Modèle du corps humain pour le suivi de gestes en monoculaire

Philippe Noriega

Département Sciences et Technologies
de l'Information et de l'Ingénierie
UFR Sciences de l'Ingénieur

École doctorale SMAE de Paris

Modèle du corps humain pour le suivi de gestes en monoculaire

THÈSE

présentée et soutenue publiquement le 11 octobre 2007

pour l'obtention du

Doctorat de l'université Pierre et Marie Curie – Paris 6

(spécialité informatique)

par

Philippe Noriega

Composition du jury

Président :		
Rapporteurs :	James Crowley	INPG
	Jack-Gérard Postaire	LAGIS
Examinateurs :	Catherine Achard	Paris VI
	Olivier Bernier	Orange Labs.
	Maurice Milgram	Paris VI

Institut des Systèmes Intelligents et Robotique — FRE 2507

Mis en page avec la classe thloria.

Remerciements

Résumé

L'estimation de la pose du corps humain ou son suivi grâce à la vision par ordinateur se heurte à la difficulté d'explorer un espace de grande dimension. Les approches par apprentissage et particulièrement celles qui font appel aux régressions vers des espaces de dimension réduits comme les LLE [RS00] ou les GPLVM [Law03] permettent de résoudre cette difficulté dans le cas de gestes cycliques [UFF06] sans parvenir à généraliser le suivi pour des poses quelconques. D'autres techniques procèdent directement par la comparaison de l'image test avec une base d'apprentissage. Dans cet esprit, le PSH [SVD03] permet d'identifier rapidement un ensemble de poses similaires dans une grande base de données. Cependant, même en intégrant des techniques d'extrapolation qui permettent de générer d'autres poses à partir de celles apprises, les approches uniquement basées sur l'apprentissage ne parviennent généralement pas à couvrir de façon assez dense l'espace des poses [TSDD06].

D'autres voies consistent à mettre en œuvre une méthode déterministe ou stochastique. Les méthodes déterministes [PF03] fournissent souvent une solution sous-optimale en restant piégées sur un optimum local du fait des ambiguïtés issues de la vision monoculaire. Les approches stochastiques tentent d'explorer la probabilité a posteriori mais là encore, la grande dimension de l'espace des poses, notamment dans le cas des méthodes à base de simulation par échantillonnage, exige de multiplier le nombre des tirages afin d'avoir une chance d'explorer le mode dominant. Une solution intéressante consiste à utiliser un modèle de corps à membres indépendants [SBR+04] pour restreindre l'exploration aux sous espaces définis par les paramètres de chacun des membres. L'influence d'un membre sur les autres s'exprime grâce à la propagation des croyances [KFL01] pour fournir une solution cohérente.

Dans ce travail de thèse, cette dernière solution est retenue en l'associant au filtre à particules pour générer un espace discret où s'effectue la propagation des croyances [BCMC06]. Ce procédé est préférable à la modélisation paramétrique des messages par un échantillonneur de Gibbs, un procédé coûteux en ressources dérivé de l'algorithme PAMPAS [Isa03]. Parallèlement à cette solution, le développement d'un suivi robuste du haut du corps, même en 2D [NB07b], exige une fusion de plusieurs indices extraits de l'image. La vraisemblance des hypothèses émises vis-à-vis de l'image est évaluée à partir d'indices tirés des gradients et de la couleur combinés avec une soustraction de fond [NB06] et une détection du mouvement.

L'interprétation de la profondeur pour le passage en 3D constitue une difficulté majeure du suivi monoculaire. La fusion d'indices évoquée précédemment devient insuffisante pour contraindre la pose. Cependant, du fait des contraintes articulaires, l'espace réel des poses occupe un sous-espace très réduit dans l'espace théorique. Le codage de ces contraintes dans l'étape de propagation des croyances associé à la fusion d'indices permet alors d'aboutir à de bonnes performances, même dans les cas d'environnements non contraints (lumière, vêtements...) [NB07a].

Une meilleure gestion des occultations est mise en œuvre en ajoutant un terme de compatibilité des hypothèses basé sur l'apprentissage. Avec le modèle utilisé [SBR+04], ce sont des membres indépendants plutôt que des poses complètes qui sont stockées dans la base d'apprentissage. Ceci permet d'obtenir une couverture satisfaisante de l'espace des poses avec un nombre raisonnable d'exemples appris. La propagation des croyances assure un assemblage cohérent des membres pour arriver au résultat et le processus de sélection des exemples dans la base peut-être accéléré grâce au PSH [SVD03].

Abstract

Human body pose estimation or tracking using computer vision is a diffucult task owing to the high dimensionality of the pose space. Learning based aproachs, especially methods using embeding spaces such as LLE [RS00] or GPLVM [Law03] can cope with this difficulty but are restricted to cyclic motions [UFF06]. Other methods proceed in comparing a test image to a learning base. Among them, PSH [SVD03] is usefull to speed up the selection of a subset of nearest neighbours from large learning bases. However, even when pose regression is used to extrapolate new poses from the learned ones, a sufficient coverage of the pose space is difficult to reach with learning-based approaches [TSDD06].

Other ways consist in using deterministic or stochastic algorithms. The former kind of methods often provide suboptimal solutions because they get stuck on local minima owing to monocular vision ambiguities [PF03]. Stochastic approachs are used to explore the posterior probability function but once again, the high dimensionality of the pose space, especially in the case of simulation-based methods using sampling, requires a huge number of samples to explore the main mode. An interesting solution consists in using a loose-limbed body model [SBR+04] where the likelihood of each limb is evaluated independently. In this manner the dimension of the explored space is reduced to the number of *dof* of a limb. Influence between limbs is estimated by belief propagation [KFL01] to provide consistent body poses.

This last solution is adopted in this thesis in association with particle filtering to provide a discrete space where the beliefs are computed [BCMC06]. This method is prefered to a parametric modelling of beliefs using Gibbs sampler, a method derived from the PAMPAS algorithm [Isa03] involving heavy computational load. However, in addition to this solution, robust human body tracking, even in two dimensions [NB07b], requires to use several images cues. Thus, hypotheses likelihood is evaluated from gradient and color based cues combined with a background subtraction [NB06] and a motion detection.

A main difficulty in monocular 3D tracking is the depth estimation making the fusioned images cues mentioned before unable to constrain sufficiently the pose. However, owing to articulation constraints, the real pose space covered by human motion is much smaller than the theoritical one. Associating the fusioned images cues with articulation constraints implemented on the belief propagation step result in suitable algorithm performances even on unconstrained environments (light, clothes...) [NB07a].

A more efficient occlusion handling is provided adding a learning-based hypotheses compatibility term. With the used body model [SBR+04], the learning base consists in limbs exemplars instead of full body poses permitting a wider coverage of the pose space with the same amount of exemplars. Belief propagation provides consistent body poses and the selection of similar limbs from the learning base can be speeded-up by PSH [SVD03].

Table des matières

Table des figures

Introduction

Les yeux qui s'ouvrent

Novembre 1996, je suis ici depuis plus d'une semaine. Bien sûr, on a beau nous dire que c'est le type d'opération qui est maintenant bien maîtrisée mais quelque chose fait qu'on se trouve être un cas à part, un cas qui exige une attention particulière faisant que je suis encore ici, chez les trois cents aveugles à en croire les lettres gravées sur la pierre du portail d'entrée. Alors je trompe l'ennui explorant mon environnement, ces quelques mètres carrés autour desquels les murs s'élèvent habillés d'un papier peint figuratif qui parvient à retenir mon attention quelques minutes. Un moulin, des fagots de pailles, ces variation champêtres sont-elles là pour faire oublier la ville ? Si c'est le cas, je ne suis pas sûr que le but soit atteint. Le fait est que ce papier à été parcouru de long en large par mes yeux enregistrant les moindres détails des motifs et leur répétition dans l'espace. Je sors dans le couloir pour trouver l'animation qui y règne habituellement à cette heure de la matinée. Les voix qui descendent de la chambre noire se font bienveillantes afin d'obtenir des patients leur coopération pendant l'examen. Les affiches ont remplacé le papier peint des chambres et devant la plus colorée d'entre elles, un petit homme approche son œuil de la surface glacée, émerveillé qu'il est à la perception de cette explosion de formes et de couleurs que je trouvais pourtant si banale il y a quelques minutes. Cette scène se passait au bout du couloir C de l'hôpital d'ophtalmologie des quinze-vingt.

Il est vrai que sa capacité informative fait de la vue un sens à part. La lecture d'une image offre un ensemble de formes et de couleurs, comme pouvaient les montrer l'affiche du petit homme ou, à l'identique du papier peint un peu daté de la chambre d'hôpital, un arrangement spatial de motifs répétées selon une fréquence donnée. Ces caractéristiques en font un canal d'information extrêmement dense surtout si on y ajoute la dimension temporelle qui laisse percevoir les mouvements ou les changement de forme. Sevré depuis longtemps, le petit homme redécouvrait l'ivresse procurée par la vue avec fascination et un immense bonheur.

Les applications

Dans un domaine scientifique où l'on cherche à rendre les objets plus intelligents, on comprends la fascination qu'exerce la perception visuelle. C'est pourquoi elle se place au centre de nombreuses recherches.

Parmi les domaines de recherche qui touchent à la vision, le suivi de personnes dans les séquences d'images monoculaires consiste à déterminer, à partir d'un ordinateur et d'une caméra, la pose d'une personne le long d'un flux vidéo (fig. 1). Cette discipline est promise à de nombreuses applications et notamment, pour rester dans le domaine médical, l'analyse des mouvements en médecine orthopédique qui permet de détecter des anomalies dans le déplacement des patients et prévenir ainsi l'usure prématurée des articulations. Dans le domaine des arts et spectacles,

FIG. 1: *Le suivi en vision monoculaire par ordinateur. De l'acquisition des images par une webcam aux coordonnées 3D de la pose estimée.*

la mise au point de chorégraphies pourra passer par l'analyse des mouvements captés par des caméras au cours des séances de répétition. Dans le domaine de l'animation cinématographique, il deviendra possible d'animer les protagonistes virtuels avec des mouvements plus naturels à partir du résultat issus d'une capture de mouvement sur un double réel du personnage. Les télécommunications sont également friands de ces technologies afin de bouleverser les protocoles de dialogue avec la machine. Les interactions homme-machine pourront se faire plus naturellement par le biais de gestes plutôt que de manière conventionnelle avec claviers et souris. À l'instar des environnements virtuels qui connaissent un succès grandissant sur internet (fig. 2), la communication entre humains fera appel à des avatars qui pourront interagir et évoluer dans l'espace virtuel à partir des gestes bien réels des participants. L'implantation de caméras dans l'espace public connaît lui aussi un succès croissant avec la vidéo surveillance. Cette discipline consiste à surveiller des lieux comme les couloirs du métro pour y détecter des comportements dits suspects. Elle permet aussi d'analyser les allées et venues des badauds dans les rues à des fins commerciales, par exemple, pour savoir si un emplacement est compatible avec l'installation d'un commerce.

Les verrous techniques

La capacité à sélectionner les informations pertinentes au milieu d'un flux vidéo puis la façon d'interpréter ces informations constitue l'une des clés de la vision par ordinateur. Les difficultés dans le domaine du suivi de personnes sont nombreuses et complexes à dénouer.

La résolution d'un problème de suivi consiste à déterminer l'ensemble des inconnues qui paramètrent la pose humaine. Le corps humain possède de nombreuses articulations et le nombre de ces inconnues se compte par dizaines. À titre d'exemple le modèle du haut du corps choisi dans le cadre de cette thèse possède 38 paramètres articulaires. Classiquement, l'estimation de la pose consiste à explorer un espace qui possède autant de dimensions que d'inconnues et si

FIG. 2: *Environnement collaboratif de type Second Life où comment incarner celui que l'on rêve d'être.*

on ajoute les paramètres qui déterminent les proportions entre les membres d'une personne, on comprends aisément que le suivi est un problème difficile à contraindre.

Le choix entre une solution multicaméra ou monocaméra pour le suivi est influencé par les difficultés de mise en œuvre des techniques multicamera. Cependant, ces dernières possèdent l'avantage de fournir une information de profondeur et peuvent aussi résoudre des problèmes d'occultations lorsque les caméras sont réparties sur des angles de vue différents. Bien moins informative, l'image monoculaire génère des ambiguïtés sur l'interprétation de la profondeur puisqu'il n'existe pas de bijection entre l'espace des poses et celui des images qu'elles génèrent. Ceci se traduit par le fait que des poses différentes peuvent donner des projections identiques à l'image. De plus, en vision monoculaire, il est difficile d'évaluer la profondeur ou la rotation d'un membre autour d'un axe parallèle au plan de l'image.

La variabilité des paramètres extérieurs fait l'objet d'une difficulté supplémentaire pour le suivi. La diversité des apparences influencées par les vêtements portés ou la nature des cheveux, ainsi que les variations de lumière ou d'environnement exigent de trouver une solution capable de s'adapter aux nombreuses situations.

Il faut ajouter à la liste des verrous à lever : les non linéarités qui découlent des déformations issues des vêtements, les mouvements rapides et nombreux du corps humain et les occultations entre les membres ou du fait d'objets pour prendre la pleine mesure de la difficulté d'un problème de suivi.

Dans ce contexte, les objectifs choisis pour la thèse vont vers un suivi du haut du corps (mains, bras, torse et tête) en temps réel dans un environnement dénué d'aides au suivi (marqueurs de couleurs, fond uniformes etc) et sans contraintes d'éclairage ou vestimentaire. Ce problème est reconnu dans la littérature scientifique pour sa difficulté.

Le plan du mémoire

Le plan s'articule autour de la levée des difficultés en regard des objectifs fixés dans la thèse. L'identification de ces difficultés et les solutions qui y ont été apportés est le fruit d'une recherche

bibliographique fouillée qui à permis d'influencer les choix envisagés par la suite.

Le chapitre bibliographique explore les trois domaines qui constituent les principales étapes de la résolution d'un problème de suivi. En premier lieu, les indices extraits de l'image sont étudiés afin de dégager, de la quantité considérable d'informations contenues dans une image, les données les plus pertinentes. Vient ensuite la modélisation du corps humain pour aborder les problèmes de représentation des membres et la précision avec laquelle ils sont rendus ainsi que le taille de l'espace généré par ces modèles. Les outils avec lesquels sont manipulés les paramètres du modèle sont décrits dans la dernière partie de la bibliographie. Il s'agit là d'analyser les avantages et les inconvénients de chaque méthode afin d'en tirer des solutions destinées à remplir les objectifs fixés.

Une fois les choix arrêtés, le chapitre théorique détaille l'algorithme utilisé afin de justifier de la validité et de la rigueur du modèle choisi à la lumière des approximations envisagées. L'aspect temps de traitement est également abordé pour tenter de dégager une solution compatible avec l'objectif de traitement en temps réel.

De manière plus pratique, le troisième chapitre aborde la description du modèle de corps et la façon d'évaluer la validité des hypothèses générées. Les indices extraits de l'image qui servent à conduire cette évaluation y sont également décrits.

Le quatrième chapitre présente les résultats expérimentaux. La pertinence des solutions adoptées est évaluée qualitativement et quantitativement à partir de nombreuses comparaisons afin de comptabiliser les objectifs tenus ou pas. Une conclusion générale vient clôturer ce mémoire.

Chapitre 1

État de l'art

Sommaire

1.1 Introduction

L'estimation de la pose d'une personne à partir d'une image fixe ou son suivi dans une séquence vidéo consiste à déterminer les coordonnées planaires (cas 2D) ou spatiales (cas 3D) de tout ou partie des membres du corps pour chacune des images. C'est un objectif primordial en vision par ordinateur et les applications en sont nombreuses. Elles concernent : les interfaces homme machine, la communication au travers d'avatars, l'analyse technique des gestes artistiques et sportifs ou la surveillance des lieux.

La paramétrisation de la pose humaine, même simplifiée comprends plusieurs dizaines de paramètres. Les gestes parfois rapides, les vêtements amples, les occultations ou encore les changements dans la scène tels que les mouvements dans l'arrière plan ou les variations de luminosité, contribuent à faire de l'estimation ou du suivi de la pose un défi scientifique.

Ce chapitre fait la synthèse des solutions précédemment mises en œuvre pour tenter de résoudre ces difficultés. Les revues d'articles [Gav99], [MG01] et récemment [MHK06] apportent une information synthétique concernant le sujet de cette thèse et plus largement la vision par ordinateur.

1.1.1 Classement des approches

Il existe de nombreuses voies de classement des différentes approches pour l'estimation de la pose ou du suivi : la nature des caractéristiques extraites de l'image (couleurs, contours, textures), le nombre de caméras utilisées (approches monoculaires, stéréoscopiques ou multi-caméra), le type de modèle du corps (2D ou 3D, patchs rectangulaires, troncs de cônes, primitives ellipsoïdales), et la nature des méthodes utilisées pour l'estimation (stochastique ou probabiliste, mono ou multi-hypothèse, par apprentissage), en fonction des niveaux d'analyse dans l'image, tantôt ascendante, tantôt descendante.

La diversité des solutions existantes dessine un paysage hétéroclite d'où il est difficile de dégager des axes de classification universels. Cependant, la résolution d'un problème de vision se décompose généralement suivant ces trois étapes dans un ordre qui n'est pas forcément celui proposé ici :

- l'extraction des caractéristiques de l'image,
- la modélisation du corps humain,
- l'estimation des coordonnées des membres.

Pour mener à bien l'estimation, la dernière étape utilise une série d'outils qui peuvent êtres stochastique, déterministes, faire appel à un apprentissage...

Il est donc intéressant de confronter les différentes approches à la lumière des choix faits pour chacune de ces trois étapes. C'est pourquoi le plan de l'état de l'art qui suit reprend ces trois thématiques.

1.2 Caractéristiques exploitées dans l'image

Une image numérique est un arrangement planaire de pixels en couleurs ou en niveaux de gris. L'analyse spatiale de ces pixels dégage des contours, des textures ou permet de segmenter des zones homogènes. Au niveau d'une séquence d'images, lorsqu'elle est disponible, l'analyse temporelle est informative sur les mouvements au cours de la séquence. Le paragraphe qui suit tente d'établir un panorama des différentes sortes d'indices utilisées par les approches analysées au cours de la thèse.

FIG. 1.1: *PFINDER [WADP97], les membres du personnage sont recherchés parmi les zones de couleur homogènes de l'image.*

1.2.1 Les couleurs

La couleur constitue une information importante de l'image. Avec PFINDER [WADP97] (fig. 1.1), l'image est divisée en zones selon leur couleur. Les parties du corps sont recherchées à l'intérieur des zones homogènes. La connaissance totale ou partielle de la couleur des vêtements permet de juger de la vraisemblance d'une pose [NTTC05] ou, en considérant l'hypothèse de la symétrie des couleurs pour la tenue vestimentaire, de chercher le membre opposé à un membre déjà connu [RBM05].

FIG. 1.2: *Soustraction de fond [ST02].*

L'utilisation de gabarits dont la forme est adaptée à chaque type de membre est une manière de retrouver les membres d'une personne dans l'image grâce à l'estimation de l'homogénéité des couleurs contenues dans le gabarit. Autrement dit : un "template matching" basé sur les couleurs [RMR04] avec l'hypothèse de vêtements uniformes. La couleur ou les niveaux de gris permettent aussi de découper la silhouette d'un personnage dans l'image grâce à un procédé de soustraction de fond [AT06b] [EL04] [LH04][LV04][ST02][TLS05] (fig. 1.2). La recherche des zones de teinte chair est un moyen de repérer les zones susceptibles de contenir la tête ou les mains [GS04][LV04]. Associés à des détecteurs de plus haut niveau comme un détecteur de visage ou des contours

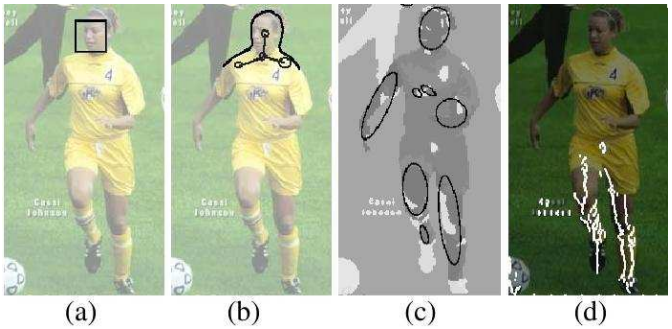


FIG. 1.3: *Observations tirées des couleurs associées à un détecteur de visage et des contours actifs : (a) détection du visage, (b) détection des épaules d'après un contour actif, (c) blobs de teinte chair, (d) axe médian des jambes [LC04].*

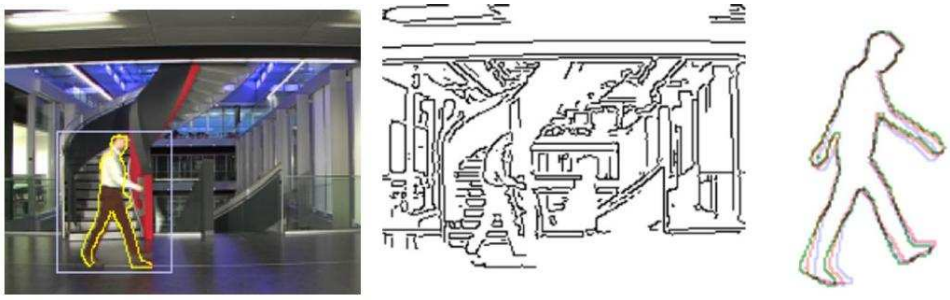


FIG. 1.4: *Template matching sur les contours [DLF05]. Image originale à gauche, au centre : contours d'après l'algorithme de Canny, à droite : un gabarit utilisé pour la détection de piétons.*

actifs, les couleurs et notamment la teinte chair [LC04] sont capables de générer un ensemble d'indices pertinents (fig. 1.3).

Si la couleur apporte beaucoup d'informations, elle varie selon les conditions d'éclairage donc en fonction de l'heure du jour ou de la météo. Dans ces conditions, il est donc souvent nécessaire d'extraire une composante moins sensible à partir des coordonnées colorimétriques. Une première approche consiste à extraire la chrominance normalisée et à admettre qu'une même zone puisse devenir plus sombre du fait d'une ombre [WADP97]. Des techniques robustes d'invariants de couleurs ont également été développées [FCF96][GS96]. Cependant, dans cette nouvelle représentation de l'espace colorimétrique, la quantité d'informations se trouve diminuée.

1.2.2 Les contours

Les contours ont l'avantage d'être plus robustes vis à vis des variations de luminance, mais ils présentent une sensibilité au bruit du fait qu'ils sont issus d'une opération de dérivation, d'où la nécessité d'un filtrage préalable. L'algorithme de Sobel pour la détection des contours associé à un calcul robuste de flot optique va permettre à Sminchisescu et Triggs d'estimer la vraisemblance de leur modèle vis-à-vis des observations [ST01]. Les contours peuvent aussi êtres utilisés pour créer des histogrammes d'orientations des gradients simples [SVD03][TSDD06] ou de type SIFT, plus robustes grâce à une pondération spatiale sur la zone de calcul et une normalisation des amplitudes [AT06a].

FIG. 1.5: *Segmentation des membres [MREM04]. image originale (a), contours (b), segmentation par coupe normalisée.*

La comparaison d'un gabarit spatio-temporel avec l'image des contours produite par l'algorithme de Canny [Can86] conduit à localiser des personnages animés d'un mouvement de marche dans une scène (fig. 1.4). Le gabarit comprends les silhouettes de contours de trois images consécutives incluant la position clé lorsque les deux pieds touchent le sol [DLF05]. Cette solution n'est pas sans poser des problèmes d'échelle lors de la scrutation de l'image avec le gabarit pour adapter la taille de celui-ci à la pose recherchée. Cette difficulté peut être résolue en échantillonnant le contour de la silhouette sur un nombre fixe de pixels et en incluant une opération de normalisation lors de la comparaison avec la base de données apprise [MM02]. Une autre utilisa-

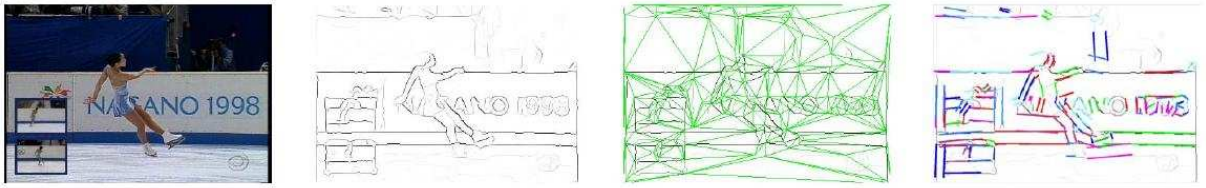


FIG. 1.6: *Détection des membres candidats [RBM05]. Image originale (a), contours (b), triangulation de Delaunay (c), affichage des membres candidats (d).*

tion de l'image des contours consiste à mesurer la distance de chanfrein [RP66] entre les pixels correspondant à la projection d'un modèle dans l'image et la silhouette issue d'une soustraction de fond [FH05]. De la même manière mais sans la soustraction de fond, le modèle peut être plus finement comparé avec l'image en tenant compte de l'orientation des contours [NTTC05].

Associé à une segmentation de l'image, l'algorithme de Canny produit un découpage en zones qui segmente grossièrement les différents membres d'un personnage [MREM04] (fig 1.5). Avec une triangulation de Delaunay, les contours de Canny soumis à des critères de sélection peuvent renvoyer une série de position candidates des membres dans l'image [RBM05] (fig 1.6). Une autre approche utilise l'algorithme de Shen Castan [SC92] pour extraire des tendances de contours par zones locales, une transformée de Hough droite se charge de tracer des directions globales le long de ces zones afin de détecter des membres candidats [LV04] (fig 1.7).

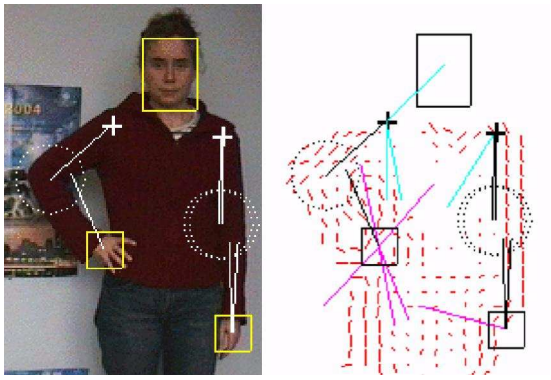


FIG. 1.7: *Placement des membres supérieurs complets candidats sur l'image originale à gauche. Image de droite : détection des zones de teinte chair (rectangles) et orientation locale des contours (tirets fins) permettant de déterminer les directions globales (segments) à partir des mains et des épaules (croix) qui se coupent au niveau des coudes (cercles) [LV04].*

Comparée à une simple extraction de contours, une analyse plus fine et multi-échelles peut être obtenue grâce aux ondelettes. Dans cet optique, Urtasun et al mettent en œuvre des filtres orientés du deuxième degré à base d'ondelettes [JFEM03] dans le but de suivre les articulations lors de la marche ou du swing de golf [UFHF05]. Ce procédé apporte une robustesse supplémentaire vis-à-vis des rotations et des homothéties [FA91].

1.2.3 Détection des mouvements, flot optique

L'estimation du mouvement d'une zone dans l'image à l'instant t permet d'estimer la position de cette zone dans l'image à $t+1$ si on suppose son mouvement uniforme sur la période d'image. Une façon d'estimer l'intensité du mouvement consiste à comparer des images consécutives par soustraction. Cette technique conduit à générer une image de l'énergie du mouvement (fig. 1.8) [BD96] exploitée pour estimer la pose 3D du haut du corps [GS04]. La vraisemblance d'un membre est fonction du nombre de pixels détectés en mouvement dans la zone où le membre est recherché.

Le principe du flot optique consiste à comparer entre elles des images issues du flux vidéo

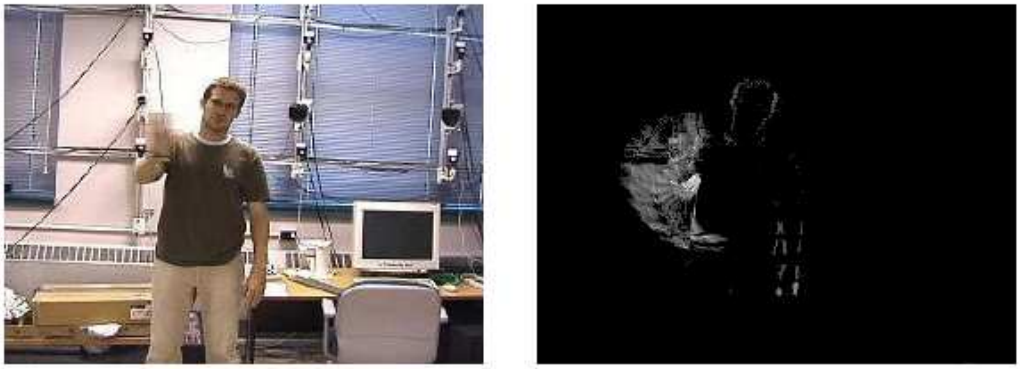


FIG. 1.8: *Image de l'énergie du mouvement [GS04].*

pour évaluer le mouvement en chaque point de l'image (fig. 1.9). Il s'appuie sur le bilan nul des dérivées spatio-temporelles le long de la ligne de mouvement dans l'image [HS80]. Une hypothèse courante consiste à considérer que la valeur d'un pixel, le long de la ligne de mouvement, est indépendante du temps [BAHH92]. Le flot optique permet de générer un modèle dynamique pour rechercher les membres dans l'image [BM98][ST01]

1.2.4 Image des disparités

La tentation d'acquérir une information de profondeur en plus des indices fournis par les couleurs, les contours ou les mouvements est grande, surtout dans les cas difficiles qui présentent un fond complexe ou des occultations de membres. Dans ce dernier cas, les approche multi-caméra [SBR+04] sont particulièrement efficaces pour désambiguïser les occultations. D'une manière générale, les procédés multi-caméra, et notamment la vision stéréoscopique [BCMC06][DTS+05][PF03] ou trinoculaire [UF04] permettent d'estimer la profondeur d'un pixel dans l'image. Cette information améliore l'estimation de la

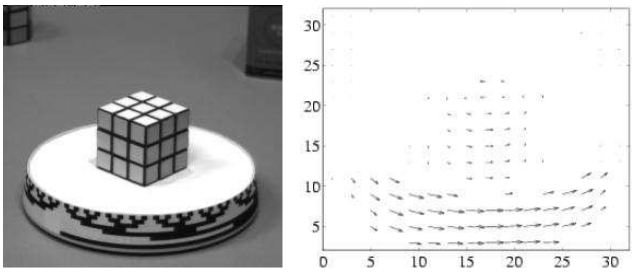


FIG. 1.9: *Mesure du flot optique (image de droite) pour un cube en rotation (image de gauche).*

FIG. 1.10: *Image des disparités (au centre) issue d'une paire d'images stéréoscopiques (gauche et droite).*

vraisemblance du modèle puisque l'erreur est mesurée dans l'espace 3D plutôt qu'en 2D avec une projection du modèle dans le plan de l'image. L'exploitation de ces informations peut se faire de manière déterministe [DTS+05][PF03][UF04] ou stochastique [BCMC06][SBR+04].

Pour fournir des informations de profondeur, les caméras doivent d'abord êtres *calibrées*. Cette opération sert à déterminer les paramètres *intrinsèques* propres aux caméras et *extrinsèques* des positions mutuelles entre les caméras. À ce niveau, la *disparité* qui est l'écart de position entre deux pixels qui "voient" le même point sur des images rectifiées issues des deux caméras calibrées permet de déterminer la profondeur. La rectification vise à faciliter l'identification de ces pixels en faisant correspondre les *lignes épipolaires* avec les lignes horizontales de pixels dans les images stéréo. L'appariement des pixels revient alors à scruter uniquement ces lignes pour trouver la disparité et donc la profondeur (fig. 1.10). La nécessité de calibrer les caméras reste l'inconvénient majeur de ces méthodes qui exigent une mise en œuvre plus lourde que les solutions monocaméra.

1.2.5 Sélection des informations pertinentes

L'image contient un grand nombre d'informations ainsi que du bruit et l'exploitation des indices bruts ne permet pas toujours d'en retirer les informations pertinentes à la détermination de la pose. Respectant ce principe, la soustraction de fond (fig. 1.2) permet de neutraliser l'influence du fond mais elle implique une caméra fixe. Ce paragraphe présente deux techniques capables de traiter l'information issue de l'image pour la rendre plus synthétique et pertinente.

Factorisation en matrices non négatives

L'utilisation d'histogrammes de contours pour identifier les articulations dans l'image [AT06a] est amélioré grâce à la méthode de factorisation des matrices non négatives ("*Non-negative Matrix Factorization - NMF*") [Hoy04]. Des bases sur lesquelles les matrices d'histogrammes sont projetées sont apprises pour chaque articulation avec un fond vierge afin d'affecter un poids faible aux contours issues du fond (fig. 1.11). Si cette technique est susceptible d'épargner le détourage de la silhouette, elle fait l'hypothèse que le personnage est centré dans l'image avec une échelle connue.

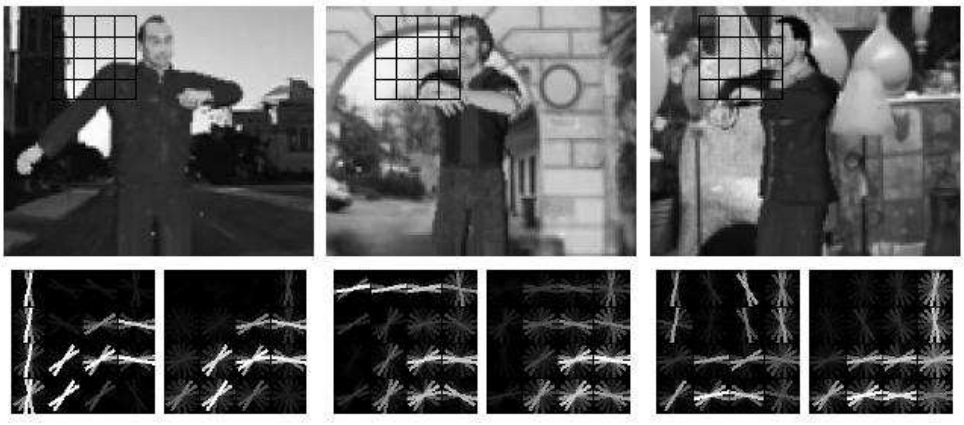


FIG. 1.11: *Factorisation en matrice non négative. Ligne du haut : images originales, ligne du bas, à gauche et en bas de chaque image : histogrammes des orientations et à droite, sous les images : histogrammes des orientations après factorisation. Les contours engendrés par le fond (arbre, arche...) n'apparaissent plus du fait d'un apprentissage sur un fond vierge [AT06a].*

Analyse en composantes principales

L'analyse en composantes principales (ACP) permet de réduire le nombre de dimensions de l'espace dans lequel les données sont exprimées en sélectionnant les dimensions les plus discriminantes à partir de la matrice de covariance calculée sur les données à traiter. Sigal et al [SBR+04] extraient des images les trois premières composantes principales pour obtenir des données plus compactes en provenance de leur système d'acquisition multi-caméra (fig. 1.12). L'ACP peut s'appliquer à d'autres types de données et notamment à un modèle de forme du corps pour réduire sa dimension et estimer la probabilité à priori sur les données réduites [LC04]. Les mouvements cycliques tels que la marche ou la course à pied peuvent faire l'objet d'une ACP sur les paramètres articulaires pour distinguer ces deux comportements sur un espace à deux dimensions seulement [UF04].

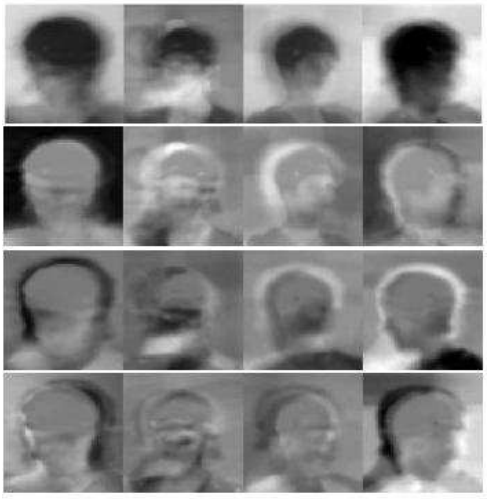


FIG. 1.12: *Reconstruction du visage (première ligne) d'après une ACP sur les trois composantes principales (trois dernières lignes)[SBR+04].*

1.3 La modélisation

L'étape de modélisation consiste à déterminer, par rapport aux caractéristiques extraites de l'image, la façon de paramétrer la pose humaine afin d'estimer la vraisemblance des hypothèses. Les approches classiques utilisant un modèle explicite de la forme consistent à modéliser le corps humain en 2D ou en 3D. Dans le premier cas, on trouve principalement des modèles sous la forme de patchs rectangulaires [FH05] [GS04] [IF01] [JBY96] [LH04] [MREM04] [RBM05][RMR04], ou d'un squelette 2D [LV04]. Le modèle 3D offre plus de variantes mais on notera, pour représenter les membres, l'utilisation d'un squelette 3D [Tay00][TLS05], de troncs de cônes [DD02][DKD03][DTS+05] [LC04][SBR+04] ou de quadriques [BM98][NTTC05][ST01][ST02][ST03]. Une solution offrant beaucoup de précision consiste à utiliser des primitives Gaussiennes en trois dimensions [PF03] [UF04]. Il est possible de mixer les types de modèles : sphère pour la tête, cylindre pour le torse et modèle patchs rectangulaires 2D pour les bras [BCMC06].

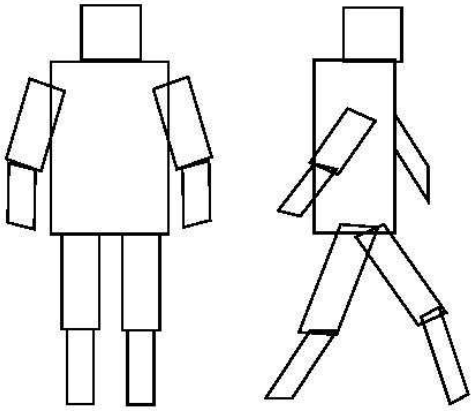


FIG. 1.13: *Modèle "cardboard" [JBY96].*

Il existe également des techniques implicites pour modéliser la pose humaine. Celles-ci font généralement appel à un apprentissage avant de comparer l'image test avec la base apprise. En 2D, la silhouette peut être associée à un descripteur de forme de type "shape context" [MM02][AT06b]. Ce procédé à pour but de coder les contours à partir d'un histogramme 2D. D'autres façons compactes de modéliser implicitement la pose humaine consistent à utiliser des variétés [EL04] [TLS05], ou un étiquetage par hachage [SVD03] [TSDD06], voire des gabarits spatio-temporels [DLF05].

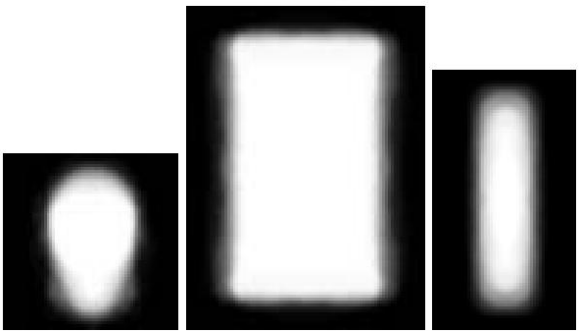


FIG. 1.14: *Modèle de masques flous destinés à détecter la tête, le torse, une jambe ou un bras [RMR04].*

Modéliser la position du corps humain exige de définir des paramètres intrinsèques et extrinsèques. Les premiers, constants pour chaque personne, contraignent la forme des membres en fixant la taille et les proportions entre eux. Les seconds sont variables et ils donnent la position des articulations dans l'espace. En réalité, les paramètres intrinsèques ne sont pas constants du fait, par exemple, des déformations engendrés par les vêtements. En pratique on adopte souvent l'approximation qui consiste à les considérer comme constants.

