Vision et filtrage particulaire pour le suivi tridimensionnel de mouvements humains: applications à la robotique

Mathias Fontmarty

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse III - Paul Sabatier
Discipline ou spécialité : Automatique

Présentée par Mathias FONTMARTY
Le 2 décembre 2008

Titre : Vision et filtrage particulaire pour le suivi tridimensionnel de mouvement humain. Applications à la Robotique

---

JURY
Marie-Odile Berger, Rapporteur
Patrick Pérez, Rapporteur
Jenny Benois Pineau, Examinateur
Patrice Dalle, Examinateur
Patrick Danès, Directeur de thèse
Frédéric Lerasle, Directeur de thèse
Michel Devy, Invité

École doctorale : EDSYS
Unité de recherche : LAAS-CNRS (Groupe RAP)
Directeur(s) de thèse : Patrick Danès, Frédéric Lerasle
Rapporteurs : Marie-Odile Berger, Patrick Pérez

Table des matières

Remerciements

Merci à Malik Ghallab et Raja Chatila, directeurs du LAAS-CNRS, de m'avoir accueilli au sein du laboratoire pour ces 3 ans de thèse.

Merci ensuite à Michel Devy, directeur du groupe de recherche RAP, de m'avoir permis de rejoindre son équipe, dans un premier temps pour un stage de M2R, puis par la suite pour ces travaux plus en profondeur sur la capture de mouvement humain.

Merci également aux rapporteurs et examinateurs de cette thèse pour l'avoir (re)lue, et être venus assister à la soutenance malgré leur emploi du temps très chargé.

Merci à Frédéric et Patrick surtout, pour leur encadrement omniprésent et de qualité quoi qu'ils en disent, mais aussi pour leur bonne humeur systématique, leur sympathie et leurs encouragements qui ont fait que cette thèse a été menée à son terme dans les meilleures conditions possibles. Merci également pour les nuits blanches passées à m'épauler pour finir les articles dans les temps... même le week end... et pardon pour avoir dégradé notoirement leur vie familiale. Merci aussi d'avoir été parmi les premières victimes à acheter notre album !

Merci toujours à tous mes collègues du laboratoire, doctorants, stagiaires, pour la bonne humeur qu'ils contribuent à répandre au travail. Il n'en est que plus agréable de prendre le vélo sous la pluie le matin !

Merci à Pier et leGroupe pour les aventures musicales qui changent les idées et qui sont toujours aussi riches en émotions. Dans le même ordre de choses, merci également aux collègues musiciens de tous bords : rock and roll d'Elvis aux pays des merveilles, variété de Graine de Sel, musique du cœur de Enfoiros de l'INSA, celtique du 7 de Trèfle, vocalises d'Agato, groove de Motown Time, et j'en oublie c'est sûr.

Merci aux copains pour leur soutien, leur bonne humeur, et parfois leur curiosité scientifique sur mon sujet pour le moins... ludique.

Merci aussi à mon père et ma mère qui m'ont épaulé et encouragé tout au long de ces 3 années de thèse, mais aussi durant toute ma scolarité, et, d'une manière plus générale, pour tous les projets dans lesquels je me suis impliqué d'une manière ou d'une autre. Je leur dois bien plus que je ne saurai jamais l'exprimer. Pour leur patience, je remercie également mes grands-parents et la famille que je ne vois malheureusement pas aussi souvent que je le souhaiterais...

Merci à mon frère, pour sa bonne humeur, son soutien et tous ces moments à parler de tout et de rien, nos rêves, nos souvenirs, nos délires,...

Et merci enfin à Anneline qui jusqu'au bout a réussi à me supporter... dans tous les sens du terme ! Et ce malgré les horaires de travail chaotiques, les séances d'autisme

prolongées et mon attitude à tendance lunaire. C'est aussi grâce à elle que cette thèse a pu aboutir.

En bref, merci à tous ! (et pardon à ceux que j'ai honteusement oubliés)

Anecdotes...

Pour la réalisation de cette thèse auront été nécessaires :

- 3 ans, 2 mois et 8 jours...
- dont 3 mois et demi de rédaction.
- plus de 500 pages de corrections et annotations (fournies gracieusement par Fred et Pat)
- quelques soirées pizzas au labo...
- une vingtaine de soirées à terminer des articles à 2h du mat'
- 10 kg de chocolat en tablette (à la réflexion, ce n'est pas énorme)

« *Hakuna Matata, mais quelle phrase magnifique !*
Hakuna Matata, quel chant fantastique !
Ces mots signifient que tu vivras ta vie
Sans aucun soucis, philosophie !
Hakuna Matata. . . »

Timon le suricate et Pumba le phacochère

Notations

Typographie

À quelques rares exceptions près, nous utilisons les conventions suivantes.
Les scalaires sont notés en minuscules : x, y, z.
Les vecteurs sont notés en minuscules et en gras : $\boldsymbol{\mu}, \boldsymbol{x}_k, \boldsymbol{z}$
Les matrices sont notées en majuscules : Δ, A.
Les fonctions scalaires sont notées en minuscules : $f : \mathbb{R} \to \mathbb{R}$.
Les fonctions vectorielles sont notées en minuscules et en gras : $\boldsymbol{g} : \mathbb{R} \to \mathbb{R}^4$

Symboles

Ceci est un petit récapitulatif des notations souvent utilisées dans ce manuscrit. Le lecteur pourra s'y référer à tout moment.

$\mathbb{R}$	Ensemble des réels
$\mathbb{N}$	Ensemble des entiers naturels
$\mathbb{Z}$	Ensemble des entiers relatifs
$\mathbf{0}$	Vecteur nul (0 sur toutes les composantes)
$\mathbf{1}$	Vecteur unité (1 sur toutes les composantes)
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Delta)$	Densité gaussienne de moyenne $\boldsymbol{\mu}$ et de covariance Δ
$\mathcal{U}_{[a,b]}$	Densité uniforme définie sur l'intervalle $[a, b]$
$P(x = y)$	Probabilité pour que $x = y$
N	Nombre de particules utilisées dans les stratégies de type SIR
$\boldsymbol{x}_k$	Vecteur d'état à l'instant k définissant une configuration du modèle de l'homme
$\boldsymbol{z}_k$	Mesure à l'instant k
$p(\boldsymbol{x}_k \vert \boldsymbol{x}_{k-1})$	Distribution décrivant la dynamique de notre système
$p(\boldsymbol{z}_k \vert \boldsymbol{x}_k)$	Distribution décrivant la fonction de vraisemblance de notre système
$D_{(.)}$	Distance de similarité exploitée au sein de la fonction de vraisemblance
d	Nombre de dimensions de notre système
$\boldsymbol{x}_k^{(i)}$	$i^{\text{ème}}$ particule représentant la distribution de $\boldsymbol{x}_k$
$\mathcal{D}$	Domaine de définition

Considérations générales

Par souci de simplicité d'écriture, nous ne distinguerons pas une variable aléatoire de sa réalisation. De même, nous commettrons l'abus de langage consistant à employer indifféremment les termes « distribution de probabilité » et « densité de probabilité ».

Chapitre I

Introduction et état de l'art

Le présent chapitre introduit le contexte de nos travaux, ses enjeux et applications, pour ensuite établir un état de l'art des techniques déjà existantes dans le domaine du suivi visuel. Après avoir positionné nos travaux par rapport à la littérature, nous détaillons le plan du manuscrit.

1 Introduction

1.1 Applications

La capture de mouvement est actuellement un défi majeur de la communauté « Vision par Ordinateur ». Elle motive de nombreuses investigations, notamment pour la grande diversité d'applications qu'elle recouvre.

Il existe déjà, à l'heure actuelle, des systèmes de capture de mouvement commerciaux précis et relativement fiables, tels les systèmes proposés par Motion Analysis [171], ou VICON [172]. Ils sont classiquement utilisés pour des applications de synthèse de mouvement dans l'industrie du cinéma et du jeu vidéo, pour l'analyse du mouvement à des fins médicales, ou pour l'étude du geste sportif. Ces systèmes sont généralement constitués d'une dizaine (ou plus) de caméras infra-rouges. Leur principe consiste à localiser dans l'espace des marqueurs réfléchissants. Toutefois, de tels systèmes s'avèrent onéreux (quelques centaines de milliers d'euros), et imposent la fixation de marqueurs sur le sujet, ce qui peut gêner certains mouvements. Enfin, le protocole d'acquisition résultant est relativement lourd à mettre en œuvre.

La capture de mouvement sur la base de caméras « classiques » constitue une alternative intéressante. En effet, le coût matériel d'une telle solution est bien plus abordable (de l'ordre du millier d'euros), et le protocole d'acquisition est plus léger, ne nécessitant pas d'instrumenter le sujet. L'ensemble des applications citées précédemment peut être abordé, mais la légèreté du système permet également d'ouvrir des champs d'application inédits tels l'interface homme-machine et/ou l'interface homme-robot, où les capteurs peuvent être embarqués sur un robot mobile. Cela nécessite généralement l'exploitation conjointe de techniques de reconnaissance de mouvement afin de permettre

une interprétation de gestes ou de postures. D'une manière générale, les applications plus orientées « grand public », comme par exemple en domotique ou dans le domaine des jeux vidéos, constituent des débouchés intéressants.

Dans le cadre de nos travaux, nous visons une application robotique et plus particulièrement un robot autonome évoluant dans des environnements humains. Ce domaine revêt de multiples enjeux. Il peut s'agir :

- **de robots assistants ou auxiliaires de service** aussi bien du grand public que du professionnel dans un domaine donné. Leurs capacités sont définies *a priori* et liées aux services à assurer. Citons par exemple les robots guides de musées, tel Rackham [144], ou dédiés aux maisons de retraites, tel RG [97].
- **de robots personnels ou compagnons**, assimilables à des robots assistants de seconde génération, destinés à des interactions et tâches plus personnelles avec l'homme. Citons ici les robots Papero de NEC et Qrio de Sony. La finalité des robots personnels est d'acquérir de nouvelles capacités, et connaissances à l'aide d'un apprentissage ouvert et actif et d'évoluer en constante interaction et coopération avec l'homme. Ces robots, aux capacités évolutives et corrélées aux besoins spécifiques de leur tuteur, sont appelés robots cognitifs. Plusieurs plates-formes expérimentales dédiées à la recherche existent actuellement, parmi lesquelles Cog [49], Biron [100] et JIDO [51].

L'énumération est loin d'être exhaustive. Un état de l'art complet sur les robots sociaux et leurs applications est accessible dans [50].

Malgré une variété d'applications attractive, l'utilisation de caméras pour le suivi de mouvement soulève un grand nombre de difficultés, limitant le développement de ces applications. Aussi les recherches dans le domaine du suivi visuel du mouvement humain prennent-elles une place de plus en plus importante dans la communauté Vision par Ordinateur, même si les premiers travaux dans le domaine remontent à quelques décennies.

1.2 Historique

Bien avant l'apparition des ordinateurs, le mouvement humain a fait l'objet d'études approfondies. Au XIX$^{\text{ème}}$ siècle, le photographe Muybridge est le premier à se pencher sur l'analyse du mouvement de l'animal et de l'homme [118, 119]. Il travaille tout d'abord sur le galop du cheval, avant de se focaliser sur des mouvements tels que la marche humaine.

Avec l'apparition des ordinateurs et des caméras numériques, le suivi visuel prend une autre dimension. Les premiers travaux dans le domaine datent des années 1980. O'Rourke et Badler [124] exploitent alors des images de synthèse. Les traitements sur des images réelles suivent avec Hogg [76] puis Rohr [137, 138] dans le début des années 90. Les systèmes sont alors limités au suivi d'un mouvement de marche à un degré de liberté (DDL). Par la suite, des problèmes de plus grande dimension sont abordés, tels le suivi d'une main [80, 135] ou du corps humain dans sa totalité [58]. Les publications

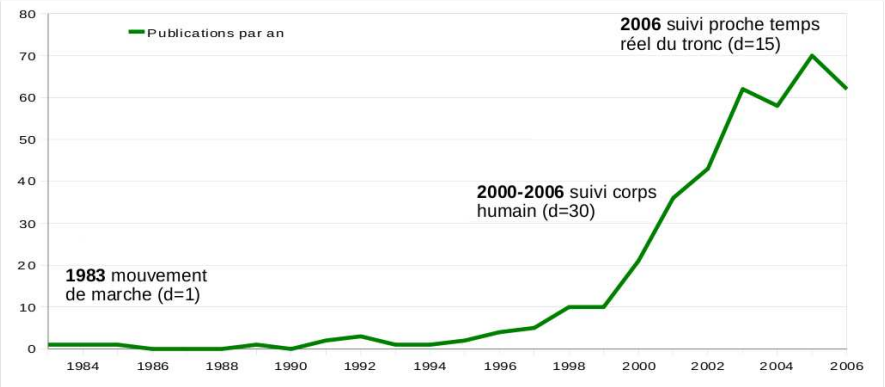


FIG. I.1 – Évolution du nombre de publications annuelles majeures dans le domaine du suivi visuel de mouvements humains. Les données sont tirées des études [113, 114].

dans le domaine du suivi par Vision par Ordinateur n'ont de cesse d'augmenter depuis, comme l'attestent les nombreuses études [2, 47, 60, 117, 127], les numéros spéciaux de journaux [73, 74], ou l'attraction grandissante de la recherche industrielle pour les applications [95, 120]. Dans [113, 114], Moeslund recense 155 publications avant l'année 2000 contre 352 entre les années 2000 et 2006, témoignant du dynamisme de la communauté. Ceci est bien visible sur la figure I.1, qui présente l'évolution du nombre de publications majeures par année depuis 1980.

Le domaine est en constante évolution et les techniques proposées sont nombreuses, si bien qu'il est difficile d'en avoir une vue d'ensemble complète. Le problème est d'une grande complexité, tant au niveau de sa formulation, de sa résolution théorique, que des aspects implémentation.

2 Contexte

2.1 Énoncé de la problématique

La capture de mouvement humain (ou HMC pour « Human Motion Capture ») consiste à reconstruire le mouvement d'un individu à partir des données fournies par un ou plusieurs capteurs. Cette reconstruction se focalise généralement sur la configuration 3D des articulations de la personne, mais peut parfois inclure l'apparence du sujet ou la forme de ses membres. Elle peut être plus ou moins fine selon la complexité du modèle choisi, et implique parfois des capteurs hétérogènes (caméras couleur ou noir et blanc, caméras 3D, caméras thermiques, ...). Ces capteurs peuvent être fixes ou mobiles, et l'environnement plus ou moins contrôlé. D'une manière générale, les systèmes de HMC doivent établir un compromis entre la précision du résultat escompté, les hypothèses simplificatrices et le coût calculatoire. Il convient donc de cibler un type

d'application précis afin de s'adapter au mieux au contexte visé. Cette problématique soulève des difficultés à plusieurs niveaux, évoqués ci-après.

2.2 Difficultés

Formulation du problème

Le problème que nous abordons ici se pose en tant que problème inverse. En effet, il consiste à exploiter les mesures délivrées par les capteurs mis en place afin d'identifier le mouvement du sujet observé. La première difficulté vient du fait qu'il s'agit d'un problème mal posé, au sens où les seules observations ne suffisent pas à caractériser entièrement une configuration du modèle de l'homme [71]. De ce fait, une approche analytique rigoureuse est impossible à mettre en œuvre dans la plupart des cas, et la grande majorité des méthodes proposées dans la littérature sont basées sur une discrétisation de l'espace des solutions [24, 45]. De plus, elles prennent en compte des contraintes et des connaissances *a priori* sur le modèle afin de permettre une recherche plus efficace des solutions.

Une autre difficulté vient de la nature des données extéroceptives, par définition 2D, alors que la configuration du sujet doit être caractérisée dans l'espace euclidien 3D. En effet, la projection perspective impliquée dans l'acquisition des données entraîne la perte de l'information de profondeur. Cette complication peut être en partie éludée par l'utilisation de plusieurs capteurs couvrant au mieux tout l'espace de travail et dont les champs de vue sont complémentaires [10, 21, 38]. Toutefois, certaines applications ne permettent pas une telle configuration des capteurs. Le suivi doit alors prendre en compte les incertitudes liées à la profondeur [146].

Les considérations précédentes sont valables quels que soient les modèles de l'homme et de l'environnement choisis, y compris s'ils sont en parfaite adéquation avec la réalité physique. À ces premières difficultés vient donc s'ajouter l'inexactitude des modèles manipulés.

Modèles de l'homme et de l'environnement

Une représentation fidèle de l'homme requiert une centaine de DDL cinématiques, chaque main en possédant une vingtaine. La représentation volumique des membres peut en outre nécessiter une dizaine de paramètres additionnels. L'espace des configurations caractérisant le modèle de l'homme est donc immense, et les solutions actuelles ne sont pas en mesure de traiter une telle quantité de données de manière efficace. En outre, l'apparence du sujet est très variable : les vêtements sont de natures très différentes (longs, courts, serrés ou amples, de couleurs plus ou moins vives, unis ou multicolores) et leurs déformations dues aux mouvements sont très complexes à modéliser. La couleur de la peau peut également affecter la définition du modèle choisi, augmentant encore la dimension de l'espace de recherche. Ceci rend la modélisation exacte quasi impossible. Il convient alors de limiter le réalisme du modèle que nous adoptons en fonction de l'application visée. Le suivi en devient moins précis, puisque le modèle est plus fruste et ne peut pas être en parfaite adéquation avec les mesures observées. Ce

compromis est omniprésent dans la littérature, néanmoins, plusieurs systèmes exploitant un modèle simple parviennent à suivre le sujet avec succès [26, 94, 179]. Notons également que l'homme est par définition une cible très dynamique avec des mouvements difficilement prédictibles. Les connaissances *a priori* sur les types de gestes réalisés sont donc très limitées.

A l'instar de la modélisation de l'homme, l'environnement non contrôlé engendre un grand nombre d'écueils. L'arrière-plan peut changer significativement (environnement intérieur ou extérieur, encombré ou dégagé), et perturber le suivi par la présence d'artefacts sources de fausses mesures. La luminosité de la scène peut également varier fortement et la présence d'individus autres que le sujet d'étude peut gravement nuire au système de suivi. La fonction de coût qui représente le lien entre une configuration du modèle et le jeu de données présente ainsi de nombreux *extrema* locaux.

Enfin, il faut garder à l'esprit que de nombreuses applications (interface homme-robot ou homme-machine) requièrent une grande réactivité. Une exécution en temps réel du suivi est alors indispensable, ce qui est peu compatible avec les modèles décrits précédemment.

Devant ces nombreuses difficultés, toutes les approches pour le suivi par Vision par Ordinateur se fondent sur des hypothèses — plus ou moins contraignantes quant au domaine d'applicabilité — que nous détaillons ci-après.

2.3 Hypothèses

Dans [113], Moeslund répartit les hypothèses régulièrement utilisées en plusieurs catégories : les hypothèses sur les mouvements du sujet, sur son apparence, et celles sur l'environnement. Elles sont détaillées dans le tableau I.1 par ordre de fréquence d'utilisation. Les hypothèses formulées dans la littérature dépendent du contexte et de l'application visée. Nous pouvons d'ores et déjà différencier deux grandes familles d'applications.

Les applications « caméras déportées »

Elles regroupent notamment la capture de mouvement pour l'animation de synthèse, l'étude du mouvement sportif ou à des fins médicales. Généralement, ces contextes exploitent plusieurs capteurs/caméras fixes et larges champs proposant des vues complémentaires. L'environnement est en majeure partie contrôlé. Le but est de réaliser une capture de mouvement dans les meilleures conditions possibles, en limitant les artefacts environnementaux. Les capacités de calcul en temps réel du système ne constituent pas un critère déterminant, même si la vitesse de traitement est toujours appréciée. Dans un tel contexte, on émet généralement les hypothèses d'éclairage constant (3.1) et d'arrière-plan connus (3.2, 3.3). On connaît les paramètres de calibrage des caméras (3.4), la posture initiale du sujet est fixée (2.1) et son apparence est en grande partie connue (2.2, 2.4, 2.5). On suppose également que le sujet évolue dans le champ de vue (1.1, 1.3), les caméras étant fixes (1.2).

Mouvements	1.1. Le sujet reste à l'intérieur de l'espace de travail 1.2. Pas de mouvement ou mouvement constant de la caméra 1.3. Une seule personne dans le champ de vue 1.4. Le sujet regarde la caméra tout le temps 1.5. Mouvements fronto-parallèles 1.6. Pas d'occlusion 1.7. Mouvement lents et continus 1.8. Un seul mouvement ou très peu de mouvements simultanés 1.9. La dynamique du mouvement est connue 1.10. Le sujet évolue sur un sol plat
Sujet	2.1. Position de départ connue 2.2. Morphologie connue 2.3. Marqueurs placés sur le sujet 2.4. Habits colorés 2.5. Habits moulants
Environnement	3.1. Éclairage constant 3.2. Fond fixe 3.3. Fond uniforme 3.4. Paramètres de calibrage des caméras connus 3.5. Matériel spécial

TAB. I.1 – Les hypothèses couramment utilisées dans les systèmes de HMC.

Les applications de surveillance, bien que plus contraignantes en terme de temps de calcul, reprennent une grande partie de ces hypothèses (1.1, 1.2, 3.1, 3.2, 3.4). Toutefois, l'environnement et le modèle de l'homme sont moins maîtrisés. La robotique ubiquiste où les réseaux de capteurs sont intégrés à l'environnement s'inscrit également dans cette problématique. Les contraintes temps-réel sont présentes et les hypothèses sont similaires à celles faites en surveillance. Notons enfin que les applications de téléprésence rentrent dans cette catégorie, bien qu'elles n'exploitent souvent qu'une caméra monoculaire, pour des raisons de simplicité de mise en œuvre.

Les applications « caméras embarquées »

Ces applications supposent des capteurs embarqués sur des plates-formes mobiles qui évoluent en intérieur ou en extérieur dans un environnement non contrôlé, variable et dynamique. C'est typiquement le cas de la robotique autonome. Nous ne pouvons donc pas — ou difficilement — émettre d'hypothèses sur les environnements (luminosité et arrière-plans changeants). Toutefois, en nous plaçant dans un contexte d'interaction, on peut s'attendre à ce que le sujet s'adresse au robot (1.1, 1.3, 1.4). Les hypothèses sur le sujet sont en revanche plus difficile à avancer : on ne connaît pas sa position initiale, ni son apparence. Sa morphologie peut cependant être supposée proche de la moyenne (2.2).

Le point le plus important des applications robotiques demeure la nécessité de trai-

ter les données en temps réel. Le système est embarqué et doit partager les ressources avec d'autres modules logiciels (déplacement, localisation, supervision des tâches,...). Il doit également présenter une grande réactivité pour être accepté par l'homme.

À l'exception de [68, 69], les approches présupposent la présence d'une seule personne dans le champ de vue (1.3). En effet, lorsque plusieurs personnes sont présentes, outre les occultations propres, viennent se greffer les occultations inter-personnes, qui rendent le suivi encore plus complexe, particulièrement dans le cas mono-caméra.

Par ailleurs, la totalité des systèmes de capture de mouvement 3D sont mono-cibles. Cette hypothèse n'apparaît d'ailleurs même pas dans la classification de Moeslund tant elle est universelle. Le suivi multi-cible reste à l'heure actuelle incompatible avec le HMC à part entière de par le nombre de paramètres à estimer. Les applications de suivi visuel multi-cible [83, 134] reposent quant à elles sur des modèles de l'homme plus simples, la finalité n'étant pas de reconstruire les mouvements avec autant de précision, mais simplement de localiser globalement les individus, typiquement dans un contexte de surveillance.

Suite à cette introduction aux enjeux et difficultés du problème de suivi visuel, nous présentons ci-après un état de l'art dans lequel nous positionnons notre approche.

3 Notre approche au sein de l'état de l'art

Devant la variété des techniques développées dans la littérature, il est difficile de proposer une classification universelle, même si quelques critères sont incontournables. Nous pouvons catégoriser les approches existantes selon le type de configuration du système visuel ainsi que les techniques utilisées (détection ou suivi, approches globales ou par parties, techniques déterministes ou stochastiques). Nous positionnons ainsi nos travaux dans l'état de l'art et terminons par un rapide survol des techniques de filtrage stochastique investies dans cette thèse et omniprésentes dans la littérature.

3.1 Vision monoculaire *vs* stéréoscopique *vs* multi-oculaire

Le nombre de caméras utilisées joue un rôle crucial dans le suivi par vision. Certaines approches requièrent une seule caméra pour le suivi [63, 126, 140, 147, 164], pour des raisons de simplicité de mise en œuvre, ou par contraintes applicatives. Le suivi est alors plus difficile, l'information de profondeur n'étant pas disponible. Les approches se focalisent donc sur l'apparence du sujet, et/ou sur son mouvement dans les images. La technique consiste généralement à effectuer la projection dans l'image du modèle de l'homme et à la confronter à l'image. Ces approches sont dites « par apparence » (ou « appearance based ») et posent le problème des parties cachées (observation partielle du corps en raison des occultations propres), qui doivent être gérées lors de la projection image du modèle [109, 135]. Notons que ce problème, inhérent à la projection mise en œuvre dans les approches par apparence, est également présent pour les systèmes stéréoscopiques ou multi-oculaires.

Les approches stéréoscopiques [115, 179] reposent sur l'utilisation de deux caméras séparées de quelques dizaines de centimètres. Cette faible base permet généralement une reconstruction dense de la scène, qui peut être exploitée *via* des mesures 3D. Notons que cette opération a cependant un coût calculatoire assez élevé. Les systèmes monoculaires et stéréoscopiques peuvent être embarqués sur des plates-formes mobiles et constituent des approches de choix pour les applications d'interaction homme-robot. L'utilisation de méthodes de soustraction de fond n'est alors pas envisageable, du fait de la mobilité de la plate-forme. Il convient d'exploiter d'autres informations dans les images, telles que les contours, les formes, ou éventuellement des connaissances *a priori* sur l'apparence du sujet observé [9].

Afin de contourner les problèmes de l'estimation des parties cachées et du manque d'indices visuels discriminants sur la profondeur, certaines approches reposent sur plusieurs caméras séparées par une large base (plusieurs mètres) [37, 38, 58, 98, 152, 160]. Ces systèmes de vision multi-oculaire sont généralement fixes et permettent une reconstruction 3D de la scène. L'utilisation de plusieurs caméras apporte une information plus complète sur l'apparence du sujet. Toutefois, l'information sur la texture et l'apparence du sujet peut varier d'un point de vue à l'autre. L'utilisation d'un système multi-oculaire se focalise alors sur des indices tels que les contours ou la silhouette du sujet dans les images [38, 142].

De manière générale, les systèmes les plus efficaces à l'heure actuelle [39, 160] reposent sur l'utilisation de plusieurs caméras en configuration multi-oculaire pour couvrir au mieux l'espace de travail et exploiter les informations riches et complémentaires incluses dans chaque plan image.

Dans notre cas, l'objectif est l'intégration d'un système de suivi sur un robot mobile. Toutefois, afin d'appréhender les difficultés séquentiellement, nous souhaitons mener à bien des études dans un contexte plus simple, *i.e.* avec des caméras fixes. C'est pourquoi nous nous focalisons sur deux contextes différents que nous abordons avec des techniques similaires :

- Un contexte de caméras déportées constitué de trois caméras fixes dans un environnement maîtrisé. Ce cadre peut être considéré comme un premier pas vers des applications de type robotique ubiquiste, thème de recherche exploré par le groupe RAP. Par la suite, nous utilisons également l'appellation de « contexte multi-oculaire ».
- Un contexte de caméras embarquées exploitant une tête stéréoscopique positionnée sur un robot mobile pour l'interaction homme-robot, qui constitue l'axe de recherche privilégié du groupe. Nous désignons ce contexte par « contexte stéréoscopique ».

3.2 Suivi *vs* détection

Dans [113, 114], Moeslund propose une taxonomie qui différencie le suivi du mouvement sur une séquence d'images et les techniques se limitant à la détection de posture sur une ou plusieurs images fixes à un instant donné. Les premières reposent sur une analyse spatio-temporelle du/des flux de données [80] tandis que la détection est

complètement indépendante du temps et se focalise sur l'analyse d'une seule et unique image [78, 133]. Toutefois, il est courant que le suivi ait recours aux techniques de détection [16, 17, 92, 142], notamment pour le difficile problème de l'initialisation (voire de la ré-initialisation automatique), ce qui est le cas de notre approche. La détection visuelle constitue une aire de recherche à part entière [75, 163].

3.3 Approche descendante *vs* approche ascendante

Le suivi du mouvement humain impose des modèles admettant un grand nombre de DDL (de 20 à 40 classiquement). La recherche d'une solution dans un espace de grande dimension est une tâche complexe. Une grande partie des approches proposées dans la littérature, dites « approches descendantes » (ou « top down »), tente une exploration de l'espace complet [38, 80, 140], ce qui est généralement très consommateur en temps de calcul. Afin de contourner ce problème, il est possible de privilégier une approche par partie se focalisant sur la localisation d'un membre ou d'un sous-ensemble des membres. On peut citer les techniques de partitionnement explicite ou implicite de l'espace [39, 103], ou les approches dites « ascendantes » (ou « bottom up ») [94, 134] qui consistent à localiser chaque partie indépendamment puis à appliquer des contraintes de cohérence sur l'ensemble. Ces techniques sont particulièrement adaptées pour les algorithmes de propagation de croyances [121]. Toutefois, la localisation indépendante des membres est parfois difficile : il convient d'utiliser des détecteurs spécifiques (têtes, mains, bras, jambes) ou de mettre en place des mesures très discriminantes sur les membres car leur apparence dans l'image est souvent très similaire (bras et avant-bras par exemple). Notre approche se focalise sur les méthodes descendantes afin de pouvoir exploiter des mesures plus génériques. En outre, les avancées récentes dans le domaine [8, 41] présentent des temps de calcul plus acceptables.

3.4 Approche stochastique *vs* déterministe

Le choix de la stratégie d'exploration de l'espace des solutions revêt une grande importance. Nous pouvons ici distinguer les approches déterministes et les approches stochastiques. Les premières proposent par définition une solution toujours identique pour un problème donné. Elles formulent le problème du suivi comme la minimisation d'une fonction de coût traduisant l'écart entre une configuration du modèle donnée et les images. Toutefois la nature complexe de ces fonctions bloque parfois les solutions dans des *minima* locaux. On y trouve les algorithmes d'optimisation numérique tels que l'ICP (« Iterative Closest Point ») [94, 116], la descente de simplexe [149] ou la méthode de Levenberg-Marquardt [125]. Le suivi par techniques déterministes reste cependant moins employé que les approches bayésiennes dans la littérature, bien que les avancées récentes soient prometteuses [15, 160].

Les techniques stochastiques [5, 80] sont quant à elles des outils de plus en plus exploités. Elles considèrent le problème de suivi comme l'estimation récursive de la densité de probabilité de la configuration de l'homme *a posteriori*, *i.e.* en tenant compte de connaissances *a priori* sur sa dynamique et en exploitant les observations à chaque

instant.

La résolution de ce problème d'estimation bayésien peut elle-même être déterministe, comme dans le cas du filtre de Kalman [89, 105]. Celui-ci n'est cependant applicable qu'aux systèmes linéaires gaussiens. Plusieurs extensions aux systèmes non linéaires et potentiellement non gaussiens existent [86, 162], bien que leur application dans le domaine du suivi visuel 3D soit assez restreint [88, 155]. D'autres approches s'orientent vers une résolution stochastique du problème. Les filtres particulaires notamment rencontrent un fort succès auprès de la communauté Vision. Ils permettent la modélisation de densité multi-modales particulièrement adaptées au problème dans des environnements non contrôlés aux arrière-plans encombrés. En outre, ils autorisent la fusion de données hétérogènes dans un cadre théoriquement étayé. Toutefois, ces méthodes ont un coût calculatoire important, dont la réduction fait l'objet de recherches approfondies [38, 39, 103]. Nous pouvons également mentionner ici les techniques par propagation de croyances [12, 69, 120, 177]. Les méthodes stochastiques posent cependant le problème de la répétabilité du suivi : pour plusieurs exécutions sur un même jeu de données, le résultat obtenu n'est pas identique, de par la nature aléatoire du procédé. La littérature est hélas peu loquace sur ce problème spécifique.

Signalons enfin que certaines approches choisissent de coupler les deux stratégies : [146] propose un algorithme où un échantillonnage stochastique est mis en place et chaque échantillon est optimisé localement de manière déterministe par rapport à une fonction de coût. Les techniques de « Stochastic Meta-Descent » (SMD) [14, 91] se basent sur la méthode de la descente de gradient dans laquelle est introduite une composante aléatoire afin de s'affranchir du problème des *extrema* locaux.

3.5 Au sein des techniques stochastiques

Les fondations du suivi visuel stochastique tel qu'il est abordé dans cette thèse furent posées par Isard et Blake en 1996. Dans [80, 81], ils proposent une approche, nommée CONDENSATION pour « CONDitional DENSity propagATION » et dérivée de la technique de « factored sampling » initialement proposée par Grenander dans [66]. Ils présentent plusieurs applications, dont le suivi 2D de personne, de visage et de main dans un flot vidéo. Ces applications sont exécutées en temps réel, même si la dimension de l'espace de recherche reste faible. Une comparaison au filtre de Kalman montre que la gestion de la multi-modalité des distributions permise par le filtre particulaire le rend plus efficace que les méthodes jusqu'alors exploitées [40]. La communauté Vision par Ordinateur ne cesse depuis de proposer de nouvelles variantes de filtres particulaires plus aptes à traiter des espaces de grandes dimensions.

MacCormick et Isard proposent dans [102, 103] une alternative au filtre particulaire classique qu'ils nomment filtre PARTITIONNÉ. Le principe sommaire consiste à diviser l'espace de recherche en parties disjointes de dimensions plus petites. La complexité exponentielle en fonction du nombre de dimensions est alors contrebalancée par la complexité linéaire en fonction du nombre de partitions. L'algorithme est exploité dans une application temps-réel de suivi de main pour l'interaction homme-machine dans des environnement encombrés.

Sminchisescu et Triggs présentent dans [146, 147] un algorithme de suivi basé sur un échantillonnage selon les covariances associées aux modes de la distribution filtrée. L'algorithme proposé, dénommé « Covariance Scaled Sampling » (CSS), modélise la densité de filtrage par des mixtures de gaussiennes.

