Suivi 2D du corps articulé avec gestion des auto-occultations

2D Body tracking with self-occlusions handling

E. Para[1] O. Bernier[1] C. Achard[2]

[1]France Telecom, Orange Labs, Recherche et Développement
[2]Université Pierre et Marie Curie

Technopole Anticipa, 2 Avenue Pierre Marzin, 22307 Lannion, France
{eric.para, olivier.bernier}@orange-ftgroup.com

Résumé

Depuis quelques années, de nombreuses approches concernant le suivi du corps humain articulé ont été proposées. Les techniques généralement utilisées nécessitent cependant beaucoup de temps de calcul, interdisant leur utilisation pour les interfaces homme-machine. Nous proposons dans cet article un suivi en temps réel de chaque membre supérieur du corps d'une personne avec gestion de leurs occultations respectives. Pour réaliser ce suivi et initialiser automatiquement la cible, nous utilisons un modèle articulé permettant de prendre en compte la déformabilité du corps humain. Après la recherche indépendante des meilleures positions possibles de chacun des membres, un algorithme de programmation dynamique est utilisé pour obtenir la meilleure configuration en tenant compte des liens entre les différents membres. La gestion des auto-occultations entre les membres d'une même personne est au cœur de l'algorithme de suivi, l'objectif étant d'utiliser ultérieurement ces techniques pour suivre plusieurs personnes en gérant également les occultations entre ces personnes.

Mots Clef

Suivi, modèle articulé du corps, gestion d'occultations, programmation dynamique, calcul temps-réel, vision par ordinateur.

Abstract

Recently many methods for human articulated body tracking were proposed in the literature. These techniques are often computationally intensive and cannot be used for Human-Computer Interface. We propose in this article a real-time algorithm for upper body tracking with occultation handling. The tracking is based on an articulated body model, also used to automatically initialize the target. After an independent search of the most likely positions of each limb, a dynamic programming algorithm is used to find the best configuration according to the links between limbs. The self-occlusions between the limbs are directly taken into account by the tracking algorithm and results show the interest of the proposed approach.

Keywords

Tracking, humanoid articulated model, occlusions handling, dynamic programming, real-time processing, computer vision.

1 Introduction

Le suivi de personnes soulève de nombreuses difficultés principalement dues à la déformabilité du corps humain. De nombreuses solutions sont ainsi proposées dans la littérature [1]. On trouve aussi bien des approches de suivi global [2], souvent limitées par une modélisation trop rigide, que des approches de suivi par morceaux [3,4] qui permettent une amélioration de robustesse par la souplesse apportée. Afin de rendre le suivi d'une personne plus robuste, des modèles géométriques du corps humain de plus en plus complexes [5,6,7] ont été proposés. Cependant, les temps de calculs souvent importants interdisent le suivi en temps réel et les applications qui en découlent. Un autre problème concerne les occultations entre les membres de la personne suivie. Certains auteurs proposent des extensions de leurs algorithmes afin d'essayer de les prendre en compte [7] mais au détriment d'un coût de calcul élevé. Nous proposons dans cet article, un suivi du corps articulé en temps réel avec une gestion des occultations entre les objets suivis (Figure 1).

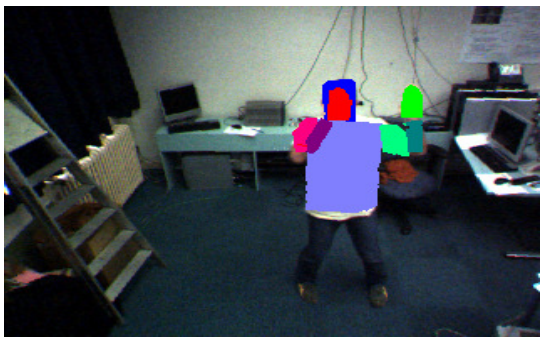


Figure 1 : exemple de résultat du système. Même en présence d'occultations, chaque membre de la personne continu à être suivi.

Dans le chapitre qui suit, nous exposerons le principe de recherche par morceaux et la stratégie adoptée pour garantir la cohérence géométrique. Le troisième chapitre sera consacré à l'initialisation automatique du modèle de la personne à suivre. Des primitives génériques seront recherchées et le modèle d'apparence de la personne trouvée sera ainsi initialisé. Dans le quatrième chapitre nous détaillerons la stratégie de suivi de cette personne par la recherche des différentes parties composant son modèle en introduisant la gestion de leurs occultations. Enfin, la dernière partie sera consacrée aux résultats obtenus avant de conclure sur les limites et les perspectives de nos travaux.

2 Principe général

Dans ce chapitre, nous nous intéressons à la description de l'architecture générale utilisée dans la suite de cet article. Le principe de la recherche par morceaux retenue implique l'utilisation d'un modèle géométrique afin de garantir une cohérence de l'ensemble de la structure. De nombreux modèles du corps humain peuvent être utilisés, du plus simple au plus complexe. L'approche *"cardboard people"* de Ju et al. [5] est basée sur des modèles 2D articulés tandis que Demirdjian et al. [6] utilisent une reconstruction volumique plus précise.

