Une stratégie hybride de filtrage particulaire pour le suivi de mouvement humain depuis un robot mobile

Mathias Fontmarty[1] Frédéric Lerasle[1,2] Patrick Danès[1,2]

[1] LAAS-CNRS, Université de Toulouse, 7 avenue du Colonel Roche, 31077 Toulouse Cédex 4
[2] Université de Toulouse, UPS, 118 route de Narbonne, 31062 Toulouse Cédex 9

prenom.nom@laas.fr

Résumé

Cet article présente un nouvel algorithme dédié au suivi tridimensionnel de mouvements humains à l'aide d'un système stéréoscopique embarqué sur un robot mobile. Cette approche combine les avantages de la stratégie ICONDENSATION classique et ceux du filtre particulaire à recuit simulé (Annealed Particle Filter) en un filtre que nous appelons "I-Annealed Particle Filter". Nous montrons que la fusion d'attributs visuels pertinents garantit une robustesse du suivi aux différents types d'environnement. Nous présentons une implantation complète du système de suivi ainsi que quelques résultats sur des séquences vidéo réalisées depuis une plate-forme robotique dédiée à l'interaction Homme-Robot. Enfin, quelques améliorations possibles sont discutées.

Mots Clef

Suivi visuel, filtrage particulaire, fusion de données, robotique mobile.

Abstract

This paper presents a new algorithm for human motion three-dimensional tracking based on a stereo camera system embedded on a mobile robot. The approach mixes advantages of the well-known ICONDENSATION strategy and of the Annealed particle filter into a more reliable "I-Annealed" particle filter based tracker. Data fusion is also studied to show that a wide variety of visual cues must be used so that the system can adapt to various backgrounds. We present a complete implementation of the proposed tracker as well as some results on indoor sequences. Finally, evolutions and future work are discussed.

Keywords

Visual tracking, particle filter, data fusion, mobile robotics.

1 Introduction et contexte

Un défi majeur de la Robotique aujourd'hui est sans doute celui du robot autonome personnel avec la perspective de voir un robot interagir à distance avec ses interlocuteurs par la perception de leurs attitudes ou la reconnaissance de gestes. L'objectif est donc de capturer les mouvements 3D humains à partir du flot vidéo issu d'une ou plusieurs caméras embarquées. Le suivi visuel depuis une plate-forme mobile autonome est une problématique complexe qui se démarque de la capture de mouvement depuis des caméras d'ambiance [6, 16, 17]. L'approche doit être à la fois : (i) peu consommatrice de ressources CPU pour ne pas compromettre l'exécution des autres fonctionnalités nécessaires à l'évolution autonome du robot, (ii) robuste aux mouvements *a priori* quelconques des membres corporels, aux conditions de prises de vues ainsi qu'aux occultations dues notamment au faible champ de perception du système extéroceptif embarqué. Il est alors opportun de gérer à chaque instant plusieurs hypothèses sur les paramètres à estimer et de doter le filtre de mécanismes d'initialisation et ré-initialisation automatiques.

A l'instar de nombreux travaux [6, 19, 11], nous modélisons le corps humain comme un ensemble de parties rigides dans leurs dimensions. La littérature distingue alors les approches basées sur une reconstruction 3D [19, 11] des approches qui synthétisent la projection du modèle 3D et ajustent les paramètres de localisation pour la faire correspondre à son apparence image [6, 16, 17]. Par le passé, nous avons privilégié une approche "appearance-based" par vision monoculaire [13], pour sa simplicité de mise en œuvre. Une difficulté majeure résidait alors dans l'estimation des mouvements non fronto-parallèles au plan image. Les développements actuels reprennent et étendent ces travaux à un système stéréoscopique. L'approche retenue se démarque de la classification énoncée au sens où elle repose sur des mesures d'apparence mais également sur des mesures/contraintes géométriques provenant d'une reconstruction 3D éparse.

Le filtrage particulaire [2] est adapté au contexte décrit précédemment. Il s'affranchit de toute hypothèse relative aux distributions mises en jeu. Un autre avantage est de pouvoir fusionner aisément les informations issues de différentes sources de mesures. Un inconvénient de taille ré-

side néanmoins dans la nécessité d'un nombre de particules croissant exponentiellement avec la dimension de l'espace d'état pour les applications considérées ici.
A l'instar de [6, 12], nous avons proposé et évalué dans [7] une stratégie permettant de limiter le nombre de particules par un échantillonnage partitionné de l'espace d'état. Une alternative, introduite dans [5], repose sur un filtre particulaire à recuit simulé, ou "Annealed Particle Filter" (APF), pour capturer les mouvements humains à partir de caméras d'ambiance et de connaissances *a priori* sur la scène statique. On notera que ces deux stratégies n'intègrent aucune capacité de ré-initialisation du filtre, indispensable dans un contexte Robotique.

Nous nous proposons donc d'améliorer la stratégie APF à deux niveaux. Tout d'abord, la fusion de différentes primitives visuelles permet une réduction progressive de l'espace de recherche. Ensuite, l'ajout de la faculté de (ré-)initialisation automatique autorise les pertes de cible temporaires. Alors que les filtres particulaires traditionnels ne peuvent rattraper ces décrochages de par la grande dimension de l'espace de recherche, la stratégie ICONDENSATION, initiée dans [9], utilise des détecteurs pouvant remédier à ce problème. À notre connaissance, cette stratégie n'a que très peu été mise en œuvre dans un contexte de suivi 3D, notamment par Giebel *et. al.* dans [8] qui utilisent une fonction d'importance 3D dans un contexte de détection grossière de piétons. Cette (ré)-initialisation s'appuie logiquement sur une détection préalable de membres corporels les plus discriminants. Si certaines parties du corps sont détectées, un algorithme de cinématique inverse peut être mis en place pour déterminer la position 3D de ces membres et ainsi réinitialiser le filtre. Nous proposons ici une nouvelle stratégie de filtrage, notée IAPF, qui combine les intérêts des stratégies ICONDENSATION et APF.

L'article est structuré comme suit. La section 2 rappelle sommairement l'algorithme APF et décrit la version modifiée permettant la fusion de données. La section 3 spécifie les mesures visuelles reposant sur des attributs 3D ou 2D. La section 4 décrit l'implémentation des différents filtres (APF, IAPF et ICONDENSATION). Notre filtre IAPF est ensuite évalué dans notre contexte applicatif et comparé avec les stratégies ICONDENSATION et APF dans la section 5. Enfin, la section 6 résume notre contribution et propose quelques extensions envisagées pour ces travaux.