Deutscher *et al.* présentent dans [38] l'« Annealed Particle Filter », ou APF, un algorithme de filtrage particulaire faisant appel à certains principes du recuit simulé. Ils reprennent et étendent cette idée dans [39] où ils travaillent sur la dynamique appliquée aux particules : en lieu et place d'un bruit gaussien identique sur chaque dimension de l'espace, ils appliquent un bruit gaussien dont la covariance est proportionnelle à celle du nuage de particules à l'étape précédente, reprenant ainsi le principe de [146]. Ils introduisent en complément l'utilisation d'un opérateur de « crossing-over », technique empruntée au domaine des algorithmes génétiques. Le filtre ainsi mis en place permet de s'affranchir de la définition d'un partitionnement explicite de l'espace d'état que l'on est contraint de mettre en place dans des techniques comme celle du filtrage PARTITIONNÉ.

De nombreuses variantes de la CONDENSATION sont proposées dans la littérature. Citons par exemple le filtre particulaire AUXILIAIRE [132] de Pitt et Shephard, et une évolution par affinages successifs de l'exploration proposée dans [104]. Wang et Rehg proposent un filtre particulaire optimisé exploitant la transformée « unscented » [86] (« Optimized Unscented Particle Filter » ou OUPF). Ils comparent diverses variations de leur algorithme avec une vérité de terrain obtenue par un système commercial dans un contexte de capture de mouvement avec caméras fixes. Le filtrage par échantillonnage de l'historique (ou « History Sampling » SIR), proposé par Torma et Szepesvári dans [158] s'applique aux systèmes stochastiques dont une partie du vecteur d'état (dite « historique ») suit une évolution déterministe. Certains filtres exploitent en outre la méthode de Rao-Blackwellisation [23], qui tire parti de la possibilité de traiter une sous-partie du vecteur à estimer par des méthodes analytiques telles que le filtre de Kalman lorsque cela est possible. Ceci résulte en une réduction de la variance du filtre [46]. Han *et al.* proposent dans [72] un filtre particulaire capable de représenter les densités de manière analytique sous forme de mixtures de gaussiennes.

Malgré l'enthousiasme de la communauté Vision pour les filtres particulaires en raison de leur efficacité pratique, et contrairement à une idée répandue, Daum met en évidence le fait que les filtres particulaires ne s'affranchissent que dans certains cas de la « malédiction de la dimensionnalité » (« curse of dimensionality ») [32]. Il propose une classification empirique de la complexité selon que le problème est « vaguement gaussien » ou non. Dans [31, 33, 35], il propose de privilégier les approches de types Quasi Monte Carlo (QMC) aux approches de type Monte Carlo classique. Les premières permettent une couverture plus homogène de l'espace de définition des fonctions que l'on souhaite échantillonner pour un même nombre d'échantillons tirés. Cette discrépance inférieure leur confère une convergence plus rapide. Philomin [130] et Guo et Wang [67] prônent également l'utilisation des techniques QMC et proposent d'intégrer l'échantillonnage quasi-aléatoire dans les schémas particulaires classiques.

Notre approche s'inspire de tout ou partie des différentes techniques présentées ci-dessus. Après avoir positionné nos travaux dans la littérature, nous listons ci-après les

caractéristiques qui s'en démarquent.

4 Notre approche en complément de l'état de l'art

4.1 Exploitation de mesures « hybrides »

Dans le domaine du suivi visuel, la littérature différencie classiquement les approches par apparence qui consistent à projeter dans l'image chaque hypothèse de configuration du modèle de l'homme et à en évaluer la pertinence par le biais de mesures 2D, et les approches 3D qui consistent à reconstruire tout ou partie de la scène observée afin de recaler le modèle grâce à des mesures 3D.

Les approches 2D sont largement utilisées dans les systèmes de suivi [38, 145, 148]. Elles reposent sur des indices visuels basés sur l'exploitation de la forme, des contours, de la couleur, ou du mouvement. Elles sont abordées en détail plus loin dans le manuscrit. La fusion de ces mesures 2D, à l'instar de [111, 128, 146], nous semble ici primordiale. La manipulation de données 3D permet quant à elle de s'affranchir des problèmes liés à la projection perspective, mais nécessite plusieurs caméras (système stéréoscopique ou multi-oculaire). Caillette *et al.* [21] notamment mettent en place un système de suivi de mouvement multi-caméras en temps réel. Ils reconstruisent la scène en voxels à partir de la silhouette segmentée dans chaque image. Cette technique est par ailleurs assez présente dans la littérature [25, 83, 91, 112, 155]. Cette reconstruction 3D est modélisée au moyen d'une mixture de gaussiennes (position et couleur des voxels sont prises en compte). Chaque particule est ensuite évaluée grâce à une distance entre la mixture de gaussiennes associée et celle extraite de l'image. L'approche de Ogawara *et al.* [122] exploite elle aussi une reconstruction 3D de la scène mais *via* la méthode de « marching cubes » permettant l'obtention d'un volume maillé. La localisation est ensuite réalisée grâce à un couplage d'ICP et d'une approche hiérarchique (« Kd-tree »).

Notre approche « hybride » tente quant à elle de fusionner plusieurs attributs 2D (forme, couleur, mouvement, ...) et une reconstruction 3D éparse afin de tirer à la fois parti de la richesse de l'apparence image et des contraintes 3D très discriminantes [8].

Nous exploitons également la méthode de [82], où Isard et Blake sont les premiers à proposer une extension de leur propre algorithme : alors que la CONDENSATION classique propose un schéma similaire au filtre de Kalman pour le suivi — propagation de la représentation particulaire à l'instant précédent selon la dynamique, puis confrontation à la mesure —, la stratégie I-CONDENSATION se propose d'exploiter les mesures au plus tôt dans l'échantillonnage des particules. Ils présentent un système de suivi temps-réel de main dans un environnement encombré en se basant sur une détection de blobs de couleur peau et une mesure de contours dans l'image. Depuis, plusieurs travaux reprennent ce même principe [61, 128], mais à notre connaissance exclusivement dans des contextes 2D. Notre approche se propose d'étendre cette exploitation de la mesure au plus tôt par l'ajout de mesures 3D.

4.2 Évaluation des performances

Les techniques proposées dans la littérature sont exploitées dans divers contextes applicatifs reposant sur l'utilisation de systèmes de perception distincts (stéréoscopique, multi-oculaire, ...). Elles affichent donc des performances assez variées et il reste parfois difficile de les comparer. En effet, l'évaluation de stratégies de suivi visuel soulève plusieurs problèmes. Le suivi 3D étant un problème inverse, nous ne disposons pas *a priori* de vérité de terrain à laquelle comparer les résultats obtenus. Dans de très nombreux cas, cette vérité de terrain est obtenue « à la main » suite à un étiquetage et une localisation manuelle des membres dans les séquences vidéos [26, 142]. Cette démarche est particulièrement longue et fastidieuse. Une autre possibilité consiste à utiliser, lorsque c'est envisageable, des systèmes de HMC commerciaux permettant une grande précision du suivi [10, 68]. La mise en œuvre est toutefois lourde, et d'autres problèmes peuvent survenir (difficulté de localisation des marqueurs, nécessité de plusieurs opérateurs, conditions d'acquisition imposant un grand espace libre, difficulté de calibrage). En outre, l'environnement est statique et généralement contrôlé, ce qui ne convient pas à toutes les applications.

La mise en place de métriques constitue également une difficulté : faut-il mesurer des erreurs sur les angles entre les membres ou sur les positions des jonctions articulaires ? Comment mettre en place un tel procédé pour les approches par parties ? Comment tenir compte des erreurs en position et en angle dans une seule et même métrique ? Plusieurs approches sont proposées dans la littérature [10,26], mais elles demeurent très attachées aux choix retenus des modèles de l'homme et de l'environnement. La répétabilité du suivi par technique stochastique reste également très peu abordée dans la littérature bien qu'elle soit un problème important. Il convient donc de mettre en place une métrique permettant de la prendre en compte.

Enfin, les environnements peuvent être très variés, tout comme les types de mouvements suivis, et il est très difficile de caractériser le comportement d'algorithmes de suivi pour un ensemble exhaustif des conditions d'application. En plus des difficultés inhérentes à l'acquisition et la mise en forme des données, les évaluations elles-mêmes représentent un travail important. Des propositions de mise en commun de base de données commencent toutefois à se mettre en place [170], ce qui constitue un premier pas vers la réalisation d'évaluations comparatives. Les données proposées ne sont toutefois pas encore étendues à un grand nombre de contextes différents.

Pour ces raisons, de nombreux travaux proposent des évaluations qualitatives [38, 41,146], même si récemment, quelques études quantitatives émergent [68]. Malheureusement elles se limitent souvent à une comparaison de la stratégie proposée à une stratégie de référence, souvent la CONDENSATION et de plus en plus l'APF [10, 26, 166]. Il reste ainsi difficile d'avoir une vue d'ensemble sur l'efficacité relative des différentes classes de filtrage particulaire. C'est pourquoi nous proposons la mise en place de métriques spécifiques permettant une comparaison quantitative des diverses stratégies particulaires étudiées au regard de plusieurs critères complémentaires.

4.3 Temps de calcul

La littérature sur le suivi visuel propose de nombreuses méthodes, mais les performances calculatoires des algorithmes ne sont pas toujours discutées. Les approches de type caméras déportées restent éloignées du temps réel (de l'ordre de la minute par image pour [68, 140, 146]). Dans le cas de caméras embarquées, notamment pour des applications de robotique autonome, les méthodes développées présentent des temps de traitement plus rapides (entre $10\ Hz$ et $15\ Hz$ pour [8, 94]).

Notre approche vise des temps de traitement similaires et donc compatibles avec le contexte applicatif d'interface homme-machine, mais également adaptés à l'intégration sur plate-forme mobile avec des ressources CPU limitées.

4.4 Intégration sur plate-forme robotique

Malgré la mise en place d'un nombre croissant de plates-formes robotiques dédiées à l'interaction homme-robot, les systèmes de suivi visuels compatibles avec les contraintes imposées par ce contexte demeurent peu nombreux.

Cielnak *et al.* proposent dans [26] un système de suivi 2D sur la base d'une caméra infra-rouge. Le modèle utilisé est composé de deux ellipses (une pour le torse, une pour la tête). Les auteurs utilisent la CONDENSATION classique avec un minimum de 300 particules. Ils proposent également un ensemble de mesures 2D permettant d'évaluer quantitativement l'efficacité d'un filtre. Ces mesures sont principalement basées sur les travaux de Doermann et Mihalcik dans [44, 173].

Azad *et al.* proposent dans [8] un système de suivi 3D du haut du corps utilisant une tête stéréoscopique. Ils exploitent une stratégie de filtrage classique (CONDENSATION) mais introduisent une fusion de données hétérogènes couplant des mesures d'apparence classiques et une mesure basée sur la reconstruction stéréoscopique de la position de la tête et des mains dans l'espace.

Dans [94], Knoop *et al.* mettent en œuvre une caméra 3D active (Swissranger) permettant une reconstruction matérielle de la scène en temps réel, gagnant ainsi un temps précieux pour le suivi. Un algorithme ICP est utilisé afin de recaler à chaque instant un modèle de l'homme basé sur des cônes tronqués dans le nuage de points 3D délivré par le capteur. Une approche ascendante est adoptée, permettant de localiser chaque membre indépendamment et de reconstruire la position globale de l'homme en appliquant des contraintes d'étirement élastique au niveau des liaisons.

Menezes *et al.* [106, 108, 109] exploitent une caméra monoculaire fixe. Un filtre particulaire AUXILIAIRE est utilisé afin de suivre un modèle à base de quadriques dégénérées dans le flot vidéo. Des contraintes de non-collision sont introduites, et l'exploitation de l'image repose sur la recherche de contours et sur l'utilisation de « patches » de couleurs associés aux différents membres.

Zhao *et al.* [178] privilégient un algorithme génétique hiérarchique à recuit simulé (« Hierarchical Annealed Genetic Algorithm » ou HAGA) couplé à une analyse en composante principale (ACP) afin de réduire la dimension de l'espace de recherche et d'améliorer son exploration. Ils exploitent des images provenant d'une caméra fixe

et présentant un fond dégagé.

Dans un contexte de collaboration par téléconférence, Mulligan, dans [115], tire parti des positions de la tête et des mains obtenues par segmentation couleur afin d'estimer la configuration d'un bras. Des considérations psychophysiques sont prises en compte afin de favoriser les poses naturelles. La méthode est appliquée dans le cas de caméras monoculaire et stéréoscopique, ce qui permet alors une reconstruction dense de la scène pour estimer l'orientation du torse.

Bien que les enjeux de la robotique soient des plus intéressants (assistance aux personnes, robot personnel), les difficultés inhérentes à l'intégration d'un système de HMC par vision peuvent freiner la mise en place de prototypes. Ainsi, parmi les travaux mentionnés, peu ont abouti au portage effectif sur un robot autonome [26, 94]. Notre approche cherche donc à combler ce manque en terme d'intégration robotique par un portage sur une plate-forme mobile.

Nous présentons dans les pages suivantes deux tableaux récapitulant les travaux principaux en suivi de mouvement humain menés dans les deux contextes « caméras déportées » et « caméras embarquées » et positionnant nos travaux dans la littérature. D'une manière générale, les approches dédiées à la robotique autonome mettent en œuvre des modèles de l'homme simplifié et affichent des performances résolument plus proches du temps-réel. De plus, elles reposent parfois sur des capteurs proposant une information plus immédiatement exploitable et moins riche que les images couleurs (caméras 3D ou infra-rouge). La signification des sigles non encore définis est présentée dans le glossaire.

Contexte « caméras déportées »

Qui ?	Où ?	Quand ?	Quoi (Capteurs, Contexte) ?	Comment (Modèles, Techniques) ?	Combien (Performance) ?
Balan *et al.*	[10]	2005	1 à 4 caméras fixes, environnement contrôlé	cônes tronqués, APF et SIR, segmentation	$1\times?\times?@1/45\ Hz$
Caillette *et al.*	[21]	2007	multi caméras fixes, environnement encombré	volume mixture de gaussiennes, 25 DDL + dérivées, modèles de dynamiques appris, SIR + VLMM	$?\times 320\times 240@25\ Hz$
Deutscher *et al.*	[38, 39]	2001	3 caméras, environnement très dégagé	34 DDL, dynamique gaussienne, APF, PAPF	$3\times?\times?@1/15\ Hz$
Fontmarty *et al.*	**[53, 54]**	**2008**	**$2/3$ caméras fixes, environnement quelconque**	**cônes tronqués, FP, fusion de mesures, $14/22$ DDL**	$\mathbf{2\times 640\times 480@1}\boldsymbol{Hz}$
Gupta *et al.*	[68]	2008	2 caméras fixes, environnements variés	cylindres, approche ascendante, pas de contraintes temporelles	$2\times?\times?@1/45\ Hz$
Kehl *et al.*	[91]	2005	$4-8$ caméras fixes, environnement dégagé	volume maillé, 24 DDL, SMD	$4\times 640\times 480@1\ Hz$
Ogawara *et al.*	[122]	2007	8 caméras fixes, environnement dégagé	29 DDL, modèle maillé déformable, Kd-Tree, ICP, reconstruction 3D	$8\times?\times?@?$
Sidenbladh *et al.*	[140]	2000	1 caméra fixe	cylindres, 25 DDL, FP, modèle générateur de l'image	$?\times?@1/300\ Hz$
Sigal *et al.*	[142]	2004	4 caméras fixes, environnement quelconque	cônes tronqués, 10×6 DDL, BP	$4\times?\times?@?$
Sminchisescu et Triggs	[146, 147]	2001	caméra mono, environnement encombré ou fond noir	quadriques, 30 DDL, CSS	$?\times?@1/180\ Hz$
Urtasun *et al.*	[160]	2004	digiclops 3 stéréo caméras fixes	squelette + surface 3D, modèle de mouvement + ACP ($80\rightarrow 6$ DDL), optimisation déterministe	$3\times 640\times 480@?$
Wang et Rehg	[166]	2006	multi-caméras, environnement dégagé	22 DDL, modèle 2D, OUPF	
Ziegler *et al.*	[179]	2006	4 caméras stéréos fixes, environnements encombrés	ICP+UKF, 14 DDL (moitié supérieure), modèle maillage 3D	$4\times?\times?@1\ Hz$

Contexte « caméras embarquées »

Qui ?	Où ?	Quand ?	Quoi (Capteurs, Contexte) ?	Comment (Modèles, Techniques) ?	Combien (Performance) ?
Azad *et al.*	[7–9]	2007	caméra stéréo fixe, environnement encombré, apparence de sujet connue	cônes tronqués, 14 DDL, marche aléatoire, SIR	$320 \times 240@15\ Hz$
Cielniak *et al.*	[26]	2005	caméra infrarouge, environnement encombré	ellipses 2D, 9 DDL, vitesse constante, SIR	$320 \times 240@(> 20\ Hz?)$
Fontmarty *et al.*	**[52]**	**2007**	**caméras stéréo, environnement quelconque**	**cônes tronqués, FP, fusion de mesures, 14 DDL**	$\mathbf{2 \times 320 \times 240@8Hz}$
Knoop *et al.*	[94]	2006	caméra 3D Swiss Ranger	10×6 DDL, cônes tronqués, ICP, contraintes géométriques, fusion de données	$150 \times 100@12\ Hz$
Menezes *et al.*	[108, 109]	2005	caméra fixe, environnements variés	8 DDL, quadriques, FP AUXILIAIRE , patches de couleurs	$320 \times 240@1\ Hz$
Mulligan	[115]	2005	caméra fixe mono ou stéréo, environnements variés	4 DDL, contraintes psychophysiques à partir de la position de la tête et des mains	$?\times?@?$
Philomin	[130]	2000	caméra embarquée sur véhicule, environnement extérieur	ACP ($\rightarrow$ 8 DDL), FP QMC, contours actifs	$?\times?@?$
Zhao *et al.*	[178]	2008	caméra fixe, fond dégagé	(H)AGA + ACP ($44 \rightarrow 6$ DDL)	$320 \times 240@?$

5 Récapitulatif

La capture de mouvement par vision est un problème intrinsèquement difficile. Malgré une littérature extrêmement riche, la démocratisation des applications est encore difficilement envisageable, tant les contextes imposent des contraintes strictes. Des compromis doivent être établis entre réalisme du modèle, qualité de l'estimation souhaitée, et temps de calcul nécessaire. Selon l'application visée, il convient donc de baser l'approche choisie sur des hypothèses simplificatrices raisonnables et adaptées.

Face à ces difficultés et afin de les appréhender graduellement, nous avons choisi de nous confronter à deux cadres différents, ici dénommés contexte multi-oculaire et contexte stéréoscopique, avec pour objectif une intégration sur un des robots du laboratoire. Nous avons situé notre approche par rapport à la littérature et proposé un état de l'art des techniques existantes. Nous choisissons d'investir les techniques stochastiques de filtrage particulaire en raison de leur capacité à modéliser des densités multi-modales. De plus, malgré une complexité parfois importante de ces méthodes, les avancées récentes semblent prometteuses quant aux possibilités d'application temps réel. Nous choisissons un modèle de l'homme compatible avec ces contraintes et présentant un nombre de DDL suffisants pour l'interaction. Nous basons nos filtres sur l'exploitation de mesures « hybrides » 2D et 3D.

Dans cette littérature sur le filtrage particulaire, il n'existe encore que peu d'évaluations quantitatives des techniques développées, et celles-ci se limitent souvent à la comparaison à une stratégie « de référence ». Dans l'optique d'une intégration sur plate-forme mobile, nous souhaitons mener une évaluation plus complète des différentes stratégies envisagées, afin de nous orienter vers la plus adaptée à chacun de nos deux contextes applicatifs. Il convient au préalable de définir plusieurs critères qui nous guideront dans notre démarche. Ces évaluations s'appuieront logiquement sur une vérité de terrain fiable délivrée par un système de HMC commercial.

6 Plan du manuscrit

Le manuscrit est structuré en cinq chapitres incluant le présent état de l'art.

Le chapitre II détaille l'approche que nous avons choisi d'investir et met en exergue quelques caractéristiques propres à nos travaux. Nous présentons le modèle de l'homme que nous adoptons pour le suivi ainsi que les informations visuelles exploitées. Nous présentons enfin le synoptique du système de suivi et abordons les points clés de son implémentation dans l'optique d'une intégration sur les plates-formes robotiques.

Le chapitre III se focalise sur la formalisation du problème de suivi, en rappelant l'approche classique par filtrage particulaire. Plusieurs évolutions de cet algorithme sont présentées, ainsi que leurs avantages et leurs inconvénients. Nous rappelons rapidement les propriétés du filtrage particulaire, et proposons plusieurs métriques inspirées de la littérature. Une première évaluation des performances intrinsèques de chaque stratégie est proposée sur des séquences de synthèse.

Le chapitre IV présente des évaluations plus conséquentes sur des séquences réelles.

Nous décrivons le protocole expérimental mis en place, qui repose sur l'utilisation d'un système commercial de capture de mouvement pour l'obtention de la vérité de terrain. Nous abordons le difficile problème du choix des mesures et mettons en avant les propriétés inhérentes à chacune des techniques exploitées. Nous présentons les difficultés auxquelles est confronté l'utilisateur lors du choix d'une stratégie de filtrage. Nous menons nos évaluations dans deux cadres différents : un contexte multi-oculaire où l'environnement est instrumenté, et un contexte stéréoscopique.

Guidé par ces évaluations, le chapitre V aborde la mise en œuvre du système de suivi sur une plate-forme robotique dédiée. Nous présentons un scénario d'exécution et quelques résultats sur des données acquises par le robot durant l'exécution du scénario en conditions réelles.

Le manuscrit se conclut par une synthèse des travaux réalisés et leurs perspectives.

Chapitre II

Suivi visuel : de la modélisation à l'implémentation

Dans ce chapitre, nous présentons le principe global du système de suivi visuel mis en place. Nous détaillons tout d'abord le modèle de l'homme que nous adoptons, puis, dans un deuxième temps, nous présentons les indices visuels que nous exploitons dans les images afin d'estimer sa configuration. Nous introduisons ensuite l'approche modulaire de l'implémentation du système, articulée autour du module de filtrage. Nous présentons l'intégration des indices visuels au sein d'une approche « hybride », qui couple des mesures 2D et 3D. Enfin, nous consacrons une section au réglage des paramètres libres du système, étape incontournable quoique souvent occultée dans la littérature.

1 Généralités

La problématique du suivi de mouvement est un sujet qui occupe une place importante dans la littérature de la communauté « Vision par Ordinateur ». Le problème est intrinsèquement difficile, et, comme mentionné dans le chapitre I, de nombreuses méthodes ont été proposées. Notre approche se classe parmi les méthodes stochastiques et repose plus précisément sur l'utilisation de filtres particulaires. Cette technique présente l'avantage de pouvoir modéliser les distributions multi-modales, et permet de fusionner des données potentiellement hétérogènes dans un cadre probabiliste. Nous considérons ici une approche « hybride » reposant sur des mesures basées apparence et des informations géométriques 3D.

Nous rappelons que nous nous confrontons à deux contextes sensiblement différents, imposant des contraintes plus ou moins marquées :

- Le contexte multi-oculaire, qui se caractérise par des caméras fixes. Les caméras ont des angles de vue complémentaires, proposant une information riche. L'arrière-plan est connu, et la luminosité varie peu.
- Le contexte de caméras stéréoscopiques embarquées, qui impose des capteurs positionnés sur une plate-forme mobile fournissant des informations plus délicates à exploiter. Les environnements changent au cours des déplacements de la

plate-forme, de même que les conditions d'éclairage.

Bien que ces deux cas demandent un temps de traitement rapide, les contraintes temporelles sont évidemment plus drastiques dans le contexte de caméras embarquées. La finalité est de proposer un cadre générique pour ces deux contextes applicatifs. La première difficulté à laquelle nous sommes alors confrontés est le choix d'un modèle de l'homme adapté à nos applications.

2 Représentation de l'homme

Le modèle de l'homme que nous choisissons ici est décomposé en trois parties : le modèle cinématique, régissant les mouvements possibles par le biais de la définition de contraintes sur chaque articulation, le modèle volumique, caractérisant l'enveloppe corporelle de l'homme, et le modèle d'apparence s'attachant à son apparence image.

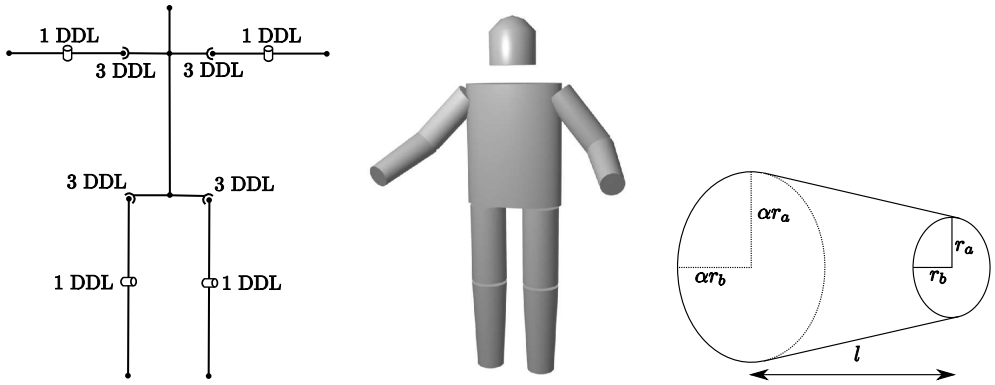
2.1 Modèle cinématique

Définition du modèle

La littérature propose des modèles cinématiques très variés, impliquant un nombre de DDL plus ou moins important. Les modèles les plus simples sont issus de la communauté Robotique, soucieuse des contraintes temps-réel imposées par les applications. Ainsi, dans [7, 9], Azad *et al.* introduisent 14 DDL pour la moitié supérieure du corps humain. La communauté Vision par Ordinateur, plus focalisée sur l'adéquation du modèle à la physionomie humaine, propose des cinématiques impliquant généralement plus de 30 DDL [146]. La plupart des modèles sont constitués de plusieurs chaînes cinématiques ouvertes où chaque membre présente un ou plusieurs DDL par rapport au membre "parent". Cependant, certaines approches [94] préfèrent aborder le suivi de chaque partie du corps de manière indépendante. Chaque membre affiche ainsi six DDL (ou un ensemble restreint) et l'exploration de l'espace des paramètres présente alors une complexité moindre. Cependant, ceci requiert l'utilisation d'indices discriminants permettant de différencier chaque partie du corps. De plus, des contraintes de non-collisions sont généralement ajoutées, impactant sur la puissance de calcul nécessaire.

Afin de rester au maximum compatible avec les contraintes temps-réel que nous nous fixons, et à l'instar de nombreux travaux [9, 160] nous avons privilégié un modèle hiérarchique présentant 22 DDL unissant 9 corps cinématiques. Ce modèle, présenté figure II.1 (a), peut être limité aux 14 DDL de la moitié supérieure pour les applications robotiques, le suivi des jambes étant moins primordial. La modélisation est faite à l'aide des paramètres de Denavit-Hartenberg modifiés [151]. Notre modèle présente trois liaisons prismatiques (q_1, q_2, q_3) — relatives à la translation du sujet dans l'espace — et 19 liaisons rotoïdes réparties en :

- 3 liaisons rotoïdes (q_4, q_5, q_6) permettant l'orientation du torse dans l'espace
- 2×3 liaisons rotoïdes constituant 2 rotules — (q_7, q_8, q_9) et (q_{10}, q_{11}, q_{12}) — pour les épaules



(a) (b) (c)

FIG. II.1 – Les modèles cinématique (a) et volumique (b) exploitant des cônes tronqués (c).

- 2×1 liaisons rotoïdes constituant 2 charnières — q_{13} et q_{14} — pour les coudes
- 2×3 liaisons rotoïdes constituant 2 rotules — (q_{15}, q_{16}, q_{17}) et (q_{18}, q_{19}, q_{20}) — pour les hanches
- 2×1 liaisons rotoïdes constituant 2 charnières — q_{21} et q_{22} — pour les genoux

Une configuration complète du modèle cinématique est alors donnée par le vecteur

$$\boldsymbol{q} = [q_1, q_2, \ldots, q_{21}, q_{22}].$$

La tête est supposée rigidement liée au tronc. En effet, il est relativement complexe d'estimer de manière suffisamment fine l'orientation de la tête pour des personnes se situant à une distance de quelques mètres des caméras, uniquement à partir de l'apparence image de celle-ci.

L'ensemble des paramètres cinématiques du modèle constitue les données que nous cherchons à estimer. En effet, nous choisissons de nous focaliser sur les mouvements effectués par le sujet sans chercher à estimer des paramètres morphologiques. (cf. section 2.2).

Dynamique du modèle

L'analyse spatio-temporelle dans un cadre stochastique suppose de caractériser le modèle de dynamique $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})$ décrivant la connaissance *a priori* que nous avons sur l'évolution temporelle du modèle de l'homme caractérisé par son état $\boldsymbol{x}_k$. La littérature différencie deux approches distinctes.

La première présuppose que l'on a accès à une base de données permettant de caractériser des modèles de mouvements canoniques (marche, course, geste de la main, ...), et que les gestes observés sont proches des mouvements pré-appris [21, 140, 160]. Le filtrage est alors effectué en exploitant cette connaissance *a priori* et en introduisant des modèles de mouvements déduits de ces séquences d'apprentissage. Certains

auteurs [160] utilisent en complément des techniques d'ACP sur ces bases de données afin de réduire la dimension de l'espace de recherche, et ainsi accélérer le processus de suivi.

A contrario, la deuxième classe de méthode ne repose sur aucune connaissance *a priori* des mouvements effectués. La seule hypothèse raisonnable que l'on est alors en droit d'avancer est que l'état $\boldsymbol{x}_k$ à l'instant k est proche de l'état $\boldsymbol{x}_{k-1}$ à l'instant $k-1$. Ceci donne lieu aux modèles bruités de type « marche aléatoire », reposant sur l'utilisation d'une dynamique gaussienne centrée sur l'état à l'instant précédent :

$$p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) = \mathcal{N}(\boldsymbol{x}_k; \boldsymbol{x}_{k-1}, \Delta_k). \tag{II.1}$$

Cette approche englobe les modèles auto-régressifs, tels ceux à « vitesse constante » qui introduisent les dérivées des variables de configuration dans le vecteur à estimer [26, 109]. Cependant, une telle formulation du problème double le nombre de dimensions de l'espace de recherche, ce qui n'est pas toujours compatible avec nos contraintes temps-réel. En conséquence, nous avons choisi de nous restreindre à un modèle de type marche aléatoire non auto-régressif pour représenter les mouvements erratiques de l'homme. Dans notre cas, l'état $\boldsymbol{x}_k$ est ainsi réduit à la configuration articulaire $\boldsymbol{q}$ à l'instant k.

Nous avons proposé un modèle cinématique de l'homme englobant les degrés de libertés articulaires et la dynamique qui leur est associée. Dans la prochaine section, nous abordons la définition de l'enveloppe corporelle qui vient compléter ce modèle.

2.2 Modèle volumique

Définition du modèle

La littérature propose ici encore un grand nombre de choix de modèles différents allant de simples « patches » définis dans un plan [175], à des modèles 3D déformables [58, 87]. Dans [21], Caillette *et al.* modélisent le corps humain comme une densité volumique définie par une mixture de gaussiennes à trois dimensions. Sminchiscescu et Triggs exploitent des volumes maillés déformables [146], proposant ainsi un modèle riche, mais assez lourd à manipuler.

La forme géométrique de notre modèle repose sur l'utilisation de cônes tronqués. Cette approche est assez répandue dans la littérature [8, 107, 160] de par la facilité de manipulation et de projection des primitives géométriques qu'elle permet [9, 109]. Chaque corps cinématique se voit rattacher un ou plusieurs cônes tronqués (figure II.1 (b)). Chaque cône est paramétré par quatre variables (figure II.1 (c)) :

- l, la longueur du tronc de cône ;
- r_a et r_b, les 2 rayons de l'extrémité elliptique de référence ;
- α, le ratio des rayons de chaque extrémité.

Certaines approches choisissent d'inclure les dimensions des primitives géométriques façonnant la silhouette du modèle dans l'ensemble des paramètres à estimer [146]. Nous avons choisi ici de fixer celles-ci afin de ne pas augmenter la dimension de l'espace à explorer durant le processus d'estimation, toujours afin de rester dans les condi-

tions les plus favorables à une application temps-réel. Les dimensions sont choisies selon des considérations anthropomorphiques moyennes afin de s'adapter au mieux au suivi des sujets.

Projection du modèle

Afin de gérer la projection des limbes dans l'image, propre à toute approche basée apparence, nous adoptons le modèle de caméra sténopé, classiquement utilisé dans la communauté Vision par Ordinateur. Le choix de la modélisation géométrique influence directement la facilité de mise en œuvre de la projection associée. Les quadriques et coniques présentent des propriétés rendant quasi immédiate la projection image, ce qui leur vaut une certaine popularité [37, 110, 152, 153]. Le choix de cônes tronqués permet également une projection aisée [9], car basée sur des calculs analytiques, contrairement aux maillages 3D. En effet, la projection image de tels maillages, plus complexe à mettre en œuvre, est classiquement faite au moyen d'un algorithme de « Z-buffer » [174]. Le principe consiste à calculer pour chaque point image la profondeur (Z) de la plus proche surface. Les cartes graphiques sont particulièrement efficaces pour traiter rapidement ce genre de données, cependant la transmission d'information entre la mémoire du GPU (Graphical Process Unit) et celle du CPU (Central Process Unit) est restée pendant longtemps un goulet d'étranglement, bien que leur utilisation soit croissante [179].

Le principe de notre méthode de projection, illustré figure II.2 (a), est inspiré de [9], dans lequel le lecteur pourra trouver les détails : pour une position spatiale donnée d'un cône, on projette les segments de droite situés sur les génératrices dont la normale est orthogonale à la droite de vue. Cette opération est réalisée sur l'ensemble des cônes décrivant le modèle, donnant lieu à la projection complète, présentée figure II.2 (b).

Notre méthode de projection permet également, le cas échéant, la prise en compte des parties cachées. Un tel calcul impose un temps de traitement supplémentaire mais peut s'avérer nécessaire pour l'exploitation de certains indices visuels (cf. section 3.1). L'algorithme est inspiré de celui proposé dans [106] : il consiste à calculer les intersections de tous les segments projetés, et à tester la visibilité de leurs points centraux, ceci garantissant la visibilité ou non du segment dans son intégralité.