En pratique, un modèle 3D est assez lourd à mettre en œuvre et à initialiser et pose le problème de minima locaux cinématiques. Il est donc intéressant d'étudier le cas 2D où la correspondance modèle-image est beaucoup plus directe. Pour cette raison, le modèle que nous utilisons prend la forme d'un ensemble d'objets 2D composant la silhouette humaine (main, avant-bras, bras, tronc et tête). Ces objets sont connectés par des articulations élastiques comme le proposent Sigal et al. [7] (Figure 2). Le haut du corps humain est alors représenté par une structure sous forme d'arbre avec le torse comme nœud central.

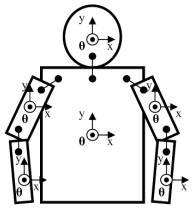


Figure 2 : modèle utilisé avec des objets 2D connectés.

L'état de chaque objet est limité à trois degrés de liberté dans le plan 2D de l'image caméra : les coordonnées de son centre (xi,yi) et son angle de rotation θi autour de l'axe perpendiculaire au plan de l'image. Le facteur de zoom global est supposé connu et les rotations autour des autres axes ne sont pas gérées. Ceci impose à l'utilisateur des mouvements dans le plan 2D de la camera pour assurer son suivi. Ce modèle algorithmique limité peut néanmoins être étendu à la troisième dimension pour prendre en compte tous les mouvements humains.

Plusieurs méthodes sont alors possibles pour trouver la meilleure configuration du modèle articulé. Ramanan et al. [4] optent pour la déformation d'un modèle géométrique afin de trouver la meilleure correspondance dans l'image. Mais le test de toutes les possibilités nécessite de lourds calculs. Pour les minimiser, Fischler et al. [8] ont introduit les *"Pictorial Structures"* qui ont été adapté au problème d'initialisation d'un corps humain par Felzenszwalb et al. [3]. Le temps de calcul nécessaire à une image est encore de plus d'une minute. Pour obtenir un traitement temps-réel, nous proposons une nouvelle procédure en trois étapes :

- La première étape consiste à calculer un score de correspondance $c_i(e_i)$ pour toutes les positions potentielles e_i de chaque objet i. Seuls les meilleures positions et leurs scores correspondants sont retenues pour la suite des calculs.
- Dans la deuxième étape, un score d'interaction $\lambda_{\{j,k\}}(e_j,e_k)$ est calculé entre chaque couple d'hypothèses retenues (e_j,e_k) de paires d'objets connectés $\{j,k\}$.
- La dernière étape permet enfin de trouver la configuration d'hypothèses C^* maximisant les scores de correspondance et d'interaction pour chaque objet :

$$C^* = \arg\max\left(\sum_{i=1}^{n} c_i(e_i) + \sum_{\{j,k\}\in L} \lambda_{\{j,k\}}(e_j,e_k)\right) \qquad (1)$$

où L représente l'ensemble des paires d'objets connectés du modèle articulé.

3 Initialisation automatique

Dans ce chapitre nous nous intéressons à l'initialisation automatique du modèle, étape indispensable à la mise en œuvre d'une interface Homme-Machine. Nous détaillons ainsi plusieurs méthodes qui seront réutilisées par la suite pour le suivi, comme le calcul des scores de correspondance des objets et de leurs liens, ou encore le choix de la meilleure configuration du modèle articulé.

3.1 Calcul des scores de correspondance

La première étape consiste à tester plusieurs hypothèses de position pour chaque objet afin de leur attribuer un score de correspondance $c_i(e_i)$. Certains auteurs utilisent uniquement un modèle générique comme la teinte chair permettant de trouver les mains et la tête mais restant fortement dépendant de l'éclairage. Nous préférons utiliser donc la fusion d'informations recueillies par une soustraction de fond [1,3], une détection des contours [1,4] et d'une détection de visage [1,2] pour initialiser le modèle. Une fois la position initiale de chaque membre de la personne déterminée, un modèle de couleurs spécifique à chacun d'eux est appris. Dans la suite nous utiliserons ce modèle pour suivre plus spécifiquement la personne (chapitre 4).

(a) *(b)* *(c)*

Figure 3 : image observée *(a)*, soustraction du fond *(b)* et transformée de distance *(c)*.

L'utilisation de caméras fixes permet d'obtenir à faible coût une détection d'objet par différence d'images entre l'image observée et une image de fond. Nous utilisons un seuil automatique pour suivre les changements globaux de luminosité et permettre l'amélioration des détections (Figure 3*(b)*). L'initialisation proposée par Felzenszwalb et al. [11] qui travaillent directement sur cette image binaire est couteuse en temps de calcul. D'autres auteurs [1,4] se limitent alors aux contours des silhouettes extraites. Ces techniques nécessitent une bonne qualité de détection qui ne peut pas toujours être obtenue.

