Suivi 3D Monoculaire pour un Système de Vidéosurveillance à l'aide d'un Modèle de Mouvement et un Modèle d'Apparence

Mourad Boufarguine, Nicolas Thome, Vincent Guitteny, Frédéric Precioso

Suivi 3D Monoculaire pour un Système de Vidéosurveillance à l'aide d'un Modèle de Mouvement et un Modèle d'Apparence

M. Boufarguine[1] N. Thome[2] V. Guitteny[1] F. Precioso[3]

[1] Thales Training & Simulation SAS
[2] UPMC - Sorbonne Universités
[3] Université Nice-Sophia Antipolis

1 rue du Général de Gaulle, 95520 Osny, France
mourad.boufarguine@thalesgroup.com

Résumé

Le besoin en méthodes non intrusives d'analyse des mouvements humains se fait sentir à travers des applications comme la vidéosurveillance intelligente, les interfaces homme-machine et l'indexation multimédia.
Dans cet article, nous proposons à une approche générative se basant sur un filtre particulaire à recuit simulé (APF) : une fonction de vraisemblance qui combine des mesures basées sur les silhouettes et sur l'apparence, et un modèle temporel se basant sur une réduction de l'espace des poses pour une activité donnée. Le filtre proposé permet d'estimer en ligne la vitesse de marche ainsi que les coordonnées du cycle dans l'espace réduit. Nous évaluons l'approche proposée sur la base de données HumanEva. Les résultats du suivi montrent que la fonction de vraisemblance mixte réduit l'erreur 3D. Le modèle temporel proposée permet d'améliorer le suivi tout en réduisant le coût calculatoire du filtre particulaire.

Mots Clef

Estimation de pose 3D, suivi monoculaire, modèle de mouvement, modèle d'apparence, histogrammes de couleur.

Abstract

A wide range of applications would benefit from a robust solution to monocular human pose estimation and tracking such as video surveillance systems, human-computer interfaces and video indexing and retrieval.
In this article, we propose improvements to a model-based approach within an Annealed Particle Filtering framework : a likelihood function that combines silhouette and visual appearance information ; a temporal prior embedded in particle states using a PCA dimension reduction of walking cycle descriptions. The latent walk motion space is used within a Bayesian framework that allows inline recovering of the walk speed and the motion coordinates. Finally, a set of experiments on the HumanEva datasets is presented. Tracking results using the proposed likelihood function show that mixing image features can increase the estimation accuracy. Additionally, the results are further improved and the computational cost is reduced when the PCA-based temporal prior is included.

Keywords

3D pose estimation, monocular tracking, temporal model, appearance model, color histograms.

FIGURE 1 – Gauche : représentation d'un système classique de vidéosurveillance. Droite : virtualisation 3D+t de la scène dynamique.

1 Introduction

Dans les systèmes classiques de vidéosurveillance (Figure 1), les flux vidéo sont représentés sous la forme d'une mosaïque sur un écran de contrôle. L'opérateur est fortement sollicité en observant et analysant la quantité considérable de données capturées par les caméras. Des études sur l'efficacité des opérateurs des systèmes de vidéosurveillance menées par l'institut national américain de justice [11] montrent qu'"*une telle tâche [détecter manuellement les évènements dans un système de vidéosurveillance], ne peut être efficace même si elle est réalisée par une personne dévouée et bien intentionnée. Au bout de seulement 20 minutes d'observation et d'évaluation des écrans de contrôle, l'attention de la plupart des personnes baisse à des niveaux au delà des niveaux acceptables. Surveiller des écrans est une tâche ennuyante et hypnotisante. Elle n'engage aucun stimuli intellectuel comme lors du visionnage d'une émission de télévision*".
Plusieurs travaux se sont intéressés à minimiser l'effort de

l'opérateur [5, 10]. Les approches existantes ont tendance à combiner la vidéo avec l'environnement 3D afin de fournir une cohérence visuelle entre les flux vidéo et le contexte spatial. Wang et al. [23] présentent une étude comparative des différentes techniques liées aux vidéos contextualisées. Dans cet article, nous proposons une intégration plus poussée des flux vidéo sous forme d'une virtualisation complète de la scène observée. La scène virtualisée est composée d'éléments statiques qui représentent l'environnement, et d'éléments humanoïdes mobiles qui représentent les personnes en mouvement détectées dans les flux des caméras. Généralement, dans les installations de vidéosurveillance existantes, il n'y a pas de recouvrement entre les champs des différentes caméras ; nous traiterons donc le problème d'estimation de pose dans le contexte monoculaire.

(a) Un système classique de vidéosurveillance. (b) Objectif final : virtualisation de mouvements humains dans un environnement intérieur

FIGURE 2 – Virtualisation d'une scène dynamique

L'estimation de pose d'une personne revient à estimer, au cours du temps, la configuration de la structure cinématique du corps humain.
La suite de cet article est organisée comme suit. Le modèle 3D est présenté au paragraphe 2. Le paragraphe 3 dresse l'état de l'art du suivi de mouvements. Le paragraphe 4 présente notre approche. Des expérimentations avec la base de HumanEva [17] sont présentées dans le paragraphe 5. Le paragraphe 6 conclut et présente le travail futur.