2.3 Modèle d'apparence

En complément des simples volumes des cônes tronqués, nous positionnons également cinq points virtuels $\boldsymbol{p}_t$, $\boldsymbol{p}_{mg}$, $\boldsymbol{p}_{md}$, $\boldsymbol{p}_{pg}$, $\boldsymbol{p}_{pd}$, caractérisant respectivement le centre de gravité de la tête, des mains gauche et droite et des pieds gauche et droit. Ces marqueurs sont exploités lors de la localisation des extrémités de membres dans les images. En effet, les mains par exemple se trouvent dans des zones de l'image présentant une teinte chair. Le visage également, et un détecteur peut aider à sa localisation. Ces points présentent donc une importance particulière dans le suivi du sujet.

Enfin, certaines zones — ou « patches »— définies sur les surfaces des cônes se voient rattacher une distribution de couleur décrivant l'apparence de leur projection

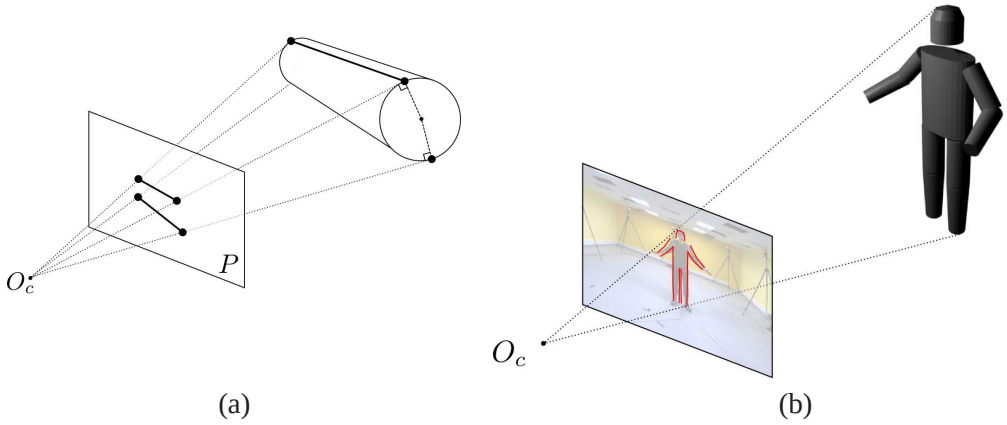


FIG. II.2 – Projection perspective d'un cône (a) et du modèle complet (b).

dans l'image. Cette distribution est modélisée par un histogramme sur les trois canaux chromatiques (rouge, vert et bleu). Ceci permet d'associer une information de couleur à différents points précis du modèle (torse, bras, ...).

2.4 Bilan et compléments

La figure II.3 présente le modèle complet de l'homme que nous utilisons (modèle cinématique, volumique et d'apparence). Afin de prendre en compte les contraintes physiques interdisant un certain nombre de mouvements, l'ensemble des DDL cinématiques est limité à une plage de valeur données. La gestion des collisions entre membres mène quant à elle généralement à des algorithmes de complexité quadratique en fonction du nombre de parties. Pour des raisons de contraintes de temps réel, nous choisissons ici de ne pas utiliser ce type d'approche. À ces fins, nous mettons en place des potentiels répulsifs entre les parties « sensibles », *i.e.* la tête et les mains dans notre cas, qui présentent des couleurs et formes relativement proches.

Au delà des mains, toutes les extrémités corporelles apportent des contraintes intéressantes. Ainsi, nous introduisons une contrainte de distance au sol minimale, imposant aux pieds de se situer à proximité du plan du sol Π_{sol}. Ceci est rendu possible grâce à une calibration préalable intrinsèque et extrinsèque des caméras.

Après avoir détaillé le modèle de l'homme et ses particularités, nous abordons les différents indices visuels reposant sur l'exploitation du modèle précédemment décrit afin d'estimer au mieux la position 3D du sujet à partir des images.

3 De la représentation vers l'analyse

Afin de suivre les mouvements de l'homme à partir des images acquises, le modèle est projeté pour différentes hypothèses de configuration que nous confrontons aux

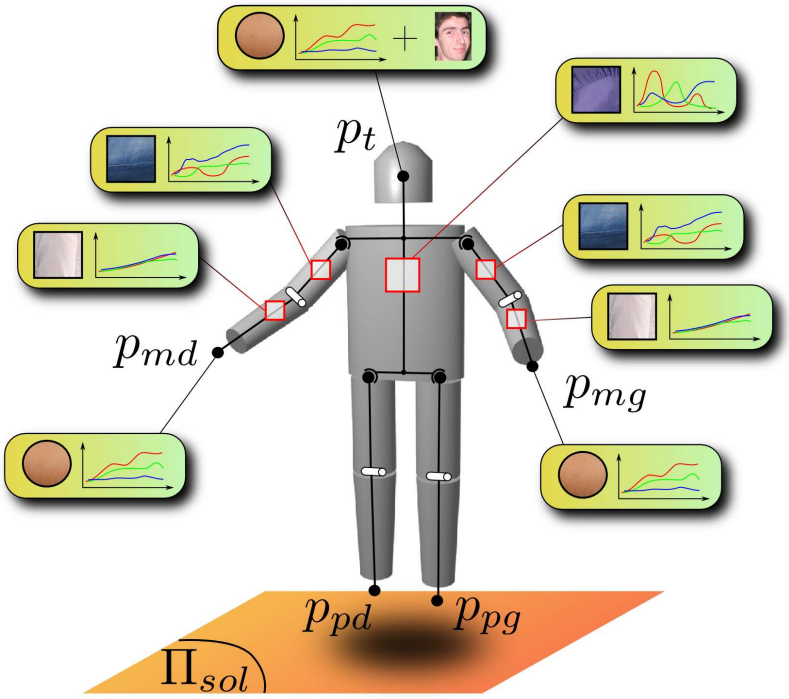


FIG. II.3 – Le modèle complet de l'homme incluant le modèle cinématique à 22 DDL, le modèle volumique à base de cônes tronqués et le modèle d'apparence constitué des « patches » de couleurs et des points virtuels identifiant la tête, les mains et les pieds.

(a) (b) (c)

FIG. II.4 – Exploitation de la silhouette : (a) image source ; (b) image du masque de segmentation (I_{sil}) dans laquelle le modèle projeté est échantillonné (en jaune) ; (c) masque de segmentation échantillonné (en vert) et configuration du modèle projeté (en rouge).

images sur la base de plusieurs mesures. Cette approche « basée apparence » exploite la richesse des informations délivrées par le flot vidéo. Nous la complétons par une utilisation plus ponctuelle de mesures 3D, rendant le suivi plus robuste. La littérature propose des indices visuels reposant sur la forme, la couleur et le mouvement. Nous présentons dans les lignes suivantes plusieurs distances notées $D_{(.)}$ mesurant l'adéquation entre une hypothèse de configuration $\boldsymbol{x}_k$ du modèle de l'homme et les images à l'instant k.

3.1 Indices visuels de forme

Silhouette

Une mesure très répandue dans le domaine de la surveillance est celle qui consiste à comparer la silhouette segmentée de la personne dans les images à la silhouette correspondant à la projection du modèle sous l'hypothèse $\boldsymbol{x}_k$ [38, 147]. Le principe est présenté en figure II.4 (a) et (b). Soient $\boldsymbol{p}^i_{sil}, i \in 1..N_{sil}$ des points échantillonnés uniformément à l'intérieur de la projection du modèle sous hypothèse $\boldsymbol{x}_k$. La distance correspondante s'écrit alors :

$$D_{sil} = \frac{1}{N_{sil}} \sum_{i=1}^{N_{sil}} (1 - I_{sil}(\boldsymbol{p}^i_{sil})), \tag{II.2}$$

où I_{sil} désigne le masque de la silhouette du sujet, avec $I_{sil}(\boldsymbol{p}) = 1$ si le point $\boldsymbol{p}$ est situé dans la silhouette, 0 sinon. Cette fonction de vraisemblance n'est cependant exploitable que dans la mesure où une segmentation de la silhouette est réalisable, typiquement en contexte multi-oculaire. Le principe repose sur un algorithme de soustraction de fond classique, et une heuristique pour gérer les phénomènes d'ombrage dus au sujet : les pixels présentant une faible baisse de luminosité identique sur les trois canaux chromatiques sont considérés comme appartenant au fond. L'image résultat est classiquement filtrée par des opérateurs de morphologie mathématique. Notons que sur toutes les figures présentées par la suite, D_{sil} varie entre 0 et 255 et non pas 0 et 1 pour des raisons d'implémentation.

(a) (b) (c)

FIG. II.5 – Étapes de la mise en place de la mesure de contours : (a) image source ; (b) image de contours obtenue à l'aide d'un filtre de Canny ; (c) projection (en rouge) et échantillonnage (en jaune) des limbes de la silhouette dans l'image de distance.

Silhouette duale

La seule mesure précédente privilégie les hypothèses de configuration telles que la silhouette du modèle se trouve à l'intérieur de la silhouette segmentée I_{sil}. Nous proposons, à l'instar de [145], une mesure duale pour contrebalancer cet effet. Le principe est d'échantillonner N_{sil2} points $\boldsymbol{p}^i_{sil2}, i \in 1..N_{sil2}$ dans l'image segmentée de la silhouette I_{sil} (cf. figure II.4 (c)). Nous utilisons pour cela la méthode d'acceptation-rejet [42] qui consiste à échantillonner toute l'image de manière uniforme et à ne conserver que les échantillons situés à l'intérieur du masque. La distance duale s'écrit alors :

$$D_{sil2} = \frac{1}{N_{sil2}} \sum_{i=1}^{N_{sil2}} (1 - f_{in}(\boldsymbol{p}^i_{sil2}, \boldsymbol{x}_k)), \tag{II.3}$$

où $f_{in}(\boldsymbol{p}^i_{sil2}, \boldsymbol{x}_k) = 1$ si le point $\boldsymbol{p}^i_{sil2}$ se situe à l'intérieur de la projection du modèle sous l'hypothèse $\boldsymbol{x}_k$, 0 sinon. L'efficacité de cette mesure, tout comme la précédente, est soumise à la qualité de la segmentation de l'objet recherché.

Image de contour

Un indice visuel relativement intéressant dans des environnements peu encombrés est l'utilisation des contours, schématisée en figure II.5. Blake *et al.* dans [13] l'utilisent pour du suivi 2D. Nous étendons ici l'approche aux images de distances, à l'instar de [59, 61]. Le principe est d'échantillonner les limbes du modèle projeté sous l'hypothèse $\boldsymbol{x}_k$ en $N_{contours}$ points $\boldsymbol{p}^i_{contours}, i \in 1..N_{contours}$ et de favoriser les configurations dont les limbes sont proches des contours images :

$$D_{contours} = \frac{1}{N_{contours}} \sum_{i=1}^{N_{contours}} I_{dist}(\boldsymbol{p}^i_{contours}), \tag{II.4}$$

où I_{dist} est l'image de distance aux contours. Un filtre de Canny [22] est appliqué sur l'image source afin de détecter les contours, puis une transformation en distance est appliquée afin d'obtenir I_{dist}. La valeur de chaque pixel est alors proportionnelle à la

distance au contour le plus proche. Notons que cet indice visuel nécessite la gestion des parties cachées afin de ne pas échantillonner des limbes non visibles dans l'image.

Ce calcul pourrait être effectué directement sans passer par l'utilisation de I_{dist}, mais dans un contexte de filtrage particulaire où la fonction de vraisemblance est évaluée plusieurs centaines/milliers de fois par image, ce principe permet un gain de temps de calcul. Dans notre cas, pour une image de 640×480 pixels, Azad [9] établit une équivalence entre les deux méthodes (image de distance et calcul de distance complet pour chaque pixel échantillonné) pour 300 hypothèses de configurations différentes par image, ses considérations reposant sur un modèle de la moitié supérieure du corps. Pour 1000 hypothèses sur des images 320×240, le gain est de 25% en temps de calcul.

En complément des indices de forme, nous utilisons également l'information de couleur.

3.2 Indices visuels de couleur

Patches de couleur

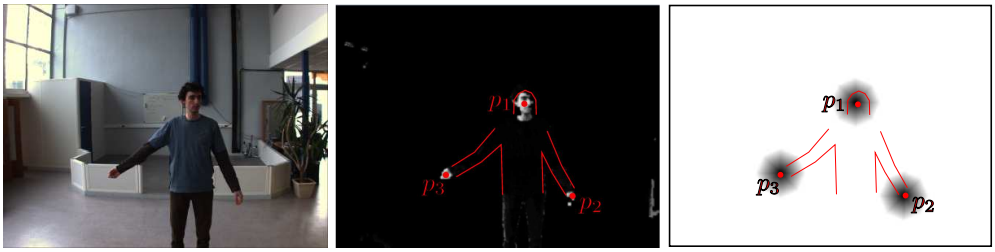
À l'instar de [17, 129], nous exploitons les zones spécifiques définies sur le modèle. Ces $N_{patches}$ « patches » codent les distributions de couleur $h_i^{ref}, i \in 1..N_{patches}$ associées à autant de zones surfaciques du modèle de l'homme et sont initialisés sur une configuration connue — typiquement sur la première image de la séquence. Pour chaque hypothèse $\boldsymbol{x}_k$, les coordonnées de ces zones sont reprojetées dans l'image et les distributions de couleur $h_i^{\boldsymbol{x}_k}$ correspondantes sont comparées au modèle pré-appris *via* :

$$D_{patches} = \frac{1}{N_{patches}} \sum_{i=1}^{N_{patches}} d_{Bhatta}(h_i^{ref}, h_i^{\boldsymbol{x}_k}), \tag{II.5}$$

où d_{Bhatta} dénote la distance de Bhattacharyya [3] permettant la comparaison de deux distributions de couleur modélisées par des histogrammes sur chaque canal R,V,B.

Distance aux blobs de couleur peau

Les mains sont d'ordinaire assez difficiles à localiser précisément, de par leur petite taille dans l'image, et leur vitesse importante par rapport au reste du corps [120]. De plus, elles constituent les extrémités des chaînes cinématiques représentant les membres supérieurs et apportent donc des contraintes sur l'ensemble de ces chaînes. Elles constituent également le moyen d'interaction privilégié de l'homme, et doivent en ce sens revêtir une grande importance dans les fonctions de vraisemblance. Une approche préliminaire [56] repose sur l'utilisation de l'image de probabilité de couleur peau I_{peau}, calculée par rétroprojection d'un histogramme de couleur pré-appris sur une base de données de teinte chair [84, 85]. Le centre de la tête et des mains $\boldsymbol{p}_t$, $\boldsymbol{p}_{mg}$ et $\boldsymbol{p}_{md}$ sont alors projetés sous l'hypothèse $\boldsymbol{x}_k$ dans l'image en $\boldsymbol{p}_1, \boldsymbol{p}_2$ et $\boldsymbol{p}_3$ respectivement (cf. figure II.6 (a) et (b)), et l'on échantillonne I_{peau} en ces points selon :



(a) (b) (c)

FIG. II.6 – Étapes de la mise en place de la mesure de couleur peau : (a) image source ; (b) image de probabilité de couleur peau I_{peau} ; (c) modèle projeté dans I_{dist_peau}.

$$D_{peau} = \frac{1}{3}\sum_{i=1}^{3} I_{peau}(\boldsymbol{p}_i). \tag{II.6}$$

Cependant, une telle approche produit une distance particulièrement piquée autour de l'optimum et par là même peu exploitable — il suffit que la projection des mains et de la tête ne se situe pas exactement sur les zones de couleur peau pour que la distance soit très grande —. Une alternative consiste donc à exploiter l'image de distance aux blobs de couleur peau I_{dist_peau} qui fournit une information plus lissée (figure II.6 (c)). On calcule alors la distance moyenne qui sépare chaque point projeté du blob de couleur peau le plus proche :

$$D_{dist\_peau} = \frac{1}{3}\sum_{i=1}^{3} I_{dist\_peau}(\boldsymbol{p}_i), \tag{II.7}$$

où $I_{dist\_peau}(\boldsymbol{p}_i)$ représente la valeur du pixel situé en $\boldsymbol{p}_i$ dans l'image de distance I_{dist_peau}.

Couleur uniforme

Cette mesure repose sur N_m ensembles disjoints $E_i, i \in 1..N_m$ de points uniformément échantillonnés dans chacun des N_m membres projetés pour une configuration $\boldsymbol{x}_k$.

Nous supposons que la personne suivie porte des vêtements présentant une couleur homogène sur chaque membre. Nous exploitons alors la mesure suivante :

$$D_{uni} = \frac{1}{N_m}\sum_{i=1}^{N_m}\left(\frac{1}{3}\sum_{c\in\{R,G,B\}}\sigma_{E_i,c}\right), \tag{II.8}$$

où $\sigma_{E_i,c}$ est 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.

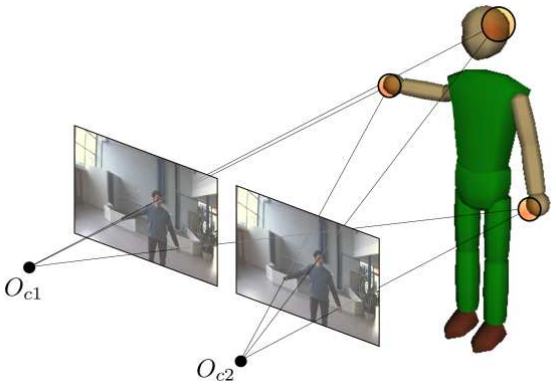


FIG. II.7 – Principe de la triangulation des blobs de couleur peau à partir de deux caméras stéréo.

Toutes les mesures présentées ci-dessus sont par définition 2D et restent donc moins discriminantes que les mesures 3D.

3.3 Contraintes géométriques

Distance aux blobs 3D

Afin de compléter cet ensemble de mesures 2D, nous introduisons une mesure 3D impliquant une reconstruction éparse, et permettant une localisation plus précise de la tête et des mains. À partir de l'image de probabilité de couleur peau I_{peau} calculée pour chaque caméra, nous obtenons par un simple seuillage plusieurs blobs 2D dans chaque image. Dans un contexte stéréoscopique, ces blobs sont ensuite appariés suivant des critères topographiques et géométriques définis dans [139], *i.e.* taille et forme des blobs, et similarités des blobs voisins. La position 3D des blobs $\boldsymbol{p}^i_{blobs}, i \in 1..N_{blobs}$ est ensuite triangulée grâce à la connaissance des paramètres intrinsèques et extrinsèques des caméras (cf. figure II.7). La distinction entre la tête et les mains est faite selon de simples heuristiques. Elle s'appuie sur un détecteur de visage basé sur l'utilisation de masques de Haar [163]. On note alors :

$$D_{blobs} = \frac{1}{3} \sum_{i=\{t,md,mg\}} ||\boldsymbol{p}_i - \boldsymbol{p}^{j^i}_{blobs}||_2. \tag{II.9}$$

Le lien entre i et j^i, permettant par exemple d'associer la tête du modèle au blob 3D triangulé correspondant, est réalisé au moyen des heuristiques précédemment décrites. Lorsque celles-ci ne sont pas applicables (mauvaise triangulation des blobs, nombre de blobs trop important, pas de détection de visage,...), le blob 3D le plus proche est utilisé. L'échec occasionnel de certaines des procédures impliquées rend également cet indice intermittent. Notons que les contraintes imposées au modèle cinématique pour prévenir les collisions entre membres et pour privilégier les configurations dont les

(a) (b) (c)

FIG. II.8 – (a) Image de contours classique ; (b) image de mouvement ; (c) image de contours pondérées par l'image de mouvement.

pieds sont proches du sol peuvent elles-mêmes être interprétées comme des mesures géométriques 3D.

3.4 Exploitation du mouvement

Les images de contours présentent parfois de nombreux artefacts. Elles peuvent alors être pondérées par un masque de mouvement I_{mouv}. Dans le cas de caméras fixes, ce masque est généralement calculé par la soustraction d'images consécutives à laquelle une opération de seuillage classique est ensuite appliquée. Ceci s'avère utile dans les cas où une segmentation par soustraction de fond ne peut pas être réalisée, comme présenté en figure II.8.

Les informations fournies par les images sont donc particulièrement riches et sont exploitables de plusieurs manières complémentaires. Toutefois, certaines dépendent du contexte d'application envisagé.

3.5 Bilan

Les différents indices visuels présentés ci-dessus ne peuvent pas toujours être exploités. Le tableau II.1 dresse un bilan de leur utilité selon le contexte. Il est à noter que les mesures reposant sur une segmentation du fond (silhouette et silhouette duale) peuvent difficilement être exploitées dans un contexte robotique où les capteurs embarqués sur le robot sont amenés à bouger. L'utilisation de patches de couleurs nécessite quant à elle de connaître avec certitude la configuration de l'homme à un instant donné afin d'initialiser l'apparence image des zones d'intérêt définies sur le modèle. Ceci pose moins de problèmes dans un contexte de capture de mouvement où l'initialisation peut être effectuée à la main, et où nous sommes assurés de la cohérence entre modèle projeté et image. Enfin, la localisation 3D des blobs de couleur peau (idéalement le visage et les mains) est applicable au contexte stéréoscopique, mais plus complexe à mettre en œuvre avec des caméras d'ambiance. En effet, l'étape d'appariement des blobs précédant la triangulation est bien plus difficile lorsque les caméras présentent des angles de vue très différents. Nous pouvons toutefois appliquer une méthode « gloutonne »

Type	Mesure	Multi-oculaire	Stéréoscopique
Forme	Silhouette	OUI	NON
	Silhouette Duale	OUI	NON
	Contours	OUI	OUI
Couleur	Patches	OUI	$\sim$
	Blobs peau 2D	OUI	OUI
	Couleur uniforme	OUI	OUI
Géométrique	Blobs peau 3D	$\sim$	OUI
	Non collision	OUI	OUI
	Pieds au sol	OUI	OUI
Mouvement	Soustraction de fond	OUI	NON
	Images successives	OUI	OUI

TAB. II.1 – Récapitulatif des fonctions de mesures et de leurs domaines d'application.

consistant à tester toutes les correspondances possibles. Ceci peut générer de fausses mesures, qui diminuent toutefois avec le nombre de caméras utilisées.

Les distances décrites précédemment permettent d'évaluer la consistance d'une hypothèse $\boldsymbol{x}_k$ du modèle vis-à-vis de l'image. Nous présentons dans la prochaine section leur intégration au système de suivi complet, ainsi que la démarche adoptée en vue d'une implémentation sur une plate-forme robotique.

4 Implémentation

4.1 Synoptique

Dans l'optique de proposer un système intégrable sur un robot, nous avons privilégié une approche modulaire. Le schéma figure II.9 en présente un synoptique. Notre système se décompose ainsi en cinq modules principaux :

- Un module de **pré-traitement des images**, qui a pour rôle de restaurer ou d'améliorer les images avant leur analyse. L'ensemble des opérations effectuées est présenté en annexe B.
- Un module de **projection** qui, pour une configuration donnée du modèle de l'homme en effectue la projection image selon la méthode présentée en section 2.2.
- Un module de **filtrage**, qui implémente plusieurs stratégies de filtrage particulaire détaillées dans le chapitre III ; il constitue le cœur du système et fait appel à tous les autres modules.
- Un module de **traitement des images**, en charge de la production d'image de probabilité, d'image de contour, ou proposant des utilitaires tels des détecteurs de visage et de blobs.

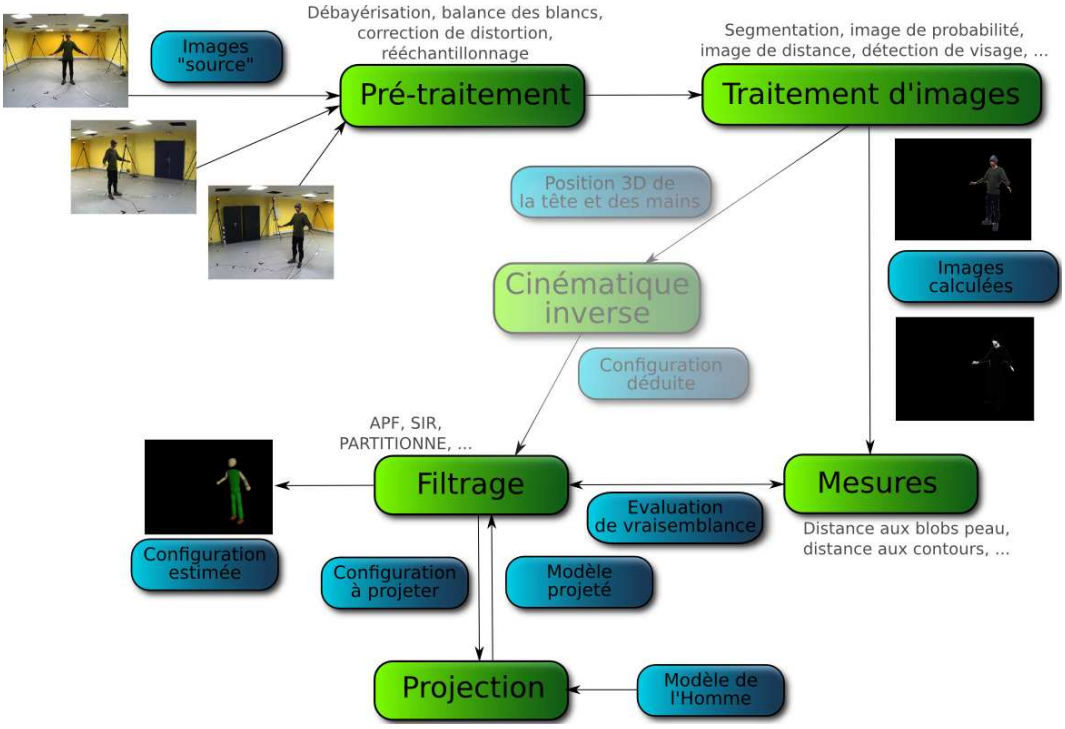


FIG. II.9 – Synoptique de notre approche.

– Un module de **mesures**, qui repose sur l'utilisation des images calculées par le module de traitement, et qui effectue l'évaluation de chaque indice visuel présenté précédemment pour une configuration donnée du modèle de l'homme.

On peut noter la présence d'un module de cinématique inverse qui n'est cependant pas systématiquement utilisé. Il a pour but de fournir une configuration plausible de la pose 3D du sujet au module de filtrage à partir d'une localisation de la tête et des mains dans l'espace lorsque celle-ci est possible.

Le principe global de l'approche, au travers du module de filtrage, consiste à proposer, à chaque instant k, plusieurs hypothèses quant à la configuration $\boldsymbol{x}_k$ du modèle de l'homme dans l'espace et à confronter leur projection à l'image (aux images), dénotées $\boldsymbol{z}_k$, *via* le module de mesure. Les configurations présentant une grande cohérence vis-à-vis des images selon les critères définis dans la section 3 se voient affectées une grande confiance. De manière complémentaire, les autres configurations moins pertinentes revêtent une importance moindre. Cette évaluation de l'adéquation entre une hypothèse et la mesure est formalisée sous le terme de « fonction de vraisemblance », et notée $p(\boldsymbol{z}_k|\boldsymbol{x}_k)$. Le choix des indices visuels impliqués dans ce procédé est capital pour assurer un comportement satisfaisant du filtre. La génération des hypothèses de configuration du modèle quant à elle n'est pas faite en « aveugle » mais en exploitant les informations de la configuration $\boldsymbol{x}_{k-1}$ liées à l'instant précédent ainsi que les mesures courantes $\boldsymbol{z}_k$: on a recours à une fonction dite « d'importance », notée $q(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{z}_k)$. L'idée sous-jacente à l'utilisation d'une fonction d'importance est de pouvoir placer les hypothèses de la manière la plus pertinente possible. La théorie du filtrage particulaire sera abordée plus en détails dans le chapitre III.

Le choix des fonctions d'importance et de vraisemblance est donc au cœur du problème de suivi par filtrage particulaire, aussi les deux prochaines sections détaillent-elles l'implémentation de ces fonctions-clés.

4.2 Fonction d'importance

Il n'existe, à notre connaissance, que peu d'approches qui tentent d'exploiter les images dans la fonction d'importance [61, 82, 128], et celles-ci restent majoritairement 2D. Pour être à même de proposer un ensemble de configurations susceptibles de se situer dans des pics de fonctions de vraisemblance, nous avons logiquement exploité l'information 3D issue des positions de la tête et des mains, respectivement dénotées $\boldsymbol{p}_t, \boldsymbol{p}_{md}$ et $\boldsymbol{p}_{mg}$.

Afin d'induire une configuration du modèle de l'homme à partir de ces points, nous mettons en place un algorithme de cinématique inverse analytique. La littérature propose classiquement des approches génériques numériques [20], parfois instables de par la présence de points singuliers autour desquels les calculs de jacobiens posent problème. Dans certains cas plus particuliers (nombre de DDL restreint), des méthodes analytiques peuvent être exploitées [157]. C'est l'approche que nous avons choisie, toutefois, la méthode utilisée ici n'a pas pour but de résoudre ce problème (difficile) de

manière complète et rigoureuse. La finalité est d'aboutir à une configuration approchée, autour de laquelle l'espace des solutions sera exploré plus finement, ceci dans le but de focaliser les efforts de calcul dans les zones pertinentes de l'espace des configurations. C'est pourquoi nous privilégions une technique simple calculant la configuration unique qui vérifie $q_1 = q_2 = q_3 = q_9 = q_{12} = 0$, *i.e.* une configuration au torse vertical et avec les coudes le plus bas possible (pliés vers le haut).

Notons que notre méthode ne permet pas de proposer une configuration pour les jambes. Nous supposons donc ici qu'elles restent droites et verticales. En outre, à l'instar de la mesure proposée en 3.3, il n'est pas toujours possible de proposer une configuration issue de la méthode de cinématique inverse ; cette fonction d'importance n'est donc pas mise en œuvre systématiquement à chaque instant image. Dans ce cas, les hypothèses de configuration sont positionnées en exploitant la seule dynamique du système. D'une manière générale, notons que les fonctions d'importance font classiquement intervenir des indices visuels discriminants mais intermittents [128], ceci dans l'optique de permettre une (ré-)initialisation du filtre lorsque celui-ci perd sa cible.

4.3 Fonctions de vraisemblance

Les fonctions de vraisemblance ont pour rôle d'évaluer la pertinence d'une hypothèse de configuration $\boldsymbol{x}_k$ vis-à-vis d'une image (ou d'un ensemble d'images) $\boldsymbol{z}_k$ à l'instant k. Elles impliquent plusieurs « distances de similarité » $D_{(.)}$ — caractérisant la distance entre l'observation courante et l'hypothèse de configuration émise —, dont il est admis qu'elles se distribuent suivant une loi normale autour de 0 :

$$p(\boldsymbol{z}_k^{(.)}|\boldsymbol{x}_k) \propto e^{\frac{-D_{(.)}^2}{2\sigma_{(.)}^2}}. \tag{II.10}$$

Elles reposent ici sur l'utilisation des distances $D_{(.)}$ décrites en section 3. Nous supposons enfin que toutes ces mesures sont indépendantes entre elles conditionnellement à $\boldsymbol{x}_k$ *i.e.* :

$$p(\boldsymbol{z}_k^1, \ldots, \boldsymbol{z}_k^Z|\boldsymbol{x}_k) = \Pi_{y=1}^Z p(\boldsymbol{z}_k^y|\boldsymbol{x}_k). \tag{II.11}$$

Cette hypothèse d'indépendance conditionnelle permet de fusionner des mesures potentiellement hétérogènes (cf. figure II.10). Il est ainsi possible de mettre en place une fonction de vraisemblance globale pertinente (figure II.10 (e)) à partir de mesures qui apportent des informations complémentaires (figures II.10 (b)-(d)). Cette propriété est particulièrement appréciable dans un cadre de capteurs embarqués, puisqu'il est alors possible de coupler les caméras avec d'autres types de capteur. Dans notre cadre, nous couplons des mesures basées apparence (silhouette D_{sil}, contours $D_{contours}$, patches de couleurs $D_{patches}$, ...) et une mesure 3D (triangulation des blobs de couleur peau D_{blobs}).

Si le choix des indices visuels impliqués s'avère primordial, le choix des paramètres $\sigma_{(.)}$ impliqués dans la définition des fonctions de vraisemblance l'est tout autant, alors que la littérature est peu loquace sur le problème [99].