1.3.1 Approches 2D

Le modèle de patchs rectangulaires ou "cardboard" [JBY96] (fig. 1.13) consiste à inscrire chaque membres dans un rectangle [FH05][GS04][IF01][LH04][MREM04][RBM05]. Aisément projeté dans l'image, ce modèle délimite une zone de l'image dans laquelle on évalue la vraisemblance des hypothèses en regard des primitives extraites. Ce modèle ne renseigne pas sur les

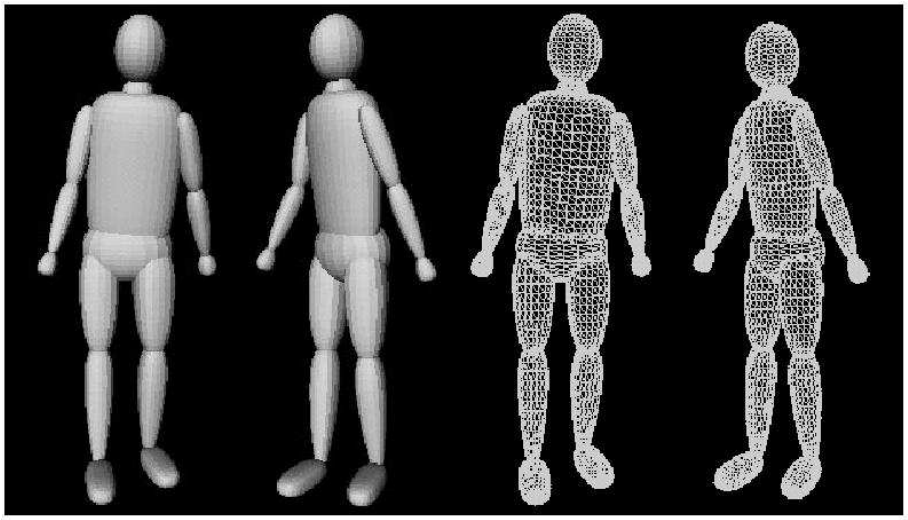


FIG. 1.16: *Modélisation des membres par des primitives ellipsoïdales (quadriques). Le passage en "fil de fer" permet de discrétiser le modèle pour faciliter sa projection dans le plan image [ST02].*

données de profondeur et les approches qui l'utilisent se bornent souvent à retrouver la pose 2D [FH05][IF01][LH04]. Le passage à l'information en 3D peut se faire en exploitant l'algorithme de Taylor mais celui-ci
exige que les proportions entre les membres soient connues [Tay00]. Cette dernière option est utilisée en mono-caméra pour faire du suivi de personnes [GS04] ou de l'estimation de pose [MM02]. Les "*pictorial structures*" constituent une solution pour résoudre des problèmes généraux de reconnaissance dans les images comme la reconnaissance de visages [FE73]. L'association de cette technique avec un modèle cardboard permet d'effectuer de l'estimation de pose [FH05]. Une variante floue du modèle cardboard à été développée pour une approche où un masque support du membre est associé au patch [RMR04] (fig. 1.14).

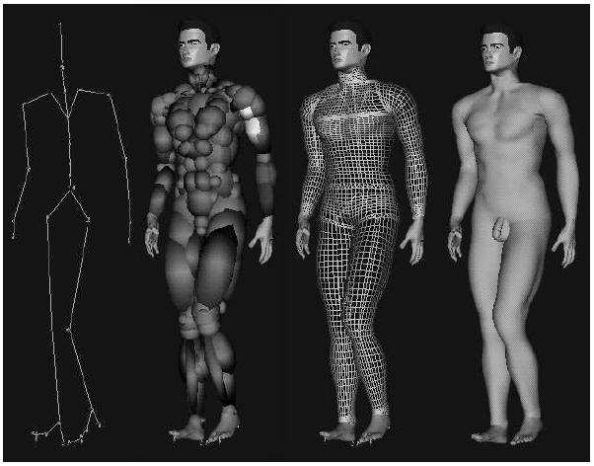


FIG. 1.15: *Modélisation des tissus musculaires avec des primitives Gaussiennes [PF01].*

Le squelette 2D permet de positionner, sous la forme d'axes, des hypothèses des membres trouvés sur l'image [LV04]. La vraisemblance, évaluée le long de ces axes, permet de valider le modèle.

1.3.2 Approches 3D

L'utilisation de quadriques [BM98][NTTC05][ST01][ST02][ST03] consiste à modéliser un membre à partir d'une ellipsoïde. Un exemple classique génère une pose paramétrée par 38 variables et la forme des membre exige 9 paramètres chacun [ST02] (fig. 1.16). Une variante de ce principe utilise des Gaussiennes 3D ou métasphères pour modéliser les membres [UF04] ou indépendamment chaque muscle du corps [PF01][PF03]. Dans ce dernier cas, le résultat est d'une

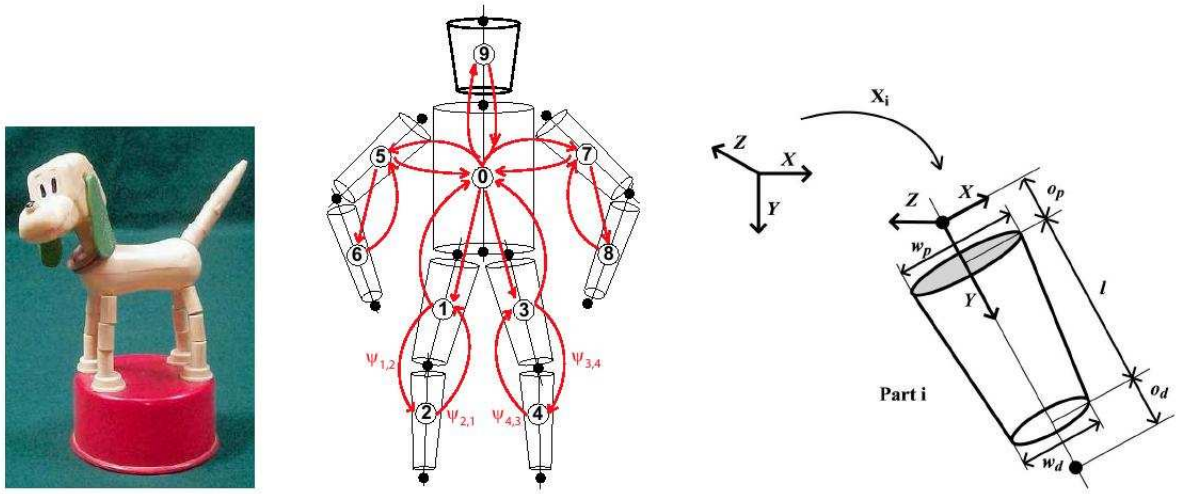


FIG. 1.17: *Modèle troncs de cônes à membres indépendants. A gauche : analogie avec un jouet à poussoir pourvu d'articulations élastiques. Au centre : modèle graphique. À droite : les 11 paramètres pour définir un membre [SBR+04].*

grande précision mais le modèle d'apparence est plus complexe (fig. 1.15).

Moins précise, la modélisation à base de troncs de cônes [DD02][DKD03][DTS+05] exige aussi moins de paramètres intrinsèques. De manière classique, cette approche utilise 3 paramètres intrinsèques par membre et 31 paramètres extrinsèques pour contraindre la pose du corps [LC04]. Dans ce dernier cas, et en supposant les proportions connues, la recherche de la meilleure pose implique d'explorer un espace de grande dimension (31 pour l'exemple cité), compliquant ainsi la convergence vers une solution optimale. La dimension de cet espace peut être réduite en adoptant un modèle de corps avec membre indépendants [SBR+04] (fig. 1.17). Certes le nombre de paramètres est plus élevé que précédemment puisque le modèle intègre cinq paramètres fixes correspondant à la longueur du membre, sa largeur au niveau de l'articulation proximale et distale et son décalage le long de ces articulation. Les paramètres variables pour estimer la pose correspondent à la position et à l'orientation de l'articulation proximale, soit six paramètres pour un membre. Le modèle entier qui comprends un total de dix membres fournit une pose paramétrée par soixante variables. Cependant, puisque les membres sont indépendants, le problème se ramène à une optimisation sur les 6 paramètres variables de chacun d'eux.

Dans les précédentes approches, les primitives volumétriques (troncs de cône, quadriques, etc) prennent support sur une structure squelettique 3D. Le squelette seul peut faire fonction de modèle et permettre la localisation des articulations, mais il ne fournit pas de modèle de forme [Tay00][TLS05].

1.3.3 Approche sans modèle explicite de la forme

Avec la "structure from motion", une forme est représentée par un ensemble de points 3D dont leur projection est suivie au cours d'une suite d'images prises sous différents angles. Cette technique permet de retrouver les coordonnées 3D de ces points à partir de leur coordonnées 2D concaténées dans une matrice de mesures qui va engendrer un système d'équations surdéterminé. Cette méthode a été mise au point pour une cible rigide non déformable [TK92], puis a été adapté aux cibles non rigides [BHB00] avec une gestion des occultations [THB03]. Cependant, le nombre de prises de vue doit être important pour contraindre le système et les performances dépendent

de la précision avec laquelle les points du modèle sont suivis dans l'image.

1.4 Outils de manipulation des paramètres du modèle

Le but de cette étape est fondamentale puisqu'elle détermine la façon dont sont optimisés les paramètres du modèle après les avoir comparés aux observations sur l'image. Parmi les méthodes parcourues dans cette bibliographie, on distingue les approches déterministes qui consistent à optimiser une fonction objectif [DD02][DKD03][FH05][PF03][ST02][UF04][BM98], ainsi que les solutions qui mettent en œuvre un apprentissage sous la forme d'une régression [AT06a][AT06b], de variétés [EL04][TLS05][UFHF05], de clés de hachage [SVD03] ou de "shape context" [MM02]. On trouve aussi les approches stochastiques qui introduisent des probabilités pour modéliser les incertitudes [BCMC06][DTS+05][GS04][IF01][LH04][LC04][NTTC05][SBR+04][ST01][TSDD06], les méthodes qui mettent en œuvre une série de règles pour constituer des assemblages cohérents entre les membres [LV04][MREM04][RBM05] et les approches "*template matching*" qui consistent à rechercher les membres à partir de gabarits [DLF05][RMR04].

1.4.1 Approches déterministes

La frontière entre déterministe et stochastique est parfois ténue et des choix ont dû être effectués pour constituer le classement proposé ici. Les approches qui modélisent la vraisemblance du modèle connaissant l'image ainsi que celles qui utilisent des outils probabilistiques (modèles graphiques probabilistes du corps, modèles de Markov cachés...) ont étés classées comme stochastiques. En revanche, sont classées comme déterministe les méthodes qui consistent adapter le modèle aux indices extraits de l'image à partir de l'optimisation d'une fonction de coût calculée selon une métrique définie. Même si cette métrique peut prendre la forme d'une probabilité [ST02] ou qu'un apprentissage y soit utilisé [DKD03], ces méthodes restent globalement déterministes. Dans ce cadre, l'optimisation du modèle utilise courament au premier ordre : le gradient local selon Levenberg Marquardt [PF03][UF04], Newton Raphson [BM98] ou la projection du gradient de Rosen [DKD03] et au second ordre le hessien [ST02]. Alternativement, une méthode d'optimisation peu gourmande en puissance de calcul, l'ICP [BM92] est retenue pour [DD02][DKD03].

Les points générés par des acquisitions stéréoscopiques [DD02][DKD03][PF03] ou trinoculaires [UF04] peuvent êtres appairés avec un modèle 3D. L'algorithme d'ICP (*Iterative closest point*) [BM92], consiste à trouver ces correspondances de manière à minimiser la distance euclidienne entre le modèle et ces points. Une transformation géométrique est calculée d'après ces paires pour affiner les paramètres du modèle au cours des itérations [DD02]. Cette étape peut précéder une optimisation dans un espace généré par une machine à vecteurs de support apprise sur les contraintes articulaires pour les prendre en compte [DKD03]. Une autre approche [UF04] affecte le membre le plus proche à chaque point issu des données de disparité. L'optimisation est menée par la méthode des moindres carrés sur la distance globale entre le modèle et ces points. La cohérence temporelle du suivi est assurée en menant l'optimisation sur l'ensemble des images d'une séquence au lieu de le faire image par image. Une ACP sur des sujets marchant ou courant sélectionne un faible nombre des paramètres articulaires les plus représentatifs de ces deux comportements. Le mouvement est modélisé selon une combinaison linéaire des composantes principale d'après laquelle l'optimisation s'opère. Cette approche se borne au suivi hors ligne d'un sujet marchant ou courant. Le bruit contenu dans l'image des disparités peut aboutir à un débordement du modèle vis à vis de la réalité, surtout lorsque ce dernier est très précis [PF03] (voir §1.3.2). Afin d'éviter cet artefact, les auteurs on recours à l'ajout pondéré d'observations issues des points de contour de la silhouette.

Une approche mono-caméra consiste à comparer une silhouette extraite de l'image à la projection d'un modèle 3D [ST02]. L'optimisation des paramètres s'appuie sur une vraisemblance

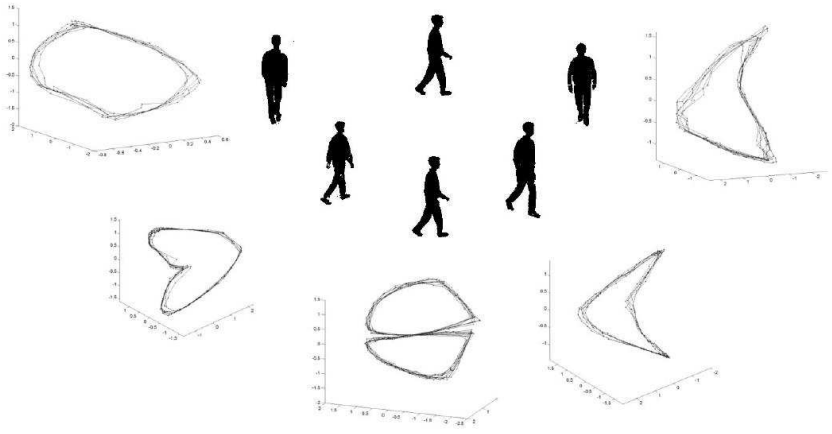


FIG. 1.18: *Variétés issues de la marche en fonction de différents angles de vue [DLF05].*

comprenant deux termes. Le premier estime la surface en commun entre la silhouette cible et la projection du modèle au sens des moindres carrés. Le second est un terme d'attraction qui "pousse" le modèle à l'intérieur de la silhouette grâce à une distance estimée par la méthode du cheminement rapide (*fast marching method*). Cette méthode consiste à déplacer à vitesse égale tous les points d'un contour dans le sens centripète à la normale du contour [Set99].

Un formalisme mathématique élégant, les produits d'exponentiels de "*twists*" [MSZ94], est mis en œuvre dans le but de faire un suivi du corps [BM98]. Comparé à la manipulation classique des matrices homogènes, cette nouvelle écriture simplifie la résolution des chaînes cinématiques. Grâce à cette technique associée à un modèle de flot optique affine [BAHH92], l'estimation du mouvement dans le plan image génère un système d'équations permettant de trouver les paramètres du mouvement dans l'espace. La projection du modèle sur l'image délimite des masques englobant chaque membre à l'intérieur desquels une sélection des pixels est opérée. Afin d'améliorer l'estimation du mouvement 3D, les pixels sélectionnés doivent valider le modèle de mouvement rigide local dans l'image. Les résultats expérimentaux montrent un suivi des membres rigoureux sur des courtes scènes de Muybridge [Muy01]. Appliquée en monoculaire, cette technique est aussi implémentable pour des captures multi-caméra.

1.4.2 Approches à base d'apprentissage

Ces méthodes utilisent un apprentissage préalable pour stocker des données inhérentes à la pose dans un but de comparaison avec les indices extraits de l'image test. Ces données peuvent prendre la forme de clés de hachage [SVD03] ou d'histogrammes log-polaires [MM02]. Une seconde famille d'approches basées sur l'apprentissage consiste à établir une relation entre les indices tirés de l'image et les paramètres de la pose. Ces derniers peuvent êtres exprimés dans une base de fonctions vectorielles à partir d'une régression [AT06a][AT06b][EL04] ou inclure une variable latente Gaussienne pour introduire des probabilités dans l'espace latent contenant les poses apprises [TLS05][UFHF05].

Régressions et espaces latents

Classiquement, l'apprentissage d'une régression se fait en recherchant les coefficients a_k d'une combinaison linéaire de fonctions base $\phi_k(.)$ faisant le lien entre les indices extraits de l'image x

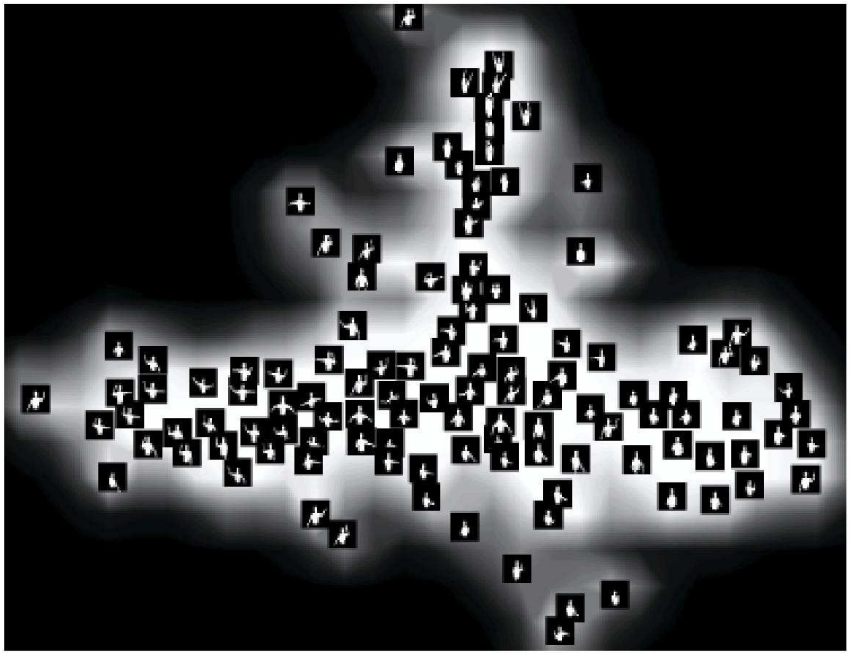


FIG. 1.19: *Variété en deux dimensions dans un espace probabiliste. Un point de la variété défini une pose associée à une probabilité (niveaux de gris du fond)[TLS05].*

et les paramètres de la pose y et minimisant l'erreur commise ϵ :

$$x = \sum_{k=1}^{p} a_k \phi_k(y) + \epsilon \tag{1.1}$$

Ce formalisme peut être utilisé avec de nombreux indices extraits de l'image, et en particulier des histogrammes d'orientations [AT06a], les points de la silhouette [EL04] ou des descripteurs de formes de type "*shape context*" [AT06b].

La recherche d'une fonction de transfert entre l'image et la pose peut s'accompagner d'une réduction de la dimension. L'analyse en composantes principales n'est pas adaptée du fait des occultations ou des tissus déformables qui introduisent des non-linéarités. Pourtant, l'idée de réduire la dimension de l'espace de recherche peut séduire vu le nombre important des paramètres à optimiser (voir §1.3.2). Avec les variétés localement linéaires (LLE) [RS00] un point de l'espace de départ est exprimé comme une combinaison linéaire de ses N plus proches voisins. Les composantes principales sont calculées sur la matrice des poids de cette combinaison linéaire. Cette opération génère une une variétés de dimension réduite. Cette solution est utilisée pour suivre la silhouette d'un sujet animé du mouvement de la marche vu sous plusieurs angles [EL04]. Les variétés obtenues se distinguent clairement les unes des autres en fonction de l'angle choisi pour la prise de vue (fig. 1.18). Cette propriété permet de trouver l'angle de vue d'une séquence en comparant la variété obtenue à une base de variétés apprises. La fonction de transfert inverse donne alors la pose du personnage. Cependant, le nombre fini d'exemples appris crée une variété discrète procurant des sauts dans le suivi si la base n'est pas suffisamment dense. L'ajout de probabilités par l'intermédiaire d'un bruit Gaussien engendre un modèle de processus à variable latente Gaussienne [Law03]. Contrairement à l'analyse en composante principale probabiliste (PPCA) où l'apprentissage se fait en optimisant la vraisemblance de la base d'après la variable latente [TB99], on cherche à optimiser la vraisemblance d'après la fonction noyau qui exprime la similitude entre les vecteurs de la base [Law03] (GPLVM et SGPLVM). Ces techniques génèrent un espace de modélisation continu et autorisent un suivi fluide du personnage. Le suivi peut

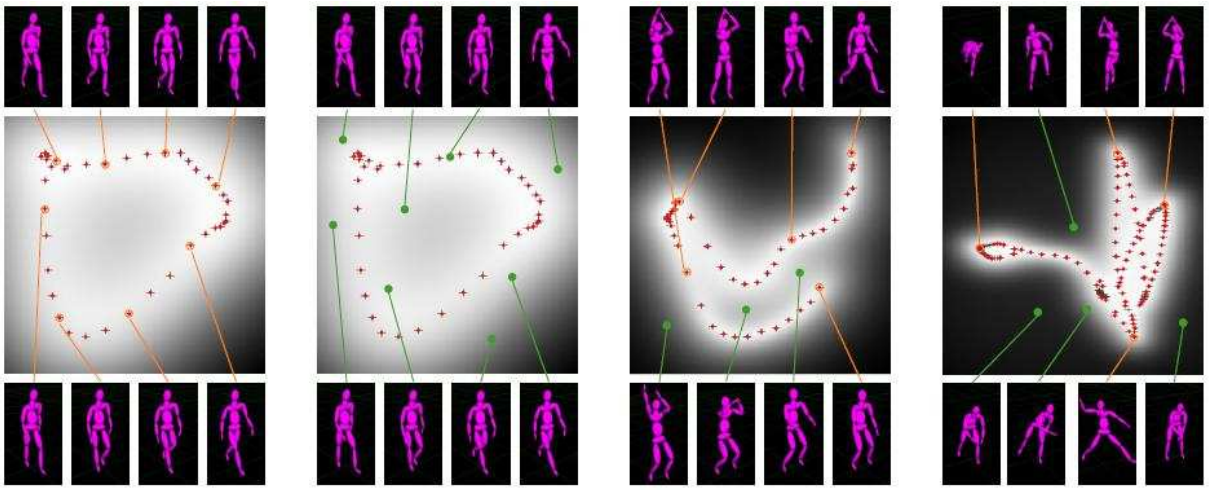


FIG. 1.20: *Espaces latents appris pour la marche, le smash de basket ou le service au base-ball. Les croix rouges correspondent aux poses apprises. Les points oranges sont reliés à des poses contenues dans la base d'apprentissage et les verts en sont des extrapolations. L'apprentissage organise l'espace latent de manière à regrouper les poses similaires dans la même zone. Les niveaux de gris sur le fond correspond à la valeur de la vraisemblance des paramètres de la pose dans l'espace latent. Celle-ci est maximum à proximité des poses apprises [GMHP04].*

être étendu à des gestes plus généraux comme les signes de guidage des avions au sol [TLS05] (fig. 1.19) ou des gestes sportifs [UFHF05] (fig. 1.20).

Comparaison à une base apprise

Quelques approches tentent de généraliser la reconnaissance aux poses quelconques en comparant les données extraites de l'image avec celles enregistrées dans une base d'apprentissage. Ces méthodes exigent des bases importantes [SVD03] ou utilisent des méthodes d'interpolation [MM02]. Pour cette dernière approche, la base est traduite en histogrammes log-polaires grâce aux "*shape context*" pour faire de l'estimation de pose (fig. 1.21). Les histogrammes issus de l'image test sont comparés à ceux de la base d'apprentissage pour trouver la meilleure correspondance [MM02]. La position 2D des membres est déterminée en estimant, par la méthode des moindres carrés, la transformation géométrique locale entre l'image test et l'image d'apprentis-

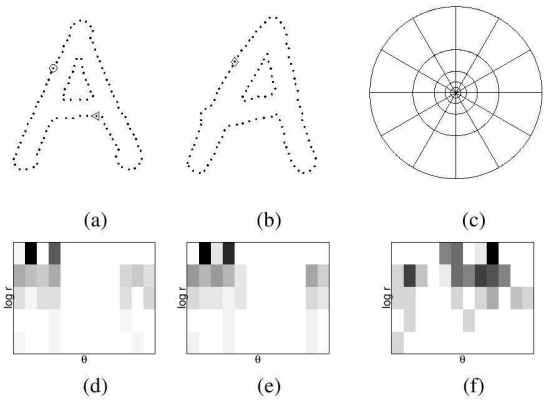


FIG. 1.21: *Shape context : (a,b) points échantillonnés le long des contours des lettres, (c) histogramme log-polaire utilisé pour calculer le descripteur de forme, (d-f) Exemple de diagrammes calculés à partir des points de référence marqués ∘ ◇ ◁ (valeurs élevées en sombre) [MM02].*

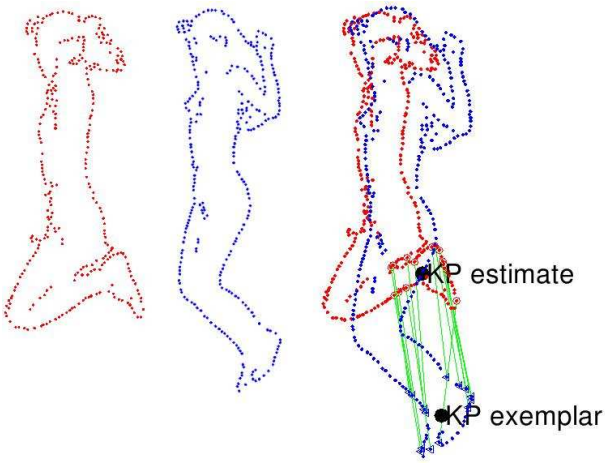


FIG. 1.22: *Shape context : localisation des articulations. Les points échantillonnés le long de la silhouette exemple (à gauche) et de test (au centre) sont mis en correspondance. Une transformation est estimée au niveau local pour retrouver le pied dans l'image test (lignes vertes dans l'image de droite) [MM02].*

sage qui lui correspond le plus (fig. 1.22). La position 3D des articulations est déduite grâce à l'algorithme de Taylor [Tay00].

Le fait de d'utiliser des bases d'apprentissage de très grande taille exige de mettre en œuvre des stratégies adaptées afin d'accélérer la recherche des plus proches voisins d'une requête. Une solution est trouvée grâce au LSH, une technique de hachage qui conserve les distances entre les exemples [GIM99]. Cette méthode appliquée à l'estimation de poses quelconques et appelée "*PSH - Pose Sensitive Hashing*" [SVD03], permet de générer des clés binaires de faible dimension qui respectent les relations métriques entre les poses. Des indices extraits de l'image sont binarisés d'après un seuil optimal calculé sur les données apprises. Un sous-ensemble d'indices binaires tirés aléatoirement génèrent les fonctions de hachage et la comparaison d'une image test avec la base consiste à sélectionner un sous-ensemble des poses apprises d'après la similitude des clés. Celle-ci est efficacement mesurée grâce à la distance de Hamming et la pose résultat est extrapolée à partir d'une régression locale pondérée sur les plus proches voisins trouvés.

1.4.3 Approches stochastiques

Le bruit induit par les caméras, les imprécisions du modèle ou les hypothèses simplificatrices génèrent des incertitudes dans les observations rendant pertinente l'expression de la validité du modèle par une fonction de probabilité. Cette fonction étant multimodale dans un espace de grande dimension, les méthodes analytiques sont peu efficaces et on préférera discrétiser l'espace des poses en le divisant par zones [LH04][TSDD06] ou grâce à un échantillonnage à base de MCMC ("*Monte Carlo Markov Chain*") [LC04], un échantillonnage d'importance (filtre à particules) [BCMC06][ST01][ST03] ou une exploration méthodique dans l'image [GS04][IF01][NTTC05]. Une représentation continue de la probabilité de la pose dans l'image sous la forme de noyaux Gaussiens peut être aussi envisagée [DTS+05][SBR+04].

Les approches stochastiques ont la capacité de fournir une approximation de la densité a posteriori et de propager les hypothèses pertinentes au cours des itérations. Ce principe paraît judicieux si on considère qu'une hypothèse non optimale à l'instant t peut être à l'origine de la bonne pose dans le futur. La propriété multi-hypothèse de ces algorithmes leur permet donc de raccrocher la bonne solution après une erreur de suivi.

La résolution probabiliste d'un problème de vision par ordinateur consiste généralement à estimer la probabilité a posteriori $p(x \mid y)$ des paramètres du modèle x en considérant les observations sur l'image y. À partir de cette estimation, il est possible d'en extraire le maximum a

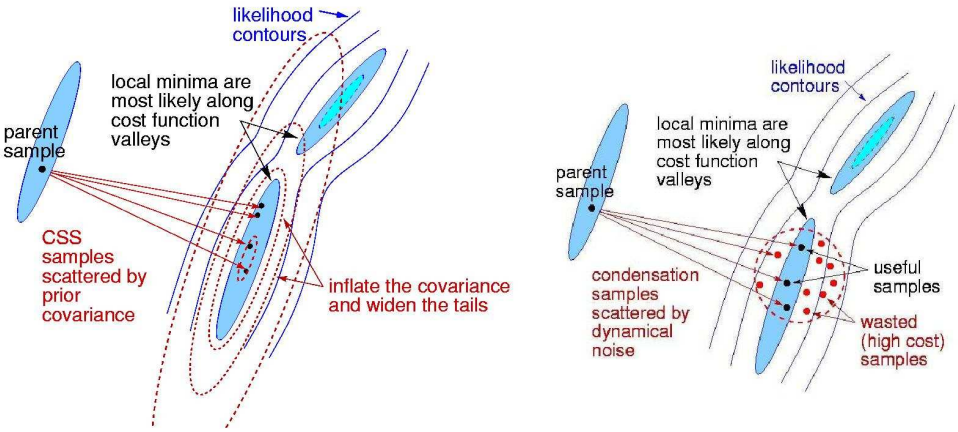


FIG. 1.23: *"Covariance scalled sampling". À gauche, le rééchantillonnage se fait d'après une ellipsoïde qui suit les vallées à fort vraisemblance. Avec l'algorithme CONDENSATION [BI96] classique, ce rééchantillonnage se fait selon un mouvement brownien isotrope (figure de droite). De nouveaux maxima peuvent êtres découverts plus probablement dans le premier cas (coutesy of Sminchisescu PhD Thesis).*

posteriori (MAP) :

$$\overset{*}{x} = \underset{x}{argmax}[p(x \mid y)], \tag{1.2}$$

ou l'espérance :

$$< x >= \int xp(x \mid y)dx. \tag{1.3}$$

L'écriture Bayésienne permet de décomposer l'estimation de la probabilité a posteriori :

$$p(x \mid y) \propto p(y \mid x)p(x). \tag{1.4}$$

La probabilité $p(y \mid x)$ et notée $L(x, y)$ est appelée vraisemblance. C'est elle qui est exprimée par un modèle génératif. La probabilité a priori sur le modèle $p(x)$ représente l'ensemble des connaissances dont on dispose sur le modèle sans tenir compte des observations. Ces connaissances peuvent porter sur les liens physiques (les articulations et leur contraintes) et/ou temporels (cohérence temporelle d'un corps en mouvement)[DTS+05], elles peuvent êtres apprises [GS04] et être représentées par un mélange de Gaussiennes [SBR+04]. Demirdjian et al [DTS+05] procèdent à une comparaison entre l'image test et une base apprise à la manière de [SVD03]. Un processus d'optimisation est initialisé à partir des meilleures poses pour modéliser la vraisemblance avec un mélange de Gaussiennes. La probabilité a posteriori est issue de la vraisemblance pondérée par un *prior* Gaussien centré sur la pose à $t-1$.

Échantillonnage de la densité a posteriori

La densité a posteriori est généralement impossible à exprimer analytiquement et une approximation par échantillonnage est souvent utilisée dans la pratique. Lee et Cohen [LC04] font appel à un échantillonnage de type Métropolis Hastings pour estimer la densité a posteriori. La densité de proposition, classiquement un mouvement brownien, est remplacée par une fonction conditionnée par les observations. Cette technique d'échantillonnage par MCMC conduite d'après les données sur l'image (data driven Monte Carlo Markov Chain) permet de faire converger l'algorithme vers un optimum global plus efficacement. Pour alimenter ce procédé, il est nécessaire

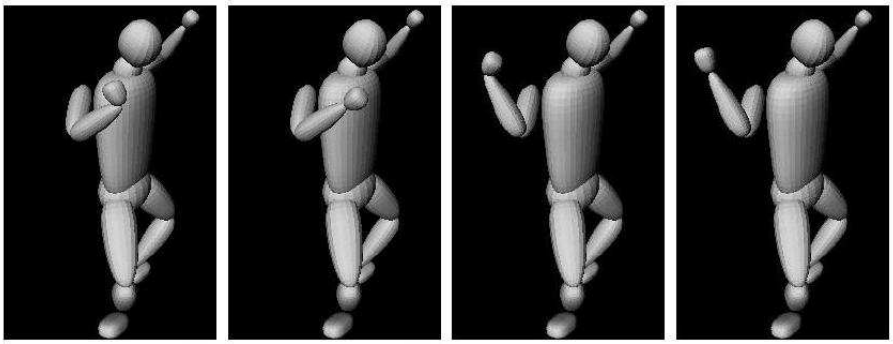


FIG. 1.24: *Ambiguïtés 3D-2D : un membre doté de deux articulations vu en 2D peut générer quatre positions 3D qui auront la même projection dans l'image. Un exemple est donné ici avec le bras complet muni des deux articulations : le coude et le poignet [ST03].*

d'extraire des cartes de probabilités pour chacun des membres recherchés. Ces "proposal maps" sont le fruit d'hypothèses pondérées par leur confiance et issues des caractéristiques extraites de l'image. Ces hypothèses sont modélisée sous la forme de Gaussiennes 2D sur l'image d'après lesquelles les échantillons sont tirés.

Dans le cadre du filtre à particules [IB96] (§ 2.3.1), l'approximation des distributions de probabilités a posteriori est conduite par un échantillonnage d'importance séquentiel suivant une distribution de proposition. Dans le cas de l'algorithme CONDENSATION [BI96], la cohérence temporelle est contrainte par une distribution de proposition Gaussienne centrée sur la pose trouvée à l'image précédente. Un rééchantillonnage est nécessaire pour empêcher la dégénérescence des échantillons vers une solution unique. Le nombre de particules (les échantillons) doit être suffisant pour couvrir la plus grande partie des hypothèses plausibles. Or, la grande dimension de l'espace de recherche pour le suivi de personnes exige un nombre de particules souvent prohibitif. Dans le but d'éviter le "street light effect" [DTS+05], c'est-à-dire le fait de rechercher un optimum dans un espace trop restreint, une amélioration de l'étape de rééchantillonnage peut offrir des solutions. Le fait de remarquer que la fonction de vraisemblance possède des maxima allongés sous la forme de vallées pousse à rééchantillonner avec une covariance qui suit ces vallées (fig. 1.23) grâce à la technique de "covariance scalled sampling" [ST01]. Une seconde amélioration consiste à générer des échantillons vers les poses 2D qui présentent une ambiguïté avec la projection du modèle dans l'image (fig. 1.24). En exploitant la propriété multi-hypothèses du filtre à particules, le procédé de saut cinématique [ST03] permet de raccrocher le suivi après une désambiguïsation de la pose 2D grâce aux contraintes temporelles.

Modélisation du corps par un modèle graphique

Un modèle graphique est un graphe incluant des variables aléatoires représentés par des nœuds. Des liaisons forment un système de voisinage entre ces nœuds pour exprimer des couplages entre les variables aléatoires. L'intérêt d'un tel modèle se trouve dans sa capacité à mettre en évidence des relations d'indépendance entre les nœuds, lorsqu'il y a absence de liens, pour factoriser l'expression de la probabilité jointe sur l'ensemble des variables aléatoires du graphe. Dans le contexte du suivi du corps humain, les nœuds modélisent souvent l'état des membres ou des articulations.

Le réseau Bayésien consiste en un graphe acyclique orienté [Pea88]. Il est commodément utilisé dans le cadre d'approches ascendantes pour modéliser les liens entre les membres du corps par une structure arborescente. Une interprétation Bayésienne des "*pictorial structures*" adaptée aux problèmes d'estimation de la pose à été décrite initialement par [FH05]. Dans [NTTC05], la recherche des membres consiste à calculer l'orientation des contours appartenant à la projection d'un modèle 3D sur l'image. Pour chaque pixel de la projection, la distance qui le sépare du plus proche contour de l'image ayant une orientation similaire est mesurée. Le résultat de la moyenne des distances par une exponentielle négative fourni une probabilité qui, mêlée à une probabilité d'apparence basée sur les couleurs et apprise sur les trente premières images, donne la probabilité de similitude du modèle. Plusieurs hypothèses sont retenues pour chaque image avant d'être départagées par l'algorithme de Viterbi. Pour une seconde approche [IF01], la recherche des membres est facilitée par le choix des photos de Muybridge [Muy01] où le corps apparaît généralement en clair sur un fond sombre. Le problème de reconnaissance suit une procédure classique inspirée du modèle de réseau Bayésien. Si le corps comprends k membres et x_i représente la configuration du membre i, la probabilité jointe $P(x_1, ..., x_i, .., x_k)$ est factorisée en produit de probabilités conditionnelles du membre x_i connaissant le membre parent x_{i-1}. La probabilité de la racine x_{root} détermine la pose globale du corps et la probabilité jointe est :

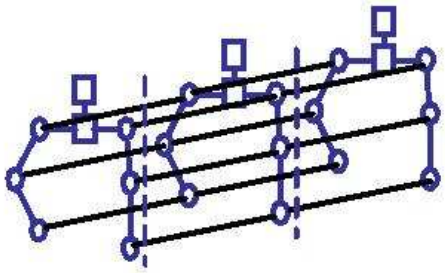


FIG. 1.25: *Champ de Markov intégrant une fenêtre temporelle sur trois images [GS04].*

$$P(x_1, ..., x_i, .., x_k) = p(x_{root}) \prod_{i \neq root} P(x_i \mid x_{i-1}) \tag{1.5}$$

Dans le but d'établir un modèle plus général, les auteurs font appel à une mixture d'arbres afin d'inclure des modèles d'arbres tronqués qui correspondent aux différents cas d'occultations.

Un champ de Markov aléatoire ou *Markov Random Field (MRF)* [Li95] est un graphe non orienté. Il est utilisé dans [GS04][SBR+04] pour modéliser l'influence des contraintes articulaires ou temporelles grâce aux liaisons (fig. 1.25). Cette modélisation permet d'exprimer la probabilité jointe comme un produit de facteurs sur les cliques maximales (théorème de Hammersley-Clifford). L'utilisation d'un graphe de facteurs [KFL01] offre la possibilité de diviser les cliques pour ne garder que des facteurs entre paires de nœuds. Cette simplification est avantageusement exploitée pour parvenir à un algorithme de suivi temps réel [BCMC06].

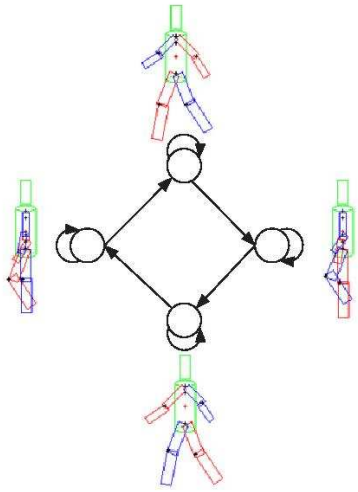


FIG. 1.26: *Modèle de Markov caché comportant les positions clés de la marche pour la vue de côté[LH04].*

Modèles de Markov cachés temporels

Les modèles de Markov cachés temporels ("*Hidden Markov Model - HMM*") [BP66] consistent en un graphe orienté dont chaque nœuds représente l'état du système à un instant donné. Ces états sont raccordés par un réseau de liens qui détermine les probabilités de passage. Ce modèle

est adopté par Lan et Huttenlocher [LH04] pour suivre un sujet animé de la marche. Il comprend les postures clés de la marche observées depuis huit angles de vue, décalés chacun de quarante-cinq degrés (fig. 1.26). Les observations permettent d'associer chaque image avec un état du graphe d'après la distance de chanfrein entre le modèle "*cardboard*" et la silhouette. Le parcours dans la chaîne de Markov est retrouvé à l'aide de l'algorithme de Viterbi pour contraindre les règles articulaires et la cohérence temporelle du résultat. Un réseau HMM est également adoptée pour assurer la cohérence temporelle au cours d'un suivi de gestes quelconques [NTTC05]. Les hypothèses sont générées à partir d'un modèle hiérarchique en arbre qui représente le corps. Le torse est recherché en premier suivi des membres qui s'y rattachent.