2 Modèle 3D Humanoïde

Dans cet article, le corps humain est modélisé par 15 parties sous forme de cônes tronqués comme dans [16]. Les articulations des épaules, des cuisses, du thorax et du cou ont 3 degrés de libertés (DDL) chacune. Les articulations des clavicules ont 2 DDL. Les autres articulations (genoux, chevilles, coudes et poignets) sont modélisées par un seul DDL chacune. Les dimensions des parties du corps sont supposées être constantes et connues *a priori*. La configuration du corps (pose) peut donc être totalement définies par *36 paramètres* : 6 paramètres définissent la position et l'orientation globales du bassin ; 30 paramètres définissent les angles relatifs entre les différents membres.
Pour estimer la position et l'orientation de chaque membre dans un repère monde, on applique la cinématique directe le long de l'arbre cinématique modélisant le corps. La pose d'un nœud de l'arbre cinématique dépend alors de celle de son nœud père ansi que les angles de rotation relative entre les deux. L'utilisation d'un tel arbre pour modéliser le corps humain, permet de formuler la pose 3D du corps par un vecteur d'état $\mathbf{x}$:

$$\mathbf{x} = [\tau_x^g, \tau_y^g, \tau_z^g, \theta_x^g, \theta_y^g, \theta_z^g, \mathbf{\Theta}],$$

où $[\tau_x^g, \tau_y^g, \tau_z^g]$ est la position globale de la racine de l'arbre (bassin) dans le repère monde global, $[\theta_x^g, \theta_y^g, \theta_z^g]$ sont les angles de la rotation globale du bassin et $\mathbf{\Theta}$ est un vecteur d'angles relatifs entre les différentes paires de nœuds adjascents dans l'arbre cinématique.

3 Contexte

Les méthodes de suivi de personnes peuvent être classées en deux catégories [13] : méthodes discriminatives et méthodes génératives. Les méthodes discriminatives établissent une relation directe entre l'espace des observations (images) et l'espace des poses 3D moyennant une base de données labélisées. Une fonction probabiliste peut être apprise à l'aide d'une régression [1] ou une recherche des plus proches voisins [8]. Les approches dicriminatives ont l'avantage d'être rapides. Par contre, elles nécessitent un ensemble d'entraînement assez représentatif des configurations possibles qu'on peut observer dans les flux vidéo, ce qui entraîne un nombre assez important d'échantillons annotés pour pouvoir construire une fonction de décision capable de retrouver les configurations possibles.
Les méthodes génératives se basent sur un modèle du corps humain permettant de générer des observations pour évaluer les hypothèses de pose. Ces approches procèdent généralement à une recherche de la pose optimale dans un espace de grande dimension, ce qui entraîne un coût calculatoire important. Dans ces méthodes, le suivi est généralement formulé dans un cadre bayésien. Dans ce contexte, plusieurs algorithmes ont été proposés pour le suivi de personnes [6, 22]. Le filtrage particulaire a été largement utilisé pour le suivi de personnes [16, 6] en raison de sa capacité à propager des hypothèses multiples. En utilisant la notation du vecteur d'état de la section précédente, le problème de suivi revient à estimer séquentiellement le vecteur d'état $\mathbf{x}_t$ à l'instant t, en utilisant les observations $\mathbf{y}_t$ fournies par la caméra ainsi que l'état précédent $\mathbf{x}_{t-1}$. Dans le cadre de l'inférence bayésienne, la distribution *a posteriori* $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ de $\mathbf{x}_t$ est calculée à partir du *posteriori* à l'instant précédent $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$, un *a priori* temporel $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ et la fonction de vraisemblance $p(\mathbf{y}_t|\mathbf{x}_t)$.
Le filtre particulaire permet d'approximer la distribution *a posteriori* à l'aide d'un ensemble de N échantillons appelés *particules*. Dans notre cas, chaque particule est composé d'un vecteur d'état $\mathbf{x}_t^{(i)}$ qui représente une configuration possible du corps et d'un poids $\pi_t^{(i)}$ qui est proportionnel à la vraisemblance évaluée à l'état de la particule, $\pi_t^{(i)} \propto p(\mathbf{y}_t|\mathbf{x}_t^{(i)})$. Malgré leur flexibilité et leur multi modalité, le coût calculatoire des filtres particulaires peut devenir très important quand un nombre élevé de particules est nécessaire. La grande dimension de l'espace des poses rend donc le suivi coûteux en temps de calcul. Pour mieux

échantillonner l'espace d'état et réduire par conséquence le nombre de particules nécessaires, le filtre particulaire avec une étape de recuit simulé (Annealed Particle Filter) [6] permet de concentrer les particules autour des pics de la distribution *a posteriori*. [6, 22, 16] montrent qu'il est plus robuste aux occultations et aux grandes variations de poses.