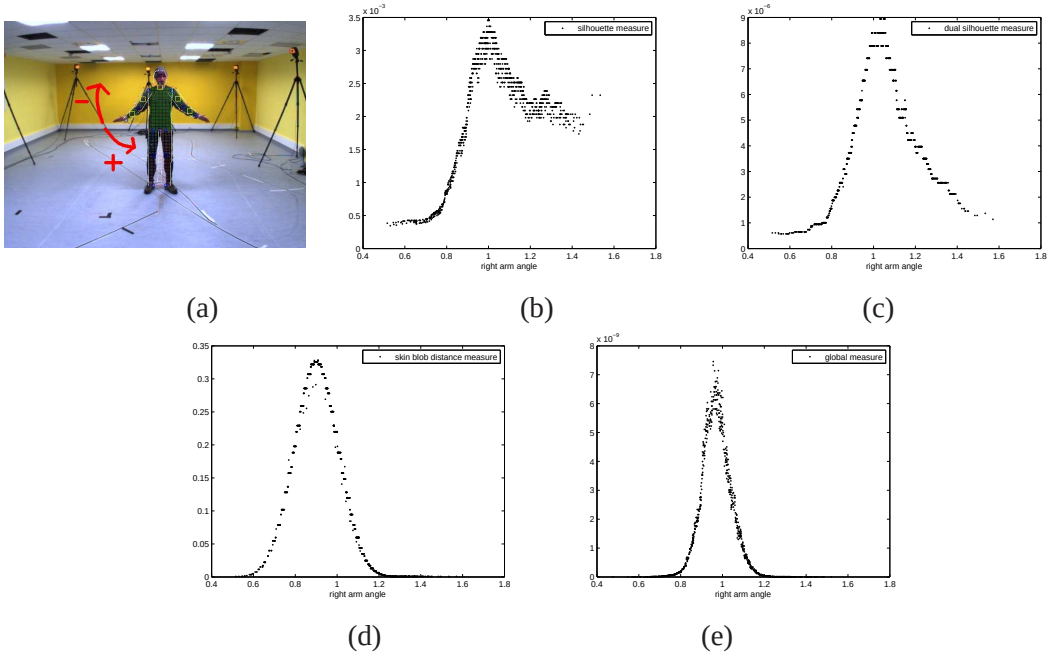
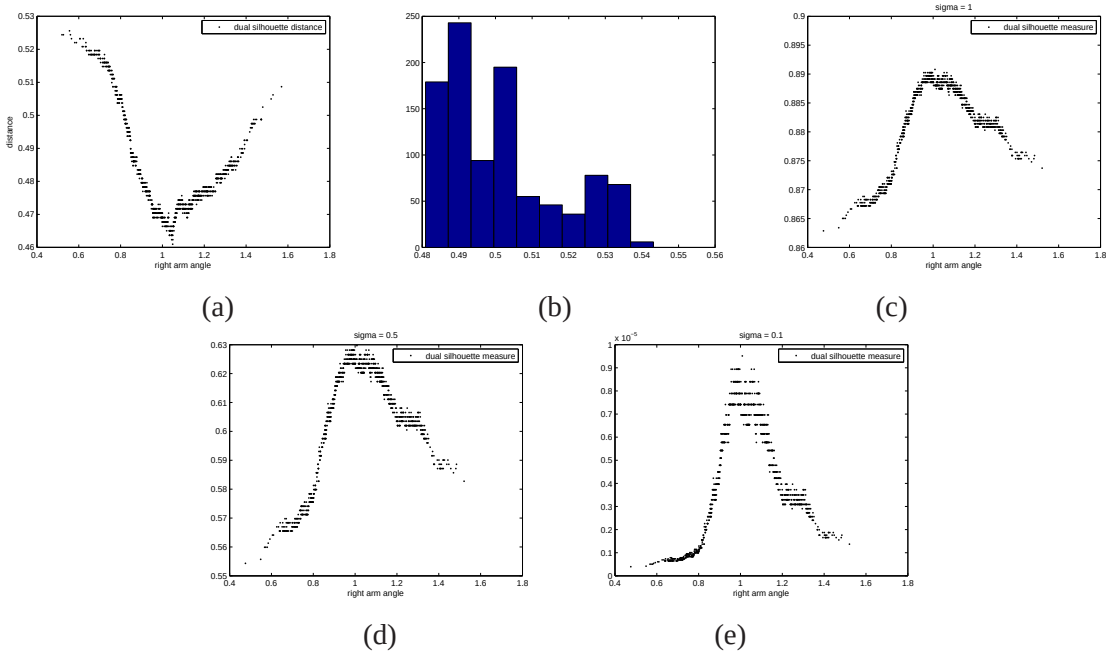


FIG. II.10 – (a) Exemple de mouvement du bras suivant 1 DDL ; (b)-(d) fonctions de vraisemblance $p(\boldsymbol{z}_k^{sil}|\boldsymbol{x}_k)$, $p(\boldsymbol{z}_k^{sil2}|\boldsymbol{x}_k)$, $p(\boldsymbol{z}_k^{dist\_peau}|\boldsymbol{x}_k)$ en fonction de l'orientation du bras ; (e) fonction de vraisemblance issue de la fusion des indices précédents $p(\boldsymbol{z}_k^{sil,sil2,dist\_peau}|\boldsymbol{x}_k)$.

Nous consacrons ainsi les paragraphes suivants à une présentation du comportement des filtres et présentons quelques heuristiques pour le choix de ces paramètres libres.

4.4 Comportement et paramétrisation des fonctions de vraisemblance



(a) (b) (c)

(d) (e)

FIG. II.11 – (a) Évolution de D_{sil2} pour un mouvement du bras droit à 1 DDL autour de la configuration réelle ; (b) histogramme associé des valeurs de D_{sil2} ; (c), (d) et (e) fonctions de vraisemblance pour les valeurs de $\sigma = 1.0$, $\sigma = 0.5$ et $\sigma = 0.1$ respectivement.

Une première sélection des paramètres $\sigma_{(.)}$ impliqués dans les fonctions de vraisemblance est effectuée sur la base d'heuristiques simples puis affinées empiriquement afin d'arriver à une configuration relativement stable. Dans le cas idéal où le modèle sous l'hypothèse de configuration $\boldsymbol{x}_k$ est en parfait accord avec les images, la distance de similarité $D_{(.)}$ « idéale » devrait être nulle. En pratique, ce cas ne se produit jamais ; par conséquent, nous devons choisir des valeurs de σ équilibrées et adaptées au contexte. De plus, il convient de pondérer chaque mesure par rapport aux autres en fonction de leur pouvoir discriminant. Quelques heuristiques guidant les choix initiaux sont ici présentées.

A - Distance de similarité – L'évolution d'une distance donnée $D_{(.)}$ en fonction des variations du vecteur d'état $\boldsymbol{x}_k$ apporte une première information sur la plage de valeurs « utiles ». La figure II.11 (a) présente l'évolution de D_{sil2} pour plusieurs orientations du bras du modèle autour de la configuration réelle (expérience décrite figure II.10 (a)).

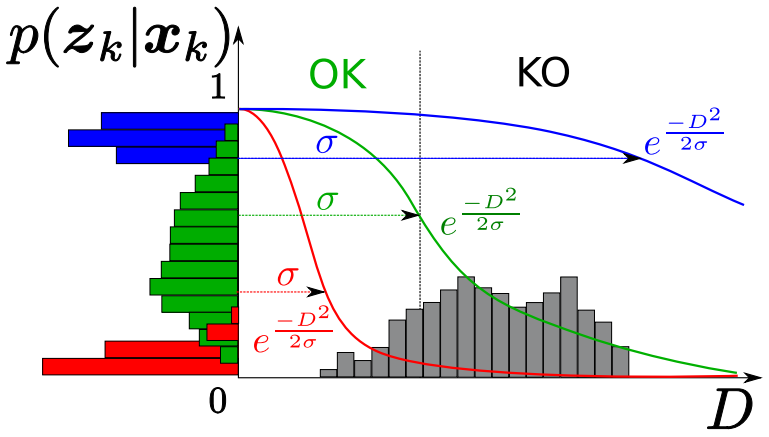


FIG. II.12 – Vraisemblances $p(\boldsymbol{z}_k|\boldsymbol{x}_k)$ associées à un ensemble d'hypothèses $\boldsymbol{x}_k$ (en couleurs) en fonction de l'histogramme des distances de similarité associées (en gris) pour différentes valeurs de $\sigma_{(.)}$.

La courbe présente un minimum pour l'orientation la plus proche de la configuration réelle. Les valeurs des distances varient entre 0.48 et 0.54 (figure II.10 (b)). On se limite ici à un mouvement à 1 DDL pour présenter les courbes de manière lisible. Toutefois, de manière plus générale, on peut également observer l'histogramme caractérisant la répartition des distances de similarité pour un nuage de particule donné quelque soit le nombre de DDL impliqués, comme schématisé en figure II.12.

Ces informations sur le domaine d'évolution des distances de similarité permettent le choix d'une première valeur de $\sigma_{(.)}$ située dans cette plage. Ainsi, la majorité des échantillons présente une vraisemblance située dans la pente de la gaussienne associée comme illustré en vert sur la figure II.12. En effet, une gaussienne $\mathcal{N}(x;\mu,\sigma^2)$ présente une pente maximale pour $x=\sigma$. Ainsi, dans notre exemple, en choisissant $\sigma_{sil2}\sim 0.5$, on façonne la fonction de vraisemblance de telle sorte qu'elle entraîne une dispersion des poids maximale compte tenu de la zone utile dans laquelle se situent les distances de similarité. Intuitivement, ce choix correspond à favoriser les particules qui affichent une distance de similarité $D_{(.)}<\sigma_{(.)}$ au détriment des autres.

B - Ajustement empirique – Nous présentons en figure II.13 un suivi réalisé avec une configuration guidée par l'heuristique précédente. Nous constatons que le comportement du système n'est pas satisfaisant. Le choix des $\sigma_{(.)}$ n'accorde pas suffisamment d'importance aux mesures. Nous devons affiner la configuration en diminuant leur valeur. Les figures II.11 (c), (d) et (e) présentent la fonction de vraisemblance $p(\boldsymbol{z}_k^{sil2}|\boldsymbol{x}_k)$ pour différentes valeurs de σ_{sil2} ($1.0, 0.5$ et 0.1). Lorsque σ_{sil2} décroît, la fonction de vraisemblance présente un pic plus marqué, ce qui résulte en une sélection plus drastique de la particule la plus vraisemblable au regard des mesures. Cependant, la valeur des poids calculés décroît également, puisque les particules se trouvent alors dans la queue de la gaussienne de la fonction de vraisemblance (cf données en rouge sur la

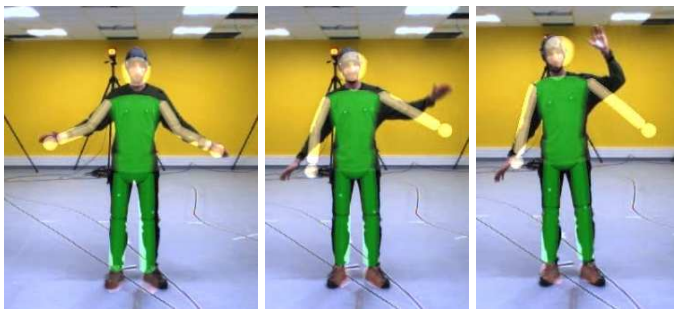


FIG. II.13 – Suivi sur une séquence simple avec $\sigma_{sil} = 130$, $\sigma_{sil2} = 0.5$ et $\sigma_{dist\_peau} = 30$. L'avatar en surimpression représente la configuration estimée. Le bras gauche décroche en milieu de séquence.

figure II.12). Ceci peut amener au problème suivant.

C - Limites calculatoires – La valeur de σ ne peut pas être diminuée indéfiniment : une trop petite valeur induit des poids nuls de par la limite d'encodage des nombres flottants. Ceci est visible sur la figure II.11 (e) où la limite de précision de calcul entraîne une discrétisation pauvre de la fonction de vraisemblance. En effet, à titre d'exemple, $\exp\left(\frac{1}{2}\frac{0.5^2}{0.01^2}\right) = 0$, c'est-à-dire que choisir une valeur de $\sigma_{sil2} = 0.01$ conduit à une fonction de vraisemblance nulle, et par la suite à une divergence du filtre.

D - Équilibre – Le phénomène mentionné ci-dessus est amplifié par le nombre d'indices visuels et de caméras impliqués puisque les fonctions de vraisemblance sont alors multipliées entre elles. Il faut ainsi trouver le juste équilibre entre le comportement que l'on souhaite avoir, et la stabilité du filtre.

Nous présentons en figure II.14 un suivi réalisé avec des valeurs de $\sigma_{(.)}$ choisies suivant nos heuristiques. Nous constatons que le système parvient à suivre le mouvement du sujet. Ces évaluations qualitatives préliminaires confirment que ces paramètres libres doivent être fixés avec la plus grande attention.

5 Conclusion

Ce chapitre a présenté notre système de suivi par vision multi-oculaire ou stéréoscopique. Nous avons tout d'abord détaillé notre modèle de l'homme. La modélisation cinématique et volumique adoptée reste assez classique au regard de la littérature dans la communauté « Vision par Ordinateur ».

Nous avons détaillé dans un deuxième temps le modèle d'observation qui se veut multi-attributs. Nous proposons ainsi de fusionner plusieurs indices visuels relatifs aux attributs de formes, couleur et mouvement du sujet observé avec pour perspective de robustifier le système de suivi. Ces indices visuels répondent aux deux contextes appli-

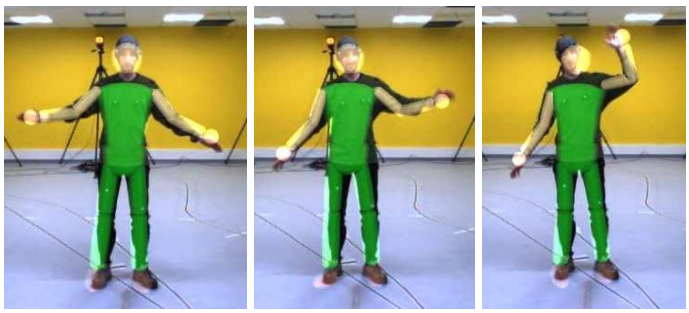


FIG. II.14 – Suivi sur une séquence simple avec $\sigma_{sil} = 20$, $\sigma_{sil2} = 0.1$ et $\sigma_{dist\_peau} = 5$. Ces valeurs ont été choisies suivant nos heuristiques. La configuration estimée est correcte.

catifs mentionnés. Ces attributs d'apparence sont mixés avec des attributs géométriques 3D dans une stratégie hybride mi-apparence, mi-reconstruction 3D (éparse) peu répandue dans la littérature [8].

Notre système de suivi est structuré en modules pour faciliter son intégration sur toute plate-forme robotique, ainsi que l'évaluation séparée de ces derniers. La notion de fonction d'importance basée mesure est peu répandue dans la littérature et est exclusivement limitée (à notre connaissance) au suivi 2D [61, 128]. Nous proposons ici son extension au suivi 3D. Ces propos seront illustrés dans le chapitre IV.

Enfin, des heuristiques guidant le choix des paramètres libres des fonctions de vraisemblance sont proposées. Ceci reste un problème très ouvert, et encore peu abordé dans la littérature, de par la complexité qu'il présente.

Après avoir présenté le système de suivi visuel dans sa globalité, nous nous focalisons dans le prochain chapitre sur les stratégies de filtrage particulaire implémentées au sein du module de filtrage.

Chapitre III

Filtrage stochastique

Alors que le chapitre précédent s'intéressait à la description du modèle de l'homme et à l'analyse des images, ce chapitre se propose de traiter plus formellement certains aspects mathématiques relatifs aux stratégies de suivi mises en place. Nous commençons par rappeler l'approche bayésienne du suivi visuel par filtrage particulaire. Par la suite, nous présentons les algorithmes de base ainsi que plusieurs de leurs évolutions, sur lesquelles nous avons choisi de focaliser notre attention. Nous proposons un certain nombre de métriques définies à partir des considérations de la littérature et terminons enfin par une première évaluation des stratégies envisagées vis-à-vis de ces critères dans le cadre de séquences vidéos de synthèse.

1 Principe et fonctionnement

1.1 Approche bayésienne

A l'instar de nombreux autres problèmes de suivi visuel, la capture de mouvement humain peut être formalisée dans un cadre stochastique bayésien. Les données fournies par les caméras sont alors fusionnées avec une information *a priori* sur la dynamique des membres du corps, permettant ainsi une analyse spatio-temporelle du mouvement.

Nous nous plaçons dans le cadre d'un système stochastique markovien caractérisé à chaque instant $k \in \mathbb{N}$ par un vecteur d'état $\boldsymbol{x}_k$. L'évolution de ce système se traduit par un ensemble de mesures $\boldsymbol{z}_k$. On suppose connue la distribution du vecteur d'état à l'instant initial $p_0(\boldsymbol{x}_0)$. Toute la connaissance *a priori* sur l'évolution temporelle du vecteur d'état $\boldsymbol{x}_k$ est modélisée par la dynamique du système $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})$. Le lien entre le vecteur de mesure $\boldsymbol{z}_k$ et le vecteur d'état $\boldsymbol{x}_k$ est régi par la densité de probabilité $p(\boldsymbol{z}_k|\boldsymbol{x}_k)$. Le système est ainsi entièrement caractérisé par la donnée de :

$$\left\{ \begin{array}{c} p_0(\boldsymbol{x}_0), \text{ la distribution du vecteur d'état à l'instant initial,} \\ p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}), \text{ la dynamique } \textit{a priori} \text{ du système,} \\ p(\boldsymbol{z}_k|\boldsymbol{x}_k), \text{ le lien état-mesure.} \end{array} \right. \tag{III.1}$$

L'objectif du filtrage (particulaire ou autre) est alors d'estimer la loi *a posteriori*

de $\boldsymbol{x}_k$ conditionnellement aux mesures $\boldsymbol{z}_{1:k} = \boldsymbol{z_1}, \ldots, \boldsymbol{z}_k$, *i.e.* la densité de probabilité $p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})$.

Dans notre contexte, le sujet dont nous devons capturer les mouvements constitue le système étudié. Les données $\boldsymbol{z}_k$ symbolisent les images transmises à chaque instant par les caméras. Le chapitre II a détaillé les différents modèles que nous avons choisi d'adopter. Nous rappelons que nous focalisons le suivi sur les mouvements du sujet uniquement, *i.e.* la configuration spatiale et articulaire définie par les paramètres de Denavit-Hartenberg modifiés, sans prendre en compte la forme des membres eux-mêmes, définie par ailleurs. L'état interne $\boldsymbol{x}_k$ intègre donc l'ensemble de ces paramètres. La dynamique du mouvement $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})$ — dépendant essentiellement de la volonté du sujet dans ce cas — est modélisée par une marche aléatoire. La fonction de vraisemblance $p(\boldsymbol{z}_k|\boldsymbol{x}_k)$ repose sur les indices visuels précédemment décrits dans le chapitre II.

La solution du problème de filtrage abordé ici peut reposer sur plusieurs méthodes. Le filtre Kalman [105] fut la première utilisée. Toutefois, bien qu'il décrive la solution exacte par des calculs analytiques, il impose des hypothèses contraignantes : la dynamique et le lien état-mesure doivent pouvoir se modéliser comme des fonctions linéaires auxquelles sont ajoutés des bruits gaussiens. La condition d'initialisation doit être gaussienne. Le principe consiste alors à propager à chaque instant les deux premiers moments statistiques de la loi *a posteriori*, elle-même gaussienne. Plusieurs alternatives ont été proposées afin d'étendre le domaine d'application aux cas non linéaires et/ou non gaussien, dont le filtre de Kalman étendu (ou EKF pour « Extended Kalman Filter ») [150] ainsi que le filtre de Kalman sans parfum [86, 161] (ou UKF pour « Unscented Kalman Filter »). Cependant, ces méthodes ne gèrent pas la multi-modalité potentielle de la loi *a posteriori*, puisque seuls les deux premiers moments statistiques sont propagés.

La deuxième classe de méthodes, qui est celle qui nous intéresse ici, est basée sur une approche numérique permettant d'approximer toute distribution par une distribution ponctuelle dont le support peut évoluer stochastiquement. Ces approches, dites « particulaires », sont détaillées ci-après.

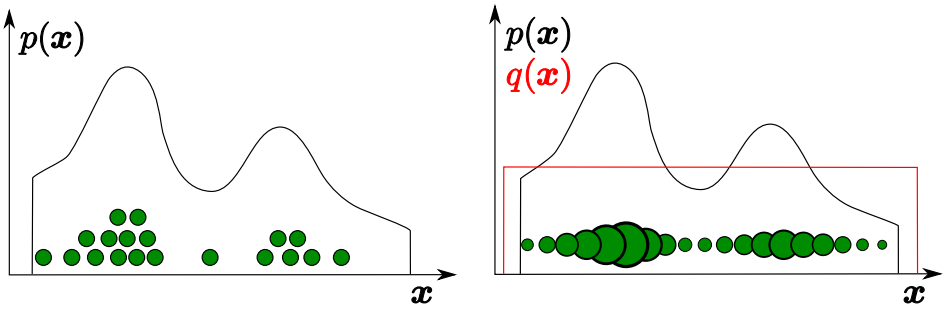
1.2 Approximation particulaire

Modélisation

Le principe consiste à approximer une distribution quelconque $p(\boldsymbol{x})$ par une distribution ponctuelle

$$\widehat{p}(\boldsymbol{x}) = \sum_{i=1}^{N} w^{(i)} \delta(\boldsymbol{x} - \boldsymbol{x}^{(i)}), \qquad \sum_{i=1}^{N} w^{(i)} = 1. \tag{III.2}$$

La densité est alors représentée par un ensemble de N échantillons — ou particules — $\boldsymbol{x}^{(i)}, i = 1, \ldots, N$ auxquels sont associées autant de valeurs — ou poids — $w^{(i)}$,



(a) (b)

FIG. III.1 – Approximation particulaire d'une distribution $p(\boldsymbol{x})$. Les points représentent les particules, et leur taille est proportionnelle à leur poids. En (a), les particules sont des échantillons i.i.d. selon $p(\boldsymbol{x})$. En (b), les particules sont placées selon une fonction d'importance $q(\boldsymbol{x})$, une densité uniforme dans ce cas, et les poids sont corrigés en conséquence afin que le nuage pondéré représente $p(\boldsymbol{x})$.

$i = 1, \ldots, N$. Dès lors, une variable aléatoire distribuée selon $p(\boldsymbol{x})$ peut être simulée en choisissant de tirer la particule $\boldsymbol{x}^{(i)}$ avec la probabilité $w^{(i)}$. L'approximation $\widehat{p}(\boldsymbol{x})$ de $p(\boldsymbol{x})$ peut être instanciée de plusieurs manières en fonction des valeurs choisies pour $\boldsymbol{x}^{(i)}$ et $w^{(i)}$. L'approche la plus immédiate consiste à tirer N échantillons $\boldsymbol{x}^{(i)}$ indépendants et identiquement distribués (i.i.d.) suivant la loi à approximer $p(\boldsymbol{x})$ et à leur attribuer des poids tous égaux à $\frac{1}{N}$ (cf. figure III.1 (a)) :

$$\boldsymbol{x}^{(i)} \sim p(\boldsymbol{x}), \qquad w^{(i)} = \frac{1}{N}. \tag{III.3}$$

Cependant, il n'est pas toujours possible ou souhaitable d'échantillonner $p(\boldsymbol{x})$. La méthode précédente n'est alors plus applicable, et on a recours à l'échantillonnage préférentiel — ou « importance sampling » —, schématisé en figure III.1 (b). La répartition des particules $\boldsymbol{x}^{(i)}$ se fait alors par le biais d'une fonction $q(\boldsymbol{x})$, dite « d'importance », définie sur un support $\mathcal{D}'$ recouvrant le support $\mathcal{D}$ de $p(\boldsymbol{x})$. Les poids sont alors affectés d'une valeur proportionnelle à $\frac{p(\boldsymbol{x}^{(i)})}{q(\boldsymbol{x}^{(i)})}$ puis normalisés :

$$\boldsymbol{x}^{(i)} \sim q(\boldsymbol{x}), \qquad w^{(i)} \propto \frac{p(\boldsymbol{x}^{(i)})}{q(\boldsymbol{x}^{(i)})}, \qquad \sum_{i=1}^{N} w^{(i)} = 1. \tag{III.4}$$

Cette technique peut être vue comme une généralisation de l'approche plus intuitive qui consiste à choisir des valeurs de $\boldsymbol{x}^{(i)}$ uniformément réparties sur $\mathcal{D}$ et à affecter aux poids $w^{(i)}$ la valeur de $p(\boldsymbol{x})$ correspondante, l'ensemble des poids étant toujours normalisé.

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

1: **SI** $k = 0$**, ALORS** Échantillonner $\boldsymbol{x}_0^{(1)}, \ldots, \boldsymbol{x}_0^{(N)}$ i.i.d. selon $p_0(\boldsymbol{x}_0)$, et poser $w_0^{(i)} = \frac{1}{N}$. **FIN SI**
2: **SI** $k \geq 1$ **ALORS** $\{$—$\{(\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)})\}_{i=1}^N$ représente $p(\boldsymbol{x}_{k-1}|\boldsymbol{z}_{1:k-1})$—$\}$
3: **POUR** $i = 1, \ldots, N$, **FAIRE**
4: Échantillonner indépendamment $\boldsymbol{x}_k^{(i)} \sim q(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}^{(i)}, \boldsymbol{z}_k)$
5: Mettre à jour les poids *via* $w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\boldsymbol{z}_k|\boldsymbol{x}_k^{(i)})p(\boldsymbol{x}_k^{(i)}|\boldsymbol{x}_{k-1}^{(i)})}{q(\boldsymbol{x}_k^{(i)}|\boldsymbol{x}_{k-1}^{(i)}, \boldsymbol{z}_k)}$
6: **FIN POUR**
7: Normaliser les poids de sorte que $\sum_i w_k^{(i)} = 1$
8: Calculer l'estimé du MMSE $\mathrm{E}_{p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})}[\boldsymbol{x}_k] = \sum_{i=1}^N w_k^{(i)} \boldsymbol{x}_k^{(i)}$
9: Rééchantillonner (cf. table III.2) $\{(\boldsymbol{x}_k^{(i)}, w_k^{(i)})\}_{i=1}^N$ en $\{(\boldsymbol{x}_k^{(s^i)}, \frac{1}{N})\}_{i=1}^N$ et renommer en $\{(\boldsymbol{x}_k^{(i)}, w_k^{(i)})\}_{i=1}^N$
10: **FIN SI**

TAB. III.1 – Algorithme générique de filtrage particulaire (SIR).

Mise en œuvre et intérêt

L'approximation particulaire permet ainsi la représentation d'une densité $p(\boldsymbol{x})$ par l'ensemble de couples $\{(\boldsymbol{x}^{(i)}, w^{(i)})\}_{i=1}^N$. Du fait de III.2, toute intégrale

$$\mu = \int f(\boldsymbol{x}) p(\boldsymbol{x}) dx \tag{III.5}$$

exprimant l'espérance de l'image par une fonction f d'une variable aléatoire distribuée selon $p(\boldsymbol{x})$ peut être approchée par

$$\widehat{\mu} = \sum_{i=1}^{N} w^{(i)} f(\boldsymbol{x}^{(i)}). \tag{III.6}$$

En particulier, la moyenne $\int \boldsymbol{x} p(\boldsymbol{x}) dx$ de la distribution $p(\boldsymbol{x})$ est approximée par $\sum_{i=1}^N w^{(i)} \boldsymbol{x}^{(i)}$. Cette représentation des densités de probabilité est exploitée dans le cadre des filtres particulaires, présentés ci-après.

1.3 Algorithme générique

La première version d'un filtre bayésien particulaire est proposée par Gordon *et al.* dans [64] sous le nom de « Bootstrap Filter ». La méthode est redécouverte indépendamment dans un contexte de Vision par Ordinateur par Isard et Blake dans [80, 81], qui introduisent le vocable CONDENSATION pour « CONditional DENSity propagATION ». L'algorithme générique SIR — pour « Sampling Importance Resampling » — qui englobe ces travaux pionniers ainsi que d'autres variantes est présenté en table III.1. Le lecteur est invité à consulter le document remarquable de Chen [24] et les tutoriaux d'Arulampalam [5] et Doucet [45, 46] pour une présentation plus détaillée.

L'algorithme SIR a pour but d'estimer récursivement la loi *a posteriori* par l'approximation particulaire

$$p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta(\boldsymbol{x}_k - \boldsymbol{x}_k^{(i)}), \qquad \sum_{i=1}^{N} w_k^{(i)} = 1. \tag{III.7}$$

Cette approximation est généralement initialisée selon $p_0(\boldsymbol{x}_0)$. Par la suite, le principe est de propager la description particulaire au cours du temps en tenant compte de la dynamique $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})$ et de la vraisemblance $p(\boldsymbol{z}_k|\boldsymbol{x}_k)$. La définition de la fonction d'importance $q(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{z}_k)$ influence quant à elle la manière dont sont positionnées les particules dans l'espace d'état. Les poids sont ensuite corrigés en conséquence. À titre d'exemple, la CONDENSATION utilise la dynamique du système comme fonction d'importance, *i.e.* $q(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{z}_k) = p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})$. L'étape de mise à jour des poids se voit alors simplifiée en $w_k^{(i)} \propto w_{k-1}^{(i)} p(\boldsymbol{z}_k|\boldsymbol{x}_k^{(i)})$.

Par la suite, on peut décider de calculer l'estimé du minimum d'erreur quadratique moyenne (ou MMSE pour « Minimum Mean Square Error »), qui est en fait la moyenne *a posteriori* du vecteur d'état :

$$\mathrm{E}_{p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})}[\boldsymbol{x}_k] = \sum_{i=1}^{N} w_k^{(i)} \boldsymbol{x}_k^{(i)}. \tag{III.8}$$

La dernière partie de l'algorithme est constituée d'une étape de rééchantillonnage que nous abordons dans la section suivante.

1.4 Étape de rééchantillonnage

Nombre de particules efficaces

La dernière étape de l'algorithme présenté ci-dessus consiste à rééchantillonner le nuage de particules afin d'éviter le phénomène de « dégénérescence ». En effet, sans cette étape, après quelques itérations du filtre, toutes les particules affichent un poids nul à l'exception d'une seule. Un indicateur permettant de vérifier le bon comportement du filtre vis-à-vis de ce problème est le nombre de particules efficaces, proposé dans [96]. Son calcul rigoureux fait intervenir la variance d'estimateurs reposant sur un échantillonnage préférentiel et sur un échantillonnage parfait de la loi *a posteriori*. Il ne peut pas être calculé de manière directe, néanmoins, une estimation couramment utilisée est donnée par :

$$\widehat{N_{eff}} = \frac{1}{\sum_{i=1}^{N} (w^{(i)})^2}. \tag{III.9}$$

Lorsque toutes les particules ont le même poids, $\widehat{N_{eff}}$ vaut N. En revanche, si une seule concentre tout le poids de la distribution, $\widehat{N_{eff}}$ prend la valeur 1.

Rééchantillonnage systématique

Afin de maintenir des poids équilibrés au sein du nuage de particules, l'objectif du rééchantillonnage est de transformer le nuage pondéré $\{(\boldsymbol{x}_k^{(i)}, w_k^{(i)})\}_{i=1}^N$ en un nuage équivalent, (au sens où il représente la même distribution) $\{(\boldsymbol{x}_k^{(s^{(i)})}, w_k^{(s^{(i)})} = \frac{1}{N})\}_{i=1}^N$. Le principe consiste à échantillonner $\boldsymbol{x}_k^{(s^{(i)})}$ (avec remise) dans $\{\boldsymbol{x}_k^{(i)}\}_{i=1}^N$ selon $P(\boldsymbol{x}_k^{(s^{(i)})} = \boldsymbol{x}_k^{(j)}) = w_k^{(j)}$.

Idéalement, chaque particule originale $\boldsymbol{x}_k^{(i)}$ est rééchantillonnée un nombre de fois proportionnel à son poids $w_k^{(i)}$. Cependant, bien que réalisable en moyenne, ce comportement n'est pas garanti sur une réalisation du procédé. En effet, les particules sont nécessairement rééchantillonnées un nombre entier de fois, et il est inévitable, que ce nombre ne soit pas exactement proportionnel à leur poids (réel par définition). Toutes les techniques de rééchantillonnage introduisent ainsi une variance supplémentaire dans la représentation de la densité approchée, dite « variance de Monte-Carlo ».

Bien que de nombreuses techniques de rééchantillonnage existent [24], la technique généralement exploitée est celle du rééchantillonnage dit « systématique », introduite par Kitagawa dans [93] et présentée table III.2. Elle a l'avantage d'introduire la variance de Monte-Carlo la plus faible, et présente une complexité $\mathcal{O}(N)$, particulièrement adaptée aux contextes contraints en temps de calcul. Notons enfin que l'estimé fourni par le filtre doit être calculé à partir de la représentation particulaire avant rééchantillonnage afin de ne pas être affecté par la variance de Monte-Carlo introduite. De même, ce rééchantillonnage n'est pas toujours souhaitable à chaque itération du filtre. C'est pourquoi il n'est généralement déclenché que lorsque le nombre de particules efficaces se situe en deçà d'un certain seuil, *e.g.* $2N/3$ [70].

Nous avons introduit le principe des filtres particulaires. La prochaine section détaille plusieurs améliorations du SIR proposées dans la littérature et que nous exploitons dans nos évaluations. Tous les algorithmes ne sont pas détaillés ici par souci de lisibilité. Certains sont reportés en annexe B.

2 Évolutions

2.1 Fonction d'importance

La stratégie de CONDENSATION choisit la dynamique du système comme fonction d'importance, ce qui a pour avantage de simplifier l'étape de mise à jour des poids. Toutefois, dans le but d'explorer les zones de l'espace de recherche présentant des pics de vraisemblance, le positionnement judicieux des particules constitue un des piliers de l'efficacité du suivi. Isard et Blake [82] ont ainsi introduit la prise en compte des mesures dans la fonction d'importance, assurant une exploration plus pertinente de l'espace. La fonction d'importance prend alors généralement la forme suivante :

$\{(\boldsymbol{x}_k^{(s^{(i)})}, w_k^{(s^{(i)})})\}_{i=1}^N = RESAMPLE(\{(\boldsymbol{x}_k^{(i)}, w_k^{(i)})\}_{i=1}^N)$

1: Initialiser la somme cumulée des poids (SCP) : $c_1 = w_k^{(1)}$
2: **POUR** $i = 2, \ldots, N$, **FAIRE**
3: Construire la SCP : $c_i = c_{i-1} + w_k^{(i)}$
4: **FIN POUR**
5: Poser $i = 1$
6: Échantillonner un point de départ : $u_1 \sim \mathcal{U}_{[0,N^{-1}]}$
7: **POUR** $j = 1, \ldots, N$, **FAIRE**
8: Se déplacer le long de la SCP : $u_j = u_1 + (j-1)N^{-1}$
9: **TANT QUE** $u_j > c_i$, **FAIRE**
10: $i = i + 1$
11: **FIN TANT QUE**
12: Recopier la particule : $\boldsymbol{x}_k^{(s^{(j)})} = \boldsymbol{x}_k^{(i)}$
13: Affecter le poids : $w_k^{(s^{(j)})} = N^{-1}$
14: **FIN POUR**

TAB. III.2 – Algorithme de « rééchantillonnage systématique ».

$$q(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{z}_k) = (1-\alpha-\beta)p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) + \alpha\pi(\boldsymbol{x}_k|\boldsymbol{z}_k) + \beta\pi_0(\boldsymbol{x}_k). \qquad \text{(III.10)}$$

où $\alpha, \beta \in [0; 1]$. En d'autres termes, une proportion α des particules est échantillonnée en fonction de l'observation selon $\pi(\boldsymbol{x}_k|\boldsymbol{z}_k)$, lorsque celle-ci est exploitable. Dès lors, lorsque le filtre perd la cible, la prise en compte de l'observation dans le positionnement des particules peut permettre une ré-initialisation automatique du filtre. Un pourcentage β des particules est échantillonné suivant une connaissance *a priori* notée $\pi_0(\boldsymbol{x}_k)$ [128]. Un choix courant consiste à poser $\pi_0(\boldsymbol{x}_k) = p_0(\boldsymbol{x}_0)$, *i.e.* la distribution *a priori* du vecteur d'état initial. Une autre approche pour les espaces d'état finis et/ou bornés est de choisir une densité uniforme $\pi_0(\boldsymbol{x}_k) = \mathcal{U}(\boldsymbol{x}_k)$ permettant de couvrir l'intégralité du domaine. Notons que cette technique s'emploie de préférence pour les espaces de petite dimension. Le reste des particules est classiquement propagé selon la dynamique du système. L'ajustement des coefficients α et β est laissé au libre choix de l'utilisateur, et il est envisageable de ne pas exploiter de connaissance *a priori* ($\beta = 0$) et/ou la mesure ($\alpha = 0$).