Une autre approche consiste à utiliser des champs aléatoires conditionnels ou *Conditional Random Fields (CRF)* [LMP01] pour estimer aisément la probabilité a posteriori [TSDD06]. Cette approche à base d'apprentissage consiste à propager temporellement les hypothèses formulées dans un espace discrétisé sur les poses apprises. Similairement à [SVD03], la vraisemblance est évaluée d'après un hachage de type LSH [GIM99] pour accélérer la comparaison des hypothèses avec l'image test. Cependant, pour considérer la fonction de probabilité constante dans les zones délimités par les poses apprises, l'apprentissage doit être dense et exhaustif.

Propagation des croyances

Appliquée aux modèles graphiques, la propagation des croyances [YFW05] permet d'estimer la probabilité marginale de chaque membre. Même dans le cas de graphes bouclés, il est montré que cet algorithme est capable de converger vers une bonne approximation probabiliste de la pose après un processus itératif [Wei00]. Cette approche montante a l'avantage de limiter la dimension de l'espace d'investigation au nombre de *ddl* du membre recherché. L'algorithme assure alors la cohérence de la solution fournie vis-à-vis des contraintes articulaires grâce à des facteurs de compatibilité entre les membres. Ces facteurs vont générer des messages qui sont propagés à travers le graphe pour exprimer l'influence de chaque membre sur le reste du corps. Cette méthode est adoptée pour assurer un suivi récursif sur une séquence d'images [BCMC06][GS04][SBR+04], ou en intégrant au modèle graphique une fenêtre temporelle [GS04][SBR+04] (fig. 1.25) et des contraintes de non-collision entre les membres [BCMC06].

La propagation des croyance s'exprimant plus aisément dans les espaces discrets, les auteurs qui y font appel utilisent un espace discrétisé par une grille [GS04] ou des filtres à particules qui fournissent un ensemble d'échantillons pondérés pour explorer la vraisemblance du modèle [BCMC06]. Dans le cas d'un espace continu, les messages peuvent être modélisés par des mélanges de Gaussiennes [SBR+04] et leur mise à jour crée une explosion du nombre de Gaussiennes du fait de la multiplication des mélanges entre eux. Pour empêcher cela, les auteurs font appel à un échantillonneur de Gibbs [Isa03]. Une approche analogue à la propagation des croyances consiste à mêler l'algorithme du champ moyen aux techniques de Monte-Carlo [WHY03]. Cette approche qui utilise les niveaux de gris et les contours est limitée aux poses en 2D du fait de son modèle.

1.4.4 Approches multi-critères à base de règles

Ces approches ce basent sur un ensemble de règles qui s'appuient sur plusieurs indices pour constituer des assemblages cohérents en estimation de pose.

La fusion des observations sur l'amplitude des contours, la forme et le gradient de luminosité créé par l'ombre sur un membre comparé à une base d'apprentissage (fig. 1.27), permet de trouver de manière opportuniste un ensemble de membres candidats sur une image segmentée [MREM04]. Le nombre de combinaisons est réduit en faisant des hypothèses sur la longueur moyenne des

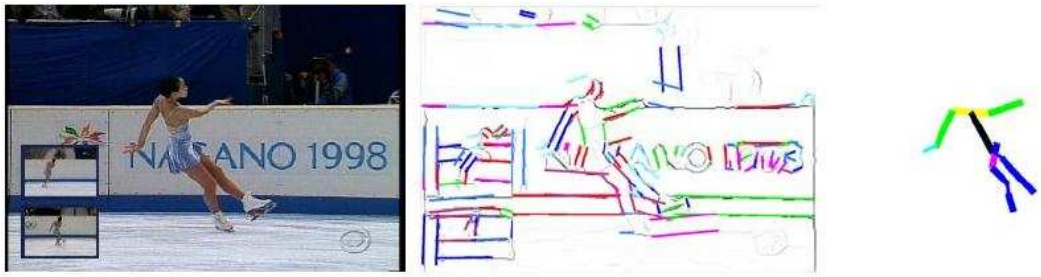


FIG. 1.28: *Reconnaissance des membres avec une approche à base de règles. De gauche à droite : image originale, sélection des membres candidats d'après des critères géométriques sur les segments extraits de l'image, sélection finale des membres et résultat de la pose en 2D d'après des critères anthropomorphiques [RBM05].*

membres, leur emplacement (pas de membres isolés) ou la symétrie des vêtements.

Une approche analogue consiste à désigner les membres candidats à partir d'hypothèses sur des lignes parallèles issues d'une extraction de contours suivie d'une triangulation de Delaunay [RBM05]. Similairement à l'algorithme précédent, dans [MREM04] les critères de sélection sur les contours incluent l'orientation, la longueur, l'amplitude ou la distance entre les centres des segments. La réduction du nombre d'hypothèses peu probables se fait sur des critères anthropomorphiques et des contraintes de connexion apprises sur une base issue de séquences de patinage artistique (fig. 1.28).

FIG. 1.27: *Base d'apprentissage sur les membres inférieurs [MREM04].*

L'intégration d'un module d'intelligence artificielle par l'intermédiaire de tableaux noirs hiérarchisés permet de rationaliser le processus de désignation et d'élimination des membres candidats [LV04]. Ce processus est conduit par trois niveaux de sources de connaissance, chacun spécialisé dans une tâche précise. Au niveau le plus bas, les spécialistes sont chargés de détecter les membres de manière opportuniste : bras, torse, tête... Le niveau intermédiaire a pour rôle de construire des membres complets : épaule + bras + avant-bras + main. Le plus haut niveau doit reconstituer un buste cohérent avec ces éléments. Les niveaux inférieurs sont sollicités par les niveaux supérieurs en fonction du contenu des tableaux noirs de manière à avoir assez d'hypothèses pour réussir la reconstruction. Un ensemble de règles d'ordre et d'adjacence sur les membres permet de mener à bien ces tâches.

1.4.5 Approches à base de "template matching"

Ces approches recherchent directement les membres [RMR04] ou le corps entier [DLF05] à partir de la scrutation de l'image avec des gabarits adaptés en utilisant des masques probabilistes adaptés aux membres recherchés [RMR04] (fig. 1.14) ou des gabarits spatio-temporels pour détecter un sujet marchant [DLF05] (fig. 1.29).

Les masques probabilistes [RMR04] consistent à calculer, au centre et sur les bords du masque,

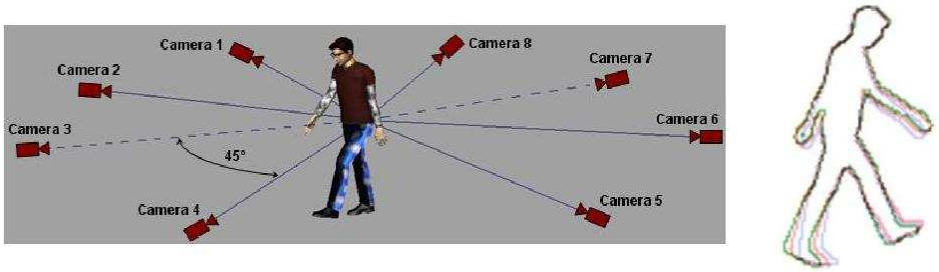


FIG. 1.29: *Création d'un gabarit spatio-temporel pour la base d'apprentissage. Image de gauche : prise de vue depuis plusieurs angles Image de droite : le gabarit inclut trois silhouettes consécutives centrées sur une position clé définie par l'instant où les deux pieds touchent le sol. Les trois silhouettes superposées mettent en évidence l'évolution temporelle du mouvement [DLF05].*

deux histogrammes de couleurs. L'homogénéité chromatique de la vignette est testée en comparant le résultat à un modèle appris pour distinguer les cas où le masque est centré sur un membre (fortement homogène) ou non (peu homogène). Des contraintes sur les liaisons entre les membres et la symétrie des vêtements permettent d'affiner le processus de décision.

Dans le cas des gabarits spatio-temporels, la base apprise est composée des contours de la silhouette d'un sujet marchant sur un tapis roulant et vu de plusieurs angles. Une séquence de trois silhouettes consécutives autour d'une position clé pour laquelle les deux pieds touchent le sol constitue un élément de la base d'apprentissage et donc un gabarit (fig. 1.29). La comparaison entre la base apprise et la séquence test est opérée à partir du calcul d'une distance de chanfrein robuste qui pénalise les contours possédant une orientation trop différente [DLF05].

1.5 Conclusions

1.5.1 Faire le chemin à l'envers

Après avoir exposé un état de l'art qui parcourt les techniques existantes en allant de l'extraction des caractéristiques aux outils de manipulation des paramètres du modèle, en passant par la modélisation elle-même, il est maintenant temps de synthétiser ces différentes approches pour en tirer des solutions qui permettront d'atteindre les objectifs fixés par la thèse. Dans cette optique, les choix les plus déterminants concernent les outils utilisés pour manipuler les paramètres du modèle et retrouver la pose à partir des caractéristiques extraites de l'image. Ces choix influencent tous les autres, c'est donc à partir de là que cette conclusion débute.

1.5.2 Choix des outils de manipulation des paramètres du modèle

Les approches déterministes

Ces méthodes sont principalement utilisées en stéréo [DD02][DKD03][PF03] ou en trinoculaire [UF04]. Dans cette configuration, elles consistent principalement à faire "coller" un modèle 3D sur l'image des disparités par des méthodes classiques d'optimisation (voir §1.4.1). Une approche originale utilise l'ICP [BM92] pour parvenir à un suivi temps réel [DD02] et tient compte des contraintes articulaire avec une seconde optimisation dans un espace de machines à vecteurs de supports apprises sur ces contraintes [DKD03].

En monoculaire, les ambiguïtés sont telles (fig 1.24) que les indices extraits de l'image, un flot optique [BM98] ou des silhouettes [ST02], sont insuffisants pour empêcher le processus d'optimisation de rester bloquer sur un des nombreux optimums locaux ou résoudre les occultations.

L'apprentissage

Exception faite de [AT06a], les approches qui utilisent des régressions [AT06b] ou des variétés [EL04][TLS05][UFHF05] se restreignent généralement à un sous-ensemble de gestes souvent cycliques. [AT06a] propose une approche capable d'estimer des poses quelconques. L'influence des distracteurs (fond, vêtements) étant en partie neutralisée par une *NMF* [Hoy04], l'apprentissage peut se concentrer sur la couverture de l'espace des poses. Le fait que la solution trouvée par l'algorithme soit parfois éloignée de la pose test montre la difficulté à couvrir exhaustivement l'espace des poses par ce type d'apprentissage.

L'extrapolation des poses apprises, pour les adapter à l'image test, permet de combler certaines lacunes dans la base. Dans un premier cas, il s'agit d'une transformation géométrique sur les points de la silhouette, mais cette approche se borne à un recueil de photographies artistiques avec des parti-pris esthétiques contraignants (fond blancs, poses standards etc) [MM02]. Une seconde approche allie l'extrapolation des poses à une très grande base dont le procédé de comparaison image/base est accéléré par *LSH* [GIM99]. Un sous ensemble de poses candidates est dégagé pour fournir une solution issue d'une extrapolation sur ces poses [SVD03]. Même avec une très grande base, ici 150 000 exemples, les meilleurs *matching* avant extrapolation sont parfois éloignés de la pose test. L'extrapolation parvient parfois à faire des petits "miracles" mais la règle générale veut que la dimension élevée de l'espace des poses, la variabilité de l'habillement, des cheveux et de l'environnement ne permette pas une généralisation suffisante à partir des bases d'apprentissage.

Les approches stochastiques

Les méthodes stochastiques tentent de modéliser la probabilité a posteriori du modèle. Le plus souvent, cette modélisation est discrète et fait appel à un échantillonnage mais dans certains cas, elle est à base de noyaux [DTS+05][SBR+04]. Pondéré par des connaissances a priori, ce modèle, en plus d'être capable de fournir un optimum, peut aussi donner une moyenne sur les hypothèses émises. Le principal avantage de ces méthodes réside dans le fait qu'elles soient multi-hypothèses et peuvent désambiguïser au cours du temps les poses issues de captures monoculaires. Au début des années 2000, une solution parvenait à un suivi de gestes quelconques dans un environnement non contraint [ST01][ST03] (fig. 1.30) grâce à un échantillonnage d'importance de type filtre à particules [BI96]. Mais l'algorithme est trop lent pour traiter les images en ligne et la robustesse est insuffisante pour permettre un suivi le long d'une scène complète. La raison de ces problèmes provient de l'impossibilité d'échantillonner un espace dont la dimension est de l'ordre de plusieurs dizaines d'unités. Des améliorations comme l'échantillonnage le long des vallées de vraisemblances [ST01] ou le saut cinématique [ST03] ne suffisent pas à faire sortir l'échantillonnage hors d'un sous espace restreint.

Les approches ascendantes sont à l'origine d'une solution à ce problème. L'idée est de rechercher les membres indépendamment en accumulant des hypothèses pour chacun d'entre-eux et, avec ces hypothèses qui seraient les pieces d'un puzzle, de reconstituer un corps entier et cohérent. Des approches arborescentes tentent de conduire cette reconstruction [IF01][NTTC05] mais avec un ordre fixe. Or, l'ordre optimal dépend de l'arrangement pictural de la scène à reconstituer, et particulièrement de ce que Mori et al appellent "*islands of saliency*" [MREM04]. C'est là où les graphes non orientés apportent un plus en levant la contrainte d'un ordre donné. Pour conduire la reconstruction, ce type de graphe peut faire appel à l'algorithme de propagation des croyances [YFW05] pour tenir compte de l'influence d'un membre sur ces voisins. Cette méthode à été implémentée avec succès avec des approches stéréo [BCMC06] ou multi-caméra [SBR+04]. Son exploitation en monoculaire a également donné des résultats remarquables [GS04] avec une gestion des occultations [SB06] mais ces deux algorithmes sont lents et ils ne font appel qu'aux indices d'énergie de mouvement et de teinte chair pour le premier. Le second exploite une soustraction de fond associée à un modèle de couleur et de teinte chair mais, à moins d'une mise à jour du modèle, le peu de constance des couleurs au fil des images peuvent le mettre en échec.

Les autres méthodes

D'autres méthodes ascendantes mettent en œuvre un ensemble de règles pour conduire l'extraction d'hypothèses sur les membres. L'assemblage d'une solution cohérente avec les observations menées sur l'image fait appel à la force brute [MREM04][RBM05]. Une alternative consiste à utiliser l'intelligence artificielle pour mener à bien cette tâche [LV04]. Cependant, l'efficacité de cette dernière approche réside principalement dans le choix des règles qui s'avère souvent complexe.

1.5.3 Choix du modèle

En vision monoculaire, les modèles 2D ont tendance à être mis en échec lorsque l'information de profondeur est déterminante. Cela se produit particulièrement quand l'axe d'un membre devient perpendiculaire au plan de l'image. Dans ce cas, la surface occupée par le membre dans l'image est restreinte et les contraintes anthropomorphiques ramenées en 2D ne sont pas suffisamment discriminantes.

Dans le cadre des approches probabilistes multi-hypothèses, le modèle 3D est couramment utilisé [SBR+04][ST03]. L'utilisation d'un modèle 2D de type "cardboard" n'est possible qu'avec l'agorithme de Taylor [Tay00] qui suppose connues la taille des membres et la configuration de la pose pour lever les ambiguïtés [GS04].

Si l'environnement n'est pas contraint, les vêtements varient beaucoup en fonction des personnes et des poses. Dans ce contexte, l'apparence des membres sur l'image est très variable et le modèle 3D ultraprécis composé de métasphères [PF03] n'apporte guère d'améliorations. En revanche, les troncs, de cônes [SBR+04], voire des cylindres semblent être un bon compromis entre précision et simplicité.

FIG. 1.30: *Sminchisescu et Triggs [ST01][ST03] : suivi monoculaire en 3D. Les améliorations sur le reéchantillonnage du filtre à particules permet d'offrir un environnement peu contraint mais la robustesse reste insuffisante.*

1.5.4 choix des caractéristiques extraites de l'image

Lorsque l'information manque, il faut savoir la trouver ailleurs. Ce principe s'applique à la vision monoculaire où, contrairement à la vision stéréoscopique ou multi-caméra, l'indice de profondeur pour la désambiguïsation des poses et des occultations fait défaut. De plus, dans le cas de l'estimation de la pose dans une image, l'information temporelle est inexistante. Ceci pousse à l'accumulation de détecteurs complémentaires pour gagner en robustesse. Dans ce cadre, les approches se basent sur une segmentation en zones homogènes de couleurs [MREM04], la recherche de lignes parallèles d'après l'extraction des contours [RBM05] l'évaluation de directions globales d'après une transformée de Hough droite sur les gradients de contours locaux [LV04] ou la concaténation d'histogrammes des orientations de contours calculés sur plusieurs échelles [AT06a][SVD03]. Dans tous ces cas, le pouvoir discriminant est influencé par la complexité du fond et le risque de se retrouver noyé sous les fausses alertes est grand. Pour éviter cela, la segmentation du personnage par soustraction du fond [LV04][ST02] permet de restreindre l'espace

de recherche et ainsi diminuer le taux des fausses alertes. Certes, la soustraction de fond implique une caméra fixe mais cette hypothèse est compatible avec les objectifs de la thèse. La silhouette, ainsi découpée et initialisée selon une pose judicieusement choisie, peut fournir des informations utiles pour constituer un modèle de couleurs des vêtements ou de la teinte du visage [LC04][LV04], ou encore déterminer des proportions entre les membres.

En incluant l'information temporelle pour le suivi, il devient possible d'obtenir des indices sur le mouvement dans l'image. Ceux-ci peuvent utilement informer sur l'intensité du mouvement grâce à la différence d'image [GS04] ou sur son intensité et sa direction [BM98][ST01] grâce au flot optique.

Parmi ces pistes, les contours, moins influencés que les couleurs par les variations lumineuses, paraissent d'un grand intérêt, notamment lorsqu'ils sont utilisés pour dégager des directions globales [LV04], ou sous la forme d'histogrammes [AT06a]. Dans ce dernier cas, les histogrammes normalisés procurent une représentation compacte et fidèle de l'image avec une robustesse supplémentaire face aux légers mouvements et aux variations de contraste. L'ajout d'un détecteur de teinte chair pour la détection du visage et des mains [GS04][LC04][LV04] permet de bien contraindre la position des membres supérieurs. Cependant, il est fréquent qu'un bras habillé de la même couleur que le torse te replié devant lui ne permette pas aux contours d'être suffisamment discriminants. Dans ce cas, l'image de l'énergie du mouvement [GS04] est une solution simple et efficace pour contraindre les membres vers les zones de mouvement.

1.5.5 choix effectués

Les exigences établies dans le cadre de cette thèse imposent un suivi quasi temps réel à partir d'une caméra monoculaire dans un environnement le moins contraint possible. À la lecture des nombreux articles, le modèle stochastique avec sa capacité de propager plusieurs hypothèses semble s'imposer pour sa capacité à désambiguïser une image monoculaire. La propagation des croyances dans les graphes de facteurs [YFW05] semble être une solution adéquate pour diminuer la dimension de l'espace exploré avec une approche ascendante. Cet algorithme s'exprimant plus facilement dans les espaces discrets, il paraît judicieux d'utiliser une méthode d'échantillonnage. De plus, afin de limiter les itérations nécessaires lors de la propagation des croyances sur une fenêtre temporelle [SBR+04], une estimation récursive de la pose [BCMC06] est préférable afin de respecter la contrainte temps réel. Ce type d'approches implémentées en monoculaire utilisent généralement un nombre d'indices restreint [GS04][SB06] alors qu'une gestion correcte des occultations et des ambiguïtés peut exiger d'étendre les indices, par exemple, à une soustraction de fond pour neutraliser les distracteurs issus du fond, ou aux indices de contours qui sont préférable aux couleurs variant excessivement avec la lumière ambiante. L'apprentissage semble être une voie à ne pas négliger à la lumière des progrès apportés par les régressions [AT06a] ou des approches mixtes intégrant le PSH [DKD03][TSDD06].

Dans le cadre de la vision 3D en monoculaire, le choix du modèle semble s'orienter naturellement vers un modèle 3D. Celui-ci est préférable aux modèles 2D de type "cardboard" [JBY96] qui doivent souvent intégrer l'algorithme de Taylor [Tay00] peu réaliste d'un point de vue pratique.

Chapitre 2

L'algorithme de suivi

Sommaire

2.1 Introduction

Le chapitre bibliographique (chapt. 1) présente les enjeux du suivi du corps humain. Ce problème générant des espaces de grande dimension avec un nombre important d'inconnues à déterminer, les techniques stochastiques utilisant la simulation par échantillonnage doivent pour être viables, s'opérer dans des sous-ensembles d'inconnues, idéalement les membres, plutôt que sur le corps entier. L'adoption d'un tel principe exige de procéder à un synthèse des résultats trouvés pour chacun des membres afin de générer une solution cohérente en regard des contraintes et des limites articulaires.

L'association du filtre à particules [BI96] et de la propagation des croyances [YFW05] se prête bien à ces exigences. Un filtre à particules par membre génère des hypothèses indépendamment pour chacun des membres, la validité de celles-ci dépendent des observations tirées de l'image (chapt. 3) et du respect des contraintes articulaires implémentées dans la propagation des croyances.

Cette association présente l'avantage de générer un espace discret dans lequel la propagation des croyances s'effectue simplement sur les échantillons tirés par les filtres à particules. Les hypothèses sont propagées temporellement par filtrage particulaire et spatialement, à travers les membres du corps, par la propagation des croyances. L'algorithme résultant est rapide et modulable, pouvant intégrer aisément de multiples indices et contraintes ainsi qu'un terme issu, par exemple, de l'apprentissage afin d'améliorer les performances lorsque les observations tirées de l'image deviennent difficiles à interpréter du fait des occultations.

Ce chapitre présente l'algorithme de suivi utilisé en exposant le modèle graphique (§2.2) dans lequel va s'effectuer la propagation des croyances (§2.2.3) puis, dans le paragraphe suivant (§2.3), comment l'utilisation du filtre à particules (§2.3.1) va permettre à cet algorithme de s'effectuer dans un espace discret (§2.3.2).

2.2 Modèles graphiques

La méthode statistique de suivi du corps articulé proposée ici comprends deux parties : le modèle statistique utilisé pour représenter la structure articulée et l'algorithme d'estimation de la pose. D'une manière synthétique, le corps articulé est représenté par un modèle graphique (§2.2.2) et l'estimation des paramètres fait appel à l'algorithme de propagation des croyances (§2.2.3) dans un espace discrétisé par l'utilisation de filtres à particules (§2.3). Les paragraphes 2.2.2, 2.2.3 et 2.3.2 sont écris d'après la méthode proposée par Bernier et Cheung-Mon-Chang dans [BCMC06]. Ce chapitre débute par un paragraphe rappelant quelques généralités (§2.2.1).

2.2.1 Généralités

Rappels sur la règle de Bayes et la marginalisation

Soient a et b deux variables aléatoires, la règle de Bayes s'écrit :

$$P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b)}. \tag{2.1}$$

En outre, la marginalisation de $P(a)$ sur la variable b consiste à calculer :

$$P(a) = \int P(a, b)db. \tag{2.2}$$

Modélisation de l'évolution temporelle par une chaîne de Markov cachée

On désigne par X_t l'état à l'instant t qui détermine la pose et par Y_t les observations sur l'image issues de cet état. L'évolution temporelle de la pose peut être modélisée par une chaîne de markov cachée (fig. 2.1) (§1.4.3). L'ensemble des états de $t = 0$ à $t = T$ est noté $X_{0:T} = \{X_0, \ldots, X_T\}$. L'ensemble des observations de $t = 0$ à $t = T$ est noté $Y_{0:T} = \{Y_0, \ldots, Y_T\}$.

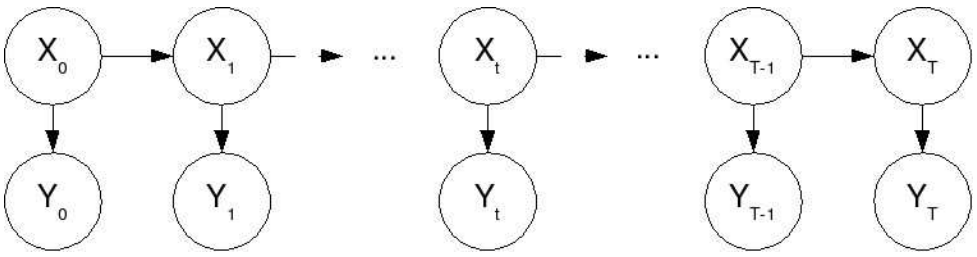


FIG. 2.1: *Chaine de Markov Cachée.*

Ce graphe implique des relations d'indépendance sur les états : $P(X_t \mid X_{0:t-1},\ Y_{0:t-1}) = P(X_t \mid X_{t-1})$ et sur les observations : $P(Y_t \mid X_{0:t},\ Y_{0:t-1}) = P(Y_t \mid X_t)$. La probabilité jointe, $P(X_{0:T})$, se factorise alors sous la forme :

$$\begin{aligned} P(X_{0:T}) &= P(X_T,\ X_{0:T-1}), \\ &= P(X_T \mid X_{0:T-1})P(X_{0:T-1}), \end{aligned} \tag{2.3}$$

or, d'après le graphe, $P(X_T)$ ne dépend que de $P(X_{T-1})$:

$$P(X_{0:T}) = P(X_T \mid X_{T-1})P(X_{0:T-1}), \tag{2.4}$$

et récursivement :

$$P(X_{0:T}) = P(X_0) \prod_{t=1}^{T} P(X_t \mid X_{t-1}). \tag{2.5}$$

Estimation de la probabilité a posteriori

À $t = T$, la probabilité d'une séquence de gestes est estimée d'après l'ensemble des observations :

$$\begin{aligned} P(X_{0:T} \mid Y_{0:T}) &= P(X_{0:T-1},\ X_T \mid Y_{0:T-1},\ Y_T), \\ &= \frac{P(X_{0:T-1},\ X_T,\ Y_T \mid Y_{0:T-1})}{P(Y_T \mid Y_{0:T-1})}, \end{aligned} \tag{2.6}$$

du fait que les observations soient connues, la probabilité $P(Y_T \mid Y_{0:T-1})$ est une constante, d'où :

$$\begin{aligned} P(X_{0:T} \mid Y_{0:T}) &\propto P(Y_T \mid X_{0:T-1},\ X_T,\ Y_{0:T-1}) P(X_{0:T-1},\ X_T \mid Y_{0:T-1}), \\ &\propto P(Y_T \mid X_T) P(X_{0:T-1},\ X_T \mid Y_{0:T-1}), \\ &\propto P(Y_T \mid X_T) P(X_T \mid X_{0:T-1},\ Y_{0:T-1}) P(X_{0:T-1} \mid Y_{0:T-1}), \\ &\propto P(Y_T \mid X_T) P(X_T \mid X_{T-1}) P(X_{0:T-1} \mid Y_{0:T-1}). \end{aligned} \tag{2.7}$$

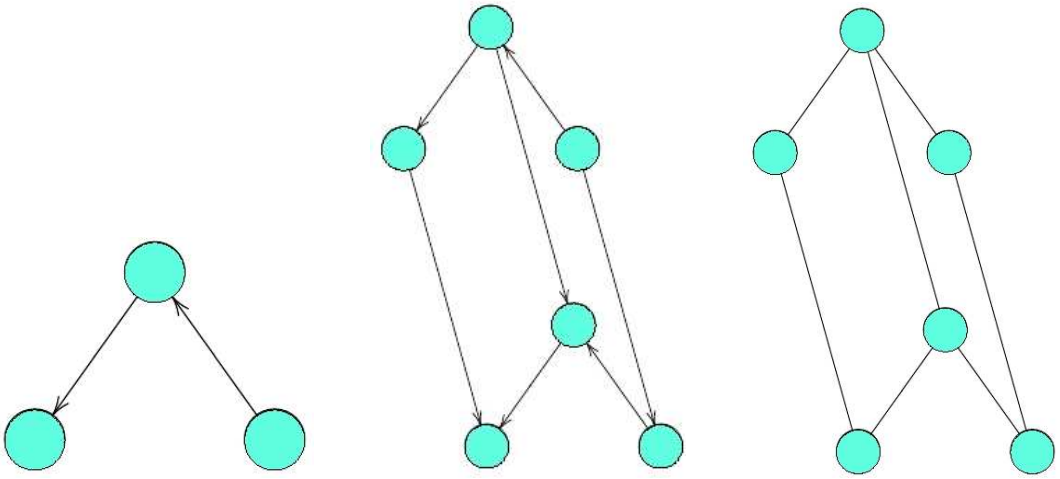
Cette expression montre que la probabilité a posteriori à l'instant T est égale à la vraisemblance du modèle $P(Y_T \mid X_T)$ pondérée par les connaissances a priori $P(X_T \mid X_{T-1})$ et la probabilité a posteriori calculée à l'instant $T-1$. Le modèle d'évolution temporel par une chaîne de Markov cachée aboutit donc une estimation récursive de la pose. Ce fait permet de rechercher la pose à partir de l'estimation précédente et non plus sur une trajectoire complète de 0 à T.

2.2.2 Modèle graphique du corps

La façon la plus simple de représenter une structure articulée par un modèle graphique consiste à utiliser un réseau Bayésien arborescent où les nœuds représentent les membres et les liens les articulations (fig. 2.2a) [Pea88]. La vraisemblance d'un tel modèle est efficacement estimée pour une forme particulière des liens (une distance entre les membres liés) [FH05] mais la nécessité de respecter la cohérence temporelle entre deux images consécutives conduit à ajouter des liens temporels pour relier un même membre entre ces deux instants (fig. 2.2b). De ce fait, le graphe ne possède plus sa structure arborescente rendant caduque la solution précédemment envisagée. Ce modèle peut être avantageusement remplacé par un champ de Markov aléatoire ou "Markov Random Field - MRF" [Li95] (fig. 2.2c) mais le fait d'intégrer des contraintes entre les membres, comme des limites articulaires ou des contraintes de non collision (§3.4.5 et fig. 2.2d) peut amener des cliques d'ordre supérieur à deux. Un graphe de facteurs (fig. 2.2e) permet de modéliser ces cliques en introduisant des facteurs reliant l'ensemble des membres d'une même clique [KFL01]. Dans le cas de cliques d'ordre supérieur à deux, la probabilité jointe, qui s'exprime comme un produit de ces facteurs positifs (rectangles noirs), est plus lourde à calculer puisque l'estimation d'un facteur implique de considérer les combinaisons sur tous les états des membres qu'il relie. Le graphe de facteurs permet d'imposer des cliques d'ordre deux afin qu'un facteur ne puisse relier que deux nœuds pour être calculé simplement (fig. 2.2f). L'estimation et les inférences sur ce type de graphe peut être mené par l'algorithme de propagation des croyances [KFL01].

Le modèle à base d'un graphe de facteurs adopté ici distingue trois types de facteurs : les facteurs reliant les états de deux membres adjacents pour une même image sont appelés *facteurs*

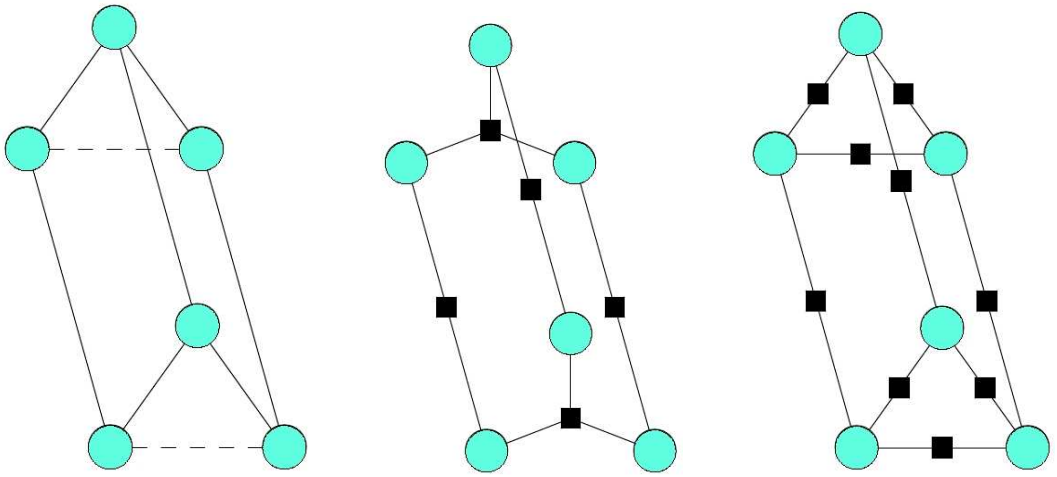
de liens, ceux entre un même membre représenté sur deux images adjacentes sont les *facteurs de cohérence temporelle* et ceux (non représentés sur la figure 2.2f) entre un membre et son observation sont les *facteurs de compatibilité à l'image*. Les observations correspondent aux indices extraits de l'image (voir chapitre 3).



(a) Réseau Bayésien arborescent.

(b) Réseau Bayésien pour deux images.

(c) MRF pour deux images.



(d) MRF pour deux images et un lien supplémentaire.

(e) Graphe de facteurs équivalent pour deux images et un lien supplémentaire.

(f) Graphe de facteurs pour deux images et un lien supplémentaire comportant uniquement des facteurs reliés à deux nœuds .

FIG. 2.2: *Modèles graphiques pour une structure à trois membres articulés. Pour la clarté de la figure, les observations associées aux états sont omises.*

Dans le cadre du suivi de pose articulée, on cherche la probabilité de la pose du corps connaissant les observations de $t = 0$ à $t = T$. En considérant l'état X_t^μ du membre μ au temps t et les observations associées Y_t^μ, la probabilité conditionnelle recherchée s'écrit à $t = T$:

$$P(\{X_t^\mu\}_{0:T}^\mu | \{Y_t^\mu\}_{0:T}^\mu) = \frac{P(\{X_t^\mu\}_{0:T}^\mu, \{Y_t^\mu\}_{0:T}^\mu)}{P(\{Y_t^\mu\}_{0:T}^\mu)} \tag{2.8}$$

Du fait que les observations Y soient connues lors du suivi, la probabilité conditionnelle est proportionnelle à la probabilité jointe des états et des observations qui leur sont associées (le numérateur de 2.8). C'est cette dernière probabilité qui est généralement considérée dans le cadre des modèles génératifs comme les MRFs. Cependant, les modèles conditionnels comme les CRFs [LMP01] sont mieux adaptés au problème de suivi connaissant les observations puisqu'ils fournissent directement l'estimation de la probabilité conditionnelle. Ce modèle peut être représenté rigoureusement par un graphe de facteurs avec l'avantage d'éluder la normalisation nécessaire à l'expression des probabilités d'observation [TSDD06]. Le graphe de facteurs permet de décomposer la probabilité conditionnelle en produit de facteurs positifs de deux variables :

$$P(\{X_t^\mu\}_{0:T}^\mu|\{Y_t^\mu\}_{0:T}^\mu) \propto \left[\prod_{t=0}^{T}\prod_{\mu} K^\mu(X_t^\mu, Y_t^\mu)\right]\left[\prod_{t=0}^{T}\prod_{(\mu,\nu)\in\Gamma} L^{\mu,\nu}(X_t^\mu, X_t^\nu)\right]\left[\prod_{t=1}^{T}\prod_{\mu} H^\mu(X_t^\mu, X_{t-1}^\mu)\right], \quad (2.9)$$

où K^μ est le facteur de compatibilité à l'image entre l'état du membre μ et les observations correspondantes, $L^{\mu,\nu}$ est le facteur de lien entre les membres μ et ν, Γ étant l'ensemble des liens entre membres et H^μ est le facteur de cohérence temporel pour le membre μ.

2.2.3 Propagation des croyances

Ce graphe de facteurs permet d'obtenir les probabilités marginales en utilisant l'algorithme de propagation des croyances [KFL01]. Cependant, comme le graphe inclut des boucles, les probabilités obtenues ne sont généralement qu'une approximation qui dépend de l'ordre dans lequel les messages sont propagés dans le graphe [YFW05]. Cette approximation, pour les MRFs, est reliée à l'approximation de Bethe de l'énergie libre ce qui peut expliquer pourquoi cette approximation est souvent valide en pratique.

Dans le contexte du suivi du corps, l'intérêt se porte particulièrement sur l'estimation marginale de la probabilités pour chaque état qui paramètre la pose d'un membre à l'instant présent, celle-ci étant conditionnée à toutes les observations cumulées jusqu'à cet instant. Pour simplifier l'algorithme et alléger la charge de calcul, les messages sont d'abord propagés dans une même tranche temporelle du graphe avec un nombre fixé d'itérations (10 dans notre cas), puis ils sont propagés unidirectionellement vers l'image suivante. Cette façon de procéder fait que l'estimation d'une probabilité marginale à $T-1$ connaissant toutes les observations jusqu'à T ne dépend pas des observations après $T-1$. L'estimation des marginales au temps T (connaissant toutes les observations jusqu'à T) peut être calculée récursivement à chaque nouvel instant T d'après la précédente approximation (celle à $T-1$ sachant les observations à $T-1$). L'approximation obtenue par la propagation unidirectionnelle des hypothèses dans le temps ne respecte pas le procédé de propagation des croyances standard dans les graphes bouclés. La validité de cette démarche suppose que la pose est estimée à l'instant présent en connaissant les observations jusqu'à cet instant, et non celles qui seront vues dans le futur. Cette hypothèse est compatible avec les objectifs temps réel fixés dans le cadre de ce travail. En conséquence, les probabilités exprimées à l'instant t seront implicitement considérées comme conditionnées aux observations passées et présentes. Dans ce qui suit, aucune distinction ne sera faite entre l'instant T, avant et pendant lequel les observations sont connues, et l'instant d'estimation t. Nous allons maintenant décrire l'algorithme de propagation des croyances appliqué à notre cas.

Supposons qu'à l'instant $t-1$, les messages sont propagés et les marginales P_{t-1}^μ sont estimées pour chaque membre. L'expression du message reçu par un membre à l'instant t de la part du

facteur le reliant à son homologue à $t-1$ est :

$$r^{\mu}_{t-1\to t}(X^{\mu}_t) = \int dX^{\mu}_{t-1} H^{\mu}(X^{\mu}_t, X^{\mu}_{t-1}) s^{\mu}_{t-1\to t}(X^{\mu}_{t-1})\,, \tag{2.10}$$

où $s^{\mu}_{t-1\to t}$ est le message émis du passé vers le facteur reliant un même membre entre $t-1$ et t :

$$s^{\mu}_{t-1\to t}(X^{\mu}_{t-1}) = \prod_{\nu'\in\Gamma_{\mu}} r^{\nu'\to\mu}_{t-1}(X^{\mu}_{t-1})\,, \tag{2.11}$$

en considérant que Γ_{μ} est l'ensemble des membres reliés à μ. D'après [YFW05], et après convergence pour une tranche temporelle, la probabilité marginale de ce membre à l'instant $t-1$ est :

$$P^{\mu}_{t-1}(X^{\mu}_{t-1}) \propto r^{\mu}_{t\to t-1}(X^{\mu}_t) \prod_{\nu'\in\Gamma_{\mu}} r^{\nu'\to\mu}_{t-1}(X^{\mu}_{t-1})\,, \tag{2.12}$$

mais comme les messages ne sont jamais propagés vers le passé, $r^{\mu}_{t\to t-1}$ peut être omis :

$$P^{\mu}_{t-1}(X^{\mu}_{t-1}) \propto \prod_{\nu'\in\Gamma_{\mu}} r^{\nu'\to\mu}_{t-1}(X^{\mu}_{t-1})\,, \tag{2.13}$$

donc, le message émis du passé après convergence est :

$$s^{\mu}_{t-1\to t}(X^{\mu}_{t-1}) \propto P^{\mu}_{t-1}(X^{\mu}_{t-1})\,, \tag{2.14}$$

en considérant les équations 2.10 et 2.14, on trouve :

$$r^{\mu}_{t-1\to t}(X^{\mu}_t) \propto \int dX^{\mu}_{t-1} H^{\mu}(X^{\mu}_t, X^{\mu}_{t-1}) P^{\mu}_{t-1}(X^{\mu}_{t-1})\,. \tag{2.15}$$

Les messages reçus depuis les facteurs vers les nœuds qui représentent les états des membres à t, sont initialisés avec une distribution uniforme (on suppose que l'espace des états est borné). Un message de ce type reçu par le membre μ est noté $r^{\nu\to\mu}_t$ lorsqu'il est issu du facteur reliant le membre μ au membre ν. Le message $s^{\mu\to\nu}_t$ envoyé par le nœud correspondant à l'état du membre μ vers le facteur qui le relie au membre ν est mis à jour suivant :

$$s^{\mu\to\nu}_t(X^{\mu}_t) \leftarrow r^{\mu}_{t-1\to t}(X^{\mu}_t) K^{\mu}(X^{\mu}_t, Y^{\mu}_t) \prod_{\nu'\in\Gamma_{\mu,\nu}} r^{\nu'\to\mu}_t(X^{\mu}_t)\,, \tag{2.16}$$

où $\Gamma_{\mu,\nu}$ est l'ensemble des membres reliés au membre μ sauf ν. Tous les messages envoyés sont mis à jour parallèlement en appliquant l'équation 2.16. Il en va de même pour les messages reçus qui obéissent à l'opération de mise à jour suivante :

$$r^{\nu\to\mu}_t(X^{\mu}_t) \leftarrow \int dX^{\nu}_t L^{\mu,\nu}(X^{\mu}_t, X^{\nu}_t) s^{\nu\to\mu}_t(X^{\nu}_t)\,. \tag{2.17}$$

Ce processus de mise à jour peut être itéré jusqu'à convergence, mais pour limiter la charge de calcul, il est répété pour un petit nombre d'itérations après lesquelles l'estimation des probabilités marginales pour l'état d'un membre μ à l'instant t est donné par :

$$P^{\mu}_t(X^{\mu}_t) \propto r^{\mu}_{t-1\to t}(X^{\mu}_t) K^{\mu}(X^{\mu}_t, Y^{\mu}_t) \prod_{\nu\in\Gamma_{\mu}} r^{\nu\to\mu}_t(X^{\mu}_t)\,. \tag{2.18}$$

La probabilité marginale est connue à un facteur multiplicatif près qui peut être facilement retrouvé en considérant que l'intégrale de la probabilité marginale $P^{\mu}_t(X^{\mu}_t)$ sur X^{μ}_t doit être égale à 1.