2 Le filtre particulaire à recuit simulé pour la fusion de données

2.1 Principe de la stratégie APF

Tout comme les autres filtres particulaires, l'APF ("Annealed Particle Filter") est une méthode de Monte Carlo destinée à l'estimation récursive du vecteur d'état d'un système stochastique markovien. Son but est d'approximer la densité *a posteriori* $p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})$ du vecteur d'état $\boldsymbol{x}_k$ à l'instant k conditionnellement aux mesures $\boldsymbol{z}_{1:k} = \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$ par la distribution ponctuelle

$$p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta(\boldsymbol{x}_k - \boldsymbol{x}_k^{(i)}),\ \sum_{i=1}^{N} w_k^{(i)} = 1,$$

qui représente la sélection d'une hypothèse – ou particule – $\boldsymbol{x}_k^{(i)}$ avec la probabilité – ou poids – $w_k^{(i)}$, $i = 1, \ldots, N$. Le calcul des moments *a posteriori* de toute fonction de $\boldsymbol{x}_k$ s'en suit immédiatement, *e.g.* l'estimé du minimum d'erreur quadratique moyenne $\mathrm{E}[\boldsymbol{x}_k|\boldsymbol{z}_{1:k}]$.
Considérons un système d'état interne $\boldsymbol{x}_k$, dont les densités de dynamique et d'observation sont définies par $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})$ et $p(\boldsymbol{z}_k|\boldsymbol{x}_k)$. Le principe de l'APF, présenté dans le tableau 1, est de scinder la boucle principale de la CONDENSATION [2] en la séquence de L couches. Ainsi, pour chaque $l \in \{1, \ldots, L\}$, nous appliquons une "sous-dynamique" $p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1})$ aux particules résultant de la $(l-1)^{\text{ième}}$ couche afin de les faire évoluer vers des régions où une "sous-vraisemblance" $p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l})$ présente de grandes valeurs.
Il convient ensuite de sélectionner chacune de ces fonctions de manière judicieuse afin d'améliorer les résultats obtenus par la stratégie CONDENSATION pour les espaces d'état de grande dimension. Deutscher *et al.*, dans [5], proposent de choisir :

$$\begin{aligned} p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1}) &= \left[p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})^{\alpha_l}\right]_{\boldsymbol{x}_k=\boldsymbol{x}_{k,l}\,;\,\boldsymbol{x}_k=\boldsymbol{x}_{k,l-1}} \\ p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}) &= \left[p(\boldsymbol{z}_k|\boldsymbol{x}_k)^{\beta_l}\right]_{\boldsymbol{x}_k=\boldsymbol{x}_{k,l}} \end{aligned}$$

où $\alpha_l \in [1, +\infty]$ et $\beta_l \in [0, 1]$ sont des suites croissantes de paramètres. Ainsi, l'exploration de l'espace d'état s'effectue tout d'abord selon la fonction de dynamique ($\alpha_l = 1$), puis s'affine au fur et à mesure que l croît (auquel cas α croît, rendant ainsi les modes de $p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k-1,l})$ plus prononcés et plus étroits). Parallèlement, les coefficients β_l augmentent, et, alors que les premières couches présentent une fonction de vraisemblance très lissée (β_l petit), les dernières exploitent des fonctions très piquées (β_l proche de 1).
Notons qu'il faut être capable d'échantillonner suivant $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})^{\alpha_l}$, ce qui peut ne pas être trivial. Dans notre contexte, nous considérons une dynamique de type "marche aléatoire" $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) = \mathcal{N}(\boldsymbol{x}_{k-1}, \Delta_k)$, avec Δ_k diagonale. Échantillonner suivant $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})^{\alpha_l}$ revient alors à échantillonner selon $\mathcal{N}(\boldsymbol{x}_{k-1}, \frac{1}{\alpha_l}\Delta_k)$.

2.2 Description de notre stratégie APF modifié

L'extension de l'APF est inspirée de l'ICONDENSATION [9]. L'idée principale consiste à explorer l'espace d'état au moyen d'une fonction d'importance $q(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{z}_k)$ au lieu de l'unique dynamique $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})$ du système. L'APF commence ses traitements sur un nuage de particules qui est affiné à chaque couche de l'algorithme ; il convient donc d'introduire l'échantillonnage d'importance dans la première étape. L'algorithme complet est présenté tableau 2. L'utilisation de l'échantillonnage d'importance

$[\{\boldsymbol{x}_k^{(i)}, w_k^{(i)}\}]_{i=1}^N = APF([\{\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)},\}]_{i=1}^N, \boldsymbol{z}_k)$

1: **SI** $k = 0$**,** échantilloner $\boldsymbol{x}_0^{(1)}, \ldots, \boldsymbol{x}_0^{(i)}, \ldots, \boldsymbol{x}_0^{(N)}$ i.i.d. selon $p(\boldsymbol{x}_0)$, et poser $w_0^{(i)} = \frac{1}{N}$. Choisir $\alpha_1, \ldots, \alpha_L \in [1, +\infty]$, la suite croissante des exposants de la fonction de dynamique ainsi que $\beta_1, \ldots, \beta_L \in [0, 1]$, la suite croissante des exposants de la fonction de vraisemblance. **FIN SI**
2: **SI** $k \geq 1$ **ALORS** {—$[\{\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N$ est une approximation particulaire de $p(\boldsymbol{x}_{k-1}|\boldsymbol{z}_{1:k-1})$—}
3: Poser $[\{\boldsymbol{x}_{k,0}^{(i)}, w_{k,0}^{(i)}\}]_{i=1}^N = [\{\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N$
4: **POUR** $l = 1, \ldots, L$**, FAIRE**
5: **POUR** $i = 1, \ldots, N$**, FAIRE**
6: Propager la particule $\boldsymbol{x}_{k,l-1}^{(i)}$ au travers de la fonction de dynamique $\boldsymbol{x}_{k,l}^{(i)} \sim p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1}^{(i)})$ avec $p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1}^{(i)}) = p(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1}^{(i)})^{\alpha_l}$
7: Mettre à jour le poids $w_{k,l}^{(i)}$ associé à $\boldsymbol{x}_{k,l}^{(i)}$ *via* $w_{k,l}^{(i)} \propto w_{k,l-1}^{(i)} p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)})$ avec $p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)}) = p(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)})^{\beta_l}$
8: Normaliser les poids $w_{k,l}^{(i)}$ de telle sorte que $\sum_i w_{k,l}^{(i)} = 1$
9: Éventuellement, rééchantillonner le nuage de particules $[\{\boldsymbol{x}_{k,l}^{(i)}, w_{k,l}^{(i)}\}]_{i=1}^N$
10: **FIN POUR**
11: **FIN POUR**
12: Poser $[\{\boldsymbol{x}_k^{(i)}, w_k^{(i)}\}]_{i=1}^N = [\{\boldsymbol{x}_{k,L}^{(i)}, w_{k,L}^{(i)}\}]_{i=1}^N$
13: Calculer la moyenne *a posteriori* $\mathrm{E}[\boldsymbol{x}_k|\boldsymbol{z}_{1:k}]$ à partir de la représentation particulaire $\sum_{i=1}^N w_k^{(i)} \delta(\boldsymbol{x}_k - \boldsymbol{x}_k^{(i)})$ de $p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})$
14: **FIN SI**