Pour lisser les erreurs de détection, nous procédons à un calcul de distance sur l'image binaire. Nous utilisons la Transformée en Distance de Chanfrein (TDC), introduite en 1977 par Barrow et al. [14], qui est une simplification de la distance euclidienne en valeurs entières. Ne nécessitant que deux balayages de l'image, elle offre comme avantage une grande rapidité de calcul souvent mise à profit dans le cadre d'applications temps-réel [15].Pour déterminer la position optimale d'un membre, nous projetons son modèle dans l'image des distances à la manière de Gavrila [15]. Le score de correspondance $c_i(e_i)$ d'une hypothèse de position e_i d'un membre i, s'obtient alors en sommant les pixels de l'image des distances $d_{(x,y)}$ correspondant à la projection de son modèle :

$$c_i(e_i) = 1 - \frac{1}{nb} \sum_{\{x,y\} \in \ell} d_{(x,y)} \qquad (2)$$

où nb est le nombre de pixels décrivant le modèle du membre ℓ constitué des arrêtes d'un rectangle. Le choix du modèle de membre est déterminé par la visibilité des arrêtes sur l'image de distance. Pour un bras par exemple, seule l'arrête supérieure est prise en compte afin d'éviter les problèmes liés au port de vêtements amples. La figure 4*(a)* illustre les trois arrêtes du rectangle utilisées pour la modélisation d'un avant-bras.

Pour tenir compte de la forme particulière de la tête, les orientations des gradients sont utilisées. Un modèle elliptique g est alors recherché et le score est calculé par soustraction angulaire :

$$c_{tête}(e_{tête}) = \sum \left(\left| \Phi_{I(x,y)} - \alpha_{g(x,y)} \right| \right) \qquad (3)$$

où $\Phi_{I(x,y)}$ correspond à l'orientation du gradient du pixel (x,y) et $\alpha_{g(x,y)}$ à celle du gradient du modèle elliptique g.

Cette méthode est plus robuste qu'une simple détection de contours pour un visage. Le poids relatif des scores obtenus à partir de différents indices peut s'avérer être un problème. En pratique, nous observons un comportement similaire entre l'utilisation de la soustraction du fond et de l'orientation des gradients. On notera qu'une pose de la personne, face à la caméra et bras visibles le long du corps, est nécessaire pour initialiser notre système.

3.2 Retenue des meilleures hypothèses

Le score $c_i(e_i)$ est calculé pour chaque hypothèse de position (x_i, y_i, θ_i) d'un objet i. La complexité algorithmique est directement proportionnelle à ce nombre de candidats potentiels. Felzenszwalb et al. [3] calculent pour chaque objet toutes les solutions possibles avec une précision angulaire en θ_i de 11° et un pas de 4 pixels pour les paramètres x_i and y_i. Pour une image de 320x240, le temps de calcul annoncé est de plus d'une minute pour tester les 1,5 millions de configurations correspondantes. Pour limiter le nombre d'hypothèses tout en garantissant une meilleure précision, nous utilisons un détecteur de visage sur la partie supérieure de l'image binaire obtenue par soustraction du fond. Cela permet de limiter l'espace de recherche de chaque membre dans une zone définie à partir de la position du visage et des poses d'initialisation possibles. L'orientation de chaque membre est ainsi limité avec par exemple $\theta_{bras} \in [10°;90°]$ par rapport à la verticale. Une recherche pyramidale en multi-résolution limite les temps de calcul en conservant les zones intéressantes aux résolutions suivantes.

Enfin, pour limiter le nombre d'hypothèses finales, un sous-échantillonnage est effectué et seules sont conservées les meilleures hypothèses pour la suite des calculs (figure 4*(b),(c)*). Une hypothèse e_j est retenue si son score $c_i(e_j)$ est supérieur au score maximum de toutes les hypothèses e_i de l'objet i, multiplié par une variable de tolérance τ :

$$c_i(e_j) > \underset{\tau \in [0;1]}{\tau} * \arg\max\left(c_i(e_i)\right) \qquad (4)$$

Nous choisissons τ*=0,90* avec un maximum de 20 hypothèses conservées par orientation, ce qui correspondant au meilleur compromis précision-performance avec le matériel dont nous disposons : un processeur simple cœur de type *Pentium 4* cadencé à 3,2 GHz.

(a) (b) (c)

Figure 4 : *(a)* modèle de l'avant-bras à 3 arrêtes, *(b)* hypothèses retenues après calcul sur l'image des distances, *(c)* positions possibles de l'avant-bras gauche.