Ces équations montrent que l'estimation des marginales à t ne dépendent pas du futur et sont possible si les marginales à $t-1$ sont disponibles. De cette façon, une estimation récursive de la pose est obtenue, ouvrant la voie à un traitement en ligne des vidéos.

2.3 Représentation particulaire

Comme l'espace d'état des membre est continu, les expressions manipulés dans le paragraphe précédent (§2.2 : densités de probabilité, messages ou facteurs) sont rarement calculables analytiquement. Une méthode d'approximation doit être choisie et la solution adoptée ici consiste à utiliser l'échantillonnage pour les approximer, en particulier le filtre à particules qui est un échantillonnage d'importance séquentiel. Ce paragraphe s'articule autour d'une explication de la méthode du filtre à particules (§2.3.1) et son adaptation à la propagation des croyances dans les espaces discrets (§2.3.2).

2.3.1 Filtre à particules

Suivi par échantillonnage d'importance séquentiel

En pratique, l'expression 2.7 n'est pas calculable analytiquement. L'utilisation du filtre à particules apporte une solution en modélisant cette distribution grâce à un échantillonnage d'importance (annexe A.2). Dans un premier temps, des échantillons $\{X^i_{0:t}\}$ de trajectoires complètes de 0 à t sont générés suivant une distribution de proposition, la probabilité a posteriori $P(X_{0:t} \mid Y_{0:t})$ est modélisée par la somme des échantillons pondérés par les poids $\{w^i_t\}$:

$$P(X_{0:t} \mid Y_{0:t}) \simeq \sum_i w^i_t \delta(X_{0:t} - X^i_{0:t})\,. \tag{2.19}$$

La distribution de proposition $q^t(X_{0:t} \mid Y_{0:t})$ est choisie en faisant l'hypothèse qu'elle obéit à un processus markovien. Cette distribution se factorise alors sous la forme :

$$\begin{aligned} q^t(X_{0:t} \mid Y_{0:t}) &= q^t(X_t,\ X_{0:t-1} \mid Y_{0:t}), \\ &= q^t(X_t \mid X_{0:t-1},\ Y_{0:t}) q^{t-1}(X_{0:t-1} \mid Y_{0:t}), \\ &= q^t(X_t \mid X_{0:t-1},\ Y_{0:t}) q^{t-1}(X_{0:t-1} \mid Y_{0:t-1}). \end{aligned} \tag{2.20}$$

Les trajectoires $X_{0:t}$ sont tirées sous la forme d'échantillons pondérés d'après cette distribution de proposition :

$$X^i_{0:t} \overset{iid}{\sim} q^t(X_{0:t},\ Y_{0:t}), \tag{2.21}$$

avec les poids :

$$w^i_t = \frac{P(X^i_{0:t} \mid Y_{0:t})}{q^t(X^i_{0:t} \mid Y_{0:t})}. \tag{2.22}$$

Le fait de tirer des trajectoires complètes va entraîner un accroissement de la dimension de l'espace des états au fur et à mesure des images. La parade consiste à tirer non plus des trajectoires mais des extensions à celles-ci pour aboutir à un suivi séquentiel de la pose. Le résultat trouvé à l'équation 2.7, montre que le numérateur de l'équation 2.22 peut s'écrire sous la forme :

$$P(X^i_{0:t} \mid Y_{0:t}) \propto P(Y_t \mid X^i_t) P(X^i_t \mid X^i_{t-1}) P(X^i_{0:t-1} \mid Y_{0:t-1}). \tag{2.23}$$

En injectant les résultats des équations 2.23 et 2.20 dans l'équation 2.22 des poids, on obtient :

$$w^i_t \propto \frac{P(Y_t \mid X^i_t) P(X^i_t \mid X^i_{t-1})}{q^t(X^i_t \mid X^i_{0:t-1},\ Y_{0:t})} \times \frac{P(X^i_{0:t-1} \mid Y_{0:t-1})}{q^{t-1}(X^i_{0:t-1} \mid Y_{0:t-1})}, \tag{2.24}$$

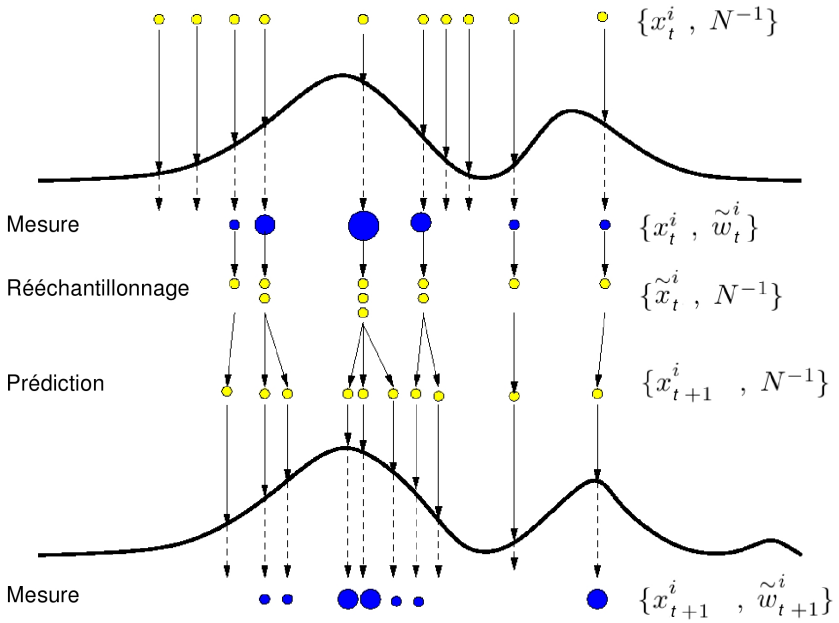


FIG. 2.3: *Échantillonnage d'importance séquentiel avec rééchantillonnage.*

d'après l'équation 2.22, le second membre obtenu ci-dessus exprime le poids w_{t-1}^i. Ce résultat permet d'aboutir à une expression récursive des poids :

$$w_t^i \propto \frac{P(Y_t \mid X_t^i)P(X_t^i \mid X_{t-1}^i)}{q^t(X_t^i \mid X_{0:t-1}^i,\ Y_{0:t})} w_{t-1}^i. \tag{2.25}$$

Cette expression montre la possibilité de tirer des extensions aux trajectoires et non plus des trajectoires entières. À l'instant t, les nouveaux poids w_t^i associés aux extensions X_t^i sont calculés d'après les poids w_{t-1}^i obtenus à $t = t-1$. L'approximation de la nouvelle densité a posteriori est alors :

$$P(X_t \mid Y_{0:t}) \simeq \sum_i \widetilde{w}_t^i\ \delta(X_t - X_t^i), \tag{2.26}$$

avec $\widetilde{w}_t^i$, les poids normalisés :

$$\widetilde{w}_t^i = \frac{w_t^i}{\sum_j w_t^j}. \tag{2.27}$$

Dégénérescence des poids

Il a été montré que l'algorithme de suivi séquentiel par échantillonnage d'importance (§2.3.1) conduit à une dégénérescence des poids qui se traduit par leur variance qui augmente au cours des itérations. Ce phénomène peut aboutir à l'ensemble des poids qui deviennent nuls sauf un [DGA00].

Une première solution au problème de dégénérescence des poids consiste à choisir une densité de proposition qui minimise leur variance. Le choix optimal consiste à retenir q^t telle que $q^t(X_t \mid$

Algorithm 1 Algorithme SIR de type Condensation

{Prédiction, mesure}
for $i = 1$ to N **do**
  tirer $X_t^i \overset{iid}{\sim} P(X_t \mid X_{t-1}^i)$
  Calculer $w_t^i = P(Y_t \mid X_t^i) w_{t-1}^i$ {Mesure}
end for

{Normalisation}
Calculer la somme des poids $W_t \leftarrow SUM[\{w_t^i\}_{i=1}^N]$
for $i = 1$ to N **do**
  Normaliser : $\tilde{w}_t^i = w_t^i / W_t$
end for

{Rééchantillonnage}
Initialiser $c_1 = \tilde{w}_t^1$
for $i = 2$ to N **do**
  Construire la fonction cumulative $c_i \leftarrow c_{i-1} + \tilde{w}_t^i$
end for
for $i = 1$ to N **do**
  générer un nombre aléatoire d'après une densité uniforme $r \sim \mathbb{U}[0,\ 1]$
  $j \leftarrow 1$
  while $(c_j < r)$ **do**
    $j \leftarrow j + 1$
  end while
  Affecter l'échantillon $\tilde{X}_t^i \leftarrow X_t^j$
  Affecter le poids $w_t^i \leftarrow 1/N$
end for

$X_{0:t-1},\ Y_{0:t}) = P(X_t \mid X_{0:t-1},\ Y_{0:t})$ [DGA00]. Une autre option qui conduit à l'algorithme *CONDENSATION* [BI96] consiste à choisir $q^t(X_t \mid X_{0:t-1},\ Y_{0:t}) = P(X_t \mid X_{t-1})$. Dans ce cas :

$$w_t^i \propto P(Y_t \mid X_t^i) w_{t-1}^i. \tag{2.28}$$

Échantillonnage d'importance séquentiel avec réechantillonnage

Les solutions portant sur le choix de la densité de proposition conduisent à retarder le phénomène de dégénérescence sans l'empêcher. Le rééchantillonnage des N particules consiste à retirer l'ensemble des échantillons pondérés $\{X_t^i,\ \tilde{w}_t^i\}$ obtenus à l'étape t suivant la densité postérieure approximée par ces échantillons :

$$\tilde{X}_t^i \overset{iid}{\sim} \sum_i^N w_t^i \delta(X_t - X_t^i). \tag{2.29}$$

Ces échantillons sont affectés du poids N^{-1} pour donner un nouvel ensemble de particules pondérées uniformément $\{\tilde{X}_t^i,\ N^{-1}\}$ (figure 2.3). Lors du rééchantillonnage, la variance des poids

retombe à 0. On obtient alors l'algorithme d'échantillonnage d'importance séquentiel avec ré-échantillonnage ("*Sequential Importance sampling with Resampling - SIR*"). Celui-ci est écrit en pseudo code (alg. 1) sous sa variante dite Condensation [BI96] puisqu'il utilise la densité de proposition $P(X_t \mid X_{t-1})$ pour tirer les échantillons.

2.3.2 Propagation des croyances discrète

Appliquée à un modèle graphique du corps, la technique du filtre à particules (§2.3.1) permet de créer un espace discret à l'intérieur duquel les probabilités marginales et les messages générés par l'algorithme de propagation des croyances sont évalués sous la forme d'échantillons pondérés. Cette approximation permet de faciliter la mise en pratique de la propagation des croyances.

À l'instant $t-1$, une approximation $\hat{P}^{\mu}_{t-1}$ de la probabilité marginale de chaque membre μ sous la forme d'une somme d'échantillons pondérés est supposée connue :

$$\hat{P}^{\mu}_{t-1}(X^{\mu}_{t-1}) = \sum_i \hat{w}^{\mu,i}_{t-1} \delta(X^{\mu}_{t-1} - X^{\mu,i}_{t-1})\,. \tag{2.30}$$

En considérant chaque membre indépendamment, une distribution de proposition Q^{μ}_t permet de générer N_μ échantillons pour l'état conjoint $(X^{\mu}_t, X^{\mu}_{t-1})$ du membre μ aux instants t et $t-1$. Cette distribution de proposition est générée en tirant d'abord un échantillon $\dot{X}^{\mu,i}_{t-1}$ à partir de l'état du membre μ à $t-1$ en utilisant l'approximation marginale obtenue à cet instant, puis en tirant un échantillon $X^{\mu,i}_t$ pour l'état du membre μ à t sachant son état à $t-1$ en utilisant un distribution conditionnelle $D^{\mu}_t(X^{\mu}_t | X^{\mu}_{t-1}, I_t)$ de l'état actuel sachant l'état passé et l'image courante I_t. Notons que les observations Y^t_μ sont attachées aux régions de l'image I_t qui correspondent au membre ou aux indices extraits de ces régions (chapt. 3). Le tirage d'échantillons indépendants à partir de l'estimation de la probabilité marginale précédente $\hat{P}^{\mu}_{t-1}(X^{\mu}_{t-1})$ représentée par une somme d'échantillons pondérés est équivalent à l'étape de rééchantillonnage d'un filtre à particules (§2.3.1). La distribution de proposition générée de cette manière est simplement :

$$Q^{\mu}_t(X^{\mu}_t, X^{\mu}_{t-1}) = D^{\mu}_t(X^{\mu}_t | X^{\mu}_{t-1}, I_t) \hat{P}^{\mu}_{t-1}(X^{\mu}_{t-1})\,. \tag{2.31}$$

À partir de ces échantillons, il est possible de générer l'approximation d'une densité R, fonction de l'état du membre μ aux instants t et $t-1$:

$$R(X^{\mu}_t, X^{\mu}_{t-1}) \approx \sum_i w_i \delta(X^{\mu}_t - X^{\mu,i}_t) \delta(X^{\mu}_{t-1} - \dot{X}^{\mu,i}_{t-1})\,, \tag{2.32}$$

avec les valeurs de poids suivantes (annexe A.2, eq. A.5) :

$$w_i = \frac{1}{N_\mu} \frac{R(X^{\mu,i}_t, \dot{X}^{\mu,i}_{t-1})}{Q^{\mu}_t(X^{\mu,i}_t, \dot{X}^{\mu,i}_{t-1})}\,. \tag{2.33}$$

Pour simplifier l'algorithme, une seule distribution de proposition Q^t_μ et un seul ensemble de N_μ échantillons tirés depuis celle-ci, pour chaque membre μ à chaque instant t, seront utilisés pour le calcul des messages et des probabilités marginales. En combinant les équations 2.15 et 2.16, puis en remplaçant la probabilité marginale à $t-1$ par son approximation $\hat{P}^{\mu}_{t-1}$, on obtient :

$$s^{\mu\to\nu}_t(X^{\mu}_t) \leftarrow \int dX^{\mu}_{t-1} U(X^{\mu}_t, X^{\mu}_{t-1}) \tag{2.34}$$

$$U(X^{\mu}_t, X^{\mu}_{t-1}) = H^{\mu}(X^{\mu}_t, X^{\mu}_{t-1}) \hat{P}^{\mu}_{t-1}(X^{\mu}_{t-1}) K^{\mu}(X^{\mu}_t, Y^{\mu}_t) \prod_{\nu' \in \Gamma_{\mu,\nu}} r^{\nu'\to\mu}_t(X^{\mu}_t)\,. \tag{2.35}$$

En utilisant les équation 2.32 et 2.33 pour fournir une approximation de $U(X_t^\mu, X_{t-1}^\mu)$, on obtient l'approximation $\tilde{s}_t^{\mu\to\nu}$ de la mise à jour des messages envoyés $s_t^{\mu\to\nu}$:

$$\tilde{s}_t^{\mu\to\nu}(X_t^\mu) \leftarrow \int dX_{t-1}^\mu \sum_i u_t^{\mu\to\nu,i} \delta(X_t^\mu - X_t^{\mu,i}) \delta(X_{t-1}^\mu - \dot{X}_{t-1}^{\mu,i}) \tag{2.36}$$

$$u_t^{\mu\to\nu,i} = \frac{1}{N_\mu} \frac{U(X_t^{\mu,i}, \dot{X}_{t-1}^{\mu,i})}{Q_t^\mu(X_t^{\mu,i}, \dot{X}_{t-1}^{\mu,i})}, \tag{2.37}$$

ce qui donne :

$$\tilde{s}_t^{\mu\to\nu}(X_t^\mu) \leftarrow \sum_i s_t^{\mu\to\nu,i} \delta(X_t^\mu - X_t^{\mu,i}) \tag{2.38}$$

$$s_t^{\mu\to\nu,i} = u_t^{\mu\to\nu,i} \tag{2.39}$$

$$= \frac{1}{N_\mu} \frac{H^\mu(X_t^{\mu,i}, \dot{X}_{t-1}^{\mu,i}) \hat{P}_{t-1}^\mu(\dot{X}_{t-1}^{\mu,i}) K^\mu(X_t^{\mu,i}, Y_t^\mu)}{Q_t^\mu(X_t^{\mu,i}, \dot{X}_{t-1}^{\mu,i})} \prod_{\nu' \in \Gamma_{\mu,\nu}} r_t^{\nu'\to\mu}(X_t^{\mu,i}). \tag{2.40}$$

Les approximations par échantillonnage des messages reçus ne sont pas directement utilisées du fait que le procédé de mise à jour des messages comprend des produits et que le produit des approximations composées d'échantillons n'est pas défini (un produit de sommes de dirac). Le terme impliquant H^μ, K^μ, $\hat{P}_{t-1}^\mu$ et Q_t^μ est fixé une fois les échantillons connus. Ce terme noté $\phi_t^{\mu,i}$ vaut :

$$\phi_t^{\mu,i} = \frac{1}{N_\mu} \frac{H^\mu(X_t^{\mu,i}, \dot{X}_{t-1}^{\mu,i}) \hat{P}_{t-1}^{\mu,i}(\dot{X}_{t-1}^{\mu,i}) K^\mu(X_t^{\mu,i}, Y_t^\mu)}{Q_t^\mu(X_t^{\mu,i}, \dot{X}_{t-1}^{\mu,i})} \tag{2.41}$$

$$= \frac{1}{N_\mu} \frac{H^\mu(X_t^{\mu,i}, \dot{X}_{t-1}^{\mu,i}) K^\mu(X_t^{\mu,i}, Y_t^\mu)}{D_t^\mu(X_t^{\mu,i} | \dot{X}_{t-1}^{\mu,i}, I_t)}. \tag{2.42}$$

Ce terme est simplement le poids de l'échantillon i composant l'approximation du produit du message provenant du passé (eq. 2.15), qui peut être interpréter comme une prédiction basée sur le facteur de cohérence temporel, avec le facteur de compatibilité à l'image. L'équation 2.38 devient :

$$s_t^{\mu\to\nu,i} \leftarrow \phi_t^{\mu,i} \prod_{\nu' \in \Gamma_{\mu,\nu}} r_t^{\nu'\to\mu}(X_t^{\mu,i}). \tag{2.43}$$

En combinant cette approximation dans l'équation 2.17, l'approximation $\tilde{r}_t^{\nu\to\mu}$ du message reçu $r_t^{\nu\to\mu}$ est mis à jour par :

$$\tilde{r}_t^{\nu\to\mu}(X_t^\mu) \leftarrow \sum_i L^{\mu,\nu}(X_t^\mu, X_t^{\nu,i}) s_t^{\nu\to\mu,i}. \tag{2.44}$$

En notant $r_t^{\nu\to\mu,i}$ la valeur de cette approximation pour l'échantillon i de X_t^μ, nous obtenons :

$$r_t^{\nu\to\mu,i} \leftarrow \sum_j L^{\mu,\nu}(X_t^{\mu,i}, X_t^{\nu,j}) s_t^{\nu\to\mu,j}. \tag{2.45}$$

Comme les échantillons sont fixés durant tout le processus de propagation des croyances, le facteur de lien $L^{\mu,\nu}(X_t^{\mu,i}, X_t^{\nu,j})$ peut être calculé préalablement pour tous les liens $(\mu,\nu) \in \Gamma$ et pour toutes les paires d'échantillons i, j de X_t^μ et X_t^ν :

$$L_t^{\mu,\nu,i,j} = L^{\mu,\nu}(X_t^{\mu,i}, X_t^{\nu,j}). \tag{2.46}$$

L'équation de mise à jour des messages reçus est finalement :

$$r_t^{\nu\to\mu,i} \leftarrow \sum_j L_t^{\mu,\nu,i,j} s_t^{\nu\to\mu,j}, \tag{2.47}$$

et l'équation 2.43 de mise à jour devient :

$$s_t^{\mu\to\nu,i} \leftarrow \phi_t^{\mu,i} \prod_{\nu'\in\Gamma_{\mu,\nu}} r_t^{\nu'\to\mu,i}. \tag{2.48}$$

Ces équations sont simplement les équations de mise à jour pour l'algorithme de propagation des croyances dans les espaces discrets, l'espace d'état de chaque membre étant limité à ses propres échantillons qui sont également ceux utilisés lors du processus de filtrage particulaire. C'est l'idée clé pour parvenir à une méthode d'estimation rapide. La probabilité marginale de chaque membre est alors représenté par une somme d'échantillons pondérés et l'utilisation d'une approximation d'après un échantillonnage d'importance utilisant les mêmes échantillons aboutit de la même façon à l'expression suivante de la probabilité marginale :

$$\hat{P}_t^{\mu}(X_t^{\mu}) \propto \sum_i \hat{w}_t^{\mu,i} \delta(X_t^{\mu} - X_t^{\mu,i}) \tag{2.49}$$

$$\hat{w}_t^{\mu,i} = \phi_t^{\mu,i} \prod_{\nu\in\Gamma_{\mu}} r_t^{\nu\to\mu,i} \tag{2.50}$$

De cette manière, une estimation pleinement récursive est obtenue. Comme pour le filtre à particule, le choix de la distribution de prédiction D_t^{μ} est crucial. L'utilisation classique du facteur de cohérence temporel H^{μ} pour D_t^{μ} (Condensation [BI96]) simplifie l'équation 2.42 mais peut entraîner des effets pervers lorsque les échantillons prédits d'après la cohérence temporelle ne sont pas dans les zones où la vraisemblance est élevée (le"street light effect" [DTS+05]). Comme l'espace d'état est réduit du fait de sa décomposition en sous-espaces pour chaque membre, le choix simple d'un facteur de cohérence temporelle est généralement suffisant. Cependant, des problèmes peuvent apparaître dans le cas de mouvements rapides et des distributions plus complexes doivent alors êtres choisie.

Chapitre 3

Indices extraits de l'image et modèle du corps

Sommaire

3.1 Introduction

Les indices extraits de l'image correspondent à ce que l'on choisi de donner à voir à la machine. Ces données doivent être suffisamment discriminantes pour parvenir à saisir les subtilités qui distinguent les différentes poses entre-elles dans une image monoculaire. Dans un contexte de déficit en informations comparé aux procédés stéréo ou multicaméra, il est nécessaire d'établir un ensemble d'indices complémentaires et robustes face aux variations de l'environnement capable de mener à bien le suivi. Le calcul de ces indices ne doit pas empêcher le fonctionnement en temps réel de l'algorithme de suivi et le choix d'utiliser plusieurs indices pousse donc à adopter des indices et des algorithmes rapides.

Conformément aux choix entrepris à l'issu du travail bibliographique (§1.5.5), les indices retenus consistent en une détection de gradient, des mouvements et de la teinte chair (§3.2) que vient compléter une soustraction de fond (§3.3). Le modèle géométrique du corps utilisé pour exploiter ces indices est décrit au paragraphe (§3.4).

L'approche proposée ici se démarque des précédentes par l'utilisation conjuguée d'un ensemble d'indices complémentaires augmentant la robustesse du suivi même dans des cas d'occultations. En plus des contraintes articulaires pour forcer les liaisons entre les membres adjacents, des limites articulaires sont directement implémentées au cœur du procédé de calcul des messages lors de la propagation des croyances (2.2.3). Ces deux nouveautés font l'objet de l'étude menée tout au long de ce chapitre.

3.2 indices extraits de l'image

3.2.1 Détection de gradient

La détection des gradients de luminosité consiste généralement à convoluer un masque qui va filtrer le bruit dans l'image avant de dériver le signal horizontalement et verticalement afin d'obtenir I_x et I_y, les magnitudes des gradients dans ces deux directions. Le module du gradient exprimé selon la norme L_N est alors :

$$\| \overrightarrow{grad(x,y)} \| = (I_x(x,y)^N + I_y(x,y)^N)^{1/N} . \tag{3.1}$$

Les contours peuvent être extraits à partir des gradients par la recherche des maxima de 3.1 après seuillage (fig. 3.1). Leur orientation est perpendiculaire à celle du gradient qui est donnée par :

$$\Phi(x,y) = arctan\left(\frac{I_y(x,y)}{I_x(x,y)}\right) \tag{3.2}$$

Les masques courament utilisés sont Gaussiens pour le détecteur de Canny [Can86] ou de type ISEF "*Infinite Symetric Exponential Filter*" pour l'algorithme de Shen-Castan [SC92]. Le choix d'un détecteur doit être guidé par sa précision et sa rapidité d'où l'intérêt pour ces deux algorithmes et leur versions récursives. Sous cette implémentation, le Shen-Castan est plus simple. C'est pourquoi il est choisi dans le cadre de ce travail. Une image des contours ainsi que les images des intensités des gradients horizontaux et verticaux obtenus après filtrage sont données à la figure 3.1.

3.2.2 Détection des mouvements

Il existe principalement deux façons de détecter le mouvement dans une image : la détection des zones contenant du mouvement peut se faire simplement par soustraction d'images adjacentes ou par flot optique. Cette dernière méthode à la capacité de fournir la direction et l'amplitude du mouvement. Ces informations peuvent être utiles pour prédire la position d'objets dans les images. Cependant, le flot optique est lourd à calculer et la soustraction d'image adjacente est préférée pour respecter un traitement en temps réel.

Une image binaire des mouvements $I_m^t(r)$ est constituée pour chaque pixel $I^t(x,y)$ de l'image I^t à l'instant t à partir de la soustraction des deux images adjacentes en niveaux de gris I^t et

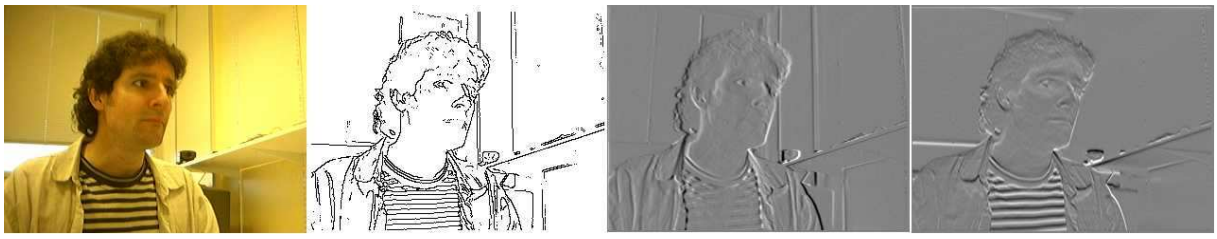


FIG. 3.1: *Détection des contours par l'algorithme de Shen-Castan [SC92]. De gauche à droite : image originale, contours par seuillage et détection des maxima du module du gradient, gradients horizontaux et verticaux.*

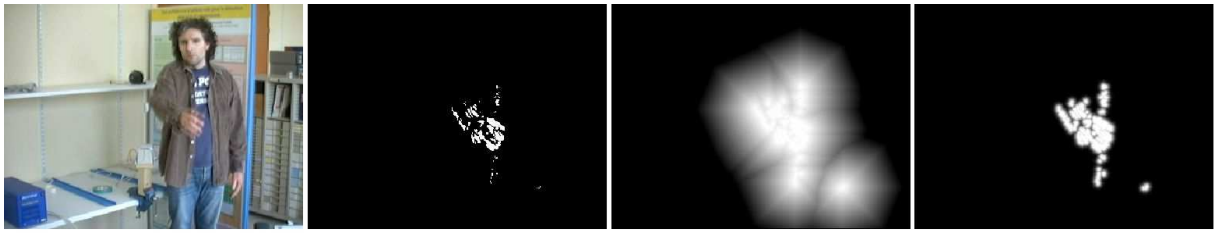


FIG. 3.2: *Carte de probabilité de l'énergie de mouvement. (a) Image originale, (b) seuillage, (c) distance de chanfrein, (d) probabilités de mouvement.*

I^{t-1}, puis d'un seuillage :

$$\begin{aligned} &\forall\,(x,y)\in I^t\,,\\ &\left(I^{t-1}(x,y)-I^t(x,y)\right)^2 > s \Rightarrow I^t_m(x,y)=1\,,\\ &\left(I^{t-1}(x,y)-I^t(x,y)\right)^2 \leqslant s \Rightarrow I^t_m(x,y)=0\,. \end{aligned} \tag{3.3}$$

FIG. 3.3: *Carte des probabilités de la teinte du visage. (a) Image originale, (b) carte des probabilités.*

À partir de l'image du mouvement I^t_m, on cherche à construire une carte des probabilités du mouvement I^t_p qui exprime la probabilité qu'un pixel soit en mouvement. Cette probabilité est calculée pour chaque pixel dans l'image et le modèle choisi suppose qu'elle est fonction de la distance qui sépare un pixel de son plus proche voisin marqué en mouvement dans I^t_m (c'est à dire un pixel (x,y) tel que $I^t_m(x,y)=1$). Le calcul intermédiaire de la carte des distances aux mouvement I^t_d fait appel à la distance de chamfrein [RP66] car elle peut être efficacement calculée par un algorithme récursif rapide qui procède en deux balayages d'image. Pour simuler un processus stochastique, la carte des probabilités I^t_p (fig. 3.2) fait intervenir la Gaussienne G_m des distances calculés sur I^t_d pour avoir :

$$P\left(I^t_m(x,y)=1\right) = G_m\left(I^t_d(x,y)\right)\,. \tag{3.4}$$

Les paramètres utilisés sont les suivants : le seuil de mouvement est réglé à $s=10\%$ de la valeur maximum du carré de la différence entre les pixels (eq. 3.3). Ce réglage permet de filtrer les petits mouvements des vêtements ou du corps pour ne garder que les plus amples souvent associés aux membres supérieurs (bras, avan-bras et mains). La variance du noyau G_m est fixé à 4 pixels. Des valeurs plus élevées aboutiraient à un positionnement trop lâche des membres.

3.2.3 Détection de la teinte chair

Durant la phase d'initialisation, la position du visage est donnée grâce à un détecteur robuste [FBVC01]. Un modèle de couleurs du visage est crée sur la zone détectée en calculant un histogramme normalisé des couleurs UV dans l'espace YUV. L'histogramme est quantifié selon 32 cases par canaux, soit un total de 32×32 cases.

Chaque pixel de l'image est comparé à ce modèle afin de déterminer la probabilité pour ce pixel d'être de la même teinte que le visage. La probabilité est simplement la valeur de la case d'histogramme dans lequel ce pixel tombe. Ce procédé permet aussi de détecter les mains qui possèdent une teinte identique à celle du visage (fig. 3.3).

3.3 Une soustraction de fond robuste

Le principe le plus basique d'une soustraction de fond consiste à soustraire l'image actuelle à une image de référence censée représenter le fond. Cette différence tend vers 0 pour les pixels identiques sur les deux images et elle est non nulle pour ceux qui vont générer une différence. Un seuillage peut être utilisé pour segmenter le fond du premier plan mais le résultat est souvent bruité.

Dans le cadre du suivi du corps, l'objectif de la soustraction de fond proposée est de segmenter un masque qui va contenir la silhouette du personnage. La détection des membres va se focaliser dans cette zone afin de neutraliser l'influence des distracteurs contenus dans le fond. Il n'est pas nécessaire que le masque épouse précisément la silhouette. Au contraire, celui-ci doit déborder afin d'inclure plus sûrement les contours extérieurs à la silhouette qui serviront à la détection des membres.

Conformément aux objectifs fixés, la soustraction de fond doit être robuste aux variation de lumière dans la scène et pouvoir fournir des résultats fiables même dans les cas d'utilisation de fonds complexes. Cette objectif est atteint en mettant en œuvre un nouvel algorithme précis, polyvalent et bien sûr rapide pour la contrainte temps réel. Cet algorithme utilise le nouveau concept d'histogramme local des orientations à noyaux Gaussiens décrit ci-après.

3.3.1 Travaux connexes et justification des choix

Dans le but de lever l'hypothèse peu réaliste d'un fond statique lors d'une soustraction de fond, plusieurs approches tentent d'inclure un module de remise à jour du fond. Wallflower [TKBM99] exploite trois niveaux d'analyse spatiales. À l'échelle du pixel, la valeur d'un pixel appartenant au fond est prédite grâce à un filtre de Wiener. Le niveau intermédiaire a pour but de détecter les zones du premier plan en mouvement pour corriger d'éventuelles erreurs. En cas de changement soudain d'illumination, le dernier niveau qui opère à l'échelle de l'image est chargé de trouver, parmi des fonds précédemment appris, un modèle qui correspond aux nouvelles conditions. Cette approche, restreinte par la quantité limitée des fonds appris, est peu généralisable. Elgammal et al. [EHD00] utilisent un modèle non paramétrique incluant un modèle de bruit Gaussien pour la caméra, où chaque pixel du fond est comparé à un échantillon comportant les dernières valeurs d'intensité acquises pour ce pixel. Les résultats d'un modèle à court et à long terme sont fusionnés et les fausses détections sont traités lors d'une seconde étape qui consiste à tester si un pixel du premier plan possède le même modèle de fond que ses voisins. Les fausses détections engendrées par le bruit de la caméra ne sont pas détectés par la précédente approche, c'est pourquoi Sminchisescu et Triggs [ST02] ont recours aux surfaces de niveau pour squelettiser leur silhouette avant de la faire croître à nouveau pour obtenir un résultat exempt de fausses détections. Cependant, le calcul des surfaces de niveau rend difficile le temps réel. Une autre solution précise mais non temps réel [SS05] consiste à modéliser le premier plan en plus du fond. Cette méthode stochastique utilise un réseau de Markov pour prendre en compte les dépendances spatiales entre les pixels voisins. La technique des graph-cut est mise en œuvre pour segmenter le fond.

Différencier le premier plan du fond dans un environnement réel où les conditions d'illumination peuvent changer instantanément, lorsque des lampes s'allument ou s'éteignent, n'est pas trivial. L'utilisation d'un invariant couleur est une solution alternative aux modèles de remise à jour du fond. Les couleurs dans une image numérique sont exprimées dans un espace donné (RGB, YUV, HSV...) [TT03]. Une méthode simple en coordonnées RGB consiste à diviser chaque canal de couleur avec la somme des autres. Un modèle plus sophistiqué [GS96] consiste calculer les composantes robustes à partir d'une normalisation de la soustraction des canaux de couleur deux à deux. D'autres paramètres robustes peuvent êtres extraits après avoir calculé les covariances des canaux de couleur normalisés et centrés en leurs moyennes. L'inverse cosinus des covariances sont des valeurs peu influencées par les changements d'illumination [FCF96]. Le principal inconvénient de ces méthodes provient du fait que les nouvelles coordonnées sont moins informatives sur le contenu de l'image.

Une alternative aux méthodes précédentes consiste à utiliser d'autres indices seuls ou en plus des couleurs. Du fait que l'orientation du gradient en un point de l'image est moins influencée par les changements lumineux, les histogrammes d'orientation des gradients (HOG) peuvent constituer une solution. Cette technique est utilisée pour détecter des piétons dans des scènes d'extérieur où les variations lumineuses sont nombreuses [DT05]. Une approche Bayesienne [LHGT04] exploite des indices spectraux et spatiaux temporels pour modéliser le fond. La fusion des indices améliore la robustesse face aux variations de lumière et aux mouvements réguliers dans le fond. Si la méthode statistique de modélisation du fond est plus fine que les méthodes à base de Gaussiennes, elle ne permet pas de s'affranchir de post traitements qui utilisent la morphologie mathématique avec une ouverture-fermeture et un remplissage des trous afin d'améliorer les résultats. L'algorithme n'est donc pas temps réel (trois images 320×240 par seconde avec un PC équipé d'un Pentium 4 1,7 GHz).

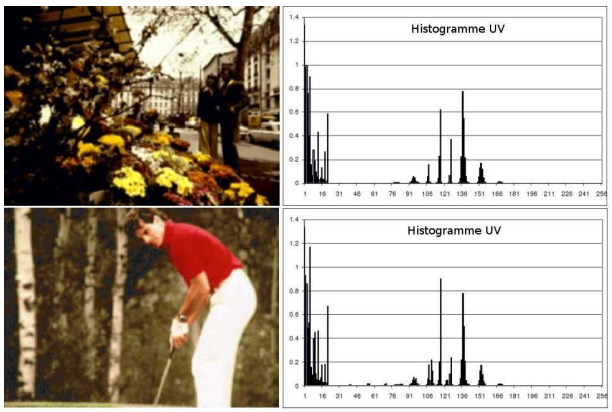


FIG. 3.4: *Deux images différentes donnent des histogrammes identiques.*

Un inconvénient majeur des histogrammes provient du fait qu'ils n'aient pas la capacité de conserver l'information spatiale dans l'image. En effet, deux images peuvent être différentes et posséder le même histogramme (fig. 3.4). En revanche, les histogrammes locaux à noyaux [NBB06] permettent de fournir une information spatiale en gardant une taille restreinte pour des calculs en temps réel. Ils ont la capacité de lisser le bruit généré par la caméra ou les petits mouvements dans le fond de l'image. Cependant l'association de cette méthode avec les caractéristiques de couleurs ne permet pas de traiter le problème des variations lumineuses. L'idée développée dans cette thèse consiste à adapter cette technique avec des indices basés sur les gradients de contours dans le but d'obtenir une robustesse satisfaisante face aux variations d'illumination tout en restant temps réel. L'algorithme développé ici se base sur les HOG [DT05] et modélise en plus l'influence du bruit émis par la caméra sur le module et l'orientation des gradients.

3.3.2 Histogramme local des orientations à noyaux Gaussiens

Pondération spatiale

La solution adoptée pour retrouver une information spatiale consiste à découper l'image en zones locales de $n \times n$ pixels se recouvrant entre elles à la manières des tuiles d'ardoise pour une plus grande précision spatiale. Ceci permet de caractériser spatialement l'image par un ensemble d'histogrammes [NBB06].

