suivi de personnes dans un groupe, est basée sur un filtrage de Kalman partiel et une poursuite de visage.

Du côté des avantages, les cadences de traitement pour cette étape de traitement sont relativement correctes (cf. table 5.1, page 140). Pour la segmentation basée sur les champs aléatoire de Markov, cette étape ne prend pas beaucoup de temps. Mais pour la segmentation optimisée en vitesse, c'est l'étape la plus coûteuse à ce stade du système. En ce qui concerne la robustesse, cette méthode s'est révélée robuste lors de déplacements localement assimilables à des translations (par exemple quand deux individus se croisent, puis tournent autour l'un

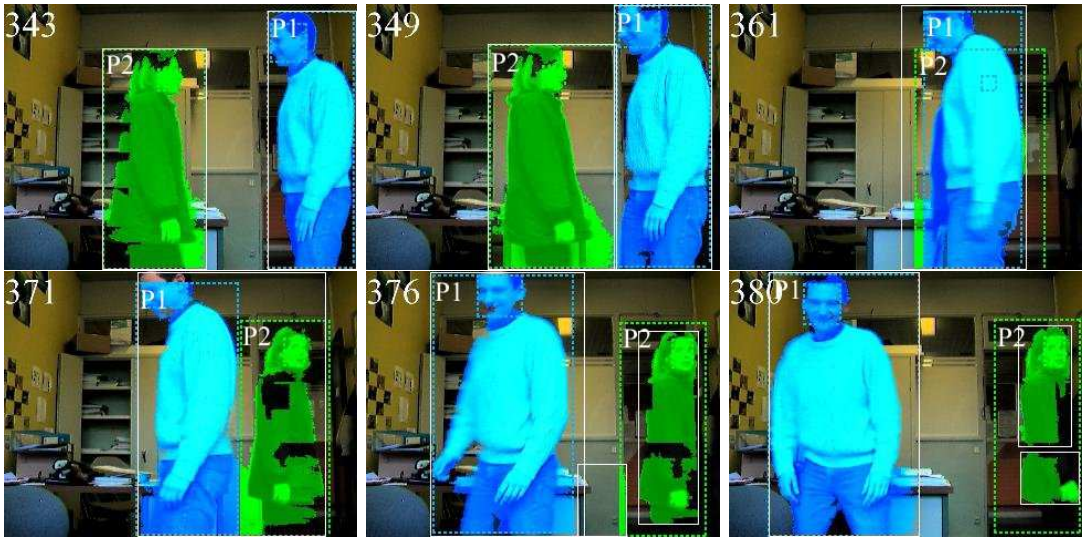


FIG. 5.11 – Exemple 2 : suivi temporel de plusieurs personnes avec occultation complète.

de l'autre). Les résultats sont illustrés dans la partie 5.4. Ce succès est dû au fait que même si les personnes ne forment toujours qu'un seul objet segmenté du point de vue de la caméra, il suffit que leurs visages respectifs soient assez éloignés pour que la seconde étape du suivi temporel soit efficace. Dans ce cas, un seul visage intersectant la BEREP, si la localisation du visage est correcte, alors l'estimation de sa vitesse permettra de mieux gérer le filtrage des paramètres du vecteur d'état, et par conséquent de mieux estimer la position de la personne à l'intérieur du groupe.

Du côté des limitations, cette méthode dépend grandement de l'étape de localisation du visage. Si cette étape amène une mauvaise localisation pour une personne, il sera difficile de la suivre lors d'un croisement. Ensuite, cette méthode est, lors d'une occultation complète, assez sensible à la dernière estimation de la vitesse du visage de la personne qui est occultée. En effet, en mode GPPre, on se base uniquement sur cette estimation. Néanmoins, lors de croisements qui ne durent pas longtemps, le suivi pourra se raccorder sur la personne même si, à cause du filtrage en mode GPPre, les prévisions se sont décalées spatialement par rapport à la personne seule.

Du côté des pourcentages de temps de calcul et des cadences de traitement atteintes, la table 5.1 présente les résultats obtenus pour l'ensemble des étapes de traitement jusqu'à la seconde étape du suivi temporel, qui est notée (2/2). Avec la segmentation optimisée en vitesse, nous sommes proches de la cadence vidéo (en résolution 640×480) ou encore assez nettement au-dessus (en résolution 320×240).

5.6 Conclusion

Ce chapitre a présenté la seconde étape du suivi temporel pour suivre des individus à l'intérieur d'un groupe. Cette seconde étape du suivi temporel est basée sur un filtrage de Kalman partiel et une poursuite du visage. Nous avons détaillé les principes de cette méthode, notamment les différents modes de filtrage. Puis nous avons décrit les données bas-niveau

TAB. 5.1 – Pourcentages de temps de calcul et cadences de traitement pour la seconde étape du suivi temporel.

Segmentation	Champs aléatoires de Markov		Optimisée en vitesse	
Résolution d'image	320 × 240	640 × 480	320 × 240	640 × 480
Acquisition	0.2%	0.3%	1.4%	3.2%
Segmentation	80.3%	87.4%	19.9%	24.7%
Suivi temporel (1/2)	9.3%	1.6%	2.5%	0.7%
Localisation et suivi du visage et des mains	8.5%	2.3%	35.3%	28.1%
Suivi temporel (2/2)	1.7%	8.4%	40.9%	43.3%
Cadences de traitement	**7.77** images/s	**1.85** images/s	**63** images/s	**18** images/s

extraites lors de cette étape de traitement. Après avoir illustré quelques résultats de suivi temporel pour cette seconde étape, nous avons donné les avantages, les limitations et les cadences de traitement atteintes. Le principal avantage de la méthode est sa robustesse et le fait qu'elle permette de suivre des personnes au cours du temps même si surviennent des occultations partielles ou complètes. Sa principale limitation est qu'elle dépend grandement de l'étape de localisation du visage et qu'elle échoue si les occultations sont de longue durée. Ce suivi temporel devrait encore être efficace même si les personnes sont occultées par des objets fixes, du moment que leur mouvement global reste cohérent avec le mouvement de leur visage et que leur visage est correctement détecté, localisé et suivi.

Au niveau des perspectives envisagées pour le suivi temporel de personnes dans un groupe par filtrage de Kalman partiel et poursuite du visage, un moyennage temporel de la vitesse du visage pourrait être ajouté afin de rendre le suivi plus robuste en fonction de la dernière estimation de la vitesse. Peu d'autres perspectives sont envisagées sans changer d'approche. Les difficultés et limitations qui restent à éliminer ne peuvent pas l'être facilement en gardant cette approche de poursuite du visage. Nous pourrions essayer de voir s'il est possible de garder en mémoire, pour un objet résultant d'une réunion temporelle, d'autres caractéristiques des objets lui ayant donné naissance, il serait alors peut-être plus simple de reconnaître les objets résultants de la séparation temporelle suivante. Il serait intéressant, dans cet état d'esprit, de mettre en œuvre la méthode basée sur les pistes élémentaires, les modèles d'apparence et la forme présentée dans [Mostafaoui05]. Cela permettrait de comparer les résultats de suivi temporel avec notre méthode.

Ce chapitre a présenté la dernière étape de la phase d'analyse qui concerne l'extraction de données bas-niveau. Nous allons maintenant présenter l'étape de traitement de la phase d'interprétation de plus haut niveau où certaines des informations bas-niveau extraites lors des étapes précédentes vont être utilisées pour réaliser une reconnaissance de postures statiques. Cette étape haut-niveau d'interprétation du comportement humain en tant que reconnaissance de postures nécessite des processus de fusion de données qui vont faire l'objet du prochain chapitre.

Chapitre 6

Fusion de données et reconnaissance de postures statiques

Sommaire

Après avoir réussi à extraire des personnes du fond de la scène (segmentation 2D spatio-temporelle), à les suivre temporellement seules ou en groupe en gérant les problèmes d'occultation (suivi temporel en deux étapes), et à détecter leurs visages et leurs mains (localisation et suivi temporel du visage et des mains), le problème de la compréhension du comportement humain se pose naturellement. Il inclut la reconnaissance de comportements (postures et / ou actions). La définition et la classification parmi un ensemble d'actions ou de postures typiques (courir, être debout, grimper, sauter, pointer, etc.) est souvent guidée par le type d'application visée. La plupart du temps, il existe aussi un compromis entre la complexité calculatoire et la précision de la reconnaissance. La reconnaissance est basée sur certaines données bas-niveau. Les données bas-niveau sont extraites en utilisant une analyse dynamique de la séquence vidéo. Selon le type d'application visée, une partie ou l'ensemble de ces données bas-niveau peut être utilisé pour réaliser une interprétation et une reconnaissance de haut-niveau du comportement humain [Canton-Ferrer06].

La majorité des travaux de recherche effectués sur le corps humain en entier concerne principalement la reconnaissance de démarche, la reconnaissance d'interactions entre des personnes, ou entre des personnes et des objets. Le système W^4 [Haritaoglu98], par exemple, est capable de reconnaître des gens en train de prendre, de déposer ou d'échanger des objets. Dans notre système, nous nous intéressons à la reconnaissance de postures statiques.

Nous commencerons par un état de l'art sur les principales méthodes de fusion de données, puis nous exposerons les principes de la théorie de l'évidence. Ensuite nous détaillerons notre méthode de reconnaissance de postures statiques qui est basée sur la théorie de l'évidence. Nous présenterons alors différents classifieurs définis dans ce but et les résultats de reconnaissance et de classification obtenus avec ces classifieurs. Quelques résultats de reconnaissance de postures seront illustrés en images issues de séquences vidéo variées. Puis nous donnerons les avantages, les limitations et les cadences de traitement pour cette étape avant de conclure ce chapitre.

6.1 État de l'art sur la fusion de données

En fusion de données, les trois grandes approches les plus souvent utilisées afin de prendre une décision parmi un ensemble d'hypothèses sont la théorie des probabilités, la théorie de l'évidence et la théorie des possibilités. Quelques articles ont essayé de comparer ces théories [Bloch96a]. Ceci est une tâche ardue principalement parce que ces trois théories ne représentent et ne traitent pas le même type d'information. Une connaissance statistique des mesures est une bonne raison d'utiliser la théorie des probabilités [Silverman86]. Si une connaissance experte est disponible, l'utilisation de la théorie des possibilités, avec des sous-ensembles flous, est généralement préférée [Dubois94]. La théorie de l'évidence permet de traiter des informations imprécises et / ou contradictoires [Smets94]. De plus, autant une connaissance statistique qu'une connaissance experte peut aider à améliorer les résultats de reconnaissance / classification.

Cet état de l'art propose un panorama des principales méthodes de fusion de données, largement inspiré de [Bloch05]. Il s'agit des méthodes classiques selon la théorie des probabilités (en particulier l'inférence bayésienne), mais aussi de méthodes non probabilistes (selon la théorie des possibilités et des ensembles flous et selon la théorie de l'évidence ou des fonctions de croyance), apparues plus récemment, mais qui connaissent un essor de plus en plus important. Pour chaque théorie, nous présentons les deux composantes essentielles des systèmes de

fusion : la représentation des connaissances et le raisonnement.

Nous définissons la fusion d'informations au sens large comme la combinaison d'informations hétérogènes issues de plusieurs sources afin d'améliorer la prise de décision. Concernant les informations (aussi appelées connaissances ou données) utilisées, elles peuvent présenter des caractéristiques générales. Principalement, les informations présentent une certaine imperfection. Celle-ci est toujours présente (sinon la fusion ne serait pas nécessaire). Cette imperfection peut prendre diverses formes que nous allons présenter brièvement. L'**incertitude** est relative à la véracité d'une information, et caractérise son degré de conformité à la réalité [Dubois88]. Elle fait référence à la nature de l'objet ou du fait concerné, à sa qualité, à son essence ou à son occurrence. L'**imprécision** concerne le contenu de l'information et mesure donc son défaut quantitatif de connaissance, sur une mesure [Dubois88]. L'**incomplétude** caractérise l'absence d'information apportée par la source sur certains aspects du problème. L'**ambiguïté** exprime la capacité d'une information de conduire à deux interprétations. Elle peut provenir des imperfections précédentes. Le **conflit** caractérise deux ou plusieurs informations conduisant à des interprétations contradictoires et donc incompatibles. Les situations conflictuelles sont fréquentes dans les problèmes de fusion et posent toujours des problèmes difficiles à résoudre. D'autres caractéristiques de l'information sont plus positives et sont exploitées pour limiter les imperfections. La **redondance** est la qualité de sources qui apportent plusieurs fois la même information. La redondance entre les sources est souvent observée dans la mesure où les sources donnent des informations sur le même phénomène. Idéalement, la redondance est exploitée pour réduire les incertitudes et les imprécisions. La **complémentarité** est la propriété des sources qui apportent des informations sur des grandeurs différentes. Elle vient du fait qu'elles ne donnent en général pas d'informations sur les mêmes caractéristiques du phénomène observé. Elle est exploitée directement dans le processus de fusion pour avoir une information globale plus complète et pour lever les ambiguïtés.

L'intérêt de la fusion de données est d'obtenir une synthèse des informations fournies par plusieurs sources. L'utilisation de plusieurs sources se justifie par l'espoir d'atteindre un résultat plus stable et plus pertinent qu'avec une source unique. Chacune des sources fournit une vision du monde observé. En utilisant un formalisme et un opérateur de combinaison qui permet d'obtenir une représentation synthétique du point de vue des sources, il est ensuite possible de prendre une décision plus fiable et plus représentative de la réalité.

En général, la fusion n'est pas une tâche simple. Elle peut se décomposer de manière schématique en deux étapes que nous allons décrire succinctement. Considérons un problème général de fusion de données pour lequel on dispose de l sources $S_1, S_2, \ldots, S_l$, et pour lequel le but est de prendre une décision parmi un ensemble de n décisions possibles $d_1, d_2, \ldots, d_n$. Les deux étapes à résoudre pour construire le processus de fusion sont les suivantes :

1. Modélisation : cette étape comporte le choix d'un formalisme, et de la représentation des informations à fusionner dans ce formalisme. Beaucoup de modélisations nécessitent une phase d'estimation (comme les méthodes utilisant les distributions). La modélisation peut par conséquent être guidée par des informations supplémentaires (sur les données et sur le contexte ou le domaine). Supposons pour fixer les idées que chaque source S_j fournisse une information M_i^j sur la décision d_i. La forme (représentation) de M_i^j dépend bien sûr du formalisme choisi.
2. Combinaison ou fusion : cette étape concerne le choix d'un opérateur, compatible avec le formalisme de modélisation retenu, et guidé par les informations supplémentaires.

En s'appuyant sur les résultats de la fusion, une prise de décision est alors effectuée par rapport aux informations fusionnées où l'incertitude est minimale.

La manière dont les deux étapes précédentes sont agencées définit le système de fusion et son architecture. En particulier, on distingue les systèmes décentralisés dans lesquels des décisions locales sont prises au niveau de chaque source séparément puis sont fusionnées en une décision globale, et les systèmes centralisés dans lesquels on combine par une fonction F tous les M_i^j relatifs à une même décision d_i, pour obtenir une forme fusionnée $M_i = F(M_i^1, M_i^2, \ldots, M_i^l)$, puis une décision est prise sur le résultat de cette combinaison. Ces deux systèmes ayant des propriétés différentes, nous ne considérerons que des systèmes centralisés dans cet état de l'art.

6.1.1 Fusion probabiliste et bayésienne

6.1.1.1 Mesures d'information

Lorsque l'on dispose d'un ensemble de l sources S_j, une première tâche consiste souvent à le transformer en un sous-ensemble plus réduit, donc de traitement plus simple, sans perdre d'information. Pour exprimer l'apport d'information dû à l'ajout d'une source S_{l+1} à un ensemble déjà connu $\{S_1, S_2, \ldots, S_l\}$, les notions d'information et d'entropie [Kullback59, Maître96] (entropie jointe et entropie conditionnelle) sont bien adaptées. On définit ainsi la redondance entre deux sources par :

$$R(S_1, S_2) = H(S_1) + H(S_2) - H(S_1, S_2),$$

où H est l'entropie ; et la complémentarité de la source S_2 par rapport à la source S_1, c'est-à-dire la quantité moyenne d'information qu'il faut ajouter à S_2 pour retrouver S_1 :

$$C(S_1/S_2) = H(S_1/S_2).$$

Des approches analogues peuvent être envisagées dans un cadre non probabiliste, en s'appuyant par exemple sur l'entropie floue [De Luca72]. Le formalisme est pour l'instant moins développé dans cette direction.

6.1.1.2 Modélisation et estimation

La théorie la plus exploitée dans la littérature est de loin la théorie des probabilités, associée à la théorie bayésienne de la décision [Duda73]. L'information y est modélisée par une probabilité conditionnelle. Ainsi la mesure s'écrit elle sous la forme :

$$M_i^j = P(d_i/S_j).$$

Cette probabilité est calculée à partir de caractéristiques de l'information extraites à partir des sources disponibles. L'apprentissage des distributions s'appuie sur des outils statistiques classiques. C'est en général $P(S_j/d_i)$ qui peut être estimée, et on en déduit la probabilité précédente par application de la règle de Bayes.

L'avantage essentiel des méthodes probabilistes vient de ce qu'elles reposent sur une base mathématique solide et ont été l'objet de nombreux travaux. Elles proposent donc un éventail d'outils très riche permettant aussi bien la modélisation que l'apprentissage des modèles. Elles proposent également des règles d'usage soit théoriques soit heuristiques. Mais malgré

ces avantages, elles présentent aussi quelques limitations. Tout d'abord, si elles représentent bien l'incertain qui entache l'information, elles ne permettent pas aisément de représenter son imprécision, et elles conduisent souvent à confondre ces deux notions. Ensuite, elles nécessitent que, lors de l'apprentissage, des contraintes très strictes soient vérifiées par les mesures (imposées par les axiomes de base des probabilités) et par l'ensemble de classes considéré (exhaustivité). Ces contraintes peuvent rendre l'apprentissage très délicat ou, si le problème à traiter est complexe, conduire pratiquement à des incohérences car l'utilisateur ne peut alors prendre en compte tout le réseau des dépendances probabilistes. L'apprentissage de lois de probabilités nécessite, outre les hypothèses, un nombre de données important.

6.1.1.3 Combinaison dans un cadre bayésien

Dans le modèle bayésien, la fusion peut être effectuée de manière équivalente à deux niveaux :

– soit au niveau de la modélisation, et on calcule alors les probabilités de la forme :

$$P(d_i/S_1, S_2, \ldots, S_l),$$

à l'aide de la règle de Bayes :

$$P(d_i/S_1, S_2, \ldots, S_l) = \frac{P(S_1, S_2, \ldots, S_l/d_i)P(d_i)}{P(S_1, S_2, \ldots, S_l)},$$

où les différents termes sont estimés par l'apprentissage ;

– soit par la règle de Bayes elle-même, où l'information issue d'une source vient mettre à jour l'information estimée d'après les sources précédentes :

$$P(d_i/S_1, S_2, \ldots, S_l) = \frac{P(S_1/d_i)P(S_2/d_i, S_1)\ldots P(S_l/d_i, S_1, \ldots, S_{l-1})P(d_i)}{P(S_1)P(S_2/S_1)\ldots P(S_l/S_1, \ldots, S_{l-1})}.$$

Très souvent, étant données la complexité de l'apprentissage à partir de plusieurs capteurs et la difficulté d'obtenir des statistiques suffisantes, ces équations sont simplifiées sous l'hypothèse d'indépendance. Là encore, des critères ont été proposés pour vérifier la validité de ces hypothèses. Les formules précédentes deviennent alors :

$$P(d_i/S_1, \ldots, S_l) = \frac{\prod_{j=1}^{l} P(S_j/d_i)P(d_i)}{P(S_1, \ldots, S_l)}.$$

Cette équation fait clairement apparaître le type de combinaison des informations, sous la forme d'un produit, donc une fusion conjonctive. Il est notable que la probabilité *a priori* joue exactement le même rôle dans la combinaison que chacune des sources, auxquelles elle est combinée également par un produit.

Malgré les avantages de cette combinaison, elle est contrainte, comme pour la modélisation, par les axiomes des probabilités, et son utilisation en pratique nécessite souvent des hypothèses simplificatrices (comme l'indépendance) rarement vérifiées. Elle nécessite de plus l'estimation des probabilités *a priori* $P(d_i)$, qui est souvent délicate et est primordiale dans les cas où l'on a peu d'informations. La forme conjonctive de la combinaison bayésienne conduit souvent en pratique à un effondrement des probabilités des événements qui sont déduits d'une longue chaîne de déduction. Enfin, elle ne permet pas de modéliser l'ignorance pour la prendre en compte dans la combinaison.

6.1.1.4 Combinaison vue comme un problème d'estimation

Une autre manière de voir la fusion probabiliste consiste à considérer que chaque source donne une probabilité (d'appartenance à une classe par exemple), et que la fusion consiste à combiner ces probabilités pour trouver la probabilité globale d'appartenance à la classe. Cette vision revient à considérer la fusion comme un problème d'estimation, et permet d'utiliser des opérateurs de combinaison différents du produit. En particulier les méthodes de moyenne ou moyenne pondérée, de médiane, de consensus sont souvent employées [French85, Cooke91]. Des estimateurs robustes peuvent également être employés, afin de limiter ou supprimer l'influence des valeurs aberrantes (*outliers*).

6.1.1.5 Décision

La dernière étape concerne la décision, par exemple le choix de la classe à laquelle appartient un point. Cette décision binaire peut être assortie d'une mesure de la qualité de cette décision, pouvant éventuellement conduire à la rejeter. La règle la plus utilisée pour la décision probabiliste est le critère *MAP (Maximum A Posteriori)* :

$$d_i \quad \text{si} \quad P(d_i/S_1,\ldots,S_l) = \max_{k=1}^{n} P(d_k/S_1,\ldots,S_l),$$

mais de très nombreux autres critères ont été développés par les probabilistes et les statisticiens, pour qu'ils s'adaptent au mieux aux besoins de l'utilisateur et au contexte de sa décision : maximum de vraisemblance, maximum d'entropie, marginale maximale, espérance maximale, risque minimal, etc. Cependant, la grande variété de ces critères laisse l'utilisateur à nouveau démuni devant la justification d'un choix et l'éloigne de l'objectivité recherchée initialement par ces méthodes.

6.1.2 Fusion floue et possibiliste

6.1.2.1 Modélisation

La théorie des ensembles flous fournit un très bon outil pour représenter explicitement des informations imprécises, sous la forme de fonctions d'appartenance [Zadeh65]. La mesure M_i^j prend alors la forme $M_i^j = \mu_i^j$, où μ_i^j désigne par exemple le degré d'appartenance d'un élément à la classe d_i selon la source S_j, ou la traduction d'une information symbolique exprimée par une variable linguistique. Ces fonctions ne souffrent pas des contraintes axiomatiques imposées aux probabilités et offrent donc une plus grande souplesse lors de la modélisation. Cette souplesse peut être considérée comme un inconvénient puisqu'elle laisse facilement l'utilisateur démuni pour définir ces fonctions. L'inconvénient des ensembles flous est qu'ils représentent essentiellement le caractère imprécis des informations, l'incertitude étant représentée de manière implicite et n'étant accessible que par déduction à partir des différentes fonctions d'appartenance. La théorie des possibilités [Zadeh78, Dubois88], dérivée des ensembles flous, permet de représenter à la fois l'imprécision et l'incertitude, par l'intermédiaire de distributions de possibilités π et de deux fonctions caractérisant les événements : la possibilité Π et la nécessité N. Dans le cadre de la fusion de données, une application possible de cette théorie consiste à définir π sur D (l'ensemble des décisions possibles) et la mesure M_i^j par $M_i^j = \pi_j(d_i)$, c'est-à-dire comme le degré de possibilité de la décision d_i selon la source S_j. Dans un problème de classification, cette modélisation suppose que les classes (ou décisions) sont nettes, alors que le modèle flou suppose que les classes sont floues.

La construction des fonctions d'appartenance ou distributions de possibilités peut être effectuée de plusieurs manières. Dans la plupart des applications, cette construction est faite en s'inspirant directement des méthodes d'apprentissages probabilistes, soit par des heuristiques, soit par des méthodes neuromimétiques permettant d'apprendre les paramètres de formes particulières de fonctions d'appartenance, soit enfin par la minimisation de critères de classification [Bezdek81]. Plusieurs méthodes ont également été proposées pour transformer une distribution de probabilités en distribution de possibilités [Dubois83, Bharati Devi85, Klir92]. D'autres méthodes cherchent à estimer directement les fonctions d'appartenance à partir de l'histogramme, afin d'optimiser des critères d'entropie [Cheng97], ou de minimum de spécificité et de cohérence [Civanlar86].

6.1.2.2 Combinaison

Un des intérêts de la théorie des ensembles flous et des possibilités, outre qu'elle impose peu de contraintes au niveau de la modélisation, est qu'elle offre une grande variété d'opérateurs de combinaison. Une caractéristique importante, commune à toutes les théories, de ces opérateurs de combinaison est qu'ils fournissent un résultat de même nature que les fonctions de départ (propriété de fermeture) et qui a donc la même interprétation en termes d'imprécision et d'incertitude. Ainsi ils permettent de ne prendre aucune décision binaire partielle avant la combinaison, ce qui pourrait conduire à des contradictions difficiles à lever. La décision n'est prise qu'en dernier lieu, sur le résultat de la combinaison.

Dans la théorie des ensembles flous et des possibilités, de multiples modes de combinaison sont possibles [Dubois85, Yager91]. Parmi les principaux opérateurs, on trouve en particulier les t-normes, les t-conormes [Menger42, Schweizer83], les moyennes [Yager88, Grabisch95], les sommes symétriques, et des opérateurs prenant en compte les mesures de conflit ou encore de fiabilité des sources [Dubois92, Deveughele93].

Le choix d'un opérateur peut se faire selon différents critères [Bloch96a]. Un premier critère est le comportement de l'opérateur. Des comportements sévères, indulgents ou prudents se traduisent sous forme mathématique de conjonction, disjonction ou de compromis. Soient x et y deux réels (dans $[0, 1]$) représentant les degrés de confiance à combiner. La combinaison de x et y par un opérateur F est dite :

- conjonctive si $F(x, y) \leq \min(x, y)$ (comportement sévère) ;
- disjonctive si $F(x, y) \geq \max(x, y)$ (comportement indulgent) ;
- de compromis si $x \leq F(x, y) \leq y$ si $x \leq y$, et $y \leq F(x, y) \leq x$ sinon (comportement prudent).

Mais un opérateur n'a pas toujours le même comportement selon les valeurs des informations à combiner. Ainsi, la classification présentée dans [Bloch96a] présente les opérateurs de la façon suivante :

1. opérateurs à comportement constant : le résultat ne dépend que des valeurs à combiner et le comportement est le même quelles que soient ces valeurs (par exemple, quelles que soient les valeurs de x et de y, le comportement est toujours sévère) ;
2. opérateurs à comportement variable : le résultat ne dépend que des valeurs à combiner mais le comportement dépend de ces valeurs (par exemple, si les valeurs de x et de y sont faibles, un comportement indulgent peut être adopté pour tenir compte de faibles degrés de confiance, et si les valeurs sont élevées, le comportement peut être choisi sévère ; cet exemple montre la différence avec un comportement prudent) ;

3. opérateurs dépendant du contexte : le résultat dépend en plus d'une connaissance plus globale telle que la fiabilité des capteurs, ou le conflit entre les sources.

Un autre critère est donné par les propriétés des opérateurs et leur interprétation en termes de fusion de données incertaines, imprécises, incomplètes ou encore ambiguës. D'autres critères peuvent encore être utilisés, comme la qualité de la décision à laquelle ils conduisent, leur pouvoir discriminant ou encore leur capacité à combiner des informations quantitatives (numériques) ou qualitatives (pour lesquelles seul un ordre est connu) [Dubois99].

6.1.2.3 Décision

La règle principalement utilisée en fusion floue est le maximum des degrés d'appartenance :

$$d_i \quad \text{si} \quad \mu_i(x) = \max_{k=1}^{n}\{\mu_k(x)\},$$

où $\mu_k(x)$ désigne la fonction d'appartenance à la classe k résultant de la combinaison. La qualité de la décision est mesurée essentiellement selon deux critères :

- le premier porte sur la "netteté" de la décision : le degré d'appartenance maximum (ou plus généralement celui correspondant à la décision) est comparé à un seuil, choisi selon les applications (et éventuellement selon l'opérateur de combinaison choisi) ;
- le deuxième porte sur le caractère "discriminant" de la décision, évalué par comparaison des deux valeurs les plus fortes.

Dans le cas où ces critères ne sont pas vérifiés pour un élément x, celui-ci est généralement placé dans une classe de rejet, ou reclassé en fonction d'autres critères.

6.1.3 Fusion dans la théorie des fonctions de croyance

La théorie des fonctions de croyance, aussi appelée théorie de l'évidence ou théorie de Dempster-Shafer, date des années 70. Nous détaillerons les principes de cette théorie dans la partie 6.2, et nous présenterons les caractéristiques de cette théorie qui justifient que l'on s'y intéresse, aussi bien du point de vue de la représentation des connaissances et de leurs imperfections (imprécision, incertitude, ambiguïté, ignorance, conflit) que de leur combinaison.

6.1.4 Approches générales utilisées pour la reconnaissance

La reconnaissance d'actions associées au comportement humain peut être vue comme un problème d'association de données variant au cours du temps. Les approches générales utilisées sont principalement la transformation dynamique temporelle *DTW (Dynamic Time Warping)*, les chaînes de Markov cachées *HMM (Hidden Markov Models)* et les réseaux de neurones *NN (Neural Networks)*. Nous ne présenterons que très brièvement ces approches, pour des raisons de place, mais nous donnons un certain nombre de références les décrivant plus en détails.

6.1.4.1 Transformation dynamique temporelle *DTW (Dynamic Time Warping)*

La transformation dynamique temporelle *DTW*, très utilisée au départ pour la reconnaissance de parole [Myers80], est une technique d'association basée sur des gabarits ou schémas et

réalisée par programmation dynamique. Elle a l'avantage d'être relativement simple conceptuellement et elle permet l'obtention de performances robustes. Elle a été utilisée en reconnaissance de schémas de comportement humain (parole, mouvement, etc.) [Darrell93, Takahashi94, Bobick95, Bobick97]. Même si les échelles de temps entre un schéma de test et un schéma de référence ne sont pas exactement les mêmes, tant que les contraintes d'ordre temporel sont respectées, la *DTW* peut toujours établir des relations cohérentes entre ces schémas.

6.1.4.2 Chaînes de Markov cachées *HMM (Hidden Markov Models)*

Une façon plus sophistiquée de réaliser l'association de données variant au cours du temps est possible grâce aux chaînes de Markov cachées, ou *HMM (Hidden Markov Models)*. Une description précise de cette technique est disponible dans [Poritz88] et [Rabiner89].