3.3 Calculs des scores d'interaction

Comme introduit dans l'équation *(1)*, deux scores entrent en jeu dans le choix de la meilleure configuration : celui de la correspondance à l'image, et celui d'interaction lié à l'articulation entre deux objets connectés. Différentes solutions existent dans la littérature pour calculer ce score d'interaction $\lambda_{\{j,k\}}(e_j,e_k)$ entre une hypothèse e_j de position d'un objet j et une hypothèse e_k de position d'un objet k (j et k étant connectés). Demirdjian et al. [6] utilisent un modèle rigide qui imite les articulations humaines avec des rotules centrées en un point. Sigal et al. [7] proposent leur *"Loose-Limbed Model"* avec des liens élastiques pour plus de souplesse. Ces liens absorbent ainsi les erreurs de positions sans les propager à toute la structure.

Nous choisissons une méthode similaire et modélisons les scores de liens avec une fonction Gaussienne permettant de traduire cette élasticité :

$$\lambda_{\{j,k\}}(e_j,e_k) = e^{-\frac{d^2(e_j,e_k)}{2\sigma^2}} \qquad (5)$$

avec $d(e_j,e_k)$ la distance euclidienne entre deux points d'articulation pour les deux hypothèses de pose e_j et e_k (Figure 5). Un score articulaire $\lambda_{\{j,k\}}(e_j,e_k)$ de 1 indique un positionnement idéal des deux membres voisins. En pratique, la constante de raideur σ qui règle la tolérance des liens est expérimentalement fixée à 1. Ce score est calculée pour chaque couple d'hypothèses retenues (voir §3.3).

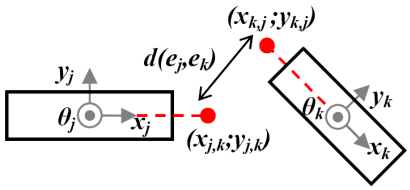


Figure 5 : illustration des hypothèses ej et ek de deux objets j et k avec leur centre articulaire respectif.

3.4 Choix de la meilleure configuration

Comme discutée au chapitre 2.2, la dernière étape consiste à trouver la meilleure configuration à partir des scores calculés précédemment comme le décrit l'équation *(1)*. Le calcul exhaustif nécessite h^n opérations avec h hypothèses retenues pour chacun des n objets. Afin de réduire ce temps de calcul, Bernier et al. [9] utilisent la propagation des croyances qui est similaire à une méthode de relaxation. Nous choisissons ici de mettre à profit la structure d'arbre utilisée pour implémenter une approche de programmation dynamique [10]. Celle-ci permet le calcul de la meilleure correspondance entre le modèle et l'image avec une complexité polynomiale réduite à $O(h^2n)$. Le modèle d'arbre est découpé en autant d'itérations que de nombre de liens, i.e. $n-1$ avec notre modèle non bouclé (Figure 6).

L'algorithme propage les scores depuis les extrémités du graphe vers le torse. A chaque étape, le score de l'objet suivant est remplacé par la meilleure somme du score de l'objet précédent et du score d'articulation entre eux. Par exemple, si l'objet *i-1* est une extrémité, le score de l'objet *i* est calculé à partir de l'équation *(6)*. L'objet *i* devient alors une extrémité pour l'objet *i+1* et ainsi de suite. A chaque étape, le score final d'un objet contient ainsi l'accumulation des scores de la meilleure branche de l'arbre qui lui précède. Les chemins explorés doivent être mémorisés afin de pouvoir retrouver la meilleure configuration globale.

Pour chaque hypothèse e_j de l'objet i-1et e_k de l'objet i :

$$\hat{c}_i(e_k) = \arg\max_j \left(\hat{c}_{i-1}(e_j) + \lambda_{\{i-1,i\}}(e_j,e_k)\right) + c_i(e_k) \qquad (6)$$

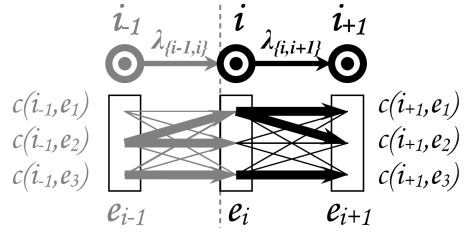


Figure 6 : représentation de deux étapes successives de l'algorithme de programmation dynamique avec les meilleures solutions en gras.

L'approche de recherche par morceaux avec un nombre limité d'hypothèses sélectionnées permet un calcul rapide tout en bénéficiant de bons résultats. Felzenszwalb et al. [3] ont réussit à limiter la complexité de leur algorithme à $O(hn)$ mais en utilisant une transformée de distance particulière sous la forme de Mahalanobis. Ils conservent ainsi beaucoup d'hypothèses mais annoncent des temps de calcul de plus d'une minute par image. Notre solution teste en 1/25e de seconde plus de 100 poses complètes d'initialisation pour un facteur d'échelle donné (avec un *P4 à* 3,2GHz). Cette rapidité permet alors la détermination optimale de la taille du modèle en fonction de l'éloignement du sujet à la caméra (Figure 7) ou d'améliorer la précision de l'initialisation. Les N meilleures configurations peuvent en outre être très facilement obtenues en utilisant la même approche avec une complexité en $O(Nh^2n)$ en retenant les N meilleurs chemins à chaque itération.