La fonction de vraisemblance permet d'évaluer les hypothèses de pose à chaque instant en fonction de l'observation courante. Plusieurs descripteurs d'image peuvent être utilisés dans la formulation de cette fonction tels que les silhouettes [6, 16], les contours [16] ou le flôt optique [4]. En se basant sur l'hypothèse d'une apparence qui varie peu, plusieurs fonctions de vraisemblance basées sur l'apparence des parties du corps ont été proposées [2, 14, 9, 13]. Ces fonctions sont souvent calculées à partir des histogrammes de couleur ou de template matching. Balan et Black [2] intègrent un modèle d'apparence décrit par des patchs d'images dans une fonction de vraisemblance basée sur les silhouettes. Ils prennent en compte l'occultation des parties de corps en utilisant des cartes de visibilité. Cette fonction de vraisemblance mixte permet d'améliorer le suivi dans le cas d'un suivi multi-vues. La mesure liée aux silhouettes ne permet pas de pénaliser les pixels de la silhouette non recouverts par la projection de l'hypothèse de pose. Cette limitation peut être adressée par une formulation bidirectionnelle de cette mesure [19, 18] permettant d'améliorer les résultats du suivi [16] aussi bien dans le cas monoculaire que multi-vues. Toutefois, dans un suivi monoculaire, les silhouettes ne permettent pas d'éviter les ambiguïtés de profondeur [12]. Pour pallier cette limitation, d'autres descripteurs plus discriminants doivent être mis à contribution dans la fonction de vraisemblance pour améliorer la qualité du suivi.

Des modèles temporels sont utilisés pour propager les particules d'un instant à un autre. Ils représentent les connaissances à propos du déplacement de la personne suivie entre deux instants consécutifs. Les modèles temporels les plus basiques sont des modèles gaussiens dans l'espace des poses et par conséquence, ne modélisent pas assez bien ni le type d'activité que la personne est en train d'entreprendre, ni sa trajectoire. Par contre, l'espace de recherche peut être significativement réduit en utilisant des modèles temporels plus riches tel que des modèles physiques [4, 21], ou des modèles se basant sur des techniques de réduction de la dimensionalité de l'espace des poses [15, 20, 7]. Malgré la grande dimension de l'espace des poses du corps humain, l'espace de recherche peut être drastiquement réduit dans le cas d'une classe d'activité observée. La connaissance *a priori* de l'activité permet donc de réduire l'espace de poses en un ensemble de poses plausibles. L'activité de marche, par exemple, est une activité cyclique et peut donc être segmentée en cycles. Dans [20, 15], la dimensionnalité de l'espace des poses pour l'activité de marche est réduite à l'aide d'une Analyse en Composantes Principales (ACP) de l'espace de mouvement. L'espace de mouvement est définie comme la concaténation de vecteurs de poses statiques pour un cycle entier. L'ACP permet de construire un espace latent réduit de dimension plus faible. Cet espace réduit est défini à partir des composantes principales de l'espace de mouvement. A partir des coordonnées dans cet espace réduit, on peut remonter l'ensemble des poses successives de tout le cycle. Pour chaque instant (ou *phase* dans [20, 15]) dans le cycle reconstruit, on récupère un vecteur de pose contraint par un type d'activité. Estimer la durée du cycle de marche lors d'un suivi en ligne est assez délicat, puisqu'elle varie d'une personne à une autre et peut aussi varier pour la même personne. Elle dépend aussi de la fréquence d'échantillonnage du matériel d'acquisition. Dans [20], la variation de la largeur de la silhouette de la personne permet d'estimer le paramètre de phase pour toute la séquence de test, avant d'estimer les paramètres du mouvement. Toutefois, l'utilisation des silhouettes dépend fortement du point de vue de la caméra. De plus, toute la séquence doit être traitée une première fois pour avoir une première estimation de la phase pour toutes les images de la séquence. Une minimisation déterministe permet alors d'estimer les coordonnées du mouvement dans l'espace réduit et affiner les estimations de la phase. Cette méthode est donc incompatible avec un traitement en ligne requis pour un système de vidéosurveillance. Dans [15], un incrément constant est utilisé pour propager le paramètre de phase dans le cadre d'un suivi. Cette approche ne convient pas pour des séquences de test où la vitesse de marche est assez différente ce celle des séquences d'apprentissage.

4 Méthode proposée

Dans cet article, nous proposons un algorithme de suivi en ligne pour un système de vidéosurveillance. Nous proposons une fonction de vraisemblance qui combine la mesure bidirectionnelle des silhouettes avec un modèle d'apparence. Nous adaptons le modèle d'apparence proposé par Gall et al. [9] pour le cas d'un suivi monoculaire en utilisant des cartes de visibilités pour le calcul du modèle d'apparence de chaque partie du corps. Nous montrons qu'un tel modèle est plus robuste que les templates. Nous améliorons aussi les précédents travaux sur les modèles de mouvements basés sur une réduction par ACP de la dimensionnalité [20, 15] en proposant une méthode d'extraction automatique de cycles de marche depuis une base de données *MoCap*, et en introduisant un paramètre *pas* dans le vecteur d'état réduit pour tenir compte des différences de vitesses et de fréquences d'échantillonnage entre les séquences *MoCap* d'entraînement et les séquences vidéo de test. Notre approche est capable de s'adapter à des vitesses ou des fréquences différentes. Contrairement à [20], nous formulons le problème de suivi comme un problème d'inférence incrémentale.