Les *HMM* sont des machines à états non déterministes qui, selon une entrée, passent d'un état à un autre selon des probabilités de transition variées. Dans chaque état, les *HMM* peuvent générer des symboles de sortie de façon probabiliste. L'utilisation des *HMM* implique une étape d'apprentissage et une étape de classification. L'étape d'apprentissage consiste à spécifier le nombre d'états (éventuellement cachés) et à optimiser les transitions entre états et les probabilités des sorties de telle façon que les symboles de sortie générés correspondent aux caractéristiques de l'image observées pendant des exemples particuliers de classes de mouvement. Le principal outil dans les *HMM* est l'algorithme de Baum-Welch avant-arrière (*forward-backward*) pour l'estimation des symboles de sortie selon un critère de maximum de vraisemblance. L'association implique le calcul des probabilités qu'un *HMM* donné puisse avoir généré le symbole de test qui correspond aux caractéristiques de l'image observée. La capacité à apprendre à partir de données d'apprentissage et à développer des représentations internes dans un cadre mathématique précis et la possibilité de traiter des données non bornées temporellement font que les *HMM* sont plus intéressants que la *DTW* [Yamato92, Starner95b, Starner95a, Bregler97, Brand99, Wren00, Rigoll00, Nair02].

6.1.4.3 Réseaux de neurones *NN (Neural Networks)*

Une approche également intéressante pour réaliser l'association de données variant dans le temps est possible grâce aux réseaux de neurones ou *NN (Neural Networks)*. Une description précise des réseaux de neurones est disponible dans [Hertz91, Bishop95, Haykin99].

Un réseau de neurones est d'abord et avant tout un graphe, avec des schémas représentés sous forme de valeurs numériques attachées aux nœuds du graphe et des transformations entre ces schémas permises grâce à des algorithmes simples de transmission de message. Certains nœuds du graphe sont généralement distingués comme étant des nœuds d'entrées ou des nœuds de sortie, et le graphe dans son entier peut être vu comme la représentation d'une fonction multivariable reliant les entrées aux sorties. Des valeurs numériques (poids) sont associées aux nœuds du graphe, paramétrant la fonction d'entrée / sortie et lui permettant d'être ajustée selon un algorithme d'apprentissage. De même que les *HMM*, les réseaux de neurones ont donc besoin d'être entraînés afin de reconnaître des schémas caractéristiques.

De nos jours, il n'est pas rare de disposer de bases de données de taille conséquente, par conséquent l'accent est mis sur les réseaux de neurones pour représenter l'information temporelle car il y a alors suffisamment de données pour les entraîner [Guo94b, Rosenblum94, Rosales00]. Par exemple, [Guo94b] *et al.* utilisent un réseau de neurones pour comprendre des

schémas de mouvement humain. Rosenblum *et al.* se servent d'un réseau de neurones pour interpréter les émotions humaines à partir de caractéristiques du mouvement [Rosenblum94].

6.1.4.4 Autres approches

En plus des trois grandes approches que nous venons de présenter très brièvement, apparaissent aussi dans la littérature l'analyse en composantes principales *PCA (Principle Component Analysis)* [Chomat98, Yacoob98], et des variantes des approches précédentes (*HMM* et *NN*) comme les chaînes de Markov cachées couplées *CHMM (Coupled Hidden Markov Models)* [Brand97], les chaînes de Markov à longueur variable *VLMM (Variable Length Markov Models)* [Galata01] et les réseaux de neurones à délai temporel *TDNN (Time Delay Neural Networks)* [Lin99].

6.2 Théorie de l'évidence

Le modèle de croyance transférable *TBM (Transferable Belief Model)* [Smets94] a été introduit par Smets et Kennes en 1994. Il poursuit les travaux de Dempster et Shafer [Dempster68, Shafer76]. Les principaux avantages de la théorie de l'évidence sont la possibilité de modéliser l'incertitude associée aux données, l'imprécision sur la cardinalité des hypothèses et le conflit (un conflit survient quand les mesures utilisées pour la reconnaissance conduisent à des résultats contradictoires).

De nombreux ouvrages et travaux présentent le formalisme de cette théorie de façon plus ou moins complexe [Dempster68, Shafer76, Smets90a, Smets90b, Smets93, Smets94, Smets98, Rombaut01]. La théorie de l'évidence a été utilisée par exemple en analyse du mouvement humain pour la reconnaissance d'expressions faciales [Hammal05] et en segmentation pour l'imagerie médicale [Rombaut02, Capelle03, Capelle04]. Capelle, Colot et Fernandez-Maloigne présentent une méthode de segmentation basée sur la théorie de l'évidence pour la reconnaissance de parties du cerveau et la détection de tumeurs [Capelle03, Capelle04]. Nous appliquerons la théorie de l'évidence à la reconnaissance de postures statiques (cf. partie 6.3). Nous rappellerons d'abord ici, dans les grandes lignes, les principales définitions et notations, proches de celles utilisées par Smets, et nous illustrerons les différents concepts par l'étude d'un exemple simple. L'exemple choisi est l'observation, grâce à un ou deux capteurs, de la face supérieure d'un dé comportant six faces, après un jet, ceci afin de déterminer le résultat du jet [Rombaut01].

6.2.1 Cadre de discernement et espace de définition

Dans le cas général, nous observons un système réel dont l'état X est inconnu. C'est cet état que l'on cherche à déterminer. Le **cadre de discernement** Ω, aussi appelé monde représente un ensemble d'états qui peuvent être atteints. Il peut être discret ou continu. Nous ne parlerons dans ce mémoire que de cadres de discernement *discrets*, c'est-à-dire dont le nombre d'éléments est fini. Ω est alors constitué de N hypothèses H_i :

$$\Omega = \{H_1, H_2, \ldots, H_N\}.$$

La seule contrainte sur Ω est qu'il doit être *exclusif*, c'est-à-dire constitué d'hypothèses mutuellement exclusives.

Si Ω est *exhaustif*, c'est-à-dire que les hypothèses sont exhaustives, le cadre de discernement est dit **fermé**. Dans ce cas, la solution X au problème est unique et est alors obligatoirement l'une des hypothèses H_i. Dans le cas contraire (Ω non exhaustif), il est dit **ouvert**.

Dans la théorie de l'évidence, le raisonnement porte sur l'**espace de définition** 2^Ω qui est l'ensemble des 2^N sous-ensembles A de Ω :

$$2^\Omega = \{A/A \subseteq \Omega\} = \{\emptyset, \{H_1\}, \{H_2\}, \ldots, \{H_N\}, \{H_1, H_2\}, \ldots, \Omega\}.$$

Supposons que $A = \{H_1, H_2, H_3\}$. La proposition $X \in A$ signifie que la valeur réelle de X correspond à l'une des trois hypothèses sans distinction, c'est-à-dire que la proposition logique $H_1 \cup H_2 \cup H_3$ est vraie. Par abus de notation, on notera indifféremment la proposition A par $A = \{H_1, H_2, H_3\}$ ou par $A = H_1 \cup H_2 \cup H_3$. De même, nous écrirons indifféremment $A \in 2^\Omega$ ou $A \subseteq \Omega$.

Parmi les éléments de l'espace de définition, nous distinguons :
- l'ensemble vide $\emptyset$;
- le cadre de discernement Ω ;
- les **propositions** : sous-ensembles qui contiennent une ou plusieurs hypothèses, parmi lesquelles nous distinguerons les **singletons** (sous-ensembles qui contiennent une seule hypothèse) et les **paires** (sous-ensembles qui contiennent deux hypothèses).

Pour l'exemple du dé, nous pouvons choisir un cadre de discernement fermé :

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

L'espace de définition est alors :

$$2^\Omega = \{\emptyset, \{1\}, \{2\}, \ldots, \{6\}, \{1, 2\}, \ldots, \Omega\}.$$

Dans le cas général, plusieurs sources ou capteurs sont utilisés et donnent accès à des observations (mesures) qui formeront la base de la reconnaissance. La théorie de l'évidence est basée sur l'utilisation de structures appelées **distributions de masses d'évidence élémentaires** ou *bba (basic belief assignment)*. Pour obtenir des distributions de masses à partir des observations, il est nécessaire de définir des modèles de conversion. Les modèles peuvent être de différentes formes et nous détaillerons ceux que nous avons utilisés ultérieurement (cf. partie 6.3.2.2). Nous allons commencer par décrire les caractéristiques d'une distribution de masses.

6.2.2 Distribution de masses

Pour exprimer un degré de confiance envers chaque proposition A de 2^Ω, il est possible de lui associer une masse d'évidence élémentaire $m^\Omega(A)$ qui indique toute la confiance que l'on peut avoir dans cette proposition sans pour autant privilégier aucune des hypothèses qui la composent. Un ensemble de masses définies pour tous les éléments de l'espace de définition 2^Ω s'appelle une **distribution de masses** (ou jeu de masses). L'obtention d'une distribution de masses se fait grâce à une fonction m^Ω, définie par :

$$\begin{array}{rcl} m^\Omega : 2^\Omega & \longrightarrow & [0;1] \\ A & \longmapsto & m^\Omega(A), \end{array}$$

avec les propriétés suivantes :

$$\begin{aligned} m^{\Omega}(\emptyset) &= 0, \\ \sum_{A \subseteq \Omega} m^{\Omega}(A) &= 1. \end{aligned}$$

Les éléments de 2^{Ω} ayant une masse d'évidence non nulle sont appelés **éléments focaux**. Par abus de notation et de langage, on parle indifféremment de la fonction m^{Ω} et de la distribution de masses résultante. Un jeu de masses peut se décrire uniquement par les masses de ses éléments focaux, les autres étant nulles. $m^{\Omega}(A)$ représente la masse d'évidence de la proposition A. Quand la source est complètement incertaine, alors il est impossible de différencier aucune des hypothèses et :

$$m^{\Omega}(\Omega) = 1.$$

Si, par contre, la source est parfaite, c'est-à-dire qu'elle donne une information précise et sûre, alors si $X = \{H_i\}$:

$$m^{\Omega}(\{H_i\}) = 1.$$

La fonction m^{Ω} découle très souvent d'un modèle d'évidence défini selon le type de capteur / source utilisé(e) (type de mesure observée), selon le cadre de discernement défini et selon l'application visée. Souvent, selon une démarche simplificatrice et basée sur des heuristiques, les masses d'évidence sont attribuées à certaines combinaisons d'hypothèses. Mais d'autres approches existent aussi. Des modifications de modèles probabilistes ont été proposées, introduisant des masses d'évidence par affaiblissement [Shafer76], en prenant en compte l'information sur tout ce qui n'est pas une hypothèse unique (singleton) [Appriou93]. D'autres approches sont fondées sur des distances à des prototypes [Denœux95], s'inspirant de ce qui est réalisé en reconnaissance des formes. Dans de nombreuses applications, il est possible de disposer d'informations *a priori* qui permettent de déterminer de manière supervisée quels sont les éléments focaux à prendre en compte [Bloch96b, Tupin99, Milisavljevic03]. Des méthodes d'apprentissage des éléments focaux ont également été proposées [Mascle97, Bloch97]. Mascle, Bloch et Vidal-Madjar utilisent pour l'apprentissage les intersections entre les classes détectées par les différentes sources individuellement [Mascle97]. Bloch présente aussi un apprentissage basé sur des opérations de morphologie mathématique [Bloch97].

Dans l'exemple du dé, si l'on n'a aucune connaissance sur le résultat, on peut juste affirmer que le résultat se trouve dans Ω sans pouvoir préciser l'une des six faces. Une distribution de masses intuitive associée peut être définie par m^{Ω} :

$$\begin{aligned} m^{\Omega}(\Omega) &= 1, \\ \forall A \subset \Omega \; m^{\Omega}(A) &= 0. \end{aligned}$$

Si l'on dispose d'un premier capteur, la source S_1, qui ne peut voir que le centre de la face supérieure du dé, alors on peut savoir s'il y a ou non un point au centre de la face et par conséquent si la face est paire ou impaire. Une distribution de masses intuitive peut être définie par $m^{\Omega}_{S_1}$ telle que :

S'il y a un point :

$$\begin{aligned} m^{\Omega}_{S_1}(\{1,3,5\}) &= 1, \\ \forall A \subseteq \Omega\backslash\{1,3,5\}\ , m^{\Omega}_{S_1}(A) &= 0. \end{aligned}$$

S'il n'y a pas de point :

$$\begin{aligned} m^{\Omega}_{S_1}(\{2,4,6\}) &= 1, \\ \forall A \subseteq \Omega\backslash\{2,4,6\}\ m^{\Omega}_{S_1}(A) &= 0. \end{aligned}$$

Si l'on utilise un deuxième capteur, la source S_2, capable de voir l'un des bords de la face supérieure du dé, le nombre de points observés peut être 0, 1, 2 ou 3. D'où une distribution de masses intuitive notée $m^{\Omega}_{S_2}$:

- dans le cas 0, la seule masse non nulle est $m^{\Omega}_{S_2}(\{1\}) = 1$;
- dans le cas 1, la seule masse non nulle est $m^{\Omega}_{S_2}(\{2,3\}) = 1$;
- dans le cas 2, la seule masse non nulle est $m^{\Omega}_{S_2}(\{4,5,6\}) = 1$;
- dans le cas 3, la seule masse non nulle est $m^{\Omega}_{S_2}(\{6\}) = 1$.

Dans le cas général, plusieurs sources ou capteurs sont utilisés, donnant accès à plusieurs mesures (observations). Une fois les distributions de masses obtenues à partir de ces mesures, il est nécessaire de les fusionner avant de pouvoir prendre une décision. Pour cela, il existe différentes règles de combinaison.

6.2.3 Règles de combinaison et notion de conflit

Le principal intérêt de la théorie de l'évidence est le fait de pouvoir effectuer une fusion de données différente de celles que l'on peut effectuer avec les théories probabiliste et possibiliste. Il est donc très rare de ne disposer que d'une seule distribution de masses. Généralement, l'utilisation de plusieurs capteurs, ou sources, donne accès à plusieurs jeux de masses. Le but de cette partie est de présenter les façons de combiner ces distributions de masses afin d'obtenir un jeu de masses final qui tienne compte de toutes les informations disponibles et grâce auquel il sera possible de prendre une décision [Smets90a].

Les distributions de masses obtenues grâce à plusieurs capteurs peuvent être définies sur le même cadre de discernement ou non. Or, pour combiner deux distributions de masses, il est nécessaire qu'elles soient définies sur le même cadre de discernement. Dans le cas de distributions de masses qui ne seraient pas définies sur le même cadre de discernement, la solution est d'étendre les jeux de masses à un cadre de discernement commun qui correspond à l'**extension commune** des cadres de discernement s'ils sont différents mais compatibles, à leur **produit cartésien** sinon.

Par exemple, prenons deux distributions m^{Ω_1} et m^{Ω_2}, m^{Ω_1} étant définie sur le cadre de discernement Ω_1 et m^{Ω_2} sur le cadre de discernement Ω_2. La solution pour pouvoir combiner m^{Ω_1} et m^{Ω_2} est d'étendre ces distributions au cadre de discernement $\Omega_1 \times \Omega_2$ si Ω_1 et Ω_2 sont incompatibles. Le problème que cette solution implique est l'**explosion combinatoire** lorsque Ω_1 et Ω_2 sont des cadres de discernement comportant de nombreux éléments. En effet, l'espace de définition résultant de leur produit cartésien possède alors $2^{N_1 \times N_2}$ éléments. Par contre, si Ω_1 et Ω_2 sont compatibles (par exemple $\Omega_1 = \{H_1, H_2, H_3\}$ et $\Omega_2 = \{H_2, H_3, H_4\}$),

l'extension commune de ces cadres de discernement (ici $\Omega_3 = \{H_1, H_2, H_3, H_4\}$) amènera un espace de définition comportant moins d'éléments que celui obtenu avec le cadre de discernement résultant de leur produit cartésien $\Omega_1 \times \Omega_2$.

6.2.3.1 Somme orthogonale conjonctive

Une fois que l'on dispose de plusieurs distributions de masses définies sur le même cadre de discernement, il est possible d'en déduire un jeu de masses qui tienne compte de toutes les informations disponibles. Il peut être obtenu en utilisant une des nombreuses règles de combinaison qui sont toutes basées sur une opération appelée **somme orthogonale**.

Soient deux distributions de masses $m^{\Omega}_{S_1}$ et $m^{\Omega}_{S_2}$, obtenues par des sources S_1 et S_2, et définies sur le même cadre de discernement Ω. Si l'on note $m^{\Omega}_{S_1 \cap S_2}$ la distribution de masses résultant de la somme orthogonale conjonctive de $m^{\Omega}_{S_1}$ et $m^{\Omega}_{S_2}$, cette distribution est définie par :

$$\forall A \subseteq \Omega, m^{\Omega}_{S_1 \cap S_2}(A) = \sum_{B \subseteq \Omega, C \subseteq \Omega, B \cap C = A} m^{\Omega}_{S_1}(B).m^{\Omega}_{S_2}(C).$$

La somme orthogonale conjonctive a pour effet d'affecter des masses d'évidence à des propositions dont le nombre d'éléments est plus faible que celui des propositions initiales. En effet, A est un sous-ensemble de B et de C car $A = B \cap C$. Le jeu de masses final $m^{\Omega}_{S_1 \cup S_2}$ est donc plus précis que chacune des distributions de masses initiales.

La somme orthogonale impose comme condition nécessaire que les sources d'information soient indépendantes, ceci pour éviter un résultat biaisé. Cette notion d'indépendance reste dans la pratique une notion ambiguë et difficile à vérifier pour des données réelles. Quand on n'est pas assuré de l'indépendance, on se contente de s'assurer que les sources sont *distinctes*.

Pour l'exemple du dé, reprenons les deux sources S_1 et S_2. La source S_1 observe un point et la source S_2 deux. Les distributions intuitives $m^{\Omega}_{S_1}$ et $m^{\Omega}_{S_2}$ qui en découlent sont très simples :

$$\begin{aligned} m^{\Omega}_{S_1}(\{1,3,5\}) &= 1, \\ m^{\Omega}_{S_2}(\{4,5\}) &= 1. \end{aligned}$$

Bien évidemment, si nous appliquons le principe de la somme orthogonale conjonctive à ces deux distributions en les fusionnant, comme l'intersection entre les deux sous-ensembles $\{1,3,5\}$ et $\{4,5\}$ est le singleton $\{5\}$ alors la distribution de masses résultante $m^{\Omega}_{S_1 \cap S_2}$ est :

$$m^{\Omega}_{S_1 \cap S_2}(\{5\}) = 1.$$

Dans ce cas extrêmement simple, la seule masse d'évidence non nulle de la distribution de masses résultante donne la solution du jet de dé, ici 5. La face numérotée 5 comporte bien un point au centre de la face supérieure et deux sur chaque bord.

Si les sources S_1 et S_2 observent respectivement zéro et deux points, les jeux deviennent :

$$\begin{aligned} m^{\Omega}_{S_1}(\{2,4,6\}) &= 1, \\ m^{\Omega}_{S_2}(\{4,5,6\}) &= 1. \end{aligned}$$

Dans ce cas, après la somme orthogonale conjonctive, la distribution de masses résultantes est :

$$m^{\Omega}_{S_1 \cap S_2}(\{4,6\}) = 1,$$

et il n'est pas possible de trouver le résultat du jet de dé sans information / observation supplémentaire. Nous pouvons juste affirmer que le résultat est soit 4, soit 6. C'est l'illustration du doute entre deux hypothèses.

6.2.3.2 Notion de conflit

Les différentes règles de combinaison existantes diffèrent surtout par leur façon de gérer le conflit. Le **conflit** est, par définition, la masse d'évidence associée à l'ensemble vide $\emptyset$, notée, dans le cadre de discernement Ω, $m^{\Omega}(\emptyset)$. Lorsque cette masse d'évidence est non nulle, elle représente le fait que les sources ou les capteurs utilisés ont amené des observations contradictoires, d'où le terme de conflit. Ceci se comprend aisément en notant que le conflit, masse d'évidence de l'ensemble vide $\emptyset$, survient lors de la somme orthogonale conjonctive quand les éléments focaux (sous-ensembles ayant des masses non nulles) dans les distributions initiales ont une intersection vide.

Le conflit peut avoir trois origines [Lefèvre02] :
- des sources ou des capteurs qui fournissent des informations erronées ;
- des modèles qui représentent mal l'information ;
- des jeux de masses identiques combinés entre eux.

La première origine du conflit est souvent due à un défaut du capteur pendant l'étape d'acquisition ou à une mauvaise calibration pendant l'étape d'apprentissage. Si le comportement du capteur est satisfaisant, cette situation peut correspondre à un cadre de discernement non-exhaustif (une classe inconnue par exemple). La seconde origine dépend des modèles d'évidence. La plupart des modèles qui vont permettre de passer des mesures aux jeux de masses initiaux dérivent d'une information sur le voisinage selon deux types d'approche : un critère de distance [Denœux95] ou des fonctions de vraisemblance [Appriou91, Smets94]. Un choix inadéquat de métrique pour le premier type d'approche ou une mauvaise estimation des fonctions de vraisemblance pour celles du second type peut induire des variations dans les modèles d'évidence. Finalement, quand les sources à fusionner sont nombreuses, une masse de conflit peut être créée même si les sources sont concordantes. Par exemple, si nous considérons un ensemble de J sources d'information ayant le même jeu de masses suivant :

$$m^{\Omega}_j(H_1) = 0.8,\ m^{\Omega}_j(H_2) = 0.15 \text{ et } m^{\Omega}_j(\Omega) = 0.05,$$

nous pouvons remarquer que la masse d'évidence la plus importante supporte l'hypothèse H_1. Quand on combine les J jeux de masses, la masse de conflit est approximativement de 25% quand deux sources ($J = 2$) sont combinées et elle est proche de 80% pour dix sources ($J = 10$) ! [Lefèvre02].

La notion de conflit est très importante dans la théorie de l'évidence, elle permet de comprendre si les modèles d'évidence permettant d'obtenir les distributions de masses initiales à partir des mesures sont fiables et cohérents. Elle permet aussi de définir éventuellement la façon de gérer des mesures contradictoires. Nous reviendrons plus en détails sur cette notion dans la partie 6.3.4.

Nous allons maintenant présenter quelques-unes des règles de combinaisons conjonctives existantes, qui diffèrent par leur façon d'utiliser ou non le conflit. Certaines l'utilisent en

pondérant les masses non nulles par le complément à 1 du conflit et d'autres en transférant ou en réallouant le conflit sur l'espace de définition ou sur les sous-ensembles ayant créé le conflit. D'autres encore ne l'utilisent pas de façon explicite mais préfèrent garder cette information comme une indication sur la fiabilité des sources ou des capteurs utilisés. Dempster [Dempster68], Yager [Yager87], Smets [Smets90a], Itoh et Inagaki [Itoh97], Dubois et Prade [Dubois98] et, plus récemment, Murphy [Murphy00] et Lefèvre [Lefèvre02] proposent, entre autres, diverses règles de combinaison. Ne pouvant faire une description exhaustive de toutes les règles existantes, de leurs avantages et de leurs inconvénients, nous avons choisi de présenter celles de Dempster, de Yager, du TBM (Smets) et de Dubois et Prade. Nous laissons le soin au lecteur intéressé de se renseigner sur les autres règles de combinaison existantes.

Dans un souci de clarté concernant les notations, nous considérerons que, sans précision supplémentaire, les propositions A, B et C sont des sous-ensembles du même cadre de discernement Ω, ceci jusqu'à la fin de ce chapitre. De même, lorsque cela sera sans équivoque, nous ne préciserons plus le cadre de discernement Ω et l'espace de définition 2^Ω.

6.2.3.3 Règle de combinaison conjonctive de Dempster

Cette règle de combinaison nécessite une étape de normalisation afin de préserver les propriétés de base des distributions de masses. Dans [Zadeh86], Zadeh souligne que cette normalisation conduit à des comportements contraires à l'intuition. Soient deux distributions de masses m_1^Ω et m_2^Ω définies sur le même cadre de discernement 2^Ω. La distribution de masses qui résulte de leur combinaison par la règle de combinaison conjonctive de Dempster est souvent notée $m_{1\oplus 2}^\Omega$ et est définie par :

$$\begin{aligned} m_{1\oplus 2}^\Omega &= m_1^\Omega \oplus m_2^\Omega, \\ m_{1\oplus 2}^\Omega(\emptyset) &= \sum_{B\cap C=\emptyset} m_1^\Omega(B).m_2^\Omega(C), \\ \forall A \subseteq \Omega\backslash\emptyset,\ m_{1\oplus 2}^\Omega(A) &= \frac{\sum_{B\cap C=A} m_1^\Omega(B).m_2^\Omega(C)}{1-m_{1\oplus 2}^\Omega(\emptyset)}. \end{aligned}$$

Avec cette règle, le conflit $m_{1\oplus 2}^\Omega(\emptyset)$ sert à pondérer les masses d'évidence non nulles de la distribution de masses résultante (facteur $\frac{1}{1-m_{1\oplus 2}^\Omega(\emptyset)}$) [Dempster68].

6.2.3.4 Règle de combinaison de Yager

Soient deux distributions de masses m_1^Ω et m_2^Ω définies sur le même cadre de discernement Ω. Si l'on note $m_{1,2}^\Omega$ la distribution de masses qui résulte de leur combinaison par la règle de combinaison de Yager, elle est définie par :

$$\begin{aligned} m_{1,2}^\Omega(\emptyset) &= 0, \\ m_{1,2}^\Omega(\Omega) &= m_1^\Omega(\Omega).m_2^\Omega(\Omega) + \sum_{B\cap C=\emptyset} m_1^\Omega(B).m_2^\Omega(C), \\ \forall A \subseteq \Omega\backslash\{\emptyset,\Omega\},\ m_{1,2}^\Omega(A) &= \sum_{B\cap C=A} m_1^\Omega(B).m_2^\Omega(C). \end{aligned}$$

Dans cette règle, le conflit est réalloué au cadre de discernement Ω, ce qui signifie que lors de mesures contradictoires, le doute entre toutes les hypothèses voit sa masse d'évidence augmenter [Yager87, Yager88].

6.2.3.5 Règle de combinaison conjonctive du *TBM*

La règle de combinaison conjonctive utilisée par Smets dans son *TBM* est identique à la somme orthogonale conjonctive définie plus haut. La précision "du *TBM*" a été ajoutée pour éviter la confusion avec d'autres règles de combinaison conjonctives qui gèrent le conflit d'une certaine façon, comme par exemple la règle de combinaison conjonctive de Dempster, souvent citée à tort comme étant **la** somme orthogonale ou **la** règle de combinaison conjonctive de la théorie de l'évidence.

Soient deux distributions de masses m_1^Ω et m_2^Ω définies sur le même cadre de discernement Ω. La distribution de masses qui résulte de leur combinaison par la règle de combinaison conjonctive du *TBM* est notée $m_{1\textcircled{\cap}2}^\Omega$ et est définie par :

$$m_{1\textcircled{\cap}2}^\Omega = m_1^\Omega \textcircled{\cap} m_2^\Omega = m_{1\cap 2}^\Omega,$$
$$\forall A \subseteq \Omega, m_{1\textcircled{\cap}2}^\Omega(A) = \sum_{B\cap C=A} m_1^\Omega(B).m_2^\Omega(C).$$

Smets propose que la masse de conflit résulte de la non-exhaustivité du cadre de discernement. Avec cette règle, la masse d'évidence du conflit n'est pas utilisée pour "améliorer" la distribution de masses résultante. De plus, quand il faut réaliser la fusion de plusieurs distributions, cette règle de combinaison est associative et commutative, contrairement à d'autres, ne conservant pas et utilisant le conflit, ne sont ni l'un ni l'autre.

6.2.3.6 Règle de combinaison de Dubois et Prade

Soient deux distributions de masses m_1^Ω et m_2^Ω définies sur le même cadre de discernement Ω. Si l'on note $m_{1,2}^\Omega$ la distribution de masses qui résulte de leur combinaison par la règle de combinaison de Dubois et Prade, elle est définie par :

$$\begin{aligned} m_{1,2}^\Omega(\emptyset) &= 0, \\ \forall A \subseteq \Omega\backslash\emptyset, m_{1,2}^\Omega(A) &= \sum_{B\cap C=A} m_1^\Omega(B).m_2^\Omega(C) + \sum_{D\cap E=\emptyset, D\cup E=A} m_1^\Omega(D).m_2^\Omega(E). \end{aligned}$$

Dans cette règle, à chaque fois que deux éléments focaux ont une intersection nulle, le produit de leur masse (qui représente une partie du conflit) est allouée à l'union formée par ces deux éléments focaux. Le conflit est donc réalloué, par parties, à l'union des hypothèses qui le créent [Dubois98, Dubois99].

6.2.4 Grandeurs de décision

Une fois l'ensemble des jeux de masses fusionné en une seule distribution de masses résultante qui tient compte de toute les informations disponibles, il est maintenant temps de prendre une décision afin d'obtenir un résultat de reconnaissance / classification. Pour cela, on peut définir, à partir de cette distribution de masses résultante, des grandeurs de décision qui pourront chacune, par la suite, servir de fonction de décision.

6.2.4.1 Masse d'évidence

Lorsque la masse d'évidence est la grandeur de décision utilisée, la décision est prise directement sur la distribution de masses résultante, sans transformation ni calcul supplémentaire.

6.2.4.2 Crédibilité

La crédibilité, ou croyance, notée *bel* (*belief* : croyance), est une autre grandeur de décision couramment utilisée. La crédibilité caractérise toute la masse de croyance placée exactement sur A. La fonction de crédibilité est définie par :

$$\forall A \subseteq \Omega, bel(A) = \sum_{\emptyset \neq B \subseteq A} m^{\Omega}(B).$$

Il est à noter que l'on peut avoir $bel(A) = 1$ et $bel(B) = 1$ avec $A \neq B$ (par exemple, on a toujours $bel(\Omega) = 1$) et donc :

$$\sum_{A \subseteq \Omega} bel(A) \geq 1.$$

6.2.4.3 Plausibilité

La plausibilité sera notée *pl* (*plausibility* : plausibilité). Cette grandeur caractérise toute la force avec laquelle on ne doute pas de A. La fonction de plausibilité est définie par :

$$\forall A \subseteq \Omega, pl(A) = \sum_{A \cap B \neq \emptyset} m^{\Omega}(B).$$

La plausibilité peut aussi être définie par rapport à la fonction de crédibilité :

$$pl(A) = 1 - bel(\overline{A}),$$

où $\overline{A}$ est le complémentaire de A.