Figure 7 : exemple de résultats d'initialisation automatique dans des conditions variables : éloignement à la caméra et postures différentes.

4 Suivi

Nous présentons, dans ce chapitre, la partie de l'algorithme relative au suivi. Certains auteurs [3,4] décident de réinitialiser la personne à chaque nouvelle image. Cette méthode, intensive en temps de calcul, ne peut se faire qu'en limitant l'espace de recherche des hypothèses, espace à priori trop important. La méthode présentée dans cet article utilise un algorithme de suivi pour réduire la dimension de cet espace de recherche. Le coût de calcul est ainsi réduit et permet une meilleure précision de positionnement des membres. A chaque nouvelle image, la zone de recherche d'un objet est limitée au voisinage de sa position précédente.

La méthode précédemment utilisée pour l'initialisation automatique est de nouveau employée. La meilleure configuration globale est trouvée à partir du calcul des scores de correspondance de chaque objet et de leurs articulations respectives. Les modèles génériques sont remplacés par un modèle d'apparence couleur, propre à chaque objet appris à partir de la phase d'initialisation automatique. Cette méthode de suivi plus précise et moins coûteuse en temps de calcul tient également compte des occultations entre les membres.

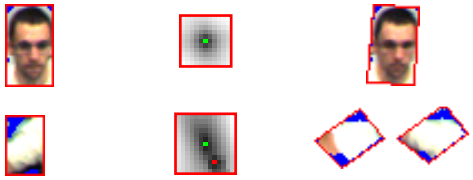
4.1 Calcul des scores

Pour le suivi, l'emploi d'indices génériques reste possible. Par exemple, Felzenszwalb et al. [3] réinitialisent perpétuellement la personne à suivre à chaque nouvelle image de la séquence depuis un modèle purement générique alors que Ramanan et al. [4] ajoutent un apprentissage progressif à leur modèle. Un tel modèle spécifique est en effet plus précis qu'un modèle générique et permet d'améliorer la discrimination des différentes parties du corps notamment en présence de plusieurs personnes. Ici, un modèle d'apparence couleur de chaque partie du corps est extrait à partir de l'initialisation automatique. Un exemple est présenté figure 8*(a)* pour un visage et un bras.

Pour trouver la nouvelle position d'un objet, son modèle est recherché dans la nouvelle image par une technique de corrélation dans l'espace de couleurs RVB. Nous avons choisi la corrélation SAD (Somme de valeur Absolue des Différences), souvent utilisée dans le calcul de disparité ou de compression vidéo MPEG pour sa grande rapidité de calcul, le gain apporté par d'autres corrélations plus complexes ne permettant pas une amélioration significative du suivi. Le score $c_i(e_i)$ est donc la somme des différences entre les pixels de l'objet projeté dans l'image $I(x,y)$ et ceux de son modèle $M(x,y)$:

$$c_i(e_i) = 1 - \frac{1}{nb} \sum_{\{x,y\} \in \ell} \left(\left| I(x,y) - M(x,y) \right| \right) \qquad (7)$$

L'utilisation de l'image précédente permet de limiter la recherche de chaque objet au voisinage de sa position précédente. Le diamètre de recherche correspond alors au déplacement maximal pour un objet entre deux images consécutives. Nous choisissons une zone carrée de 20 pixels de côté, taille suffisante pour le suivi de mouvements naturels humain à 25 images par seconde. La recherche du paramètre θ_i est quant à lui limité à 90° autour de sa valeur précédente avec une précision de 1°. La figure 8 illustre la recherche de ces modèles d'apparence. Une carte dense de scores est établie à partir du calcul des scores de chaque hypothèse testée afin d'en extraire les plus représentatives.



(a) *(b)* *(c)*

Figure 8 : *(a)* modèles d'apparence couleur du visage (en haut) et d'un bras (en bas), *(b)* carte dense des scores avec la position des hypothèses retenues en couleur, *(c)* illustration des hypothèses retenues (deux dans le cas du bras).

Comme pour la phase d'initialisation automatique, les scores d'interaction entre les différents objets liés sont également calculés. L'algorithme de programmation dynamique est ensuite de nouveau mis à contribution pour choisir une hypothèse par membre maximisant la cohérence globale de la structure. Enfin, le modèle d'apparence de chaque partie du corps est mis à jour. Un choix doit ainsi se faire entre la conservation de l'apparence définie à l'initialisation ou son remplacement par la nouvelle apparence de l'hypothèse retenue. Dans le premier cas, un conservatisme trop prononcé peut poser des problèmes si les conditions de luminosité ou de pose évoluent fortement alors qu'une mise à jour systématique peut engendrer rapidement une divergence du modèle. Nous choisissons de ne mettre à jour que l'intensité des modèles d'apparence couleur en utilisant le changement global d'illumination observé pour chaque objet.