Moins robustes vis-à-vis des variations lumineuses, les couleurs sont laissées de côté pour utiliser les contours. Il s'agit de calculer, pour chaque zone locale dans l'image, un histogramme d'orientations des contours pondéré spatialement par un noyau Gaussien. Le poids spatial d'un pixel $r_k(x_k, y_k)$ appartenant à la zone locale Z_l est fonction de la distance Euclidienne qui le sépare du centre de cette zone. En considérant le noyau Gaussien spatial $G_s^l(r_k;\, \mu_s^l, \sigma_s)$ dont la moyenne $\mu_s^l = (x_l, y_l)$ se trouve centré sur Z_l (fig. 3.5), K_s étant un coefficient de normalisation, on a :

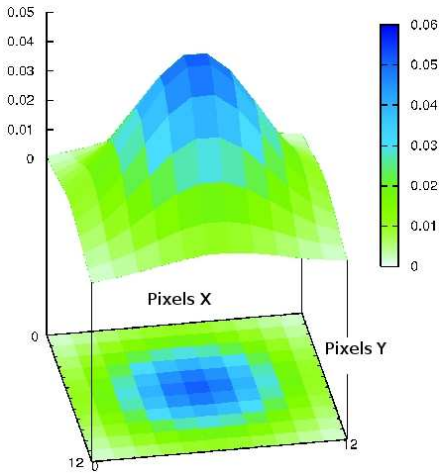


FIG. 3.5: *Noyau Gaussien spatial sur une zone locale.*

$$
\begin{aligned}
d_x &= x_k - x_l\,, \\
d_y &= y_k - y_l\,, \\
K_s &= \frac{1}{\sqrt{2\pi}\sigma_s} \\
G_s^l(r_k;\, \mu_s^l, \sigma_s) &= K_s.exp\left(-\frac{d_x^2 + d_y^2}{2\sigma_s^2}\right)\,.
\end{aligned}
\tag{3.5}
$$

Calcul des histogrammes

Un histogramme polaire H_l discrétisé suivant N orientations est calculé pour chaque zone locale Z_l. Les cases de H_l sont notées h_l^n, avec $n \in \{1, \ldots, N\}$. Elles représentent chacune une orientation o qui vaut :

$$o = 2\pi\frac{n}{N}\,, \tag{3.6}$$

Pour chacun des pixels r_k appartenant à la zone Z_l, l'algorithme de Shen-Castan [SC92] fournit une norme $\|\overrightarrow{E_k}\|$ et une orientation $dir(\overrightarrow{E_k})$ du vecteur gradient $\overrightarrow{E_k}$. La norme correspond au contraste et la direction est orthogonale à l'orientation du contour. La variance du bruit sur l'orientation du gradient est mesurée durant une courte séquence en utilisant une mire qui contient des lignes de contraste différent. Le résultat de ce test est représenté à la figure 3.6. Il montre que le bruit sur l'orientation du gradient peut être modélisé par un noyaux Gaussien G_o dont la variance est linéairement décroissante en fonction de la norme du contour :

$$\sigma_o(\|\overrightarrow{E_k}\|) = -a_o\|\overrightarrow{E_k}\| + b_o\,. \tag{3.7}$$

Pour modéliser les erreurs de quantification dûes au bruit sur l'orientation des contours, un pixel r_k va contribuer à toutes les cases de l'histogramme H_l. Une case h_l^n sera remplie selon le noyau

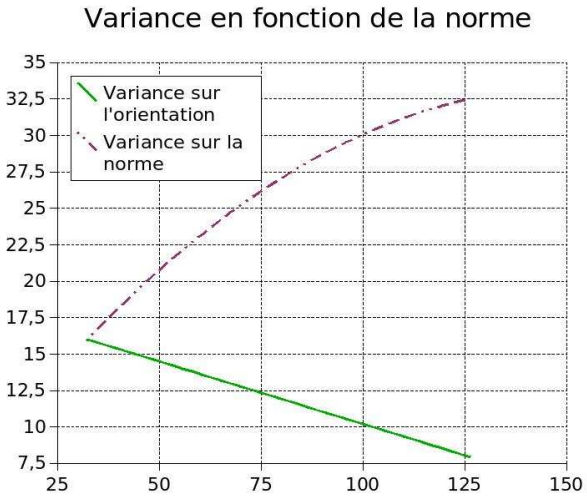


FIG. 3.6: *Variance du bruit sur la norme et l'orientation du vecteur gradient en fonction de sa norme.*

Go centré sur l'orientation o (éq. 3.6) qui représente cette case :

$$G_o\left(r_k;\, \mu_o = o, \sigma_o(\|\overrightarrow{E_k}\|)\right) = \frac{1}{\sqrt{2\pi}\sigma_o} exp\left(-\frac{\left(dir(\overrightarrow{E_k}) - o\right)^2}{2\sigma_o^2}\right). \tag{3.8}$$

En tenant compte de tous les pixels appartenant à la zone locale Z_l et du noyau spatial G_s^l (eq. 3.5), la case d'histogramme h_l^n vaut :

$$h_l^n = \sum_{k \in l} G_o\big(r_k;\, \mu_o, \sigma_o(\|\overrightarrow{E_k}\|)\big) G_s^l(r_k;\, \mu_s^l, \sigma_s) f(\|\overrightarrow{E_k}\|)\,, \tag{3.9}$$

avec :

$$f(\|\overrightarrow{E}\|) = \frac{1}{\lambda}\|\overrightarrow{E}\| tanh(\frac{\lambda}{\|\overrightarrow{E}\|}), \tag{3.10}$$

$f(\|\overrightarrow{E}\|)$ est une fonction qui pénalise les modules élevés de $\overrightarrow{E}$ et λ est un coefficient qui permet de moduler cette pénalisation. Dans la pratique on choisira λ approximativement égal à la moyenne des modules des gradients dans l'image. Ce réglage permet de saturer les fortes valeurs de gradients tout en offrant une bonne discrimination pour les valeurs situées autour de la moyenne. Un histogramme polaire H_l est calculé pour la zone locale Z_l d'après (3.9) :

$$H_l = \{h_l^n\}^{n \in N}. \tag{3.11}$$

Comparaison des histogrammes

Les histogrammes du fond sont calculés sur une première image d'initialisation ne montrant que le fond. Pour les autres images, les histogrammes nouvellement calculés sont comparés avec ceux du fond pour une même zone locale. Les histogrammes n'étant pas normalisés, la distance de Bhattacharyya utilisée dans [NBB06] ne peut fournir une mesure de similitude entre eux. D'autres mesures telles que l'intersection d'histogrammes ont étés abandonnées au profit d'une

méthode prenant en compte le bruit sur la norme du gradient. Comme précédemment, ce bruit à été mesuré grâce à une mire comprenant différents niveaux de contraste. L'expérience montre (fig. 3.6) que le bruit est modélisable par une loi normale G_m présentant une variance affine, croissante en fonction du module du gradient.

D'après ce modèle, le gradient mesuré sur l'image $\|\widetilde{\overrightarrow{E_k}}\|$ correspond à une valeur aléatoire tirée suivant la Gaussienne G_m. Or, l'estimation de la moyenne et de la variance de G_m doit faire appel à la vraie valeur du gradient $\|\overrightarrow{E_k}\|$ qui est inconnue. Pour simplifier, on supposera ces deux valeurs proches l'une de l'autre :

$$\begin{aligned} &\|\overrightarrow{E_k}\| \backsimeq \|\widetilde{\overrightarrow{E_k}}\|\,, \\ &G_m\left(\|\widetilde{\overrightarrow{E_k}}\|;\, \mu_m(\|\overrightarrow{E_k}\|), \sigma_m(\|\overrightarrow{E_k}\|)\right) \begin{cases} \mu_m(\|\overrightarrow{E_k}\|) \backsimeq \|\widetilde{\overrightarrow{E_k}}\| \\ \sigma_m(\|\overrightarrow{E_k}\|) \backsimeq a_m\|\widetilde{\overrightarrow{E_k}}\| + b_m \end{cases} \end{aligned} \tag{3.12}$$

Pour une même zone locale Z_l, la similitude de l'histogramme de référence du fond H_l^{ref} et de l'histogramme actuel H_l^{act} est évaluée en comparant leurs cases deux à deux. Sachant que la valeur d'une case est homogène à une norme de gradient, l'idée est de mesurer la similitude des deux Gaussiennes qui modélisent le bruit sur les amplitudes $h_n^{l_{ref}}$ et $h_n^{l_{act}}$ résultantes de ces cases pour les comparer. Cette mesure qui s'appuie sur la surface des zones de recouvrement des deux Gaussiennes fait appel à la fonction de densité cumulée $\aleph_{cdf}\left(G_m(\mu_m, \sigma_m)\right)$. De manière heuristique, le point de calcul de $\aleph_{cdf}$ correspond à la moyenne des deux amplitudes :

$$\mu_n = \frac{h_n^{l_{ref}} + h_n^{l_{act}}}{2}\,. \tag{3.13}$$

Les probabilités que l'amplitude des cases d'histogrammes soit inférieure à μ_n sont calculées :

$$\begin{aligned} P_n^{l_{ref}} &= P(h_n^{l_{ref}} < \mu_n) = \aleph_{cdf}\left(\mu_n;\, G_m\left(h_n^{l_{ref}}, \sigma_m(h_n^{l_{ref}})\right)\right)\,, \\ P_n^{l_{act}} &= P(h_n^{l_{act}} < \mu_n) = \aleph_{cdf}\left(\mu_n;\, G_m\left(h_n^{l_{act}}, \sigma_m(h_n^{l_{act}})\right)\right)\,, \end{aligned} \tag{3.14}$$

ainsi que leurs compléments :

$$\begin{aligned} \overline{P}_n^{l_{ref}} &= 1 - P_n^{l_{ref}}\,, \\ \overline{P}_n^{l_{act}} &= 1 - P_n^{l_{act}}\,. \end{aligned} \tag{3.15}$$

La distance de Bhattacharyya est utilisée pour donner un score de similitude compris entre 0 et 1 :

$$P_n^l = \sqrt{P_n^{l_{ref}} P_n^{l_{act}}} + \sqrt{\overline{P}_n^{l_{ref}} \overline{P}_n^{l_{act}}}\,. \tag{3.16}$$

Si cette mesure est interprêtée comme une probabilité, la probabilité $P(Z_l \in fond)$ que Z_l appartienne au fond correspond alors au produit des P_n^l sur les N cases de l'histogramme :

$$P(Z_l \in fond) = \prod_{n \in N} P_n^l\,. \tag{3.17}$$

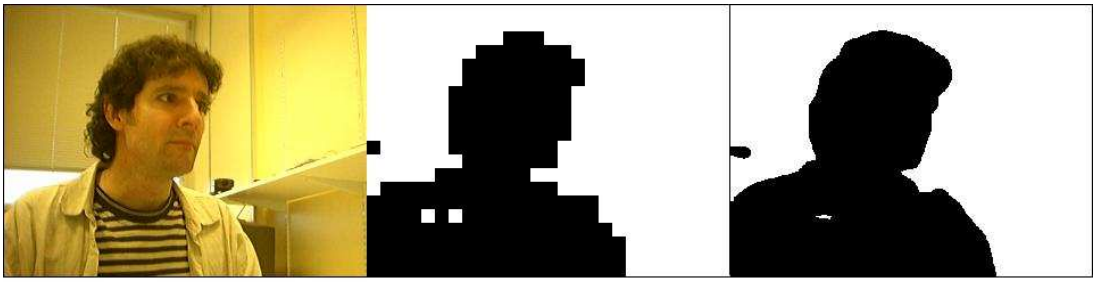


FIG. 3.7: *Lissage des probabilités au niveau des pixels. (a) : image originale, (b) : probabilités au niveau des zones locales, (c) : probabilités au niveau des pixels.*

Probabilité à l'échelle d'un pixel

La comparaison des histogrammes attachés à une zone locale donne une probabilité identique pour tous les pixels de cette zone. Ce procédé aboutit à une précision spatiale grossière de la carte des probabilités du fond (fig. 3.7b). La précision peut être améliorée en augmentant le taux de recouvrement des zones entre-elles mais les temps de traitement s'alongent puisque le nombre de zones requis pour couvrir l'image augmente avec le taux de recouvrement. Alternativement, une probabilité $P(r \in fond)$ peut être affectée individuellement à chaque pixel r en calculant une moyenne sur les N_l zones locales auxquelle appartient ce pixel (fig. 3.7c). Cette moyenne est pondérée par la distance au centre du pixel par rapport aux zones considérées :

$$P(r \in fond) = \frac{1}{\sum_{l=1}^{N_l} G_s^l(r)} \sum_{l=1}^{N_l} G_s^l(r) P(Z_l \in fond) \; . \tag{3.18}$$

3.3.3 Résultats expérimentaux

Les histogrammes d'orientations locaux à noyaux Gaussiens sont comparés à sept autres algorithmes représentatifs de l'état de l'art. Ceux basés sur les couleurs utilisent la chrominance UV extraite d'une image YUV. Pour les autres, les gradients sont calculés sur le canal Y. Les huit algorithmes testés sont :

Moyenne & Seuillage : les pixels du fond sont modélisés par leur moyenne calculée durant une phase d'apprentissage. Durant les tests, les pixels dont la valeur dépasse un seuil autour de la moyenne apprise sont considérés comme appartenant au premier plan.

Moyenne & Variance : les pixels du fond sont modélisés par leur moyenne et leur variance apprises lors d'une phase d'initialisation. Un seuillage en fonction de la variance détermine les pixels appartenant au premier plan.

Histogramme Local Classique des Couleurs [MD01] : les images sont segmentées en deux grilles de 20×20 pixels décalées 10 pixels en hauteur et en largeur de manière à se recouvrir partiellement. Un histogramme des couleurs est calculé pour chaque zone dans l'image de référence et l'image actuelle. La mesure de similitude est mesurée par intersection d'histogrammes et un seuil classe les zones du fond ou du premier plan.

Histogramme Local des Couleurs à Noyaux Gaussiens [NBB06] : l'image est partitionnée en zones locales qui se recouvrent. Les histogrammes sont calculés en appliquant un noyau Gaussien dans l'espace des images et dans l'espace des couleurs.

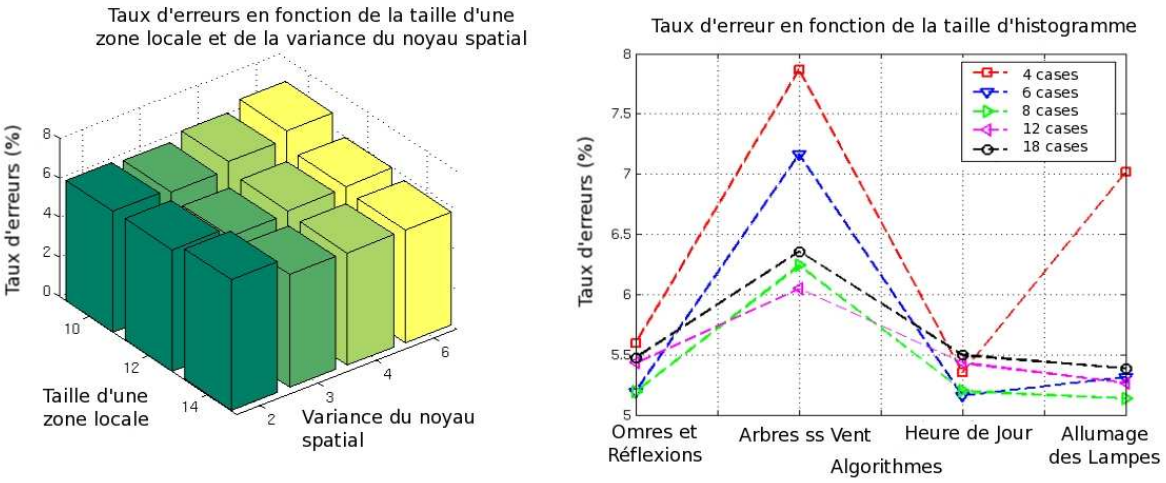


FIG. 3.8: *Influence de la taille des zones locales et de la variance du noyau Gaussien spatial dans le pourcentage de pixels erronés. A droite, le pourcentage d'erreurs commises pour différentes scènes avec différentes taille de noyaux Gaussiens spatiaux. Les meilleures performances sont atteintes pour* 8 *et* 12 *cases. Le réglage retenu sera* 8 *puisque les petites tailles d'histogrammes favorisent la vitesse de calcul.*

Conservation du Signe du Gradient [XRB04] : les séquences spatiales dans les quatre directions cardinales du signe du gradient autour du voisinage de chaque pixel sont apprises sur l'image de référence. Elles sont comparées à l'image courante pour détecter les pixels appartenant au premier plan.
Histogramme Local Classique des Orientations des Gradients : si le gradient d'un pixel dépasse un seuil, il est classé dans un histogramme polaire des orientations. Les autres pixels ne sont pas pris en compte pour diminuer l'influence du bruit généré par la caméra. Les histogrammes sont calculés pour des zones locales se recouvrant sur toute l'image. Ils sont comparés avec la distance de Bhattacharyya après normalisation. Un seuil détermine la nature de la zone (fond ou premier plan).
Histogramme d'Orientation des Gradients (HOG) [DT05] : les histogrammes d'orientation des gradients normalisés selon la norme L2 sont calculés à partir d'une grille dense pondérée spatialement par des noyaux Gaussiens qui partitionnent l'image. L'image des probabilités fournie par la distance de Bhattacharyya est seuillée pour obtenir un résultat binaire.
Histogramme Local des Orientations à Noyaux Gaussiens [NB06] : C'est la méthode utilisée ici. Des histogrammes d'orientation des gradients sont calculés d'après des zones locales pondérées par un noyau Gaussien spatial. Le bruit dans l'image affectant la norme et l'orientation des gradients est modélisé par deux autres noyaux Gaussiens.

Les séquences de test incluent des scènes d'intérieur et d'extérieur (voir figure 3.9). Certaines d'entre elles : "Occultation d'un Moniteur", "Arbre sous le Vent" et "Heure du Jour" sont utilisées dans Wallflower [TKBM99]. Elles sont disponibles sur internet[1]. Les cinq séquences incluent les problèmes caractéristiques rencontrés en détourage de silhouettes :
Ombres et Réflexions : une personne se tient debout entre une fenêtre et la porte. L'ombre et les réflexions modifient légèrement la partie gauche de l'image.
Occultation d'un Moniteur : un écran cathodique d'ordinateur se trouve sur un bureau. Une

[1] http ://research.microsoft.com/users/jckrumm/WallFlower/TestImages.htm

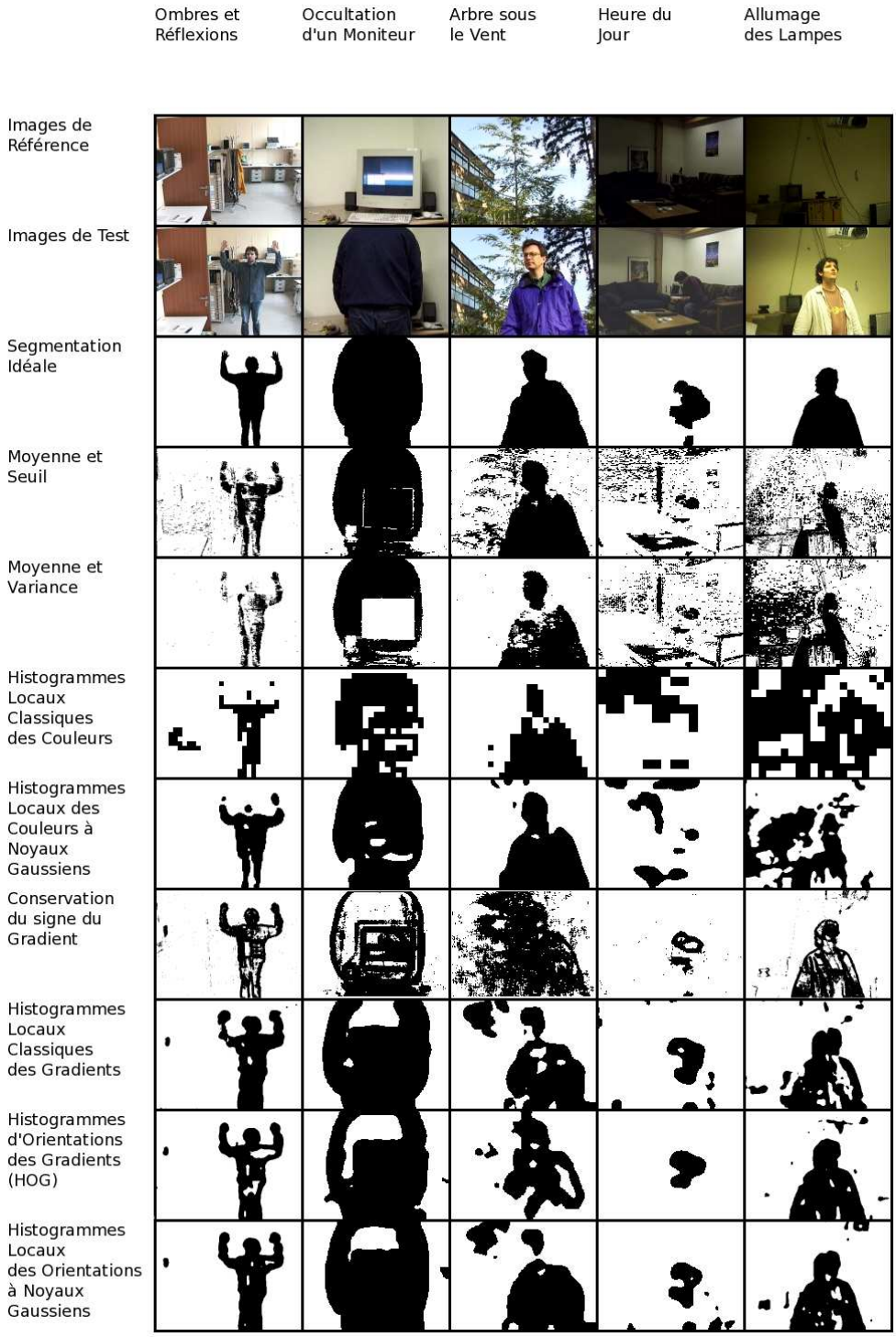


FIG. 3.9: *Comparaison des algorithmes de détourage pour différentes scènes. La première ligne montre les images utilisées lors de d'initialisation, la seconde ligne montre les images de test, la troisième représente la vérité de terrain segmentée à la main. Les autres lignes montrent les résultats pour chaque algorithme.*

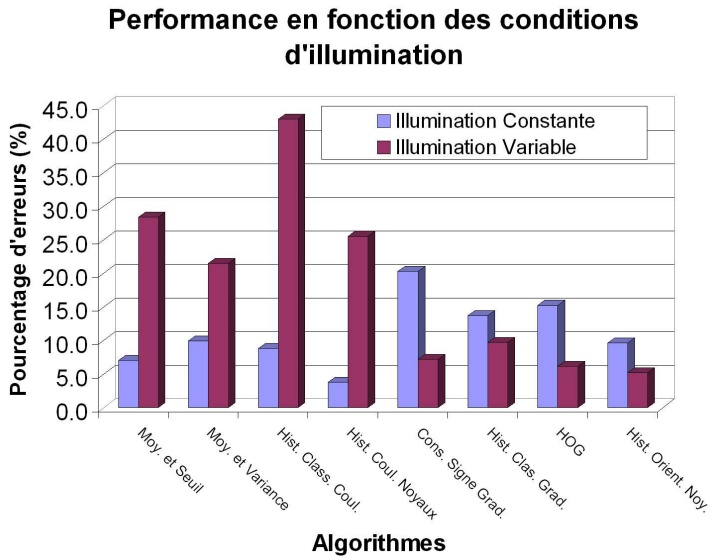


FIG. 3.10: *Performances dans le cas d'illuminations constantes ou variables.*

personne entre et se place devant le moniteur.
Arbre Sous le Vent : une personne marche près d'un arbre qui de balance sous le vent.
Heure du Jour : Cette séquence montre une pièce sombre qui s'éclaire graduellement. Une personne entre et s'installe sur le canapé.
Allumage des Lampes : au début de cette séquence de test, la pièce est baignée d'une lumière faible. Après quelques minutes, un personnage entre et commande l'éclairage principal.

L'apprentissage de l'algorithme "Moyenne et Seuil" ainsi que "Moyenne et Variance" se fait

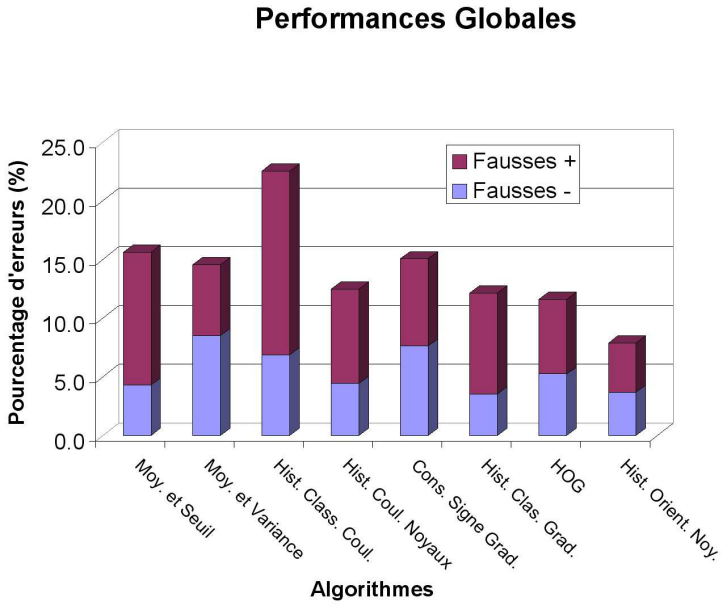


FIG. 3.11: *Performances globales des algorithmes.*

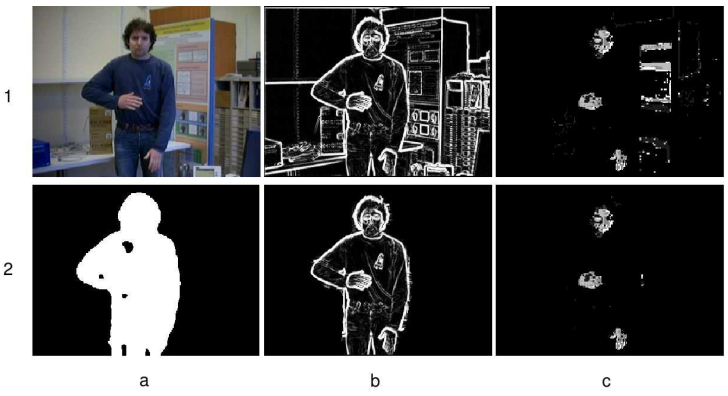


FIG. 3.12: *Fusion de la soustraction de fond (2a) et des indices de contours (2b) et de teinte visage (2c). Image originale (1a), contours (1b) et teinte visage (1c) sur toute l'image.*

sur les vingt premières images. Pour les autres, le fond est appris uniquement sur la première image. Pour la scène "Heure du Jour", la phase d'apprentissage débute à la $800^{\text{ème}}$ image, lorsque la scène n'est plus plongée dans une pénombre totale.

La taille d'une zone locale $n \times n$, celle des histogrammes d'orientation et la variance σ_S du noyau spatial ont étés réglés selon plusieurs valeurs lors de tests sur les scènes "Ombres et Réflexions" et "Arbre Sous le Vent". Le pourcentage d'erreurs a été évalué à chaque fois pour déterminer le meilleur réglage (voir figure 3.8) : $n = 12$ pixels avec une taille d'histogramme de 8 cases et $\sigma_s = 3$.

Les résultats qualitatifs sont donnés à la figure 3.9. Les résultats quantitatifs des tests se trouvent à la figure 3.10 et 3.11. Ils fournissent le taux d'erreurs en pourcentage par rapport à une segmentation manuelle qui donne la vérité de terrain. Ils montrent que si les histogrammes d'orientation locaux à noyaux Gaussiens donnent les meilleures performances pour la globalité des tests, les histogrammes locaux des couleurs à noyaux Gaussiens [NBB06] sont meilleurs dans les scènes à illumination constante. En considérant uniquement les scènes où la lumière varie, les histogrammes d'orientation des gradients (HOG) [DT05] donnent de bons résultats. Cet algorithme est utilisé pour apprendre un classifieur à base d'une machine à vecteurs de support dédié à la détection de personnes. La conservation du signe du gradient [XRB04] donne des résultats bruités, surtout dans les scènes comprenant des mouvements dans le fond, là où les algorithmes à base d'histogrammes démontrent leurs capacités de lissage. Dans ces mêmes conditions, l'algorithme "Moyenne et Seuil" échoue par de trop nombreuses fausses détections. Si le "Moyenne et Variance" parvient à filtrer le mouvement, il échoue également dans les cas de mouvements imprévisibles et donc non appris (scène "Waving Flowers" dans [NBB06]).

3.3.4 Conclusion

L'algorithme proposé fournit les meilleurs performances sur l'ensemble des tests. Ce résultats montre sa capacité à segmenter correctement une silhouette dans des cas qui correspondent à de

Moy. Seuil.	Moy. Var.	Hist. Coul.	Hist. Coul. à Noy.	Cons. Gradient	Hist. Class. des Grad.	HOG	Hist. Orient. à Noyaux.
23	23	17	16	7	18	20	16

TAB. 3.1: Rapidité des algorithmes testés (en images par seconde) sur une Machine bi-processeur Xeon 3.4 GHz.

nombreuses situations réelles : petits objets en mouvement dans le fond et changement brutaux d'illumination. Le noyau Gaussien spatial permet de filtrer les petits mouvements en donnant moins de poids aux objets qui entrent sur le bord d'une zone locale. Le noyau Gaussien sur l'orientation des contours résout le problème des erreurs de quantification qui est à l'origine d'un nombre important de fausses détections dans les algorithmes à base d'histogrammes. De cette manière, une étape de filtrage des erreurs n'est pas nécessaire et ceci permet à l'algorithme présenté ici de fonctionner en temps réel sur une machine bi-processeur Xeon $3,4$ GHz (voir tableau 3.1).

La figure 3.12 montre l'implémentation de la soustraction de fond en association avec les indices utilisés dans l'algorithme de suivi. Les distracteurs issus du fond sont neutralisés en grande partie. Ce procédé permet de superposer la silhouette et les contours issus des membres (fig. 3.12-2b) pour fournir un contenu plus informative que la silhouette seule utilisée par exemple par [AT06b] ou [ST02].

3.4 Modèle du corps pour la détection des membres

Deux modèles sont utilisés pour modéliser le corps. Le modèle graphique représente les facteurs de compatiblité entre les membres pour servir de support à l'algorithme de propagation des croyances (§2.2.3) associé au filtrage particulaire (§2.3.1). Le second est un modèle géométrique du corps qui va permettre d'évaluer la compatibilité des hypothèses par rapports aux indices extraits de l'image.

3.4.1 Modèle graphique

Conformément au paragraphe §2.2.2, la modélisation des interactions entre les membres utilise un graphe de facteurs où chaque facteur relie deux états modélisés par des nœuds qui représentent les membres. Le graphe de facteurs utilisé pour modéliser le corps est représenté en figure 3.13. Il inclut des liens articulaires entre les membres adjacents (traits continus), des liens destinés à empêcher les collisions entre la tête et les mains (traits interrompus) et des liaisons temporelles (traits pointillés).

Un facteur exprime la compatibilité entre deux membres. Pour un lien articulaire, la compatibilité est élevée si deux membres adjacents se touchent au niveau de leur point d'articulation. En revanche, pour les deux mains qui ne doivent pas rentrer en collision, le facteur sera nul pour une telle configuration. Il en va de même pour la tête et les mains.

L'originalité du modèle mis en œuvre ici provient du fait qu'il intègre au calcul des facteurs, des limites articulaires entre les membres afin de contraindre la pose sur des configurations cohérentes du corps humain. Par exemple les clavicules, en liaison avec le torse, échangent avec lui des informations sur l'inclinaison du torse et évaluent ainsi un encadrement de l'angle maximum qu'elle peuvent faire avec lui.

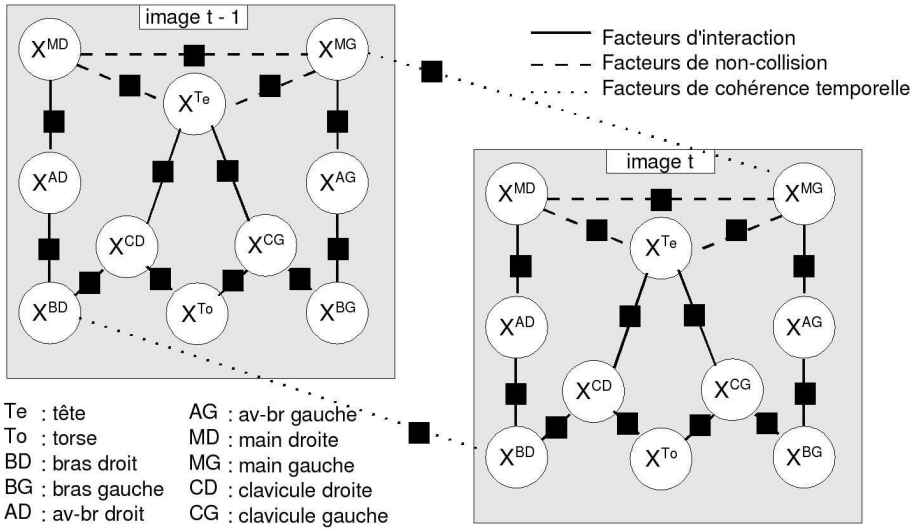


FIG. 3.13: *Graphe de facteurs utilisé pour modéliser le haut du corps humain.*

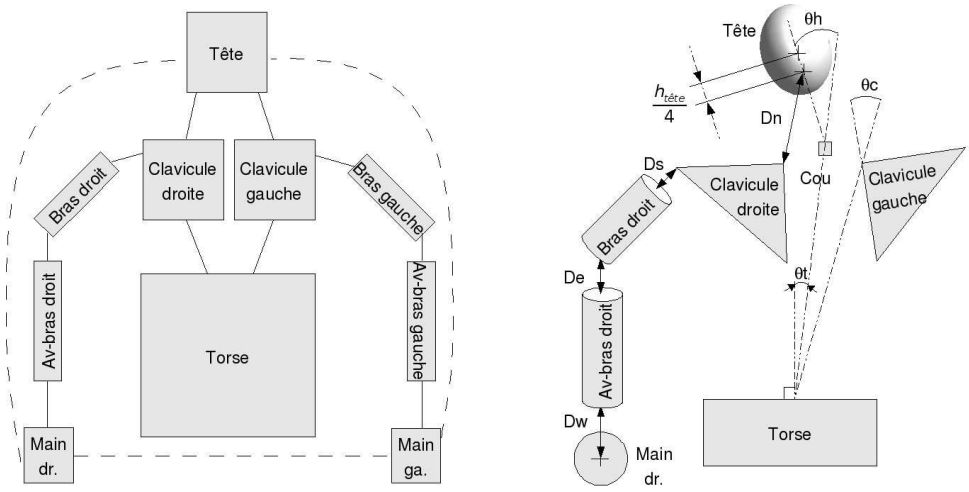


FIG. 3.15: Interactions entre les membres (figure de gauche) : les noeuds correspondent aux membres, les contraintes articulaires sont représentées par les traits pleins et les pointillés représentent des contraintes de non-collision entre la tête et les mains. Modèle du haut du corps (figure de droite) : les interactions entre les membres sont calculées à partir des distances (Dn, Ds, De et Dw) qui les séparent. Les autres contraintes articulaires sont déduites des angles θh, θc et θt.

3.4.2 modèle géométrique du corps

Alors que les conclusions sur l'état de l'art s'orientaient vers un modèle purement 3D (§1.5.3), le modèle de corps utilisé représente le haut du corps avec un mélange de modèle cardboard [JBY96] et 3D (fig. 3.15). Les membres capables de mouvements amples sur plusieurs degrés de liberté sont représenté en 3D en utilisant une sphère pour la tête et des cylindres pour les bras et les avant-bras. La précision fournie par les indices étant insuffisante pour distinguer l'orientation des mains ainsi que l'angle des clavicules par rapport au plan image, ces membres sont modélisés par des éléments 2D (cercles pour les mains et triangles pour les clavicules). Un procédé de détection original du torse (§3.4.3) utilise une modélisation sous la forme d'un rectangle faisant face à la caméra. Les membres sont discrétisés dans l'espace à l'aide de points régulièrement distribués autour d'eux. La figure 3.14 montre la projection sur le plan image des points discrétisant chaque membre :

FIG. 3.14: *Projection des points qui discrétisent les membres sur le plan image. De haut en bas : en bleu, les points de la tête, les lignes vertes et rouges pour les clavicules, la projection des spirales correspondent aux bras et avant-bras et les mains sont matérialisées par des disques oranges.*

- tête : une distribution régulière de points répartis sur l'hemisphère visible par la caméra,
- bras et avant-bras : des points distribués le long d'une spirale épousant leur surface,
- mains : des points répartis sur la surface du disque qui les modélisent,
- clavicules : des points régulièrement espacés sur les segments supérieurs des clavicules,
- torse : il est modélisé par une grille rectangulaire de points située au bas de l'image.

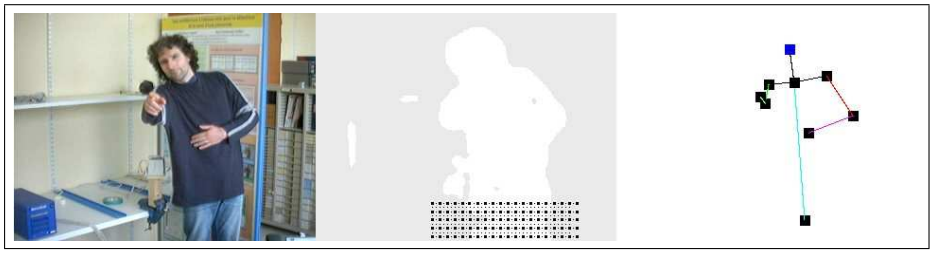


FIG. 3.16: Position du torse. La grille de points noirs sur le bord inférieur de l'image modélise la position du bassin. Elle se déplace horizontalement pour maximiser la correspondance entre les points de la grille et les pixels détectés positifs par la soustraction de fond (pixels blancs). L'énergie est maximum lorsque la grille est centrée sur la zone inférieure de l'image marquée positivement par la soustraction de fond. Le haut du torse est situé entre les deux clavicules.

3.4.3 Exploitation des indices extraits de l'image

Les facteurs de compatibilité par rapport aux observations extraites de l'image $\phi^{\mu}(X_t^{\mu}, Y_t^{\mu})$ sont calculés à partir des scores S_f^{μ} représentant la compatibilité entre l'hypothèse d'un membre μ et l'indice f extrait de l'image. La vérification des hypothèses se fait d'après la projection sur le plan image des points de discrétisation qui appartiennent au membre. Cette projection notée $proj(\mu)$ respecte la transformation géométrique paramétrée par la position hypothétique du membre par rapport à la caméra (fig. 3.14).

Contrairement aux images stéréo [BCMC06], la vision monoculaire exige plus d'indices pour atteindre un niveau de robustesse satisfaisant. Une compatibilité basée sur une fusion multi-indices procure un score global : $S^{\mu} = \prod_f S_f^{\mu}$. Pour éviter les effets néfastes dûs aux distracteurs issus du fond de l'image, la soustraction de fond robuste aux variation lumineuses et aux mouvements des petits objets décrite précédemment est utilisée pour filtrer les indices (§ 3.3).

Suivi des mains et du visage

Le score de teinte chair S_c^{μ} est calculé d'après le modèle d'histogramme (§3.2.3) calculé durant la phase d'initialisation (§4.2.1). Les pixels r qui appartiennent à la projection $proj(\mu)$ des points appartenant au membre μ pour la pose candidate (§3.4.3) sont comparés à ce modèle afin de déterminer le score de couleur :

$$S_c^{\mu} = \sum_{r \in proj(\mu)} H(r) \tag{3.19}$$

La fonction $H(r)$ renvoie la valeur de la case d'histogramme dans laquelle le pixel r tombe.

Suivi du torse

Du fait de la déformation des vêtements ou des occultations qui se produisent lorsqu'une personne est en mouvement, le torse est une partie du corps difficile à localiser avec précision. Cependant, la position du bassin peut être estimée de manière précise si on suppose qu'il se trouve dans la zone inférieur de l'image. Le bassin est modélisé par un rectangle glissant horizontalement sur le bord inférieur de l'image. Il est discrétisé selon une grille régulière de pixels pondérés d'après la Gaussienne de la distance d'un point au centre $b(x_b, y_b)$ de la grille. Un pixel $r(x, y)$

appartenant au bassin est donc affecté d'un poids $w(r)$ tel que :

$$\begin{aligned} &\forall\, r(x,y) \in bassin\,, \\ &dx = x - x_b\,, \\ &dy = y - y_b\,, \\ &w(r) = \frac{1}{\sqrt{2\pi}\sigma_b} exp\left(-\frac{d_x^2 + d_y^2}{2\sigma_b^2}\right)\,. \end{aligned} \tag{3.20}$$

Où σ_b vaut la demi largeur du rectangle. Le bassin interagit avec la soustraction de fond qui fournit la probabilité $P(r \in fond)$ que le pixel r appartienne au fond (§ 3.3). Le score du bassin S^b est :

$$\begin{aligned} &\overline{P}(r \in fond) = 1 - P(r \in fond)\,, \\ &S^b = \sum_{r \in bassin} w(r)\overline{P}(r \in fond)\,. \end{aligned} \tag{3.21}$$

Le score est maximum lorsque le bassin est centré sur la zone inférieure de l'image détectée positivement par la soustraction de fond (figure 3.16). La position du bassin donne la base du torse. Le haut du torse est situé à la base du cou, à mi-distance des deux clavicules.