TAB. 1 – Filtre particulaire APF.

permet l'initialisation automatique ou la ré-initialisation dans les cas de décrochage du filtre, ce qui est capital dans un contexte de robotique.

La fonction d'importance $q(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{z}_k)$ que l'on utilise dans notre version modifiée de l'APF met en jeu à la fois la mesure et la dynamique, à l'instar de l'ICONDENSATION.

Notons que, comme pour l'APF, dont les auteurs précisent que "son unique inconvénient est de ne pas opérer dans un contexte bayésien rigoureux" [5], la stratégie IAPF souffre d'un manque de rigueur mathématique.

3 Description des indices visuels

Par le positionnement des particules conformément à des mesures image discriminantes mais possiblement intermittentes dans le flot vidéo, la fonction d'importance $q(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{z}_k)$ permet de ré-initialiser le filtre. La mise à jour s'effectue par le modèle de mesure ; ces mesures sont ici persistantes mais hélas pas toujours suffisamment discriminantes en présence d'environnements encombrés. Il est alors judicieux de fusionner des mesures 2D et 3D dans le filtre et ainsi apporter des contraintes de différentes natures. A notre connaissance, peu de travaux [3] combinent simultanément des informations d'apparence et géométriques dans leurs filtres. De plus, la fusion de mesures persistantes dans une fonction de mesure unifiée permet de suppléer le faible pouvoir discriminant d'une seule mesure pour un contexte environnemental donné. Nous spécifions ci-après la fonction d'importance $q(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{z}_k)$ et les mesures utilisées par la fonction de vraisemblance $p(\boldsymbol{z}_k|\boldsymbol{x}_k)$.

3.1 Fonction d'importance

La fonction d'importance que nous avons élaboré place une proportion α des particules suivant la dynamique, une proportion β suivant la mesure, et le reste (dans le cas où $\alpha + \beta < 1$) suivant une connaissance $p_0(\boldsymbol{x})$ que l'on se donne *a priori*. Les particules échantillonnées selon la mesure sont tirées suivant une gaussienne centrée sur une configuration $\boldsymbol{x}_k^D$ calculée à partir des positions 3D de la tête et des mains du sujet suivi grâce à un algorithme de cinématique inverse analytique. Les positions 2D de la tête et des mains sont extraites de chaque image par segmentation de blobs de couleur peau. Les blobs sont ensuite mis en correspondance entre les 2 plans images suivant des critères définis dans [15, Chap. 4]. Enfin, les centres des régions sont triangulés en exploitant les paramètres de calibration du banc stéréo. Notre système considère ainsi la tête et les mains comme trois marqueurs "naturels".
Échantillonner $\boldsymbol{x}$ suivant $q(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{z}_k)$ revient alors à tirer $u \sim \mathcal{U}(0, 1)$ puis à échantillonner :
– $\boldsymbol{x} \sim p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})$ si $u < \alpha$
– $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}_k^D, \Delta_k)$ si $\alpha \leq u < \alpha + \beta$
– $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}_0, \Delta_k)$ si $u \geq \alpha + \beta$
où Δ_k est une matrice de covariance. Dans notre cas, il s'agit de la covariance de la gaussienne utilisée pour modéliser la dynamique *a priori* de type "marche aléatoire".

3.2 Fonction de vraisemblance

Distance aux contours. Cette distance nécessite la projection du modèle 3D avec gestion des parties cachées [4]. La vraisemblance vis-à-vis de la forme est traditionnellement calculée par la somme des carrés des distances entre des points situés sur les limbes du modèle projeté et les contours les plus proches dans l'image. Les N_p points de

$[\{\boldsymbol{x}_k^{(i)}, w_k^{(i)}\}]_{i=1}^N = IAPF([\{\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)},\}]_{i=1}^N, \boldsymbol{z}_k)$