On peut noter enfin que beaucoup d'auteurs préconisent la prédiction d'ordre 1 ou 2 de la position d'un membre à partir de filtres de Kalman par exemple. Du fait de la présence de ces articulations souples, nous nous sommes aperçu qu'une telle anticipation sur la position probable du membre était trop souvent erronée. En effet, du fait de certaines limitations du modèle 2D, des sauts entre les différents minima locaux viennent perturber la prédiction de vitesse et d'accélération. Pour cette raison nous ne faisons pas intervenir de prédiction lors du suivi. La souplesse des articulations privilégie ainsi le poids des observations sur l'image.

4.2 Gestion des occultations

Un des problèmes majeurs des systèmes de suivi de corps articulés réside dans la présence d'occultations qui engendrent des pertes de suivi [1,2,5,6]. Deux cas peuvent alors se présenter :

- les inter-occultations entre deux personnes différentes qui se superposent, rendant invisibles certaines parties de leurs corps,
- et les auto-occultations où un membre d'une même personne vient en occulter un autre.

Beaucoup de systèmes permettent de suivre globalement des silhouettes séparées mais ne les différencient plus lors d'interactions avec d'autres personnes. Ces méthodes, dites de *"Split and Merge"*, nécessitent alors la dissociation des silhouettes afin d'en réattribuer l'identité [11]. De plus, le suivi étant global, seule l'information de position des personnes est extraite. Une autre approche consiste à continuer de suivre les personnes même pendant leurs occultations. Un modèle plus complet est alors nécessaire et chaque pixel de l'image doit être classé à tout instant [12,13]. Nous avons choisi cette approche car elle permet de gérer les auto-occultations même dans le cas d'un suivi d'une personne unique.

Pour limiter les problèmes d'occultations entre la tête et les mains, Bernier et al. [8] ajoutent des contraintes de non recouvrement 3D entre ces objets. Sigal et al. [7] introduisent leurs vraisemblances locales *"occlusion-sensitive"* avec des masques binaires qui tiennent compte des occultations entre les membres. Ces deux solutions ajoutent des liens non physiques entre des objets non voisins qui engendrent des graphes bouclés plus complexes à traiter. Pour approcher un traitement temps-réel, nous proposons de conserver l'architecture d'arbre proposée tout en y intégrant une gestion des occultations entre les différents objets suivis. Comme Sigal et al. [7], nous introduisons un masque d'occultation binaire pour chaque objet occulté (figure 9).

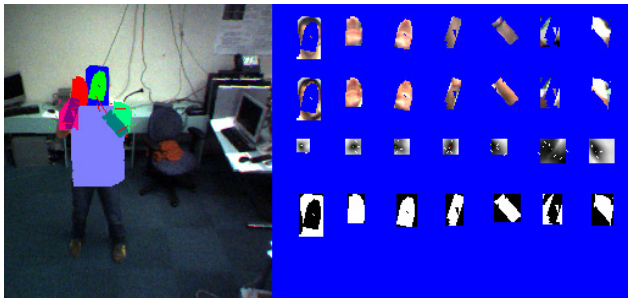


Figure 9 : la partie gauche représente la projection de la configuration précédente sur la nouvelle image. La partie droite illustre les modèles d'apparence des différents objets. La première ligne correspond aux patchs couleur recherchés qui tiennent compte des masques d'occultations et la seconde aux patchs retenus en fin de traitement. La troisième ligne montre les cartes denses des scores de correspondance et les hypothèses retenues. Enfin, en dernière ligne sont représentés les masques d'occultation utilisés pour la recherche de chaque objet.

Ces masques sont initialisés en même temps que les modèles d'apparence couleur. Une pose d'initialisation sans occultation est nécessaire pour obtenir un modèle exact de chaque objet. Lors de la recherche d'un membre, ce masque intervient directement dans le calcul de la corrélation. L'équation (7) est de nouveau utilisée pour calculer les scores de correspondance des objets mais seulement pour les pixels non occultés. Pour les autres, un score neutre leur est attribué correspondant au bruit dans l'image, ce qui permet de retenir des positions potentielles fortement, voire totalement occultées.