Le choix d'une fonction d'importance « mixte » telle que celles décrites précédemment soulève cependant un problème au niveau de la mise à jour des poids : l'association d'une particule $\boldsymbol{x}_{k-1}^{(i)}$ aux particules $\boldsymbol{x}_k^{(i)}$ échantillonnées suivant $\pi(\boldsymbol{x}_k|\boldsymbol{z}_k)$ ou $\pi_0(\boldsymbol{x}_k)$ se fait de manière arbitraire. Cela peut conduire à un écrasement des poids des particules qui ne sont pas cohérentes avec leurs particules prédécesseurs du point de vue de la dynamique. Cette difficulté est abordée par Torma et Szepesvári dans [158].

L'exploration de l'espace d'état peut ainsi être améliorée grâce à un choix judicieux de la fonction d'importance. Dans notre cas, nous adoptons la fonction d'importance

décrite au chapitre II, section 4.2. Il existe toutefois d'autres leviers qui affectent ce critère.

2.2 Partitionnement de l'espace

Afin d'optimiser l'exploration de l'espace d'état lorsque celui-ci présente une structure particulière, MacCormick et Blake proposent dans [103] un filtre particulaire « partitionné ». Ils supposent que le vecteur d'état peut se décomposer en M sous-vecteurs $\boldsymbol{x}_k^m$, $m \in 1, \ldots, M$, de telle sorte que la dynamique et la vraisemblance se factorisent respectivement en $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) \propto \Pi_{m=1}^M p(\boldsymbol{x}_k^m|\boldsymbol{x}_{k-1}^m)$ et $p(\boldsymbol{z}_k|\boldsymbol{x}_k) \propto \Pi_{m=1}^M l_m(\boldsymbol{z}_k|\boldsymbol{x}_k^1, \ldots, \boldsymbol{x}_k^m)$, où les vraisemblances intermédiaires l_m exploitent un nombre croissant de sous-vecteurs $\boldsymbol{x}_k^m$ au fur et à mesure que m croît vers M. L'algorithme complet, résumé en table III.3, est présenté de manière plus complète dans la thèse de MacCormick [101]. Sa complexité est linéaire en le nombre de partitions du vecteur d'état, alors que le SIR est de complexité exponentielle en le nombre de DDL. En outre, le nombre de particules utilisées peut être différent pour chaque partition traitée, limitant ainsi le coût calculatoire. Le suivi du corps humain présente des dynamiques découplées pour chaque articulation paramétrée par le vecteur d'état. L'algorithme de filtrage partitionné est donc applicable dans notre contexte (la première application au suivi monocible, proposée par MacCormick et Isard [103], se focalisait sur le suivi 2D d'une main). Notons cependant que l'application d'une telle stratégie au contexte de HMC impose de pouvoir localiser plusieurs parties du corps humain de manière indépendante. Certains auteurs [41] avancent que les informations purement géométriques doivent être complétées par de la couleur ou un système d'étiquettes afin de permettre un suivi correct.

Une alternative de cette méthode exploitant une fonction d'importance $q(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{z}_k)$ en lieu et place de la simple dynamique du système pour le placement des particules est proposée par Pérez dans [128].

2.3 Approche par affinement

L'« Annealed Particle Filter » (APF), ou filtre particulaire à recuit simulé [38], est basé sur une exploration de l'espace d'état par affinements successifs. Le principe, présenté en table III.4, est de décomposer la boucle principale de la CONDENSATION en L sous-étapes. À chaque étape $l \in 1, \ldots, L$, les particules $\boldsymbol{x}_{k,l}^{(i)}$ sont propagées selon une « fonction de dynamique élémentaire » $p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1})$ puis confrontées à la mesure *via* une « fonction de vraisemblance élémentaire » $p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l})$. Il s'agit alors de définir les fonctions $p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1})$ et $p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l})$ de manière judicieuse afin d'améliorer le processus de suivi. Deutscher *et al.* [38] préconisent l'utilisation de

$$\left\{ \begin{array}{rcl} p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1}) & = & [p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})^{\alpha_l}]_{\boldsymbol{x}_k=\boldsymbol{x}_{k,l},\boldsymbol{x}_{k-1}=\boldsymbol{x}_{k,l-1}} \\ p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}) & = & [p(\boldsymbol{z}_k|\boldsymbol{x}_k)^{\beta_l}]_{\boldsymbol{x}_k=\boldsymbol{x}_{k,l}} \end{array} \right. \quad \text{(III.11)}$$

où α_l et β_l sont des paramètres croissants lorsque $l \rightarrow L$ avec $\alpha_l \geq 1$ et $\beta_l \leq 1$, $l \in 1, \ldots, L$. Ce choix permet une exploration de plus en plus ciblée de l'espace de

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

SI $k = 0$**, ALORS** Échantillonner $\boldsymbol{x}_0^{(1)}, \ldots, \boldsymbol{x}_0^{(N)}$ i.i.d. selon $p_0(\boldsymbol{x}_0)$, et poser $w_0^{(i)} = \frac{1}{N}$ **FIN SI**
SI $k \geq 1$ **ALORS** $\{$—$\{(\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)})\}_{i=1}^N$ représente $p(\boldsymbol{x}_{k-1}|\boldsymbol{z}_{1:k-1})$—$\}$
  Poser $\tau_0^{(i)} = w_{k-1}^{(i)}$ pour $i = 1, \ldots, N$
  POUR $m = 1, \ldots, M$, **FAIRE**
    POUR $i = 1, \ldots, N$, **FAIRE**
      Échantillonner indépendamment $\boldsymbol{x}_k^{m,(i)} \sim p_m(\boldsymbol{x}_k^m|\boldsymbol{x}_{k-1}^{m,(i)})$
      Associer le poids $\tau_m^{(i)} \propto \tau_{m-1}^{(i)} l_m(\boldsymbol{z}_k|\boldsymbol{x}_k^{1,(i)}, \ldots, \boldsymbol{x}_k^{m,(i)})$
    FIN POUR
    Normaliser les poids de sorte que $\sum_{i=1}^N \tau_m^{(i)} = 1$
    SI $m < M$ **ALORS**
      Rééchantillonner $\{((\boldsymbol{x}_k^{1,(i)}, \ldots, \boldsymbol{x}_k^{M,(i)}), \tau_m^{(i)})\}_{i=1}^N$
    FIN SI
  FIN POUR
  Renommer $\{((\boldsymbol{x}_k^{1,(i)}, \ldots, \boldsymbol{x}_k^{M,(i)}), \tau_M^{(i)})\}_{i=1}^N$ en $\{(\boldsymbol{x}_k^{(i)}, w_k^{(i)})\}_{i=1}^N$
  Calculer l'estimé du MMSE $\mathrm{E}_{p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})}[\boldsymbol{x}_k] = \sum_{i=1}^N w_k^{(i)} \boldsymbol{x}_k^{(i)}$
FIN SI

TAB. III.3 – Algorithme du filtrage partitionné.

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

SI $k = 0$**, ALORS** Échantillonner $\boldsymbol{x}_0^{(1)}, \ldots, \boldsymbol{x}_0^{(N)}$ i.i.d. selon $p_0(\boldsymbol{x}_0)$, et poser $w_0^{(i)} = \frac{1}{N}$ **FIN SI**
SI $k \geq 1$ **ALORS** $\{$—$\{(\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)})\}_{i=1}^N$ représente $p(\boldsymbol{x}_{k-1}|\boldsymbol{z}_{1:k-1})$—$\}$
  Poser $\{(\boldsymbol{x}_{k,0}^{(i)}, w_{k,0}^{(i)})\}_{i=1}^N = \{(\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)})\}_{i=1}^N$ et choisir $1 \leq \alpha_1 < \cdots < \alpha_L$ et $\beta_1 < \cdots < \beta_L \leq 1$
  POUR $l = 1, \ldots, L$, **FAIRE**
    POUR $i = 1, \ldots, N$, **FAIRE**
      Échantillonner indépendamment $\boldsymbol{x}_{k,l}^{(i)} \sim p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1}^{(i)})$
      Associer le poids $w_{k,l}^{(i)} \propto w_{k,l-1}^{(i)} p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)})$
    FIN POUR
    Normaliser les poids de sorte que $\sum_{i=1}^N w_{k,l}^{(i)} = 1$
    SI $l < L$ **ALORS**
      Rééchantillonner $\{(\boldsymbol{x}_{k,l}^{(i)}, w_{k,l}^{(i)})\}_{i=1}^N$
    FIN SI
  FIN POUR
  Renommer $\{(\boldsymbol{x}_{k,L}^{(i)}, w_{k,L}^{(i)})\}_{i=1}^N$ en $\{(\boldsymbol{x}_k^{(i)}, w_k^{(i)})\}_{i=1}^N$
  Calculer l'estimé du MMSE $\mathrm{E}_{p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})}[\boldsymbol{x}_k] = \sum_{i=1}^N w_k^{(i)} \boldsymbol{x}_k^{(i)}$
FIN SI

TAB. III.4 – Algorithme de l'« Annealed Particle Filter » ou filtrage particulaire à recuit simulé.

recherche. En effet, la fonction de dynamique élémentaire est au départ proche de la dynamique du système de telle sorte qu'elle dissémine largement les particules. Au fur et à mesure, elle se rétrécit de manière à les placer de plus en plus finement. En parallèle et de manière complémentaire, la fonction de vraisemblance élémentaire est initialisée avec un profil très lissé (β petit) et se rapproche peu à peu de la vraisemblance $p(\boldsymbol{z}_k|\boldsymbol{x}_k)$ ($\beta \rightarrow 1$) de manière à faire progresser les particules vers les pics de la distribution *a posteriori*, proposant donc une meilleure localisation de ses modes.

Cette méthode s'avère en pratique plus efficace que la stratégie conventionnelle, même si, comme le précisent ses auteurs, elle n'est pas définie dans un cadre stochastique rigoureux [38]. Notons également qu'il n'est pas toujours trivial d'échantillonner $p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1})$. Toutefois, dans notre cas, la dynamique est de type marche aléatoire, *i.e.* $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) = \mathcal{N}(\boldsymbol{x}_k; \boldsymbol{x}_{k-1}, \Delta_k)$ où Δ_k est diagonale. Échantillonner $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})^{\alpha_l}$ revient alors à échantillonner $\mathcal{N}(\boldsymbol{x}_k; \boldsymbol{x}_{k-1}, \frac{1}{\alpha_l}\Delta_k)$. Le choix définitif des valeurs de α_l et β_l reste cependant difficile. Dans la veine de [38], nous choisissons $\alpha_l = 2^l, l \in 1, \ldots, L-1$ et $\alpha_L = 2^{L-1}$. Notons que dans la version légèrement modifiée — et mathématiquement correcte — de l'APF que nous proposons en annexe C.1, la dynamique effective résultante présente alors une covariance Δ'_k identique à la covariance Δ_k de la dynamique cible :

$$\Delta'_k = \sum_{i=1}^{L-1} \frac{1}{2^l}\Delta_k + \frac{1}{2^{L-1}} = \Delta_k. \tag{III.12}$$

Nous appliquons cette heuristique pour la version classique de l'APF avec laquelle nous effectuons nos tests dans le but de nous comparer à la littérature. Les fonctions de vraisemblance n'étant pas aussi facilement paramétrables, le choix des β_l peut se faire par la pratique, ou selon les heuristiques proposées dans [38].

L'algorithme IAPF, pour « Importance Annealed Particle Filter », proposé dans [52, 55] et détaillé en table C.2, combine l'approche précédente avec l'utilisation d'une fonction d'importance dans l'optique de tirer parti de la mesure au plus tôt. Une autre évolution de l'APF est proposée par Deutscher *et al.* dans [39, 41]. En remplacement de la fonction de dynamique élémentaire, elle exploite la covariance estimée du nuage de particules à l'étape $l-1$ pour placer les particules à l'étape l, privilégiant ainsi les zones fortement vraisemblables *a posteriori*. En complément, un opérateur de « crossing-over » dérivé des algorithmes génétiques est introduit. Il permet un partitionnement automatique de l'espace d'état en combinant aléatoirement les particules présentant des poids importants. Comme mentionné dans le chapitre I, notons que l'exploration de l'espace de recherche selon la covariance de la loi *a posteriori* est également exploitée par Sminchicescu et Triggs dans [146]. L'approche est alors complétée par une optimisation locale afin de repérer plus précisément les *maxima* de $p(\boldsymbol{z}_k|\boldsymbol{x}_k)$.

2.4 Échantillonnage de Quasi Monte Carlo

Dans les stratégies de filtrage particulaire, la nature stochastique du positionnement des particules permet d'explorer les zones de l'espace d'état où la distribution *a*

posteriori est dense. Toutefois, l'échantillonnage aléatoire (ou un algorithme approché pseudo-aléatoire) est sujet au phénomène dit « d'amas et de trous » (« gaps and clusters »), notamment dans les espaces de grande dimension. Une variation excessive de l'estimé calculé peut s'en suivre, rendant le filtre peu fiable et aboutissant parfois à des échecs de suivi.

Une alternative est de mettre en œuvre des méthodes de Quasi Monte Carlo (QMC). Le principe consiste à échantillonner les particules à l'aide d'une séquence déterministe à faible discrépance, qui garantit une certaine homogénéité dans le recouvrement de l'espace. Il a été montré [48] que pour un espace de dimension d, l'erreur d'approximation d'une intégrale par des méthodes de QMC converge à la vitesse $\mathcal{O}(N^{-1}log^{d}N)$ en le nombre de particules N, qui est plus rapide que $\mathcal{O}(N^{-\frac{1}{2}})$ pour les méthodes de Monte Carlo (MC) classiques. Cependant, l'analyse de la précision des approximations QMC déterministes est difficile et exclut toute approche statistique. Des méthodes QMC randomisées ont donc été proposées, afin de générer des séquences quasi-aléatoires à faible discrépance selon une distribution donnée. Ceci a permis la mise en place d'estimateurs non biaisés et à variance réduite, dont l'erreur d'approximation peut converger à une vitesse en $\mathcal{O}(N^{-\frac{3}{2}}log^{\frac{d-1}{2}})$.

Dans le cadre du suivi visuel, très peu de comparaisons entre filtres particulaires MC et QMC ont été proposées [123, 130]. Dans ces références, plusieurs réalisations des processus de filtrage sur des séquences de synthèse dans un espace de faible dimension pour un même nombre N de particules ont montré que pour des méthodes QMC déterministes ou randomisées : (1) la racine de l'erreur quadratique moyenne (ou RMSE pour « Root Mean Square Error ») de l'estimé par rapport à la vérité de terrain est toujours plus faible, (2) l'écart-type de l'estimé sur la durée de la séquence est toujours plus faible, (3) lorsque N augmente, le RMSE diminue d'autant plus vite que N est faible ou que le contexte est bruité, (4) pour une erreur donnée, les méthodes QMC ne nécessitent qu'entre la moitié et le tiers du nombre de particules, ce qui permet un temps d'exécution moindre. Des expériences de suivi visuel sur séquences réelles avec un espace de dimension $d \sim 10$ montrent que le filtre QMC déterministe [130] affiche un comportement satisfaisant sur l'ensemble des séquences et peut retrouver une cible perdue suite à des changements soudains ou des occlusions partielles. Quant au filtre QMC randomisé [123], il est non biaisé par rapport à la moyenne de la densité *a posteriori* réelle et présente des dispersions sur chaque composante du vecteur d'état estimé de 5% à 20% en deçà de celles du filtre MC. Les auteurs affichent un gain de particules allant de 20% à 60%, l'amélioration restant plus prononcée pour un nombre faible de particules, conjointement à une réduction du volume exploré à chaque pas (en dimension 10).

Parmi les problèmes principaux des filtres QMC, nous pouvons citer la difficulté de mise en œuvre de séquences à faible discrépance en dépit des étapes de rééchantillonnage, l'exploitation de la mesure courante dans la définition de telles séquences, et le compromis éventuel entre la réduction de la complexité quadratique des algorithmes et leur rigueur mathématique. L'équivalent quasi-aléatoire de la CONDENSATION, ici dénommé QRS pour « Quasi Random Sampling » [67], est présenté Table III.5. Nous

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

1: **SI** $k = 0$**, ALORS** Échantillonner une séquence QMC randomisée de Sobol $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(N)}$ selon $\mathcal{U}_{[0,1)^d}(\boldsymbol{u})$, la convertir en $\boldsymbol{x}_0^{(1)}, \ldots, \boldsymbol{x}_0^{(N)}$ distribués selon $p_0(\boldsymbol{x}_0)$, et poser $w_0^{(i)} = \frac{1}{N}$. **FIN SI**
2: **SI** $k \geq 1$ **ALORS** $\{$—$\{(\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)})\}_{i=1}^N$ représente $p(\boldsymbol{x}_{k-1}|\boldsymbol{z}_{1:k-1})$—$\}$
3: Sélectionner avec remise $s^{(1)}, \ldots, s^{(N)}$ dans $\{1, \ldots, N\}$ tels que $P(s^{(i)} = j) = w_{k-1}^{(j)}$
4: Poser $C_j = card(\{i|s^{(i)} = j\})$
5: **POUR** $j = 1, \ldots, N$, **FAIRE**
6: Échantillonner une séquence QMC randomisée de Sobol $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(C_j)}$ selon $\mathcal{U}_{[0,1)^d}(\boldsymbol{u})$ et la convertir en $\boldsymbol{x}_k^{(\sum_{l=1}^{j-1} C_l+1)}, \ldots, \boldsymbol{x}_k^{(\sum_{l=1}^{j-1} C_l+C_j)}$ distribués selon $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}^{(j)})$
7: Mettre à jour les poids *via* $w_k^{(i)} \propto p(\boldsymbol{z}_k|\boldsymbol{x}_k^{(i)})$
8: **FIN POUR**
9: Normaliser les poids de sorte que $\sum_i w_k^{(i)} = 1$
10: Calculer le MMSE $\mathrm{E}_{p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})}[\boldsymbol{x}_k] = \sum_{i=1}^N w_k^{(i)} \boldsymbol{x}_k^{(i)}$
11: **FIN SI**

TAB. III.5 – Algorithme de filtrage particulaire QRS.

l'exploiterons dans nos évaluations, conjointement à son homologue partitionné, ici appelé PARTITIONNÉ QRS et exposé en table C.3.

2.5 Stratégie hybride

Dans l'optique de fusionner les atouts de ces différentes stratégies, nous évaluons également un algorithme, présenté en table III.6, qui a recours aux différents procédés précédemment décrits. Le principe est de tout d'abord procéder à une approximation de la loi *a posteriori* par sa moyenne et sa covariance (étapes 3 à 5), reprenant l'approche de [39], afin de placer par la suite les particules dans les zones pertinentes de l'espace d'état par un échantillonnage QMC (étapes 6 et 7).

3 Caractéristiques et comportements

Les techniques de filtrage particulaire font l'objet de nombreux tutoriels [5]. Leur étude théorique a prouvé la convergence pour $N \to \infty$ de l'approximation particulaire vers la loi réelle [30], cependant ces résultats ont été développés dans le cadre de l'approximation particulaire de l'espérance de fonctions bornées, ce qui reste très gênant d'un point de vue pratique. En effet, ces résultats ne concernent pas le calcul de l'estimé du MMSE qui est le plus couramment utilisé. Toutefois, Hu *et al.* ont proposé dans [77] une extension de certaines preuves de convergence pour des fonctions non bornées sous l'hypothèse d'une légère modification de l'algorithme classique. Les résultats en sont améliorés en pratique et les preuves deviennent applicables dans un grand nombre de domaines.

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

SI $k = 0$**, ALORS** Échantillonner une séquence QMC randomisée de Sobol $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(N)}$ selon $\mathcal{U}_{[0,1)^D}(\boldsymbol{u})$ et la convertir en $\boldsymbol{x}_0^{(1)}, \ldots, \boldsymbol{x}_0^{(N)} \sim p_0(\boldsymbol{x}_0)$, et poser $w_0^{(i)} = \frac{1}{N}$. **FIN SI**
SI $k \geq 1$ **ALORS** $\{$—$[\{\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N$ représente $p(\boldsymbol{x}_{k-1}|\boldsymbol{z}_{1:k-1})$—$\}$
  POUR $i = 1, \ldots, N$**, FAIRE**
    Échantillonner $\boldsymbol{x}_k^{(i)} \sim q(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}^{(i)}, \boldsymbol{z}_k)$
    Mettre à jour les poids *via* $w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\boldsymbol{z}_k|\boldsymbol{x}_k^{(i)})p(\boldsymbol{x}_k^{(i)}|\boldsymbol{x}_{k-1}^{(i)})}{q(\boldsymbol{x}_k^{(i)}|\boldsymbol{x}_{k-1}^{(i)}, \boldsymbol{z}_k)}$
  FIN POUR
  Calculer les approximations de la moyenne $\boldsymbol{\mu}_k$ et de la covariance Σ_k de la distribution *a posteriori* $\boldsymbol{\mu}_k = \sum_{i=1}^N w_k^{(i)} \boldsymbol{x}_k^{(i)}, \quad \Sigma_k = \sum_{i=1}^N w_k^{(i)} (\boldsymbol{x}_k^{(i)} - \boldsymbol{\mu}_k)(\boldsymbol{x}_k^{(i)} - \boldsymbol{\mu}_k)^T$
  POUR $i = 1, \ldots, N$**, FAIRE**
    Rééchantillonner les particules par QMC selon $\mathcal{N}(\boldsymbol{\mu}_k, \lambda\Sigma_k)$, recouvrant l'approximation gaussienne $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ de la loi *a posteriori* : $\boldsymbol{x}_k^{(i)} \sim_{QMC} \mathcal{N}(\boldsymbol{\mu}_k, \lambda\Sigma_k)$
    Mettre à jour les poids *via* $w_k^{(i)} = \frac{p(\boldsymbol{z}_k|\boldsymbol{x}_k^{(i)})p(\boldsymbol{x}_k^{(i)}|\boldsymbol{x}_{k-1}^{(i)})}{\mathcal{N}(\boldsymbol{x}_k^{(i)}; \boldsymbol{\mu}_k, \Sigma_k)}$
  FIN POUR
FIN SI

TAB. III.6 – Algorithme de filtrage particulaire HYBRID.

La complexité des filtres particulaires est également un sujet très étudié. Crisan et Doucet avancent dans [30] que la convergence de l'estimé du MMSE est proportionnelle à $\frac{1}{N}$ et qu'elle est indépendante de la dimension de l'espace de recherche. Daum apporte un éclairage différent dans [32, 34]. Il argumente le fait que la complexité n'est indépendante de la dimension de l'espace de recherche que dans certains cas particuliers où le problème est « vaguement gaussien » et dans la mesure où les filtres sont « soigneusement implémentés ». En effet, toute approximation d'une distribution gaussienne nécessite un nombre d'échantillons permettant d'explorer une boule multi-dimensionnelle centrée sur sa moyenne. Aussi surprenant que cela puisse paraître, le volume d'une boule de dimension d décroît au delà de $d = 5$, limitant ainsi drastiquement le nombre d'échantillons nécessaires à la caractérisation d'une distribution gaussienne. En revanche, pour une distribution quelconque, le nombre d'échantillons nécessaires est lié au volume de l'espace de définition, qui lui, croît de manière exponentielle, à la manière du volume d'un hypercube.

L'étude de la complexité des filtres particulaires est donc problématique. Ainsi, afin d'évaluer l'estimé fourni par chaque filtre (MMSE), nous mettons en place les critères détaillés ci-après. Dans notre cas, nous nous attachons à étudier le comportement des stratégies sur plusieurs réalisations, par rapport à une vérité de terrain établie par ailleurs.

3.1 Racine de l'erreur quadratique moyenne

La littérature propose des métriques assez variées, de par le contexte d'étude et le type de vérité de terrain exploité. Elles peuvent reposer sur des mesures images 2D [26, 166] ou sur des données 3D, souvent plus complexes à obtenir [10, 68]. Dans notre cas, à l'instar de [10] et de [166], nous évaluons la précision de l'estimé par le calcul d'une distance moyenne entre les positions réelles des J liaisons $\boldsymbol{m}_j^{\boldsymbol{t}_k}, j \in 1, \ldots, J$ au fil du temps et leurs positions estimées $\boldsymbol{m}_i^{\widehat{\boldsymbol{x}_{k,r}}}$:

$$\frac{1}{J}\sum_{j=1}^{J}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{R}\sum_{r=1}^{R}||\boldsymbol{m}_j^{\boldsymbol{t}_k} - \boldsymbol{m}_j^{\widehat{\boldsymbol{x}_{k,r}}}||_2^2}, \tag{III.13}$$

où $r = 1, \ldots, R$ désigne une réalisation du filtre sur une séquence donnée. Ceci est donc la moyenne sur les articulations de la racine de l'erreur quadratique moyenne calculée sur les réalisations et les images. Nous la dénoterons RMSE par la suite.

Notons que le calcul de l'erreur repose sur une différence de positions d'articulations, et non pas sur des distances entre vecteurs d'état. En effet, un tel calcul impliquerait des différences d'angles, plus difficiles à manipuler, plusieurs vecteurs de configuration différents pouvant donner lieu à la même pose 3D.

3.2 Biais

Nous souhaitons également vérifier que les différentes réalisations des algorithmes de filtrage fournissent effectivement un estimé qui est en moyenne centré sur la vérité de terrain. Dans ce but, nous mettons en place le critère suivant :

$$\frac{1}{J}\sum_{j=1}^{J}||\frac{1}{K}\sum_{k=1}^{K}(\boldsymbol{m}_j^{\boldsymbol{t}_k} - \frac{1}{R}\sum_{r=1}^{R}\boldsymbol{m}_j^{\widehat{\boldsymbol{x}_{k,r}}})||_2, \tag{III.14}$$

qui correspond à la moyenne sur les articulations de la norme de l'écart par rapport à la vérité de terrain. Par abus de langage nous désignerons ce critère par le vocable de « biais ».

3.3 Erreur de Mahalanobis normalisée

Daum propose dans [32] une mesure de l'erreur par rapport à la vérité de terrain indépendante de la dimension de l'espace d'état et que nous adaptons à notre contexte :

$$\frac{1}{J}\sum_{j=1}^{J}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{R}\sum_{r=1}^{R}\frac{1}{3}(\boldsymbol{m}_j^{\boldsymbol{t}_k} - \boldsymbol{m}_j^{\widehat{\boldsymbol{x}_{k,r}}})^T\Delta_{j,k,r}^{-1}(\boldsymbol{m}_j^{\boldsymbol{t}_k} - \boldsymbol{m}_j^{\widehat{\boldsymbol{x}_{k,r}}})}, \tag{III.15}$$

où $\Delta_{j,k,r}$ est la covariance *a posteriori* de la position de l'articulation j à l'image k pour la réalisation r. Cette métrique présente l'avantage de prendre en compte le moment d'ordre 2 de la distribution *a posteriori* afin de ne pas accorder une importance

trop grande aux erreurs suivant les dimensions peu informatives. Toutefois, elle n'est généralement pas applicable dans notre contexte car il arrive que les matrices de covariance soient singulières et donc non inversibles de par la limite de précision de calcul sur machine. Signalons que si toutes les distributions étaient gaussiennes, le radicande de III.15 pourrait être considéré comme un estimateur du tiers de la moyenne d'une loi χ^2 à trois degrés de liberté, de sorte que III.15 prendrait une valeur proche de 1.

3.4 Variance de l'estimateur

Pour une même séquence de mesures, la variance de l'estimateur doit être analysée sur l'ensemble de ses réalisations en complément des autres critères. En effet, une variance excessive remettrait en cause la confiance que l'on peut accorder à une réalisation du filtre. Ceci est un point clé dans la mesure où tout système de HMC se doit de proposer un estimé peu dispersé pour une séquence donnée. La moyenne sur les articulations de la racine de la variance de l'estimateur est estimée par :

$$\frac{1}{J}\sum_{j=1}^{J}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{R}\sum_{r=1}^{R}||\boldsymbol{m}_j^{\widehat{\boldsymbol{x}_{k,r}}}-\frac{1}{R}\sum_{r=1}^{R}\boldsymbol{m}_j^{\widehat{\boldsymbol{x}_{k,r}}}||_2^2}. \tag{III.16}$$

De manière surprenante, ce critère reste peu étudié dans la littérature. Nous l'évoquerons par la suite sous le nom de « dispersion de l'estimateur ».

3.5 Taux d'échec

Afin de compléter les critères précédents, nous proposons de comptabiliser le nombre d'échecs du filtre. Nous considérons que le suivi « décroche » dès lors que la distance d'une articulation estimée à la vérité terrain est supérieure à un seuil S_{echec}. Ce nombre d'échecs calculé sur l'ensemble des images et des réalisations des filtres est alors ramené à un pourcentage *via* :

$$\frac{1}{R}\frac{1}{J}\frac{1}{K}\sum_{r=1}^{R}\sum_{j=1}^{J}\sum_{k=1}^{K}\text{fails}(\boldsymbol{m}_j^{\widehat{\boldsymbol{x}_{k,r}}}), \tag{III.17}$$

où $\text{fails}(\boldsymbol{m}_j^{\widehat{\boldsymbol{x}_{k,(r)}}}) = 1$ si $||\boldsymbol{m}_j^{\boldsymbol{t}_k} - \boldsymbol{m}_j^{\widehat{\boldsymbol{x}_{k,(r)}}}||^2 > S_{echec}^2$, 0 sinon.

Ces différentes métriques sont utilisées pour caractériser quantitativement le comportement des filtres. Le protocole expérimental mis en place, et les résultats préliminaires obtenus sur séquences de synthèses sont présentés dans la section suivante.

4 Évaluations préliminaires

4.1 Protocole expérimental

Afin de comparer les performances de chaque stratégie, nous avons tout d'abord choisi d'exploiter des séquences de synthèse sur lesquelles nous contrôlons tous les

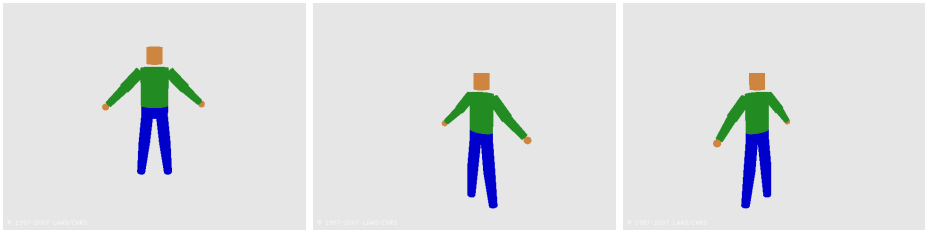


FIG. III.2 – 3 images provenant d'une séquence de synthèse présentant l'avatar de l'homme.

paramètres. Nous mettons ainsi en place un avatar 3D parfaitement conforme à notre modèle cinématique et géométrique de l'homme (figure III.2). Nous choisissons un environnement entièrement dégagé, afin de nous affranchir des fausses mesures et des artefacts de l'arrière-plan. Les paramètres intrinsèques et extrinsèques des caméras sont connus précisément. Nous nous plaçons donc dans un cas « idéal », où la réalité simulée est en parfait accord avec nos modèles.

Nous focalisons nos évaluations sur les stratégies suivantes : CONDENSATION, I-CONDENSATION, APF, IAPF, PARTITIONNÉ, HYBRID, QRS, PARTITIONNÉ QRS. Nous exploitons les deux contextes définis précédemment, *i.e.* multi-oculaire, avec 3 caméras, et stéréoscopique. L'évaluation des stratégies avec échantillonnage préférentiel (IAPF, I-CONDENSATION, HYBRID) se fait uniquement dans le cadre du contexte stéréoscopique qui permet la mise en place d'une fonction d'importance exploitant les mesures. Nous définissons cette fonction par :

$$q(\boldsymbol{x}_k|\boldsymbol{x}_{k-1},\boldsymbol{z}_k) = 0.85 \times p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) + 0.15 \times \pi(\boldsymbol{x}_k|\boldsymbol{z}_k) \tag{III.18}$$

Les valeurs $\alpha = 0.15$ et $\beta = 0$ sont choisies expérimentalement, au vu du meilleur résultat obtenu. Rappelons que la dynamique est une marche aléatoire, *i.e.* $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) = \mathcal{N}(\boldsymbol{x}_k;\boldsymbol{x}_{k-1},\Delta_k)$, de covariance diagonale Δ_k. Les écarts-types du bruit sur chacune des composantes du vecteur $\boldsymbol{x}_k$ sont respectivement $0.07\ m$ et $0.1\ rad$ pour les translations et les rotations, à l'exception des rotations propres du bras qui présentent un écart-type de $0.3\ rad$. Ce choix a été fait par l'expérience, et permet une meilleure robustesse du suivi lorsque le sujet a le bras temporairement tendu (impossibilité de caractériser avec précision l'angle de rotation propre du bras). Notons également qu'en pratique, les valeurs de chacune des articulations sont tronquées aux domaines humainement acceptables. Ceci résulte en une dynamique de type gaussienne tronquée, sans que l'exactitude mathématique des algorithmes mis en place n'en soit affectée. L'échantillonnage selon la mesure se fait suivant $\pi(\boldsymbol{x}_k|\boldsymbol{z}_k) = \mathcal{N}(\boldsymbol{x}_{IK},\Delta_k)$, où $\boldsymbol{x}_{IK}$ est la configuration du modèle déduite des images par cinématique inverse. La densité exploitée présente la même covariance Δ_k que la dynamique *a priori* du système.

50 réalisations des filtres sont lancées sur deux séquences vidéos de 200 images chacune présentant des mouvements simples afin de rester dans un cas très favorable. Nous étudions leur comportement au regard des critères définis dans la section 3 en

faisant varier le nombre de particules utilisées N de 200 à 2000. Nous choisissons d'exploiter deux partitions pour les algorithmes de types PARTITIONNÉ. La première est relative au placement du torse et la suivante affecte celui des membres, reprenant le principe appliqué par MacCormick dans [103]. À l'instar de la littérature [10, 38], nous choisissons d'utiliser trois couches au sein de l'APF, de manière à manipuler généralement entre 100 et 500 particules par couche dans l'optique d'une intégration en temps réel. Une évaluation future pourrait consister en une analyse plus fine de ces paramètres. Nous comparons les performances pour un nombre d'évaluations donné de la fonction de vraisemblance, celle-ci étant la plus consommatrice de temps. Ainsi, si N particules sont utilisées pour la CONDENSATION, alors l'APF est lancé avec $N/3$ particules, et le PARTITIONNÉ avec $N/2$ particules. Nous choisissons d'exploiter les mesures sil, $sil2$ et $dist\_peau$ dans le cas du contexte multi-oculaire, et $dist\_peau$, $dist$ et $blobs$ dans le cas du contexte stéréoscopique. Nous rappelons que ces mesures sont détaillées dans le chapitre II. Nous discutons et justifions ce choix de mesure dans le chapitre IV.