4.1 Fonction de vraisemblance mixte

Nous proposons de combiner une mesure basée sur les silhouettes avec une mesure basée sur un modèle d'appa-

rence. La mesure basée sur les silhouettes utilise une formulation bidirectionnelle qui pénalise les régions de non recouvrement entre la silhouette de la personne et la projection d'une hypothèse de pose :

$$-\ln p_s(\mathbf{y}_t|\mathbf{x}_t) \propto \frac{1}{2}\left[\frac{N_w}{\sum_{\mathbf{p}} M_s(\mathbf{p})} + \frac{N_y}{\sum_{\mathbf{p}} M_m(\mathbf{p})}\right] \quad (1)$$

où M_s est le masque binaire de la silhouette et M_m est celui de la projection d'une hypothèse de pose. N_w est le nombre de pixels de M_m non recouverts par la silhouette, N_y est le nombre de pixels de M_s non recouverts par la projection de l'hypothèse de pose. L'apparence de la personne suivie est modélisée par des histogrammes couleur dans l'espace CIE Lab. Seules les composantes de chrominances *a* et *b* sont utilisées en raison de leur robustesse au changement de luminosité. Lors de l'initialisation, un modèle d'apparence composé de plusieurs histogrammes de couleur est calculé : le torse, le bassin, la tête et les paires symétriques des membres sont modélisés par un histogramme chacun en supposant que les membres symétriques ont la même apparence.

La mesure de vraisemblance liée à l'apparence d'une particule est égale à la distance entre les histogrammes du modèle d'apparence calculés lors de l'initialisation du suivi et ceux calculés à partir de la projection d'une particule sur l'image courante. Soit $\hat{A}_p = \{\hat{h}_b\}_{b=1}^{B}$ le modèle d'apparence composé de $B = 9$ histogrammes $\hat{h}_b$ calculés lors de l'initialisation, et $\{h_b\}_{b=1}^{B}$ l'ensemble des histogrammes extraits de l'image courante et pour une particule donnée. La fonction de vraisemblance s'écrit donc comme une distances entre ces deux ensembles d'histogrammes :

$$-\ln p_a(\mathbf{y}_t|\mathbf{x}_t) \propto \sum_{s=1}^{B} w_s \left[BC(\hat{h}_s, h_s)\right], \quad (2)$$

où w_s sont des poids normalisés proportionnels aux dimensions des parties du corps. Ils sont calculés lors de l'initialisation en fonction du ratio du nombre de points échantillonnés sur chaque partie du corps par rapport au nombre total de points échantillonnés. $BC(\hat{h}_s, h_s)$ est le cœfficient de *Bhattacharyya* [9]. La fonction de vraisemblance (2) ne prend pas en compte le nombre de pixels utilisés pour calculer l'apparence d'une particule. Pour cette raison, une particule peut avoir un poids important même si la silhouette de la personne n'est pas entièrement couverte par la projection du modèle 3D.

Ce problème peut être adressé en intégrant une mesure basée sur les silhouettes à la fonction de vraisemblance (2). Soit $p_s(\mathbf{y}_t|\mathbf{x}_t)$ la log-vraisemblance basée sur les silhouettes, et $p_a(\mathbf{y}_t|\mathbf{x}_t)$ la log-vraisemblance basée sur l'apparence, la fonction de vraisemblance mixte s'écrit donc :

$$-\ln p(\mathbf{y}_t|\mathbf{x}_t) \propto (1-\alpha)p_s(\mathbf{y}_t|\mathbf{x}_t) + \alpha p_a(\mathbf{y}_t|\mathbf{x}_t), \quad (3)$$

où α est un poids pondérant l'importance de chaque terme de la fonction de vraisemblance mixte. Nous montrons dans la section 5 que la fonction de vraisemblance mixte améliore considérablement le suivi.

4.2 Modèle temporel

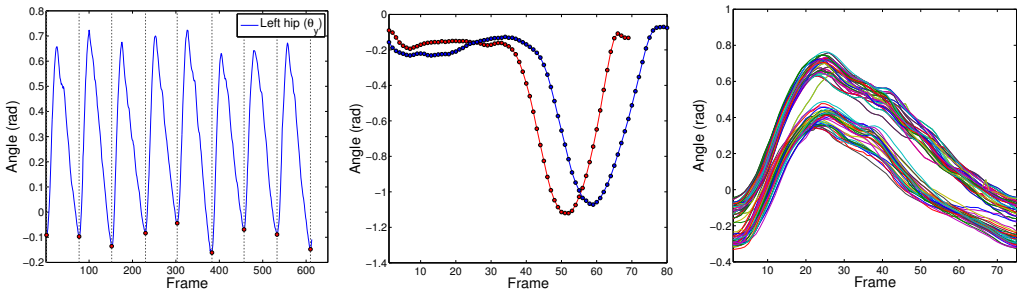
Comme discuté précédemment, le modèle de déplacement $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ modélise l'information *a priori* sur le mouvement de la personne suivie entre deux instants successifs. Nous proposons un modèle de mouvement appris à partir de données de capture de mouvement (*MoCap*) permettant de synthétiser des cycles de marche. Ce modèle est utilisé dans le suivi pour contraindre la recherche de poses à celles correspondant à un cycle de marche. Un modèle de mouvement de marche peut être appris à partir de séquences de capture de mouvement en utilisant une Analyse en Composantes Principales (ACP) [15, 20]. Les données doivent être tout d'abord segmentées en cycles et mis à l'échelle. Etant donnée une séquence de marche décrite par l'évolution au cours des temps des différents angles de DDL des articulations, les minima de l'angle de la cuisse (gauche dans notre cas) correspondent au début de chaque cycle, comme le montre la figure 3(a). Les intervalles de temps ainsi obtenus permettent de segmenter tous les angles de la séquence.