SI $k = 0$**,** échantillonner $\boldsymbol{x}_0^{(1)}, \ldots, \boldsymbol{x}_0^{(i)}, \ldots, \boldsymbol{x}_0^{(N)}$ i.i.d. selon $p(\boldsymbol{x}_0)$, et poser $w_0^{(i)} = \frac{1}{N}$. Choisir $\alpha_1, \ldots, \alpha_L \in [1, +\infty]$, la suite croissante des exposants de la fonction de dynamique ainsi que $\beta_1, \ldots, \beta_L \in [0, 1]$, la suite croissante des exposants de la fonction de vraisemblance. **FIN SI**
SI $k \geq 1$ **ALORS** {—$[\{\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N$ est une approximation particulaire de $p(\boldsymbol{x}_{k-1}|\boldsymbol{z}_{1:k-1})$—}
  Poser $[\{\boldsymbol{x}_{k,0}^{(i)}, w_{k,0}^{(i)}\}]_{i=1}^N = [\{\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N$
  POUR $l = 1, \ldots, L$**, FAIRE**
    SI $l == 1$ **ALORS**
      POUR $i = 1, \ldots, N$**, FAIRE**
        Propager la particule $\boldsymbol{x}_{k,l-1}^{(i)}$ au travers de la fonction d'importance $\boldsymbol{x}_{k,l}^{(i)} \sim q(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1}^{(i)}, \boldsymbol{z}_k)$
        Mettre à jour le poids $w_{k,l}^{(i)}$ associé à $\boldsymbol{x}_{k,l}^{(i)}$ *via* $w_{k,l}^{(i)} \propto w_{k,l-1}^{(i)} \frac{p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)}) p_l(\boldsymbol{x}_{k,l}^{(i)}|\boldsymbol{x}_{k,l-1}^{(i)})}{q(\boldsymbol{x}_{k,l}^{(i)}|\boldsymbol{x}_{k,l-1}^{(i)}, \boldsymbol{z}_k)}$ avec $p_l(\boldsymbol{x}_{k,l}^{(i)}|\boldsymbol{x}_{k,l-1}^{(i)}) = p(\boldsymbol{x}_{k,l}^{(i)}|\boldsymbol{x}_{k,l-1}^{(i)})^{\alpha_l}$ et $p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)}) = p(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)})^{\beta_l}$
        Normaliser les poids $w_{k,l}^{(i)}$ de telle sorte que $\sum_i w_{k,l}^{(i)} = 1$
        Éventuellement, rééchantillonner le nuage de particules $[\{\boldsymbol{x}_{k,l}^{(i)}, w_{k,l}^{(i)}\}]_{i=1}^N$
      FIN POUR
    SINON
      POUR $i = 1, \ldots, N$**, FAIRE**
        Propager la particule $\boldsymbol{x}_{k,l-1}^{(i)}$ au travers de la fonction de dynamique $\boldsymbol{x}_{k,l}^{(i)} \sim p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1}^{(i)})$ avec $p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1}^{(i)}) = p(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1}^{(i)})^{\alpha_l}$
        Mettre à jour le poids $w_{k,l}^{(i)}$ associé à $\boldsymbol{x}_{k,l}^{(i)}$ *via* $w_{k,l}^{(i)} \propto w_{k,l-1}^{(i)} p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)})$ avec $p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)}) = p(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)})^{\beta_l}$
        Normaliser les poids $w_{k,l}^{(i)}$ de telle sorte que $\sum_i w_{k,l}^{(i)} = 1$
        Éventuellement, rééchantillonner le nuage de particules $[\{\boldsymbol{x}_{k,l}^{(i)}, w_{k,l}^{(i)}\}]_{i=1}^N$
      FIN POUR
    FIN SI
  FIN POUR
  Poser $[\{\boldsymbol{x}_k^{(i)}, w_k^{(i)}\}]_{i=1}^N = [\{\boldsymbol{x}_{k,L}^{(i)}, w_{k,L}^{(i)}\}]_{i=1}^N$
  Calculer la moyenne *a posteriori* $\mathrm{E}[\boldsymbol{x}_k|\boldsymbol{z}_{1:k}]$ à partir de la représentation particulaire $\sum_{i=1}^N w_k^{(i)} \delta(\boldsymbol{x}_k - \boldsymbol{x}_k^{(i)})$ de $p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})$
FIN SI

TAB. 2 – Filtre particulaire IAPF.

contours $p_i, i \in \{1, \ldots, N_p\}$ du modèle pour une configuration $\boldsymbol{x}_k$ sont répartis uniformément le long des segments projetés. Dans notre implantation, l'image de contours est transformée en image de distance, notée I_{DT}, dans laquelle sont lues les valeurs de distance aux contours [8]. L'utilisation d'une image de distance permet un gain de temps de calcul pour un grand nombre de particules, puisque l'image n'est calculée qu'une fois, alors qu'un calcul de distance pour chaque particule alourdirait le traitement.
La vraisemblance est alors donnée par :

$$p(\boldsymbol{z}_k^e|\boldsymbol{x}_k) \propto \exp\left(-\frac{D^2}{2\sigma_e^2}\right),\ D = \frac{1}{N_p}\sum_{i=1}^{N_p} I_{DT}(p_i),$$

où i parcourt les N_p points du modèle, $I_{DT}(p_i)$ est la valeur associée de l'image de distance, et σ_e est l'écart-type *a priori* de notre modèle de mesure gaussien. Cette mesure, bien que persistante, reste très sensible aux conditions d'éclairement *a priori* quelconques et surtout peu discriminantes en présence de scènes encombrées.

Distance aux histogrammes couleurs. La couleur des vêtements portés permet généralement de distinguer les différents membres de la personne observée (manches, jambes, pieds, ...). L'utilisation de patches de couleur décrits par des histogrammes permet de tirer partie de cette information. Sur le modèle, N_{ROI} zones d'intérêts sont définies, auxquelles on associe une distribution de couleur de référence. Après projection, la distance aux histogrammes de référence pour une configuration $\boldsymbol{x}_k$ donnée est alors décrite par :

$$p(\boldsymbol{z}_k^{ROI}|\boldsymbol{x}_k) \propto \exp\left(-\frac{D^2}{2\sigma_{ROI}^2}\right),$$

$$D = \frac{1}{N_{ROI}}\sum_{i=1}^{N_{ROI}} (D_B(h_{\boldsymbol{x}_k,i}, h_{ref,i}))$$

où D_B est la distance de Bhattacharyya [1] utilisée pour comparer les histogrammes normalisés $(h_{ref,i}, h_{\boldsymbol{x}_k,i})$. Les histogrammes de référence $h_{ref,i}$ sont appris sur la première image de la séquence.

Distance aux blobs 3D. Dans l'esprit de la fonction d'importance $q(.)$, cette mesure implique la position 3D $\widehat{P}_j = (X_j, Y_j, Z_j)'$ de la tête et des mains ($j \in \{1, 2, 3\}$) après triangulation. On pose :

$$p(\boldsymbol{z}_k^{3d}|\boldsymbol{x}_k) \propto \exp\left(-\frac{D^2}{2\sigma_{3d}^2}\right), D = \frac{1}{3}\sum_{i=1}^{3} D_E(P_{\boldsymbol{x}_k,i}, \widehat{P}_{j_i}),$$

où $D_E(P_{\boldsymbol{x}_k,i}, \widehat{P}_{j_i})$ est la distance euclidienne entre le barycentre $\widehat{P}_{j_i}$ du blob j_i ($j_i \in \{1, \ldots, N_{Blob}\}$) et $P_{\boldsymbol{x}_k,i}$ la position 3D d'une main ou de la tête sous l'hypothèse $\boldsymbol{x}_k$. La correspondance entre i et j_i est établie par le biais d'heuristiques simples mettant en jeu la position 3D des blobs et l'utilisation d'un détecteur de visage [18].