Notons ici que nous évaluons le comportement de chaque filtre pour une séquence de mesures donnée. Une analyse plus exhaustive demanderait l'exploitation de réalisations de mesures différentes. Ceci constitue cependant un objectif difficilement atteignable tant le nombre de contextes possibles est important. Nous supposons donc qu'une trajectoire de mesure est suffisamment représentative des différents contextes, cette hypothèse d'ergodicité permettant de remplacer des moyennes d'ensemble par des moyennes temporelles.

4.2 Résultats préliminaires

Les résultats sont présentés de la manière suivante :

	Multi-oculaire	**Stéréoscopique**
Séquence 1	Figure III.3	Figure III.6
Séquence 2	Figure III.4	Figure III.7

Les considérations ci-après se rapportent à ces figures. Notons d'emblée que plusieurs « classes » de méthodes semblent émerger : les stratégies de type PARTITIONNÉ, qui semblent être les plus efficaces au regard de nos critères, les techniques classiques de type CONDENSATION, et les stratégies reposant sur l'échantillonnage préférentiel. L'APF quant à lui semble avoir un comportement plus difficile à cerner.

Erreur quadratique moyenne

Les graphiques (a) des figures III.3, III.4 III.6, III.7 présentent la racine de l'erreur quadratique moyenne calculée selon la formule III.13. Selon [38], l'APF est statistiquement supérieur aux autres stratégies dès qu'un nombre minimum de particules est atteint. En deçà de ce nombre, le filtre ne présente pas de meilleurs résultats que les stratégies classiques. C'est le cas pour nos évaluations en contexte multi-oculaire. Les stratégies de type PARTITIONNÉ sont également plus efficaces en terme d'écart à la

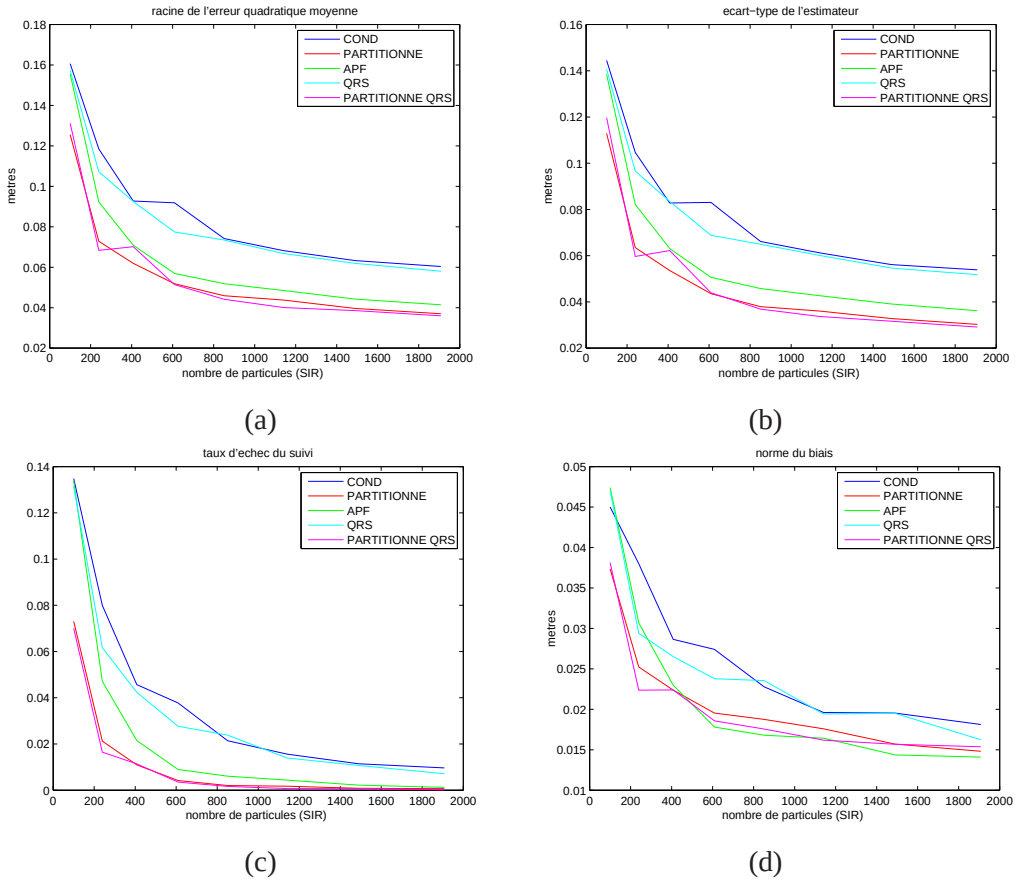


FIG. III.3 – Résultats des suivis sur la **séquence 1** pour un contexte de système **multi-oculaire** : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec du suivi et (d) norme du biais.

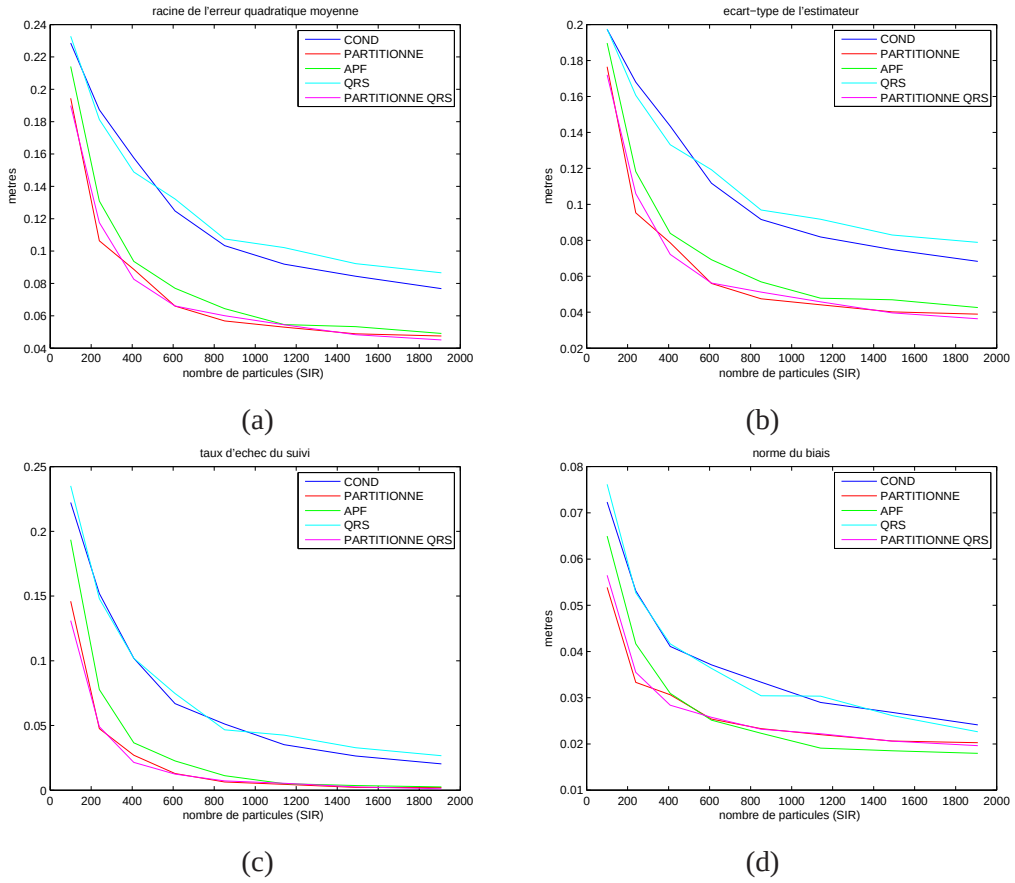


FIG. III.4 – Résultats des suivis sur la **séquence 2** pour un contexte de système **multi-oculaire** : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec du suivi et (d) norme du biais.

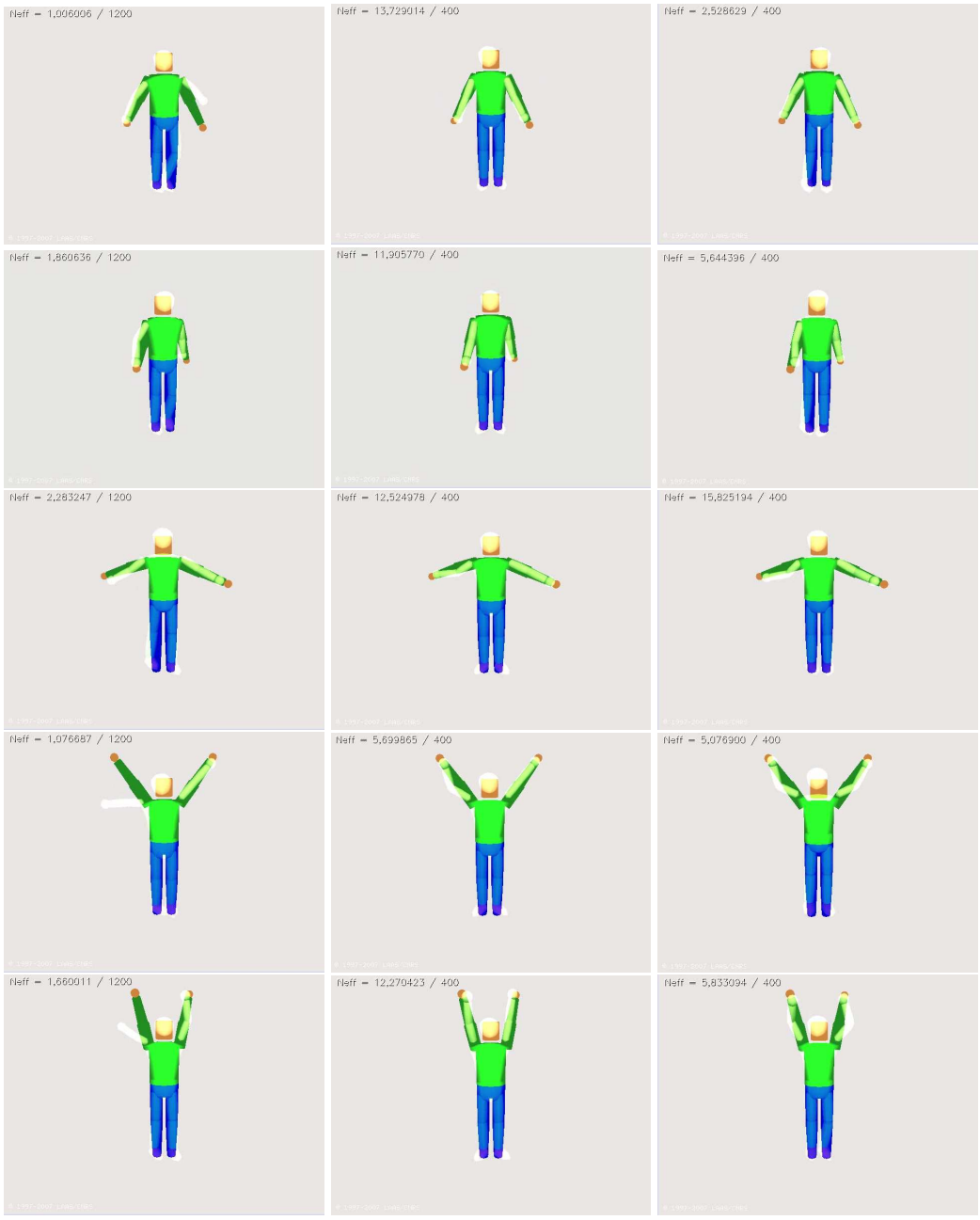


FIG. III.5 – Réalisation de différentes stratégies de filtrage pour la **séquence 1** dans le contexte **multi-oculaire** : CONDENSATION (gauche), PARTITIONNÉ QRS (milieu), APF (droite). L'avatar 3D représente l'état estimé. Une seule vue sur les trois exploitées est présentée.

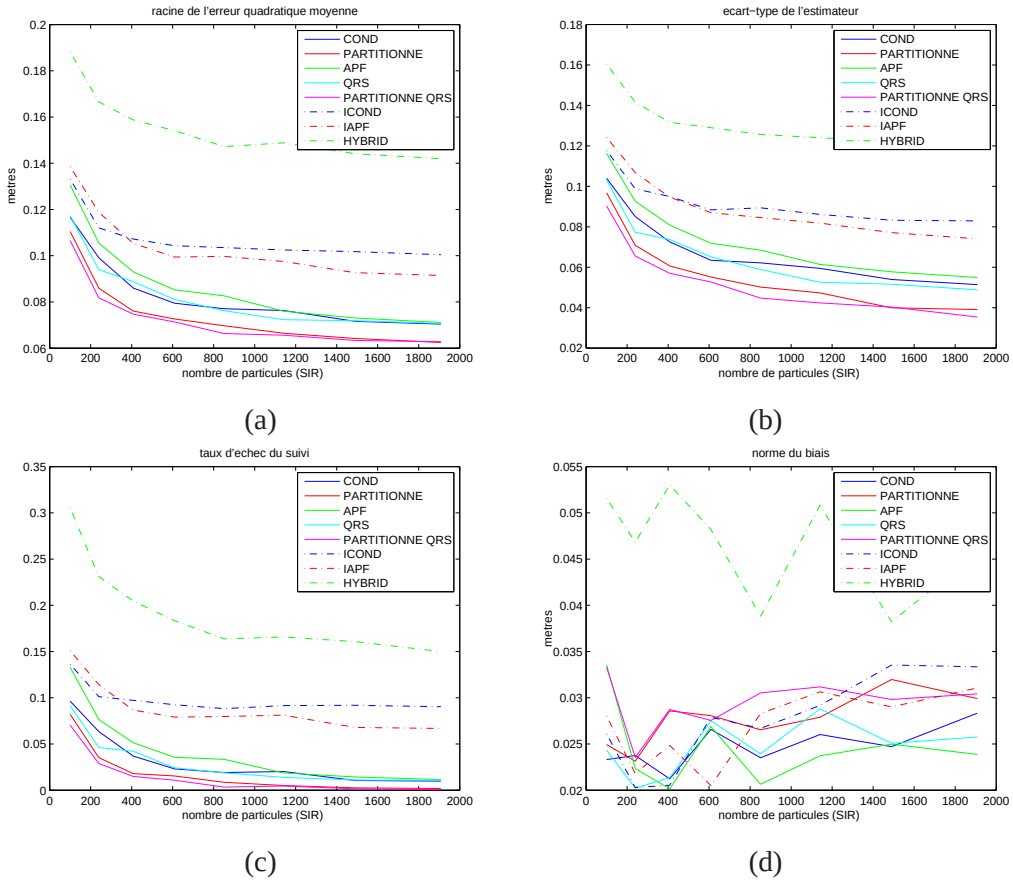


FIG. III.6 – Résultats des suivis sur la **séquence 1** pour un contexte de système **stéréoscopique** : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec du suivi et (d) norme du biais.

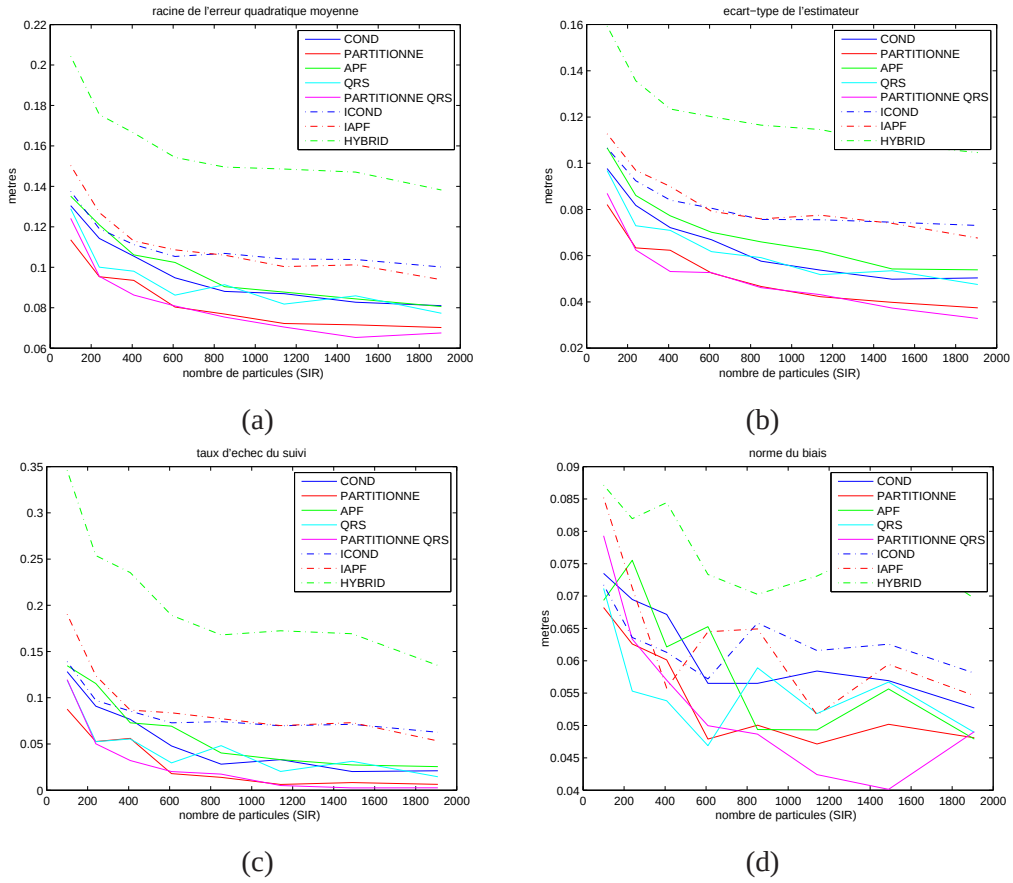


FIG. III.7 – Résultats des suivis sur la **séquence 2** pour un contexte de système **stéréoscopique** : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec du suivi et (d) norme du biais.

vérité terrain par rapport à l'algorithme pionnier CONDENSATION. Ceci paraît assez naturel dans la mesure où l'espace de recherche est divisé en sous-espaces de plus faibles dimensions pour lesquels le nombre de particules est suffisant pour améliorer la précision.

Les approches QMC semblent en moyenne au moins aussi efficaces que leurs homologues MC, même si l'amélioration est parfois très faible. Ceci est une conséquence de l'échantillonnage à faible discrépance des séquences QMC aléatoires. L'espace de recherche est exploré de manière plus uniforme ce qui résulte en un estimé en moyenne sensiblement meilleur. Notons que nous exploitons un algorithme optimisé afin de satisfaire des contraintes temps-réel au détriment de certaines caractéristiques inhérentes au filtrage QMC. Ses avantages sont donc moins marqués.

Alors que ces résultats apparaissent de manière claire pour un système multi-oculaire, ils sont plus atténués pour un système stéréoscopique. L'APF notamment présente une erreur qui reste du même ordre que la CONDENSATION. De même, les stratégies à échantillonnage préférentiel affichent des performances moins bonnes. Toutefois, elles ne sont pas à leur avantage dans ce type d'évaluation. En effet, leur but est de permettre le raccrochage des filtres lorsque la cible est complètement perdue. Ce cas n'arrive jamais dans notre contexte d'étude nominal puisque le sujet ne sort jamais du champ de vue et que nous avons choisi des indices visuels permettant un suivi satisfaisant. Notons cependant la performance sensiblement meilleure de l'IAPF par rapport à l'I-CONDENSATION. La stratégie HYBRID, quant à elle, ne s'avère pas aussi efficace que ses concurrentes.

La figure III.5 confirme ces résultats. Nous voyons que la CONDENSATION affiche un suivi des bras peu précis alors que le PARTITIONNÉ QRS est plus satisfaisant. L'estimé de l'APF est relativement acceptable bien que moins satisfaisant que le PARTITIONNÉ QRS. Toutefois, nous devons garder à l'esprit que les résultats présentés sur ces images sont issus d'une seule réalisation des filtres. Ceci nous conduit donc à l'analyse de la variance des estimateurs afin de caractériser leur dispersion.

Dispersion des estimateurs

Les graphiques (b) présentent la dispersion des estimateurs selon la formule III.16. Les stratégies de type PARTITIONNÉ présentent une dispersion plus faible que les autres méthodes. Une fois encore, nous présumons que cette propriété découle du découpage de l'espace en sous-espaces de dimensions plus faibles. Ainsi, le nombre de particules efficaces calculé selon III.9 et visible figure III.5 est plus élevé. L'estimé n'en sera que plus stable d'une réalisation sur l'autre. L'APF quant à lui présente une dispersion de l'estimé plus faible que la CONDENSATION mais toujours plus élevée que le PARTITIONNÉ pour un nombre raisonnable de particules. Notons par ailleurs que cette dispersion semble dépendante du contexte. Nous postulons, à l'instar de [10], que cette différence de comportement vient de la tendance du filtre APF à localiser un seul des modes de la distribution *a posteriori*. Dans le cas stéréoscopique, il est vraisemblable que celle-ci soit multi-modale, résultant en une estimation plus difficile pour un filtre tel que l'APF. Ainsi, il est plausible que la qualité des mesures exploitées influence le

choix de la stratégie de filtrage la plus adaptée au contexte. Les méthodes QMC présentent en moyenne une dispersion sensiblement moindre de l'estimé par rapport aux méthodes MC. Les stratégies d'échantillonnage préférentiel affichent les mêmes tendances pour la dispersion que pour la précision. Dans un cas nominal, l'introduction de l'échantillonnage d'importance semble donc légèrement dégrader les performances de ces filtres. L'APF et le PARTITIONNÉ QRS constituent des alternatives intéressantes au regard de ces deux critères.

Taux d'échec

Le taux d'échec, ici calculé pour $S_{echec} = 0.2$, est bien évidemment dépendant du seuil choisi, mais la robustesse relative des stratégies est globalement indépendante de cette valeur. Le résultat est présenté sur les graphiques (c) de chaque figure. Ces résultats sont intimement liés aux deux métriques précédentes. Ceci peut être vu comme le pourcentage de « perte de cible » des filtres.

Les tendances restent globalement les mêmes que précédemment. Le faible taux d'échec des stratégies de type PARTITIONNÉ et de l'APF font définitivement de ces deux méthodes un choix intéressant au regard de nos critères. Les stratégies QMC se révèlent une fois encore meilleures du point de vue du taux d'échec.

Biais

Les graphiques (d) présentent le biais calculé selon la formule III.14. Nous pouvons constater que celui-ci reste faible (de l'ordre de quelques centimètres) et qu'il tend vers 0 lorsque le nombre de particules augmente, même si cela semble parfois un peu plus irrégulier dans le cas de caméras stéréoscopiques. Nous pouvons en déduire que les filtres fournissent un estimé qui est en moyenne centré sur la vérité de terrain dans un contexte d'images de synthèse. Les mesures choisies sont manifestement suffisamment informatives pour permettre un suivi correct.

Les tableaux III.7 et III.8 résument les considérations ci-dessus pour un contexte d'image de synthèse. L'algorithme très utilisé d'APF, même si peu d'études quantitatives existent dans la littérature, propose de bons résultats, mais le PARTITIONNÉ QRS affiche des performances quasi identiques. Bien que les résultats soient satisfaisants dans un tel contexte, ils sont loin d'être parfaits alors que les conditions de suivi sont les plus favorables possibles. Gleicher et Ferrier mentionnent et développent ce problème dans [62]. Les résultats ci-dessus doivent également être nuancés par des évaluations en contexte réel.

5 Conclusion

Nous avons présenté dans ce chapitre de nombreuses stratégies de filtrage particulaire. Elles sont toutes basées sur l'algorithme générique SIR et exploitent différentes

Nom	Erreur	Dispersion	Taux d'échec	Biais
CONDENSATION	5	5	5	5
QRS	4	4	4	4
PARTITIONNÉ	2	2	2	3
PARTITIONNÉ QRS	**1**	**1**	**1**	**2**
APF	3	3	3	1

TAB. III.7 – Classement des différentes stratégies par critères pour un **contexte multi-oculaire**.

Nom	Erreur	Dispersion	Taux d'échec	Biais
CONDENSATION	5	4	5	5
QRS	3	3	3	4
PARTITIONNÉ	2	2	2	3
PARTITIONNÉ QRS	**1**	**1**	**1**	**2**
APF	4	5	4	1
I-CONDENSATION	7	7	7	7
IAPF	**6**	**6**	**6**	**6**
HYBRID	8	8	8	8

TAB. III.8 – Classement des différentes stratégies par critères pour un **contexte stéréoscopique**.

idées afin d'améliorer son fonctionnement : échantillonnage préférentiel, partitionnement de l'espace d'état, approche par raffinement, méthodes de QMC. Nous avons abordé les atouts et faiblesses de chacune, ainsi que les points difficiles caractéristiques de leur mise en œuvre relatés dans la littérature. Nous avons présenté des métriques dérivées de ces observations que nous avons exploitées dans une première évaluation du système de suivi sur séquences de synthèse.

Il en ressort que dans un contexte multi-oculaire le comportement des stratégies est celui que l'on est en droit d'attendre, désignant les stratégies PARTITIONNÉ QRS et APF comme les plus efficaces. En contexte stéréoscopique, les mesures potentiellement moins informatives que nous exploitons dégradent sensiblement le fonctionnement nominal de chacun des filtres, et il est plus difficile de parvenir à des résultats aussi tranchés. L'APF notamment affiche un comportement décevant, probablement dû à sa tendance à modéliser le pic proéminent de la distribution *a posteriori*, perdant ainsi l'intérêt de la représentation potentiellement multimodale offerte par le support particulaire. On peut en déduire, à l'instar de [10], que les mesures exploitées et le profil de la distribution de filtrage jouent un rôle important dans le choix d'une stratégie adaptée au contexte.

Dans la suite du manuscrit, nous apportons un éclairage nouveau sur ces évaluations dans le cadre de séquences réelles.

Chapitre IV

Évaluation sur séquences réelles

Ce chapitre présente et discute les performances de stratégies et mesures implémentées sur des données réelles. La vérité de terrain est ici obtenue à l'aide d'un système HMC commercial. Nous détaillons le protocole expérimental mis en place, puis dans une deuxième partie, nous déroulons un ensemble d'évaluations basées sur les critères présentés dans le chapitre III afin de proposer une association de mesures et technique de filtrage. Plus spécifiquement, nous discutons le comportement des différentes stratégies et l'importance du choix des indices visuels impliqués.

1 Protocole expérimental

Évaluer un système de capture de mouvement par vision nécessite un protocole lourd et complexe. Dans ce contexte, la littérature propose de nombreuses évaluations qualitatives [38, 103, 146], tandis que les évaluations quantitatives sont encore marginales. Wang et Rehg [166] mentionnent plusieurs difficultés inhérentes aux évaluations quantitatives. En premier lieu, il n'existe pas à l'heure actuelle de séquences de tests exhaustives publiques proposant différents contextes avec des difficultés graduelles. Notons toutefois que, de par l'essor que connaît le domaine de suivi visuel de mouvement humain, plusieurs bases de données communes commencent à voir le jour [143, 167, 169, 170], mais elles restent encore peu utilisées. Ensuite, la mise en place d'une vérité de terrain est également un problème complexe. Elle peut être faite « à la main » par repérage sur chaque image de l'objet à suivre [26, 142]. Ceci est particulièrement long et fastidieux, mais ne requiert pas de matériel spécial supplémentaire. Une alternative consiste à utiliser un système de HMC commercial [10, 68] proposant des données plus précises, mais nécessitant une mise en œuvre plus lourde. En outre, ce type d'équipement, en plus de son coût important, demande un espace de travail dégagé qui ne convient pas à toutes les applications. Une autre difficulté est la comparaison des résultats obtenus à la vérité de terrain, qui nécessite la mise en place de métriques. Dans notre cas, nos évaluations s'appuient sur les critères présentés en chapitre III. Enfin, afin de comparer les performances des différentes techniques envisagées, les résultats doivent être normalisés par rapport au temps de calcul effectif de chaque stratégie.

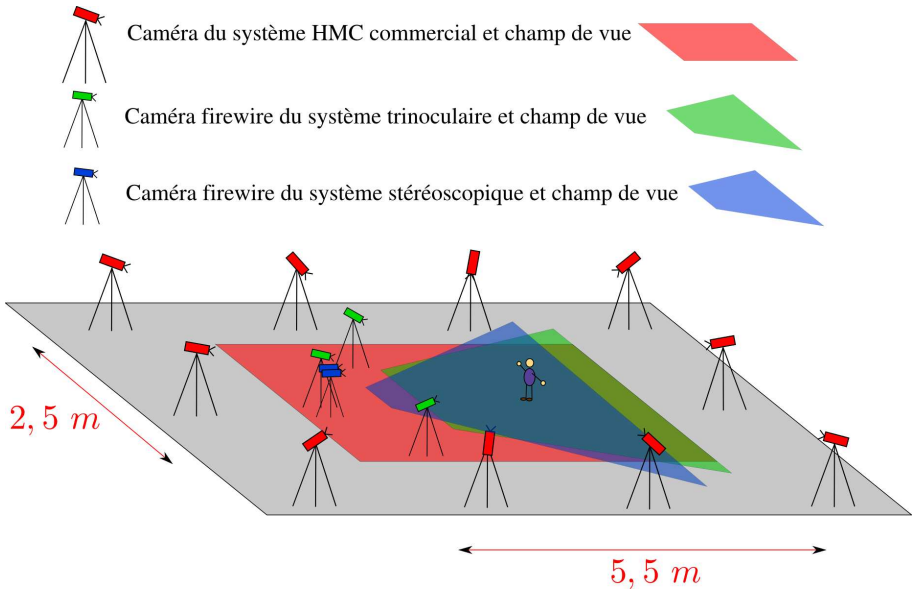


FIG. IV.1 – Configuration de nos deux systèmes de suivi visuel et du système de HMC commercial.

Une méthode, que nous adoptons ici, consiste alors à normaliser par rapport au nombre d'évaluations de la fonction de vraisemblance [10].

1.1 Acquisition de la vérité de terrain

Afin d'établir une vérité de terrain pour les acquisitions vidéos, nous exploitons un système HMC commercial développé par Motion Analysis [171]. Le système est constitué de 10 caméras infra-rouges dont 6 caméras Eagle à 1280×1024 pixels et 4 caméras Hawk à 640×480 pixels qui acquièrent les données à $100\ Hz$. Ces dernières présentant une résolution moindre, elles sont placées au plus près de l'espace de travail. La zone utile recouvre une surface de $2,5 \times 5,5\ m$ (figure IV.1). Le sujet porte des marqueurs réfléchissant la lumière infra-rouge émise par les matrices de DEL (Diodes Électro-Luminescentes) situées sous les caméras. La position des marqueurs dans les images est localisée, et la position 3D correspondante est triangulée. Une fois les positions 3D des marqueurs calculées, un traitement hors-ligne est appliqué pour corriger les données. En effet, quelques erreurs dues à des pertes de marqueurs dans les images peuvent survenir, et certains artefacts peuvent apparaître, en raison d'objets réfléchissant où émettant de la lumière infra-rouge dans la scène (voyants de matériel électronique ou surface particulièrement lisse). Par la suite, une identification semi-automatique des marqueurs est effectuée et le post-traitement est appliqué. Dans notre cas, nous utilisons une méthode de cinématique inverse permettant le recalage du modèle de l'homme à chaque instant à partir de la position 3D des marqueurs. À partir de chaque configuration calculée, nous extrayons les positions associées des articula-

FIG. IV.2 – Reprojection des articulations calculées par le système de HMC commercial dans les images acquises (en rouge).

tions qui constitueront notre vérité de terrain. La figure IV.2 montre la reprojection de la vérité terrain dans des images acquises depuis les caméras standards.

Notons que malgré la précision du système utilisé, le résultat n'est pas parfait et quelques légères incohérences peuvent apparaître. En effet, les marqueurs ne sont pas rigidement liés aux membres du sujet et les techniques d'interpolation appliquées pour reconstruire les données manquantes sont sources d'imprécision dans la reconstruction de la posture 3D du sujet. Enfin, les différentes étapes du processus d'acquisition de la vérité de terrain peuvent s'avérer relativement fastidieuses, ce qui fait de ce système une méthode lourde et complexe à mettre en œuvre.

1.2 Acquisition des images

Nous couplons ce dispositif avec les caméras couleurs de notre système de vision comme indiqué en figure IV.1. La littérature fait état de nombreux systèmes de suivi multi-oculaires exploitant entre deux et une dizaine de caméras [21, 41, 91, 122]. Toutefois, Balan *et al.* montrent dans [10] que 3 caméras sont suffisantes pour mettre en œuvre un système de suivi efficace. Dans le cas du contexte multi-oculaire, les caméras sont placées du même côté de la scène avec des angles de vues différents. Le choix d'une telle disposition des caméras est ici imposée par des contraintes physiques : la station centralisant les données ne doit pas être distante de plus de $4\ m$ de toutes les caméras pour des raisons de stabilité de l'alimentation électrique. Pour le contexte stéréoscopique, les caméras sont placées au centre de la scène. La zone utile exploitable par les 3 systèmes utilisés est d'environ $2,5\ m$ sur $3\ m$.

L'acquisition s'effectue au moyen de caméras IEEE1394b Flea 2 Color [168] « progressive scan » de 640×480 pixels à la cadence de $4\ Hz$ pour le cas multi-oculaire et $6\ Hz$ pour le cas stéréoscopique. Le débit est essentiellement limité par le temps des accès au disque pour l'archivage des données, la taille des images et le nombre de caméras utilisées. La synchronisation des caméras est faite de manière logicielle car ce procédé est plus simple à mettre en œuvre dans notre cas.