Après la segmentation, on obtient un ensemble de cycles de marche ayant des poses de début et de fin similaires. Toutefois, comme le montre la figure 3(b), les cycles sont de longueurs différentes en raison des différences de vitesse entre les personnes. Ils sont alors mis à une même échelle pour avoir des cycles contenant le même nombre d'échantillons. Pour la mise à l'échelle, les valeurs des angles sont interpolées pour chaque échantillon. Nous avons appliqué la méthode décrite plus haut aux séquences de marche de la base HumanEva-I [17] (base d'entraînement). La base contient des séquences de marche effectuées par 3 personnes différentes à des vitesses différentes. La figure 3(c) montre l'évolution de l'angle de la cuisse gauche pour tous les 64 cycles extraits de la base d'entraînement.

Pour la réduction en ACP, on retient seulement q composantes principales. Soit $\tilde{U} = [\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_q}]$ la matrice composée des q vecteurs propres associés aux valeurs propres les plus grandes. Un cycle de marche $\tilde{\mathbf{a}}$ peut donc être synthétisé en utilisant un point $\mathbf{c} = [c_1, c_2, \ldots, c_q]$ du sous espace de dimension q :

$$\tilde{\mathbf{a}} = \bar{\mathbf{a}} + \tilde{U}\mathbf{c}. \quad (4)$$

Tous les angles des articulations de notre modèle 3D ont été inclus dans les cycles de marche sauf les angles du cou. La position et orientation globale ne sont pas incluses car généralement elles ne sont pas cycliques dans un mouvement de marche. Le problème d'estimation de pose revient donc à estimer un cycle de mouvement défini par ses q coordonnées dans l'espace réduit appris à partir des données *MoCap* d'entraînement. Un cycle de mouvement étant une concaténation de vecteurs de pose, une pose 3D peut alors être extraite à partir du cycle en l'évaluant à un instant donné. On introduit alors un paramètre temporel appelé *phase* dans le vecteur d'état avec les q coordonnées d'un cycle de mouvement dans l'espace réduit. L'introduction des coordonnées dans le sous espace réduit et le paramètre de phase $\mu \in [0, 1]$ permet de contraindre le mouvement en



(a) (b) (c)

FIGURE 3 – Segmentation d'une séquence de marche en cycles. La séquence de marche dans (a) est segmentée (ligne discontinue) en utilisant les minima de l'angle de la cuisse gauche (pointillés rouges) ; tous les angles sont segmentés selon les intervalles obtenus. (b) montre deux cycles (ici seulement l'angle correspondant au genoux) correspondant à différentes personnes, (c) montre la variation de l'angle de la cuisse gauche pour plusieurs cycles extraits et mis à l'échelle.

un mouvement de marche et par conséquence de réduire la dimension du vecteur d'état.

Le vecteur d'état $\mathbf{x}$ est alors reformulé en un vecteur réduit $\mathbf{x}^r$ qui se compose de la position et rotation globales du bassin (τ^g and θ^g), l'angle du cou (θ^h_y), les coordonnées d'un point dans l'espace réduit de mouvement ($\mathbf{c}$) et la phase (μ) :

$$\mathbf{x}^r = [\tau^g_x, \tau^g_y, \tau^g_z, \theta^g_x, \theta^g_y, \theta^g_z, \theta^h_y, \mu, \mathbf{c}]. \quad (5)$$

Il est nécessaire de pouvoir reconstruire, à partir de ce vecteur réduit, la pose 3D du modèle humanoïde pour pouvoir générer des observations et l'évaluer à l'aide de la fonction de vraisemblance. Pour un vecteur réduit donnée $\mathbf{x}^r$, le cycle de mouvement $\tilde{\mathbf{a}}$ est d'abord calculé selon l'équation (4). Ensuite, les angles des articulation Θ sont extraits du cycle évalué à l'instant μ : $\Theta = \tilde{\mathbf{a}}\,[\mu T]$, où T est le nombre d'échantillons dans un seul cycle. $\tilde{\mathbf{a}}[\mu T]$ est le vecteur des angles des articulations du cycle $\tilde{\mathbf{a}}$ à l'instant μ.