Distance à la couleur peau. Dans certains cas, nous sommes dans l'impossibilité de trianguler la position de la tête et des mains (nombre de blobs détectés insuffisant, erreur de triangulation trop grande, ...). Toutefois, nous pouvons encore exploiter l'image segmentée de probabilité de couleur peau I_S et ainsi définir une autre vraisemblance basée sur la couleur. Pour une hypothèse $\boldsymbol{x}_k$ donnée, les coordonnées 2D $p_{\boldsymbol{x}_k,i}, i \in \{1, 2, 3\}$ de la tête et des mains après projection du modèle sont censées se situer dans une zone de forte probabilité de couleur peau :

$$p(\boldsymbol{z}_k^{s}|\boldsymbol{x}_k) \propto \exp\left(-\frac{D^2}{2\sigma_{s}^2}\right), D = \frac{1}{3}\sum_{i=1}^{3}(1 - I_S(p_{\boldsymbol{x}_k,i}))$$

Distance à une couleur homogène. Cette mesure fait intervenir N_m ensembles disjoints $E_i, i \in \{1, \ldots, N_m\}$ de points uniformément répartis à l'intérieur de chacun des N_m membres du modèle projeté pour une configuration $\boldsymbol{x}_k$.
Nous supposons ici que la personne porte des vêtements présentant une couleur homogène sur chacun des membres. Nous mettons alors en œuvre la mesure suivante :

$$p(\boldsymbol{z}_k^{m}|\boldsymbol{x}_k) \propto \exp\left(-\frac{D^2}{2\sigma_{m}^2}\right),$$

$$D = \frac{1}{N_m}\sum_{i=1}^{N_m}\left(\frac{1}{3}\sum_{c\in\{R,G,B\}}\sigma_{E_i,c}\right),$$

avec $\sigma_{E_i,c}$ l'écart-type de la distribution de couleur sur le canal $c \in \{R, G, B\}$ associée à l'ensemble de points E_i du membre i.

3.3 Etude des fonctions de mesure

On suppose que les mesures précédentes sont indépendantes entre elles conditionnellement à l'état. La fonction de mesure unifiée peut alors s'écrire :

$$\begin{aligned} p(\boldsymbol{z}_k|\boldsymbol{x}_k) &= p(\boldsymbol{z}_k^{e}, \boldsymbol{z}_k^{ROI}, \boldsymbol{z}_k^{3d}, \boldsymbol{z}_k^{s}, \boldsymbol{z}_k^{m}|\boldsymbol{x}_k) \\ &= p(\boldsymbol{z}_k^{e}|\boldsymbol{x}_k) \times p(\boldsymbol{z}_k^{ROI}|\boldsymbol{x}_k) \times p(\boldsymbol{z}_k^{3d}|\boldsymbol{x}_k) \\ &\quad \times p(\boldsymbol{z}_k^{s}|\boldsymbol{x}_k) \times p(\boldsymbol{z}_k^{m}|\boldsymbol{x}_k) \end{aligned}$$

La figure 1 expose l'évolution des distances obtenues en balayant un sous-espace de l'espace des configurations formé par l'orientation du bras droit sur des arrières-plans dégagé et encombré. Ces graphiques mettent en évidence le fait que les mesures "basées apparence" sont moins discriminantes que celle exploitant une information 3D (figure 1 (d)). Sur fond encombré, la distance aux contours (b) n'est pas suffisamment informative (de par la présence de nombreux *minima* locaux) alors que l'utilisation d'histogrammes de couleurs (c) s'avère plus robuste à ce type de scène, mais reste cependant très sensible aux changement de luminosité. La couleur peau (e) est un attribut discriminant sous réserve que l'arrière-plan n'inclut pas de couleur chair. La distance à une couleur uniforme (f) affiche de bons résultats sur fond encombré, sous les hypothèses justifiant son utilisation ("t-shirt" de couleur uniforme), mais reste très peu discriminante dans les autres cas.

Signalons qu'aucune des fonctions de mesure précédentes ne met en jeu un attribut de mouvement, ce qui est cohérent avec notre contexte applicatif où les déplacements du robot induisent une mobilité de l'arrière-plan dans l'image.

4 Implantation du système de suivi

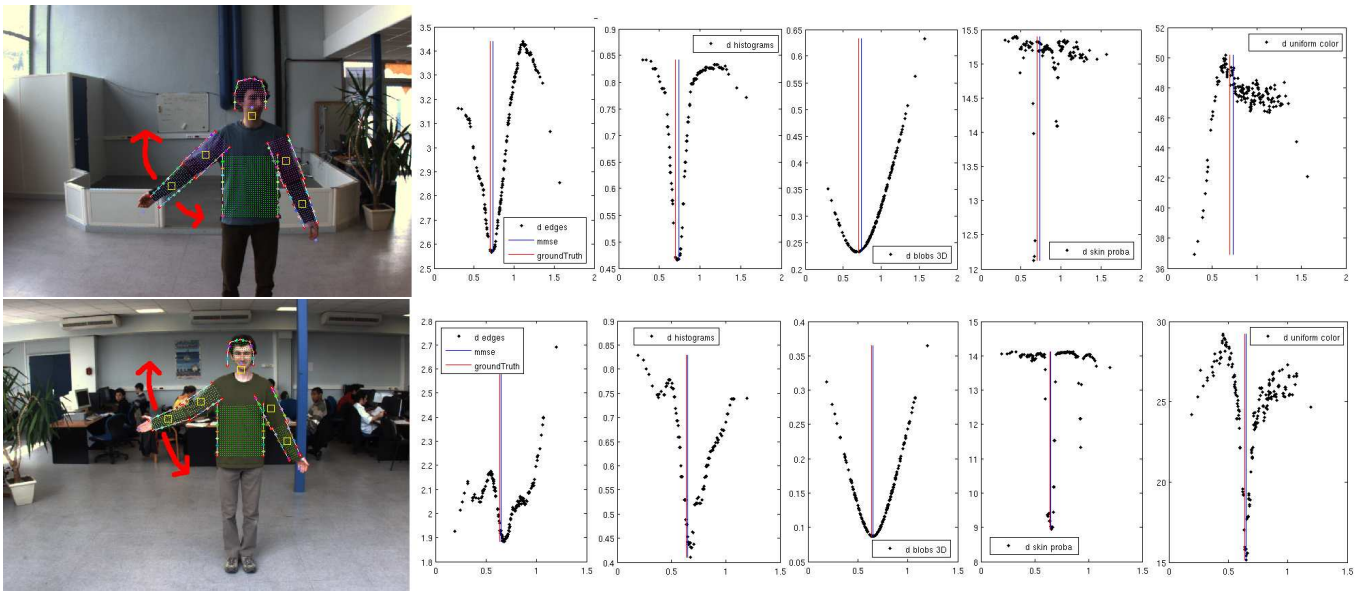
4.1 Modèle cinématique et géométrique des membres corporels