Il est possible de montrer que les deux grandeurs précédemment présentées, la crédibilité et la plausibilité, encadrent la probabilité inconnue $P(A)$ d'un évènement A, si celle-ci existe :

$$bel(A) \leq P(A) \leq pl(A).$$

6.2.4.4 Probabilité pignistique

Certains auteurs, comme Smets, préfèrent utiliser une fonction de probabilité pour ensuite appliquer les méthodes classiques de décision dans le cadre probabiliste [Smets90b].

La fonction pour construire cette probabilité, dite pignistique et notée $BetP$, est définie par :

$$BetP(A) = \sum_{B \subseteq \Omega} \frac{|A \cap B|}{|B|} m^{*\Omega}(B),$$

où $|A|$ est le cardinal (nombre d'éléments) de A et $m^{*\Omega}$ la distribution de masses normalisée par le complément à 1 de la masse du conflit :

$$\forall A \subseteq \Omega, m^{*\Omega}(A) = \frac{m^{\Omega}(A)}{1 - m^{\Omega}(\emptyset)}.$$

Contrairement au calcul de la crédibilité et de la plausibilité, cette opération n'est pas réversible, c'est-à-dire qu'il est impossible de retrouver le jeu de masses à partir de la probabilité pignistique. Ceci est dû au fait que cette opération repose sur l'hypothèse suivante : *une masse d'évidence affectée à une proposition peut être équitablement répartie entre les différentes hypothèses la composant.*

6.2.5 Prise de décision

Ces différentes fonctions de décision peuvent être utilisées pour prendre une décision et ainsi obtenir le résultat de la reconnaissance / classification après l'étape de fusion de données. La proposition qui présente un maximum pour la grandeur de décision choisie est généralement retenue comme résultat de reconnaissance. Plusieurs remarques sont néanmoins à faire concernant cette étape de prise de décision.

Tout d'abord, il faut bien noter que toute prise de décision (choix) implique un risque : le risque de se tromper. Comme les principaux avantages de la théorie de l'évidence sont, d'une part, la modélisation du doute et, d'autre part, la gestion de données contradictoires grâce au conflit, le risque est moindre qu'avec les autres approches théoriques (probabiliste ou possibiliste).

Ensuite, suivant la fonction de décision utilisée, il peut être nécessaire de réduire le nombre de propositions sur lesquelles la décision sera prise. En effet, si l'on utilise la masse d'évidence comme grandeur, le choix peut porter sur l'ensemble des propositions du cadre de discernement. Mais si l'on utilise la crédibilité ou la plausibilité, d'après les formules qui les définissent, la plus grande valeur sera forcément attribuée à la proposition Ω (sauf si le conflit est maximum dans la distribution de masse résultante, mais dans ce cas, il faut revoir les modèles d'évidence). Or Ω représente le doute total sur l'espace de définition. Pour obtenir une reconnaissance autre que le doute total, il est alors nécessaire de "forcer" le classifieur à ne considérer la grandeur de décision que sur un sous-ensemble de propositions. Le sous-ensemble de propositions généralement choisi est alors constitué des singletons et de l'ensemble vide.

Beaucoup de travaux existent pour décrire comment prendre la "meilleure" décision. En particulier, Denœux décrit dans [Denœux97] comment construire la prise de décision en fonction des distributions de masses.

6.3 Application : reconnaissance de postures statiques

Après cette présentation et ces rappels sur la théorie de l'évidence, nous allons maintenant présenter les différentes étapes de cette approche appliquée à la reconnaissance de postures statiques afin de réaliser une interprétation haut-niveau du comportement humain. Les quatre postures considérées sont les suivantes : "debout", "assis", "accroupi" et "couché".

Nous avons choisi de réaliser une reconnaissance de postures statiques selon la théorie de l'évidence pour les raisons suivantes :

- La théorie de l'évidence n'avait pas été appliquée à la reconnaissance de postures.
- Les avantages de cette théorie (la possibilité de modéliser l'incertitude associée aux données et de traiter les cas d'informations contradictoires) sont censés conduire à un taux d'erreurs relativement bas.

Pour cette étape de reconnaissance de postures statiques, trois hypothèses, spécifiques à cette étape de traitement, ont été rajoutées à celles de notre système, nous les rappelons ici :

5 Chaque personne doit être au moins une fois dans une **posture de référence**, debout avec les bras étendus horizontalement. Cette posture est celle effectuée par l'Homme de Vitruve dans le dessin de Léonard De Vinci présenté figure 6.1(c).

6 Chaque personne est supposée être filmée **entièrement**, c'est-à-dire qu'elle doit rester dans le champ de la caméra et ne pas être occultée par des objets fixes. Elle peut cependant être occultée partiellement ou complètement par une autre personne.

7 Chaque personne est supposée rester à une **distance à peu près constante** de la caméra.

6.3.1 Cadre de discernement et espace de définition

L'approche selon la théorie de l'évidence nécessite de définir un cadre de discernement Ω composé de N hypothèses exclusives H_i. Dans notre cas, ces hypothèses sont les quatre postures statiques considérées : "debout" (H_1), "assis" (H_2), "accroupi" (H_3) et "couché" (H_4). Ces hypothèses sont bien exclusives.

Si les hypothèses étaient exhaustives, c'est-à-dire que la vérité soit forcément dans Ω, alors Ω serait un cadre de discernement fermé. Avec nos choix d'hypothèses, en revanche, le cadre de discernement est ouvert car l'ensemble des postures du corps humain (même statiques) ne peut se réduire à cet ensemble relativement restreint. Pour se rapprocher du cas d'un cadre de discernement fermé, nous ajoutons une hypothèse pour l'ensemble des postures non reconnues. Au cadre de discernement Ω, nous ajoutons une **classe de rejet** notée H_0 qui permettra de favoriser cette hypothèse (posture inconnue) quand le conflit est maximal. Nous verrons plus tard, dans la partie 6.3.4.2, les avantages et les inconvénients de ce choix. Si nous ne pouvons reconnaître une posture unique (singleton) ou un doute entre ces différentes postures (proposition formée de plusieurs hypothèses), alors nous reconnaîtrons une posture inconnue. Par conséquent, nous avons $\Omega = \{H_1, H_2, H_3, H_4\}$ et H_0.

L'espace de définition est formé des 2^N sous-ensembles de Ω. Nous considérons donc ici $2^4 = 16$ propositions.

6.3.2 Mesures, modèles d'évidence et distributions de masses

6.3.2.1 Mesures

Pour réaliser la reconnaissance de postures statiques, nous avons besoin de mesures, d'observations. L'idée est d'utiliser un nombre réduit de distances normalisées pour caractériser la taille relative, et l'élongation / la compacité de la forme de la personne. Pour obtenir ces distances, nous utilisons les données bas-niveau suivantes :

- le centre de gravité ;
- la BAPS (Boîte par Axes Principaux issue de la Segmentation) (cf. figure 6.1(b,d)) ;
- la BERS (Boîte Englobante Rectangulaire de la personne issue de la Segmentation) ;
- la BERV (Boîte Englobante Rectangulaire du Visage).

Si l'on veut reconnaître les postures statiques de plusieurs personnes en tenant compte des occultations entre personnes, il faut utiliser les estimations des boîtes obtenues grâce au filtrage de Kalman lors de la deuxième étape du suivi temporel. La BERS et la BERV sont alors remplacées respectivement par la BEREP et la BEREV. Pour la BAPS et le centre de

gravité (qui est le centre de la BAPS), il est possible de définir une boîte par axes principaux estimée pour la personne (BAPEP) de la même manière que la BEREP (ajout d'une boîte dans le filtrage de Kalman) ou de calculer cette BAPS sur les pixels à l'intérieur de la BEREP.

Trois distances sont utilisées (cf. figure 6.1(b,d)) :

- D_1 distance verticale entre le centre de la BERV et le bas de la BERS ;
- D_2 distance entre le centre de la BERV et le centre de la BAPS (centre de gravité) ;
- D_3 longueur du demi grand axe de la BAPS.

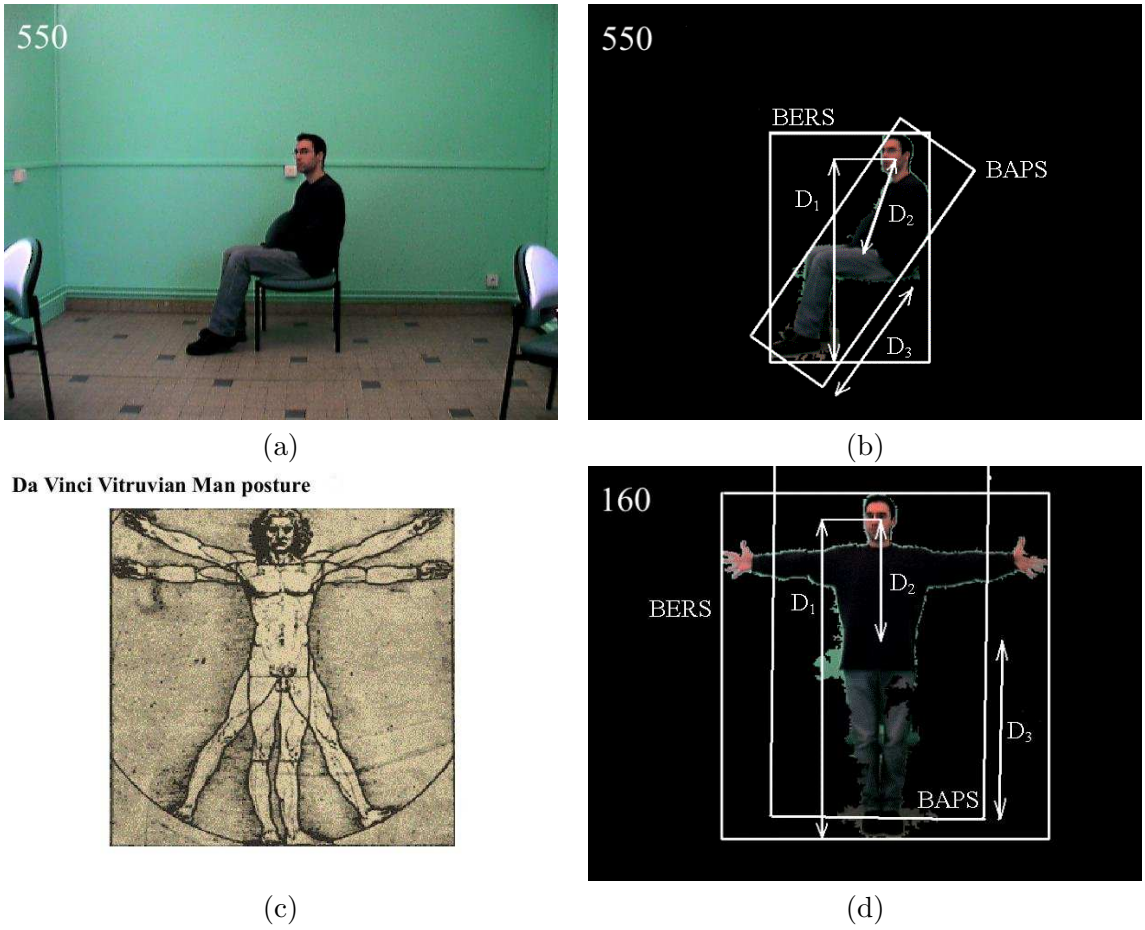


FIG. 6.1 – Exemples de distances D_i pour deux postures. (a,b) assis, (c,d) posture de référence.

Afin de normaliser les distances utilisées par rapport à leurs valeurs obtenues pour une posture connue, nous définissons une posture de référence (cf. figure 6.1(c)). L'utilisation de cette posture de référence a de nombreux avantages, comme par exemple réduire la variabilité des distances suivant la taille de la personne considérée, ou encore permettre de réinitialiser les localisations du visage et des mains, mais c'est également une contrainte.

Les distances D_i utilisées sont considérées comme indépendantes puisque issues d'étapes de traitement différentes. Cependant, il faut noter qu'une erreur de segmentation ou de localisation du visage peut avoir des répercussions sur plusieurs mesures.

La distance D_1 caractérise la **taille relative** de la silhouette de la personne (c'est-à-dire la distance entre son visage et le sol) alors que les distances D_2 et D_3 caractérisent plutôt

l'**élongation** ou la **compacité** de la silhouette de la personne. Pour la distance D_2, nous utilisons la localisation du visage car cette dernière est indépendante de la position et de l'orientation des bras.

Les trois distances D_i $(i = 1, 2, 3)$ sont normalisées par rapport à celles obtenues quand la personne est dans la posture de référence, distances notées D_i^{ref} $(i = 1, 2, 3)$. Ceci permet de prendre en compte les variations interindividuelles de tailles.

Les valeurs des distances normalisées forment les mesures utilisées, notées r_i.

Nous avons donc $r_i = \frac{D_i}{D_i^{ref}}$ $(i = 1, 2, 3)$.

La figure 6.2 illustre les variations temporelles des mesures r_i pour plusieurs personnes qui réalisent la même succession de postures : posture de référence, "assis", "debout", "accroupi", "debout", "couché", "debout", "assis", "debout" et "couché". En (a), nous avons la variation temporelle de r_1, en (b) celle de r_2 et en (c) celle de r_3. La liste des hypothèses successives H_i $(i = 1, \ldots, 4)$ correspondant à la succession de postures de la troisième personne (Vincent) est donnée en bas de la figure 6.2. H_1, H_2, H_3 et H_4 correspondent respectivement aux quatre postures "debout", "assis", "accroupi" et "couché". H_0 correspond aux postures qui surviennent pendant les étapes de transition qui sont considérées comme des postures inconnues.

Bien que le temps mis pour s'asseoir, s'accroupir, tomber, se relever, ou juste pour rester dans la même posture soit différent pour chaque personne, les variations temporelles montrent des allures et des niveaux typiques pour chaque posture.

6.3.2.2 Modèles d'évidence

À partir des mesures considérées r_i, il est nécessaire de réaliser une conversion entre les valeurs numériques des mesures pour obtenir les jeux de masses d'évidence élémentaires correspondant à cette mesure.

Cette conversion est effectuée grâce à des modèles d'évidence basés sur des sous-ensembles flous. Ces modèles d'évidence se présentent sous deux types (cf. figure 6.3). Le premier type de modèle (a) est utilisé pour r_1 et le second (b) pour r_2 et r_3. Grâce à ces modèles d'évidence, à partir des valeurs numériques de r_1, r_2 et r_3, trois distributions de masses d'évidence élémentaires sont créées, où les masses d'évidence sont réparties sur un nombre réduit de propositions appartenant au cadre de discernement Ω.

La mesure r_1 caractérise la taille relative de la silhouette de la personne par rapport à celle qu'elle a dans la posture de référence. Par rapport à cette mesure, le premier type de modèle est basé sur l'idée que plus le visage d'une personne est proche du sol, plus cette personne est proche de la posture "couché". À l'opposé, plus le visage d'une personne est éloigné du sol, plus la personne est proche de la posture "debout". Dans le modèle 6.3(a), selon la valeur numérique de r_1, soit une posture unique est reconnue, soit la combinaison d'une posture seule et de l'union de deux postures est reconnue. Dans ce dernier cas, l'union de deux postures modélise à la fois le doute entre les deux postures et la zone de transition entre ces deux postures. Par exemple (cf. figure 6.3(a)) :

Valeur de r_1	H_i reconnue(s)	somme des masse(s) d'évidence (non nulle(s))
$f < r_1$	H_1	$m_{r_1}(H_1) = 1$
$\frac{e+f}{2} < r_1 < f$	H_1, $H_1 \cup H_2$	$m_{r_1}(H_1) + m_{r_1}(H_1 \cup H_2) = 1$
$e < r_1 < \frac{e+f}{2}$	H_2, $H_1 \cup H_2$	$m_{r_1}(H_2) + m_{r_1}(H_1 \cup H_2) = 1$
etc.	etc.	etc.

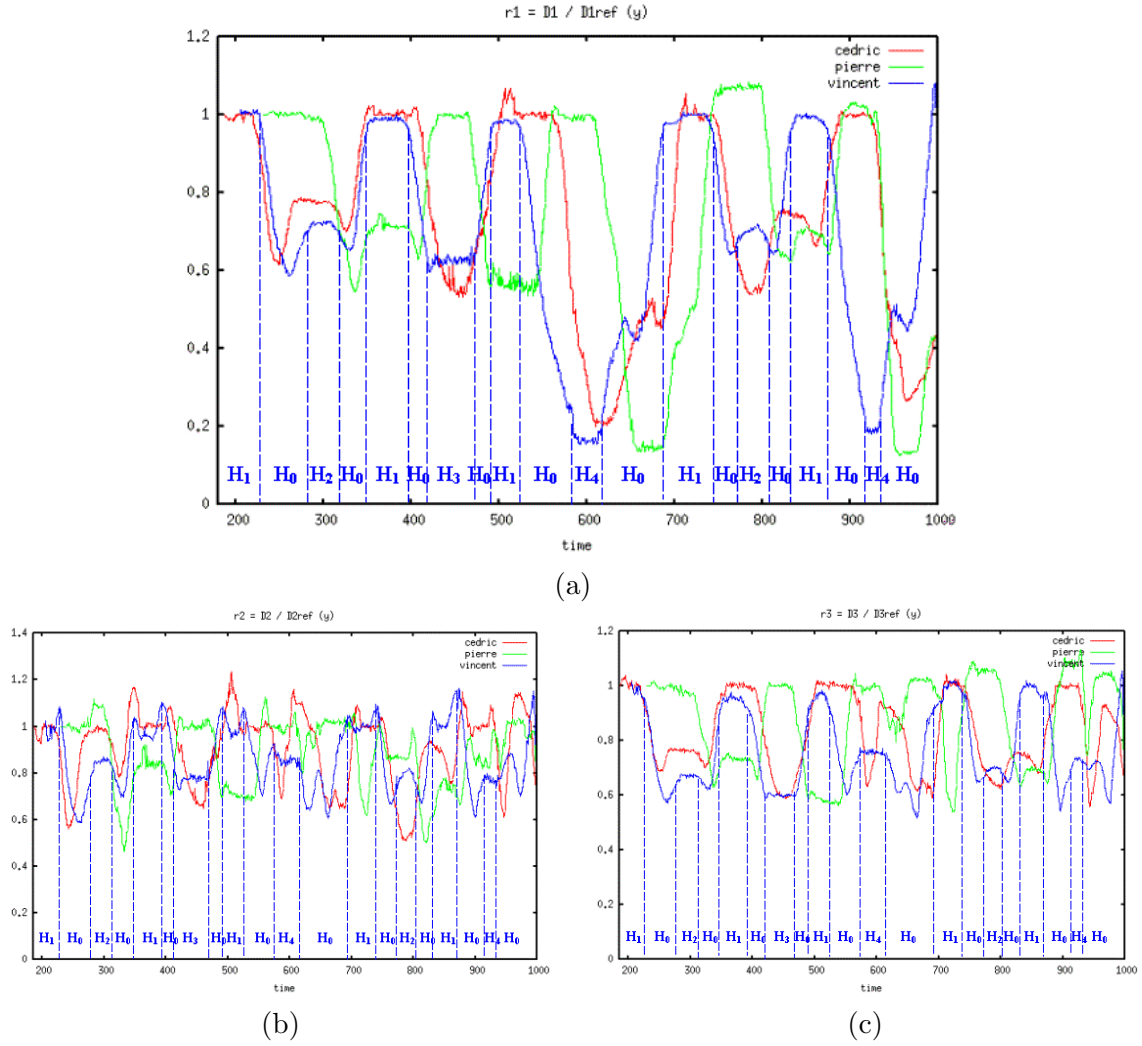


FIG. 6.2 – Variations temporelles de r_1, r_2 et r_3 pour trois personnes différentes.

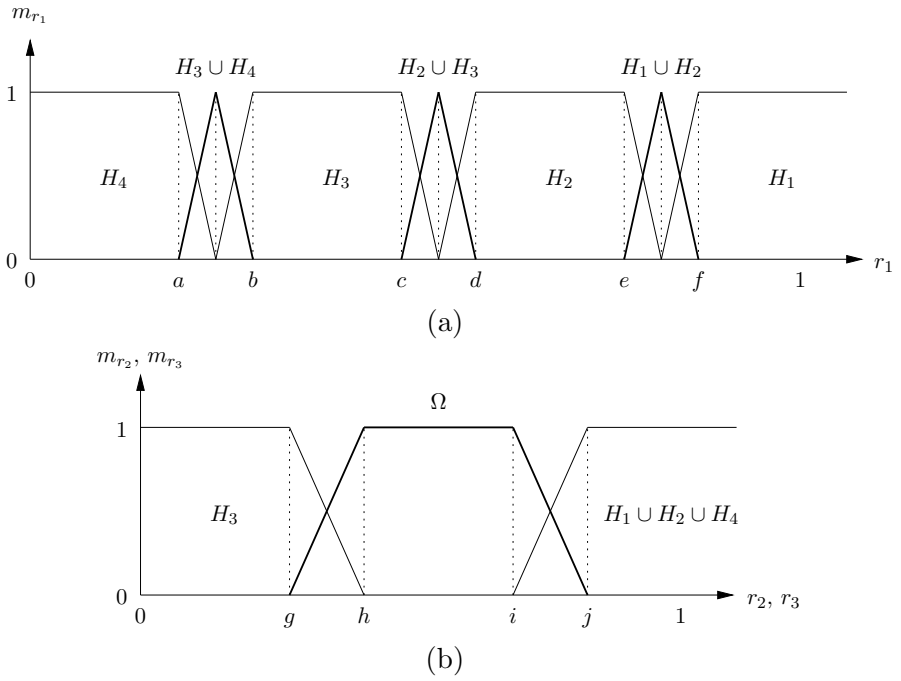


FIG. 6.3 – Modèles d'évidence. (a) Type pour r_1, (b) Type pour r_2 et r_3. Les H_i définissent les postures reconnues.

Les zones où la combinaison d'une posture seule et de l'union de deux postures est reconnue illustrent l'imprécision et l'incertitude des modèles. L'utilisation de formes "en triangle" pour la masse d'évidence attribuée à l'union de deux postures permet d'avoir toujours une posture plus favorisée que l'autre, sauf au niveau du pic du triangle en question. Par exemple, supposons que $\frac{c+d}{2} < r_1 < \frac{\frac{c+d}{2}+d}{2}$, alors une masse d'évidence assez forte sera placée sur $H_2 \cup H_3$, mais une masse d'évidence non nulle sera aussi attribuée à H_2, montrant ainsi que l'on doute entre les deux postures, mais qu'entre les deux, la préférence est donnée à H_2 plutôt qu'à H_3.

Les mesures r_2 et r_3 caractérisent l'élongation / la compacité de la silhouette de la personne. Par rapport à ces mesures, le second type de modèle est basé sur l'idée que la posture "accroupi" est une forme de corps humain très compacte, alors que les postures "assis", "debout" et "couché" sont des formes moins compactes et de plus en plus allongées. Dans le modèle 6.3(b), selon les valeurs numériques de r_2 ou de r_3, le modèle peut attribuer des masses d'évidence non nulles soit à la posture H_3 seule, soit à l'union de toute les postures (Ω correspond à $H_1 \cup H_2 \cup H_3 \cup H_4$), soit au sous-ensemble formé de l'union des postures "debout", "assis" et "couché" ($H_1 \cup H_2 \cup H_4$), ou encore à la combinaison de deux des sous-ensembles précédents.

Le principe d'attribution des masses d'évidence est le même que pour le modèle 6.3(a).

Suivant le type de modèle considéré, des seuils sont nécessaires pour définir et limiter les zones pour l'attribution des masses d'évidence. Ces seuils sont au nombre de 6 pour le modèle 6.3(a), et de 4 pour le modèle 6.3(b). Pour r_2 et r_3, les seuils ne sont pas les mêmes, même si ces mesures ont le même sens de variation. Par conséquent, il a été nécessaire de

définir 14 valeurs numériques fixant les seuils pour l'ensemble des modèles. Beaucoup de tests ont été réalisés pour trouver les 14 seuils les plus adaptés ($a_1 - f_1$ pour r_1, $g_2 - j_2$ pour r_2 et $g_3 - j_3$ pour r_3). Les seuils les plus adaptés sont ceux qui amènent le minimum de conflit sur les parties "statiques" des séquences vidéo.

Afin d'obtenir ces seuils de façon fiable, les histogrammes et les statistiques des quatre mesures r_i ont été calculées (minima min, maxima max, moyennes μ et écart types σ) sur un ensemble de séquences vidéo. Des détails sur les caractéristiques de ces séquences vidéo d'apprentissage sont donnés dans la partie 6.4.2. La figure 6.4 montre les histogrammes des valeurs numériques des mesures r_1, r_2 et r_3 pour l'ensemble des séquences vidéo d'apprentissage. Après cette étape de calcul, les seuils peuvent être choisis en comparant les mimima et les maxima, en calculant $\mu \pm 1, 2, 3\sigma$ ou en faisant un mélange de ces deux méthodes (par exemple, e_1 peut correspondre au maximum de r_1 pour la posture "assis", alors que d_1 est obtenu pour $\mu_{r_1}^{H_2} - 2\sigma_{r_1}^{H_2}$). Les meilleurs seuils (les plus adaptés) ont été obtenus pour un mélange des méthodes. Cette étape d'expertise a été réalisée par un opérateur humain. Les meilleurs seuils sont les suivants :

$a_1 = 0.41$		
$b_1 = 0.48$	$g_2 = 0.69$	$g_3 = 0.63$
$c_1 = 0.62$	$h_2 = 0.77$	$h_3 = 0.64$
$d_1 = 0.67$	$i_2 = 0.94$	$i_3 = 0.82$
$e_1 = 0.8$	$j_2 = 1.0$	$j_3 = 0.88$
$f_1 = 0.9$		

En fait, une des étapes les plus difficiles quand on utilise la théorie de l'évidence est de trouver des modèles d'évidence qui donnent des jeux de masses conduisant à un minimum de conflit lors de l'étape de fusion de données.

6.3.2.3 Distributions de masses

Une fois les seuils des modèles d'évidence choisis, chaque mesure r_i donne accès à une distribution de masses correspondante notée $m_{r_i}^{\Omega}$ ou plus simplement m_{r_i}. Chacune de ces distributions de masses exprime un certain degré de confiance dans chaque sous-ensemble $A \subseteq \Omega$ sans favoriser l'une des hypothèses qui le composent. Les distributions de masses ainsi obtenues m_{r_i} présentent bien les caractéristiques suivantes :

$$\begin{aligned} m_{r_i} : 2^{\Omega} &\longrightarrow [0;1] \\ A &\longmapsto m_{r_i}(A), \end{aligned}$$

avec les propriétés :

$$\begin{aligned} m_{r_i}(\emptyset) &= 0, \\ \sum_{A \subseteq \Omega} m_{r_i}(A) &= 1. \end{aligned}$$

Nous disposons donc de trois distributions de masses, notées m_{r_1}, m_{r_2} et m_{r_3}.

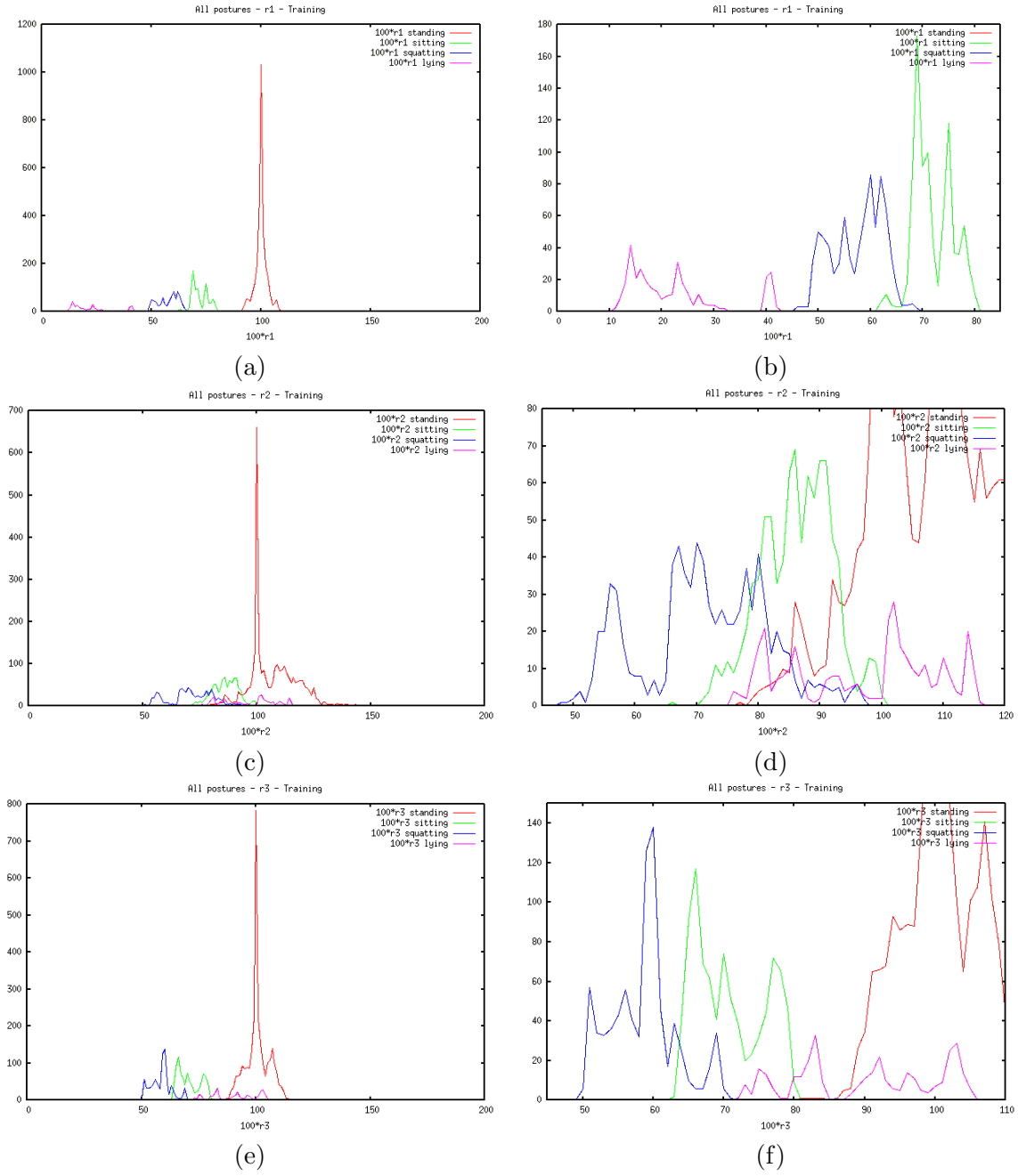


FIG. 6.4 – Histogrammes des valeurs numériques de r_1 (a), r_2 (c) et r_3 (e) pour l'ensemble des séquences vidéo d'apprentissage, (b), (d) et (f) zoom pour les postures "assis", "accroupi" et "couché".