Suivi des bras, des avant-bras, et des clavicules

Les bras ont tendance à bouger rapidement et sont fréquemment sujet aux occultations. Dans ces conditions, une fusion d'indices basée sur les gradients et l'énergie de mouvement permet d'obtenir un meilleur degré de robustesse.

Un score de gradients est estimé en prenant non seulement en compte l'intensité $\|\overrightarrow{E}\|$ du gradient (§3.2.1) $\overrightarrow{E}$ mais aussi son orientation $dir(\overrightarrow{E_r})$. Ce score d'orientation des gradients est noté S^μ_{or} pour une hypothèse faite sur le membre μ. Comme pour la soustraction de fond (§3.3.2), l'amplitude des gradients dans l'image est modulée par la fonction $f(\|\overrightarrow{E}\|)$ définie à l'équation 3.10 pour pénaliser les forts modules de gradients. Le score est calculé en considérant la Gaussienne $G_\theta(.)$ de la différence entre l'orientation θ_{limb} du membre (axe du bras ou de l'avant bras et ligne de l'épaule pour la clavicule) et l'orientation $dir(\overrightarrow{E_r})$ du contour obtenue en chaque pixel r qui correspond à la projection $proj(\mu)$ dans l'image de tous les points qui discrétisant le membre hypothèse (fig. 3.14) :

$$S^\mu_{or} = \sum_{r \in proj(\mu)} f(\|\overrightarrow{E}\|) G_\theta[\theta_{limb} - dir(\overrightarrow{E_r})]\,. \tag{3.22}$$

La variance du noyau G_θ doit être assez faible pour fournir une bonne sélectivité sur l'orientation des gradients. En pratique, cette valeur est fixée à 10°.

Le score de l'énergie de mouvement avantage les hypothèses de membres situés dans des zones contenant du mouvement (§3.2.2). Il est calculé selon la somme des probabilités de mouvement (eq. 3.4) pour tous les pixels r résultant de la projection des points appartenant aux membres sur le plan de l'image (fig. 3.14) :

$$S^\mu_m = 1 + \sum_{r \in proj(\mu)} P\left(I_m(r) = 1\right)\,. \tag{3.23}$$

La constante égale à 1 additionnée au score permet de neutraliser celui-ci quand il n'y a pas de mouvement dans l'image. Pour les clavicules, seul le score de gradients est activé car elles bougent peu et sont suffisamment contraintes par la position de la tête.

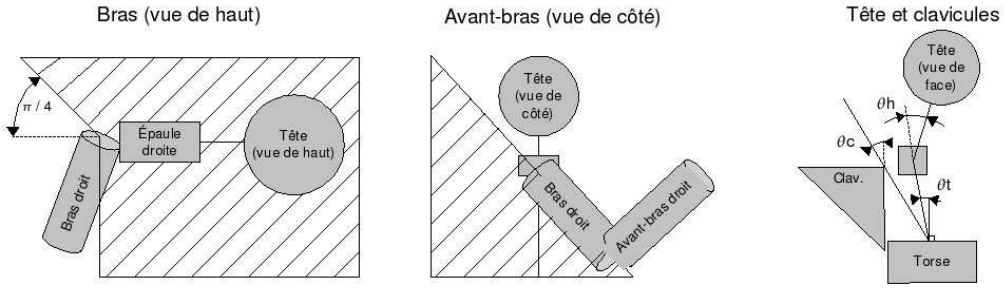


FIG. 3.17: Contraintes articulaires. Bras et avant bras : les parties hachurées montrent les zones interdites. Les contraintes angulaires sont : $|\theta_c| \leq 15°$ pour les clavicules et $|\theta_h| \leq 25°$ pour la tête. L'inclinaison du torse θ_t n'est pas limitée.

3.4.4 Indice basé sur l'apprentissage

Le modèle de corps à membres indépendants s'associe avantageusement à un apprentissage. L'idée est de ne plus apprendre des poses entières mais seulement l'apparence d'un membre dans l'image. De cette façon, la dimension de l'espace des poses à couvrir s'en trouve considérablement réduit. Une base de données de taille réduite est associée à chacun des membres pour donner un grand nombre de poses complètes possibles. La cohérence des poses est assurée par la propagation des croyances (§2.2.3).

Le procédé d'apprentissage utilise un capteur magnétique de mouvements associé à une webcam calibrée de manière à connaître les paramètres intrinsèques de la caméra et sa position par rapport au capteur (ses paramètres extrinsèques). Ce procédé a l'avantage d'enregistrer des nouveaux exemples dans la base de données à la cadence permise par la caméra (30 images par secondes).

Lors de l'apprentissage, les membres sont projetés dans l'image à partir des coordonnées fournies par le capteur de mouvement en respectant le modèle *pin hole* estimé pour la caméra. Un membre est codé sous la forme d'histogrammes locaux des orientations à noyaux Gaussien (§3.3.2) ordonnés d'après la même séquence des points qui constitue le contour du membre et calculés sur la zone qui entoure la projection dans l'image de chacun de ces points.

Lors du suivi, le score issu de l'apprentissage est donné par la comparaison des histogrammes calculés d'après l'hypothèse d'un membre avec ceux contenus dans la base de données. Les histogrammes d'orientation locaux à noyaux Gaussiens respectent un protocole de comparaison qui tient compte de l'influence du bruit de la caméra sur la valeur des gradients (§3.3.2). Le score de compatibilité par rapport à l'image est la valeur maximum obtenue lors de la comparaison des histogrammes avec la base. Sur les M points qui discrétisent un membre μ, et sachant les valeurs P_m, $(m \in M)$ retournées par la comparaison des M histogrammes, le score est donné par :

$$S^{\mu}_{app} = \prod_{m \in M} (1 + P_m) \ . \tag{3.24}$$

Ce critère s'interprête comme le codage de l'apparence d'un membre dans l'image selon une suite d'histogrammes calculés autour des points qui discrétisent les contours du membre. La position dans l'espace des exemples appris n'est pas incluse dans la base et on cherche simplement à savoir si une hypothèse possède une ressemblance avec un membre de la base en comparant ce codage généré de la même façon pour la base d'apprentissage et pour les hypothèses. Ce critère est appliqué aux bras et aux avant-bras.

3.4.5 Contraintes articulaires

Un ensemble de règles sur les contraintes articulaires sont intégrés au calcul des facteurs d'interaction. Un facteur d'attirance entre les membres adjacents prend la forme d'une Gaussienne de la distance entre ces membres (voir la figure 3.15 pour les distances Dn, Ds, De et Dw). Cette Gaussienne est centrée sur une valeur égale à la distance qui sépare la tête des clavicules pour le cou et centrée en zéro pour les épaules les coudes et les poignets. Pour respecter les limites articulaires, les facteurs d'interaction torse-tête ou torse-clavicules sont nuls si les angles θh ou θc dépassent un seuil fixé (figure 3.17). Des zones interdites (figure 3.17) et des liens additionnels qui affectent une probabilité nulle à une solution qui fait entrer les mains et la tête en collision (contraintes de non-collision, figure 3.15) complètent cet ensemble de règles destiné à fournir des poses cohérentes.

Chapitre 4

Résultats expérimentaux

Sommaire

4.1 Introduction

La validation des choix opérés au cours des chapitres précédents doit passer par une série de test, tant quantitatifs que qualitatifs, qui incluent des comparaisons portant sur les performances obtenues afin d'améliorer l'algorithme et son paramétrage et de placer ses performances par rapport à l'état de l'art. Si la comparaison d'un même algorithme dans différentes configurations ne pose pas de problèmes majeurs, la confrontation de celui-ci avec les résultats issus de l'état de l'art est moins aisée. En effet, il existe pas à ma connaissance de bases destinée à la comparaison des différents travaux portant sur le suivi de gestes. En conséquence, et même avec la batterie de tests mise en œuvre dans ce chapitre, l'établissement d'une comparaison objective avec les approches similaires reste hasardeux. Cependant, une description précise des conditions d'essai et une analyse quantitative des résultats permet de dégager des tendances fortes et met en lumière l'aboutissement ou non des objectifs fixés ainsi que les points qui sont susceptibles d'être améliorés.

Les objectifs établis à l'origine de ce travail exigeaient un suivi du haut du corps temps réel utilisant un matériel grand public pour la prise de vue dans un environnement non contraint par les vêtements ou l'environnement. La conformité des résultats obtenus avec ces objectifs a été testée qualitativement dans différents environnements avec des sujet, des vêtements, des lieux et des éclairages variés. La précision du suivi à fait l'objet de plusieurs test quantitatifs utilisant un capteur magnétique de mouvements pour établir la vérité de terrain.

Les différentes configurations testées de l'algorithme suit l'ordre chronologique de son développement. Une première version consiste à suivre les gestes parallèles au plan de l'image en 2D (§4.2). L'introduction de contraintes articulaires à permis de contraindre la pose dans les trois dimensions (§4.3). Enfin, les derniers temps de la thèse ont étés consacrés à l'étude préliminaire d'un indice lié à l'apprentissage. Le dernier paragraphe (§4.4) de ce chapitre donne quelques conclusions tirées de cette étude.

4.2 Suivi en 2D

Dans cette configuration, le système est destiné à suivre des gestes sur le plan 2D de l'image, le personnage faisant face à la caméra. Plusieurs scènes ont été testées dans des conditions variées. Les images sont acquise grâce à une caméra de type webcam avec un format couleur 320×240 pixels à une cadence de 30 images par secondes.

Il n'est pas nécessaire de changer la taille des membres pour chaque scène (tab. 4.1). Celle-ci est fixée à l'avance grâce à des données anthropomorphiques corrigées pour tenir compte des différences entre la modélisation des articulations (fig. 3.15) et la réalité. Les différences entre les morphologies sont le plus souvent rattrapées par le taux d'élasticité entre les membres (§3.4.5). Si ce dernier est trop lâche, la cohérence du corps n'est pas suffisamment contrainte et le suivi perd en précision. Dans le cas contraire, un nombre trop restreint d'hypothèses tirées lors de la phase de prédiction (alg. 1) sont viables ce qui aboutit à une perte du suivi. La variance de

	tête	clavicule	bras	avant-bras	main
longueur (cm)	22	20	23	26	10
largeur (cm)	15	0	8	7	10

Tab. 4.1: Taille des membres pour le modèle de corps.

l'élastique Gaussien entre les membres est fixé à $4cm$ pour mener à un bon compromis.

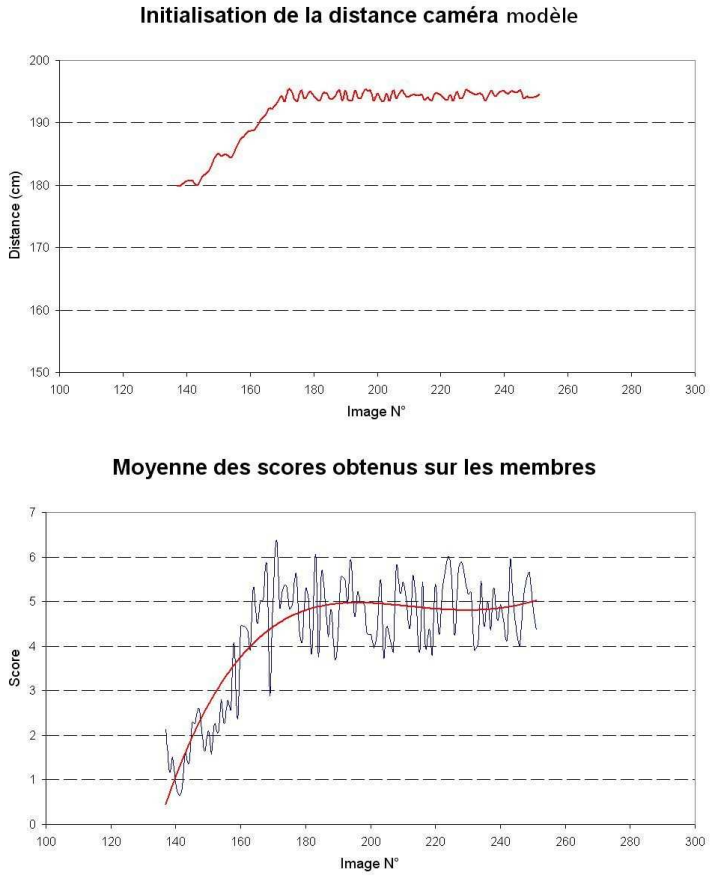


FIG. 4.1: *Initialisation automatique de la distance du modèle à la caméra. En haut : évolution de cette distance durant une courte scène d'initialisation. Graphe du bas : une moyenne des scores sur les membres du corps est donné pour la même scène. La distance évolue pour se stabiliser lorsque le score global est maximum au bout d'une cinquantaine d'images.*

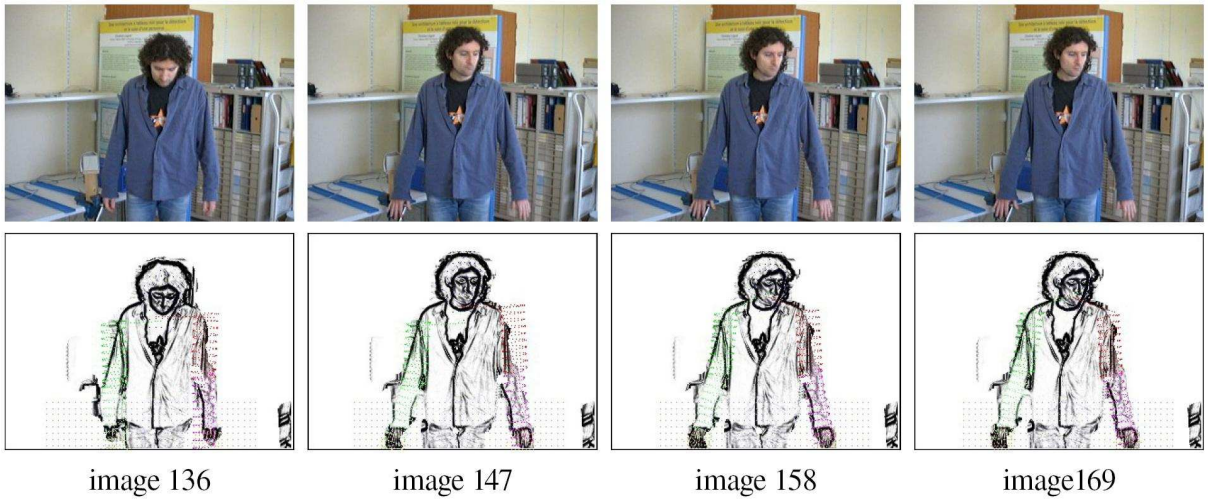


FIG. 4.2: *Initialisation automatique de la distance du modèle à la caméra. La première ligne montre les images originales, la seconde ligne contient la projection du modèle dans l'image des contours.*

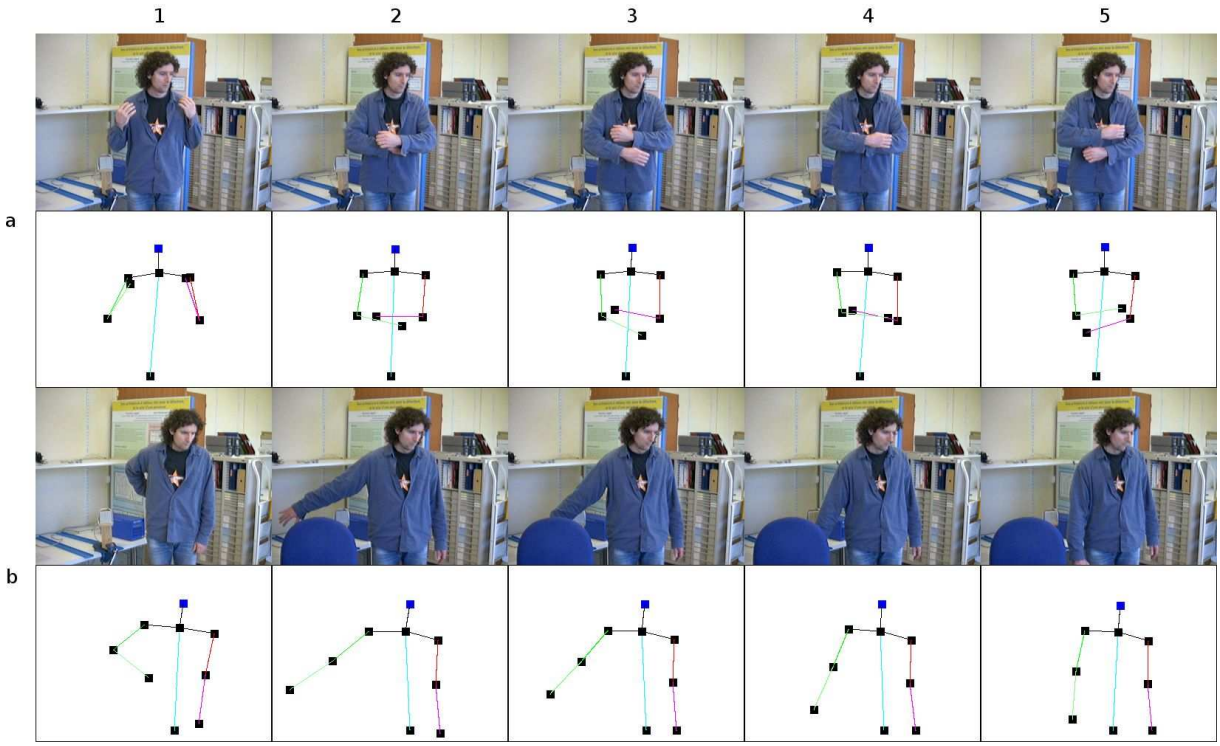


FIG. 4.3: *Cas d'auto-occultations des bras (a1) et de la main par l'avant-bras (a2, a4), une main derrière le dos (b1) et occultation de la main par le dossier d'un fauteuil (b2 à b5).*

4.2.1 Initialisation

La pose de départ suppose que les bras se trouvent le long du corps et le torse en position verticale faisant face à la caméra. Le suivi peut aisément raccrocher la pose réelle tant que celle-ci n'est pas trop éloignée de cette hypothèse.

L'estimation en profondeur du modèle va influencer la taille de sa projection dans l'image. Le système va adapter celle-ci pour minimiser les contraintes sur les liaisons entre les membres adjacents et maximiser leur scores (fig. 4.1). Cela signifie que l'algorithme ne nécessite pas d'entrer manuellement la distance exacte entre le sujet et la caméra, ce paramètre ce règle automatiquement dès les premières images. L' exemple donné (fig. 4.2) illustre bien le comportement général du système à l'initialisation. Sur la figure, on distingue aisément la correction de profondeur effectuée qui amène à adapter l'échelle de la projection du modèle à l'image. Au départ, sur l'image 136, le modèle se trouve trop avancé et les bras débordent des épaules sur l'image des contours. Au bout d'une trentaines d'images, la distance optimale du modèle à la caméra est trouvée.

4.2.2 Résultats qualitatifs en suivi 2D

Ces résultats ont pour but de démontrer les capacités de l'algorithme à s'adapter aux variations de l'environnement. Diverses personnes font l'objet de tests dans différents lieux. Mais dans un premier temps, il est intéressant de soumettre le système à quelques occultations.

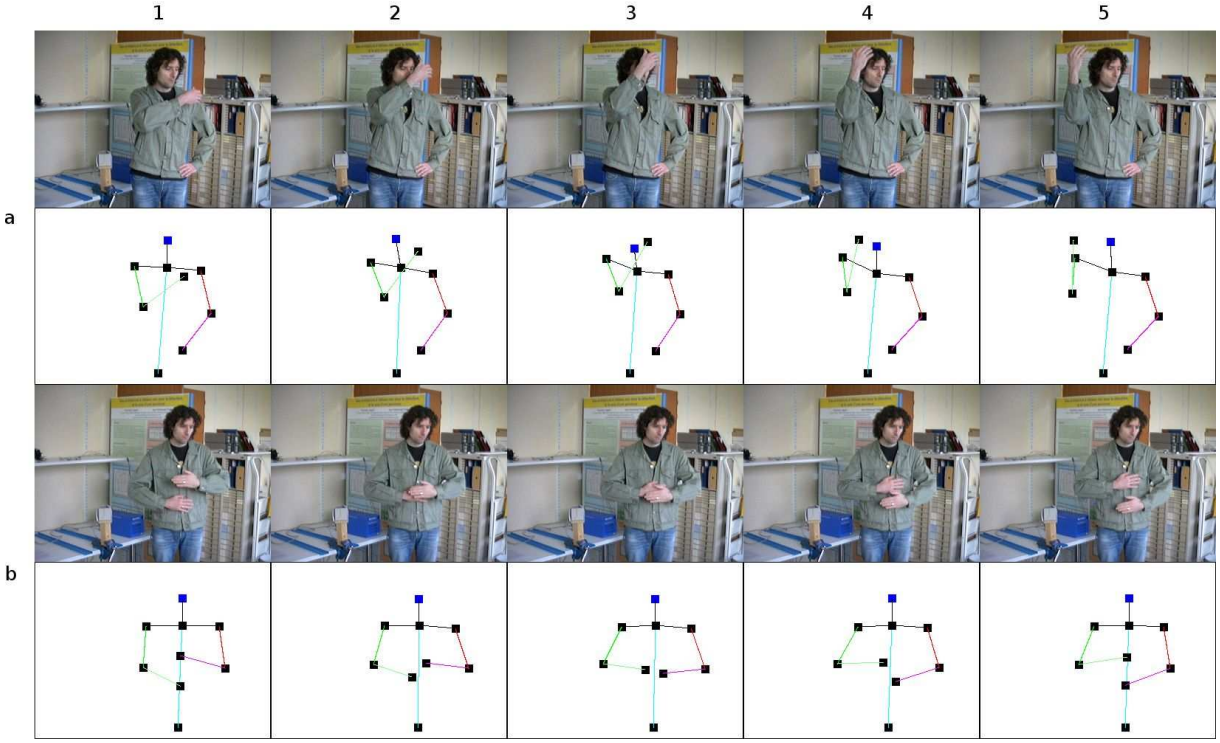


FIG. 4.4: *Cas d'auto-occultations de la tête avec une main (a1 à a5) et des mains entre-elles (b1 à b5).*

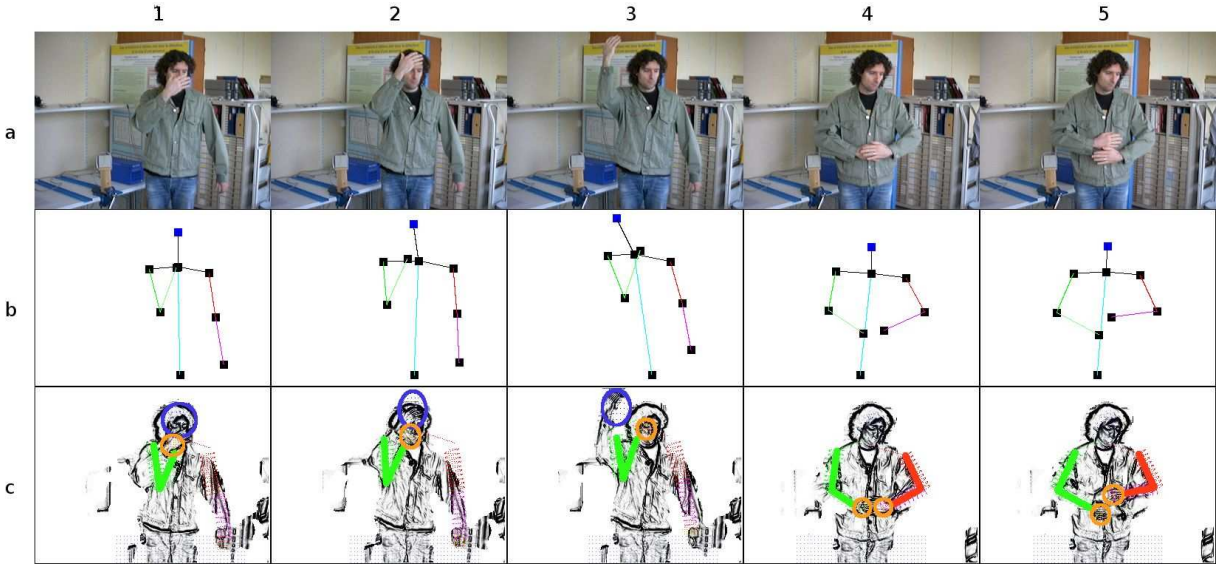


FIG. 4.5: *Cas d'échecs dûs à l'échange entre les membres occultés. La ligne c correspond à la projection du modèle sur l'image des contours. Pour plus de clarté, la position des membres à été mise en évidence par une ligne verte ou rouge pour les bras et un cercle bleu pour le visage ou orange pour les mains.*

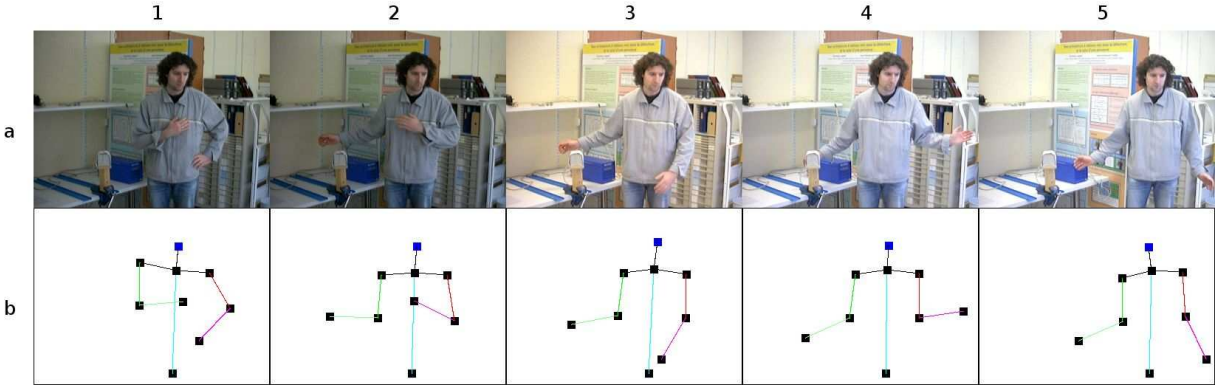


FIG. 4.6: *Cas des changements lumineux. L'intensité et la couleur changent brusquement après l'allumage des lampes dans la pièce.*

Le cas des occultations

Les mouvements naturels dans le plan 2D de l'image ont tendance à générer des auto-occultations, à savoir le masquage d'un membre par un autre membre. Les poses présentées à la figure 4.3 ont été tournées en temps réel et montrent quelques cas courants d'occultations. Dans le cas des occultations entre une main et la tête ou entre les mains, la contrainte de non collision entre ces membres font que les membres ont tendance à se contourner plutôt que se superposer (fig. 4.4). Cet inconvénient peut conduire à une perte de suivi par inversion de l'affectation des membres occultés (fig. 4.5).

Ces résultats montrent que l'algorithme développé est capable de gérer une partie des occultations en inférant la position du membre occulté à partir de l'estimation des membres adjacents non occultés. Le passage des messages entre les membres lors de la propagation des croyances (§2.2.3) permet de mener à bien ces inférences. Par exemple, dans le cas de la main qui passe derrière le dossier d'un fauteuil (fig. 4.3), la position de celle-ci est contrainte par la partie visible de l'avant-bras, elle-même contrainte par l'estimation du bras qui ne pose pas de problèmes majeurs.

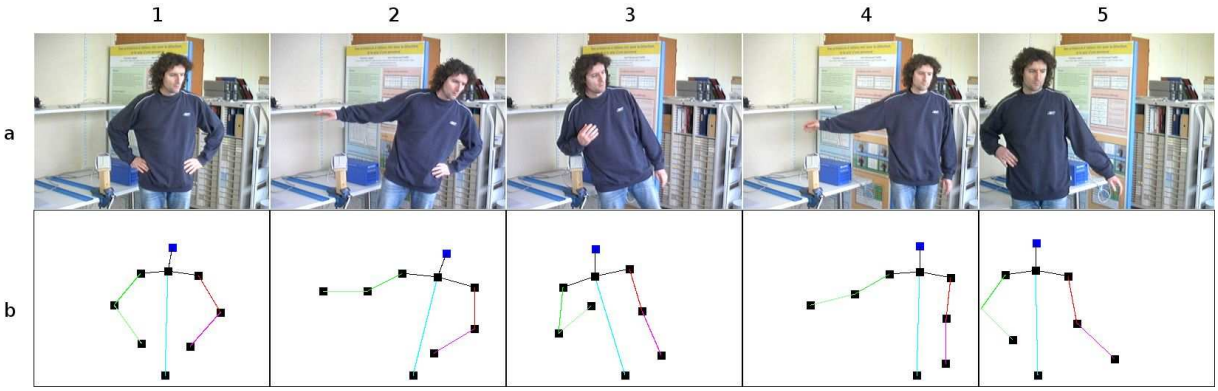


FIG. 4.7: *Suivi du torse au cours de mouvement latéraux incluant l'inclinaison du buste.*

Le cas des changements lumineux

Le choix des gradients (§3.2.1) pour construire les indices de soustraction de fond (§3.3) et d'orientation des contours pour le suivi des bras et des avant-bras (§3.4.3) procure une robustesse du suivi vis-à-vis des changements lumineux dans la scène.

La figure 4.6 montre un changement brutal en intensité et en couleur de la lumière ambiante sans que cela ne perturbe le suivi. On remarquera particulièrement la bonne gestion de l'occultation de la main (image 4) ainsi que le placement correct du torse lors d'un léger déplacement du personnage vers la droite de l'image (images 5 et 6).

Suivi du torse

Le suivi du torse n'est pas trivial et le dispositif mis en œuvre pour accomplir cette tâche (§3.4.3) est testé au cours d'une scène où le sujet incline le buste (fig. 4.7, images 2 et 3) et se déplace latéralement (fig. 4.7, images 4 et 5). L'indice de soustraction de fond utilisé pour localiser le bassin peut être mis en échec par un pardessus ou un imperméable qui déforme la silhouette (fig. 4.8).

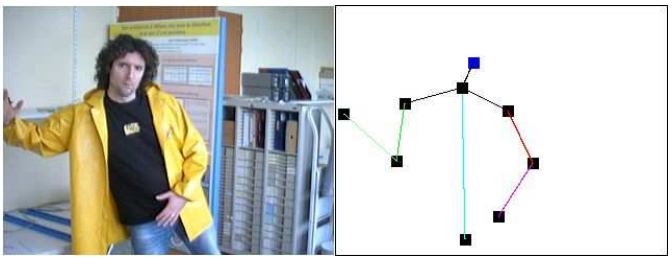


FIG. 4.8: *L'inclinaison du torse n'est pas correctement estimée du fait du pan du ciré qui pend à droite.*

Suivi de l'orientation des épaules et de l'inclinaison de la tête

Les épaules présentent également des difficultés du fait de leur petite taille et de leur variation d'apparence entre, par exemple, le bras le long du corps ou le bras levé qui escamote la clavicule (fig. 4.9, image 1). L'épaule est principalement contrainte par le bras lorsque les gradients générés par son image sont de faible amplitude. Dans ce cas, une imprécision dans le suivi du bras peut fausser l'estimation de la clavicule. Grâce à sa forme en ellipsoïde (fig. 3.15), le modèle de tête est capable de suivre l'inclinaison du visage à condition que celui-ci ne soit pas circulaire. C'est généralement le cas lorsque le cou est visible ou le visage est allongé. La figure 4.9 montre différentes configurations pour la tête et les épaules.

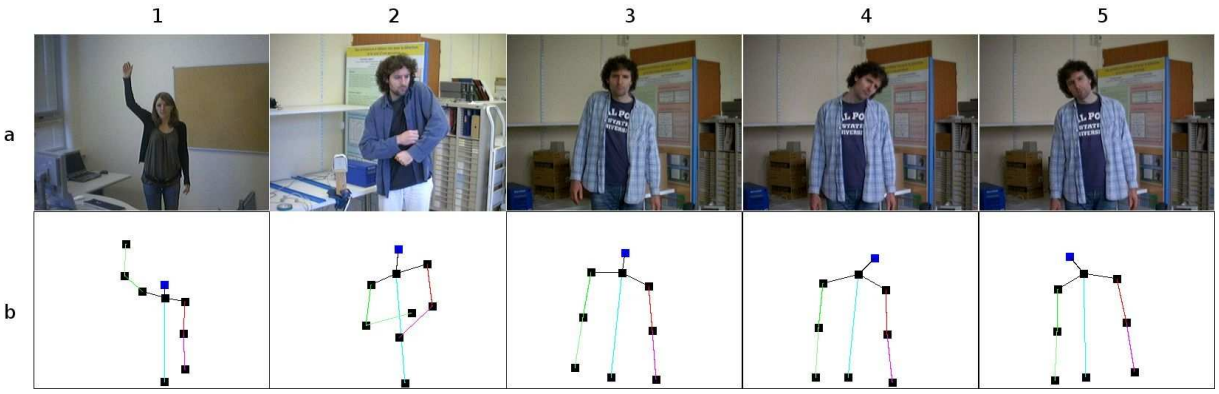


FIG. 4.9: *Suivi de l'orientation des épaules et de l'inclinaison de la tête.*

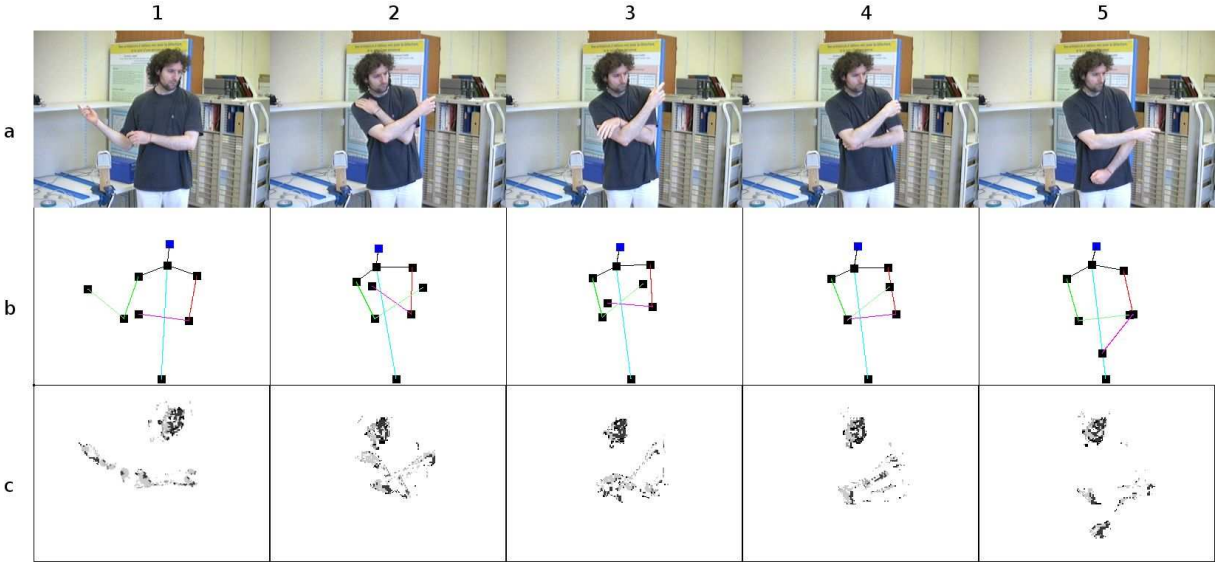


FIG. 4.10: *Le suivi des bras et des mains n'est pas perturbé par le fait que les bras soient découverts et apparaissent sur la carte de teinte chair, la modifiant significativement (ligne c).*

Suivi des bras

Les résultats exposés au cours des paragraphes précédents montrent la capacité de l'algorithme à suivre les bras de manière satisfaisante dans des situations variées. Cependant, il semble raisonnable de penser que l'utilisation de la teinte chair pour suivre le visage et les mains puisse être une cause de perturbation lorsque le sujet porte des vêtements à manches courtes. Le phénomène à craindre est de voir la main remonter le long de l'avant bras. Cependant, et en dépit de la teinte chair qui caractérise ce dernier, le bras et l'avant-bras parviennent à contraindre correctement la main à sa position optimale dans la plupart des cas.

Tests dans des situations diverses

Les tests qui suivent vont clore ce paragraphe en montrant le suivi des différents sujets (fig. 4.11) afin d'évaluer l'adaptation de l'algorithme à des morphologies différentes. Ces images

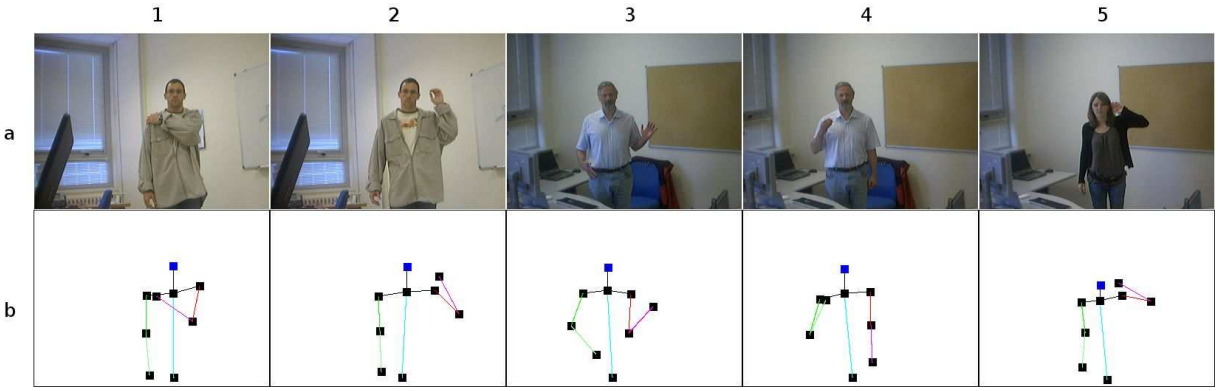


FIG. 4.11: *Trois sujets différents sont suivis sans changer les tailles paramétrant les membres.*

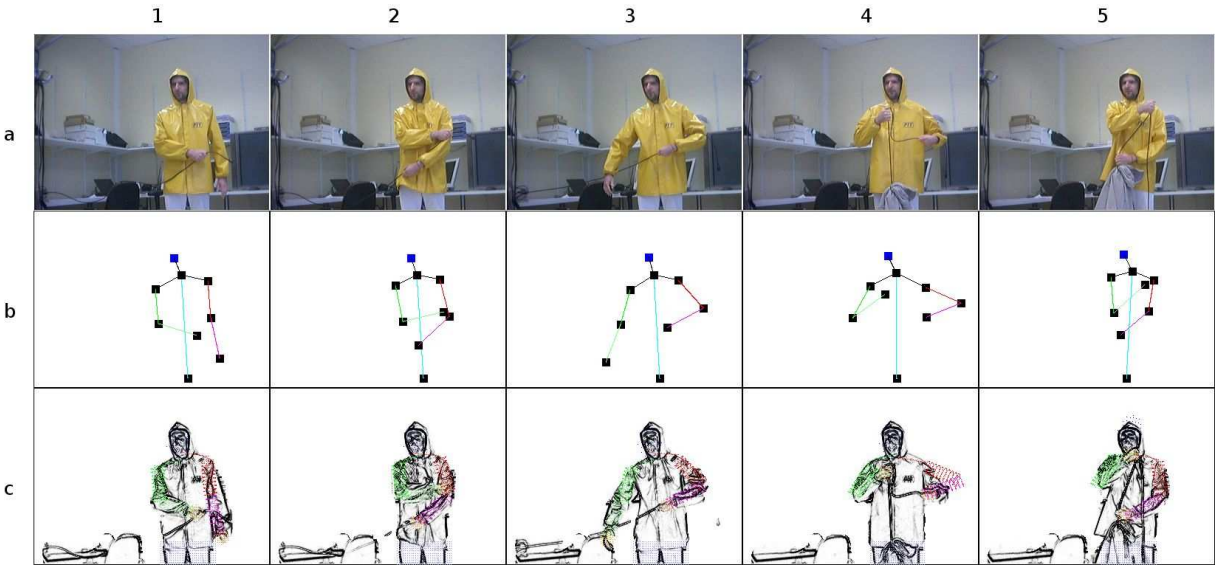


FIG. 4.12: *Scène de "pêche en haute mer". Le suivi 2D limite la précision sur l'estimation du coude à l'image* 4. *La ligne c représente la projection du modèle dans l'image des contours. La projection des points discrétisant le bras gauche est en rouge et magenta, le bras droit en vert et vert clair, la tête en bleu et les mains en orange.*

prises dans des scènes complètes, montrent la capacité de l'algorithme à suivre rigoureusement le personnage, même lorsque celui-ci est éloigné de la caméra et sa taille à l'image s'en trouve réduite. Sans changer les paramètres de taille des membres au cours des trois essais, l'algorithme réussit avec des vêtements et des morphologies différentes (étudiant sportif, homme mûr, jeune fille svelte).