Notre modèle 3D articulé est constitué de cônes tronqués. Ces primitives géométriques restent assez simples à manipuler tandis que la géométrie projective en permet (de façon élégante) la projection image et la gestion des occultations. Pour la visualisation 3D, nous utilisons néanmoins des parallélépipèdes ce qui permet de mieux appréhender les degrés de liberté (DDL) de type rotation. Le modèle complet présente 22 DDL, que nous limitons pour l'instant aux 14 DDL des membres corporels supérieurs :
- 6 pour la localisation globale (3 pour la position et 3 pour l'orientation du torse),
- 3 rotations pour chaque épaule
- 1 rotation pour chaque coude.

La tête est supposée rigidement liée au torse car l'estimation de cette rotation impliquerait de caractériser l'orientation de la tête.

4.2 Architecture

L'architecture de notre "tracker" est présentée figure 2. Elle se décompose en cinq modules :
- Le module de prétraitement des permet l'amélioration des images acquises. Celles-ci sont issues de deux caméras stéréo couleurs mono-CCD firewire, puis sont converties en image RGB par "débayérisation". Une balance des blancs est appliquée à l'image puis celle-ci est ré-échantillonnée pour obtenir la résolution souhaitée.
- Le module de traitement construit différentes images qui seront exploitées par les fonctions d'importance et de mesure. La première est l'image de distance I_{DT}, obtenue en appliquant une transformée en distance (euclidienne) sur une image de contours classiquement calculée à l'aide d'un filtre de Canny. Le module fournit également une image de probabilité de couleur peau I_S



(a) (b) (c) (d) (e) (f)

FIG. 1 – Évolution des distances par rapport à la position d'un bras à 1 DDL sur des arrières-plans dégagé ou encombré. Les figures (b), (c), (d), (e) et (f) présentent respectivement les distances aux contours, aux histogrammes, aux blobs 3D, à la couleur peau, et à une couleur uniforme. Les lignes rouges et bleues représentent respectivement la vérité terrain et l'estimé du MMSE fourni par le filtre.

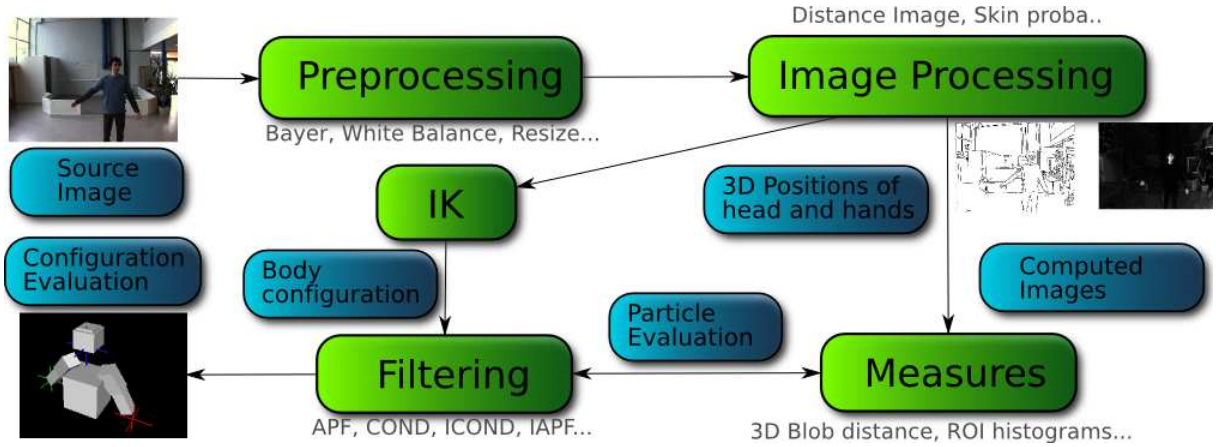


FIG. 2 – Architecture logicielle de notre système de suivi.

construite par rétroprojection d'un histogramme de référence de couleur peau appris précédemment sur une base de données [10]. Elle sera utilisée pour la localisation des blobs 2D préalablement à la triangulation. Un détecteur de visage est également implanté parmi d'autres fonctionnalités.

– Le module de filtrage propose plusieurs schémas classiques de filtres particulaires, parmi lesquels les stratégies CONDENSATION, ICONDENSATION, APF et notre variante notée ici IAPF. Les filtres faisant appel à une fonction d'importance utilisent les données produites par l'algorithme de cinématique inverse analytique pour générer des hypothèses à partir de cette mesure. Ce module estime le vecteur d'état relatif aux 14 DDL pré-cités.
– Le module de mesure calcule, pour une stratégie de filtrage donnée, les vraisemblances relatives aux particules. Ces vraisemblances, caractérisées en section 3.2, sont calculées sur la base des images générées par le module de traitement.
– Le module de cinématique inverse analytique calcule une configuration 3D du modèle articulé à partir des positions 3D de la tête et des mains lorsque les régions images associées sont segmentées par le module de traitement. La configuration renvoyée est celle qui est la plus proche de la configuration de repos (*i.e.* la plus naturelle). La configuration calculée ne se veut en aucun cas exacte — les positions de la tête et des mains n'étant pas elles-même très précises. Le but est ici de fournir une configuration approchée, mais rapidement calculable, de manière à permettre au filtre, *via* la fonction d'importance, de positionner les particules dans les zones pertinentes l'espace d'état.