La méthode de recherche indépendante par morceaux doit évoluer, les objets occultant devant être cherchés en premier pour créer les masques d'occultations des objets occultés. La figure 10 illustre ce principe. La main droite *(Md)* qui occulte la tête, doit être recherchée en premier. Il en va de même avec l'avant-bras droit *(ABd)* qui occulte à la fois le bras droit *(Bd)* et le torse. Le graphe doit donc être découpé en deux étages. Premièrement, les cinq objets non occultés (la main droite *(Md)*, l'avant-bras droit *(ABd)*, le bras gauche *(Bg)*, l'avant-bras gauche *(ABg)* et la main gauche *(Mg)*) peuvent être recherchés de manière indépendante pour obtenir leurs scores de correspondance respectifs. Comme lors de l'initialisation, le calcul des scores d'interaction entre les objets non occultés liés physiquement est ensuite effectué et l'algorithme de programmation dynamique est appliqué pour réduire le nombre d'hypothèses. Plusieurs hypothèses de position peuvent être retenues pour un membre occultant et il faut alors créer autant de masques d'occultations différents pour chaque objet qu'ils occultent d'où l'importance de limiter ce nombre. Dans la seconde étape, la tête, le bras droit *(Bd)* et le torse peuvent à leur tour être recherchés, leur masque d'occultation respectif étant mis à jour par le choix des hypothèses de position des membres qui les occultent.

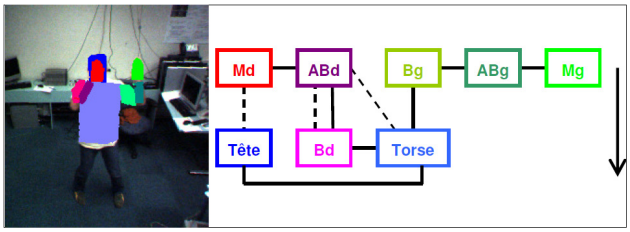


Figure 10 : les lignes continues correspondent aux liens physiques reliant deux membres entre eux et les lignes en pointillés représentent les occultations entre certains membres. Les objets occultant se trouvent alors au-dessus des occultés ce qui impose un ordre de recherche des membres (flèche verticale) et un découpage de la structure d'arbre en plusieurs étages.

Pour déterminer l'ordre de visibilité des objets, les masques d'occultations finaux sont recalculés après le choix de la meilleure configuration. En cas de nouvelle occultation entre deux objets jusque-là non superposés, leur score de correspondance permet de définir l'ordre d'occultation. Comme le font Sigal et al. [7], un ordre a priori des membres est supposé connu par l'algorithme actuel. Ainsi les mains sont systématiquement placées devant les avant-bras, eux-mêmes placés devant les bras. Enfin, le torse et la tête ne pourront jamais être des objets occultant. Cette contrainte a pour but de restreindre le nombre de solutions possibles pour l'ordre des objets en limitant les erreurs. Une unique configuration étant conservée à la fin du traitement de l'image, une erreur d'ordre se propagerait alors dans la suite de la séquence.

4.3 Résultats

Nous présentons dans cette dernière partie les résultats d'une séquence de 14 secondes composée de 350 images. Les mouvements naturels du corps d'une personne faisant face à la caméra sont filmés et traités hors ligne par notre algorithme à une moyenne de 15 images par seconde. Des situations variées comportant des variations d'illumination, des changements d'apparence des membres et d'importantes occultations sont correctement traitées.

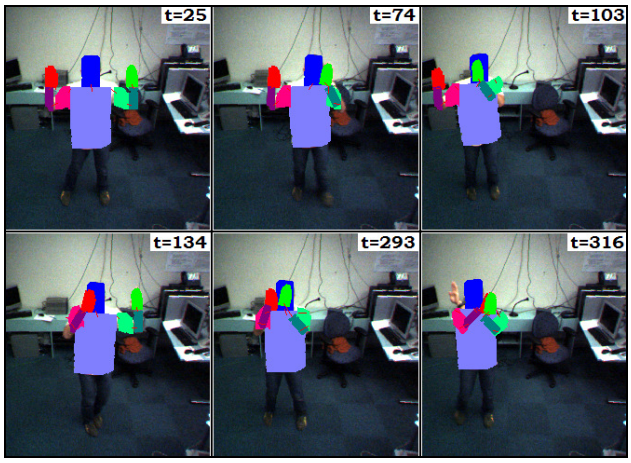


Figure 11 : résultats du suivi sur une séquence test de mouvements naturels variés en présence d'occultations.

Suite à une importante modification d'orientation hors du plan 2D d'initialisation, la projection de l'avant-bras gauche à l'image 293 ne correspond plus à sa taille réelle. Malgré cette erreur, l'algorithme de recherche par morceaux et la flexibilité des liens articulaires permet de trouver une solution qui conserve la cohérence globale de la structure. La meilleure configuration est sélectionnée et les membres continuent d'être suivis. L'algorithme perd finalement la main gauche à l'image 316 confondue avec la droite. Bien que les observations de l'avant-bras et du bras soient erronées, cette configuration est choisie comme étant la meilleure pour cette image. L'ordre d'occultation est alors mis à jour avec la main droite sous la gauche ce qui conduit à la perte de celle-ci dans la suite de la séquence. Un algorithme multi-hypothèses conservant plusieurs configurations possibles au cours du temps devrait permettre de résoudre de tels problèmes.