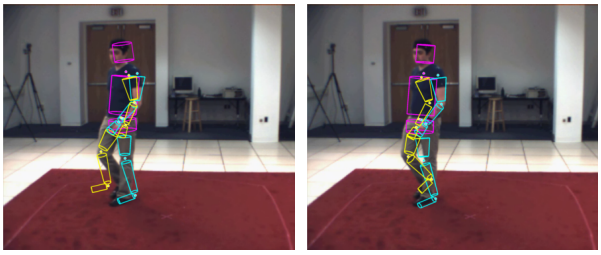
Dans [15], un incrément constant est utilisé pour propager le paramètre de phase (équation 6).

$$\mu_t = \mu_{t-1} + \frac{1}{T} + B_t \quad (6)$$

où B_t est un bruit gaussien centré. La phase est initialisée manuellement, les paramètres ACP sont initialement nuls (le cycle moyen).

Ce modèle de mouvement est capable de contraindre un mouvement de marche lors du suivi, ce qui améliore la qualité du suivi tout en réduisant le nombre de particules nécessaires. Toutefois, la propagation de la phase selon l'équation (6) ne convient pas pour des séquences où la vitesse de marche est assez différente de celles des séquences d'entraînement. La figure 4(a) montre l'image 93 d'une séquence du sujet 2 de la base HumanEva-I [17]. La figure montre que la pose n'est pas bien estimée en raison d'un décalage de la phase.

Pour adresser cette limitation, on introduit un paramètre *pas* noté s_t au vecteur d'état. Il est initialisé à $^1/_T$ et propagé dans le temps en utilisant un bruit gaussien. La phase est alors incrémentée à chaque instant en utilisant le paramètre s_t :



(a) (b)

FIGURE 4 – Résultat de l'estimation de la phase μ à l'image 93 : (a) la phase est propagée selon l'équation (6) utilisant un bruit gaussien additif. (b) En introduisant un paramètre de pas dans le vecteur d'état, le résultat du suivi est amélioré.

$$\begin{aligned} \mu_t &= \mu_{t-1} + s_{t-1} \\ s_t &= s_{t-1} + B_t \end{aligned} \quad (7)$$

Le suivi du paramètre de pas est similaire au suivi de la vitesse de marche. La figure 4(b) montre que suivi est amélioré.

5 Evaluation

Notre méthode de suivi monoculaire a été évaluée sur des séquences de la base HumanEva-I [17] et sur des vidéos provenant d'un système réel de vidéosurveillance. Nous séparons les séquences multi caméras en séquences monoculaires que nous traitons séparément. Nous nous limitons aux séquences de marche filmées par des caméras couleur. Les séquences de la base d'entraînement pour les personnes 1 à 3 sont utilisées pour l'apprentissage du sous espace. Les séquences de la base de validation du sujet 2 sont utilisées pour une évaluation quantitative.

Pour évaluer les performances des différents algorithmes, nous utilisons la mesure d'erreur absolue proposée par [17], qui calcule la moyenne des distances entre 15 marqueurs virtuels sur le corps.

D'abord, nous évaluons la fonction de vraisemblance mixte proposée en utilisant un modèle *a priori* de degré zéro et à

bruit gaussien additif en considérant des limites d'angles apprises à partir de la base d'entraînement. Ensuite, nous montrons que le modèle de mouvement proposé permet d'avoir des résultats similaires voire meilleures que le modèle *a priori* de degré zéro et à bruit gaussien additif tout en réduisant le nombre de particules et par conséquence, le coût calculatoire. Finalement, la combinaison du modèle temporel et de la fonction de vraisemblance est évaluée sur des vidéos d'un système de surveillance.

5.1 Performances du modèle d'apparence

Nous avons comparé la fonction de vraisemblance mixte proposée à une fonction basée seulement sur les silhouettes [17], et à une fonction de vraisemblance basée sur les silhouettes et une apparence basée sur les templates. Pour la comparaison, nous utilisons un modèle *a priori* de degré zéro et à bruit gaussien additif.

Pondération des deux termes de vraisemblance. Notre fonction de vraisemblance est une somme pondérée de deux termes basés sur les silhouettes et les histogrammes de couleur. Le paramètre pondérant α est déterminé empiriquement en comparant l'erreur 3D moyenne pour différentes valeurs de α. La figure 5 montre l'erreur 3D moyenne pour la séquence du sujet 2 filmé par la caméra 2 (S2/C2) en fonction du paramètre α. Pour $\alpha = 0$, la fonction de vraisemblance se réduit au terme des silhouettes. On note qu'introduire les histogrammes de couleur améliore le résultat du suivi. L'erreur minimum ($195.16mm$)

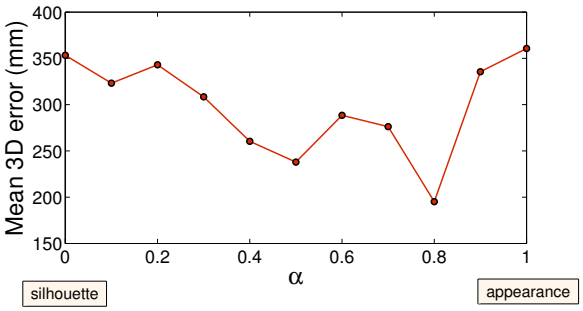


FIGURE 5 – Erreur 3D moyenne pour la séquence S2/C2 de HumanEva-I pour différentes valeurs d'α.

pour cette séquence est obtenue pour $\alpha = 0.8$. Cette valeur sera utilisée pour la suite.