6.3.3 Fusion de données

L'étape suivante consiste à fusionner les trois distributions de masses en une seule, suivant une règle de combinaison présentée plus haut. Le but est d'obtenir une distribution de masses résultante $m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}$ qui prend en compte toute l'information disponible.

Cette distribution de masses résultante est calculée en utilisant la règle de combinaison conjonctive du *TBM*.

La fusion de deux distributions de masses m_{r_i} et m_{r_j} en un jeu de masses $m_{r_i \textcircled{\cap} r_j}$ est définie, pour chaque sous-ensemble $A \subseteq \Omega$, de la manière suivante :

$$m_{r_i \textcircled{\cap} r_j} = m_{r_i} \textcircled{\cap} m_{r_j}, \tag{6.1}$$

$$m_{r_i \textcircled{\cap} r_j}(A) = \sum_{B \cap C = A} m_{r_i}(B).m_{r_j}(C). \tag{6.2}$$

Au cas où $m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(\emptyset) \neq 0$, $\emptyset$ représentant l'ensemble vide, il survient un **conflit**, ce qui signifie que les modèles d'évidence choisis conduisent à des résultats contradictoires. Cela arrive généralement quand quelques unes des mesures, parmi les r_i, sont dans les zones de transition des modèles. Pour rappel, nous avons choisi de préserver le conflit afin de reconnaître les postures inconnues (cf. parties 6.2.3.2 et 6.3.4.2).

La règle de combinaison du *TBM* utilisée est associative et commutative. Par conséquent l'ordre choisi pour fusionner les jeux de masses d'évidence n'a pas d'importance. Nous avons choisi de calculer $m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}$ dans l'ordre suivant :

$$\begin{aligned} m_{r_2 \textcircled{\cap} r_3} &= m_{r_2} \textcircled{\cap} m_{r_3}, \\ m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3} &= m_{r_1} \textcircled{\cap} m_{r_2 \textcircled{\cap} r_3}. \end{aligned}$$

6.3.4 Prise de décision

La décision est l'étape finale du processus. Une fois que toutes les distributions de masses d'évidence ont été fusionnées en une seule, il y a un choix à faire, une décision à prendre, parmi l'ensemble des hypothèses H_i et leurs combinaisons possibles. La décision est basée, bien sûr, sur la distribution de masses résultante, notée $m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}$. Une critère $Crit$ défini sur cette distribution de masses résultante est généralement maximisé pour choisir le résultat de reconnaissance $\hat{A} \subseteq \Omega$:

$$\hat{A} = \arg\max_{A \subseteq \Omega} Crit(A)$$

Avec les modèles d'évidence définis dans la partie 6.3.2.2, il est possible de déterminer le nombre d'hypothèses composant le résultat de reconnaissance. Le premier modèle d'évidence, utilisé pour la mesure r_1, peut amener, selon la valeur de r_1 :

- une seule hypothèse (singleton) ;
- une hypothèse (singleton) et le doute entre deux hypothèses (paire) ;
- le doute entre deux hypothèses (paire).

Le second modèle d'évidence, utilisé pour r_2 et r_3, peut amener, selon les valeurs de r_2 ou de r_3 :

- une seule hypothèse (singleton) ;

– une hypothèse (singleton) et le doute entre toutes les hypothèses (quadruplet) ;
– le doute entre toutes les hypothèses (quadruplet) ;
– le doute entre toutes les hypothèses (quadruplet) et le doute entre trois hypothèses (triplet) ;
– le doute entre trois hypothèses (triplet).

Le premier modèle d'évidence fixe le nombre d'éléments maximum du résultat de reconnaissance. En effet, la fusion de données étant réalisée avec la règle de combinaison conjonctive du *TBM*, les sous-ensembles ayant une masse d'évidence non nulle dans la distribution de masses résultante, qui ont été obtenus par intersection de deux sous-ensembles parmi les possibilités précédentes, auront au maximum deux éléments. C'est-à-dire que dans le pire des cas, le résultat de reconnaissance sera le doute entre deux hypothèses. Ceci est compatible par rapport aux postures statiques considérées : il est difficile d'imaginer qu'une personne puisse être à la fois "debout" ou "accroupi" ou "couché", par exemple. À l'inverse, on peut très bien s'imaginer douter entre "assis" ou "accroupi".

Après la fusion de données, les éléments focaux de la distribution de masses $m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}$, de masses d'évidence non nulles, peuvent donc être :
– l'ensemble vide $\emptyset$;
– des hypothèses (singletons) ;
– des doutes entre deux hypothèses (paires).

Le résultat de reconnaissance peut donc être un singleton, une paire d'hypothèses ou l'ensemble vide. Dans le premier cas, $\hat{A}$ étant un singleton, le résultat de reconnaissance est une posture unique. Dans le second cas, $\hat{A}$ étant une paire d'hypothèses, le résultat de reconnaissance est le doute entre deux postures. Dans le dernier cas, $\hat{A}$ étant l'ensemble vide, cela signifie que les mesures ont amené des distributions de masses contradictoires, et le résultat de reconnaissance est alors la posture inconnue (hypothèse H_0).

6.3.4.1 Fonctions de décision

Les trois grandeurs : masse d'évidence, crédibilité et plausibilité, notée respectivement m, *bel* et *pl*, définissent trois fonctions ou critères de décision possibles :

$$Crit_m(A) = m(A) = m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(A), \tag{6.3}$$

$$Crit_{bel}(A) = bel(A) = \sum_{\emptyset \neq B \subset A} m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(B), \tag{6.4}$$

$$Crit_{pl}(A) = pl(A) = \sum_{A \cap B \neq \emptyset} m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(B). \tag{6.5}$$

Comme expliqué précédemment dans la partie 6.2.5, deux choix sont possibles pour les sous-ensembles considérés lors de la prise de décision :

1. soit considérer l'ensemble des sous-ensembles de Ω, ce qui est possible quand on utilise comme grandeur de décision la masse d'évidence maximale, critère (6.3) ;
2. soit considérer uniquement l'ensemble formé par les singletons et l'ensemble vide, ce qui est nécessaire quand on utilise comme grandeur de décision la crédibilité, critère (6.4), ou la plausibilité, critère (6.5).

Pour rappel, la crédibilité et la plausibilité, respectivement notée *bel* et *pl*, amènent, par définition, des valeurs plus grandes pour les sous-ensembles de Ω avec de nombreux éléments. Par conséquent, il est nécessaire de ne calculer ces grandeurs de décision uniquement pour les singletons et pour l'ensemble vide afin de limiter le résultat de reconnaissance a une seule hypothèse H_i $(i = 0, \ldots, 4)$. Le classifieur correspondant est ainsi forcé de choisir entre une posture unique et la posture inconnue.

Le résultat de reconnaissance, quand on ne prend la décision que sur les singletons et sur l'ensemble vide, est alors $\hat{A}$ tel que :

$$\hat{A} = \arg \max_{A \subseteq \Omega, |A|<2} Crit(A),$$

où $|A|$ correspond au nombre d'éléments (cardinal) de A, A étant vu(e) comme un ensemble ou comme une proposition, par abus de notation et de langage.

Nous avons défini pour l'ensemble vide $\emptyset$, par rapport aux définitions originales, une crédibilité et une plausibilité données par :

$$\begin{aligned} bel(\emptyset) &= m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(\emptyset), \\ pl(\emptyset) &= m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(\emptyset). \end{aligned}$$

6.3.4.2 Gestion du conflit

Dans notre système, le conflit, masse d'évidence non nulle pour l'ensemble vide $\emptyset$, qui correspond à des données contradictoires, sert à déterminer si la posture que l'on cherche à reconnaître est proche ou non des postures "debout, "assis", "accroupi" ou "couché".

Si la masse d'évidence du conflit est faible, voire nulle, c'est que la posture réalisée est proche de l'une des postures statiques que l'on cherche à reconnaître. Si la masse d'évidence du conflit est forte, c'est que, d'une part, les mesures ont amené des valeurs contradictoires et, d'autre part, que la posture réalisée a peu de chances d'être l'une de celles que l'on cherche à reconnaître.

Nous associons donc le conflit à une classe de rejet, laquelle est définie par l'hypothèse H_0, et correspond à une posture inconnue.

Pour toutes les grandeurs de décision et les critères correspondants, quand la valeur considérée est maximale pour l'ensemble vide, c'est-à-dire le conflit, alors le résultat de reconnaissance est la posture inconnue (hypothèse H_0). Dans le cas où il y a égalité pour le critère choisi entre l'ensemble vide et un autre ensemble, nous adoptons un comportement prudent en reconnaissant une posture inconnue. L'avantage de cette approche, qui utilise le conflit pour une classe de rejet est que cela nous permet de détecter si la posture actuelle de la personne est une posture statique proche de celles que l'on cherche à reconnaître ou si c'est plutôt une posture de transition ou une posture véritablement inconnue.

L'inconvénient est que l'information donnée par le conflit n'est pas utilisée, comme par exemple dans la règle de combinaison conjonctive de Yager ou celle de Dubois et Prade. Par exemple, soient les deux distributions de masses suivantes, m_{r_1} et $m_{r_2 \textcircled{\cap} r_3}$:

$$\begin{aligned} m_{r_1} \quad &\text{telle que :} \quad m_{r_1}(H_2 \cup H_3) = 0.8 \text{ et } m_{r_1}(H_2) = 0.2, \\ m_{r_2 \textcircled{\cap} r_3} \quad &\text{telle que :} \quad m_{r_2 \textcircled{\cap} r_3}(H_1) = 1. \end{aligned}$$

La distribution de masses m_{r_1} indique que l'on doute fortement entre les postures "assis" et "accroupi" ($H_2 \cup H_3$), mais que l'on privilégie légèrement la posture "assis" (H_2). La distribution de masses $m_{r_2 \textcircled{\cap} r_3}$ indique que l'on est sûr que la posture est "debout" (H_1). Si l'on fusionne ces distributions de masses, alors la distribution de masses résultante est telle que $m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(\emptyset) = 1$. Les données sont contradictoires et l'ensemble vide a une masse d'évidence maximale (conflit maximal). Le conflit aurait pu être réalloué, par exemple à la proposition "assis" ou "accroupi" ou "debout", selon la règle de Dubois et Prade. Mais nous voulons reconnaître des postures statiques, et nous pensons que nos modèles d'évidence ont été conçus de telle façon que lorsque les données sont contradictoires (conflit maximal), c'est que la personne est dans une posture qui n'est ni "debout", ni "assis", ni "accroupi", ni "couché". Ce peut être une posture de transition ou réellement une posture statique inconnue. Nous préférons alors reconnaître une posture inconnue et douter au maximum entre deux postures que reconnaître souvent, en réallouant la masse du conflit, du doute entre trois ou quatre postures. Le résultat de reconnaissance pour notre exemple est alors l'ensemble vide $\emptyset$ donc la posture inconnue (hypothèse H_0). En effet, il est normal de considérer que si l'on doute entre les postures "assis" et "accroupi" d'une part et la posture "debout" d'autre part, la posture véritable a de fortes chances d'être une posture de transition ou une posture inconnue.

6.3.5 Exemple complet de reconnaissance

6.3.5.1 Distributions de masses élémentaires

Prenons par exemple les distributions de masses élémentaires suivantes, obtenues à partir des valeurs numériques des mesures r_i, grâce aux modèles d'évidence définis plus haut :

$$\begin{array}{cc} m_{r_1}(H_2) = 0.2 & m_{r_1}(H_2 \cup H_3) = 0.8, \\ m_{r_2}(H_3) = 0.9 & m_{r_2}(\Omega) = 0.1, \\ m_{r_3}(H_3) = 0.95 & m_{r_3}(\Omega) = 0.05. \end{array}$$

Concernant la mesure r_1, elle a conduit à une distribution de masses élémentaires m_{r_1} qui doute fortement entre les postures "assis" (H_2) et "accroupi" (H_3), mais favorise aussi légèrement la posture "assis" (H_2). Concernant les mesures r_2 et r_3, elles ont conduit à des jeux de masses élémentaires m_{r_2} et m_{r_3} qui privilégient fortement la posture "accroupi" (H_3), mais doutent aussi légèrement entre l'ensemble des postures ($\Omega = H_1 \cup H_2 \cup H_3 \cup H_4$).

6.3.5.2 Fusion des données

Réalisons maintenant la fusion des données, c'est-à-dire des distributions de masses. Pour fusionner deux distributions de masses selon la règle de combinaison conjonctive du *TBM*, il faut trouver les sous-ensembles de Ω qui correspondent aux intersections des sous-ensembles ayant une masse d'évidence non nulle (éléments focaux) dans chacune des distributions.

Fusionnons tout d'abord m_{r_2} et m_{r_3} en $m_{r_2 \textcircled{\cap} r_3}$. Voici la table d'intersection donnant les sous-ensembles résultants de cette fusion :

$m_{r_2} \backslash m_{r_3}$	H_3	$\Omega = H_1 \cup H_2 \cup H_3 \cup H_4$
H_3	H_3	H_3
$\Omega = H_1 \cup H_2 \cup H_3 \cup H_4$	H_3	Ω

La masse d'évidence d'un sous-ensemble résultant de la règle de combinaison conjonctive du *TBM* est la somme des produits des masses d'évidence des sous-ensembles ayant ce sous-ensemble comme intersection, par conséquent :

$$\begin{aligned} m_{r_2 \textcircled{\cap} r_3}(H_3) &= m_{r_2}(H_3) \times m_{r_3}(H_3) + m_{r_2}(H_3) \times m_{r_3}(\Omega) + m_{r_2}(\Omega) \times m_{r_3}(H_3), \\ m_{r_2 \textcircled{\cap} r_3}(\Omega) &= m_{r_2}(\Omega) \times m_{r_3}(\Omega), \end{aligned}$$

d'où le jeu de masses $m_{r_2 \textcircled{\cap} r_3}$:

$$\begin{aligned} m_{r_2 \textcircled{\cap} r_3}(H_3) &= 0.9 \times 0.95 + 0.9 \times 0.05 + 0.1 \times 0.95 = 0.995, \\ m_{r_2 \textcircled{\cap} r_3}(\Omega) &= 0.05 \times 0.1 = 0.005. \end{aligned}$$

Lors de cette fusion, il n'y a pas eu de conflit, puisqu'aucune intersection de sous-ensembles ayant une masse d'évidence non nulle n'a amené l'ensemble vide. La masse d'évidence a été renforcée sur la posture "accroupi" et le doute entre l'ensemble des postures a diminué.

Il ne reste qu'à fusionner m_{r_1} et $m_{r_2 \textcircled{\cap} r_3}$ en $m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}$. La table d'intersection est :

$m_{r_1} \backslash m_{r_2 \textcircled{\cap} r_3}$	H_3	$\Omega = H_1 \cup H_2 \cup H_3 \cup H_4$
H_2	$\emptyset$	H_2
$H_2 \cup H_3$	H_3	$H_2 \cup H_3$

De même que pour la première fusion, les masses d'évidence des sous-ensembles obtenus par intersection sont :

$$\begin{aligned} m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(\emptyset) &= m_{r_1}(H_2) \times m_{r_2 \textcircled{\cap} r_3}(H_3), \\ m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(H_2) &= m_{r_1}(H_2) \times m_{r_2 \textcircled{\cap} r_3}(\Omega), \\ m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(H_3) &= m_{r_1}(H_2 \cup H_3) \times m_{r_2 \textcircled{\cap} r_3}(H_3), \\ m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(H_2 \cup H_3) &= m_{r_1}(H_2 \cup H_3) \times m_{r_2 \textcircled{\cap} r_3}(\Omega), \end{aligned}$$

d'où la distribution de masses résultante $m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}$:

$$\begin{aligned} m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(\emptyset) &= 0.2 \times 0.995 = 0.199 = m(\emptyset), \\ m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(H_2) &= 0.2 \times 0.005 = 0.001 = m(H_2), \\ m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(H_3) &= 0.8 \times 0.995 = 0.796 = m(H_3), \\ m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}(H_2 \cup H_3) &= 0.8 \times 0.995 = 0.004 = m(H_2 \cup H_3). \end{aligned}$$

Cette fois, la fusion a amené du conflit, pour une masse d'évidence non nulle de 0.199 pour l'ensemble vide $\emptyset$. Ce conflit correspond bien à la contradiction des données avant la fusion entre les singletons représentant les postures "assis" et "accroupi" qui avaient chacun une masse d'évidence non nulle.

6.3.5.3 Prise de décision

Une fois obtenue cette distribution de masses résultante $m_{r_1 \textcircled{\cap} r_2 \textcircled{\cap} r_3}$, nous pouvons maintenant regarder les différents résultats de reconnaissance suivant la grandeur de décision utilisée et si l'on considère des hypothèses singletons, l'ensemble vide ou des hypothèses composites.

Si la grandeur de décision utilisée est la masse d'évidence maximale, critère (6.3), que l'on considère l'ensemble des sous-ensembles de Ω ou seulement les singletons et l'ensemble vide, le résultat de reconnaissance est la posture "accroupi" (H_3). En effet :

$$\begin{aligned} m(\emptyset) &= 0.199, \\ m(H_2) &= 0.001, \\ m(H_3) &= 0.796, \\ m(H_2 \cup H_3) &= 0.004. \end{aligned}$$

Si la grandeur de décision utilisée est la crédibilité, critère (6.4), après calcul sur les singletons et l'ensemble vide, nous obtenons une distribution de crédibilité identique :

$$\begin{aligned} bel(\emptyset) &= 0.199, \\ bel(H_2) &= 0.001, \\ bel(H_3) &= 0.796. \end{aligned}$$

En effet, par définition, la crédibilité, pour les singletons et avec notre définition de la crédibilité pour l'ensemble vide, est égale à la masse d'évidence. La crédibilité amène donc le même résultat de reconnaissance : posture "accroupi" (H_3). Si nous avions calculé la crédibilité de la proposition $H_2 \cup H_3$, elle serait de $bel(H_2 \cup H_3) = 0.004 + 0.001 + 0.796 = 0.801$. Ce calcul montre bien que la crédibilité est plus grande pour des sous-ensembles ayant de nombreux éléments.

Si la grandeur de décision utilisée est la plausibilité, critère (6.5), après calcul sur les singletons et l'ensemble vide, nous obtenons :

$$\begin{aligned} pl(\emptyset) &= 0.199, \\ pl(H_2) &= 0.001 + 0.004 = 0.005, \\ pl(H_3) &= 0.796 + 0.004 = 0.8, \end{aligned}$$

qui amène, une fois encore, le même résultat de reconnaissance : posture "accroupi" (H_3).

Cet exemple relativement simple a illustré le principe de la théorie de l'évidence appliqué avec nos modèles d'évidence à la reconnaissance de postures statiques.

6.4 Résultats

Nous allons maintenant présenter les résultats obtenus avec différents classifieurs. Afin de pouvoir comparer les résultats de ces classifieurs, nous définissons deux ensembles de séquences vidéo. Le premier ensemble de séquences vidéo est utilisé pour l'étape d'apprentissage, le deuxième ensemble pour l'étape de test. L'étape d'apprentissage consiste à calculer

les statistiques des mesures de distances normalisées (r_i) afin d'en déduire les seuils les plus adaptés pour nos modèles d'évidence sur des séquences vidéo où les postures sont réalisées de façon stéréotypées. L'étape de test consiste à observer les résultats de reconnaissance sur des séquences vidéo qui n'ont pas été utilisées lors de l'étape d'apprentissage. Elle permet de tester la robustesse et les performances du système de reconnaissance sur des séquences vidéo où les postures réalisées sont plus libres. L'ensemble total d'images traitées est d'environ 16000 images.

6.4.1 Classifieurs

Nous présentons les résultats de reconnaissance obtenus pour trois classifieurs différents. Les deux premiers, C_1 et C_2, sont basés sur nos modèles d'évidence. Le troisième, C_3, est un classifieur naïf afin de voir l'apport de la théorie de l'évidence au niveau de la reconnaissance.

6.4.1.1 Classifieurs C_1 et C_2 basés sur la théorie de l'évidence

Le résultat de reconnaissance pour C_1 est le sous-ensemble de Ω qui possède une **masse d'évidence maximale**.

Le résultat de reconnaissance pour C_2 est, parmi les singletons et l'ensemble vide, le sous-ensemble qui possède une **plausibilité maximale**.

Nous n'avons pas testé le classifieur utilisant le critère de crédibilité. Si l'on calcule ce critère pour les singletons et l'ensemble vide, cela revient à prendre le sous-ensemble de masse d'évidence maximale parmi les singletons et l'ensemble vide. Ceci amène des résultats moins bons que ceux du classifieur C_1 puisqu'il n'est alors pas possible d'obtenir le doute entre deux postures. De plus, les résultats obtenus sont aussi moins bons que ceux obtenus avec le classifieur C_2 puisque ce dernier réalloue en quelque sorte les masses d'évidence des paires d'hypothèses aux singletons correspondants et n'en ajoute pas à l'ensemble vide.

6.4.1.2 Classifieur naïf C_3

Le classifieur C_3 s'apparente à un détecteur majoritaire et les modèles naïfs qu'il nécessite ont surtout été définis pour caractériser l'apport de la théorie de l'évidence par rapport à une méthode naïve de fusion de données (comptage).

Les modèles naïfs découlent des modèles d'évidence. Ils se présentent donc, eux aussi, sous deux types (cf. figure 6.5) qui utilisent des ensembles nets. Le premier type de modèle est utilisé pour r_1 et le second pour r_2 et r_3.

Les seuils définissant les modèles naïfs sont directement issus des seuils des modèles d'évidence. L'utilisation de ces modèles naïfs est très simple. Selon la mesure considérée et, par conséquent, le type de modèle naïf utilisé, un décompte est réalisée pour les trois mesures. Pour chaque mesure r_i, le nombre d'occurrences de la (des) posture(s) reconnues est incrémenté. La posture qui possède le maximum d'occurrences est la posture reconnue. En cas d'égalité dans le décompte, la priorité est donnée aux postures dans l'ordre de définition des hypothèses de l'espace de définition : "debout", "assis", "accroupi" et "couché". Cela correspond, de façon très générale, à un *a priori* sur les occurrences des postures statiques. On considère que "debout" est la posture la plus fréquente, et par ordre décroissant d'occurrences : "assis", "accroupi" et "couché". Comme nous allons le voir, cet *a priori* amène des résultats de reconnaissance étonnamment bons pour ce classifieur, mais comme il ne permet

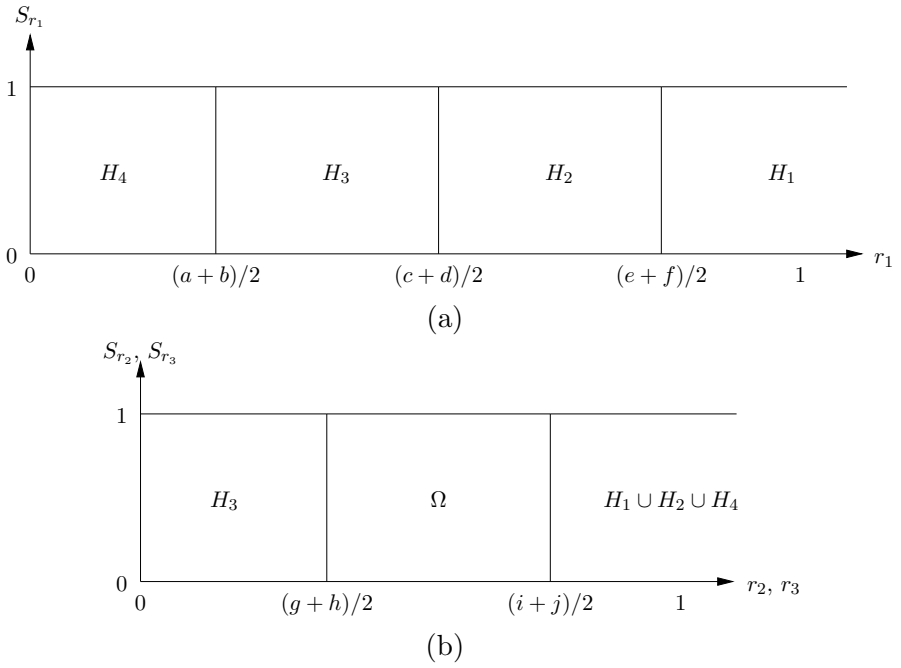


FIG. 6.5 – Modèles naïfs. (a) Type pour r_1, (b) Type pour r_2 et r_3. Les H_i définissent les postures reconnues.

pas la reconnaissance de la posture inconnue, il est assez peu comparable avec les autres classifieurs. Il faut cependant noter que même s'il aurait été possible de reconnaître la posture inconnue quand plusieurs postures sont à égalité dans le décompte, les résultats de reconnaissance obtenus auraient été très dégradés et n'auraient pas mérités d'être présentés dans ce mémoire.

6.4.2 Étape d'apprentissage

6.4.2.1 Ensemble d'apprentissage

L'ensemble d'apprentissage est constitué de 12 séquences vidéo qui représentent environ 5000 images. 6 personnes différentes sont filmées, 2 fois chacune, effectuant la même succession de postures statiques entrecoupées de postures de référence. La succession de postures est la suivante : posture de référence ("debout"), "assis", posture de référence ("debout"), "accroupi", posture de référence ("debout"), "couché", posture de référence ("debout"), "assis", posture de référence ("debout"), "couché". Les personnes sont de tailles différentes, entre 1,55 m et 1,95 m, ceci afin de prendre en compte la variabilité interindividuelle des tailles et d'améliorer la robustesse de l'algorithme. Les contraintes pour cet ensemble d'apprentissage sont triples :

- Réaliser les postures face à la caméra.
- Réaliser des postures naturelles, sans mouvement des bras par exemple.
- Réaliser les postures à la même distance de la caméra, dans un plan.

La première contrainte a pour raison de faciliter la détection et la localisation du visage, qui est nécessaire à l'obtention de deux mesures. La deuxième contrainte a pour raison d'avoir

des postures stéréotypées, et véritablement statiques. La troisième a pour raison d'éviter de trop grandes variations des mesures par rapport à la posture de référence. En effet, pour un grand changement de la distance entre la personne et la caméra, si la posture de référence n'est pas réalisée de nouveau, les variations des distances D_i $(i = 1, 2, 3)$ sont trop grandes par rapport aux distances D_i^{ref} de la posture de référence.

6.4.2.2 Taux de reconnaissance

Nous présentons les résultats obtenus avec les meilleurs jeux de seuils trouvés pour les modèles d'évidence. Les résultats sont calculés sur les images des séquences vidéo où les personnes sont considérées comme statiques. Une personne statique est une personne dont la majorité du corps (tronc et jambes) est immobile.

Les taux de reconnaissance de l'étape d'apprentissage pour les classifieurs C_1, C_2 et C_3 sont donnés respectivement dans les tables 6.1, 6.2 et 6.3. Ces tables représentent les matrices de confusion des classifieurs, les colonnes indiquant la posture réelle et les lignes la posture reconnue par le système. Les pourcentages représentent le taux de reconnaissance calculé sur l'ensemble des séquences vidéo d'apprentissage. La dernière ligne indique le taux d'erreur. Pour C_1, on considère que le doute entre deux postures est une reconnaissance correcte du moment que la véritable posture est comprise dans le doute.

TAB. 6.1 – Matrice de confusion du classifieur C_1 pour l'étape d'apprentissage.

Système\Réalité	H_1	H_2	H_3	H_4
H_0	0%	0.1%	0%	0%
H_1	**100%**	0%	0%	0%
$H_1 \cup H_2$	**0%**	**0%**	0%	0%
H_2	0%	**95.9%**	1.0%	0%
$H_2 \cup H_3$	0%	**2.1%**	**4.0%**	0%
H_3	0%	1.9%	**95.0%**	0%
$H_3 \cup H_4$	0%	0%	**0%**	**0%**
H_4	0%	0%	0%	**100%**
$\%err$	0%	2%	1%	0%

Comme les seuils des modèles d'évidence découlent des statistiques des mesures r_i calculées sur les séquences vidéo de l'ensemble d'apprentissage, les résultats sont excellents. Il n'y a que 0.1% de conflits sur plus de 5000 images de postures statiques. Il n'y a aucun problème pour reconnaître les postures "debout" et "couché". Les postures "assis" et "accroupi" sont aussi très bien reconnues même s'il existe parfois des doutes entre les deux. Le taux moyen d'erreurs de reconnaissance est de **0.8%**.

Dans le cas du classifieur C_2, les résultats sont aussi excellents. Il n'y a aucun problème pour reconnaître les postures "debout" et "couché". Le fait de calculer les plausibilités uniquement pour les singletons et pour l'ensemble vide force le classifieur à choisir entre H_2 et H_3 au lieu de choisir $H_2 \cup H_3$. Ceci amène une meilleure reconnaissance dans plus de la moitié des cas. Dans le reste des cas, il y a alors erreur sur la posture, mais dans le cas de l'étape d'apprentissage, ces erreurs représentent un faible pourcentage. Le taux moyen d'erreurs de

TAB. 6.2 – Matrice de confusion du classifieur C_2 pour l'étape d'apprentissage.

Système\Réalité	H_1	H_2	H_3	H_4
H_0	0%	0.1%	0%	0%
H_1	**100%**	0%	0%	0%
H_2	0%	**97.2%**	1.5%	0%
H_3	0%	2.7%	**98.5%**	0%
H_4	0%	0%	0%	**100%**
$\%err$	0%	2.8%	1.5%	0%

reconnaissance est d'environ **1.1%**.

TAB. 6.3 – Matrice de confusion du classifieur C_3 pour l'étape d'apprentissage.

Système\Réalité	H_1	H_2	H_3	H_4
H_1	**100%**	1.4%	0%	0%
H_2	0%	**97.6%**	14.9%	0%
H_3	0%	1.0%	**85.1%**	0%
H_4	0%	0%	0%	**100%**
$\%err$	0%	2.4%	14.9%	0%

Concernant le classifieur C_3, nous pouvons constater que les taux de reconnaissance sont légèrement moins bons que ceux de C_2 et qu'à part la posture "accroupi" (H_3), ce classifieur naïf fait peu d'erreurs de reconnaissance. Ceci est lié à l'*a priori* sur les occurrences des postures statiques (cf. partie 6.4.1.2), car les postures ont globalement cet ordre d'occurrences dans nos séquences vidéo. Il ne fait aucun doute qu'un autre choix aurait amené de moins bons résultats pour C_3. Même si le taux de reconnaissance pour la posture "assis" (H_2 *vs* H_2) est très légèrement meilleur que celui du classifeur C_2 (C_1 atteignant 98% pour cette valeur), le taux de reconnaissance pour la posture "accroupi" (H_3 *vs* H_3) est nettement moins bon. De plus, vues les contraintes des séquences vidéo de l'ensemble d'apprentissage, nous pouvons d'ores et déjà penser que les résultats lors de l'étape de test seront beaucoup plus dégradés. Le taux moyen d'erreurs de reconnaissance est d'environ **4.3%**.