5 Évaluations et discussion

Notre stratégie de filtrage IAPF a été testée sur plusieurs séquences acquises depuis le robot. Ces séquences présentent des occlusions temporaires, des sauts dans la dynamique de la cible, et de fortes variabilités environnementales. Quelques vidéos sont accessibles à l'URL `www.laas/~mfontmar`. Actuellement, le "tracker" opère avec une fréquence variant de $1\ Hz$ à $4\ Hz$ sur un Pentium IV Centrino $1.8\ GHz$ en fonction des types de filtres choisis, des mesures impliquées ou du nombre de particules. L'utilisation d'une fonction d'importance permet de réduire la taille de l'espace de recherche, et ainsi de restreindre le nombre de particules entre $[200; 1000]$. Les filtres APF et IAPF utilisent 3 couches.

5.1 Temps de calcul

Pour les évaluations présentées ici, le filtre ICONDENSATION considère N particules tandis que les filtres APF et IAPF s'appuient sur une stratégie hiérarchique exploitant N_L couches et N/N_L particules. Ce choix s'explique par les ressources CPU embarquées donc limitées : toutes les stratégies de filtrage s'exécutent à chaque instant image, avec la même ressource CPU. Nous comparons donc des résultats pour un même temps de traitement. Ainsi, à précision équivalente de l'estimé, la stratégie IAPF requiert moins de ressources que la stratégie ICONDENSATION.

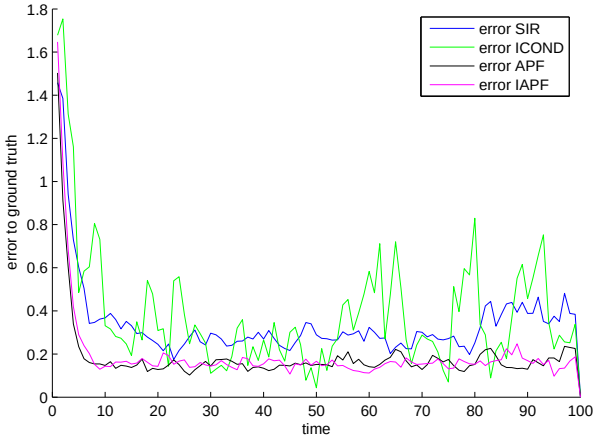


FIG. 4 – Erreur entre l'état réel et les estimés fournis par différentes stratégies de filtrage : SIR, ICONDENSATION, APF, et IAPF.

5.2 Précision

Le contexte applicatif ne permet pas de disposer de réelle vérité terrain. Aussi, nous avons tout d'abord évalué les différents filtres sur des données synthétiques. Les résultats associés sont montrés sur la figure 4. 15 réalisations de suivi par stratégie de filtrage sont effectuées à partir d'un système linéaire gaussien utilisant une dynamique de type "marche aléatoire". Pour chaque filtre, des statistiques sur la précision sont alors calculées *via* la distance moyenne entre les estimés fournis par chaque filtre et l'état réel pour un espace d'état de dimension 10. Nous observons que notre filtre IAPF donne un comportement proche du filtre APF, mais une meilleure précision que les filtres CONDENSATION ou ICONDENSATION. Ainsi, le filtre IAPF réduit l'erreur de $70\ \%$ par rapport au filtre ICONDENSATION.

Sur des flux vidéo réels, notre filtre IAPF donne des résultats au moins équivalents à ceux du filtre ICONDENSATION. Ici encore, 15 réalisations de suivi sont effectuées par séquence afin de caractériser le comportement moyen de chaque filtre. La figure 3 présente l'évolution de la distance entre une vérité terrain obtenue manuellement et les estimés calculés par les stratégies ICONDENSATION et IAPF sur une séquence vidéo représentative de leur comportement moyen. On note que l'erreur sur les rotations estimées diminue (15° au maximum). Ces dernières observations restent cependant à entériner par l'utilisation d'un système de capture de mouvement (du type VICON par exemple) qui permettra de définir une vérité terrain plus précise et fiable.

FIG. 3 – Déroulement des stratégies ICONDENSATION et IAPF sur deux séquences (haut et bas). Les graphiques de gauche présentent l'erreur quadratique (carré de la distance euclidienne) entre l'estimé délivré par chaque filtre et la vérité terrain. Pour chaque graphe, les séquences de droites montrent les résultats fournis par chaque stratégie.

FIG. 5 – De gauche à droite et de haut en bas : initialisation automatique du “tracker” par une configuration par défaut. Divergence du filtre. Accrochage du filtre sur la cible après détection et repositionnement adapté des particules.

5.3 Robustesse

Une propriété intéressante de notre filtre IAPF comparativement au filtre APF, pour des espaces d’état de grande dimension, est sa capacité à s’auto-(ré)initialiser, et donc à “raccrocher” la cible après des pertes temporaires dans un contexte applicatif complexe. L’initialisation du filtre est donc automatique, la détection 3D de la tête et des mains permettant de déduire la configuration initiale du modèle 3D articulé. L’initialisation du filtre requiert aussi la caractérisation automatique des modèles d’histogrammes de référence sur la première image de la séquence. L’initialisation automatique du “tracker” est illustrée par la séquence de la figure 5. Rappelons ici qu’aucune connaissance *a priori* sur la position et l’apparence de la personne n’est utilisée dans le processus.

5.4 Considérations pratiques

La résolution actuelle des images traitées est de 640×480 pixels. Une grande partie du temps de calcul – environ $200\ ms$ – est consacrée aux prétraitements et traitements de l’image (“débayérisation”, calculs des images de probabilité peau, de distance, ...) qui peuvent être encore optimisés. En outre, la fréquence de traitement peut être largement accrue en diminuant la résolution des images. L’ensemble des paramètres utilisés pour notre filtre IAPF est listé dans le tableau 3. Les valeurs sont fixées empiriquement ou par l’utilisation d’heuristiques simples, notamment pour les coefficients α_l et β_l, ainsi que pour les écarts-types $(\sigma_e, \sigma_{ROI}, \sigma_{3D}, \sigma_s, \sigma_m)$ fixés *a priori*.