Fonction de vraisemblance. Pour l'évaluation, nous avons utilisé un filtre avec 200 particules pour chacune des 5 couches. Comme précisé précédemment, nous propageons les particules avec un bruit gaussien centré additif pour pouvoir évaluer les fonctions de vraisemblance indépendamment du modèle *a priori*. Le tableau 1 montre l'erreur 3D moyenne pour chaque séquence de validation de la base HumanEva-I. La fonction proposée améliore les performances du suivi comparée à une fonction ne prenant en compte que les silhouettes.

Méthode	S2/C1(mm)	S2/C2(mm)	S2/C3(mm)
Silhouette	498.13	353.33	454.32
Mixte	**359.04**	**195.16**	**306.11**

TABLE 1 – Comparaison de l'erreur 3D moyenne entre la fonction de vraisemblance basée sur les silhouettes (Silhouette) et la fonction mixte (Mixed).

5.2 Performance du modèle *a priori*

Comparaison entre le modèle proposé et le modèle à bruit gaussien additif. Pour évaluer les performances du modèle *a priori* proposé par rapport au modèle classique à bruit gaussien additif, nous utilisons la fonction mixte de vraisemblance. Nous retenons les q premières composantes principales expliquant 95% de l'information extraite de la base d'entraînement. Nous obtenons $q = 11$.
Les résultats sont regroupés dans le tableau 2. Le modèle *a priori* permet de réduire la dimension de l'espace de recherche de pose en le limitant aux poses correspondant à un mouvement de marche. Ainsi, le nombre de particules nécessaires est réduit. Pour les deux séquences S2/C1 et S2/C3, nous obtenons de meilleurs résultats avec 7 fois moins de particules. Le temps de calcul est fortement réduit pour toutes les séquences. En utilisant le même nombre de particules pour les deux modèles *a priori*, l'erreur 3D moyenne est 10 fois plus faible.

Modèle (particles×couches)	S2/C1	S2/C2	S2/C3	Facteur d'accélération
ordre 0 (200×5)	359.04	**195.16**	306.11	1
ordre 0 (50×3)	3644.89	3248.24	3274.85	6.2
Notre modèle (50×3)	**202.07**	271.07	**290.51**	**7.4**

TABLE 2 – Comparaison des erreurs 3D moyennes (mm) entre un modèle *a priori* d'ordre zéro et le modèle proposé.

Propagation de la phase. L'introduction du paramètre de pas (équation (7)) permet d'améliorer les performances du modèle *a priori* proposé et son adaptation à des vitesses de marche différentes de celles de la base d'entraînement. Nous avons comparé la méthode proposé de propagation de la phase avec la méthode utilisée dans [15]. Les erreurs 3D moyennes pour les séquences HumanEva sont reportées dans le tableau 3. Le tableau 3 montre que le paramètre de

Method	S2/C1	S2/C2	S2/C3
Constant increment [15]	355.96	307.98	390.80
Our method	**202.07**	**271.07**	**290.51**

TABLE 3 – Comparaison des erreurs 3D moyennes (mm) des méthodes de propagation de la phase.

pas améliore la précision du suivi 3D en mieux estimant la phase au sein du cycle de mouvement (figure 4). L'introduction du paramètre de pas est très important aussi pour

des cas où la fréquence d'acquisition des capteurs est assez différente de celle des vidéos d'entraînement.

5.3 Evaluation globale

La figure 6 montre l'erreur 3D pour chaque image de la séquence S2/C1 pour trois différentes méthodes de suivi. L'erreur est plus faible lorsque le modèle d'apparence proposé est incorporé dans la fonction de vraisemblance (courbe bleue). L'utilisation du modèle de mouvement proposé pour propager les particules réduit davantage cette erreur (courbe verte). La figure 7 montre le résultat du suivi sur quelques images de cette séquence.

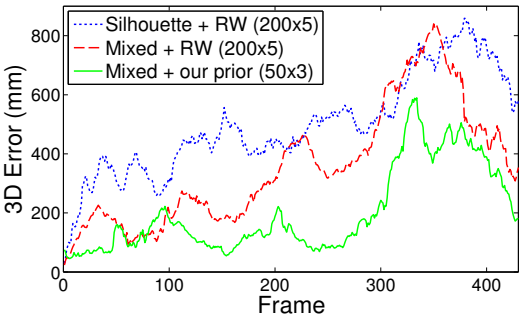


FIGURE 6 – Erreurs 3D des 3 différentes méthodes pour la séquence S2/C1.

5.4 Test sur une séquence réelle

Nous avons aussi évalué l'approche proposée sur une séquence enregistrée depuis une caméra de surveillance. Les silhouettes sont extraites en utilisant un algorithme en ligne de segmentation de fond [3]. La séquence présente divers difficultés : une faible fréquence d'échantillonnage et un arrière plan surchargé. Malgré ces difficultés, la méthode proposée permet d'estimer correctement la pose de la personne suivi (deuxième ligne dans la figure 8). Les poses ainsi estimées sont alors utilisées pour animer un modèle humanoïde virtuel (3^e et 4^e lignes de la figure 8).