Les taux moyens de reconnaissance pour les trois classifieurs sont les suivants : C_1 : **99.2%**, C_2 : **98.9%**, C_3 : **95.7%**.

6.4.3 Étape de test

6.4.3.1 Ensemble de test

Après des tests préliminaires dans des conditions et des postures très proches de celles de l'étape d'apprentissage qui ont donné de très bons résultats, nous avons voulu "modifier" les postures et nous mettre dans des conditions plus difficiles afin de tester la robustesse et les performances de notre algorithme de reconnaissance.

L'ensemble de test consiste donc en 12 autres séquences vidéo qui représentent environ 11000 images. 6 autres personnes sont filmées, 2 fois chacune, effectuant cette fois différentes successions de postures qui comprennent bien sûr les quatre que nous cherchons à reconnaître. Une seule posture de référence est effectuée en début de séquence vidéo. Les personnes choisies ne sont pas les mêmes que celles présentes dans l'ensemble d'apprentissage et sont aussi de tailles variées. Afin de tester les limites du système, les personnes sont autorisées à bouger les bras, à se mettre de profil par rapport à la caméra et à se déplacer dans la salle en restant quand même dans une gamme de distance raisonnable par rapport à la caméra. Les personnes peuvent s'asseoir de côté et même réaliser des postures qui ne surviennent pas souvent dans la vie de tous les jours, debout les deux bras levés par exemple, ou accroupi avec les bras levés.

Les contraintes pour cet ensemble de test sont donc largement réduites par rapport aux contraintes de l'ensemble d'apprentissage :

– Réaliser les postures de profil ou face à la caméra.
– Réaliser les postures à peu près à la même distance de la caméra.

Le fait de réduire les contraintes pour cet ensemble de test permet de tester véritablement la robustesse de l'algorithme dans des conditions beaucoup plus difficiles, même si relativement peu réalistes. Les résultats de reconnaissance peuvent être très affectés par des perturbations survenant à d'autres étapes de traitement. Par exemple, le fait de ne plus être contraint de réaliser les postures face à la caméra peut amener une mauvaise localisation du visage, ce qui aura pour effet de fausser deux mesures.

6.4.3.2 Taux de reconnaissance

Les taux de reconnaissance de l'étape de test pour les classifieurs C_1, C_2 et C_3 sont disponibles respectivement dans les tables 6.4, 6.5 et 6.6. Les conventions (jeux de seuils, images de personnes statiques et présentation des tables) sont les mêmes que celles de l'étape d'apprentissage et des tables 6.1, 6.2 et 6.3.

Tab. 6.4 – Matrice de confusion du classifieur C_1 pour l'étape de test.

Système\Réalité	H_1	H_2	H_3	H_4
H_0	0%	10.3%	5.0%	0%
H_1	**99.5%**	0.4%	0%	0%
$H_1 \cup H_2$	**0.5%**	**0%**	0%	0%
H_2	0%	**56.3%**	20.3%	0%
$H_2 \cup H_3$	0%	**27.1%**	**18.0%**	0%
H_3	0%	5.9%	**56.7%**	0%
$H_3 \cup H_4$	0%	0%	**0%**	**0%**
H_4	0%	0%	0%	**100%**
$\%err$	0%	16.6%	25.3%	0%

Pour le classifieur C_1, il y a, pour cette étape, plus d'erreurs de reconnaissance, mais les résultats montrent un bon taux de reconnaissance global. Il n'y a toujours aucun problème pour reconnaître les postures extrêmes "debout" et "couché". Pour les postures "assis" et

"accroupi", il y a plus d'erreurs, surtout quand les personnes ont les bras levés au-dessus de la tête, ou s'asseoient de côté. La raison en est que chacun n'a pas la même manière de s'asseoir et / ou de s'accroupir, les mains sur les genoux ou touchant le sol, le dos droit ou courbé etc. Ce fait amène plus de conflits, autour de 15%. Il y aussi plus de postures qui conduisent au doute $H_2 \cup H_3$. Néanmoins, les taux de reconnaissance sont très proches entre H_2 *vs* H_2 et H_3 *vs* H_3, ce qui tendrait à montrer que nos modèles d'évidence sont équilibrés. Le taux moyen d'erreurs de reconnaissance est d'environ **10.5%**.

TAB. 6.5 – Matrice de confusion du classifieur C_2 pour l'étape de test.

Système\Réalité	H_1	H_2	H_3	H_4
H_0	0%	10.2%	5.0%	0%
H_1	**99.9%**	0.4%	0%	0%
H_2	0.1%	**71.6%**	30.9%	0%
H_3	0%	17.8%	**64.1%**	0%
H_4	0%	0%	0%	**100%**
$\%err$	0.1%	28.4%	35.9%	0%

Pour le classifieur C_2, les résultats sont bons aussi, d'un côté les taux de reconnaissance H_i *vs* H_i sont meilleurs, mais d'un autre côté, il y a plus d'erreurs de reconnaissance entre H_2 et H_3. Nous pouvons remarquer que lorsque le classifieur doit choisir entre les postures "assis" (H_2) et "accroupi" (H_3), il a plus tendance à choisir la posture "assis" (H_2). Il confond très peu les postures "assis" et "accroupi" avec la posture "debout"(H_2 et H_3 avec H_1) mais amène plutôt la reconnaissance de la posture "inconnue" (H_0). Le taux moyen d'erreurs de reconnaissance est de **16.1%**.

TAB. 6.6 – Matrice de confusion du classifieur C_3 pour l'étape de test.

Système\Réalité	H_1	H_2	H_3	H_4
H_1	**100%**	7.4%	17.4%	0%
H_2	0%	**85.8%**	27%	0%
H_3	0%	6.7%	**55.6%**	0%
H_4	0%	0%	0%	**100%**
$\%err$	0%	14.1%	44.4%	0%

Concernant le classifieur C_3, les taux de reconnaissance sont mitigés. Les postures "debout" (H_1) et "couché" (H_4) ne posent aucun problème, et la posture "assis" (H_2) est bien reconnue. Néanmoins, il y a beaucoup d'erreurs sur la posture "accroupi". Le système confond même quelquefois la posture "accroupi" (H_3) avec la posture "debout" (H_1). Les deux inconvénients principaux du classifieur naïf est qu'il ne permet pas d'obtenir de doute entre postures, et qu'il ne permet pas non plus de reconnaître la posture inconnue. C'est pourquoi il est plus raisonnable de le comparer avec le classifieur C_2 qu'avec le classifieur C_1, qui lorsqu'il doute entre deux postures, possède très souvent la bonne parmi les deux. Au niveau des taux de reconnaissance moyens, le classifieur naïf C_3 est légèrement meilleur que

le classifieur C_2, mais la principale raison est le choix de favoriser les postures au niveau de leurs occurrences *a priori*. Comme nous l'avons dit précédemment, un autre *a priori* sur les occurrences des postures ou la reconnaissance de la posture inconnue en cas d'égalité dans le décompte pour chaque posture aurait nettement dégradé les résultats. Le taux moyen d'erreurs de reconnaissance est d'environ **14.7%**.

Les taux moyens de reconnaissance pour les trois classifieurs sont les suivants : C_1 : **89.5%**, C_2 : **83.9%**, C_3 : **85.3%**. Les meilleurs résultats sont donc obtenus avec le classifieur C_1.

Les figures 6.6, 6.7, 6.8 et 6.9, illustrent quelques résultats de reconnaissance de postures statiques. La BERS, la BAPS, la BERV et la distance D_2 sont dessinées en blanc sur l'image segmentée. Chaque figure présente six postures de chaque type, dans l'ordre suivant : "debout", "assis", "accroupi" et "couché".

FIG. 6.6 – Exemples de reconnaissance de postures statiques : "debout".

La figure 6.10 illustre une posture inconnue en (a) et un doute entre deux postures "assis" ou "accroupi" en (b). Pour la première, la personne est assise mais par terre, et pour la seconde, la personne est en fait accroupie mais avec le dos très droit.

6.5 Avantages, limitations et cadences de traitement

La méthode de reconnaissance de postures statiques, basée sur la théorie de l'évidence, a conduit à de bons taux de reconnaissance. L'approche utilisée est comparable à une méthode basée sur les formes, car nous considérons la taille relative, et l'élongation / la compacité de la silhouette de la personne. Néanmoins, aucune comparaison explicite avec une méthode basée sur les formes n'a été réalisée.

L'avantage de cette méthode est l'approche selon la théorie de l'évidence qui permet de modéliser le doute dans la reconnaissance, conduisant à la reconnaissance d'une posture ou du doute entre deux postures, avec une masse d'évidence plus grande pour l'une d'elles.

La limitation de cette méthode est qu'elle ajoute trois hypothèses de plus au système global, le fait que chaque personne soit au moins une fois dans une posture de référence, le

FIG. 6.7 – Exemples de reconnaissance de postures statiques : “assis”.

FIG. 6.8 – Exemples de reconnaissance de postures statiques : “accroupi”.

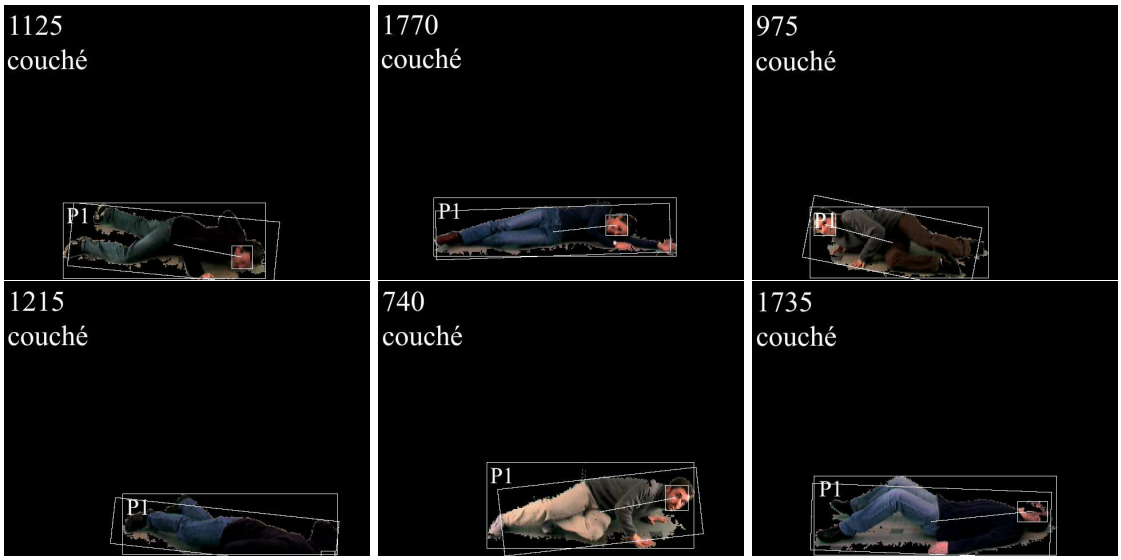
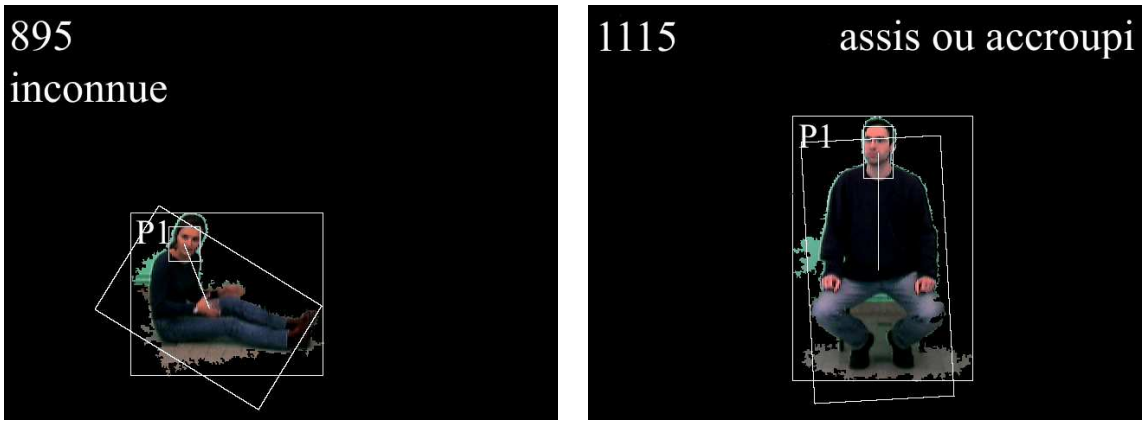


FIG. 6.9 – Exemples de reconnaissance de postures statiques : “couché”.



(a) (b)

FIG. 6.10 – Exemple de posture inconnue (a) et d’un doute entre deux postures (b).

fait qu'elle ne soit pas occultée par des objets fixes et le fait qu'elle reste à peu près à la même distance de la caméra.

Néanmoins, l'utilisation d'une posture de référence permet, par exemple, de corriger de façon simple l'étape de localisation du visage dans le cas d'une mauvaise initialisation ou d'un échec du suivi temporel.

Au niveau des pourcentages de temps de calcul et des cadences de traitement atteintes pour cette étape de traitement, la table 6.7 présente les résultats obtenus pour l'ensemble des étapes de traitement du système. Avec la segmentation basée sur les champs aléatoires de Markov, les pourcentages de temps de calcul des différentes étapes sont négligeables par rapport à celui de la segmentation elle-même. Les cadences de traitement reflètent le fait que la complexité de l'algorithme de segmentation limite les performances du système entier. Avec la segmentation optimisée en vitesse, au contraire, le système atteint des cadences de traitement relativement élevées. Les étapes les plus coûteuses en temps de calcul sont d'abord la seconde étape du suivi temporel, puis la localisation et le suivi temporel du visage et des mains et enfin la segmentation. Nous pouvons constater que les étapes les plus rapides sont la première étape du suivi temporel et la reconnaissance de postures. La cadence de traitement en résolution 640×480 n'est pas très éloignée de la cadence vidéo et en résolution d'image 320×240 permet de faire du traitement à la cadence vidéo. L'objectif du système concernant une cadence de traitement proche de la cadence vidéo est donc satisfaite.

Tab. 6.7 – Pourcentages de temps de calcul et cadences de traitement pour la reconnaissance de postures.

Segmentation	Champs aléatoires de Markov		Optimisée en vitesse	
Résolution d'image	320×240	640×480	320×240	640×480
Acquisition	0.2%	0.3%	1.1%	2.8%
Segmentation	80.1%	86.5%	17.8%	21.8%
Suivi temporel (1/2)	8.9%	1.2%	2.1%	0.6%
Localisation et suivi du visage et des mains	8.1%	2.1%	34.4%	26.3%
Suivi temporel (2/2)	1.6%	7.9%	39.7%	41.3%
Reconnaissance de postures	1.1%	2%	4.9%	7.2%
Cadences de traitement	**7.56** images/s	**1.8** images/s	**60** images/s	**16** images/s

6.6 Conclusion

Ce chapitre a présenté la dernière étape de traitement de notre système d'analyse et d'interprétation du mouvement humain. Cette étape utilise les données bas-niveau extraites lors des étapes de traitement précédentes pour réaliser une interprétation haut-niveau du comportement humain. Cette interprétation consiste à reconnaître les postures statiques parmi les quatre postures suivantes : "debout", "assis", "accroupi" et "couché". Elle est effectuée grâce à une fusion de données basée sur la théorie de l'évidence. Le système développé donne

de bons résultats de reconnaissance et est plus performant qu'un classifieur naïf qui ne peut déterminer de postures inconnues ou donner un doute entre deux ou plusieurs postures.

Nous allons maintenant donner quelques perspectives concernant les travaux présentés dans ce chapitre.

Les principales limitations de l'étape de reconnaissance de postures statiques sont qu'une personne doit refaire la posture de référence si la distance à la caméra change de façon significative et qu'elle ne peut être occultée par des objets fixes. Une solution pour lever ces hypothèses serait d'utiliser une caméra stéréo, d'une part pour estimer la profondeur, et utiliser cette information pour normaliser les distances calculées par rapport à une estimation de la taille de la personne en fonction de la distance à la caméra et, d'autre part, pour estimer la position de la personne dans l'espace 3D. La personne peut alors être occultée, du moment qu'une estimation de sa taille est disponible.

Une autre perspective est la reconnaissance d'actions pendant les transitions entre deux postures statiques. Pendant ces phases de transition, la reconnaissance amène soit du doute entre deux postures, soit la posture inconnue. Le plus souvent, c'est la posture inconnue qui est reconnue. Nous projetons d'améliorer notre méthode en ajoutant une analyse dynamique des variations temporelles des mesures. Cela devrait améliorer les taux de reconnaissance. Pour justifier cette affirmation, un point intéressant peut être vu sur la variation temporelle de la mesure r_1 (cf. figure 6.2, page 164). Quand une personne s'assied, la variation temporelle de la mesure r_1 suit un schéma caractéristique : elle décroît avant de croître de nouveau parce que la personne se penche en avant et se redresse au lieu de s'asseoir directement sans se pencher (un schéma similaire, mais opposé, se produit quand une personne assise se lève). C'est un point important pour une analyse dynamique qui pourrait conduire à la reconnaissance d'actions comme se lever, se coucher, s'asseoir, tomber, s'accroupir etc.

Conclusion et perspectives

Conclusion

Nous avons présenté dans ce mémoire de thèse un système temps-réel, dont la cadence est peu éloignée de la cadence vidéo, permettant de réaliser l'analyse et l'interprétation du mouvement humain pour une ou plusieurs personnes dans des séquences vidéo. Dans ce système, des données bas-niveau sont extraites au cours de diverses étapes de traitement. La segmentation 2D spatio-temporelle de personnes réalise l'extraction des objets en mouvement par rapport au fond de la scène. Parmi les deux méthodes disponibles dans notre système, l'une d'entre elles, basée sur les champs aléatoires de Markov, a été présentée. Les deux segmentations permettent l'obtention de bons résultats par rapport aux caractéristiques des objets vidéo obtenus. La segmentation est une étape de traitement importante car les résultats des étapes de traitement ultérieures dépendent de la qualité des résultats de cette étape. Après la segmentation, la première étape de suivi temporel est très rapide mais ne gère pas les phénomènes d'occultation. Puis le processus de détection de peau, première étape de la localisation du visage et des mains, fournit de très bons résultats, grâce à la segmentation et à l'adaptation automatique des seuils de détection, même sur des fonds relativement complexes ou dont la couleur est proche des couleurs de peau. Par conséquent, la localisation du visage et des mains est généralement précise et fiable. Dans la seconde étape du suivi temporel, l'utilisation d'un filtrage de Kalman partiel et d'une poursuite du visage permet de gérer les réunions et les séparations temporelles de personnes segmentées et par conséquent les phénomènes d'occultation entre personnes. En utilisant une partie des données bas-niveau extraites pendant ces étapes de traitement (segmentation, suivi temporel et localisation du visage et des mains), il est possible de réaliser une interprétation haut-niveau du comportement humain. Nous avons présenté une méthode basée sur la théorie de l'évidence pour reconnaître les postures statiques de personnes. Quatre postures peuvent être reconnues ("debout", "assis", "accroupi" et "couché") grâce à un faible nombre de mesures de distances normalisées. Cette méthode a donné de bons résultats de classification et est assez rapide pour être intégrée dans un système temps-réel.

Il y a deux types d'applications considérées pour ce système. Il peut être utilisé pour des applications de réalité mixte avec des interfaces homme-machine avancées. En face d'une unique caméra statique, dans un environnement intérieur, une ou plusieurs personnes peuvent interagir avec un environnement virtuel et / ou ses objets par l'intermédiaire de leurs mouvements. Le système proposé pour mélanger les mondes réel et virtuel par traitement d'image sans systèmes invasifs comme les marqueurs etc. atteint des performances respectables avec une bonne précision. Il est assez rapide pour un système interactif incluant un échange d'information homme-machine et est relativement facile d'utilisation. L'autre application possible est la vidéosurveillance de personnes âgées à la maison ou dans un milieu hospitalier. À la

condition d'être inclus dans un système autonome, afin de respecter une éthique par rapport à la vie privée des personnes filmées, ce système pourrait par exemple déclencher un signal visuel ou sonore s'il détecte une personne restée trop longtemps assise ou si elle a fait une chute.

Comparé à d'autres systèmes comme *Pfinder*, W^4 et le système issu du projet *DARPA VSAM* (respectivement [Wren97b], [Haritaoglu98] et [Collins00]) notre système propose des approches relativement différentes pour traiter les différentes étapes de traitement et leurs difficultés inhérentes, par exemple la détection et le suivi temporel d'une personne seule ou d'un groupe de personnes, ou encore des parties de leurs corps (visages et mains), ou même de l'interprétation de leurs comportements. Notre système présente certains avantages par rapport à ces systèmes.

- Par rapport au système *Pfinder*, notre système a l'avantage principal de pouvoir analyser et interpréter le mouvement de plusieurs personnes. De plus, par rapport à la localisation et le suivi du visage et des mains, nous disposons des dernières positions connues contrairement au système *Pfinder* qui efface les *blobs* correspondants.
- Par rapport au système W^4, notre système utilise les informations de couleur, notamment pour la détection du visage et des mains. De plus, les modèles utilisés dans W^4 restreignent l'interprétation du comportement à des positions verticales, ce qui n'est pas le cas de notre système.
- Par rapport au système issu du projet *DARPA VSAM*, notre système présente l'avantage d'analyser plus en détails le corps humain, ce qui n'est pas possible dans le système de Collins *et al.* car les personnes ne sont pas les objets prépondérants dans les séquences. Les vues des caméras sont en effet trop lointaines.

Ces systèmes possèdent sans aucun doute les capacités de réaliser des tâches similaires, mais avec des approches et des méthodes différentes. Il serait intéressant néanmoins de mettre en œuvre certaines de ces méthodes, par exemple l'approche par *blob* utilisée dans *Pfinder*, afin de comparer les performances et la robustesse des systèmes.

Perspectives

De nombreuses perspectives concernant chaque étape de traitement ont été présentées dans les chapitres correspondants. Elles concernent généralement les limitations de ces étapes et décrivent des solutions ou des approches qui pourraient améliorer les résultats relatifs à chaque étape. Nous ne les rappellerons donc pas ici. Nous allons présenter quelques perspectives concernant le système dans son ensemble.

Les premières perspectives concernent les hypothèses de notre système et les solutions possibles pour éviter d'avoir à les supposer. Les deux premières hypothèses sont le fait que l'environnement est filmé par une **caméra fixe** (hypothèse n°1) et que chaque personne **entre seule** dans la scène (hypothèse n°2). Ces deux hypothèses ne sont pas particulièrement contraignantes. Notons qu'avec une méthode de compensation du mouvement de la caméra, il serait possible de s'affranchir de l'hypothèse n°1. Les hypothèses n°3 et n°4, environnement **intérieur** et séquence vidéo qui commence par une **scène vide**, sont optionnelles et ont été ajoutées pour faciliter l'étape de segmentation 2D spatio-temporelle. En ce qui concerne l'extension de notre système à un environnement extérieur, nous pensons que c'est une tâche possible, même si certains algorithmes auraient besoin d'être revus et améliorés, par rapport aux difficultés pouvant survenir quand on considère les variations des conditions d'acquisition

en environnement extérieur par rapport à celles, plutôt bien contrôlées, en environnement intérieur. Du moment que les personnes sont en nombre restreint et restent les principaux objets mobiles dans la scène, les résultats devraient être relativement fiables. Les trois dernières hypothèses de notre système sont liées à l'étape de reconnaissance de postures statiques. Dans le chapitre 6, nous avons proposé la solution d'utiliser une caméra stéréoscopique afin de lever ces hypothèses. Le fait de devoir réaliser une **posture de référence** (hypothèse n°5), d'être filmé **entièrement** (hypothèse n°6) et de rester à une distance **à peu près constante** de la caméra (hypothèse n°7) pourraient être évités. En effet, grâce à une méthode stéréoscopique, il est possible d'obtenir une estimation de la distance d'un objet dans la scène à la caméra. Il est aussi possible de déterminer la position des pieds d'une personne par exemple avec une calibration de la scène. Avec cette distance, ou profondeur, et la position des pieds d'une personne, nous pourrions normaliser les distances utilisées pour la reconnaissance de posture par une estimation de la taille de la personne (la distance des pieds au visage par exemple) et ainsi s'affranchir des hypothèses n°5, n°6 et n°7, dans l'ordre croissant de contrainte.

Maintenant, nous allons présenter les perspectives qui découlent des travaux présentés dans ce mémoire. Tout d'abord, afin de poursuivre ces travaux, une perspective directe est la reconnaissance d'actions pendant les transitions entre deux postures statiques. Dans [Mokhber05], huit classes d'actions peuvent être reconnues : s'accroupir, se relever de la position accroupie, s'asseoir, se relever de la position assise, marcher, se pencher, se relever de la position penchée, sauter. Ces actions représentent une activité dynamique et il serait possible avec notre système de reconnaître la majorité de ces actions grâce à l'étude, par exemple, des variations temporelles des mesures de distances ou des transitions entre postures statiques. Nous pourrions ainsi reconnaître les classes suivantes : se lever, se coucher, s'asseoir, tomber, s'accroupir etc. Les activités de démarche (marche, course, saut etc.) pourraient aussi être reconnues grâce à l'ajout par exemple de données bas-niveau telles que la vitesse et la direction de déplacement du corps. Les interactions avec un objet (prise, dépôt, lancer etc.) constituent également une des perspectives d'étude et d'amélioration du système.

Ensuite, il serait intéressant d'extraire d'autres données bas-niveau au niveau de certaines étapes de traitement. Ces données pourraient servir à améliorer et à décrire plus précisément les *ROI*. Par exemple, lors de la segmentation, la combinaison d'informations de contour et de couleur affinerait et améliorerait les masques des *ROI*. Des modèles d'ombre basés sur la couleur avec des techniques d'invariance pourraient aussi être utilisés [Salvador01]. Les positions des pieds pourraient être estimées après la segmentation grâce à l'utilisation de cartes de distances géodésiques [Hernandez03]. Ces positions seraient utiles pour tester par exemple des modèles de squelettes de corps humain.

Puis, une autre perspective serait le passage à la 3D, en combinant les informations obtenues pour deux vues différentes de la scène obtenues par une vision stéréoscopique. La raison principale est qu'en vision par ordinateur, l'analyse et l'interprétation du comportement humain (postures, actions etc.) en tenant compte des occultations est un problème très difficile. Le passage à un traitement 3D permettrait de tester différentes méthodes robustes aux occultations, que ce soit en segmentation, en suivi temporel ou en reconnaissance d'actions et de postures et améliorerait sans doute les résultats des diverses étapes de traitement.

D'autres perspectives concernant le système complet sont basées sur la combinaison de ce travail avec d'autres travaux actuellement en cours dans le laboratoire. Il est déjà possible de réaliser l'animation d'un avatar en temps-réel en utilisant le résultat de la reconnaissance de postures statiques. Ensuite, il serait intéressant d'avoir une caméra filmant la scène en général et une autre zoomant sur le visage d'une personne filmée, après la localisation du visage. Ainsi

nous pourrions estimer la direction du regard et les expressions faciales afin de reconnaître les émotions et augmenter ainsi l'interactivité [Hammal05]. Enfin, une dernière perspective concerne l'intégration dans notre système d'autres modalités comme par exemple l'utilisation de plusieurs caméras avec des micros pour ajouter la parole à des vues multiples. Ceci pourrait conduire à des interfaces homme-machine multimodales avancées et à de nombreuses autres applications.

Annexe A

Projets *Art.live* et ARTUS

A.1 Le Projet *Art.live*

Les lignes de code qui ont servi de base à ce travail de thèse proviennent d'un projet qui est maintenant achevé et que nous allons présenter maintenant.

Le projet *Art.live (IST Project 10942, ARchitecture and authoring Tools prototype for Living Images and new Video Experiments)* [Art.live02] est un projet européen de l'*IST (Information Society Technology)*, organisme dont le but est de promouvoir le développement des technologies de l'information à travers l'Europe.

Ce projet, terminé en 2002, a eu pour but la mise en place d'une architecture et d'un ensemble d'outils, à la fois génériques et orientés application, pour l'amélioration des espaces narratifs dans lesquels réalité et monde virtuel étaient mélangés. Cet ensemble d'outils devait permettre aux artistes et aux utilisateurs de créer facilement des espaces narratifs où se mélangent les mondes réel et virtuel et de les disséminer en temps-réel à travers Internet (où n'importe quel réseau *TCP / IP*).