6 Conclusion

Cet article présente une méthode de capture de mouvement des membres corporels supérieurs à partir d’une tête stéréoscopique embarquée sur un robot mobile. Deux lignes directrices ont motivé ces travaux.
Tout d’abord, nous avons combiné les avantages des stratégies ICONDENSATION et APF dans une nouvelle stratégie de filtrage notée IAPF. Celle-ci permet, par repositionnement des particules, de ré-initialiser le filtre après des décrochages inhérents à des occultations persistantes et/ou des fausses mesures. En outre, cette stratégie s’avère plus précise que la stratégie ICONDENSATION pour des temps de calcul équivalents.
Ensuite, la fusion d’informations visuelles hétérogènes, ici 2D et 3D, permet à notre système d’être plus robuste aux environnements *a priori* encombrés et évolutifs rencontrés par notre robot mobile. La maîtrise accrue des techniques de vision corrélée aux ressources CPU sans cesse croissantes des plate-formes autonomes rendent possible aujourd’hui la capture de mouvement à partir de capteurs vision embarqués. L’intégration de cette fonctionnalité est un enjeu majeur en Robotique car elle sous-tend des tâches primordiales telles que la manipulation et le déplacement en présence de l’homme ou l’interprétation de gestes ou activités.

Les résultats obtenus sont prometteurs, mais des investigations en termes d’implémentation et d’évaluation restent à poursuivre. Tout d’abord, nous pensons étendre nos détecteurs à des membres corporels autres que les mains/tête, typiquement par des techniques de “boosting” [16, 14] et non une segmentation peau qui reste très sensible aux conditions ambiantes d’illumination. Ensuite, nous pensons combiner plus largement les mesures 3D avec les mesures d’apparence en exploitant un processus de stéréo-corrélation éparse.

Références

[1] F. Aherne, N. Thacker, and P. Rockett. The bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, 32(4) :1–7, 1997.

[2] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *Transactions on Signal Processing*, 2(50) :174–188, 2002.

[3] P. Azad, A. Ude, T. Asfour, and R. Dillmann. Stereo-based markerless human motion capture for humanoid robot systems. In *International Conference on Robotics and Automation (ICRA’07)*, pages 3951–3956, Roma, Italy, 2007.

[4] P. Azad, A. Ude, R. Dillmann, and G. Cheng. A full body human motion capture system using particle filtering and on-the-fly edge detection. In *4th IEEE/RAS International Conference on Humanoid Robots (ICHR’O4)*, volume 2, pages 941–959, California, USA, November 2004.

Nom	Description	Valeur
(W, H)	résolution image	$(640, 480)$
N	nombre de particules	600
N_{layers}	nombre de couches des stratégies APF et IAPF	3
$(\alpha_1, \alpha_2, \alpha_3)$	exposants de la dynamique	$(1, 4, 25)$
$(\beta_1, \beta_2, \beta_3)$	exposants de la vraisemblance	$(0.1, 0.4, 1)$
(α, β)	paramètres de la fonction d'importance $q(.)$	$(0.8, 0.2)$
$(\sigma_e, \sigma_{ROI}, \sigma_{3D}, \sigma_s, \sigma_m)$	paramètres de la fonction de vraisemblance	(10, 0.3, 0.15, 0.02, 10)
$p(\boldsymbol{x}_k \vert \boldsymbol{x}_{k-1})$	dynamique du système	$\mathcal{N}(\boldsymbol{x}_{k-1}, \Delta_k)$
Δ_k	covariance de la dynamique	$Diag(\delta_i), i \in \{1..14\}$
$\delta_i, i \in \{1..3\}$	bruit sur la position globale	$0,07$
$\delta_i, i \in \{11, 12\}$	bruit sur les rotations dans l'axe épaule-coude	$0,3$
$\delta_i, i \in \{4..10, 13, 14\}$	bruit sur toutes les autres rotations	$0,1$

TAB. 3 – Valeurs des paramètres utilisés dans notre filtre IAPF.

[5] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *International Conference on Computer Vision and Pattern Recognition (CVPR'00)*, volume 2, pages 126–133, Hilton Head Island, South Carolina, USA, 2000.

[6] J. Deutscher, A. Davison, and I. Reid. Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. In *International Conference on Computer Vision and Pattern Recognition (CVPR'01)*, pages 669–676, Kauaii Marriott, Hawaii, USA, 2001.

[7] M. Fontmarty, F. Lerasle, P. Danès, and P. Menezes. Filtrage particulaire pour la capture de mouvement dediée à l'interaction homme-robot. In *11ème congrès francophone des jeunes chercheurs en vision par ordinateur (ORASIS'07)*, Obernai, June 2007.

[8] J. Giebel, D. M. Gavrila, and C. Schnorr. A bayesian framework for multi-cue 3D object. In *European Conference on Computer Vision (ECCV'04)*, Prague, 2004.

[9] M. Isard and A. Blake. I-CONDENSATION : Unifying low-level and high-level tracking in a stochastic framework. In *European Conference on Computer Vision (ECCV'98)*, pages 893–908, Freiburg, Germany, 1998.

[10] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1) :81–96, 2002.

[11] S. Knoop, S. Vacek, and R. Dillman. Sensor fusion for 3D human body tracking with an articulated3D body model. In *International Conference on Robotics and Automation (ICRA'06)*, pages 1686–1691, Orlando (USA), May 2006.

[12] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *European Conference on Computer Vision (ECCV'00)*, pages 3–19, Dublin, Ireland, 2000.

[13] P. Menezes, F. Lerasle, and J. Dias. Visual tracking modalities for a companion robot. In *International Conference on Intelligent Robots and Systems (IROS'06)*, Beijing, China, 2006.

[14] D. Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. In *International Conference on Computer Vision and Pattern Recognition (CVPR'03)*, pages 467–474, Madison, USA, June 2003.

[15] C. Schmid. *Appariement d'images par invariants locaux de niveaux de gris*. PhD thesis, Institut National Polytechnique de Grenoble, 1996.

[16] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *International Conference on Computer Vision and Pattern Recognition (CVPR'04)*, pages 421–428, Washington, DC, USA, 2004.

[17] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal on Robotic Research*, 6(22) :371–393, May 2003.

[18] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *International Conference on Computer Vision and Pattern Recognition (CVPR'01)*, Kauaii Marriott, Hawaii, USA, 2001.

[19] J. Ziegler, K. Nickel, and R. Stiefehagen. Tracking of the articulated upper body on multi-view stereo image sequences. In *International Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 774–781, New York, USA, 2006.