6 Conclusion

Dans cet article, nous avons proposé une fonction vraisemblance qui combine l'information silhouette avec un modèle d'apparence prenant en compte les auto-occultations des différentes partie du corps. Le modèle d'apparence consiste en des histogrammes de couleur calculés dans l'espace couleur CieLab. Nous avons utilisé un modèle *a priori* pour propager les particules du filtre. Le modèle se base sur la construction d'un espace de poses réduit aux poses d'un mouvement de marche en faisant une ACP sur des séquences d'apprentissage. Nous avons introduit un paramètre qui modélise l'incrément temporel dans le cycle de marche pour prendre en compte la variation de vitesse entre individus et la différence de la fréquence d'échantillonnage des capteurs. Les résultats sur les séquences HumanEva ainsi que sur une séquence réelle d'un système de vidé surveillance, montrent que la qualité du suivi est améliorée tout en réduisant le coût calculatoire.

Références

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *IEEE PAMI*, 28 :44–58, 2006.

[2] A. O. Balan and M. J. Black. An adaptive appearance model approach for model-based articulated object tracking. In *IEEE CVPR*, 2006.

[3] M. Boufarguine, M. Baklouti, F. Precioso, and V. Guitteny. Virtu4d : a realtime virtualization of reality. In *3DPVT*, 2010.

[4] M. A. Brubaker, D. J. Fleet, and A. Hertzmann. Physics-based person tracking using the anthropomorphic walker. *IJCV*, 87 :1–8, 2010.

[5] G. de Haan, J. Scheuer, R. de Vries, and F. Post. Egocentric navigation for video surveillance in 3d virtual environments. In *IEEE 3DUI*, 2009.

[6] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61 :185–205, 2005.

[7] A. Elgammal and C.-S. Lee. Tracking people on a torus. *IEEE PAMI*, 31 :520–538, 2009.

[8] A. Fathi and G. Mori. Human pose estimation using motion exemplars. In *IEEE ICCV*, 2007.

[9] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture. *International Journal of Computer Vision*, 87 :75–92, 2010. 10.1007/s11263-008-0173-1.

[10] A. Girgensohn, D. Kimber, J. Vaughan, T. Yang, F. Shipman, T. Turner, E. Rieffel, L. Wilcox, F. Chen, , and T. Dunnigan. Dots : support for effective video surveillance. In *ACM MM*, 2007.

[11] M. W. Green. Appropriate and effective use of security technologies in u.s. schools. Technical Report 97-IJ-R-072, National Institute of Justice, September 1999.

[12] N. R. Howe. Silhouette lookup for automatic pose tracking. In *IEEE CVPR*, 2004.

[13] R. Poppe. Vision-based human motion analysis : An overview. *CVIU*, 108 :4–18, 2007.

[14] T. J. Roberts, S. J. McKenna, and I. W. Ricketts. Human tracking using 3d surface colour distributions. *IVC*, 24(12) :1332–1342, 2006.

[15] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV*, 2000.

[16] L. Sigal, A. Balan, and M. Black. Humaneva : Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87 :4–27, 2010.

FIGURE 7 – Résultats du suivi pour la séquence S2/C1. Trois différentes méthodes sont comparées (figure 6) : fonction de vraisemblance basée seulement sur les silhouettes (première rangée) , fonction de vraisemblance mixte combinant silhouettes et modèle d'apparence (rangée centrale) et un suivi utilisant de plus le modèle de mouvement proposé (troisième rangée).

FIGURE 8 – Virtualisation 3D d'une scène dynamique filmée par une caméra de surveillance. Première ligne : des images extraites de la séquence. Deuxième rangée : résultat du suivi 3D projeté sur les images. 3^e et 4^e lignes : animation d'un modèle humanoïde réaliste moyennant le résultat du suivi.

[17] L. Sigal, A. Balan, and M. Black. HumanEva : Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1) :4–27, Mar. 2010.

[18] C. Sminchisescu. Consistency and coupling in human model likelihoods. In *IEEE FG*, 2002.

[19] C. Sminchisescu and A. Telea. Human pose estimation from silhouettes. a consistent approach using distance level sets. In *WSCG*, 2002.

[20] R. Urtasun and P. Fua. 3d human body tracking using deterministic temporal motion models. In *ECCV*, 2004.

[21] M. Vondrak, L. Sigal, and O. C. Jenkins. Physical simulation for probabilistic motion tracking. In *IEEE CVPR*, 2008.

[22] P. Wang and J. M. Rehg. A modular approach to the analysis and evaluation of particle filters for figure tracking. In *IEEE CVPR*, 2006.

[23] Y. Wang, D. Krum, E. Coelho, and D. Bowman. Contextualized videos : Combining videos with environment models to support situational understanding. In *IEEE Transactions on Visualization and Computer Graphics*, Oct. 2007.