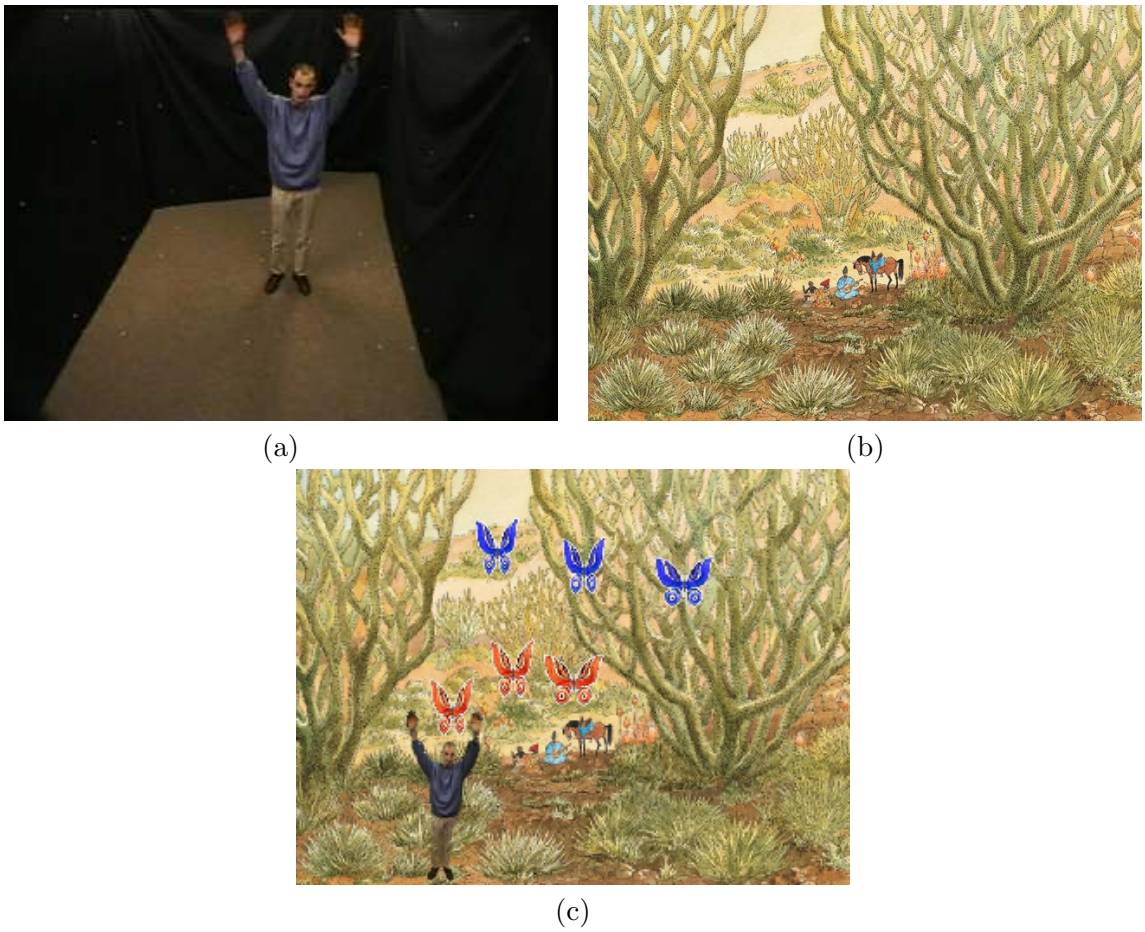
A.1.1 Objectif

De façon générale, le but de ce projet était d'expérimenter quelques pistes dans le vaste champ de la réalité mixte, qui provoque la rencontre des mondes réel et virtuel. Ceci résulte en des ambiances visuelles où les gens et les objets qui entrent dans le champ de la caméra sont incrustés dans un environnement virtuel sur des écrans géants et / ou sur Internet. Ces personnes se voient ensuite offrir la possibilité d'interagir avec l'histoire et avec d'autres personnes en utilisant une autre instance du système (contrôlée à distance). Par exemple, dans le cas de la figure A.1, le joueur essaie d'attraper des papillons virtuels. Quand tous les papillons ont été attrapés, le joueur lui-même se transforme en papillon. De plus complexes scénarios ont été mis en œuvre où plusieurs personnes en face de plusieurs caméras sont simultanément extraits du fond et incrustés dans le même milieu virtuel. C'était un projet européen de type applicatif puisque l'objectif était de réaliser deux démonstrateurs. Ces démonstrateurs devaient être capables d'extraire en temps-réel des personnes en mouvement dans une scène réelle, afin de les replacer dans un environnement virtuel. En 2001, a eu lieu une démonstration à Paris-Bercy à l'occasion du festival "Les Jardins et la Bande Dessinée". En 2002, le deuxième démonstrateur, consistant en un système interactif complet à deux caméras, a été présenté aux durant une exposition aux Salines Royales d'Arc-et-Senans.

Dans le cadre de ce projet, le LIS a apporté sa contribution en ce qui concerne la segmentation d'objets en mouvement et le codage de formes animées.

Plus précisément, le LIS a réalisé des algorithmes de :

- segmentation de personnes en vue de leur incrustation dans un monde virtuel ;
- suivi temporel de personnes, ce qui est utile, en particulier, quand il y a plusieurs personnes dans l'image.



(a) (b)

(c)

FIG. A.1 – Exemple d'incrustation d'un sujet humain dans un monde virtuel. (a) Image originale, (b) fond (dessin) et (c) image de réalité mixte. Les images sont copyright Casterman, F. Place et projet *Art.live*.

A.1.2 Partenaires du projet

Pour atteindre l'objectif fixé, *Art.live* rassemblait des ingénieurs en traitement du signal, des chercheurs en informatique et des auteurs multimédia. Ce projet européen faisait donc intervenir des partenaires de divers horizons dans les trois domaines de compétences suivants :

1. Traitement des images :
   - UCL (Université Catholique de Louvain) [Belgique] ;
   - UJF (Université Joseph Fourier), LIS [France] ;

  - EPFL (École Polytechnique Fédérale de Lausanne) [Suisse] ;
  - ADETTI / TMC [Portugal].
2. Support industriel :
  - ADERSA (traitement d'images industrielles) ;
  - FASTCOM (caméra).
3. Artistique et Multimédia :
  - Casterman.

A.2 Le Projet ARTUS

A.2.1 Objectif

Dans l'application cible, le codeur virtuel ARTUS (Animation Réaliste par Tatouage audiovisuel à l'Usage des Sourds) se substitue au télétexte à la demande de l'usager. Ce personnage se présente comme un petit clone virtuel 3D incrusté dans des émissions télévisuelles dont les mouvements du visage et de la main reproduisent les gestes du Langage Parlé Complété, langage des sourds conçu comme un complément à la lecture labiale. Les gestes de ce codeur virtuel sont calculés à partir du télétexte par un système de synthèse de gestes articulatoires à partir du texte développé dans le cadre de ce projet et sont transmis au terminal cible - typiquement un ordinateur personnel équipé d'une carte de réception numérique - en étant synchronisé avec l'émission. L'illustration du principe est visible sur la figure A.2.

La transmission effective de ces gestes s'effectue par tatouage des images de la séquence audiovisuelle originale. Si la transparence obtenue est suffisante, le flux de données sera augmenté d'instructions de positionnement du codeur ARTUS à l'écran voire de mouvements de visages liés aux expressions faciales.

A.2.2 Partenaires du projet

Ce projet fait intervenir divers partenaires dans les domaines de compétences suivants :

1. Traitement du Signal et des Images :
  - INPG (Institut National Polytechnique de Grenoble) LIS ;
  - INPG (Institut National Polytechnique de Grenoble) ICP ;
  - UTC (Université Technologique de Compiègne) HEUDIASYC ;
  - ENST (École Nationale Supérieure des Télécommunications) TSI.
2. Support industriel, Artistique et Multimédia :
  - ARTE ;
  - ATTITUDE STUDIO ;
  - THALES.

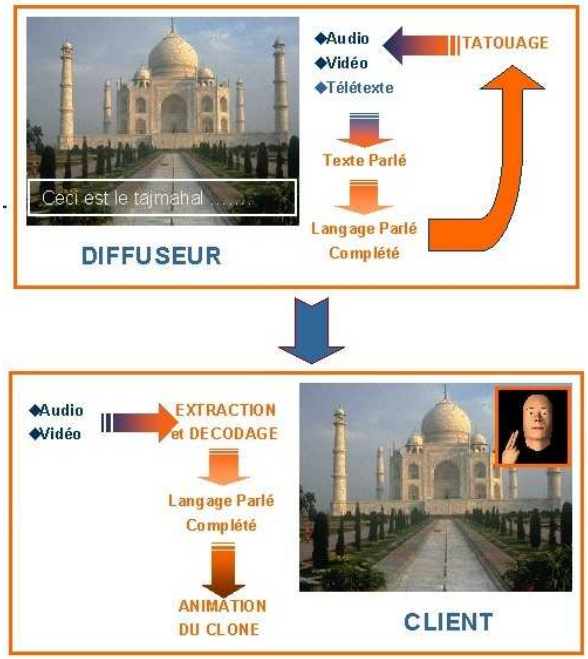


FIG. A.2 – Illustration du principe du codeur ARTUS.

Annexe B

Théorie sur le filtrage de Kalman

Ce chapitre est largement inspiré de [Maybeck79], qui comprend une présentation très pédagogique du filtrage de Kalman. Le filtre de Kalman a été nommé d'après Rudolph E. Kalman, qui décrivit en 1960 une solution récursive au problème de filtrage linéaire de données discrètes [Kalman60]. Une introduction plus complète peut être trouvée dans [Sorenson70], qui contient aussi plusieurs aspects historiques intéressants. Pour plus de références, le lecteur peut se reporter à [Gelb74, Lewis86, Jacobs93, Brown96, Grewal01].

B.1 Introduction

Le filtre de Kalman est un algorithme mathématique, et plus précisément un algorithme optimal et récursif de traitement du signal.

La notion d'optimalité peut être définie de multiple façons. On peut montrer que, sous certaines hypothèses qui seront faites plus loin, le filtre de Kalman est optimal vis-à-vis de quasiment tous les critères de mesure de performance auxquels on peut penser. Un des aspects de cette optimalité est que le filtre de Kalman utilise toute l'information qui lui est fournie. Il utilise toutes les mesures disponibles, quelle que soit leur précision, pour **estimer les valeurs des variables recherchées** en faisant usage :

1. de la dynamique connue du système et des dispositifs de mesure ;
2. des descriptions statistiques des bruits de mesure, de processus et de modèle ;
3. de toute information disponible sur les conditions initiales des variables recherchées.

Par exemple, pour estimer la vitesse d'un avion, on peut utiliser un radar à effet Doppler, les indications d'un système de navigation à inertie ou les mesures de pression statique et de vent relatif. Toutes ces mesures proviennent de dispositifs présentant un bruit de mesure, c'est-à-dire une imprécision plus ou moins grande sur la valeur mesurée. Plutôt que d'ignorer certaines mesures (par exemple celles dont l'imprécision est élevée), un filtre de Kalman pourrait être conçu de façon à les utiliser toutes et à estimer au mieux la vitesse de l'avion.

Le mot "récursif" signifie que le filtre de Kalman n'a pas besoin de toute l'information sur une plage de temps pour être lancé. Il peut fonctionner au fur et à mesure que le temps passe et que les mesures sont disponibles. De plus il ne nécessite pas de mémoriser toutes les mesures passées. Cela sera d'une grande importance pour l'implémentation du filtre.

Le filtre est bien un algorithme de traitement du signal. Contrairement à ce qu'évoque le mot "filtre" il ne s'agit pas d'un dispositif électronique ou mécanique mais bien d'un

algorithme en général implémenté sur calculateur *via* un programme informatique. Il en résulte qu'il utilise généralement des mesures en temps discret plutôt qu'en temps continu, bien que les deux approches soient possibles.

B.1.1 Un exemple simple

Supposons qu'un navigateur soit perdu en pleine mer. Par souci de simplicité, on considère que la position équivaut à une variable unidimensionnelle x. Une première visée sur les étoiles permet à l'instant t_1 de mesurer la position comme étant x_1. Toutefois, à cause des imprécisions des appareils de mesure et de l'œil humain, cette mesure est entachée d'imprécision et son écart-type est σ_1 (ou encore sa variance est σ_1^2). Ainsi on peut obtenir le tracé de la densité de probabilité $f(x|x_1)$ qui représente la probabilité (conditionnelle) que la localisation exacte soit x étant donnée la mesure x_1. σ_1 est une mesure directe de l'incertitude ou de l'imprécision sur la mesure. Plus grand est σ_1, plus étendue est la plage dans laquelle la probabilité de trouver x n'est pas négligeable. Rappelons que, pour une densité de probabilité gaussienne, il y a $68,3\%$ de chances de trouver la variable dans une bande d'une largeur de deux écart-types centrée autour de sa valeur moyenne. La figure B.1 illustre la densité de probabilité gaussienne $f(x|x_1)$, de moyenne x_1 et d'écart-type σ_1.

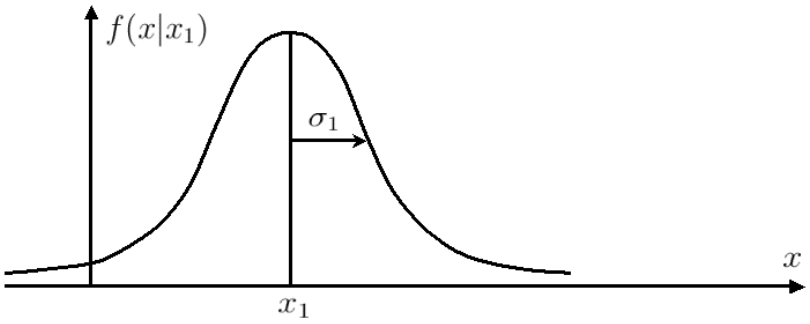


FIG. B.1 – Densité de probabilité gaussienne $f(x|x_1)$.

Supposons maintenant qu'un autre membre de l'équipage, utilisant des instruments plus précis, comme un *GPS* par exemple, fasse une autre mesure x_2 de la position à un instant t_2 quasiment identique à t_1. L'écart-type de cette deuxième mesure étant σ_2, supposons que σ_2 soit inférieur à σ_1, ce qui est cohérent avec l'idée d'une mesure plus précise. La figure B.2 illustre la densité de probabilité gaussienne $f(x|x_2)$, de moyenne x_2 et d'écart-type σ_2 superposée à celle de la figure B.1.

On dispose désormais de deux mesures faites au même instant de la position du navire. Il s'agit maintenant de déterminer une façon de combiner ces mesures afin d'obtenir une estimation plus précise de la position du navire. La solution naïve, qui consiste à moyenner "simplement" les deux valeurs mesurées n'a que peu de chances de donner une estimation optimale de la position. En effet, il paraît logique de tenir compte des imprécisions des mesures utilisées.

Il faut par conséquent tenir compte des valeurs mesurées **et de leurs variances**.

Il sera montré ci-après que, sous certaines hypothèses, on peut prendre :

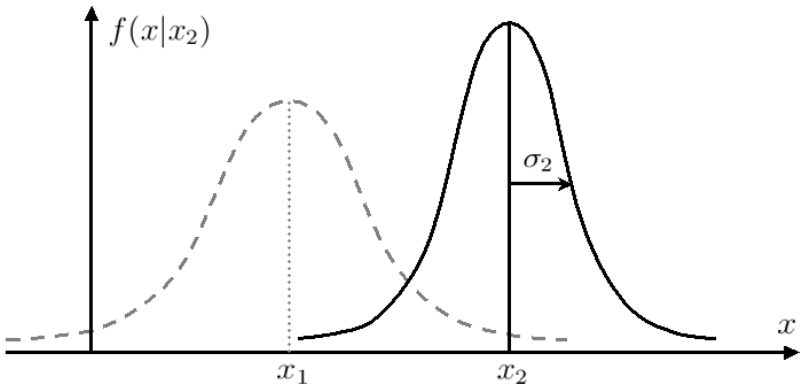


FIG. B.2 – Densité de probabilité gaussienne $f(x|x_2)$.

$$\begin{aligned} \tilde{x} &= \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}x_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}x_2, \\ \frac{1}{\sigma^2} &= \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}, \end{aligned}$$

où $\tilde{x}$ est l'estimée de x et σ^2 sa variance. On réalise donc une moyenne pondérée des valeurs mesurées, la pondération étant basée sur une normalisation des variances. La figure B.3 illustre la densité de probabilité gaussienne $f(x|x_1, x_2)$, de moyenne $\tilde{x}$ et d'écart-type σ superposée à celles de la figure B.2.

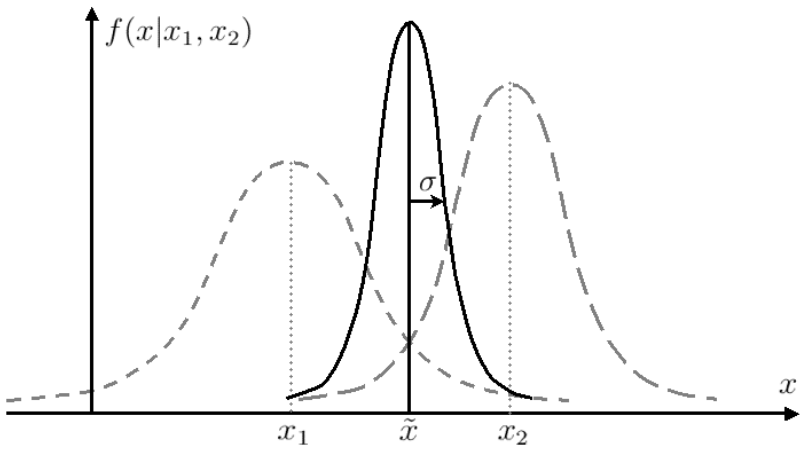


FIG. B.3 – Densité de probabilité gaussienne $f(x|x_1, x_2)$.

L'examen montre que les formules précédentes pour $\tilde{x}$ et σ sont tout à fait conformes au sens commun.

Si σ_1^2 et σ_2^2 étaient identiques, c'est-à-dire que les mesures sont de précision égales, alors la première formule montre tout simplement que $\tilde{x}$ serait la moyenne des deux mesures. En

revanche, si nous supposons que σ_1^2 est plus grande que σ_2^2, par hypothèse, cette formule propose de donner plus de poids à x_2 qu'à x_1. Enfin, la deuxième formule montre que σ^2 est plus petite que σ_1^2 même si σ_2^2 est grande ce qui montre que même les mesures de faible qualité apportent de l'information et qu'elle est utilisable.

Pour le démontrer, on a successivement :

$$\begin{aligned} \frac{1}{\sigma^2} &= \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} = \frac{\sigma_2^2 + \sigma_1^2}{\sigma_1^2 \sigma_2^2}, \\ \sigma^2 &= \frac{\sigma_1^2 \sigma_2^2}{\sigma_2^2 + \sigma_1^2} = \frac{\sigma_1^2}{1 + \frac{\sigma_1^2}{\sigma_2^2}}, \end{aligned}$$

d'où $\sigma^2 < \sigma_1^2$ (de même on montre que $\sigma^2 < \sigma_2^2$).

La formule en $\tilde{x}$ peut être réécrite en :

$$\begin{aligned} \tilde{x} &= \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} x_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} x_2 = \frac{\sigma_1^2 + \sigma_2^2 - \sigma_1^2}{\sigma_1^2 + \sigma_2^2} x_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} x_2, \\ \tilde{x} &= (1 - \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}) x_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} x_2 = x_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(x_2 - x_1), \\ \tilde{x} &= x_1 + G_2(x_2 - x_1) \text{ où } G_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}. \end{aligned}$$

Sous cette forme, cette formule peut s'interpréter de la façon suivante :

Tant que x_1 est la seule mesure disponible, l'estimation de x, $\tilde{x}$, est égale à x_1. Dès que x_2 est disponible alors x_1 est corrigée proportionnellement à l'écart entre l'ancienne et la nouvelle mesure. Ce schéma est du type "prédicteur-correcteur". La signification en deviendra plus claire encore dans le développement complet du filtre de Kalman.

Remarquons enfin que la variance σ^2 peut aussi s'écrire :

$$\sigma^2 = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{\sigma_1^2 + \sigma_2^2 - \sigma_1^2}{\sigma_1^2 + \sigma_2^2} \sigma_1^2 = (1 - \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}) \sigma_1^2 = (1 - G_2) \sigma_1^2.$$

Nous retiendrons donc, comme nouvelles expressions des formules de $\tilde{x}$ et de σ^2 :

$$\begin{aligned} \tilde{x} &= x_1 + G_2(x_2 - x_1) \text{ où } G_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \\ \sigma^2 &= (1 - G_2)\sigma_1^2. \end{aligned}$$

Pour l'instant, nous avons traité le cas stationnaire où le système n'évolue pas entre deux mesures. Supposons maintenant qu'un écart de temps significatif s'écoule entre les mesures x_1 et x_2. Le problème change de nature puisque le x qu'il faut estimer n'est plus le même aux instants t_1 et t_2. Faisons l'hypothèse que le déplacement du navire est linéaire en fonction du temps. En d'autres termes : $\frac{\delta x}{\delta t} = u + w$ où u est une vitesse nominale et w un bruit représentant les aléas sur cette vitesse ou notre méconnaissance de sa valeur exacte. On suppose aussi que w est un bruit blanc, gaussien de variance σ_w^2. La figure B.4 ci-dessous montre ce que devient la densité de probabilité de la position au cours du temps. En t_2 on a la situation précédente

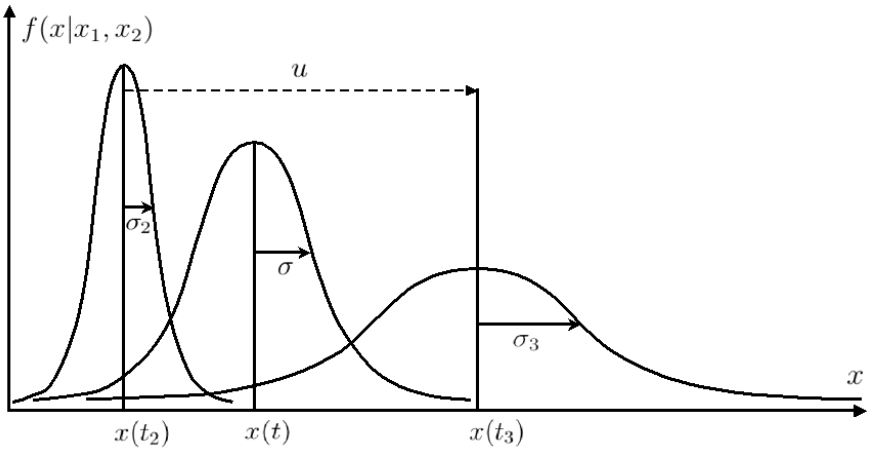


FIG. B.4 – Évolution temporelle de la densité de probabilité.

avec $x_2 = \tilde{x}_{2/2} = x(t_2)$ de variance $\sigma_2^2 = \sigma_{\tilde{x}_{2/2}}^2$. Au fur et à mesure que le temps s'écoule, elle se propage le long de l'axe des x et sa densité de probabilité s'étale.

En effet, à l'instant t_3, juste avant que la mesure x_3 ne soit faite, on a :

$$\begin{aligned} \hat{x}_{3/2} &= \tilde{x}_{2/2} + u(t_3 - t_2) \text{ (valeur prédite de l'état),} \\ \sigma_{\hat{x}_{3/2}}^2 &= \sigma_{\tilde{x}_{2/2}}^2 + \sigma_w^2(t_3 - t_2) \text{ (variance de l'état prédit).} \end{aligned}$$

Dès que la mesure x_3, d'écart-type σ_3, est disponible, on peut écrire, conformément aux formules du cas stationnaire, en comparant la valeur prédite $\hat{x}_{3/2}$ et la valeur mesurée x_3 :

$$\begin{aligned} \tilde{x}_{3/3} &= \hat{x}_{3/2} + G_3(x_3 - \hat{x}_{3/2}) \text{ où } G_3 = \frac{\sigma_{\hat{x}_{3/2}}^2}{\sigma_{\hat{x}_{3/2}}^2 + \sigma_3^2}, \\ \sigma_{\tilde{x}_{3/3}}^2 &= (1 - G_3)\sigma_{\hat{x}_{3/2}}^2. \end{aligned}$$

On constate, d'après la forme de G_3, que si la variance σ_3^2 du bruit de mesure est grande alors G_3 est petit. Cela signifie simplement que le filtrage accorde peu de confiance à des mesures très bruitées, ce qui est conforme au bon sens. À l'extrême, si σ_3^2 était infiniment grand, G_3 deviendrait nul et on aurait $\tilde{x}_{3/3} = \hat{x}_{3/2}$, c'est-à-dire qu'une valeur infiniment bruitée serait ignorée. De même, si σ_w^2 était grand, alors $\sigma_{\hat{x}_{3/2}}^2$ le serait également et donc G_3 aussi. Cela traduit l'idée intuitive que si le modèle de prédiction de l'état n'est pas fiable alors il faut faire plus confiance aux mesures. À l'extrême, si σ_w^2 est proche de l'infini alors $\sigma_{\hat{x}_{3/2}}^2$ l'est aussi et G_3 tend vers 1 d'où $\tilde{x}_{3/3} = x_3$. La valeur prédite de l'état est alors complètement ignorée.

Tous ces résultats dont la signification pratique a été soulignée par l'intuition vont maintenant être démontrés.

B.2 Le filtre de Kalman à état discret

Considérons un système stochastique dont la représentation d'état discrète s'écrit :

$$\begin{aligned} x_{t+1/t} &= A_t x_t + B_t u_t + w_t, \\ s_t &= C_t x_t + v_t, \end{aligned}$$

avec v_t et w_t pseudo-bruits blancs indépendants :

$$\begin{aligned} E[v_t] &= 0, \\ E[w_t] &= 0, \\ E[v_t v_t^{'T}] &= R\delta_{tt'}, \\ E[w_t w_t^{'T}] &= Q\delta_{tt'}, \\ E[v_t w_t^{'T}] &= 0, \end{aligned}$$

R et Q étant des matrices symétriques définies positives.

Nous supposerons que les matrices A_t, B_t et C_t sont connues : le problème est alors de trouver la meilleure estimation de x_t, que nous noterons $\underline{\tilde{x}}_{t/t}$. Nous noterons $P_{t/t}$ la matrice de covariance de $\underline{\tilde{x}}_{t/t}$.

Soit x_0 la valeur initiale de x_t : x_0 peut être soit observée, soit estimée. La matrice de covariance de x_0 sera notée P_0.

Ainsi le schéma global du filtre est celui illustré par la figure B.5 : à partir de s_t et de u_t, nous recherchons $\underline{\tilde{x}}_{t/t}$.

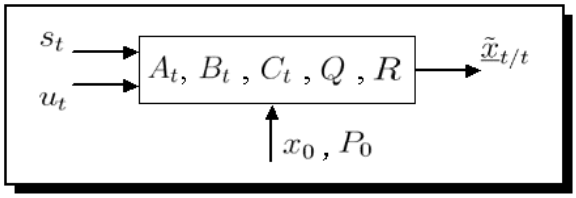


FIG. B.5 – Schéma global du filtre de Kalman.

Par ailleurs, la dynamique du système peut être représentée par le schéma de la figure B.6, où z^{-1} représente l'opérateur retard.

À l'instant t, nous recueillons l'information s_t que nous allons pouvoir utiliser pour améliorer l'estimation de x_t que l'on pouvait avoir *a priori*.

B.3 Équations de prédiction et de filtrage

B.3.1 Équations de prédiction

À l'instant t, nous appelons **prédiction**, la détermination d'un **estimateur** *a priori* de x_t que nous noterons $\underline{\hat{x}}_{t/t-1}$. Nous supposons donc connu le meilleur estimateur de x_{t-1}, soit

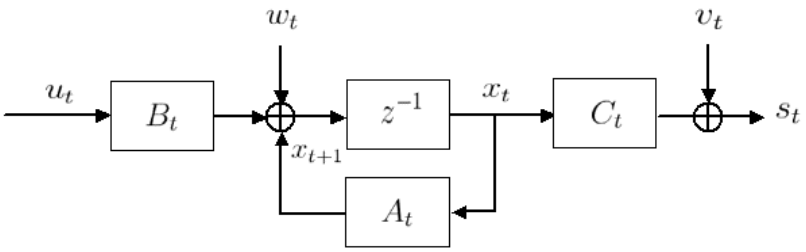


FIG. B.6 – Schéma détaillé de la dynamique du filtre de Kalman.

$\underline{\tilde{x}}_{t-1/t-1}$ et l'équation d'état va nous renseigner sur l'évolution du système. C'est pourquoi nous choisissons l'estimateur *a priori* donné par :

$$\underline{\hat{x}}_{t/t-1} = A_{t-1}\underline{\tilde{x}}_{t-1/t-1} + B_{t-1}u_{t-1}. \tag{B.1}$$

Nous pouvons aussi calculer la matrice de covariance $P_{t/t-1}$ de $\underline{\hat{x}}_{t/t-1}$. En soustrayant l'équation d'état à l'équation ci-dessus, nous obtenons :

$$\underline{\hat{x}}_{t/t-1} - x_t = A_{t-1}(\underline{\tilde{x}}_{t-1/t-1} - x_{t-1}) - w_{t-1}.$$

Développons alors $P_{t/t-1}$:

$$\begin{aligned} P_{t/t-1} &= E[(\underline{\hat{x}}_{t/t-1} - x_t)(\underline{\hat{x}}_{t/t-1} - x_t)^T] \\ &= A_{t-1}E[(\underline{\tilde{x}}_{t-1/t-1} - x_{t-1})(\underline{\tilde{x}}_{t-1/t-1} - x_{t-1})^T]A_{t-1}^T + E[w_{t-1}w_{t-1}^T] \\ &\quad -A_{t-1}E[(\underline{\tilde{x}}_{t-1/t-1} - x_{t-1})w_{t-1}^T] - E[w_{t-1}(\underline{\tilde{x}}_{t-1/t-1} - x_{t-1})^T]A_{t-1}^T. \end{aligned}$$

Le bruit w_{t-1} qui intervient à l'instant $t-1$ est indépendant de l'estimation à l'instant $t-1$. Par conséquent :

$$\begin{aligned} E[(\underline{\tilde{x}}_{t-1/t-1} - x_{t-1})w_{t-1}^T] &= 0, \\ E[w_{t-1}(\underline{\tilde{x}}_{t-1/t-1} - x_{t-1})^T] &= 0. \end{aligned}$$

Comme $E[(\underline{\tilde{x}}_{t-1/t-1} - x_{t-1})(\underline{\tilde{x}}_{t-1/t-1} - x_{t-1})^T] = P_{t-1/t-1}$ et que $E[w_{t-1}w_{t-1}^T] = Q$, nous obtenons alors la relation entre $P_{t/t-1}$ et $P_{t-1/t-1}$:

$$P_{t/t-1} = A_{t-1}P_{t-1/t-1}A_{t-1}^T + Q. \tag{B.2}$$

Les équations (B.1) et (B.2) constituent les équations de prédiction du filtre de Kalman.