La scène suivante (fig. 4.12) reproduit la remontée d'un filet de pêche. Lorsque les bras se trouvent devant le buste (images 2 et 5), le port du ciré complique leur segmentation. Cependant, le suivi reste fidèle si on exclu l'image 4 où l'avant-bras gauche est perpendiculaire au plan de la caméra et provoque une erreur d'estimation du coude gauche bien visible sur l'image des contours (image 4*c*). Cette pose non recherchée pour le tournage de cette scène de test censé rester en 2D, s'est produite d'autant plus facilement que les mouvements sur un plan 2D ne sont pas naturels à l'homme, d'où la nécessité d'aborder le suivi en trois dimensions (§4.3).

4.2.3 Résultats quantitatifs en suivi 2D

Après avoir exploré les points forts et les limites de l'algorithme de suivi en 2D, la question de la précision doit maintenant se poser. La vérité de terrain est donnée par un capteur magnétique qui mesure la position de l'épaule $E(E_x, E_y, E_z)$, du coude $C(C_x, C_y, C_z)$ et du poignet $P(P_x, P_y, P_z)$ (voir repères en annexe B). Pour mener à bien les tests quantitatifs, les paramètres du modèle de la caméra et les coordonnées de celle-ci par rapport au repère du capteur magnétique de mouvement doivent être estimés. Ce problème courrament rencontré pour le calibrage des caméra dans le cadre des procédés stéréos ou multicaméra (§1.2.4) est résolu de manière automatique à partir d'une acquisition au cours de laquelle la main équipée d'un capteur balaye l'espace vu par la caméra. Une optimisation numérique est menée sur les paires de coordonnées 3D-image enregistrées pour déterminer les paramètres intrinsèques et extrinsèques de la caméra.

Les paramètres intrinsèques de la caméra sont utilisées pour projeter les hypothèses de po-

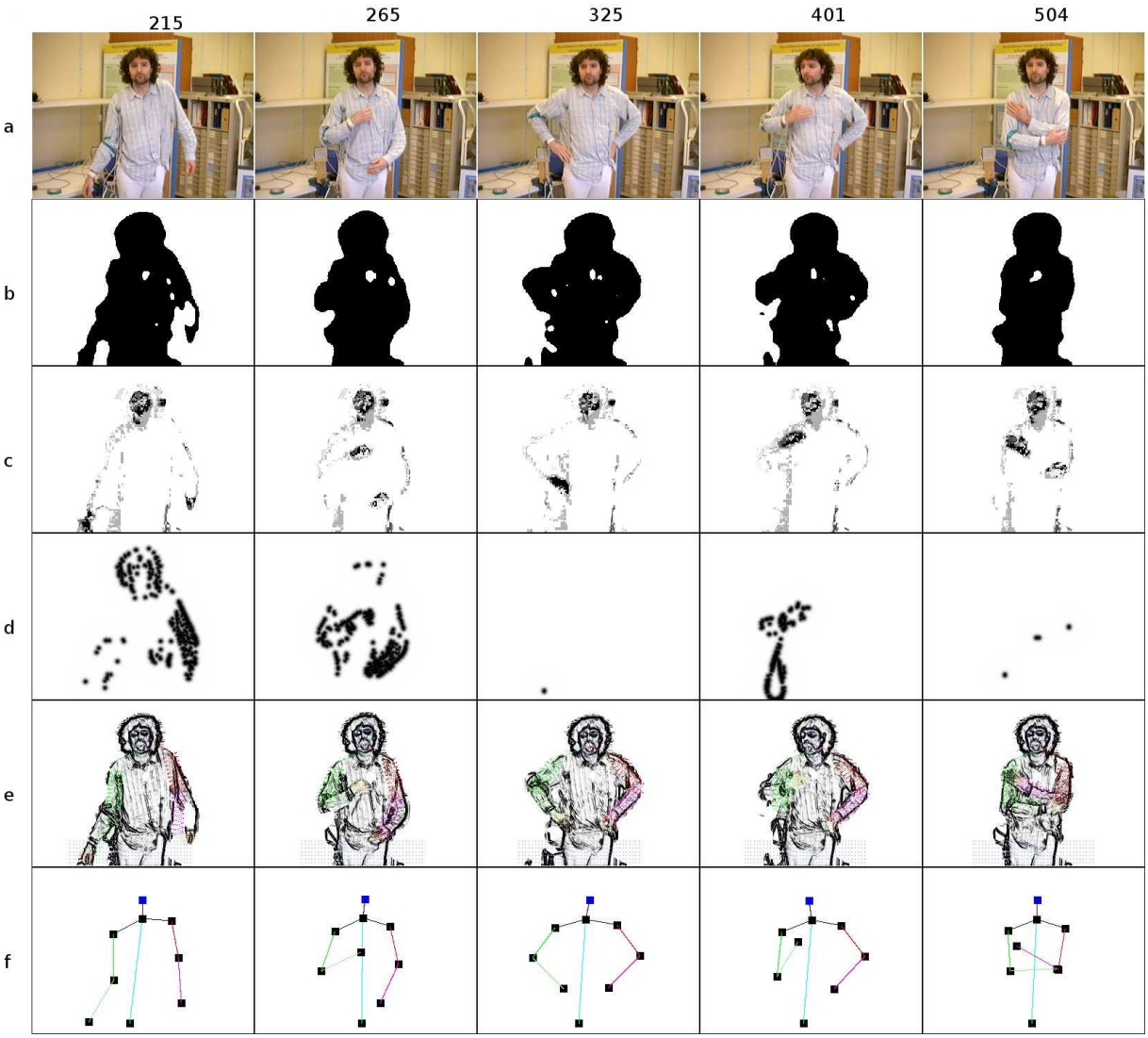


FIG. 4.13: *Suivi 2D. Quelques images de la séquence de test utilisée pour établir la vérité de terrain. Tous les indices représentés : (a) images originales, (b) soustraction de fond, (c) carte de probabilité de la teinte du visage, (d) carte de l'énergie du mouvement et (e) contours. La pose résultat est donnée à la ligne (f). On remarquera la faute de suivi sur le bras droit pour l'image 401.*

sition des membres dans l'image et les coordonnées fournies par le capteur magnétique sont exprimées dans le repère de la caméra grâce aux paramètres extrinsèques.

Calcul de l'erreur commise lors de l'estimation de la position d'un membre

En monoculaire, la profondeur est une donnée relative qui ne peut être estimée qu'à un facteur d'échelle près. Le protocole de calcul de l'erreur commise consiste tout d'abord à calculer $\overline{Z}$, la profondeur moyenne du personnage pour chaque image :

$$\overline{Z} = \frac{E_z + C_z + P_z}{3} . \tag{4.1}$$

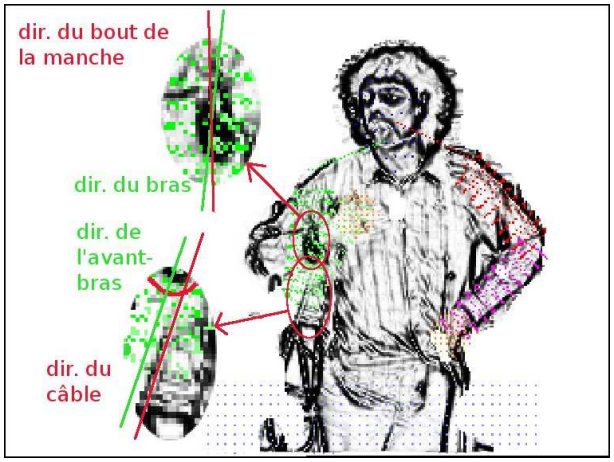


FIG. 4.14: *Agrandissement de l'image 401 des contours. La mauvaise estimation du bras et de l'avant bras droit est en partie dûe au bout de la manche et à un câble du capteur de mouvement utilisé pour établir la vérité de terrain.*

Avant d'êtres comparées à la vérité de terrain, les coordonnées estimées doivent être corrigés en tenant compte du rapport entre la profondeur vraie et estimée. Une première évaluation de l'erreur sur les coordonnées estimées de l'épaule, $(\tilde{E}_x, \tilde{E}_y)$ (les calcul pour le coude et le poignet sont identiques) donne :

$$\begin{aligned} \Delta E_x &= \alpha_u \frac{\overline{Z}}{\tilde{E}_z} \tilde{E}_x - E_x \,, \\ \Delta E_y &= \alpha_v \frac{\overline{Z}}{\tilde{E}_z} \tilde{E}_y - E_y \,. \end{aligned} \tag{4.2}$$

Avec α_u et α_v définis par le produit de la focale de la caméra et des facteurs d'échelles horizontaux k_u et verticaux k_v en $(pixels/mm)$ du capteur CCD :

$$\begin{aligned} \alpha_u &= -k_u f \,, \\ \alpha_v &= -k_v f \,. \end{aligned} \tag{4.3}$$

Il reste maintenant à corriger les erreurs dues au placement des capteurs sur le corps qui ne correspondent pas forcément avec les points articulaires théoriques du modèle. Pour ce faire, il

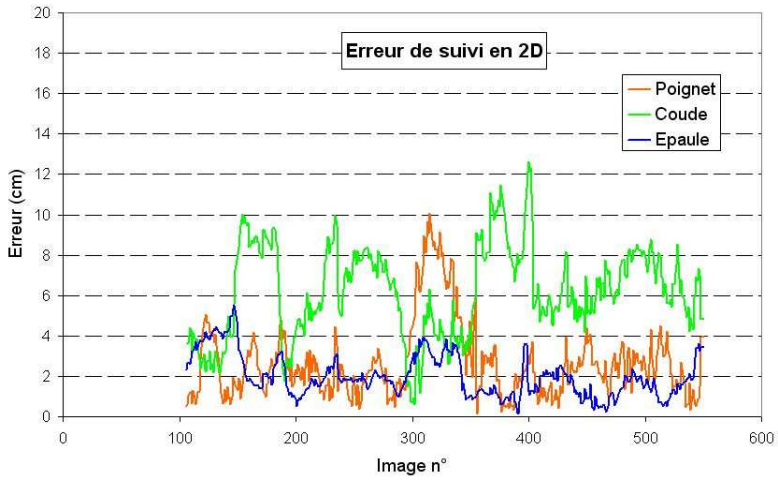


FIG. 4.15: *Résultats quantitatifs obtenus sur l'estimation de l'épaule, le coude et le poignet. La faute de suivi sur le bras droit apparaît autour de l'image 401..*

FIG. 4.16: *Suivi 3D. Résultats qualitatifs. La ligne (a) montre les images originales tirées de la scène de test, les trois autres lignes donnent respectivement la pose estimée en vue de face, de haut et de coté. Les occultations sont bien gérées et il en va de même pour le rendu de l'inclinaison du torse*

s'agit de soustraire la moyenne de l'erreur en x sur toute la séquence aux coordonnées estimées :

$$\begin{aligned} \Delta E_x^{cor} &= \Delta E_x - Moy\{\Delta E_x\}\,, \\ \Delta E_y^{cor} &= \Delta E_y - Moy\{\Delta E_y\}\,. \end{aligned} \tag{4.4}$$

L'erreur corrigée ΔE^{cor} pour l'épaule consiste à calculer la norme de l'erreur obtenue à partir de l'équation précédente (eq. 4.4) :

$$\Delta E^{cor} = \sqrt{(E_x^{cor})^2 + \left(E_y^{cor}\right)^2} \tag{4.5}$$

Scène de test utilisée

Similairement aux scènes précédentes, la scène de test comprends des mouvements exécutés parallèlement au plan de la caméra (fig. 4.13). Cette scène présente une erreur de suivi pour le bras et l'avant bras droit qui place le coude à une position erronée même si la clavicule et la main sont correctement estimées (fig. 4.14). La position des épaules est fortement influencée par la tête elle-même aisément localisée grâce à la teinte chair. Il en va de même pour le poignet qui est contraint par le main. Entre ces deux articulations, le coude à la croisée du bras et de l'avant bras, est source de nombreuses hypothèses plausibles. Celles-ci sont difficilement départageables lorsque les indices sont trompeurs comme ici où le bout de la manche et un câble provenant d'un capteur attirent le bras et l'avant bras dans une position trop verticale. Cette anomalie du suivi se répercute sur le graphe de quantification des erreurs (fig. 4.15). Les erreurs les plus importantes sont commises en estimant le coude ce qui corrobore l'idée que la localisation du coude pose souvent des problèmes.

4.3 Suivi 3D

Le fait de passer à un suivi 3D présente de nombreuses difficultés en monoculaire du fait des incertitudes portant sur la profondeur des objets observés. En plus de lier les membres adjacents par un facteur Gaussien qui diminue lorsque les deux membres s'éloignent l'un de l'autre, il paraît nécessaire de compenser le déficit d'informations sur la profondeur par un ensemble de règles qui modélisent les limites articulaires humaines (§3.4.5). Ce modèle est mis à l'épreuve au cours de scènes, où un sujet faisant face à la caméra exécute des gestes dans l'espace. Comme pour le chapitre précédent, les tests se décomposent en une évaluation qualitative (§4.3.1) qui utilise des scènes variées et une évaluation quantitative (§4.3.2). Ces deux paragraphes ont pour but d'évaluer les apports proposés dans le cadre de ce travail concernant l'intégration de la fusion d'indices et des règles articulaires à un algorithme qui à donné des résultats encourageants en vision stéréo [BCMC06].

4.3.1 Résultats qualitatifs en suivi 3D

Les poses montrées au cours de ce paragraphe occasionnent des occultations et présentent des fonds complexes. Les gestes sont exécutés de manière naturelle sans ralentir la vitesse des mouvements lors de la prise de vue.

Une première série (fig. 4.16) montre que le système a conservé ses capacités de gestion des occultations. Celles-ci sont plus nombreuses qu'en 2D du fait des bras qui peuvent maintenant pointer vers la caméra. L'image 1 montre l'avant bras droit perpendiculaire au plan de la caméra, l'image 5 reproduit une pose avec les bras croisés et la main droite cachée derrière l'avant-bras gauche.

La série suivante (fig. 4.17) consiste à faire varier les personnes et les environnements. L'image 2 montre une position difficile à rendre à cause du bras tendu partiellement occulté par la main droite. Un cas similaire avec les avant-bras apparaît à l'image 3, la 5 montre une estimation correcte avec les bras nus et des cheveux longs. et, sur l'image 6, le système parvient à estimer une pose inhabituelle.

Un troisième test utilise la séquence de pêche déjà présentée lors du suivi 2D (fig. 4.12). Cette scène semble plus adaptée à un algorithme 3D pour palier à l'erreur de suivi sur le coude. Le résultat est effectivement meilleur en 3D et l'estimation du coude gauche dans l'image 4

Erreur (cm)	En. Mvt.	Épaule	Coude	Poignet	Erreur moyenne globale
Moyenne	avec	2.9	10.4	13.7	9.0
	sans	2.7	10.4	17.0	10.1
Maximum	avec	9.7	26.4	25.9	18.3
	sans	6.9	25.0	33.4	18.4
Écart-type	avec	1.6	4.3	4.9	2.8
	sans	1.1	4.7	7.0	3.0
Vitesse moyenne ($cm.s^{-1}$)		16.0	19.5	41.6	
Vitesse maximum ($cm.s^{-1}$)		17.8	42.9	95.1	

TAB. 4.2: Moyenne, maximum et écart-type de l'erreur commise sur l'épaule, le coude et le poignet sur la séquence de test. L'erreur moyenne globale est la moyenne des erreurs commises en estimant la position de ces trois articulations. La vitesse moyenne est calculée sur toute la séquence pour chaque articulation.

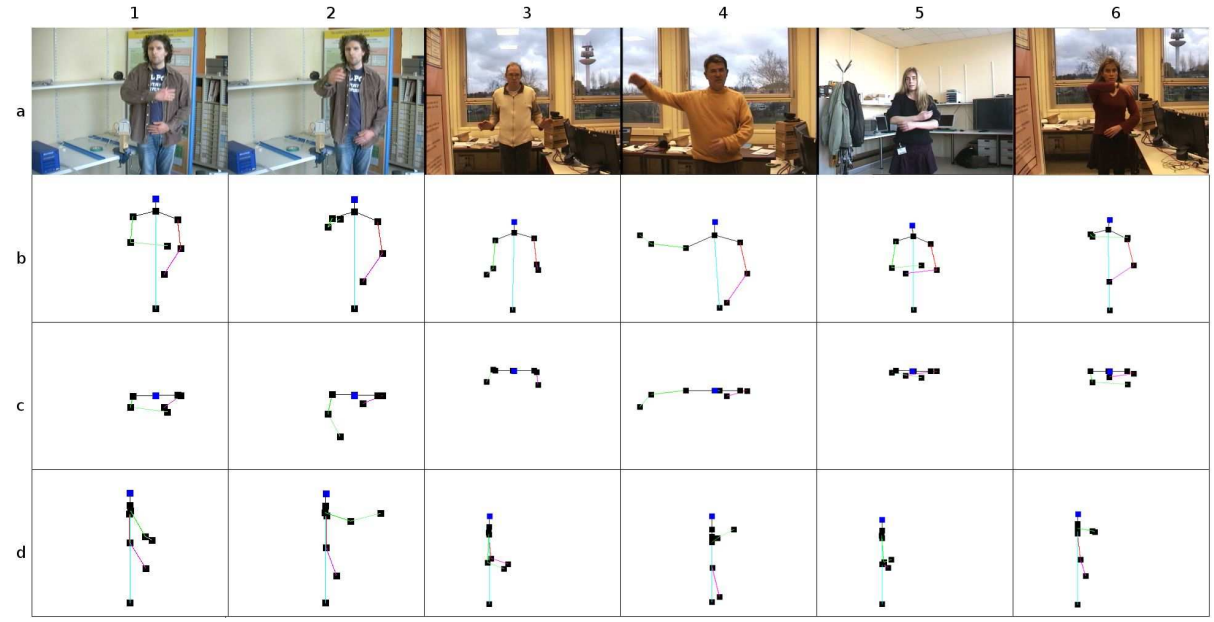


FIG. 4.17: *Suivi 3D. Quelques poses présentant des difficultés comme des occultations et des fonds complexes dans des environnements non contraints (conditions lumineuses et vêtements variés).*

est maintenant correcte (fig. 4.18). Cependant, l'estimation de la profondeur constitue le talon

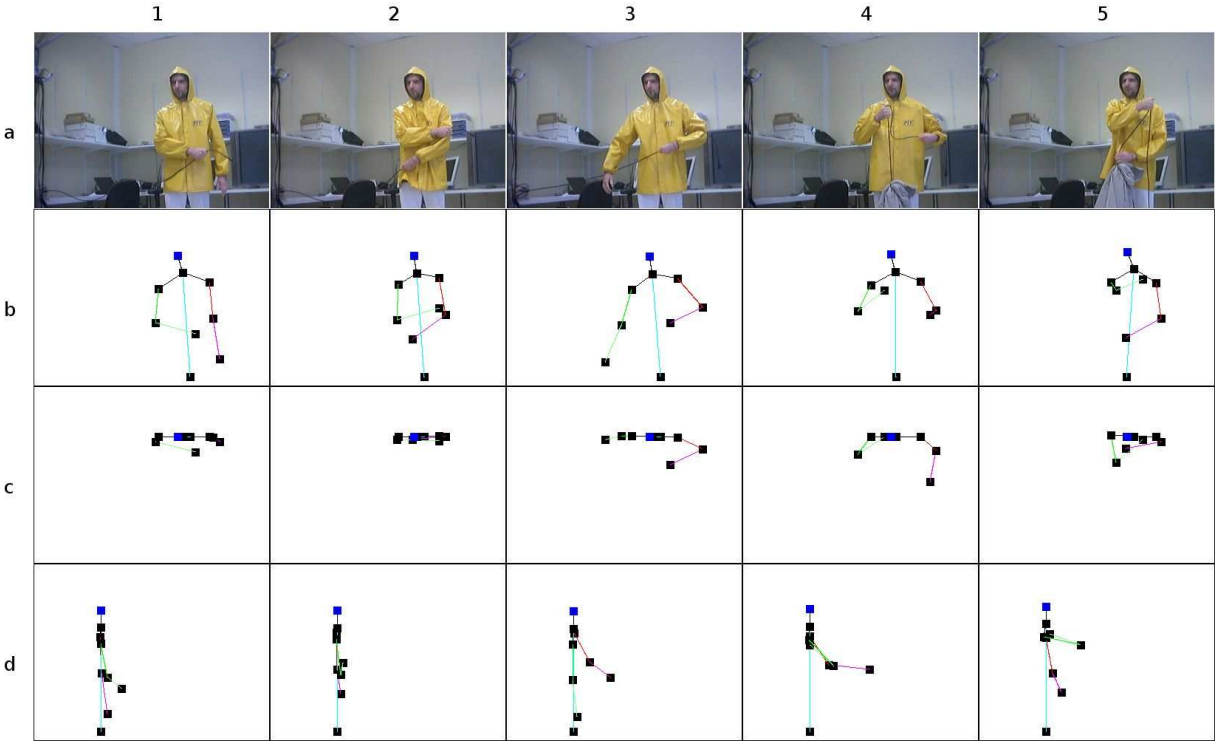


FIG. 4.18: *Scène de "pêche en haute mer" en suivi 3D. Les poses de face sont correctement estimées mais la profondeur de la main droite au cours des images 4 et 5 est fausse.*

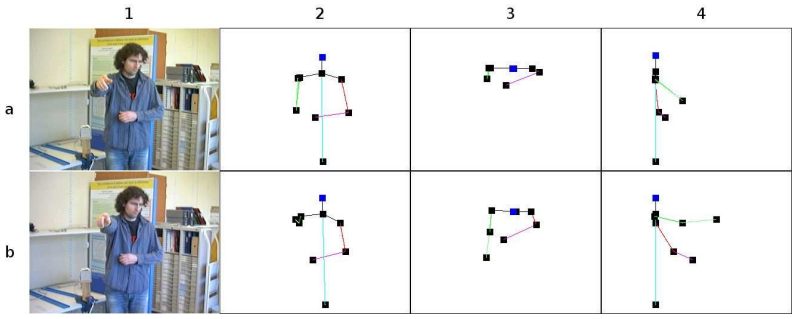


FIG. 4.19: *Tentative de suivi d'un geste de pointage. La première tentative est infructueuse (a) et le bras se superpose à l'avant-bras en position verticale. La seconde tentative est réussie (b).*

d'achille des systèmes monoculaires et l'illustration de ce principe se trouve à l'image 4 et 5 où la main droite est estimée trop près du torse même si sa position dans l'image de face (ligne b) est correcte.

La profondeur globale du modèle étant déjà variable dans la version 2D, le passage en suivi 3D consiste à ajouter un paramètre de rotation supplémentaire sur les bras et les avant-bras pour autoriser les gestes hors du plan image. Ce faisant, les bras qui comptaient trois paramètres de translation et un paramètre de rotation vont seulement gagner une cinquième inconnue pour la version 3D. De ce fait, la portée de l'échantillonnage donc les capacité exploratrice de l'algorithme dans ce nouvel espace sont peu altérées. Cependant, à la lumière de ces tests, le suivi 3D est apparu moins robuste qu'en 2D. Ce résultat n'est pas surprenant puisque l'inconnue supplémentaire se traduit par des effets de perspective difficilement observables. De ce fait, les principales causes de décrochement concerne l'évaluation du déploiement d'un membre perpendiculairement au plan de l'image (fig. 4.19).

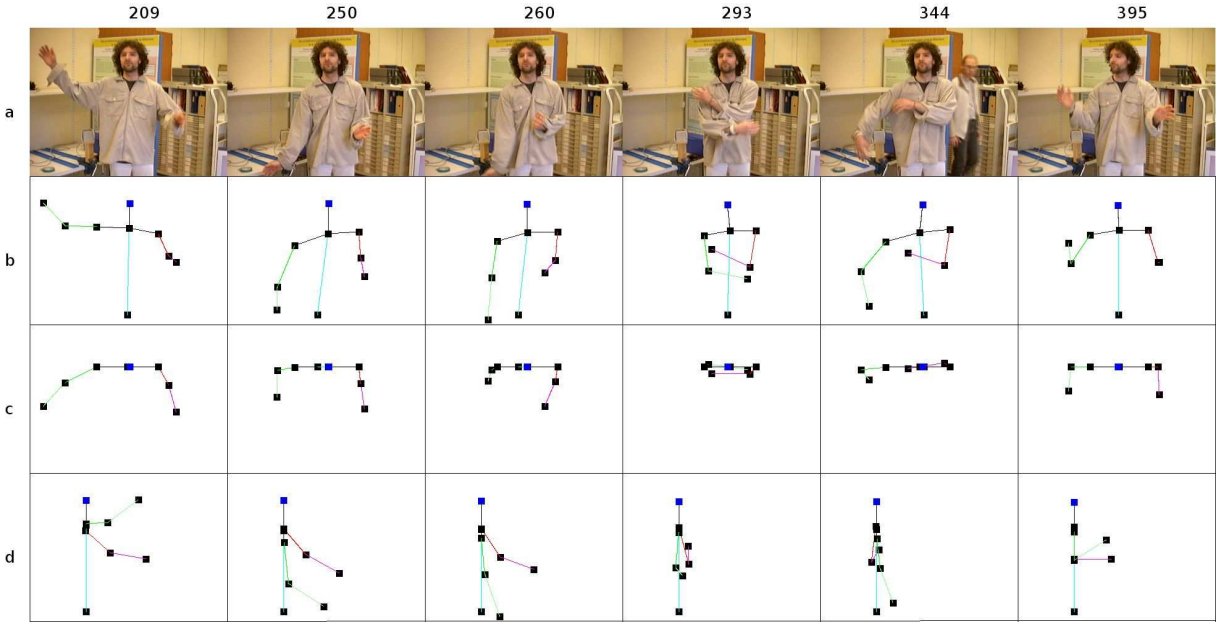
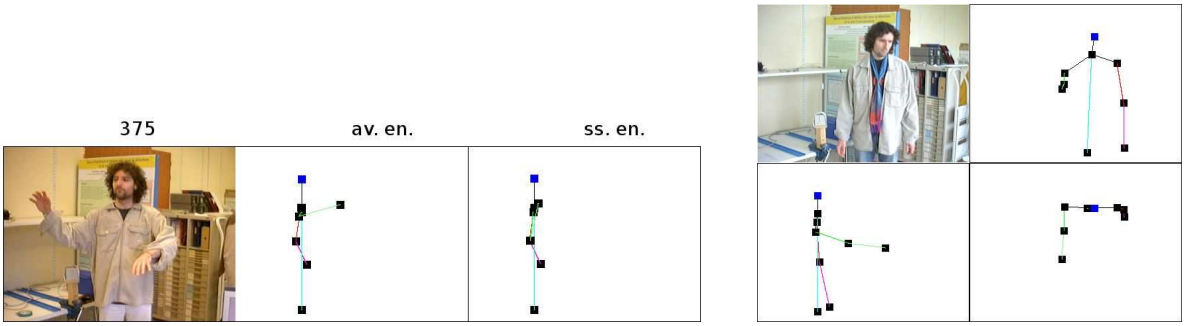


FIG. 4.20: *Quelques images de la scène de test pour le calcul des erreurs d'estimation. Les lignes représentent l'image originale (a), l'estimation de la pose vue de face (b), de haut (c) et de profil (d). Les images* 209, 260 *et* 395 *sont estimées de manière satisfaisante. Les autres montrent des erreurs de suivi caractéristiques qui se produisent principalement sur l'évaluation de la profondeur des membres.*



(a) *L'image* 375 *avec et sans l'énergie du mouvement. Dans le second cas le coude est correctement estimé mais la profondeur du poignet est mieux rendue avec l'énergie du mouvement.*

(b) *À la suite de la perte du suivi de la main droite le bras se tend naturellement face à la caméra.*

FIG. 4.21: *Fautes de suivi.*

4.3.2 Résultats quantitatifs en suivi 3D

Le protocole de mesure de l'erreur commise sur l'estimation de la position diffère de celui utilisé pour la version 2D. L'erreur calculée doit tenir compte des différences sur les trois axes tout en corrigeant l'estimation de profondeur globale du modèle.

Erreur commise en profondeur

La correction de la profondeur estimée du modèle se fait en confondant les coordonnées z (voir repères en annexe B) de l'épaule vraie et son estimation :

$$\Delta Z = \tilde{E}_z - E_z \,. \tag{4.6}$$

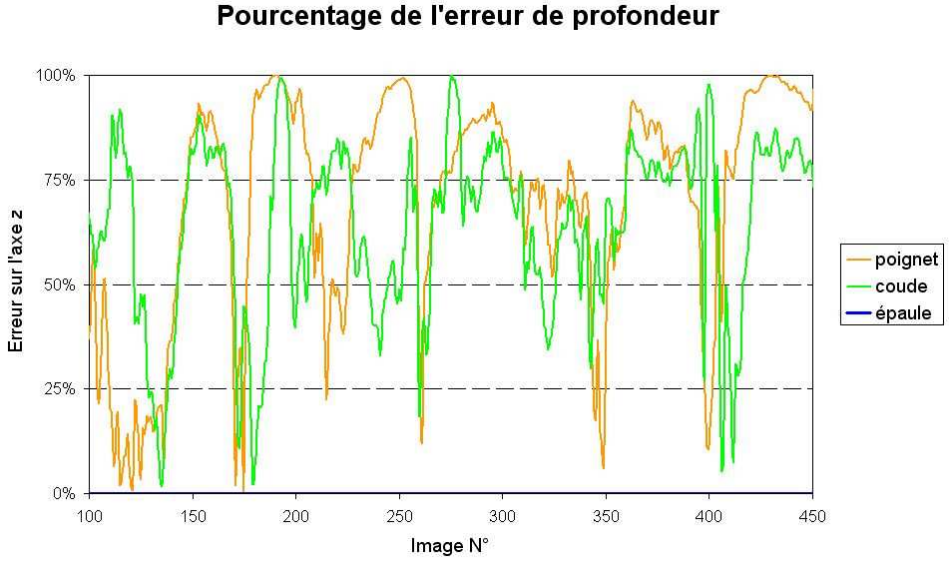


FIG. 4.22: *Pourcentage de l'erreur commise sur l'estimation de la profondeur des membres par rapport à l'erreur totale. Le pourcentage pour l'épaule est nul du fait du protocole de mesure choisi (§4.3.2).*

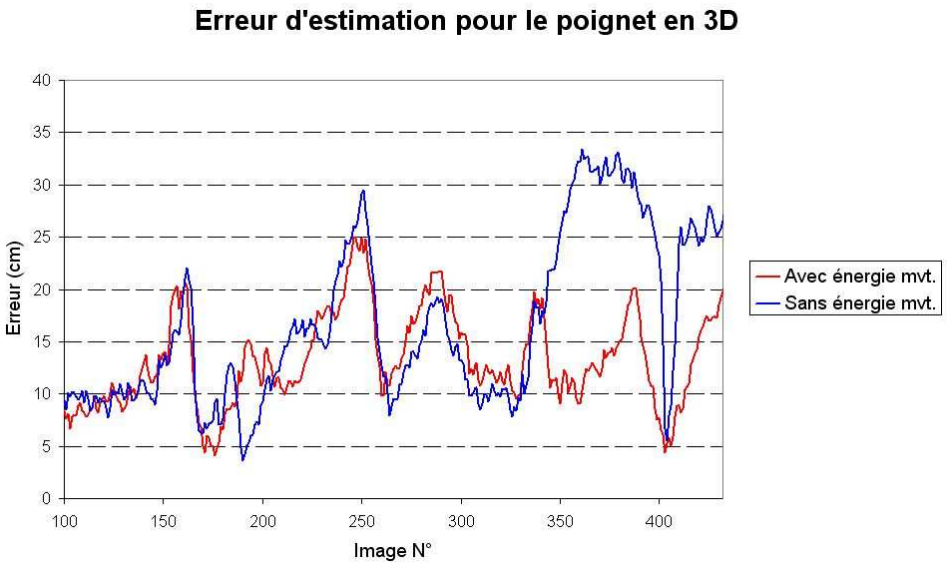


FIG. 4.23: *Erreur d'évaluation sur la position du poignet avec et sans l'indice d'énergie de mouvement.*

Les coordonnées de profondeur estimées sont soustraites avec cette valeur pour corriger l'estimation de la profondeur globale du modèle :

$$\begin{aligned} \tilde{E}_z^{cor} &= \tilde{E}_z - \Delta Z\,, \\ \tilde{C}_z^{cor} &= \tilde{C}_z - \Delta Z\,, \\ \tilde{P}_z^{cor} &= \tilde{P}_z - \Delta Z\,. \end{aligned} \tag{4.7}$$

Les erreurs commises en profondeur ΔE_z, ΔC_z, ΔP_z pour les trois articulations sont calculées à partir de ces données corrigées :

$$\begin{aligned} \Delta E_z &= \tilde{E}_z^{cor} - E_z = 0\,, \\ \Delta C_z &= \tilde{C}_z^{cor} - E_z\,, \\ \Delta P_z &= \tilde{P}_z^{cor} - E_z\,. \end{aligned} \tag{4.8}$$

Par définition, la valeur de ΔE_z est toujours nulle. Il reste maintenant à corriger le biais que l'on suppose induit par les différences entre la position des capteurs sur le corps et la définition des points articulaires sur le modèle. Pour cela, on soustrait aux coordonnées estimées la moyenne sur toute la scène de test des valeurs calculées à l'équation 4.8 :

$$\begin{aligned} \Delta E_z^{cor} &= \Delta E_z = 0\,, \\ \Delta C_z^{cor} &= \Delta C_z - Moy(\Delta C_z)\,, \\ \Delta P_z^{cor} &= \Delta P_z - Moy(\Delta P_z)\,. \end{aligned} \tag{4.9}$$

Erreurs dans le plan parallèle à l'image

La valeur de l'erreur sur les autres axes ne nécessite pas le recalage en profondeur fait à l'équation 4.7. En revanche il faut corriger le facteur d'échelle sur les axes x et y qui tient compte

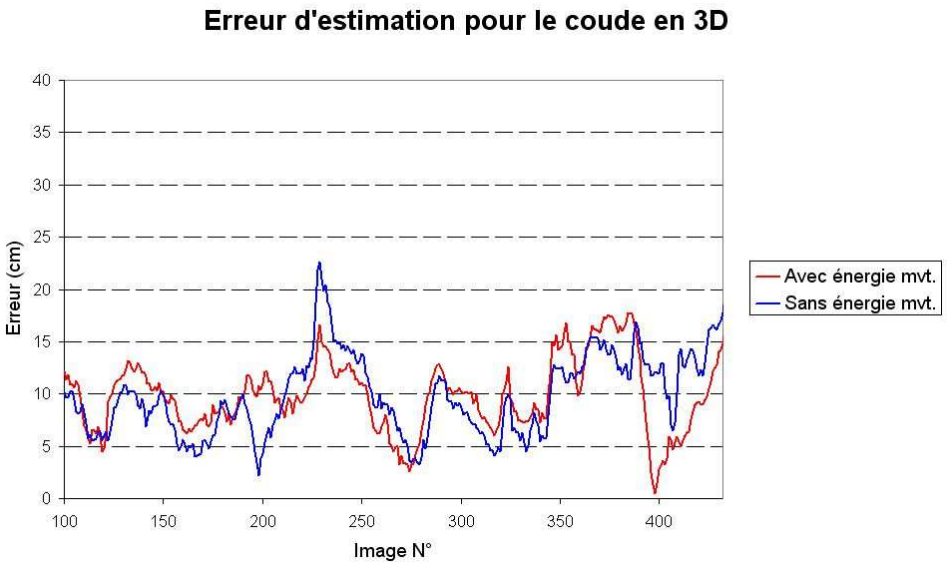


FIG. 4.24: *Erreur d'évaluation sur la position du coude avec et sans l'indice d'énergie de mouvement.*

du rapport des profondeurs vraies et estimées. Ce rapport est mesuré au niveau de l'épaule :

$$\begin{aligned} \tilde{E}_x^{cor} &= \alpha_u \frac{E_z}{\tilde{E}_z} \tilde{E}_x \,, \\ \tilde{E}_y^{cor} &= \alpha_v \frac{E_z}{\tilde{E}_z} \tilde{E}_y \,. \end{aligned} \tag{4.10}$$

Les coefficients α_u et α_b sont définis à l'équation 4.3. Une première estimation de l'erreur donne :

$$\begin{aligned} \Delta E_x &= \tilde{E}_x^{cor} - E_x \,, \\ \Delta E_y &= \tilde{E}_y^{cor} - E_y \,, \end{aligned} \tag{4.11}$$

et après correction sur le placement des capteurs en calculant la moyenne des erreurs sur toutes les images, on obtient :

$$\begin{aligned} \Delta E_x^{cor} &= \Delta E_x - Moy(\Delta E_x) \,, \\ \Delta E_y^{cor} &= \Delta E_y - Moy(\Delta E_y) \,, \end{aligned} \tag{4.12}$$

Le calcul des erreurs pour le coude (C_x^{cor}, C_y^{cor}) et le poignet (P_x^{cor}, P_y^{cor}) sont identiques.