1.3 Calibration des systèmes

Afin d'exploiter la vérité de terrain fournie par le système de HMC commercial et de projeter l'état hypothétique $\boldsymbol{x}_k$ dans les images, nous devons calibrer les systèmes

spatialement. À ces fins, nous utilisons une mire classique pour les caméras couleurs sur laquelle nous positionnons des marqueurs réfléchissants pour les caméras infrarouges. On peut ainsi obtenir la transformation rigide permettant le passage d'un système à l'autre. Les caméras couleurs sont calibrées sous MATLAB [1] et les caméras du système HMC commercial sont calibrées avec le logiciel fourni par Motion Analysis. Toutes les données sont étiquetées temporellement, et nous utilisons un signal de départ et de fin d'acquisition visible par tous les systèmes afin de s'assurer de la synchronisation. La synchronisation matérielle et/ou logicielle du système de HMC avec les caméras couleurs n'est pas envisageable dans notre cas. La fréquence du système HMC commercial reste cependant bien plus élevée que celle de nos systèmes visuels. La vérité de terrain associée à un ensemble d'images prises par les caméras couleurs est alors celle dont l'étiquette temporelle est la plus proche, la différence restant inférieure à $10\,ms$. Notons cependant que ce processus n'est pas parfait et peut introduire un léger biais par rapport à la vérité terrain.

2 Contexte multi-oculaire

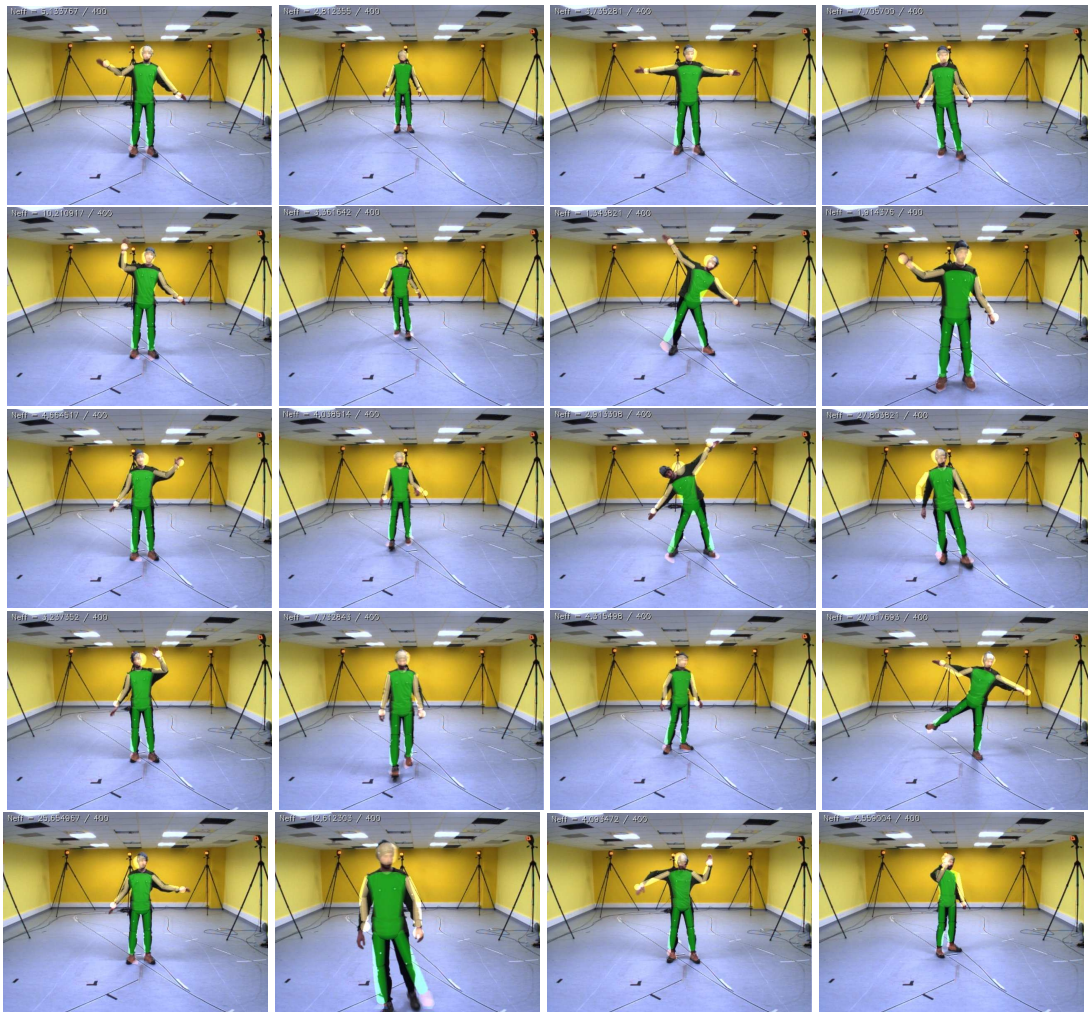
Nous reprenons le même protocole que celui décrit dans le chapitre précédent. Les algorithmes sont exécutés $R = 30$ fois sur différentes séquences. Nous comparons les stratégies CONDENSATION, QRS, PARTITIONNÉ, PARTITIONNÉ QRS et APF, qui ne reposent pas sur un échantillonnage préférentiel et qui peuvent donc être appliquées dans ce contexte. Nous poursuivons les évaluations avec la même configuration des filtres PARTITIONNÉ et APF que celle décrite au chapitre III. Notons que peu de travaux comparent différentes stratégies de filtrage ; ils se limitent très souvent à une comparaison avec la CONDENSATION et/ou l'APF [10, 41].

Nous exploitons ici quatre séquences dont les caractéristiques sont résumées dans le tableau IV.1. Selon la classification énoncée par Gupta *et al.* dans [68], elles sont de classes 1 et 2, *i.e.* présentent respectivement un seul sujet effectuant des mouvements sans et avec occultations propres. L'exécution d'un suivi sur ces séquences est présentée en figure IV.3. Notons que la littérature présente classiquement des résultats issus de l'étude d'un nombre de séquences assez faible (une seule séquence de marche pour [10], deux séquences pour [29]).

Afin de trouver les meilleures associations filtres/mesures, nous proposons en premier lieu une méthode pour choisir et configurer de manière optimale les indices visuels que nous allons exploiter.

2.1 Choix des mesures et configuration optimale

Le choix des mesures et leur mise en œuvre ont un impact important sur le comportement des stratégies de filtrage. Dans cette section, nous tirons parti de nos évaluations pour discuter leurs influences sur les filtres. Le descriptif des mesures considérées est détaillé dans le chapitre II. Les évaluations sont faites pour le cas de la CONDENSATION, qui est la stratégie générique sur laquelle reposent les algorithmes plus avan-



séquence 1 **séquence 2** **séquence 3** **séquence 4**

FIG. IV.3 – Déroulement de la stratégie APF avec 3 couches et 500 particules sur les **séquences 1 à 4** (une sur chaque colonne) en contexte **multi-oculaire**. Les images sont issues de la caméra centrale. Nous utilisons les mesures sil, $sil2$ et $dist\_peau$.

	Caractéristiques	Nombre d'images	Durée
Séquence 1	Mouvement fronto-parallèle des bras, buste fixe	78	19 s
Séquence 2	Marche	50	12 s
Séquence 3	Gymnastique, mouvement des bras, du buste et des jambes	100	25 s
Séquence 4	Marche et mouvement non fronto-parallèle des bras	80	22 s

TAB. IV.1 – Séquences multi-oculaires étudiées.

cés. Nous n'aborderons pas l'efficacité d'une mesure donnée en fonction du bruit inhérent [10, 99], mais nous nous focalisons sur la combinaison des indices visuels.

Dans ce contexte multi-oculaire, la segmentation de silhouette est très répandue dans la littérature [38,91,122]. Balan *et al.* montrent par ailleurs que son exploitation est plus intéressante que celles des contours [10]. Toutefois, l'exploitation de ce seul indice visuel n'est parfois pas suffisante pour assurer un suivi correct [145]. Afin d'exploiter au mieux cette mesure (précédemment dénotée sil), nous menons des évaluations pour différentes valeurs de σ_{sil}. Les résultats sont présentés en figure IV.4. En premier lieu, nous pouvons constater que la valeur optimale n'est pas la même en fonction du critère que l'on souhaite minimiser. Ainsi un RMSE optimal conduit à $\sigma_{sil} = 22$, la dispersion minimum nous indique $\sigma_{sil} = 255$, alors qu'un biais optimal sera atteint pour $\sigma_{sil} = 1$. Notons qu'en deçà de cette valeur, le suivi diverge car les poids prennent des valeurs nulles de par la limite de précision du calcul sur machine. De manière intuitive, il paraît logique que la dispersion soit minimale pour un écart-type maximal : les particules ont alors des poids plus équilibrés conduisant à un estimé plus lissé. Le comportement du biais semble cohérent avec l'intuition que l'on aurait à prendre une valeur de $\sigma_{(.)}$ très faible afin de ne favoriser que les configurations qui affichent une distance de similarité D la plus proche de 0 possible. Il y a ainsi manifestement un compromis à faire entre la dispersion des estimés (qui se traduit visuellement par un tressautement du modèle recalé) et le biais ou l'erreur moyenne. Ceci n'est, à notre connaissance, jamais mentionné dans la littérature.

Toutefois, indépendamment du choix de σ_{sil}, le suivi n'est pas satisfaisant (figure IV.5, où nous avons choisi $\sigma_{sil} = 30$). Le problème majeur est généralement la localisation des bras, d'apparence plus fine que les membres inférieurs. La fonction de vraisemblance favorise alors les configurations où les bras apparaissent à l'intérieur de la silhouette du torse, plus large. Nous pouvons en déduire que l'indice visuel sil n'est pas suffisant dans notre contexte. Le filtre doit logiquement fusionner plusieurs mesures.

Afin de résoudre ce problème, on met en place la mesure de silhouette duale $sil2$ qui favorise les configurations recouvrant l'ensemble de la silhouette segmentée (Figure IV.6). De la même manière, nous étudions son comportement pour différentes

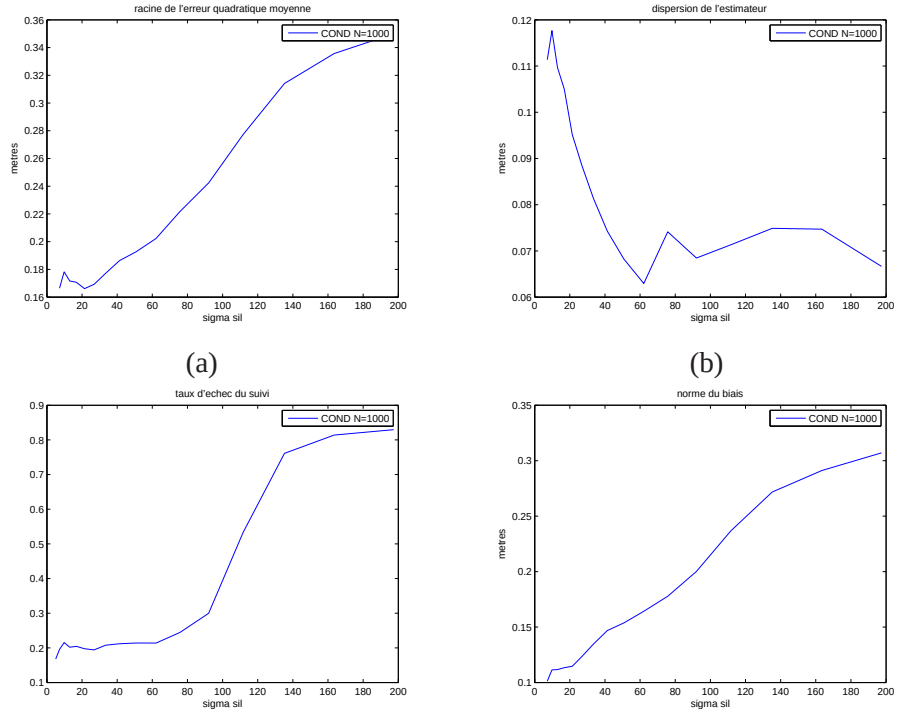


FIG. IV.4 – Influence du paramètre σ_{sil} en contexte **multi-oculaire** sur la **séquence 1** pour $R = 20$ réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais.

FIG. IV.5 – Suivi du sujet avec pour seule mesure la segmentation de la silhouette (sil). Au fur et à mesure du suivi, les bras ont tendance à se positionner devant ou derrière le corps.

FIG. IV.6 – Suivi du sujet exploitant les mesure sil et $sil2$. Le suivi des bras est plus satisfaisant, mais encore instable lorsque ceux-ci se rapprochent du corps.

valeurs de σ_{sil2} et choisissons le résultat optimal $\sigma_{sil2} = 0.07$. Toutefois, cette mesure supplémentaire n'est parfois pas suffisante pour assurer un meilleur suivi des bras, notamment lorsque ceux-ci sont près du corps. En complément, il convient d'utiliser des mesures dédiées à la localisation de ces parties spécifiques. C'est pourquoi nous combinons ces deux indices visuels avec une localisation des blobs de couleur peau (dénotée $dist\_peau$), avec $\sigma_{dist\_peau} = 5$, ce qui conduit au comportement satisfaisant présenté figure IV.3.

Une fusion de données judicieuses apporte une efficacité supplémentaire au suivi. Dans notre cas, nous avons réduit l'erreur moyenne de 17 cm (pour sil) à 9 cm (pour sil, $sil2$ et $dist\_peau$) pour $N = 1000$ particules avec la CONDENSATION (cf. Figure IV.8) tout en arrivant à un suivi visuellement satisfaisant. Toutefois, nous pouvons nous demander si les mesures sil et $sil2$ ne sont pas redondantes et si la deuxième ne serait pas suffisante dans un tel contexte. Nous évaluons de nouveau le comportement du suivi pour différentes valeurs de σ_{sil}, mais cette fois en exploitant également les mesures $sil2$ et $dist\_peau$ (Figure IV.7). Le comportement du filtre pour les grandes valeurs de σ_{sil} nous informe sur la pertinence de la mesure. En effet, un $\sigma_{(.)}$ élevé n'accorde que très peu d'importance à l'indice considéré, *i.e.* le filtre a tendance à se comporter comme si la mesure n'était pas prise en compte. L'ensemble des graphiques présentés en figure IV.7 nous indique que la mesure sil améliore sensiblement le RMSE (a) pour $\sigma_{sil} = 30$ mais très peu les autres indices.

Ceci nous amène à deux conclusions : d'une part, il est légitime de remettre en cause l'utilité réelle de la mesure sil conjointement à $sil2$ et $dist\_peau$ au vu de son amélioration médiocre de la performance dans un contexte où le temps de calcul est un critère important. Elle semble redondante avec la mesure $sil2$ et on peut donc envisager de n'exploiter que l'une des deux. Ensuite, ce comportement témoigne d'une corrélation (que l'on peut intuitivement deviner) entre cette mesure et les autres. Ceci peut partiellement remettre en cause l'hypothèse d'indépendance des mesures conditionnellement au vecteur d'état, car elle semble imparfaitement vérifiée.

Afin de proposer un suivi le plus efficace possible, il convient donc de mettre en œuvre une fusion de données complémentaires afin d'assurer un comportement du filtre satisfaisant. Nous pouvons toutefois supposer que la technique d'optimisation des valeurs de $\sigma_{(.)}$ ici proposée n'est pas parfaite. Il faudrait en effet explorer l'ensemble des

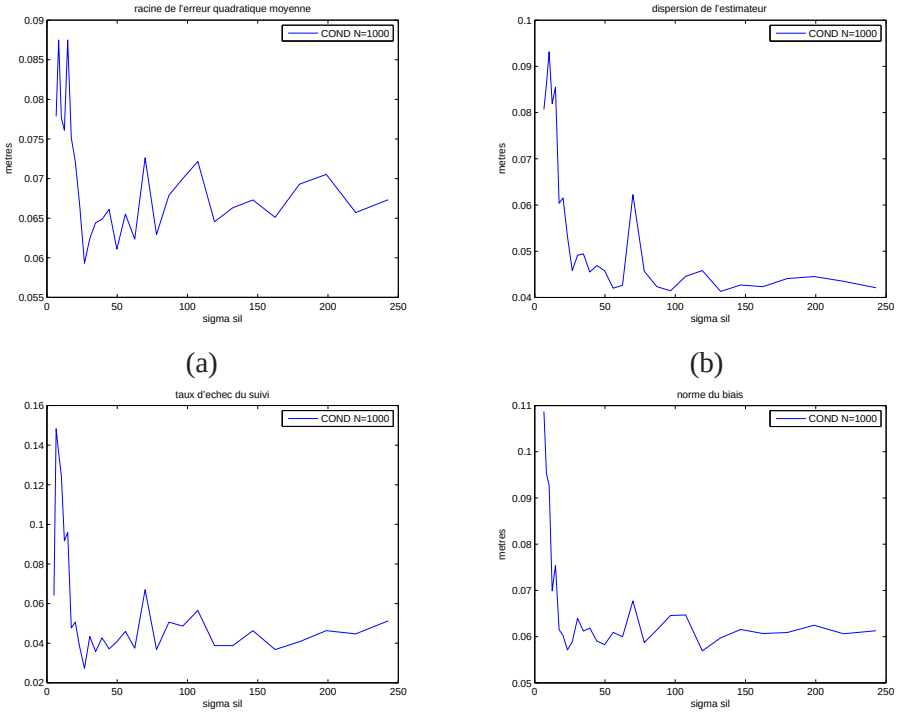


FIG. IV.7 – Influence du paramètre σ_{sil} en utilisation conjointe avec les mesures $sil2$ et $dist\_peau$ en contexte **multi-oculaire** sur la **séquence 1** pour $R = 20$ réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais.

configurations possibles pour $(\sigma_{sil}, \sigma_{sil2}, \sigma_{dist\_peau})$. Une telle approche est cependant complexe à mettre en œuvre en pratique. L'approche que nous proposons permet d'appréhender de manière plus pragmatique l'influence de ces paramètres. Toutefois, elle reste relativement lourde si l'on considère plusieurs fonctions de mesure.

2.2 Stratégies de filtrage et nombre de particules

Après avoir proposé un jeu de mesures qui permet un suivi correct, nous menons diverses évaluations afin de mieux cerner le comportement des techniques de filtrage particulaire pour le suivi visuel sur des séquences réelles. Nous exécutons les différentes stratégies $R = 30$ fois sur chaque séquence étudiée pour un nombre de particules allant de 200 à 2000. Au delà, le suivi en temps réel n'est plus envisageable. Pour éviter une surabondance de figures, seuls les résultats sur une ou deux séquences représentatives du comportement global des filtres sont présentés. Les autres graphiques peuvent être trouvés en annexe C. Les figures IV.8 et IV.9 exposent les résultats obtenus sur les séquences 1 et 2 respectivement. .

Dans l'ensemble, les performances obtenues sur séquences réelles et synthétiques (chapitre III) sont concordantes. Les erreurs sont inférieures à 12 centimètres en moyenne. Nous restons sensiblement moins précis que les résultats de la littérature, généralement inférieurs à 10 centimètres (entre 5 et 15 centimètres pour [68]). Ceci peut être le fait de notre modèle fruste du corps humain et de nos mesures simples. En outre, pour améliorer la précision des filtres, le nombre de particules à utiliser semble croître exponentiellement [10]. Nos contraintes de temps-réel limitent le nombre de particules résultant en une précision moindre. Les erreurs de localisation sont également plus importantes lorsque le sujet est éloigné des caméras. Toutefois, nous nous intéressons ici aux performances relatives, et le comportement qualitatif des filtres est satisfaisant, comme illustré sur la figure IV.3.

Notons également le biais plus important que sur les séquences de synthèse, qui indique que l'inadéquation entre nos modèles simples (de mesure, de l'homme) dégrade notablement les performances des filtres. Ce biais diminue cependant avec le nombre de particules. Notons que la limite du biais pour $N \rightarrow \infty$ fixerait la limite physique de précision que les algorithmes ne pourraient pas dépasser pour un ensemble de mesures données quelle que soit la stratégie envisagée. Nous pouvons également remarquer que l'efficacité relative des différentes stratégies vis-à-vis de ce critère est moins constante d'une séquence à l'autre (figure IV.8 (d) et IV.9 (d)).

Au vu du taux d'échec, les méthodes QMC peuvent apporter un gain de particules non négligeable dans certains cas (jusqu'à $25\ \%$ environ dans le cas de la séquence 3). Elles apportent manifestement une efficacité supplémentaire sur les critères de précision et/ou de dispersion de l'estimé. Ce constat est plus marqué que sur les séquences de synthèse : les méthodes QMC sont pertinentes dans des contextes où les fonctions de vraisemblance sont *a priori* multimodales.

Le tableau IV.2 résume ces considérations. Dans la pratique, le PARTITIONNÉ QRS semble être un choix tout indiqué pour le suivi en contexte multi-oculaire. L'APF est également une alternative intéressante pour un nombre de particules suffisant.

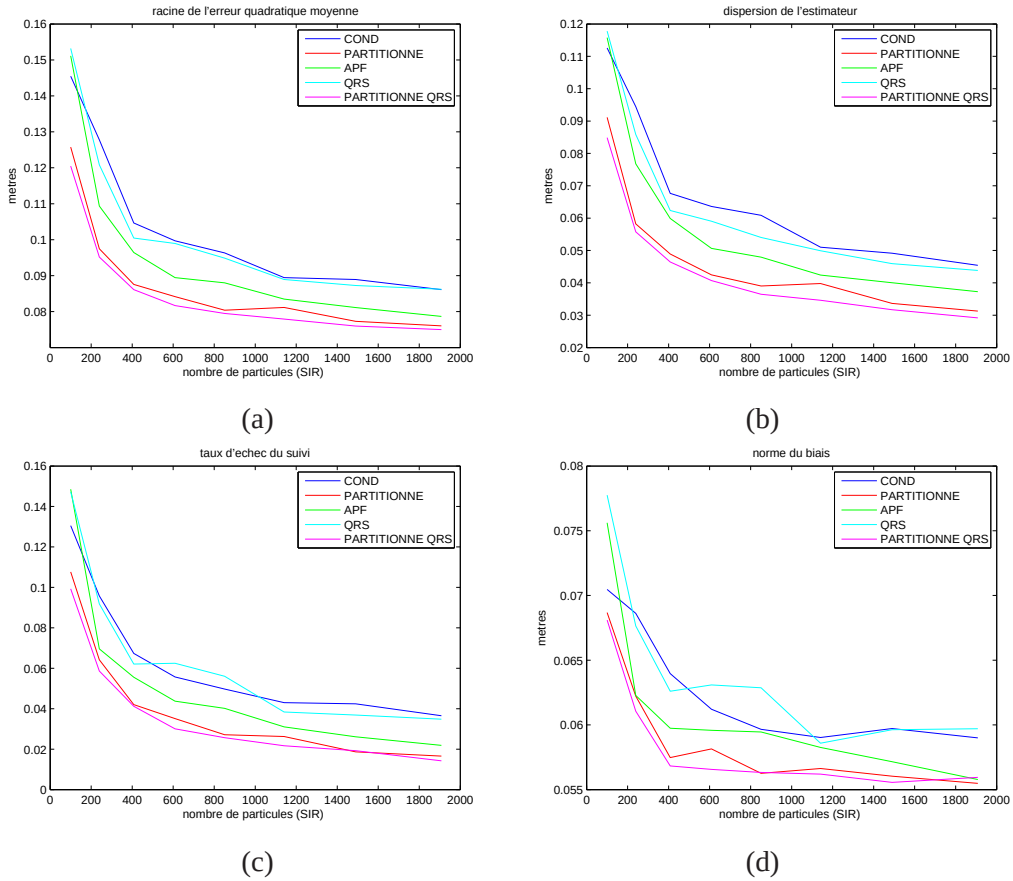


FIG. IV.8 – Influence du nombre de particules N et de la stratégie de filtrage en contexte **multi-oculaire** sur la **séquence 1** pour $R = 30$ réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais.

Nom	**Erreur**	**Dispersion**	**Taux d'échec**	**Biais**
CONDENSATION	5	5	5	5
QRS	4	4	4	4
PARTITIONNÉ	3	3	3	3
PARTITIONNÉ QRS	**2**	**1**	**2**	**1**
APF	**1**	**2**	**1**	**2**

TAB. IV.2 – Classement des différentes stratégies par critères pour un contexte **multi-oculaire**. En gras sont précisés les meilleurs compromis.

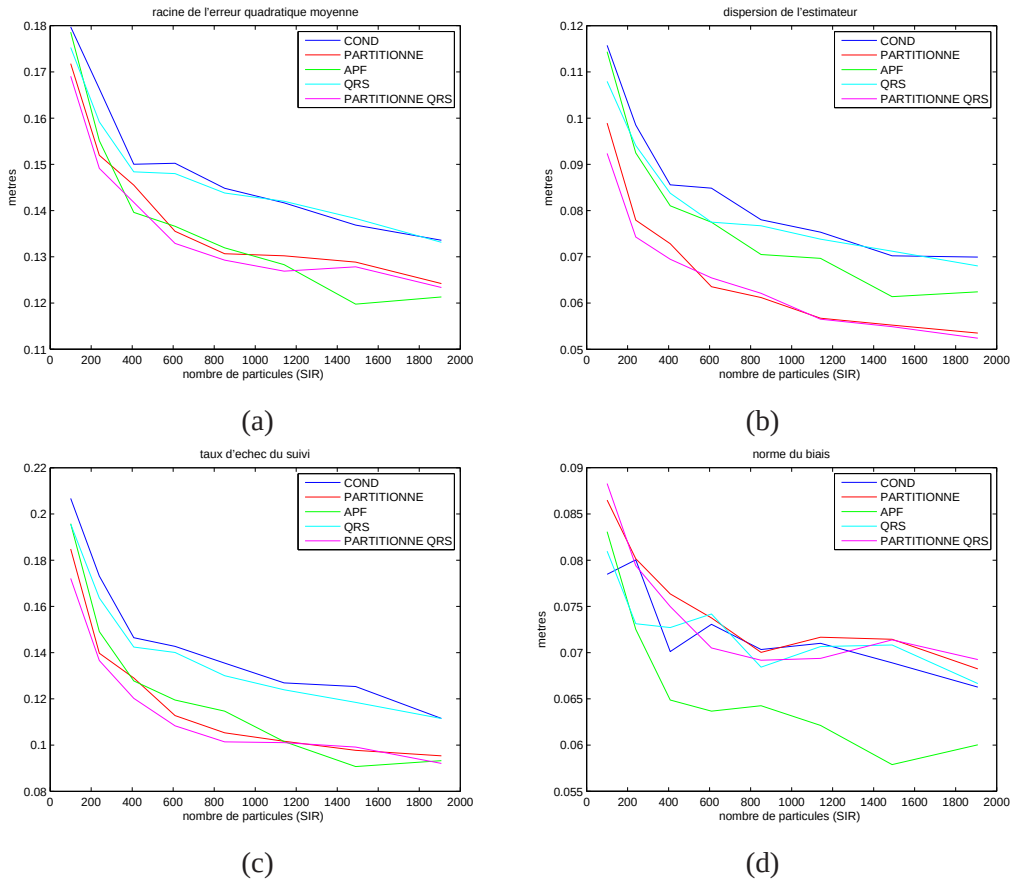


FIG. IV.9 – Influence du nombre de particules N et de la stratégie de filtrage en contexte **multi-oculaire** sur la **séquence 2** pour $R = 30$ réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais.

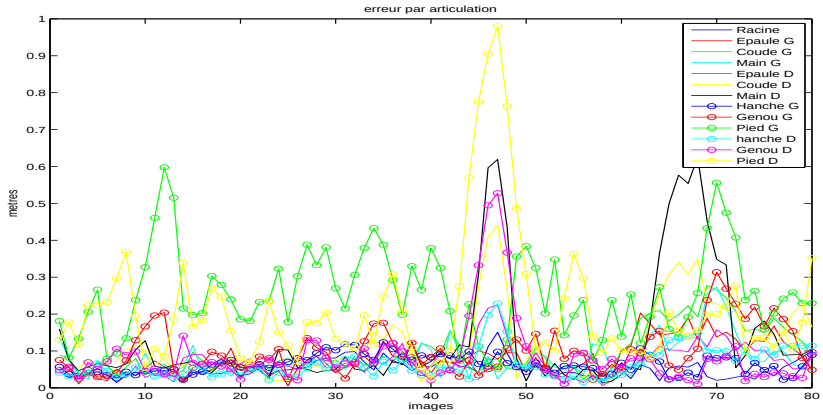


FIG. IV.10 – Erreur par articulation au cours d'une réalisation du suivi pour la **séquence 4**.

2.3 Erreur par articulation

Le RMSE renseigne sur le comportement global du suivi, mais il convient également de s'intéresser à la localisation de chaque articulation de manière indépendante [10, 68]. Nous présentons en figure IV.10 une analyse de l'erreur par articulation au cours d'une séquence représentative du comportement moyen des différents filtres. A l'instar de [10], nous pouvons voir que les erreurs sont essentiellement localisées sur les extrémités des membres, notamment les pieds. Le bassin est localisé de manière précise et l'erreur sur les mains est largement réduite par la mise en place de la mesure $dist\_peau$. La difficulté de localisation des pieds est due à la segmentation difficile de la silhouette au niveau du sol de par la présence de l'ombre. Ceci est cohérent avec l'analyse faite par Balan *et al.* dans [10] qui mentionnent que la clef d'un système de suivi précis est la qualité de la segmentation de la silhouette. Dès lors que cet indice visuel n'est plus utilisé, le filtre perd une information importante et la localisation est plus complexe. Ceci souligne l'importance de l'exploitation de mesures dédiées à certains membres corporels. Il convient ainsi d'exploiter au maximum des mesures « spécifiques » à chaque membre en complément des mesures plus génériques (telles que la segmentation de la silhouette), ces dernières demandant un nombre de particules important pour explorer l'espace d'état de manière exhaustive afin de compenser cette généricité.

Après ces évaluations menées dans le cadre de suivi multi-oculaire, nous nous focalisons dans la suite sur le contexte stéréoscopique.

	Caractéristiques	Nombre d'images	Durée
Séquence 1	Mouvement fronto-parallèle des bras, buste fixe	100	17 s
Séquence 2	Mouvement non fronto-parallèle des bras et déplacement du buste	100	16 s
Séquence 3	Mouvement complexe des bras, flexion des jambes, déplacements et inclinaison du buste	336	54 s
Séquence 4	Mouvement complexe des bras, déplacement du buste	186	30 s

TAB. IV.3 – Séquences stéréoscopiques étudiées.

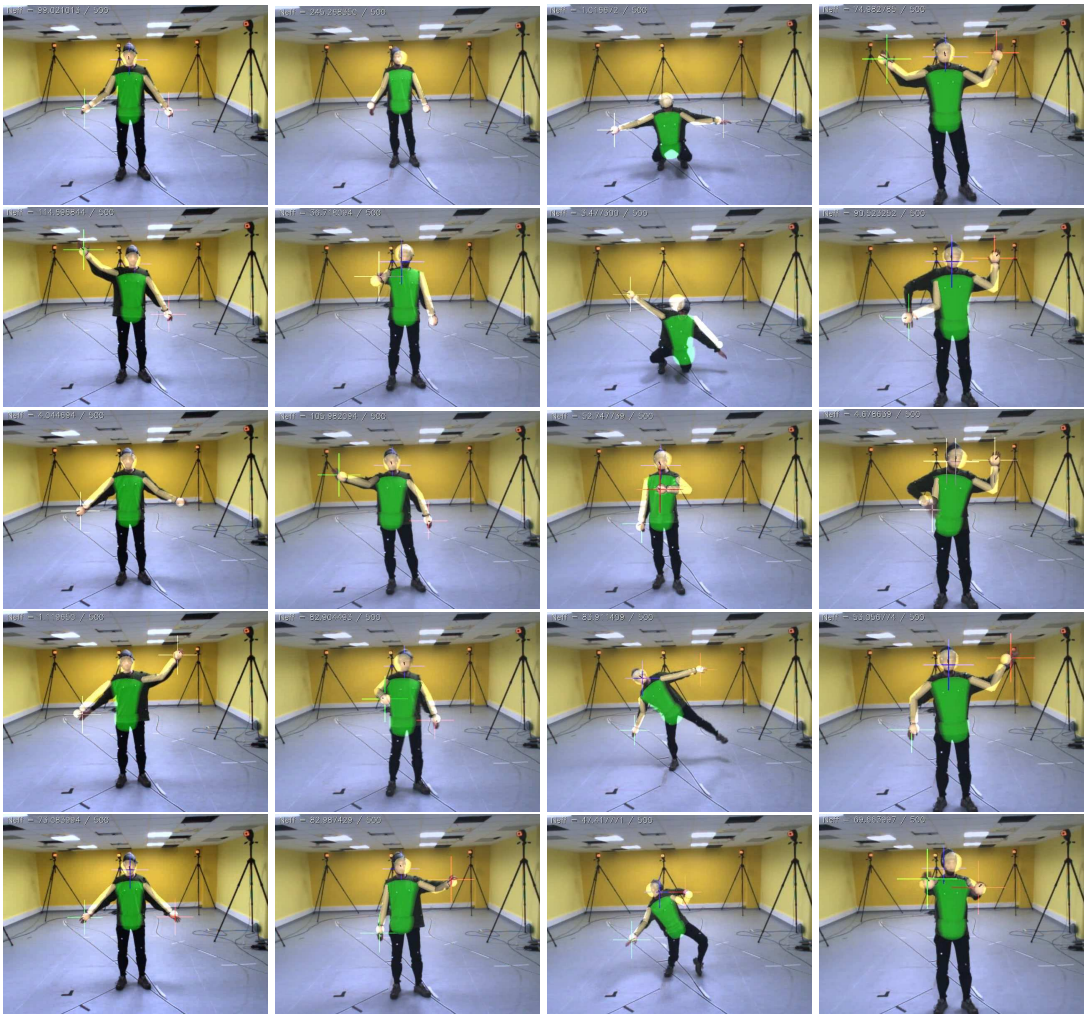
3 Contexte stéréoscopique

De la même manière que précédemment, nous évaluons mesures et stratégies dans un contexte stéréoscopique. Nous exploitons les séquences détaillées dans le tableau IV.3. Une réalisation du filtre APF sur chaque séquence est présentée en figure IV.11. Nous nous sommes restreints à des cas où l'utilisateur est majoritairement face à la caméra, puisque nous envisageons une application d'interaction homme-robot. Les séquences étudiées ne sont pas exactement identiques à celles du contexte multi-oculaire car l'expérimentation a nécessité des prises différentes. Précisons également que dans ce contexte d'interaction homme-robot, nous nous focalisons sur le suivi des membres supérieurs. Nous restreignons donc notre modèle à la moitié supérieure du corps humain. Le contexte stéréoscopique nous permet également d'évaluer les stratégies avec échantillonnage préférentiel.