B.3.2 Équations de filtrage

Les équations de filtrage permettent de calculer le meilleur estimateur $\underline{\tilde{x}}_{t/t}$, ou **estimateur** *a posteriori*, en fonction de l'estimateur *a priori* $\underline{\hat{x}}_{t/t-1}$. La correction va être effectuée

en fonction de l'**observation** s_t. Nous allons rechercher $\underline{\tilde{x}}_{t/t}$ sous forme d'une combinaison linéaire de l'estimateur *a priori* et de l'observation, soit :

$$\underline{\tilde{x}}_{t/t} = L_t \underline{\hat{x}}_{t/t-1} + G_t s_t.$$

Nous recherchons un estimateur non biaisé. Par conséquent, nous devons avoir, pour tout t : $E[\underline{\tilde{x}}_{t/t}] = x_t$. Alors l'équation (B.1) nous indique que :

$$E[\underline{\hat{x}}_{t/t-1}] = A_{t-1} E[\underline{\tilde{x}}_{t-1/t-1}] + B_{t-1} u_{t-1} = A_{t-1} x_{t-1} + B_{t-1} u_{t-1} = x_t.$$

Pour que $\underline{\tilde{x}}_{t/t}$ soit non biaisé, L_t et G_t doivent donc vérifier :

$$x_t = L_t x_t + G_t s_t = L_t x_t + G_t C_t x_t = (L_t + G_t C_t) x_t,$$

soit encore : $L_t + G_t C_t = Id$.
En remplaçant L_t par $Id - G_t C_t$ nous obtenons :

$$\underline{\tilde{x}}_{t/t} = \underline{\hat{x}}_{t/t-1} + G_t(s_t - C_t \underline{\hat{x}}_{t/t-1}). \tag{B.3}$$

On appelle G_t le **gain de Kalman** et $s_t - C_t \underline{\hat{x}}_{t/t-1}$ l'**innovation** : c'est la différence entre l'observation et l'estimée *a priori* de l'observation. Nous allons calculer la matrice de covariance $P_{t/t}(G_t)$ et nous choisirons le gain G_t qui minimise $P_{t/t}$. Nous avons :

$$\underline{\tilde{x}}_{t/t} - x_t = \underline{\hat{x}}_{t/t-1} - x_t + G_t(s_t - C_t \underline{\hat{x}}_{t/t-1}) = (Id - G_t C_t)(\underline{\hat{x}}_{t/t-1} - x_t) + G_t v_t.$$

En remarquant que v_t et $\underline{\hat{x}}_{t/t-1}$ sont indépendants, nous obtenons :

$$\begin{aligned} P_{t/t} &= E[(\underline{\tilde{x}}_{t/t} - x_t)(\underline{\tilde{x}}_{t/t} - x_t)^T] \\ &= (Id - G_t C_t) P_{t/t-1} (Id - G_t C_t)^T + G_t R G_t^T \\ &= P_{t/t-1} + G_t(R + C_t P_{t/t-1} Ct^T) G_t^T - G_t C_t P_{t/t-1} - P_{t/t-1} C_t^T G_t^T. \end{aligned}$$

Posons $D_t = R + C_t P_{t/t-1} C_t^T$. En regroupant les termes en G_t, l'équation précédente s'écrit :

$$P_{t/t} = (G_t - P_{t/t-1} C_t^T D_t^{-1}) D_t (G_t - P_{t/t-1} C_t^T D_t^{-1})^T + P_{t/t-1} - P_{t/t-1} C_t^T D_t^{-1} C_t P_{t/t-1}.$$

Puisque D_t est symétrique définie positive, elle est donc inversible. En effet, s'il n'en était pas ainsi, il y aurait un x tel que :

$$x R x^T = -x C_t P_{t/t-1} C_t^T x^T,$$

donc R ne serait pas positive, ce qui est pourtant le cas.

Pour minimiser la matrice $P_{t/t}$, il nous faut d'abord définir un critère. On pourrait chercher à minimiser une norme de $P_{t/t}$, par exemple : $||P_{t/t}^2|| = trace[P_{t/t}^T P_{t/t}]$.

Le critère le plus simple est la minimisation directe de $trace(P_{t/t})$, en effet, cette application possède toutes les propriétés des normes pour des matrices définies positives, mais ce

n'est pas une norme puisque cet ensemble n'a pas une structure d'espace vectoriel. Le choix de ce critère est assez intuitif si l'on remarque que dans la base propre de $P_{t/t}$, la trace s'écrit :

$$trace(P_{t/t}) = \sum_j (\sigma_{t/t}^j)^2,$$

où $\sigma_{t/t}^j$ est l'écart-type de $\underline{\tilde{x}}_{t/t}$ dans la $j^{ième}$ direction propre.
Nous allons donc rechercher le gain G_t solution de :

$$\min_{G_t} trace[P_{t/t}(G_t)].$$

Et la trace étant un opérateur linéaire, il nous suffit de résoudre :

$$\min_{G_t} trace[(G_t - P_{t/t-1}C_t^T D_t^{-1}) D_t (G_t - P_{t/t-1}C_t^T D_t^{-1})^T].$$

Une solution évidente pour annuler un critère positif est de choisir la valeur qui annule ce critère, soit ici :

$$\begin{aligned} G_t &= P_{t/t-1}C_t^T D_t^{-1}, \\ G_t &= P_{t/t-1}C_t^T (R + C_t P_{t/t-1} C_t^T)^{-1}. \end{aligned} \tag{B.4}$$

Alors la matrice $P_{t/t}$ vérifie :

$$\begin{aligned} P_{t/t} &= P_{t/t-1} - P_{t/t-1}C_t^T D_t^{-1} C_t P_{t/t-1}, \\ P_{t/t} &= (Id - G_t C_t) P_{t/t-1}. \end{aligned} \tag{B.5}$$

Les équations (B.3), (B.4) et (B.5) constituent les équations de filtrage du filtre de Kalman.

B.4 Conclusion

En résumé, le filtre de Kalman a pour équations :
équations de prédiction :

$$\begin{aligned} \underline{\hat{x}}_{t/t-1} &= A_{t-1}\underline{\tilde{x}}_{t-1/t-1} + B_{t-1}u_{t-1}, \\ P_{t/t-1} &= A_{t-1}P_{t-1/t-1}A_{t-1}^T + Q, \end{aligned}$$

équations de filtrage :

$$\begin{aligned} \underline{\tilde{x}}_{t/t} &= \underline{\hat{x}}_{t/t-1} + G_t(s_t - C_t\underline{\hat{x}}_{t/t-1}), \\ G_t &= P_{t/t-1}C_t^T(R + C_t P_{t/t-1}C_t^T)^{-1}, \\ P_{t/t} &= (Id - G_t C_t)P_{t/t-1}, \end{aligned}$$

conditions initiales :

$$\begin{aligned} \underline{\tilde{x}}_{0/-1} &= \underline{x}_0, \\ P_{0/-1} &= P_0. \end{aligned}$$

Annexe C

Plate-forme AIM

Le LIS dispose d'une plate-forme interactive AIM (Analyse Interprétation Multimodalités) pour tester les algorithmes développés. La plate-forme se compose de matériels standards et d'éléments assez spécifiques, adaptés à sa finalité.

C.1 Éléments de la plate-forme

Voici la liste complète des éléments composant la plate-forme. Les éléments **en gras** ont été principalement utilisés pour la réalisation du travail présenté dans ce mémoire.

1. Éclairage
   - **3 projecteurs de 500 Watts**, non fixés ;
   - 4 projecteurs de 1000 Watts, fixés aux murs et orientables ;
   - 1 cadre $3,60m \times 3,60m$, fixé au plafond, avec une toile permettant une diffusion homogène de la lumière par réflexion, ceci afin de limiter les ombres.
2. Acquisition
   - **2 caméras numériques monoculaires** *Sony DFW-VL500* ;
   - 1 caméra numérique stéréoscopique *Bumblebee BB-COL-40* ;
   - 2 tourelles contrôlables pour supporter des caméras ou un écran plat ;
   - 1 micro cravate sans fil ;
   - 1 station d'acquisition de signaux biologiques (ECG, EEG etc.).

3 Traitement
   - **2 ordinateurs bi-processeur à 3.2 GHz** sous environnement *UNIX*, afin de réaliser un traitement temps-réel et d'effectuer des mesures de cadence de traitement.

4 Affichage / Visualisation
   - **2 écrans** (moniteurs des ordinateurs de traitement) ;
   - 1 écran plat orientable en pan et tilt ;
   - 1 vidéo projecteur ;
   - 1 écran de projection mural.

C.2 Caractéristiques des caméras

Les deux types de caméras numériques, *Sony DFW-VL500* et *Bumblebee BB-COL-40*, ont les caractéristiques suivantes :

Vitesse d'acquisition : 30 images/s.
Résolution des images : 640×480 pixels ou 320×240 pixels.
Format des images : Couleur $YCbCr$ (format $4:2:0$) ou Niveaux de Gris Y.

Publications

Revues internationales avec comité de lecture

[Girondel03] V. Girondel, L. Bonnaud, and A. Caplier. Hands detection and tracking for interactive multimedia applications, *Archives of Theoretical and Applied Informatics*, 2003.

[Girondel05b] V. Girondel, A. Caplier, L. Bonnaud, and M. Rombaut. Belief theory-based classifiers comparison for static human body postures recognition in video, *International Journal of Signal Processing IJSP*, 2(1) :29-33, March 2005.

[Girondel06] V. Girondel, L. Bonnaud, and A. Caplier. A human body analysis system, *European Journal on Applied Signal Processing - EURASIP-JASP*, Article ID 61927, 18 pages, to appear, 2006.

Conférences internationales avec actes et comité de lecture

[Girondel02] V. Girondel, L. Bonnaud, and A. Caplier. Hands detection and tracking for interactive multimedia applications, *Proceedings of the International Conference on Computer Vision and Graphics*, pages 282-287, September 2002, Zakopane, Poland.

[Girondel04] V. Girondel, A. Caplier, and L. Bonnaud. Real-time tracking of multiple persons by Kalman filtering and face pursuit for multimedia applications, *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 201-205, March 2004, South Lake Tahoe, Nevada, USA.

[Girondel05a] V. Girondel, A. Caplier, L. Bonnaud, and M. Rombaut. Belief theory-based classifiers comparison for static human body postures recognition in video, *Proceedings of the International Conference on Pattern Recognition and Computer Vision*, (Part of the Second World Enformatika Congress WEC 2005), pages 237-240, February 2005, Istanbul, Turkey.

[Girondel05c] V. Girondel, L. Bonnaud, A. Caplier, and M. Rombaut. Static human body postures recognition in video sequences using the belief theory, *Proceedings of the IEEE International Conference on Image Processing*, 2 :45-48, September 2005, Genoa, Italy.

[Girondel05d] V. Girondel, A. Caplier, and L. Bonnaud. A belief theory-based static posture recognition system for videosurveillance applications, *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 10-15, September 2005, Como, Italy.

Bibliographie

[Aach93] T. Aach, A. Kaup, and R. Mester. Statistical model-based detection in moving videos. *Signal Processing*, 31(2) :165–180, March 1993.

[Aggarwal98] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata. Non-rigid motion analysis : articulated and elastic motion. *Computer Vision and Image Understanding*, 70(2) :142–156, May 1998.

[Aggarwal99] J. K. Aggarwal and Q. Cai. Human motion analysis : a review. *Computer Vision and Image Understanding*, 73(3) :428–440, March 1999.

[Ahlberg99] J. Ahlberg. *Extraction and coding of face model parameters.* PhD thesis, Licentiate Thesis No. 747, ISBN 9172194251, Linköping University, Sweden, March 1999.

[Aizawa95] K. Aizawa and T. S. Huang. Model-based image-coding : advanced video coding techniques for very-low bit-rate applications. *Proceedings of the IEEE*, 83(2) :259–271, February 1995.

[Alterface06] Alterface website. http ://www.alterface.com/. *ALTERFACE*, 2006.

[Amat99] J. Amat, A. Casals, and M. Frigola. Stereoscopic system for human body tracking in natural scenes. *Proceedings of the IEEE International Workshop on Modeling People*, pages 70–78, September 1999.

[Appriou91] A. Appriou. Probabilités et incertitude en fusion de données multisenseurs. *Revue Scientifique et Technique de la Défense*, 11 :27–40, 1991.

[Appriou93] A. Appriou. Formulation et traitement de l'incertain en analyse multisenseurs. *Proceedings of the Colloque GRETSI*, pages 951–954, September 1993.

[Art.live02] Art.live project website. http ://www.tele.ucl.ac.be/PROJECTS/art.live/. *Art.live - IST 10942*, 2002.

[Barralon05] P. Barralon. *Classification et fusion de données actimétriques pour la télévigilance médicale : élaboration et validation expérimentale chez la personne jeune et agée.* PhD thesis, Université Joseph Fourier, Grenoble, 2005.

[Barrow77] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf. Parametric correspondence and chamfer matching : two new techniques for image matching. *Proceedings of the International Joint Conference on Artificial Intelligence*, 2 :659–663, August 1977.

[Bartoli02] A. Bartoli, N. Dalal, B. Bose, and R. Horaud. From video sequences to motion panoramas. *Proceedings of the IEEE International Workshop on Motion and Video Computing*, pages 201–207, December 2002.

[Bartoli04] A. Bartoli, N. Dalal, and R. Horaud. Motion panoramas. *Computer Animation and Virtual Worlds*, 15(5) :501–517, November 2004.

[Baumberg94] A. M. Baumberg and D. C. Hogg. An efficient method for contour tracking using active shape models. *Proceedings of the IEEE International Workshop on Motion of Non-Rigid and Articulated Objects*, pages 194–199, November 1994.

[Benoit05] A. Benoit and A. Caplier. Head nods analysis : interpretation of non verbal communication gestures. *Proceedings of the IEEE International Conference on Image Processing*, 3 :425–428, September 2005.

[Besag86] J. E. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, B-48(3) :259–302, 1986.

[Bezdek81] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms.* ISBN 0306406713, 256 pages, Kluwer Academic, Norwell, MA, USA, 1981.

[Bharati Devi85] B. Bharati Devi and V. V. S. Sarma. Estimation of fuzzy memberships from histograms. *Information Sciences*, 35(1) :43–59, March 1985.

[Bharatkumar94] A. G. Bharatkumar, K. E. Daigle, M. G. Pandy, Q. Cai, and J. K. Aggarwal. Lower limb kinematics of human walking with the medial axis transformation. *Proceedings of the IEEE International Workshop on Motion of Non-Rigid and Articulated Objects*, pages 70–76, November 1994.

[Birchfield98] S. T. Birchfield. Elliptical head tracking using intensity gradients and color histograms. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 232–237, June 1998.

[Bishop95] C. M. Bishop. *Neural networks for pattern recognition.* ISBN 0198538642, 500 pages, Oxford University Press, New York, NY, USA, 1995.

[Black95] M. J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motion using local parametric models of image motion. *Proceedings of the IEEE International Conference on Computer Vision*, pages 374–381, June 1995.

[Bloch96a] I. Bloch. Information combination operators for data fusion : a comparative review with classification. *IEEE Transactions on Systems, Man and Cybernetics*, A-26(1) :52–67, January 1996.

[Bloch96b] I. Bloch. Some aspects of Dempster-Shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account. *Pattern Recognition Letters*, 17(8) :905–919, July 1996.

[Bloch97] I. Bloch. Using fuzzy mathematical morphology in the Dempster-Shafer framework for image fusion under imprecision. *Proceedings of the Special Interest Group for Fuzzy Set Theory and Applications*, pages 209–214, June 1997.

[Bloch05] I. Bloch. Fusion d'informations numériques : panorama méthodologique. *Proceedings of the Journées Nationales de la Recherche en Robotique*, pages 79–88, October 2005.

[Bobick95] A. F. Bobick and A. D. Wilson. A state-based technique for the summarization and recognition of gesture. *Proceedings of the IEEE International Conference on Computer Vision*, pages 382–388, June 1995.

[Bobick96] A. F. Bobick and J. W. Davis. Real-time recognition of activity using temporal templates. *Proceedings of the IEEE International Workshop on Applications of Computer Vision*, pages 39–42, December 1996.

[Bobick97] A. F. Bobick and A. D. Wilson. A state based approach to the representation and recognition of gesture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(12) :1325–1337, December 1997.

[Bobick99] A. F. Bobick, S. S. Intille, J. W. Davis, F. Baird, C. S. Pinhanez, L. W. Campbell, Y. A. Ivanov, A. Schutte, and A. Wilson. The KidsRoom : a perceptually-based interactive and immersive story environment. *Presence : Teleoperators and Virtual Environments*, 8(4) :367–391, August 1999.

[Boghossian99a] B. A. Boghossian and S. A. Velastin. Image processing system for pedestrian monitoring using neural classification of normal motion patterns. *Measurement and Control (Special Issue on Intelligent Vision Systems)*, 32(9) :261–264, 1999.

[Boghossian99b] B. A. Boghossian and S. A. Velastin. Motion-based machine vision techniques for the management of large crowds. *Proceedings of the IEEE International Conference on Electronics, Circuits and Systems*, 2 :961–964, September 1999.

[Braffort04] A. Braffort, A. Choisier, C. Collet, P. Dalle, F. Gianni, B. Lenseigne, and J. Segouat. Toward an annotation software for video of sign language, including image processing tools and signing space modelling. *Proceedings of the International Conference on Language Resources and Evaluation*, 1 :201–203, May 2004.

[Brand97] M. Brand, N. Oliver, and A. P. Pentland. Coupled hidden Markov models for complex action recognition. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 994–999, June 1997.

[Brand99] M. Brand. Shadow puppetry. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2 :1237–1244, September 1999.

[Bregler97] C. Bregler. Learning and recognizing human dynamics in video sequences. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 568–574, June 1997.

[Brown96] R. G. Brown and P. Y. Hwang. *Introduction to random signals and applied Kalman filtering with MATLAB exercises and solutions (3rd Ed.)*. ISBN 0471128392, 496 pages, John Wiley and Sons, Inc, New York, NY, USA, 1996.

[Bruce00] J. Bruce, T. Balch, and M. Veloso. Fast and inexpensive color image segmentation for interactive robots. *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, 3 :2061–2066, October 2000.

[Burger06] T. Burger, A. Benoit, and A. Caplier. Intercepting static hand gestures in dynamic context. *Proceedings of the IEEE International Conference on Image Processing*, to appear, October 2006.

[Campbell95a] L. W. Campbell and A. F. Bobick. Recognition of human body motion using phase space constraints. *Proceedings of the IEEE International Conference on Computer Vision*, pages 624–630, June 1995.

[Campbell95b] L. W. Campbell and A. F. Bobick. Using phase space constraints to represent human body motion. *Proceedings of the IEEE International Workshop on Automatic Face and Gesture Recognition*, pages 338–343, June 1995.

[Canton-Ferrer05] C. Canton-Ferrer, J. R. Casas, and M. Pardas. Fusion of multiple viewpoint information towards 3D face robust orientation detection. *Proceedings of the IEEE International Conference on Image Processing*, 2 :366–369, September 2005.

[Canton-Ferrer06] C. Canton-Ferrer, J. R. Casas, and M. Pardas. Human model and motion based 3D action recognition in multiple view scenarios. *Proceedings of the European Signal Processing Conference*, to appear, September 2006.

[Capellades03] 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*, 2 :85–88, September 2003.

[Capelle03] A. S. Capelle, O. Colot, and C. Fernandez-Maloigne. 3D segmentation of MR brain images into white matter, gray matter and cerebro-spinal fluid by means of evidence theory. *Proceedings of the European Conference on Artificial Intelligence in Medecine*, pages 112–116, October 2003.

[Capelle04] A. S. Capelle, O. Colot, and C. Fernandez-Maloigne. Evidential segmentation scheme of multi-echo MR images for the detection of brain tumors using neighborhood information. *Information Fusion*, 5(3) :203–216, October 2004.

[Capin00a] T. K. Capin, E. Petajan, and J. Ostermann. Efficient modeling of virtual humans in MPEG-4. *Proceedings of the IEEE International Conference on Multimedia and Exposition*, 2 :1103–1106, July 2000.

[Capin00b] T. K. Capin, E. Petajan, and J. Ostermann. Very low bitrate coding of virtual human animation in MPEG-4. *Proceedings of the IEEE International Conference on Multimedia and Exposition*, 2 :1107–1110, July 2000.

[Caplier95] A. Caplier. *Modèles markoviens de détection de mouvement dans les séquences d'images : approche spatio-temporelle et mises en oeuvre temps-réel.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, 1995.

[Caplier01] A. Caplier, L. Bonnaud, and J-M. Chassery. Robust fast extraction of video objects combining frame differences and adaptative reference image. *Proceedings of the IEEE International Conference on Image Processing*, 2 :785–788, September 2001.

[Cavallaro01] A. Cavallaro and T. Ebrahimi. Video objects extraction based on adaptative background and statistical change detection. *Proceedings of the SPIE Electronic Imaging*, pages 465–475, January 2001.

[Cedras95] C. Cedras and M. Shah. Motion-based recognition : a survey. *Image Vision Computing*, 13(2) :129–155, March 1995.

[Chai99] D. Chai and K. N. Ngan. Face segmentation using skin-color map in videophone applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(4) :551–564, June 1999.

[Chellappa95] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces : a survey. *Proceedings of the IEEE*, 83(5) :705–740, May 1995.

[Chen92] Z. Chen and H. J. Lee. Knowledge-guided visual perception of 3D human gait from a single image sequence. *IEEE Transactions on Systems, Man and Cybernetics*, 22(2) :336–342, April 1992.

[Cheng97] H. D. Cheng and J. R. Chen. Automatically determine the membership function based on the maximum entropy principle. *Information Sciences*, 96(3-4) :163–182, February 1997.

[Chomat98] O. Chomat and J. L. Crowley. Recognizing motion using local appearance. *Proceedings of the International Symposium on Intelligent Robotic Systems*, pages 271–279, 1998.

[Civanlar86] M. R. Civanlar and H. J. Trussel. Constructing membership functions using statistical data. *Fuzzy Sets and Systems*, 18(1) :1–13, January 1986.

[Collins00] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa. A system for video surveillance and monitoring. *Tech. report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon university*, May 2000.

[Cooke91] R. M. Cooke. *Experts in uncertainty.* ISBN 0195064658, 336 pages, Oxford University Press, New York, NY, USA, 1991.

[Couleur04] Ouvrage collectif du groupe couleur du GDR-PRC ISIS. *Image couleur : de l'acquisition au traitement.* ISBN 2100068431, 460 pages, Sous la direction d'A. Trémeau, C. Fernandez-Maloigne et P. Bonton, Ed. Dunod, 2004.

[Crowley97] J. L. Crowley and F. Berard. Multi-modal tracking of faces for video communications. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 640–645, June 1997.

[Cui96] Y. Cui and J. J. Weng. Hand segmentation using learning-based prediction and verification for hand sign recognition. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 88–93, June 1996.

[Cunado99] D. Cunado, M. S. Nixon, and J. N. Carter. Automatic gait recognition via model-based evidence gathering. *Proceedings of the IEEE International Workshop on Automatic Identification Advanced Technologies*, pages 27–30, October 1999.

[Dalal02] N. Dalal and R. Horaud. Indexing key positions between multiple videos. *Proceedings of the IEEE International Workshop on Motion and Video Computing*, pages 65–71, December 2002.

[Darrell93] T. J. Darrell and A. P. Pentland. Space-time gestures. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 335–340, June 1993.

[Darrell94] T. J. Darrell, P. Maes, B. Blumberg, and A. Pentland. A novel environment for situated vision and behavior. *Proceedings of the IEEE International Workshop for Visual Behavior*, pages 68–72, June 1994.

[Darrell95] T. J. Darrell, B. Blumberg, S. Daniel, B. Rhodes, P. Maes, and A. P. Pentland. Alive : dreams and illusions. *Visual Proceedings of the ACM SIGGraph Conference on Computer Graphics*, pages 267–284, July 1995.

[Darrell96] T. J. Darrell, B. Moghaddam, and A. P. Pentland. Active face tracking and pose estimation in an interactive room. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 67–72, June 1996.

[De Luca72] A. De Luca and S. Termini. A definition of non-probabilistic entropy in the setting of fuzzy set theory. *Information and Control*, 20 :301–312, 1972.

[Dempster68] A. Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society*, B-30 :205–247, 1968.

[Denœux95] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25(5) :804–813, May 1995.

[Denœux97] T. Denœux. Analysis of evidence-theory decision rules for pattern classification. *Pattern Recognition*, 30(7) :1095–1107, 1997.

[Deveughele93] S. Deveughele and B. Dubuisson. Using possibility theory in perception : an application in artificial vision. *Proceedings of the IEEE International Conference on Fuzzy Systems*, pages 821–826, 1993.

[Dockstader01] S. L. Dockstader and A. M. Tekalp. On the tracking of articulated and occluded video object motion. *Real Time Imaging*, 7(5) :415–432, October 2001.

[Dubois83] D. Dubois and H. Prade. Unfair coins and necessity measures : towards a possibilistic interpretation of histograms. *Fuzzy Sets and Systems*, 10(1) :15–20, 1983.

[Dubois85] D. Dubois and H. Prade. A review of fuzzy sets aggregation connectives. *Information Sciences*, 36 :85–121, 1985.

[Dubois88] D. Dubois and H. Prade. *Possibility Theory : an approach to computerized processing of uncertainty.* ISBN 0306425203, 280 pages, Plenum Press, New York, NY, USA, 1988.

[Dubois92] D. Dubois and H. Prade. *Combination of information in the framework of possibility theory, in Data Fusion in Robotics and Machine Intelligence.* ISBN 0120421208, 560 pages, M. Al Abidi Eds., Academic Press, 1992.

[Dubois94] D. Dubois and H. Prade. Fuzzy sets : a convenient fiction for modeling vagueness and possibility. *IEEE Transactions on Fuzzy Systems*, 2(1) :16–21, February 1994.

[Dubois98] D. Dubois and H. Prade. Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, 4 :244–264, 1998.

[Dubois99] D. Dubois, H. Prade, and R. R. Yager. *Merging fuzzy information, Chap. 6 in Fuzzy Sets in Approximate Reasoning and Information Systems, Vol. 5 of the Handbooks of Fuzzy Sets*. ISBN 0792385845, 536 pages, J. C. Bezdek, D. Dubois and H. Prade Eds., Kluwer, Dordrecht, Netherlands, 1999.

[Duda73] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. ISBN 0471223611, 482 pages, John Wiley and Sons, Inc, New York, NY, USA, 1973.

[Enterface06] Enterface website : the SIMILAR NoE Summer Workshop on Multimodal Interfaces. http ://www.enterface.net/. *ENTERFACE*, 2006.

[Essa94] I. A. Essa, T. J. Darrell, and A. P. Pentland. Tracking facial motion. *Proceedings of the IEEE International Workshop on Motion of Non-Rigid and Articulated Objects*, pages 36–42, November 1994.

[Fernandez01] C. Fernandez, C. Larabi, and N. Richard. Influence des espaces de représentation de la couleur et du système de codage dans le cadre du développement de JPEG2000. *SIG-Vision, École de printemps à Pau*, 2001.

[Ford98] A. Ford and A. Roberts. Colour space conversions. *Equations FAQ, Available online from ftp ://wmin.ac.uk/put/itrg/*, August 1998.

[Freeman96] W. T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma. Computer vision for computer games. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 100–105, October 1996.

[French85] S. French. *Group consensus probability distributions : a critical survey, in Bayesian Statistics*, volume 2. ISBN 0340814055, 352 pages, J. Bernardo et al. Eds., Elsevier, Amsterdam, Netherlands, 1985.

[Fujiyoshi98] H. Fujiyoshi and A. J. Lipton. Real-time human motion analysis by image skeletonisation. *Proceedings of the IEEE International Workshop on Applications of Computer Vision*, pages 15–21, October 1998.

[Galata01] A. Galata, N. Johnson, and D. Hogg. Learning variable-length Markov models of behavior. *Computer Vision and Image Understanding*, 81(3) :398–413, March 2001.

[Gavrila95] D. M. Gavrila and L. S. Davis. Towards 3D model-based tracking and recognition of human movement. *Proceedings of the IEEE International Workshop on Automatic Face and Gesture Recognition*, pages 272–277, June 1995.

[Gavrila96] D. M. Gavrila and L. S. Davis. 3D model-based tracking of humans in action : a multi-view approach. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 73–80, June 1996.

[Gavrila99] D. M. Gavrila. The visual analysis of human movement : a survey. *Computer Vision and Image Understanding*, 73(1) :82–98, January 1999.

[Gehrig03] N. Gehrig, V. Lepetit, and P. Fua. Golf club visual tracking for enhanced swing analysis. *Proceedings of the British Machine Vision Conference*, September 2003.

[Gelb74] A. Gelb. *Applied optimal estimation.* ISBN 0262570483, 382 pages, MIT Press, Cambridge, MA, USA, 1974.

[Geman84] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6) :721–741, November 1984.

[Gianni03] F. Gianni, B. Lenseigne, and P. Dalle. Estimation mono-vue de la posture du bras en utilisant un modèle biomécanique. *Proceedings of ORASIS, Congrès francophone de vision par ordinateur et de traitement d'images*, pages 127–136, May 2003.

[Girondel02] V. Girondel. Détection de peau, suivi de tête et de mains pour des applications multimédia. Master's thesis, Institut National Polytechnique de Grenoble, July 2002.

[Girondel06] V. Girondel, A. Caplier, and L. Bonnaud. A human body analysis system. *European Journal on Applied Signal Processing - EURASIP-JASP*, Article ID 61927, 18 pages, to appear, 2006.

[Goncalves95] L. Goncalves, E. D. Bernardo, E. Ursulla, and P. Perona. Monocular tracking of the human arm in 3D. *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–770, June 1995.

[Goncalves98] L. Goncalves, E. D. Bernardo, and P. Perona. Reach out and touch space (motion learning). *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 234–239, April 1998.

[Grabisch95] M. Grabisch. Fuzzy integral in multicriteria decision making. *Fuzzy Sets and Systems*, 69(3) :279–298, February 1995.

[Grewal01] M. S. Grewal and A. P. Andrews. *Kalman filtering theory and practise using MATLAB (2nd Ed.).* ISBN 0471392545, 416 pages, John Wiley and Sons, Inc, New York, NY, USA, 2001.

[Guo94a] Y. Guo, G. Xu, and S. Tsuji. Tracking human body motion based on a stick figure model. *Journal of Visual Communications and Image Representation*, 5(1) :1–9, 1994.

[Guo94b] Y. Guo, G. Xu, and S. Tsuji. Understanding human motion patterns. *Proceedings of the IEEE International Conference on Pattern Recognition*, 2 :325–329, January 1994.

[Hammal05] Z. Hammal, L. Couvreur, A. Caplier, and M. Rombaut. Facial expression recognition based on the belief theory : comparison with different classifiers. *Proceedings of the International Conference on Image Analysis and Processing*, pages 743–752, September 2005.

[Harasse06] S. Harasse and L. Bonnaud. Human model for people detection in dynamic scenes. *Proceedings of the IEEE International Conference on Pattern Recognition*, to appear, August 2006.

[Haritaoglu98] I. Haritaoglu, D. Harwood, and L. S. Davis. W^4 : Who ? When ? Where ? What ? A real time system for detecting and tracking people. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 222–227, April 1998.

[Haykin99] S. Haykin. *Neural networks : a comprehensive introduction*. ISBN 0139083855, 842 pages, Prentice Hall Eds., Englewood Cliffs, NJ, USA, 1999.

[Heap96] T. Heap and D. Hogg. 3D deformable hand models. *Proceedings of the Gesture Workshop*, pages 131–139, March 1996.

[Hernandez03] P. C. Hernandez, F. Marques, and X. Marichal. 3D posture estimation using geodesic distance maps. *Proceedings of the International Workshop on Gesture and Sign Language based Human-Computer Interaction*, April 2003.

[Hertz91] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the theory of neural computation*. ISBN 0201515601, 352 pages, Addison-Wesley Eds., Reading, MA, USA, 1991.

[Hjelmäs01] E. Hjelmäs and B. K. Low. Face detection : a survey. *Computer Vision and Image Understanding*, 83(3) :236–274, September 2001.

[Hogg83] D. Hogg. Model-based vision : a program to see a walking person. *Image and Vision Computing*, 1(1) :5–20, February 1983.