5 Conclusion

L'un des obstacles majeurs au suivi de corps articulés réside dans les temps de calcul souvent trop élevés. Peu d'algorithmes en temps-réel sont proposés dans la littérature. Par exemple, Demirdjian et al. [6] atteignent le temps-réel grâce à une approche déterministe. Bernier et al. [2] s'en rapprochent fortement avec une fréquence de 12Hz alors que Ramanan et al. [4] annoncent entre 7 et 10 secondes par image. Des méthodes beaucoup plus coûteuses ont été proposées par Felzenszwalb et al. [3] avec leurs *"Pictorial Structures"* (plus d'une minute par image) ou par Sigal et al. [7] (plusieurs minutes pour un algorithme avec la gestion des occultations). L'algorithme que nous proposons dans cet article fonctionne à 15-20Hz suivant la complexité de la scène et tient compte des occultations pour suivre les différents membres d'une personne.

La connaissance à priori de l'ordre d'occultations des objets comme dans [7] limite actuellement nos résultats. Cette approche peut être améliorée en conservant les N meilleures configurations à chaque image, ce qui devrait permettre de traiter les auto-occultations complètes de membres et d'améliorer la robustesse du suivi. Un avantage de l'algorithme proposé est qu'il peut être facilement étendu au suivi multi-personnes grâce au modèle d'apparence couleur. Celui-ci, qui est appris à partir de l'initialisation automatique amène d'une part à des résultats plus précis que l'utilisation d'indices génériques et permet d'autre part de traiter les interactions entre différentes personnes. Enfin, le passage à un modèle 3D ainsi que l'utilisation d'une caméra binoculaire pour le calcul de la disparité devraient également améliorer le suivi [6,9] et permettre de gérer les mouvements hors du plan de l'image.

Bibliographie

[1] T.B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, *Comput. Vis. Image Underst.*, Vol. 104, Issue 2, pp. 90-126, 2006.

[2] I. Haritaoglu, D. Harwood, L.S. Davis, W4: Real-time surveillance of people and their actions, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 22, pp. 809-830, 2000.

[3] P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial Structures for Object Recognition, *Int. J. Comput. Vis.*, Vol. 61, Issue 1, pp. 55-79, 2005.

[4] D. Ramanan, D.A. Forsyth, A. Zisserman, Tracking People by Learning Their Appearance, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 29, Issue 1, pp. 65-81, 2007.

[5] S. Ju, M. Black, Y. Yacoob, Cardboard people: A parameterized model of articulated motion, *Int. Conf. on Auto. Face and Gesture Recognition*, pp. 38-44, 1996.

[6] D. Demirdjian, L. Taycher, G. Shakkhnarovich, K. Grauman, T. Darrell, Avoiding the streetlight effect: Tracking by exploiting likelihood modes, *IEEE Int. Conf. on Comput. Vis.*, Vol. 1, pp. 357-364, 2005.

[7] L. Sigal, M.J. Black, Measure Locally, Reason Globally: Occlusion-sensitive Articuled Pose Estimation, *IEEE Comput. Society Conf. on Comput. Vis. and Pattern Recognition*, Vol. 2, pp. 2041-2048, 2006.

[8] M.A. Fischler, R.A. Elschlager, The representation and matching of pictorial structures, *IEEE Transactions on Computer*, Vol. 22, pp. 67-92, 1973.

[9] O. Bernier, P. Cheung-Mon-Chan, A. Bouguet, Fast Nonparametric Belief Propagation for Real-Time Stereo Articulated Body Tracking, *Comput. Vis. Image Underst.*, Vol. 113, no. 1, pp. 29-47, 2009.

[10] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient matching of pictorial structures, *Comput. Vis. and Pattern Recognition*, 2000.

[11] S. McKenna, S. Jabri, Z. Duric, H. Wechsler, Tracking Groups of People, *Comput. Vis. Image Underst.*, 2000.

[12] A. Elgammal, L. Davis, Probabilistic framework for segmenting people under occlusion, *IEEE Int. Conf. on Comput. Vis.*, Vol. 2, pp. 145-152, 2001.

[13] A. W. Senior, A. Hampapur, L. M. Brown, Y. Tian, S. Pankanti, R. M. Bolle, Apparence Models for Occlusion Handling, *PETS*, 2001.

[14] H.G. Barrow, J.M. Tenenbaum, R.C. Bolles, H.C. Wolf, Parametric correspondence and chamfer matching: Two new techniques for image matching, *Int. Joint Conf. on Artificial Intell.*, pp. 659-663,1977.

[15] D. Gavrila, Pedestrian detection from a moving vehicle, *Europ. Conf. Comput. Vis.*, 2000.