Erreurs d'estimation commise

L'erreur totale sur un membre consiste simplement à calculer la norme des erreurs commise sur chaque axe :

$$\Delta E^{cor} = \sqrt{(\Delta E_x^{cor})^2 + \left(\Delta E_y^{cor}\right)^2 + (\Delta E_z^{cor})^2} \,. \tag{4.13}$$

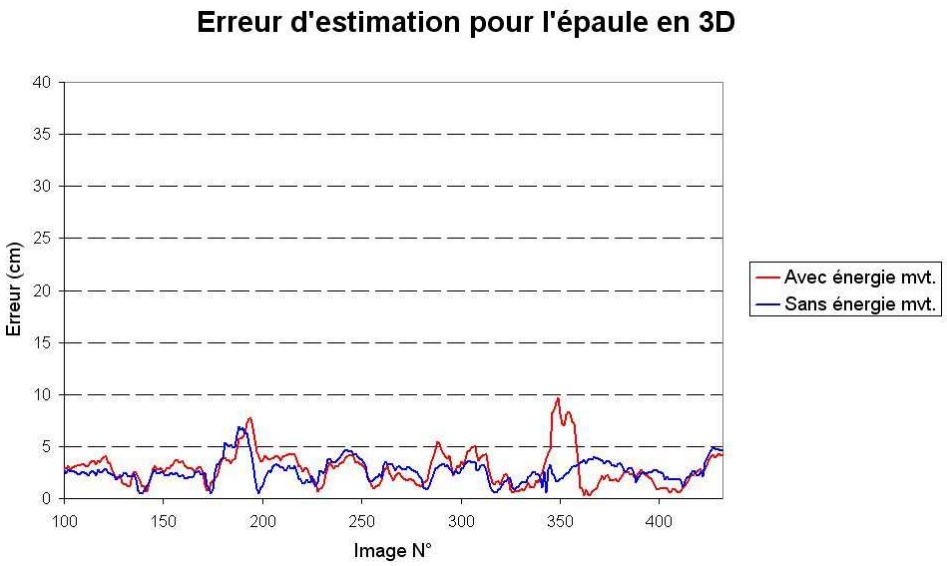


FIG. 4.25: *Erreur d'évaluation sur la position de l'épaule avec et sans l'indice d'énergie de mouvement.*

Résultats sur une scène de test

Les gestes de la scène test sont pleinement exécutés dans les trois dimensions avec une vitesse naturelle (tab. 4.2), La seule contrainte étant de faire approximativement face à la caméra (fig. 4.20). Les erreurs sont calculées pour le suivi avec et sans l'indice d'énergie du mouvement. Dans les deux cas, les résultats sont similaires jusqu'à l'image 350 où une erreur de suivi en

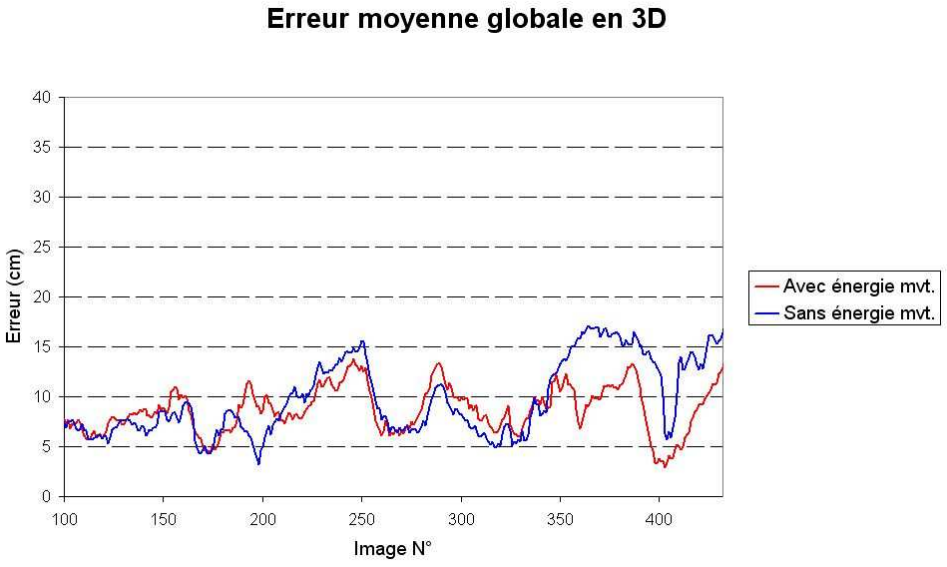


FIG. 4.26: *Moyenne de l'erreur d'évaluation sur la position de l'épaule, du coude et du poignet avec et sans l'indice d'énergie de mouvement.*

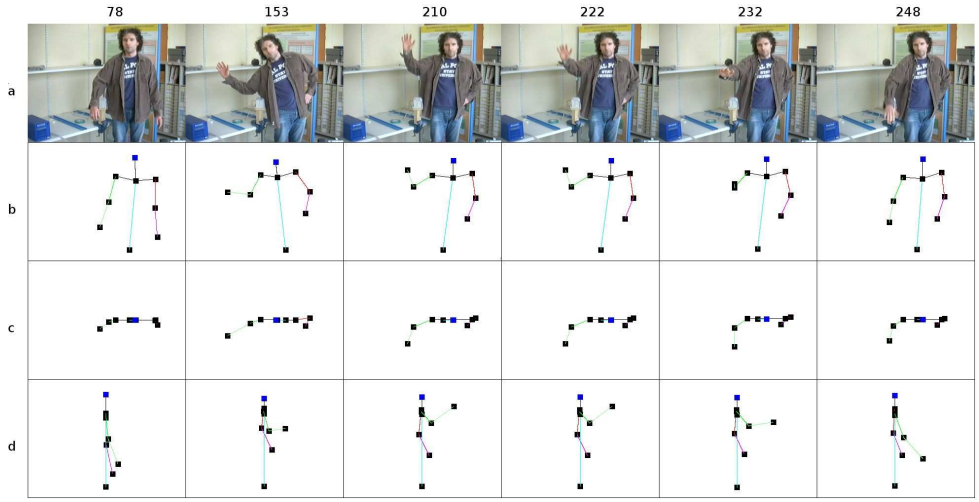


FIG. 4.27: *Scene du test quantitatif pour l'indice d'apprentissage.*

profondeur se produit au niveau du poignet lorsque l'énergie du mouvement n'a pas été utilisée (fig. 4.21a).

En suivi monoculaire du corps, l'estimation de la profondeur entraîne les erreurs les plus importantes (fig. 4.22). De façon générale, l'hypothèse de position d'un membre verra ses chances d'obtenir un meilleur score si sa surface visible à l'image est réduite. Pour maximiser sa vraisemblance, un bras aura donc tendance à se tendre perpendiculairement au plan image et créer une perte de la main (fig. 4.21b). Une pondération des scores par la surface visible des membres doit être mise en place pour corriger ce défaut. Dans certains cas, cette pondération peut avoir un effet trop fort et pousser les membres à coller au plan du torse (fig. 4.20, images 293 et 344).

Ces problèmes sur l'estimation de la profondeur entraîne des erreurs plus fortes sur le poignet

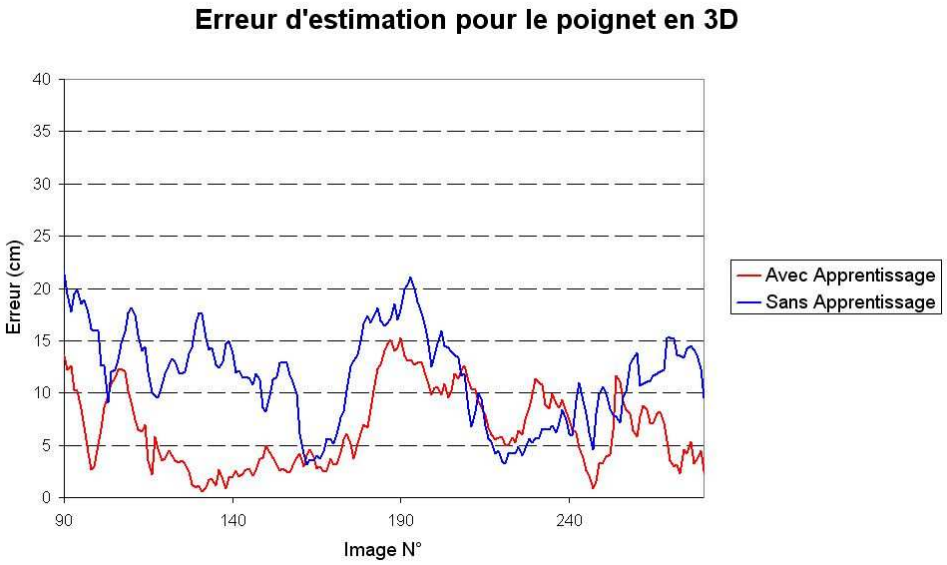


FIG. 4.28: *Erreur d'évaluation sur la position du poignet avec et sans l'indice d'apprentissage.*

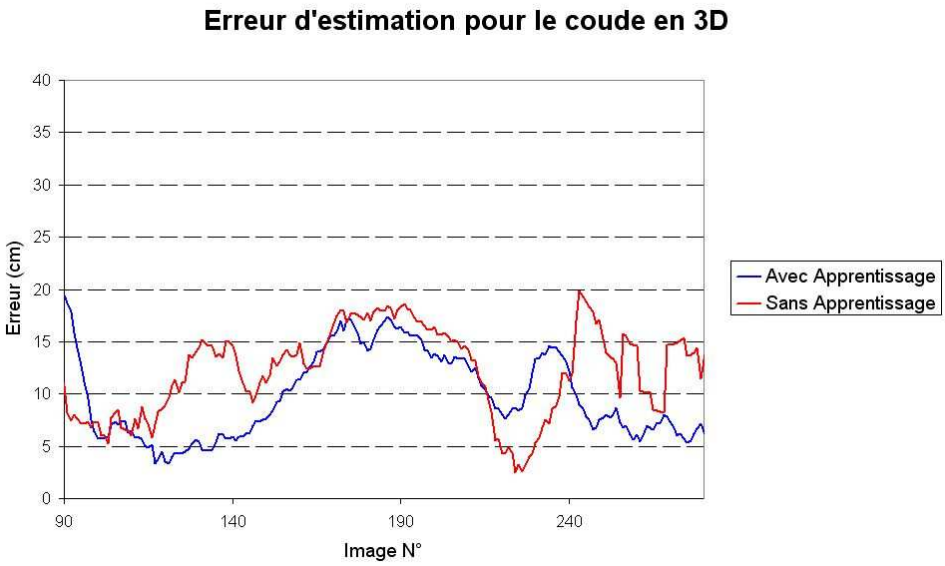


FIG. 4.29: *Erreur d'évaluation sur la position du coude avec et sans l'indice d'apprentissage.*

(fig. 4.23) du fait que ce dernier présente la plus grande amplitude de déplacement sur l'axe z. À la lumière de ce principe, on comprends que le coude (fig. 4.24), à l'origine des plus fortes erreurs en 2D, soit dépassé pas le poignet et que l'épaule continue d'afficher des erreurs faibles (fig. 4.25) vu qu'elle subit un recalage en z.

Exception faite de l'épaule (fig. 4.25), les résultats quantitatifs sont meilleurs lorsque l'énergie du mouvement est utilisée. Le moins bon score sur l'épaule s'explique en partie par le fait que cette articulation, peu mobile, génère peu d'énergie de mouvement et est donc peu influencée par elle.

Globalement, l'erreur estimée tout au long de la scène test (fig : 4.26) ne dépasse pas 25 cm et reste en dessous de 20 cm si on adopte le protocole de mesures globales utilisé dans [SB06, TSDD06] (tab. 4.2). En l'absence de scènes test standards, la comparaison des algorithmes de suivi entre-eux reste une tâche difficile du fait des dissemblances évidentes entre les scènes de test utilisées pour estimer les performances des différentes approches. Cependant, en s'en tenant aux résultats quantitatifs, la précision obtenue est aussi bonne voire meilleure que celle atteinte par des approches à la pointe de l'état de l'art [SB06, TSDD06].

4.4 Suivi intégrant un apprentissage

Nous allons maintenant tester les performances de l'indice basé sur l'apprentissage (§3.4.4). Les tests qui suivent se focalisent sur une difficulté particulière rencontrée fréquemment au cours du suivi. Il s'agit de suivre une personne qui exécute un geste de pointage vers la caméra. Ce type de scène pose souvent des problèmes de décrochage (fig. 4.19) et un apprentissage ciblé sur ce type de geste à été réalisé avec des personnes différament vêtues. La base d'apprentissage contient 1400 exemples répartis sur trois personnes.

La scène utilisée pour le test montre un personnage la main levée, qui tend le bras vers l'avant tout en le ramenant le long du corps (fig. 4.27). Des mesures d'erreurs sur l'estimation des articulation ont été enregistrées durant la scène avec et sans l'indice d'apprentissage. Les

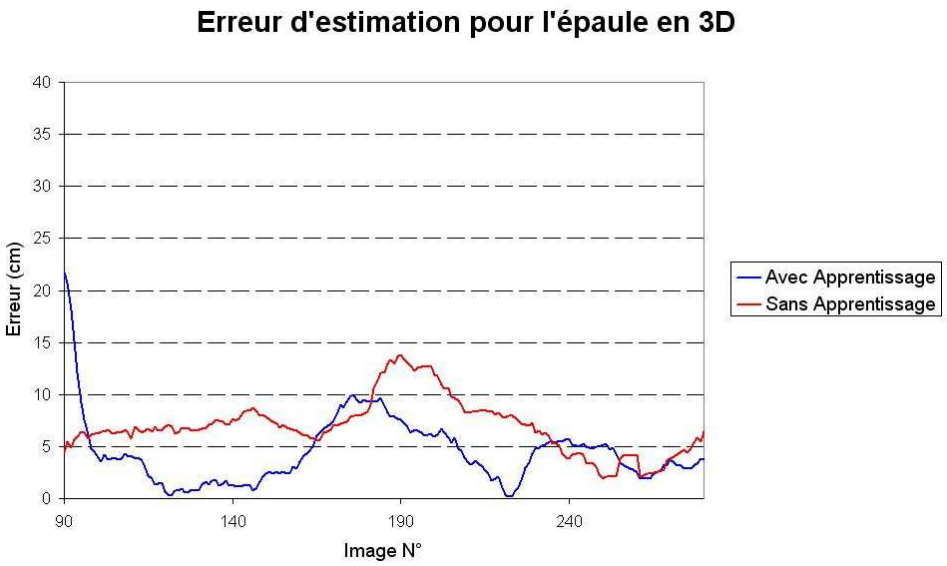


FIG. 4.30: *Erreur d'évaluation sur la position de l'épaule avec et sans l'indice d'apprentissage.*

résultats montrent clairement les améliorations, particulièrement sur le poignet (fig. 4.28) et l'épaule (fig. 4.30) mais le coude n'est pas en reste (fig. 4.29).

La force de cet indice provient du fait que les histogrammes locaux calculés aux extrémités du membre capturent aussi des informations sur les membres adjacents. Ceci permet d'intégrer une partie fort utile du contexte dans lequel se trouve le membre.

La scène testée ici illustre la tendance générale amenant une amélioration des résultats sur des scènes apprises de manière ciblée lorsque l'indice d'apprentissage est utilisé. Cet indice peut donc s'avérer très utile pour régler des problèmes de suivi au cours de scènes qui comportent des poses reconnues pour leur difficulté à être correctement perçues par l'algorithme.

4.5 Conclusion

Ce chapitre est destiné à mettre en lumière les performances de l'algorithme en général, et en particulier la pertinence des indices utilisés. Les tests ont été réalisés auprès de personnes variées dans des environnements différents. Ils se composent de scènes destinées à tester l'algorithme sur des gestes restreints aux deux dimensions du plan de l'image dans un premier temps, puis des tests plus généraux intégrant des gestes dans les trois dimensions de l'espace. Pour ces derniers, une comparaison à été faite avec et sans l'utilisation de l'énergie de mouvement avec et sans l'indice basé sur l'apprentissage.

Concernant le suivi de gestes 2D, l'algorithme a présenté une robustesse tout à fait satisfaisante lors des tests, parvenant à maintenir le suivi sur des images contenant des poses difficiles (bras croisés) ou des occultations.

Dans le cas 3D, le suivi est fidèle sur le plan de l'image mais les profondeurs ne sont pas toujours évaluées avec précision. Cet état de fait est chronique sur les algorithmes de suivi monoculaires qui doivent inférer l'information de profondeur à partir de l'image 2D uniquement. La fusion des indices utilisée permet d'améliorer la robustesse. En particulier, l'indice d'énergie de mouvement, s'il n'améliore pas significativement la précision du suivi, permet d'éviter des fautes

ou des décrochements. Enfin, l'indice basé sur l'apprentissage a montré sa capacité à résoudre des problèmes de perte de suivi récurrentes liées à des poses bien identifiées qui doivent être apprises.

L'algorithme développé, grâce à une vitesse de traitement quasi temps réel (7 images par secondes sur une machine bi-processeur Xeon 3.4 GHz) et des contraintes faibles sur l'environnement (le personnage doit faire approximativement face à la caméra fixe), permet une interaction en directe avec la machine.

L'analyse globale des résultats montre des qualités de robustesse et de précision qui démarquent l'algorithme développé durant cette thèse des approches comparables issues de l'état de l'art récent [TSDD06] [SB06].

Conclusion et perspectives

Au cours de ces trois ans de thèse, nous avons développé un algorithme destiné à assurer le suivi du haut du corps d'une personne à partir de l'acquisition d'une scène par une caméra monoculaire. L'algorithme doit fonctionner dans un milieu naturel dénué d'aides au suivi, c'est à dire des vêtements spéciaux, des gants de couleur, des marqueurs posés sur le corps...

Les principales solutions envisageables consistent en des modèles déterministes, stochastiques ou des approches par apprentissage. Dans les deux premiers cas, la résolution d'un problème de suivi consiste généralement à trouver l'optimum de la fonction qui traduit la viabilité du modèle vis à vis des informations disponibles et particulièrement vis à vis des observations issues de l'image. Cette fonction a la particularité d'être très bruitée et présente de nombreux optimums locaux rendant de ce fait les méthodes déterministes peu performantes vu le risque de blocage sur un optimum local.

Le choix d'une méthode stochastique possède l'avantage de fournir une approximation de la probabilité de validité du modèle plutôt que de renvoyer seulement un optimum aveugle. La connaissance, même par le biais d'une approximation, de cette probabilité est souvent utile pour corriger efficacement d'éventuelles erreurs au cours du processus de suivi et notamment pour lever les ambiguïtés 2D/3D au cours du temps par la propagation des modes au long de la scène. L'approximation des probabilités par échantillonnage générant des espace discrets est bien adapté à l'outil informatique. C'est pourquoi l'algorithme du filtre à particules a été préférée aux estimations à base de noyaux plus lourdes à manipuler.

La seconde difficulté rencontrée provient de l'espace de très grande dimension généré par les problèmes de suivi d'objets articulés. Tenter d'approximer une fonction exprimée dans un espace à plusieurs dizaines de dimensions paraît vain ou alors l'approximation est si locale que le risque de passer outre le mode dominant devient trop grand. Les approches montantes, avec une recherche indépendante pour chacun des membres, permet de réduire considérablement la taille de l'espace exploré. Une étape de sélection des poses cohérente permet d'avantager les membres qui respectent les liaisons d'adjacence et pénaliser les autres. Ce processus est assuré par la propagation des croyances qui s'associe avantageusement au filtre à particules générant un espace discret dans lequel les messages aisément propagés

La nouveauté proposée au cours de ce travail consiste à contraindre un peu plus le modèle pour non seulement avoir des membres adjacents reliés entre-eux, mais aussi des poses qui ne violent pas les limites articulaires du corps humain. Un ensemble de règles possédant une bonne capacité à généraliser ces limites s'intègrent facilement à la propagation des croyances et au calcul des messages propagés à travers les autres membres.

Le fait que l'environnement soit peu contraint oblige l'algorithme à s'adapter à la scène et non le contraire. La seconde contribution de ce travail consiste à donner à l'algorithme plusieurs indices de manière à augmenter ses chances d'en trouver au moins un capable de discriminer le membre recherché. Les indices ne pouvant s'accrocher à l'image faute d'informations sont tout simplement neutralisés. Les indices consistent en une soustraction de fond et une image

de l'énergie des mouvements basées sur les gradients qui servent aussi à vérifier la validité de l'orientation des hypothèses. Il faut aussi ajouter la carte de la teinte chair du visage pour le suivi de la tête et des mains.

La troisième contribution consiste à incorporer un nouvel indice lié à l'apprentissage. L'idée ne consiste pas à accumuler les indices pour dire deux valent mieux qu'un. L'apprentissage permet de fournir une solution à des problèmes complexes à condition d'avoir préalablement appris la solution. Le fait que dans l'expérience, les mêmes pertes de suivi soient souvent dues aux mêmes type de poses permet d'envisager un apprentissage ciblé sur ces poses qui génèrent ces problèmes. De plus, grâce à l'approche montante adoptée il est seulement nécessaire d'apprendre des membres indépendants plutôt que des poses entières. Un base de faible dimension suffit alors pour envisager le règlement ciblé des poses problématiques.

Les essais ont montré l'efficacité et la complémentarité des indices choisis. Les poses issues des scènes test sont généralement bien estimées même dans des conditions difficiles avec des gestes rapides dans les trois dimensions. L'algorithme est capable de fonctionner en quasi temps réel sur une machine standard dans un environnement pouvant présenter des variantions lumineuses avec des fonds quelconques. Les contraintes consistent seulement à garder la caméra fixe pour la soustraction de fond et l'utilisateur doit faire face à la caméra. Seul bémol au tableau, la gestion des auto-occultations qui ne tient pas compte de l'ordre des membres occultés. Pour ce faire, il faudrait rendre le modèle plus complexe en créant des nouvelles liaisons entre les membres occultables. Cela se ferait au prix d'une forte dégradation des performances en vitesse de calcul. En revanche, l'apprentissage peut aussi être ciblé sur des auto-occultations courantes afin de les distinguer correctement.

L'amélioration de l'apprentissage se place au centre des perspectives. La comparaison d'un exemple dans la base d'apprentissage peut être sensiblement accélérée par une technique telle que le LSH [GIM99] et un traitement du type factorisation de matrices non-négatives [AT06a] peut aussi être envisagé pour rejeter les données non discriminantes de la base.

Le suivi monoculaire tente de se placer auprès du grand public grâce à une mise en œuvre facile puisqu'une simple webcam reliée à un ordinateur suffit. Si l'algorithme développé au cours de cette thèse parvient à se démarquer de la plupart des solutions proposée dans l'état de l'art, grâce notamment à ses performances et sa vitesse d'exécution, il n'en reste pas moins que des progrès substantiels restent à accomplir en matière de robustesse pour pouvoir un jour interagir plus naturellement avec la machine.

Annexe A

Approximation d'une fonction de densité de probabilité par échantillonnage

A.1 Échantillonnage par la méthode de Monte Carlo

On considère un tirage de N échantillons indépendants et identiquement distribué de la variable aléatoire x selon la densité de probabilité $p(x)$. Le tirage d'un échantillon indicé par i est noté $x_i \overset{iid}{\sim} p(x)$. On utilise ce tirage pour approximer la valeur de $p(x)$:

$$p(x) \simeq \frac{1}{N} \sum_i \delta(x - x_i), \tag{A.1}$$

cette approximation devient une égalité à la limite pour une infinité de tirages :

$$p(x) = \lim_{N \to \infty} \frac{1}{N} \sum_i \delta(x - x_i). \tag{A.2}$$

La figure A.1 donne le résultats de différentes simulations d'une Gaussienne de moyenne 4 et d'écart type 1. Pour $N = 200\,000$ échantillons, la simulation obtenue est très proche de la Gaussienne originale.

Dans les même conditions que précédemment, l'espérance d'une fonction $f(x)$ peut être obtenue par la relation :

$$E\left[f(x)\right] = \int f(x)p(x)dx = \lim_{N \to \infty} \frac{1}{N} \sum_i f(x_i) \tag{A.3}$$

A.2 Échantillonnage d'importance

Si la valeur de $p(x)$ est calculable numériquement, il n'est pas toujours possible de tirer des échantillons suivant cette loi. L'échantillonnage d'importance consiste à introduire une densité de proposition $q(x)$ selon laquelle on sait tirer les échantillons. L'espérance d'une fonction $f(x)$ est alors donnée par la relation :

$$E\left[f(x)\right] = \int f(x)\frac{p(x)}{q(x)}q(x)dx = \lim_{N \to \infty} \frac{1}{N} \sum_i \frac{p(x_i)}{q(x_i)}q(x_i), \tag{A.4}$$

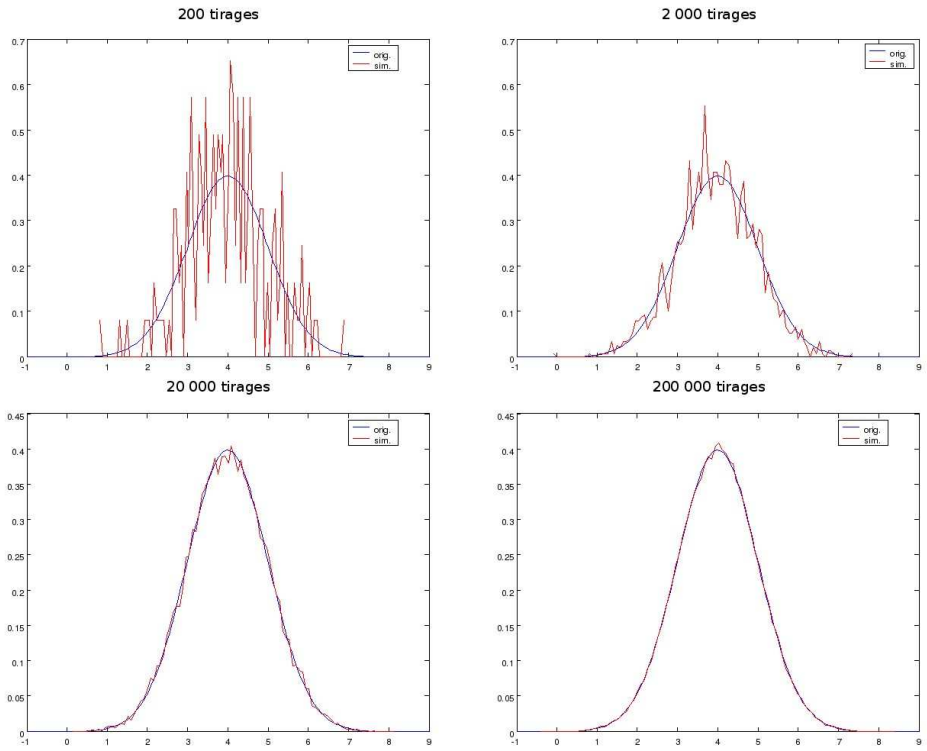


FIG. A.1: *Approximation d'une fonction de densité de probabilité Gaussienne par échantillonnage de Monte Carlo.*

en posant :

$$w_i = \frac{1}{N} \frac{p(x_i)}{q(x_i)}, \tag{A.5}$$

on approxime la densité p(x) par une somme d'échantillons pondérés :

$$p(x) = \lim_{N \to \infty} \sum_i w_i \delta(x - x_i). \tag{A.6}$$

Annexe B

Repères utilisés

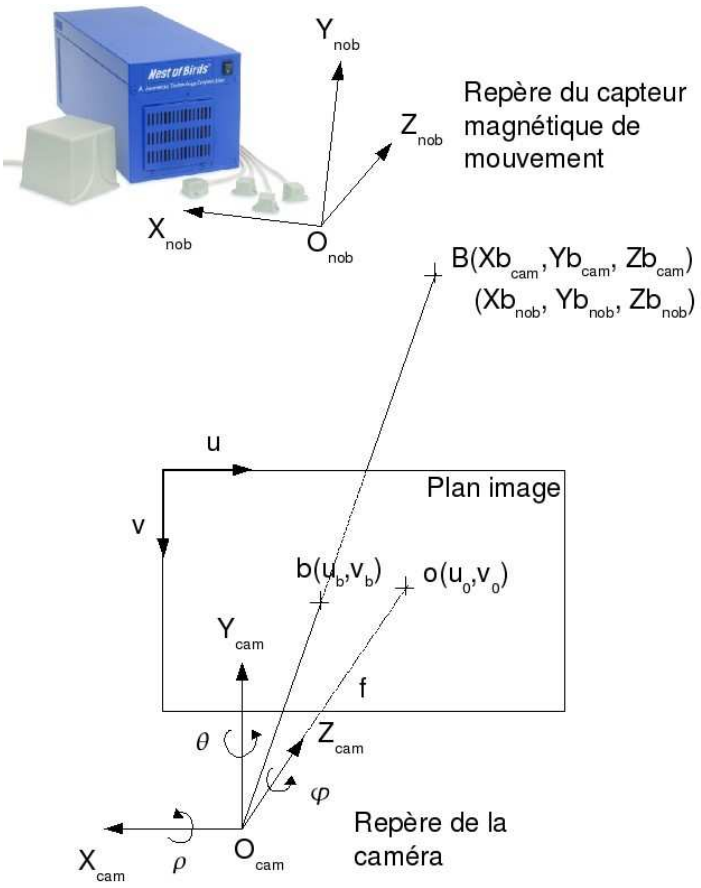


FIG. B.1: *Repères utilisés.*

Le repère de référence est le repère de la caméra $(O_{cam}, X_{cam}, Y_{cam}, Z_{cam})$. Les coordonnées de la capture magnétique de mouvement sont données par rapport au repère $(O_{nob}, X_{nob}, Y_{nob}, Z_{nob})$. Un point B de l'espace possède les coordonnées $B(Xb_{cam}, Yb_{cam}, Zb_{cam})$ dans l'espace et $b(u_b, v_b)$ dans le plan de l'image. Le point $O(u_0, v_0)$ est le point principal de la caméra et f la distance focale.

Index

Bibliographie

[AT06a] Ankur Agarwal and Bill Triggs. A local basis representation for estimating human pose from cluttered images. In *ACCV (1)*, pages 50–59, 2006.

[AT06b] Ankur Agarwal and Bill Triggs. Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 28(1), January 2006.

[BAHH92] James R. Bergen, P. Anandan, Keith J. Hanna, and Rajesh Hingorani. Hierarchical model-based motion estimation. In *ECCV '92 : Proceedings of the Second European Conference on Computer Vision*, pages 237–252, London, UK, 1992. Springer-Verlag.

[BCMC06] Olivier Bernier and Pascal Cheung-Mon-Chang. Real-time 3d articulated pose tracking using particle filtering and belief propagation on factor graphs. In *British Machine Vision Conference*, volume 01, pages 005–008, 2006.

[BD96] A. Bobick and J. Davis. An appearance-based representation of action. In *ICPR '96 : Proceedings of the 1996 International Conference on Pattern Recognition (ICPR '96) Volume I*, page 307, Washington, DC, USA, 1996. IEEE Computer Society.

[BHB00] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, pages 2690–2696, 2000.

[BI96] Andrew Blake and Michael Isard. The condensation algorithm - conditional density propagation and applications to visual tracking. In *NIPS*, pages 361–367, 1996.

[BM92] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2) :239–256, 1992.

[BM98] Christoph Bregler and Jitendra Malik. Tracking people with twists and exponential maps. In *CVPR*, pages 8–15, 1998.

[BP66] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 41, 1966.

[Can86] John Canny. A computational approach to edge detection. PAMI-8, No. 6 :679–698, 1986.

[DD02] David Demirdjian and Trevor Darrell. 3-D articulated pose tracking for untethered diectic reference. In *ICMI*, pages 267–272. IEEE Computer Society, 2002.

[DGA00] A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10 :197–208, 2000.

[DKD03] David Demirdjian, T. Ko, and Trevor Darrell. Constraining human body tracking. In *ICCV '03 : Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1071, Washington, DC, USA, 2003. IEEE Computer Society.

[DLF05] M. Dimitrijevic, V. Lepetit, and P. Fua. Human body pose recognition using spatio-temporal templates. In *ICCV*, pages 000–000, 2005.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, June 2005.

[DTS+05] David Demirdjian, Leonid Taycher, Gregory Shakhnarovich, Kristen Grauman, and Trevor Darrell. Avoiding the "streetlight effect" : Tracking by exploring likelihood modes. In *ICCV*, pages 357–364, 2005.

[EHD00] Ahmed M. Elgammal, David Harwood, and Larry S. Davis. Non-parametric model for background subtraction. In *ECCV '00 : Proceedings of the 6th European Conference on Computer Vision-Part II*, pages 751–767, London, UK, 2000. Springer-Verlag.

[EL04] Ahmed M. Elgammal and Chan-Su Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *CVPR (2)*, pages 681–688, 2004.

[FA91] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(9) :891–906, 1991.

[FBVC01] Raphaël Féraud, Olivier Bernier, Jean Emmanuel Viallet, and Michel Collobert. A fast and accurate face detector based on neural networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 23(1) :42–53, 2001.

[FCF96] Graham D. Finlayson, Subho S. Chatterjee, and Brian V. Funt. Color angular indexing. In *ECCV (2)*, pages 16–27, 1996.

[FE73] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. Computer*, 22(1) :67–92, January 1973.

[FH05] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61(1) :55–79, 2005.

[Gav99] D. M. Gavrila. The visual analysis of human movement : A survey. *Comput. Vis. Image Underst.*, 73(1) :82–98, jan 1999.

[GIM99] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In *VLDB '99 : Proceedings of the 25th International Conference on Very Large Data Bases*, pages 518–529, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[GMHP04] Keith Grochow, Steven L. Martin, Aaron Hertzmann, and Zoran Popović. Style-based inverse kinematics. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 23(3) :522–531, 2004.

[GS96] T. Gevers and A. Smeulders. A comparative study of several color models for color image invariant retreival. In *Proc. 1st Int. Workshop on Image Databases & Multimedia Search, Amsterdam, Netherlands.*, pages 17–24, 1996.

[GS04] Jiang Gao and Jianbo Shi. Multiple frame motion inference using belief propagation. In *FGR*, pages 875–882, 2004.

[Hoy04] Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, 5 :1457–1469, 2004.

[HS80] Berthold K.P. Horn and Brian G. Schunck. Determining optical flow. Technical report, Cambridge, MA, USA, 1980.

[IB96] M. Isard and A. Blake. Visual tracking by stochastic propagation of conditional density. In *Proc. Fourth European Conf. Computer Vision*, pages 343–356, 1996.

[IF01] Sergey Ioffe and David A. Forsyth. Human tracking with mixtures of trees. In *ICCV*, pages 690–695, 2001.

[Isa03] Michael Isard. Pampas : Real-valued graphical models for computer vision. In *CVPR (1)*, pages 613–620, 2003.

[JBY96] Shanon X. Ju, Michael J. Black, and Yaser Yacoob. Cardboard people : A parameterized model of articulated image motion. In *FG '96 : Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*, page 38, Washington, DC, USA, 1996. IEEE Computer Society.

[JFEM03] Allan D. Jepson, David J. Fleet, and Thomas F. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(10) :1296–1311, 2003.

[KFL01] Frank R. Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2) :498–519, 2001.

[Law03] Neil D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *NIPS*, 2003.

[LC04] Mun Wai Lee and Isaac Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *CVPR (2)*, pages 334–341, 2004.

[LH04] Xiangyang Lan and Daniel P. Huttenlocher. A unified spatio-temporal articulated model for tracking. *cvpr*, 01 :722–729, 2004.

[LHGT04] Liyuan Li, Weimin Huang, Irene Y. H. Gu, and Qi Tian. Statistical modeling of complex backgrounds for foreground object detection. In *IEEE Transactions on Image Processing*, volume 13, pages 1459–1472, 2004.

[Li95] S. Z. Li. *Markov Random Field Modeling in Computer Vision*. Springer-Verlag, 1995.

[LMP01] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proc. of 18th Int. Conf. on Machine Learning*, pages 282–289, 2001.

[LV04] Christine Leignel and Jean Emmanuel Viallet. A blackboard architecture for the detection and tracking of a person. In *RFIA*, 2004.

[MD01] Michael Mason and Zoran Duric. Using histograms to detect and track objects in color video. In *AIPR*, pages 154–162, 2001.

[MG01] Thomas B. Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Comput. Vis. Image Underst.*, 81(3) :231–268, mar 2001.

[MHK06] Thomas B. Moeslund, Adrian Hilton, and Volker Kr&#252 ;ger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 104(2) :90–126, 2006.

[MM02] Greg Mori and Jitendra Malik. Estimating human body configurations using shape context matching. In *ECCV (3)*, pages 666–680, 2002.

[MREM04] Greg Mori, Xiaofeng Ren, Alexei A. Efros, and Jitendra Malik. Recovering human body configurations : Combining segmentation and recognition. In *CVPR (2)*, pages 326–333, 2004.

[MSZ94] Richard M. Murray, S. Shankar Sastry, and Li Zexiang. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, Inc., Boca Raton, FL, USA, 1994.

[Muy01] Eadweard Muybridge. *The Human Figure in Motion*. Dover Publications, 1901.

[NB06] Philippe Noriega and Olivier Bernier. Real time illumination invariant background subtraction using local kernel histograms. In *British Machine Vision Conference*, volume 3, pages 979–988, 2006.

[NB07a] Philippe Noriega and Olivier Bernier. Multicues 3d monocular upper body tracking using constrained belief propagation. In *To appear in BMVC*, volume 00, pages 000–000, 2007.

[NB07b] Philippe Noriega and Olivier Bernier. Multicues 2d articulated pose tracking using particle filtering and belief propagation on factor graphs. In *To appear in ICIP*, volume 00, pages 000–000, 2007.

[NBB06] Philippe Noriega, Benedicte Bascle, and Olivier Bernier. Local kernel color histograms for background subtraction. In Alpesh Ranchordas, Helder Araújo, and Bruno Encarnação, editors, *VISAPP*, pages 213–219. INSTICC - Institute for Systems and Technologies of Information, Control and Communication, 2006.

[NTTC05] Ramanan Navaratnam, Arasanathan Thayananthan, Philip H. S. Torr, and Roberto Cipolla. Hierarchical part-based human body pose estimation. In *British Machine Vision Conference*, volume 00, pages 000–000, 2005.

[Pea88] Judea Pearl. *Probabilistic reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[PF01] Ralf Plänkers and Pascal Fua. Articulated soft objects for video-based body modeling. In *ICCV*, pages 394–401, 2001.

[PF03] Ralf Plänkers and Pascal Fua. Articulated soft objects for multiview shape and motion capture. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9) :1182–1187, 2003.

[RBM05] Xiaofeng Ren, Alexander C. Berg, and Jitendra Malik. Recovering human body configurations using pairwise constraints between parts. In *Proc. 10th Int'l. Conf. Computer Vision*, volume 1, pages 824–831, 2005.

[RMR04] Timothy J. Roberts, Stephen J. McKenna, and Ian W. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. In *ECCV (4)*, pages 291–303, 2004.

[RP66] Azriel Rosenfeld and John L. Pfaltz. Sequential operations in digital picture processing. *J. ACM*, 13(4) :471–494, 1966.

[RS00] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 :2323–2326, December 2000.

[SB06] Leonid Sigal and Michael J. Black. Measure locally, reason globally : Occlusion-sensitive articulated pose estimation. In *CVPR (1)*, pages 421–428, 2006.

[SBR+04] Leonid Sigal, Sidharth Bhatia, Stefan Roth, Michael J. Black, and Michael Isard. Tracking loose-limbed people. In *CVPR (1)*, pages 421–428, 2004.

[SC92] Jun Shen and Serge Castan. An optimal linear operator for step edge detection. *CVGIP : Graph. Models Image Process.*, 54(2) :112–133, 1992.

[Set99] J. A. Sethian. Fast marching methods. *SIAM Rev.*, 41(2) :199–235, 1999.

[SS05] Y. Sheikh and M. Shah. Bayesian modeling of dynamic scenes for object detection. *PAMI*, 27(11) :1778–1792, November 2005.

[ST01] Cristian Sminchisescu and Bill Triggs. Covariance scaled sampling for monocular 3d body tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, USA*, volume 1, pages 447–454. IEEE Computer Society Press, Dec 2001.

[ST02] Cristian Sminchisescu and Alexandru Telea. Human pose estimation from silhouettes - a consistent approach using distance level sets. In *WSCG*, pages 413–420, 2002.

[ST03] Cristian Sminchisescu and Bill Triggs. Kinematic jump processes for monocular 3d human tracking. In *CVPR (1)*, pages 69–76, 2003.

[SVD03] Gregory Shakhnarovich, Paul Viola, and Trevor Darrell. Fast pose estimation with parameter-sensitive hashing. In *ICCV '03 : Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 750, Washington, DC, USA, 2003. IEEE Computer Society.

[Tay00] Camillo J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Comput. Vis. Image Underst.*, 80(3) :349–363, 2000.

[TB99] Michael E. Tipping and Christopher M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2) :443–482, 1999.

[THB03] Lorenzo Torresani, Aaron Hertzmann, and Christoph Bregler. Learning non-rigid 3d shape from 2d motion. In *NIPS*, 2003.

[TK92] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams : a factorization method parts 2,8,10 full report on the orthographic case. Technical report, Pittsburgh, PA, USA, 1992.

[TKBM99] Kentaro Toyama, John Krumm, Barry Brumitt, and Brian Meyers. Wallflower : Principles and practice of background maintenance. In *ICCV*, pages 255–261, 1999.

[TLS05] Tai-Peng Tian, Rui Li, and Stan Sclaroff. Articulated pose estimation in a learned smooth space of feasible solutions. In *CVPR '05 : Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, page 50, Washington, DC, USA, 2005. IEEE Computer Society.

[TSDD06] Leonid Taycher, Gregory Shakhnarovich, David Demirdjian, and Trevor Darrell. Conditional random people : Tracking humans with crfs and grid filters. In *CVPR '06 : Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 222–229, Washington, DC, USA, 2006. IEEE Computer Society.

[TT03] M. Tkalcic and J. F. Tasic. Colour spaces : perceptual, historical and applicational background. In *EUROCON*, pages 304–308, 2003.

[UF04] Raquel Urtasun and Pascal Fua. 3d Human Body Tracking using Deterministic Temporal Motion Models. Technical report, 2004.

[UFF06] Raquel Urtasun, David J. Fleet, and Pascal Fua. 3d people tracking with gaussian process dynamical models. In *CVPR (1)*, pages 238–245, 2006.

[UFHF05] Raquel Urtasun, David J. Fleet, Aaron Hertzmann, and Pascal Fua. Priors for people tracking from small training sets. In *ICCV*, pages 403–410, 2005.

[WADP97] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. Pfinder : Real-time tracking of the human body. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7) :780–785, 1997.

[Wei00] Yair Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12(1) :1–41, 2000.

[WHY03] Y. Wu, G. Hua, and T. Yu. Tracking articulated body by dynamic markov network. In *ICCV*, pages 1094–1101, 2003.

[XRB04] Binglong Xie, Visvanathan Ramesh, and Terrance E. Boult. Sudden illumination change detection using order consistency. *Image Vision Comput.*, 22(2) :117–125, 2004.

[YFW05] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. on Information Theory*, 51(7) :2282–2312, July 2005.