[Horain02] P. Horain and M. Bomb. 3D model based gesture acquisition using a single camera. *Proceedings of the IEEE International Workshop on Applications of Computer Vision*, pages 158–162, December 2002.

[Huang99] J. Huang, S. R. Kumar, M. Mitra, W. J. Zhu, and R. Zabih. Spatial color indexing and applications. *International Journal of Computer Vision*, 35(3) :245–268, December 1999.

[Huber96] E. Huber. 3D real-time gesture recognition using proximity spaces. *Proceedings of the IEEE International Workshop on Applications of Computer Vision*, pages 136–141, December 1996.

[Hunter97] E. A. Hunter, P. H. Kelly, and R. C. Jain. Estimation of articulated motion using kinematically constrained mixture densities. *Proceedings of the IEEE International Workshop on Motion of Non-Rigid and Articulated Objects*, pages 10–17, June 1997.

[Isard96] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. *Proceedings of the European Conference on Computer Vision*, 1 :343–356, April 1996.

[Isard98] M. Isard and A. Blake. Condensation : unifying low-level and high-level tracking in a stochastic framework. *Proceedings of the European Conference on Computer Vision*, pages 893–908, June 1998.

[Itoh97] M. Itoh and T. Inagaki. Combination and updating for belief revision in the theory of evidence. *Proceedings of the Scandinavian Conference on Artificial Intelligence*, pages 71–82, August 1997.

[Iwai99] Y. Iwai, K. Ogaki, and M. Yachida. Posture estimation using structure and motion models. *Proceedings of the IEEE International Conference on Computer Vision*, 1 :214–219, September 1999.

[Iwasawa97] S. Iwasawa, K. Ebihara, J. Ohya, and S. Morishima. Real-time estimation of human body posture from monocular thermal images. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 15–20, June 1997.

[Jacobs93] O. L. Jacobs. *Introduction to control theory (2nd Ed.)*. ISBN 0198562497, 402 pages, Oxford University Press, New York , NY, USA, 1993.

[Jacquin95] A. Jacquin and A. Eleftheriadis. Automatic face location detection and tracking for model-assisted coding of video teleconferencing sequences at low bit-rates. *Signal Processing : Image Communication*, 7(3) :231–248, September 1995.

[Jain79] R. C. Jain and H. H. Nagel. On the analysis of accumulative difference pictures from image sequences of real world scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2) :206–213, April 1979.

[Jang00] D. S. Jang and H. I. Choi. Active models for tracking moving objects. *Pattern Recognition*, 33(7) :1135–1146, July 2000.

[Jebara97] T. S. Jebara and A. P. Pentland. Parameterized structure from motion for 3D adaptive feedback tracking of faces. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 144–150, June 1997.

[Johansson76] G. Johansson. Visual motion perception. *Scientific American*, 232(6) :75–88, June 1976.

[Ju96] S. Ju, M. J. Black, and Y. Yacoob. Cardboard people : a parameterized model of articulated image motion. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 38–44, October 1996.

[Kakadiaris94] I. A. Kakadiaris, D. Metaxas, and R. Bajcsy. Active part-decomposition, shape and motion estimation of articulated objects : a physics-based approach. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 980–984, June 1994.

[Kakadiaris95] I. A. Kakadiaris and D. Metaxas. 3D human body model acquisition from multiple views. *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–623, June 1995.

[Kakadiaris96] I. A. Kakadiaris and D. Metaxas. Model based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 81–87, June 1996.

[Kakadiaris98] I. A. Kakadiaris and D. Metaxas. Vision based animations of digital humans. *Proceedings of the IEEE International Conference on Computer Animation*, pages 144–152, June 1998.

[Kalman60] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering*, D-82 :35–45, March 1960.

[Kameda96] Y. Kameda and M. Minoh. A human motion estimation method using 3 successive video frames. *Proceedings of the International Conference on Virtual Systems and Multimedia*, pages 135–140, September 1996.

[Karaulova00] I. A. Karaulova, P. M. Hall, and A. D. Marshall. A hierarchical model of dynamics for tracking people with a single video camera. *Proceedings of the British Machine Vision Conference*, pages 352–361, September 2000.

[Kim01] M. Kim, J. B. G. Jeon, J. S. Kwak, M. H. Lee, and C. Ahn. Moving object segmentation in video sequences by user interaction and automatic object tracking. *Image and Vision Computing*, 19(5) :245–260, April 2001.

[Klir92] G. J. Klir and B. Parviz. Probability-possibility transformations : a comparison. *International Journal on General Systems*, 21(1) :291–312, 1992.

[Kojima01] K. Kojima, T. Otobe, M. Hironaga, and S. Nagae. Human motion analysis using the rhythm - An estimate method of dance motion with multivariate autoregressive model. *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication*, pages 194–199, September 2001.

[Koprinska01] I. Koprinska and S Carrato. Temporal video segmentation : a survey. *Signal Processing : Image Communication*, 16(5) :477–500, January 2001.

[Kullback59] S. Kullback. *Information Theory and Statistics*. ISBN 0486696847, 416 pages, John Wiley and Sons, Inc, New York, NY, USA, 1959.

[Kurakake92] S. Kurakake and R. Nevatia. Description and tracking of moving articulated objects. *Proceedings of the IEEE International Conference on Pattern Recognition*, 1 :491–495, August 1992.

[Köhle97] M. Köhle, D. Merkl, and J. Kastner. Clinical gait analysis by neural networks : issues and experiences. *Proceedings of the IEEE International Symposium on Computer-Based Medical Systems*, pages 138–143, June 1997.

[Lee05] D. S. Lee. Effective gaussian mixture learning for video background subtraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5) :827–832, May 2005.

[Lefèvre02] E. Lefèvre, O. Colot, and P. Vannoorenberghe. Belief function combination and conflict management. *Information Fusion*, 3(2) :149–162, June 2002.

[Lefèvre03] S. Lefèvre, J. Holler, and N. Vincent. A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval. *Real Time Imaging*, 9(1) :73–98, February 2003.

[Lepetit03] V. Lepetit, A. Shahrokni, and P. Fua. Robust data association for online applications. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 1 :281–288, June 2003.

[Leung94] M. K. Leung and Y. H. Yang. An empirical approach to human body motion analysis. *Tech. report VR-94-1, Dept. of Computer Science, University of Saskatchewan*, February 1994.

[Leung95] M. K. Leung and Y. H. Yang. First sight : a human body outline labeling system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(4) :359–377, April 1995.

[Lewis86] F. L. Lewis. *Optimal estimation with an introductory to stochastic control theory*. ISBN 0471837415, 400 pages, John Wiley and Sons, Inc, New York, NY, USA, 1986.

[Li98] Y. Li, S. Ma, and H. Lu. Human posture recognition using multi-scale morphological method and Kalman motion estimation. *Proceedings of the IEEE International Conference on Pattern Recognition*, 1 :175–177, August 1998.

[Lin99] C. T. Lin, H. W. Nein, and W. C. Lin. A space-time delay neural network for motion recognition and its application to lipreading. *International Journal on Neural Systems*, 9(4) :311–334, August 1999.

[Lipton98] A. J. Lipton, H. Fujiyoshi, and R. S. Patil. Moving target detection and classification from real-time video. *Proceedings of the IEEE International Workshop on the Applications of Computer Vision*, pages 8–14, October 1998.

[Long90] W. Long and Y. H. Yang. Stationary background generation : an alternative to the difference of two images. *Pattern Recognition*, 23(12) :1351–1359, October 1990.

[Long91] W. Long and Y. H. Yang. Log-tracker : an attribute-based approach to tracking human motion. *Pattern Recognition and Artificial Intelligence*, 5 :439–458, 1991.

[Maes97] P. Maes, T. J. Darrell, B. Blumberg, and A. P. Pentland. The Alive system : wireless, full-body interaction with autonomous agents. *ACM Multimedia Systems*, 5(2) :105–112, March 1997.

[Marcel00a] S. Marcel, O. Bernier, J. E. Viallet, and D. Collobert. Hand gesture recognition using input-output hidden Markov models. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 456–461, March 2000.

[Marcel00b] S. Marcel, O. Bernier, and D. Collobert. Approche EM pour la construction de régions de teinte homogènes : application au suivi du visage et des mains d'une personne. *CORESA2000*, October 2000.

[Marichal03] X. Marichal and T. Umeda. Real-time segmentation of video objects for mixed-reality interactive applications. *Proceedings of the SPIE International Conference on Visual Communication and Image Processing*, 5150 :41–50, July 2003.

[Mascle97] S. Mascle, I. Bloch, and D. Vidal-Madjar. Application of Dempster-Shafer evidence theory to unsupervised classification in multisource remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 35(4) :1018–1031, July 1997.

[Maybeck79] P. S. Maybeck. *Stochastic models, estimation, and control (mathematics in science and Engineering)*, volume 1. ISBN 0124807011, 442 pages, Academic Press, 1979.

[Maître96] H. Maître. *Entropy, Information and Image in Progress in Picture Processing*. ISBN 0444824073, 376 pages, H. Maître and J. Zinn-Justin Eds., Springer Verlag, Les Houches Session LVIII, 1996.

[McKenna98] S. J. McKenna, S. Gong, and Y. Raja. Modelling facial colour and identity with gaussian mixtures. *Pattern Recognition*, 31(12) :1883–1892, December 1998.

[McKenna99] S. J. McKenna, Y. Raja, and S. Gong. Tracking colour objects using adaptive mixture models. *Image and Vision Computing*, 17(3-4) :225–231, March 1999.

[McKenna00a] S. J. McKenna, S. Jabri, Z. Duric, and H. Wechsler. Tracking interacting people. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 348–353, March 2000.

[McKenna00b] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80(1) :42–56, October 2000.

[McKenna03] S. J. McKenna, F. Marquis-Faulkes, A. F. Newell, and P. Gregor. Scenario-based drama as a tool for investigating user requirements with application to home monitoring for elderly people. *Proceedings of the International Conference on Human-Computer Interaction*, pages 512–516, 2003.

[Menger42] K. Menger. Statistical metrics. *Proceedings of the National Academy of Sciences, USA*, 28 :535–537, December 1942.

[Meyer97] D. Meyer, J. Denzler, and H. Niemann. Model based extraction of articulated objects in image sequences for gait analysis. *Proceedings of the IEEE International Conference on Image Processing*, 3 :78–81, October 1997.

[Milisavljevic03] N. Milisavljevic and I. Bloch. Sensor fusion in anti-personnel mine detection using a two-level belief function model. *IEEE Transactions on Systems, Man and Cybernetics*, 33(2) :269–283, May 2003.

[Mitiche96] A. Mitiche and P. Bouthemy. Computation and analysis of image motion : a synopsis of current problems and methods. *International Journal of Computer Vision*, 19(1) :29–55, July 1996.

[Moeslund01] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3) :231–268, March 2001.

[Moghaddam98] B. Moghaddam, W. Wahid, and A. P. Pentland. Beyond eigenfaces : probabilistic matching for face recognition. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 30–35, April 1998.

[Mokhber05] A. Mokhber, C. Achard, X. Qu, and M. Milgram. Action recognition with global features. *Proceedings of the IEEE International Workshop on Human-Computer Interaction*, pages 110–119, October 2005.

[Mostafaoui05] G. Mostafaoui, C. Achard, and M. Milgram. Real-time tracking of multiple persons on colour image sequences. *Proceedings of the Conference on Advanced Concepts for Intelligent Vision Systems*, pages 44–51, September 2005.

[Mottin00] N. Mottin. Localisation de visages dans des images acquises avec différents cadrages de caméra. Application à l'indexation. Master's thesis, Université de Nice Sophia-Antipolis, June 2000.

[Murphy00] C. K. Murphy. Combining belief functions when evidence conflicts. *Decision Support Systems*, 29(1) :1–9, July 2000.

[Myers80] C. Myers, L. R. Rabiner, and A. Rosenberg. Performance tradeoffs in dynamic time warping algorithms for word recognition. *IEEE Transactions on Applied Speech and Signal Processing*, 28(6) :623–635, 1980.

[Nagel78] H. H. Nagel. Formation of an object concept by analysis of systematic time variations in the optically perceptible environment. *Computer Graphics and Image Processing*, 7(2) :149–194, April 1978.

[Nair02] V. Nair and J. J. Clark. Automated visual surveillance using hidden Markov models. *Vision Interface*, pages 88–94, May 2002.

[Nait-Charif04] H. Nait-Charif and S. J. McKenna. Activity summarisation and fall detection in a supportive home environment. *Proceedings of the IEEE International Conference on Pattern Recognition*, 4 :323–326, August 2004.

[Nakazawa98] A. Nakazawa, H. Kato, and S. Inokuchi. Human tracking using distributed vision systems. *Proceedings of the IEEE International Conference on Pattern Recognition*, 1 :593–596, August 1998.

[Nirei96] K. Nirei, H. Saito, M. Mochimaru, and S. Ozawa. Human hand tracking from binocular image sequences. *Proceedings of the International Conference on Industrial Electronics, Control and Instrumentation*, pages 297–302, August 1996.

[Niyogi94a] S. A. Niyogi and E. H. Adelson. Analysing and recognizing walking figures in XYT. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 469–474, June 1994.

[Niyogi94b] S. A. Niyogi and E. H. Adelson. Analysing gait with spatiotemporal surfaces. *Proceedings of the IEEE International Workshop on Motion of Non-Rigid and Articulated Objects*, pages 64–69, November 1994.

[Noury03] N. Noury, P. Barralon, G. Virone, P. Rumeau, and P. Boissy. Maison intelligente pour personnes âgées. *Journée Image et Signal pour le Handicap*, October 2003.

[Noury04] N. Noury, P. Barralon, G. Virone, P. Rumeau, and P. Boissy. Un capteur intelligent pour détecter la chute - Fusion multicapteurs et détection à base de règles. *C2I*, January 2004.

[Ong05] S. C. W. Ong and S. Ranganath. Automatic sign language analysis : a survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6) :873–891, June 2005.

[O'Rourke80] J. O'Rourke and N. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(6) :522–536, November 1980.

[Paragios00] N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3) :266–280, March 2000.

[Pavlovic97] V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction : a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7) :677–695, July 1997.

[Peng05a] K. Peng, L. Chen, and S. Ruan. A novel scheme of face verification using active appearance models. *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 247–252, September 2005.

[Peng05b] K. Peng, L. Chen, S. Ruan, and G. Kukharev. A robust algorithm for eye detection on gray intensity face without spectacles. *Journal of Computer Science and Technology*, 5(3) :127–132, October 2005.

[Pentland00] A. P. Pentland. Looking at people : sensing for ubiquitous and wearable computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1) :107–119, January 2000.

[Peterfreund99] N. Peterfreund. Robust tracking of position and velocity with Kalman snakes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6) :564–569, June 1999.

[Polana94] R. Polana and R. Nelson. Low level recognition of human motion. *Proceedings of the IEEE International Workshop on Motion of Non-Rigid and Articulated Objects*, pages 77–82, October 1994.

[Poritz88] A. B. Poritz. Hidden Markov models : a guided tour. *Proceedings of the IEEE International Conference on Acoustic Speech and Signal Processing*, pages 7–13, May 1988.

[Price06] K. Price. http ://iris.usc.edu/Vision-Notes/bibliography/contents.html. *Annotated Computer Vision Bibliography*, 2006.

[Rabiner89] L. R. Rabiner. A tutorial on hidden Markov models and selected applications. *Proceedings of the IEEE*, 77(2) :257–285, February 1989.

[Rashid80] R. F. Rashid. Towards a system for the interpretation of moving light displays. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(6) :574–581, November 1980.

[Rehg93] J. M. Rehg and T. Kanade. DigitEyes : vision-based human hand tracking. *Tech. report CMU-CS-93-220, Dept. of Computer Science, Carnegie Mellon University*, December 1993.

[Rehg94] J. M. Rehg and T. Kanade. Visual tracking of high dof articulated structures : an application to human hand tracking. *Proceedings of the European Conference on Computer Vision*, B :35–46, May 1994.

[Rehg95] J. M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. *Proceedings of the IEEE International Conference on Computer Vision*, pages 612–617, June 1995.

[Rigoll00] G. Rigoll, S. Eickeler, and S. Müller. Person tracking in real world scenarios using statistical methods. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 342–347, March 2000.

[Rodriguez05] T. Rodriguez, I. Reid, R. Horaud, N. Dalal, and M. Goetz. Image interpolation for virtual sports scenarios. *Machine Vision and Applications*, 16(4) :236–245, September 2005.

[Rohr94] K. Rohr. Towards model-based recognition of human movement in image sequences. *Computer Vision Graphics and Image Processing : Image Understanding*, 59(1) :94–115, January 1994.

[Rombaut01] M. Rombaut. Fusion : état de l'art et perspectives. *Tech. report DSP 99.60.078, Rapport DGA*, October 2001.

[Rombaut02] M. Rombaut and Y. M. Zhu. Study of Dempster-Shafer theory for image segmentation applications. *Image and Vision Computing*, 20(1) :15–23, 2002.

[Rosales98] R. Rosales. Recognition of human action using moment-based features. *Tech. report BU 98-020, Dept. of Computer Science, Boston University*, November 1998.

[Rosales00] R. Rosales and S. Sclaroff. Learning and synthesizing human body motion and posture. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 506–511, March 2000.

[Rosenblum94] M. Rosenblum, Y. Yacoob, and L. Davis. Human emotion recognition from motion using a radial basis function network architecture. *Proceedings of the IEEE International Workshop on Motion of Non-Rigid and Articulated Objects*, pages 43–49, November 1994.

[Rowley97] H. A. Rowley and J. M. Rehg. Analyzing articulated motion using Expectation-Maximization. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 935–941, June 1997.

[Salvador01] E. Salvador, A. Cavalarro, and T. Ebrahimi. Shadow identification and classification using invariant color models. *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, pages 1545–1548, May 2001.

[Sangi04] P. Sangi, J. Heikkila, and O. Silven. Motion analysis using frame differences with spatial gradient measures. *Proceedings of the IEEE International Conference on Pattern Recognition*, 4 :733–736, August 2004.

[Saporta90] G. Saporta. *Probabilités, Analyse de données statistiques.* ISBN 2710805650, 193 pages, Ed. Technip, Paris, 1990.

[Schweizer83] B. Schweizer and A. Sklar. *Probabilistic metric spaces.* ISBN 0486445143, 313 pages, North Holland, Amsterdam, 1983.

[Schwerdt00] K. Schwerdt and J. L. Crowley. Robust face tracking using color. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 90–95, March 2000.

[Segen96] J. Segen and G. S. Pingali. A camera-based system for tracking people in real time. *Proceedings of the IEEE International Conference on Pattern Recognition*, 3 :63–67, August 1996.

[Seki03] M. Seki, T. Wada, H. Fujiwara, and K. Sumi. Background subtraction based on cooccurrence of image variations. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2 :65–72, June 2003.

[Shafer76] G. Shafer. *A mathematical theory of evidence*, volume 2702. ISBN 069110042X, 297 pages, Princeton University Press, Princeton, NJ, USA, 1976.

[Shio91] A. Shio and J. Sklansky. Segmentation of people in motion. *Proceedings of the IEEE International Workshop on Visual Motion*, pages 325–332, October 1991.

[Siebel04] N. T. Siebel and S. J. Maybank. The Advisor visual surveillance system. *Proceedings of the European Workshop on the Applications of Computer Vision*, pages 103–111, 2004.

[Silverman86] B. W. Silverman. *Density estimation for statistics and data analysis, Monographs on statistics and applied probability*, volume 26. ISBN 0412246201, 176 pages, Chapman and Hall Eds., Springer Verlag, 1986.

[Similar05] SIMILAR Network of excellence website : the European taskforce creating human-machine interfaces similar to human-human communication. http ://www.similar.cc/. *SIMILAR*, 2005.

[Smets90a] P. Smets. The combination of evidence in the transferable belief model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5) :447–458, May 1990.

[Smets90b] P. Smets. Constructing the pignistic probability function in a context of uncertainty. *Uncertainty in Artificial Intelligence*, 5 :29–40, 1990.

[Smets93] P. Smets. Belief functions : the disjonctive rule of combination and the generalized bayesian theorem. *International Journal of Approximate Reasoning*, 9 :1–35, 1993.

[Smets94] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66(2) :191–234, 1994.

[Smets98] P. Smets. *The transferable belief model for quantified belief representation, in Handbook of Defeasible Reasoning and Uncertainty Management Systems*, volume 1. ISBN 0792351002, 484 pages, D. M. Gabbay and P. Smets Eds., Kluwer, Doordrecht, The Netherlands, 1998.

[Sminchisescu03] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3D human tracking. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 1 :69–76, June 2003.

[Sobottka96] K. Sobottka and I. Pitas. Extraction of facial regions and features using color and shape information. *Proceedings of the IEEE International Conference on Pattern Recognition*, 3 :421–425, August 1996.

[Sobottka98] K. Sobottka and I. Pittas. A novel method for automatic face segmentation, facial features extraction and tracking. *Signal Processing : Image Communication*, 12(3) :236–281, June 1998.

[Sorenson70] H. W. Sorenson. Least-squares estimation : from Gauss to Kalman. *IEEE Spectrum*, 7 :63–68, July 1970.

[Starner95a] T. E. Starner and A. P. Pentland. Visual recognition of american sign language recognition using hidden Markov models. *Proceedings of the IEEE International Workshop on Automatic Face and Gesture Recognition*, pages 189–194, June 1995.

[Starner95b] T. E. Starner and A. P. Pentland. Real-time american sign language recognition from video using hidden Markov models. *Proceedings of the IEEE International Symposium on Computer Vision*, pages 265–270, November 1995.

[Takahashi94] K. Takahashi, S. Seki, H. Kojima, and R. Oka. Recognition of dexterous manipulations from time-varying images. *Proceedings of the IEEE International Workshop on Motion of Non-Rigid and Articulated Objects*, pages 23–28, November 1994.

[Terrillon99] J. C. Terrillon and S. Akamatsu. Comparative performance of different chrominance spaces for colour segmentation and detection of human faces in complex scene images. *Proceedings of the Conference on Vision Interface*, 2 :180–187, May 1999.

[Thirde05] D. Thirde, M. Borg, J. M. Ferryman, V. Valentin, F. Fusier, F. Brémond, M. Thonnat, J. Aguilera, and M. Kampel. Visual surveillance for aircraft activity monitoring. *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 255–262, October 2005.

[Tian99] Y. L. Tian, T. Kanade, and J. F. Cohn. Multi-state based facial feature tracking and detection. *Tech. report CMU-RI-TR-99-18, Robotics Institute, Carnegie Mellon University*, August 1999.

[Tsekeridou01] S. Tsekeridou and I. Pitas. Content-based video parsing and indexing based on audio-visual interaction. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(4) :522–535, April 2001.

[Tupin99] F. Tupin, I. Bloch, and H. Maître. A first step towards automatic interpretation of SAR images using evidential fusion of several structures detectors. *IEEE Transactions on Geoscience and Remote Sensing*, 37(3) :1327–1343, May 1999.

[Umeda01] T. Umeda, D. Douxchamps, X. Wielemans, N. Alok, and X. Marichal. Real-time interactive immersion. *Proceedings of the International Symposium on Mixed Reality*, pages 177–178, March 2001.

[Umeda04] T. Umeda, P. C. Hernandez, F. Marques, and X. Marichal. A real-time body analysis for mixed reality application. *Proceedings of the Korea-Japan Joint Workshop on Frontiers of Computer Vision*, February 2004.

[Viallet98] J. E. Viallet, M. Collobert, R. Féraud, and O. Bernier. Panorama : a what I see is what I want contactless visual interface. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 568–572, April 1998.

[Von Luschan27] F. Von Luschan. *Voelker, Rassen, Sprachen : Anthropologische Betrachtungen*. 382 pages, Deutsche Buchgemeinshaft, Berlin, 1927.

[Wachter97] S. Wachter and H. H. Nagel. Tracking of persons in monocular image sequences. *Proceedings of the IEEE International Workshop on Motion of Non-Rigid and Articulated Objects*, pages 2–9, June 1997.

[Wachter99] S. Wachter and H. H. Nagel. Tracking persons in monocular image sequences. *Computer Vision and Image Understanding*, 74(3) :174–192, June 1999.

[Wang03a] L. Wang, W. M. Hu, and T. N. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36(3) :585–601, March 2003.

[Wang03b] J. J. Wang and S. Singh. Video analysis of human dynamics : a survey. *Real-Time Imaging*, 9(5) :321–346, October 2003.

[Webb81] J. A. Webb and J. K. Aggarwal. Visually interpreting the motion of objects in space. *IEEE Computer*, 14(8) :40–46, August 1981.

[Webb82] J. A. Webb and J. K. Aggarwal. Structure from motion of rigid and jointed objects. *Artificial Intelligence*, 19(1) :107–130, September 1982.

[Weik00] S. Weik, J. Wingbermühle, and W. Niem. Creation of flexible anthropomorphic models for 3D videoconferencing using shape from silhouettes. *Journal of Visualization and Computer Animation*, 11(3) :145–154, July 2000.

[Wingbermühle98] J. Wingbermühle and S. Weik. Towards automatic creation of realistic anthropomorphic models for real time 3D telecommunication. *Journal of VLSI Signal Processing Systems*, 20(1-2) :81–96, October 1998.

[Wren97a] C. R. Wren, F. Sparacino, A. J. Azarbayejani, T. J. Darrell, T. E. Starner, A. Kotani, C. M. Chao, M. Hlavac, K. B. Russell, and A. P. Pentland. Perceptive spaces for performance and entertainment : untethered interaction using computer vision and audition. *Applied Artificial Intelligence*, 11(4) :267–284, June 1997.

[Wren97b] C. R. Wren, A. Azarbayejani, T. J. Darrell, and A. P. Pentland. Pfinder : real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7) :780–785, July 1997.

[Wren00] C. R. Wren, B. P. Clarkson, and A. P. Pentland. Understanding purposeful human motion. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 378–383, March 2000.

[Yacoob98] Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. *Proceedings of the IEEE International Conference on Computer Vision*, pages 120–127, January 1998.

[Yager87] R. R. Yager. On the Dempster-Shafer framework and new combination rules. *Information Sciences*, 41(2) :93–138, March 1987.

[Yager88] R. R. Yager. On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics*, 18(1) :183–190, January 1988.

[Yager91] R. R. Yager. Connectives and quantifiers in fuzzy sets. *Fuzzy Sets and Systems*, 40(1) :39–75, March 1991.

[Yalamanchili82] S. Yalamanchili, W. N. Martin, and J. K. Aggarwal. Extraction of moving object descriptions via differencing. *Computer Graphics and Image Processing*, 18(2) :188–201, February 1982.

[Yamada98] M. Yamada, K. Ebihara, and J. Ohya. A new robust real-time method for extracting human silhouettes from color images. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 528–533, April 1998.

[Yamato92] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov model. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 379–385, June 1992.

[Yang95] J. Yang and A. Waibel. Tracking human faces in real time. *Tech. report CMU-CS-TR-95-210, Dept. of Computer Science, Carnegie Mellon University*, November 1995.

[Yang96] J. Yang and A. Waibel. A real-time face tracker. *Proceedings of the IEEE International Workshop on Applications of Computer Vision*, pages 142–147, December 1996.

[Yang98] J. Yang, W. Lu, and A. Waibel. Skin color modeling and adaptation. *Proceedings of the Asian Conference on Computer Vision*, 2 :687–694, January 1998.

[Yang02] M. H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images : a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1) :34–58, January 2002.

[Yoo99] T. W. Yoo and I. S. Oh. A fast algorithm for tracking human faces based on chromatic histograms. *Pattern Recognition Letters*, 20 :967–978, October 1999.

[Zadeh65] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8 :338–353, June 1965.

[Zadeh78] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1 :3–28, 1978.

[Zadeh86] L. A. Zadeh. A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. *Artificial Intelligence Magazine*, 7(2) :85–90, Summer 1986.

[Zhang01] D. Zhang and G. Lu. Segmentation of moving objects in image sequence : a review. *Circuits, Systems and Signal Processing*, 20(2) :143–183, March 2001.

[Zhong00] Y. Zhong, A. K. Jain, and M. P. Dubuisson-Jolly. Object tracking using deformable templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(5) :544–549, May 2000.

Résumé

Le travail de recherche présenté dans ce mémoire de thèse est dédié à l'analyse et à l'interprétation du mouvement humain avec application à la reconnaissance de postures. L'analyse et l'interprétation du mouvement humain en vision par ordinateur ont de nombreux domaines d'applications tels que la vidéosurveillance, les applications de réalité mixte et les interfaces homme-machine avancées. Nous proposons ici un système temps-réel permettant une analyse et une interprétation du mouvement humain.

L'analyse du mouvement humain fait intervenir plusieurs processus de traitement d'images tels que la segmentation d'objets en mouvement, le suivi temporel, la détection de peau, les modèles de corps humain et la reconnaissance d'actions ou de postures. Nous proposons une méthode de suivi temporel en deux étapes permettant de suivre au cours du temps une ou plusieurs personnes même si elles s'occultent entre elles. Cette méthode est basée sur un calcul d'intersection de boîtes englobantes rectangulaires et sur un filtrage partiel de Kalman. Puis nous explicitons une méthode de détection de peau par une approche couleur afin de localiser leurs visages et leurs mains. Toutes ces étapes préliminaires donnent accès à de nombreuses informations bas-niveau. Dans une dernière partie, nous utilisons une partie de ces informations pour reconnaître les postures statiques de personnes parmi les quatre postures suivantes : debout, assis, accroupi et couché. De nombreux résultats illustrent les avantages et les limitations des méthodes proposées, ainsi que leur efficacité et robustesse.

Abstract

This Ph.D. thesis research work is dedicated to the analysis and the interpretation of human motion with an application to posture recognition. Human motion analysis and interpretation in computer vision have numerous applications domains such as videosurveillance, mixed-reality applications and advanced man-machine interfaces. We propose here a real-time system that allows human motion analysis and interpretation.

Human motion analysis includes several processing steps of image processing such as segmentation of moving objects, temporal tracking, skin detection, human body models, and actions or pose recognition. We propose a temporal tracking method in two stages that allows to track one or several persons even if they occlude each other. This method is based on the computation of bounding boxes overlap and a partial Kalman filtering. Then we explicit a skin detection method by a color approach in order to localize their faces and hands. All these preliminary steps give access to a lot of low-level data. In a last part, we use some of these data to perform static human body posture recognition of people among the four following postures : standing, sitting, squatting and lying. Several results illustrate the advantages and limitations of the proposed methods, as their efficiency and robustness.

Mots-clés : Analyse, détection de visage, filtrage de Kalman, interprétation, mouvement humain, reconnaissance de postures, suivi temporel, temps-réel, théorie de l'évidence.

Laboratoire des Images et des Signaux
ENSIEG, Domaine Universitaire, BP 46,
38402 St-Martin-d'Hères Cedex, France