3.1 Mesures hybrides et échantillonnage préférentiel

Les algorithmes exploitant l'échantillonnage préférentiel, *i.e.* I-CONDENSATION, IAPF et la stratégie HYBRID, reposent classiquement sur des indices visuels discriminants mais intermittents afin de guider l'exploration de l'espace d'état vers les zones présentant une forte vraisemblance. Dans notre cas, nous utilisons l'indice visuel 3D de triangulation de blobs de couleur peau. À partir de cette information intermittente, l'espace d'état est échantillonné autour de la configuration construite par cinématique inverse à partir de la position de la tête et des mains dans l'espace.

Les figures IV.12 montrent qu'une telle approche peut conférer au suivi une robustesse supplémentaire. Nous constatons que la stratégie classique CONDENSATION décroche en milieu de suivi (la main vient se positionner devant la tête). L'I-CONDENSATION parvient à éviter ce problème et à raccrocher la cible grâce à une réinitialisation automatique permise par l'échantillonnage préférentiel. De manière surprenante, ce comportement n'est cependant pas systématique, et la pratique montre qu'il est dépendant du choix de la dynamique du système. En effet, le problème de l'histo-



séquence 1 **séquence 2** **séquence 3** **séquence 4**

FIG. IV.11 – Déroulement de la stratégie APF sur les **séquences 1 à 4** (une sur chaque colonne) en contexte **stéréoscopique**. Les images sont issues de la caméra centrale. Nous exploitons les mesures $dist\_peau$, $dist$ et $blobs$.

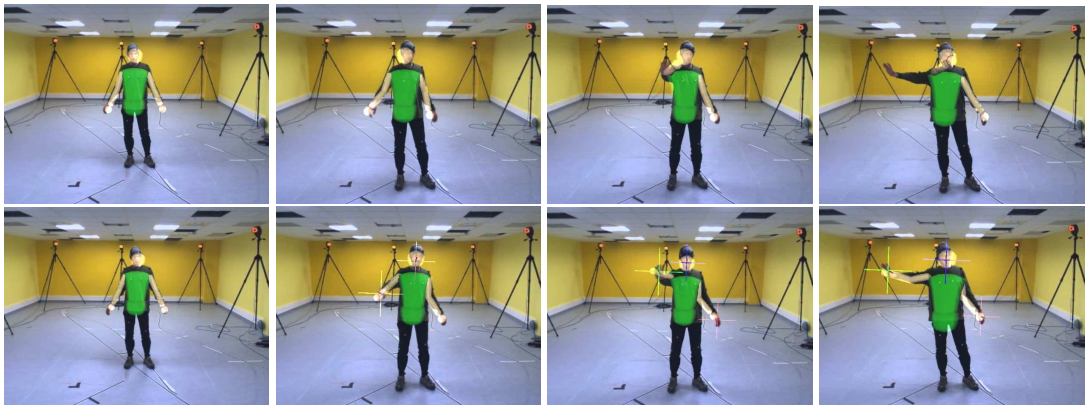


FIG. IV.12 – Déroulement des stratégies CONDENSATION (haut) et I-CONDENSATION (bas) pour 1000 particules avec les mesures $dist\_peau$ et $dist$.

rique des particules tirées suivant la mesure peut conduire à des poids écrasés par l'évaluation de la dynamique. Dans notre contexte où le système est de grande dimension, la dynamique ne peut pas être excessivement permissive sans quoi l'échantillonnage de l'espace n'est pas suffisamment fin pour produire un résultat satisfaisant. Ce phénomène est encore accentué par le faible nombre de particules utilisées. Ceci nous conduit à utiliser une dynamique relativement faible qui peut nuire à l'échantillonnage suivant la mesure.

Les mesures 3D sont par définition plus discriminantes (*i.e.* moins ambiguës) que les mesures 2D, *i.e.* basées apparence. Afin d'exploiter la force de notre mesure 3D au sein des stratégies sans échantillonnage préférentiel (CONDENSATION, PARTITIONNÉ, PARTITIONNÉ QRS, QRS, APF), nous l'introduisons dans la fonction de vraisemblance en complément des mesures 2D. En pratique nous observons que cette fonction de vraisemblance hybride induit un comportement similaire avec ou sans échantillonnage préférentiel. L'introduction de cette mesure 3D dans la fonction de vraisemblance devient alors une alternative lorsque l'échantillonnage préférentiel n'est pas souhaitable/envisageable (*i.e.* choix de la dynamique délicat). La figure IV.13 montre la propriété d'initialisation automatique, classiquement attribuée aux stratégies avec échantillonnage préférentiel, pour la stratégie CONDENSATION.

Nous constatons ainsi l'apport d'une mesure 3D en complément des indices visuels 2D. C'est pourquoi, afin de favoriser toutes les stratégies, nous choisissons d'intégrer la mesure 3D dans les fonctions de vraisemblance.

3.2 Choix des mesures et configuration optimale

De la même manière que précédemment, nous exploitons les types de mesures qui nous paraissent les plus adaptés. Les mesures sil et $sil2$ reposant sur une segmentation

FIG. IV.13 – Déroulement de la stratégie CONDENSATION exploitant les mesures $dist\_peau$, $dist$ et $blobs$ avec 1000 particules. L'initialisation est volontairement incorrecte. Le filtre parvient cependant à raccrocher la cible après quelques images.

de la silhouette sont inadaptées dans ce contexte car nous supposons le fond *a priori* inconnu. Nous choisissons donc de commencer par intégrer la mesure $dist\_peau$ (définie équation II.7) qui favorise la localisation de la tête et des mains, la caractérisation de ces membres étant vitale dans un contexte d'interaction homme-robot. La figure IV.14 présente les résultats obtenus pour différentes valeurs de σ_{dist_peau}. Afin de montrer l'apport d'une telle mesure de distance par rapport à la mesure de probabilité brute $peau$ (définie équation II.6), nous présentons en figure IV.15 l'évaluation de cette dernière. Nous constatons que le comportement de la mesure $peau$ est assez chaotique et la prise en compte d'une distance aux blobs de couleur peau améliore grandement le comportement du filtre quel que soit le critère étudié. Ceci s'explique par le fait que les zones de grande probabilité de couleur peau sont très petites dans l'image et ne favorisent que très peu de particules. La distance de similarité D_{dist_peau} est quant à elle naturellement plus lissée et permet de ne pas trop défavoriser les particules dont la projection de la tête et des mains ne se situe pas exactement dans sur les zones de teinte chair. Ceci plaide en faveur des mesures dont la courbure autour de l'*optimum* est plus douce.

De manière complémentaire, nous exploitons également la mesure de distance aux contours (cf équation II.4), utilisée sous différentes variantes dans la littérature [146, 166] et la distance aux blobs 3D dénotée $blob$. L'influence de cette dernière est présentée en figure IV.16. Nous constatons que son apport est évident puisqu'elle améliore les quatre critères d'étude, y compris la dispersion de l'estimateur. Dans notre choix, elle diminue le RMSE de $10\ cm$, la dispersion de $3\ cm$ et le biais de $5\ cm$. À l'instar des évaluations qualitatives de [8], il apparaît clairement que l'introduction d'une mesure 3D en complément des mesures 2D classiques améliore les performances du suivi.

Nous procédons de manière similaire pour les autres indices visuels, qui n'apportent pas d'information supplémentaire dans ce contexte. Les évaluations sur les filtres reposent donc sur les indices visuels $dist\_peau$, $dist$ et $blobs$ avec $\sigma_{dist\_peau} = 5, \sigma_{dist} = 5, \sigma_{blobs} = 0.07$.

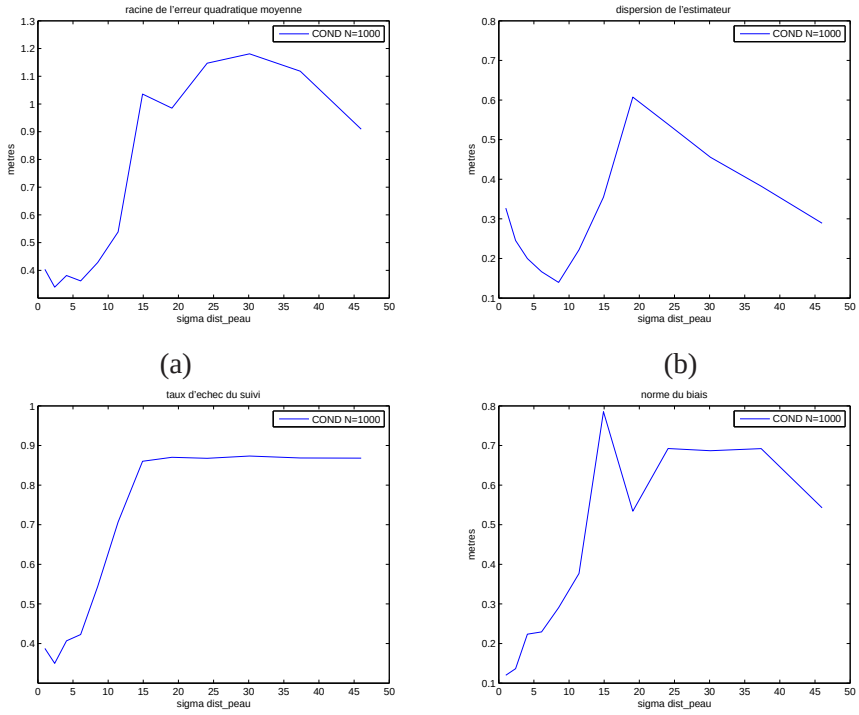
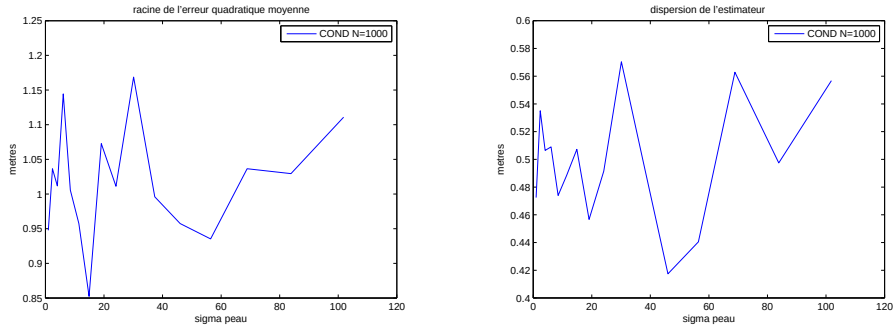
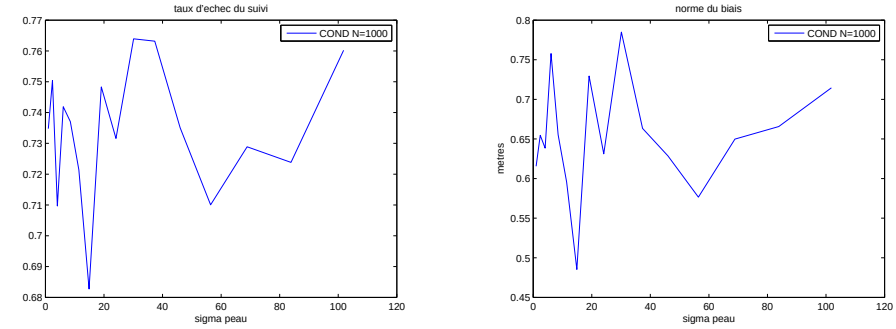


FIG. IV.14 – Étude de l'influence du paramètre σ_{dist_peau} en contexte **stéréoscopique** sur la **séquence 2** pour 20 réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais.



(a) (b)



(c) (d)

FIG. IV.15 – Étude de l'influence du paramètre σ_{peau} en contexte **stéréoscopique** sur la **séquence 2** pour 20 réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais.

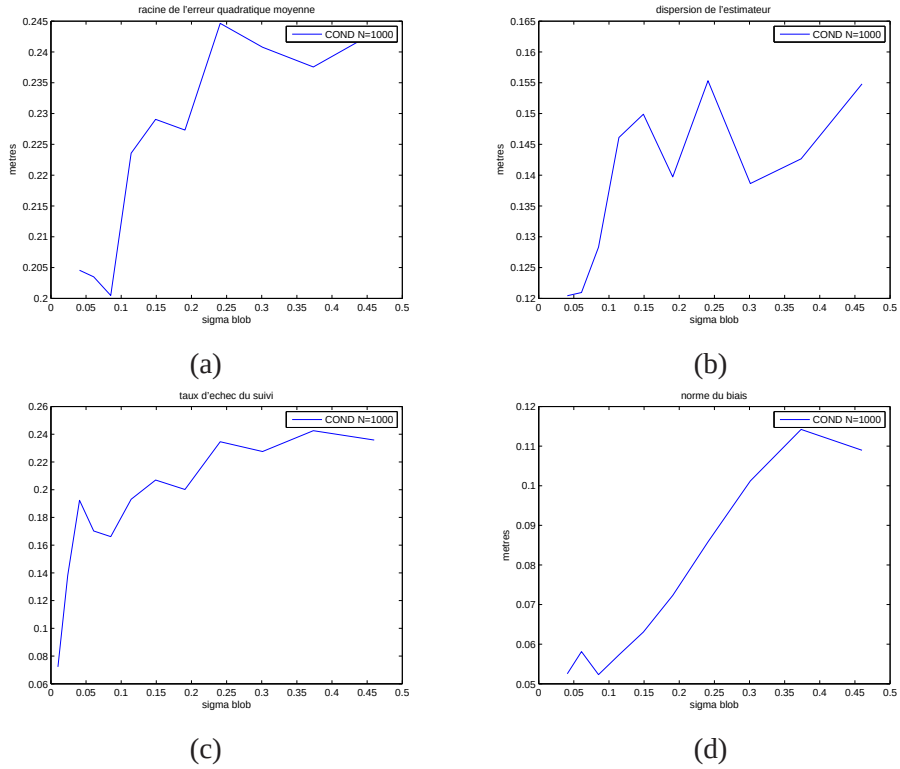


FIG. IV.16 – Étude de l'influence du paramètre σ_{blob} conjointement aux mesures $contours$ et $dist\_peau$ en contexte **stéréoscopique** sur la **séquence 2** pour 20 réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais.

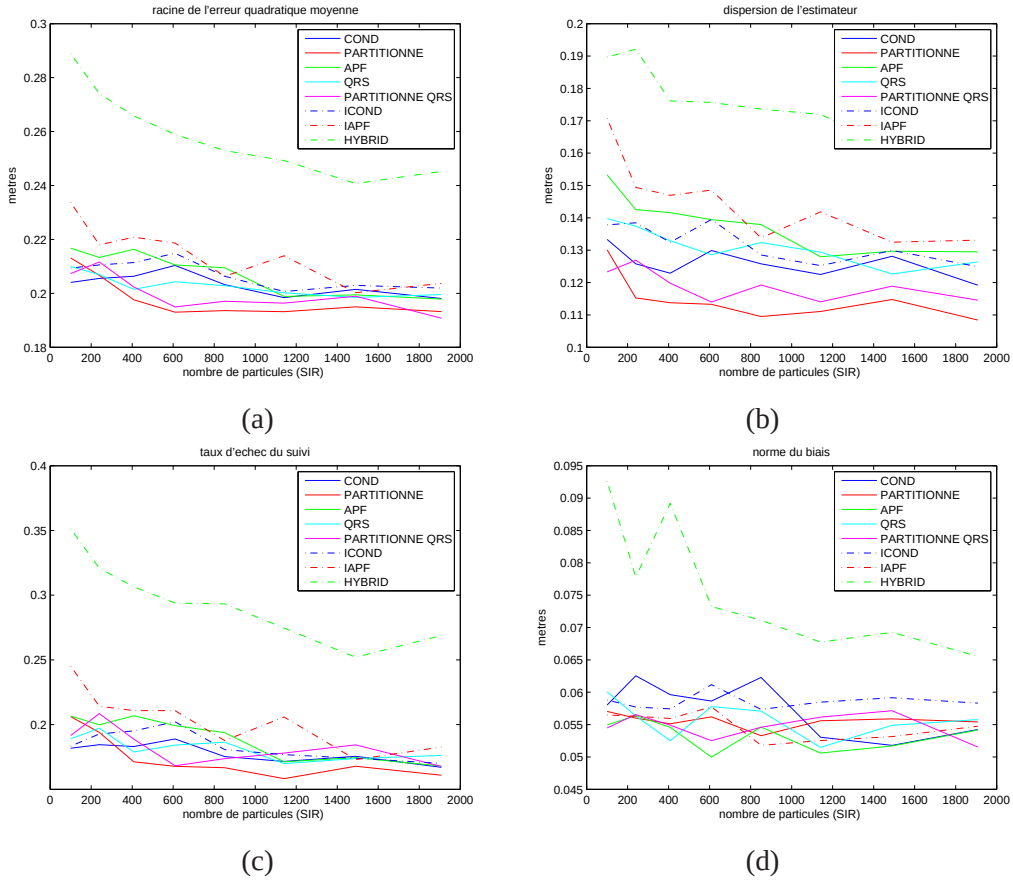


FIG. IV.17 – Influence du nombre de particules N et de la stratégie de filtrage en contexte **stéréoscopique** sur la **séquence 2** pour $R = 30$ réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais.

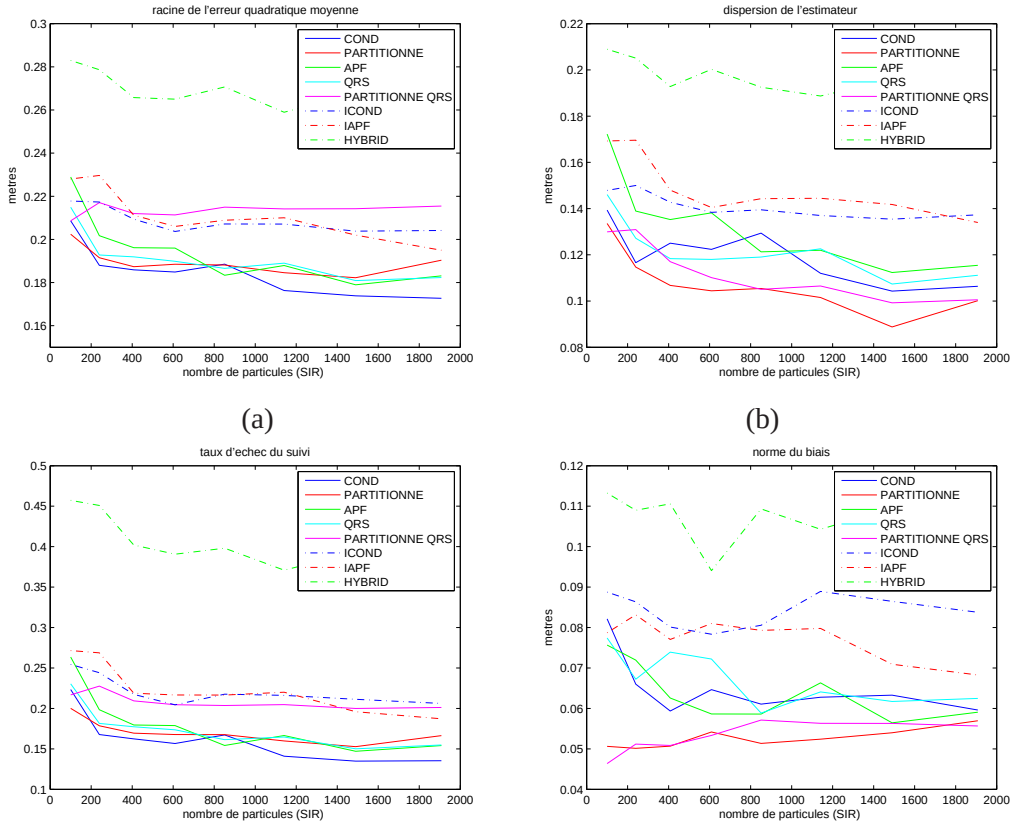


FIG. IV.18 – Influence du nombre de particules N et de la stratégie de filtrage en contexte **stéréoscopique** sur la **séquence 4** pour $R = 30$ réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais.

Erreur	X	Y	Z
Séquence 1	0.04	0.05	**0.10**
Séquence 2	0.07	0.08	**0.17**
Séquence 3	0.06	0.07	**0.15**

TAB. IV.4 – Racine de l'erreur quadratique moyenne par dimension pour $N = 1800$ toutes stratégies confondues.

3.3 Stratégie de filtrage et nombre de particules

Nous poursuivons la démarche d'évaluation par l'étude des filtres pour un nombre variable de particules. Nous présentons les résultats pour les séquences 2 et 4 respectivement sur les figures IV.17 et IV.18 (les autres courbes sont visibles en annexe sur les figures D.3 et D.4). Alors que le comportement des stratégies en contexte multi-oculaire est assez tranché, le contexte stéréoscopique affiche des résultats moins intuitifs car dépendants de la séquence. Sur certaines séquences simples (figure IV.17), on retrouve les tendances énoncées précédemment, mais sur des séquences plus complexes (figure IV.18), il est plus difficile de tirer des enseignements. La tendance qui désignait les techniques avancées comme étant plus efficaces s'inverse et la CONDENSATION parvient à de meilleurs résultats en termes de précision ou de dispersion de l'estimé. Balan *et al.* arrivent également à cette conclusion lorsque l'environnement est moins maîtrisé [10]. La stratégie classique parvient à de meilleurs résultats de par son aptitude à modéliser la multi-modalité de la distribution *a posteriori* due à nos mesures relativement peu informatives.

La précision est moindre dans ce contexte car un système stéréoscopique est moins informatif sur la profondeur de la scène qu'un système multi-oculaire. Ceci peut se vérifier sur le tableau IV.4 qui présente les racines des erreurs quadratiques moyennes associées à la localisation de chaque articulation. L'erreur suivant la profondeur (Z) — peu informative — prend alors le pas sur les erreurs sur les axes du plan image.

Dans la figure IV.19, nous présentons les mêmes évaluations mais limitées aux axes X et Y pour la séquence 4, représentative de l'ensemble des évaluations. Nous y retrouvons un comportement plus classique, bien que les différences ne soient pas aussi marquées que dans le contexte multi-oculaire. On peut en déduire que le bon comportement des filtres et leur comparaison relative n'a lieu d'être que pour une fonction de mesure suffisamment informative. Lorsque celle-ci est mal conditionnée (majoritairement suivant la profondeur ici), il faut être très prudent dans la comparaison des différentes stratégies, et manifestement, il est difficile de faire pencher la balance en faveur d'une stratégie plutôt qu'une autre. Nous pouvons noter la performance en demi-teinte de l'APF, probablement imputable à la multi-modalité de la densité filtrée dans ce contexte stéréoscopique. Ceci montre qu'il n'existe pas *a priori* de stratégie optimale quel que soit le contexte, mais que ce dernier joue un rôle important dans les choix à faire.

Une deuxième conclusion intéressante, et que l'on pouvait déjà soupçonner sur les

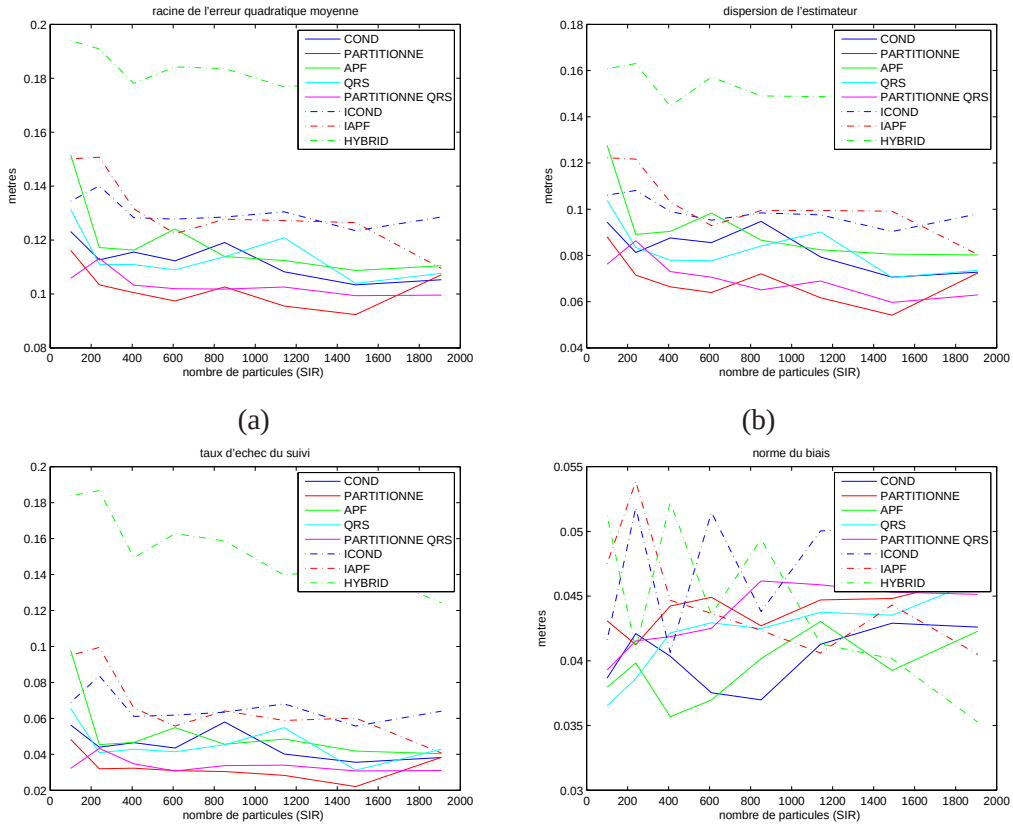


FIG. IV.19 – Influence du nombre de particules N et de la stratégie de filtrage en contexte **stéréoscopique** sur la **séquence 4** pour $R = 30$ réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais. Seules les dimensions X et Y sont considérées.

Nom	Erreur	Dispersion	Taux d'échec	Biais
CONDENSATION	5	4	5	5
QRS	3	3	3	4
PARTITIONNÉ	2	2	2	3
PARTITIONNÉ QRS	**1**	**1**	**1**	**2**
APF	4	5	4	1
I-CONDENSATION	**6**	**6**	**6**	**7**
IAPF	7	7	7	6
HYBRID	8	8	8	8

TAB. IV.5 – Classement des différentes stratégies par critères pour un contexte stéréoscopique. En gras sont précisés les meilleurs compromis sans et avec échantillonnage préférentiel.

figures IV.17 et IV.18, est que les filtres ne semblent pas améliorer leurs résultats au-delà de $N = 600$ particules. Parallèlement à ce constat, le biais ne semble pas descendre en dessous de $5\ cm$, ce qui constitue manifestement la limite de précision. Nous pouvons en déduire qu'il est inutile d'utiliser plus de 600 particules dans ce contexte. D'une manière plus générale, nous pouvons alors affirmer qu'il existe une limite en nombre de particules au-delà de laquelle les résultats ne seront pas améliorés pour un choix de mesures données. Dans le but d'intégrer ces algorithmes sur une plate-forme robotique où les capacités calculatoires sont très limitées, ce résultat est primordial.

Le tableau IV.5 récapitule les performances de chaque algorithme en ne prenant en compte que les dimensions X et Y lors du suivi.

3.4 Erreur par articulation

La figure IV.20 montre l'erreur de localisation sur chaque articulation lors d'une réalisation du suivi. Par opposition au contexte multi-oculaire, aucune articulation ne présente une erreur anormalement élevée par rapport aux autres. Les erreurs par articulation sont du même ordre de grandeur. La mise en œuvre de mesures dédiées à la localisation des mains et de la tête maintient ainsi un suivi globalement efficace.

De ces évaluations, nous pouvons tirer plusieurs règles générales pouvant guider un choix plus rapide des valeurs des $\sigma_{(.)}$, une étude exhaustive étant en pratique souvent irréalisable. Sur l'ensemble des expériences menées pour le choix des mesures, nous avons constaté que la dispersion est très souvent minimale pour la plus grande valeur de $\sigma_{(.)}$ possible. Un filtre fournit ainsi toujours un estimé plus lissé sans prise en compte de mesure supplémentaire. Il convient donc de limiter le nombre d'indices visuels exploités afin de ne pas arriver au seuil critique où $N_{eff} = 1$, voire, dans le pire des cas, à la divergence du filtre si tous les poids des particules sont nuls. Ceci est en partie dû au modèle erroné de la distribution des distances de similarités $D_{(.)}$ que nous adoptons

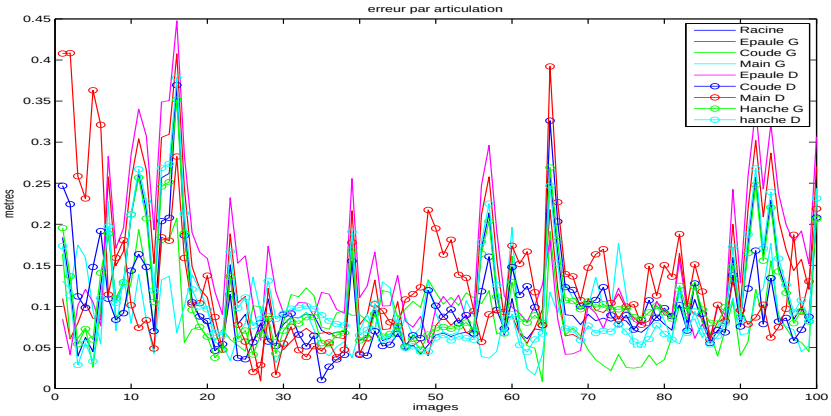


FIG. IV.20 – Erreur par articulation au cours d'une réalisation du suivi pour la **séquence 1**.

(modèle d'une distribution gaussienne centrée en 0).

D'un autre côté, afin d'obtenir un meilleur estimé en terme de distance à la vérité de terrain, il convient généralement d'adopter des valeurs de $\sigma_{(.)}$ plus petites pour permettre une convergence plus rapide. Ceci se fait généralement au détriment de la dispersion de l'estimé. De la même manière, le choix d'une valeur trop petite de $\sigma_{(.)}$ conduit à la divergence du filtre. À notre connaissance, très peu d'études exhaustives [99] ont été publiées dans la littérature, bien que le choix des mesures et leur configuration semble primordiale afin d'atteindre un comportement correct du filtre. Nous retrouvons ici le comportement que nous avons présenté de manière intuitive en fin de chapitre II. Le choix de la stratégie de filtrage semble parfois conditionné par le jeu de mesures choisi (APF) ainsi que le contexte d'application. Notons toutefois que les approches de type PARTITIONNÉ présentent globalement les résultats les plus satisfaisants.

4 Complexité et temps de calcul

À l'instar de nombreux travaux [10, 38, 166], l'ensemble des résultats présentés sur les graphiques impliquant plusieurs stratégies de filtrage sont « normalisés » par rapport au nombre d'évaluations de la fonction de vraisemblance, classiquement considérée comme le goulet d'étranglement des performances des filtres. Nous présentons dans les tableaux IV.6 et IV.7 les résultats en terme de performance pure de ces mêmes algorithmes non optimisés codés en C/C++. L'évaluation est faite sur un $Pentium\ M$ cadencé à $1.8\ GHz$.

Le choix des mesures influence fortement le temps de calcul final. En effet, après des expériences plus poussées, il s'avère que dans notre implémentation, l'évaluation des fonctions de vraisemblance pour l'algorithme CONDENSATION avec 1000 parti-

Filtre	$sil + dist\_peau$	$sil2 + dist\_peau$
CONDENSATION	1.72	0.64
QRS	1.70	0.64
APF	1.71	0.63
PARTITIONNÉ	**2.14**	**0.90**
PARTITIONNÉ QRS	2.13	0.89

TAB. IV.6 – Fréquence d'exécution en Hz des différentes méthodes proposées en contexte **multi-oculaire** pour $N = 1000$ particules et différentes mesures avec des images de 640×480 pixels.

Filtre	FPS (640×480)	FPS (320×240)
CONDENSATION	4.03	7.83
QRS	4.07	7.93
APF	4.06	7.80
PARTITIONNÉ	**4.24**	**8.56**
PARTITIONNÉ QRS	4.24	8.51
I-CONDENSATION	4.00	7.81
IAPF	4.03	7.83
HYBRID	3.98	7.80

TAB. IV.7 – Fréquence d'exécution en Hz des différentes méthodes proposées en contexte **stéréoscopique** pour $N = 600$ particules en exploitant les mesures $dist\_peau$, $blobs$ et $dist$.

cules représente environ $70\ \%$ du temps d'exécution total, contre $20\ \%$ pour l'acquisition et le prétraitement des images, $8\ \%$ pour la projection du modèle dans les images et $2\ \%$ pour l'algorithme de filtrage lui-même. Ainsi, la mesure de silhouette duale $sil2$, qui s'avère plus longue en temps de calcul que la mesure de silhouette sil, a un impact important sur la fréquence d'exécution des filtres. Nous constatons que pour un même nombre d'évaluations de la fonction de vraisemblance, le choix de la stratégie influence peu le temps de calcul nécessaire, excepté pour le PARTITIONNÉ. En effet, une partie des fonctions de vraisemblance utilisées ne concerne que la première partition exploitée, *i.e.* le buste dans notre le cas. Les mesures sont donc évaluées plus rapidement et confèrent un avantage intrinsèque en terme de temps de calcul aux stratégies de type PARTITIONNÉ. Les stratégies QRS demandent classiquement plus de ressources que les techniques de Monte-Carlo classiques pour la génération de séquences quasi-aléatoires, mais cela est à peine perceptible dans ces résultats.

La littérature présente assez peu de temps de calcul pour comparatif. Dans le contexte multi-oculaire, les systèmes affichent des performances relativement lentes pour traiter une image : $45\ s$ pour [10, 68], $15\ s$ pour [39], $1\ s$ pour [91, 179]. Nous nous situons donc au niveau des approches proposant les temps de calcul les plus compatibles avec

des contraintes temps-réel.

En contexte stéréoscopique, la littérature affiche des fréquences de traitement bien plus importantes : $15\ Hz$ pour [8], $20\ Hz$ pour [26], $12\ Hz$ pour [94]. Notons toutefois que toutes ces approches traitent des images de taille inférieure ou égale à 320×240 pixels. Dans de telles conditions, notre algorithme s'exécute à $8.56\ Hz$ dans un cas favorable.

5 Conclusion

Alors que dans le cas idéal le comportement des filtres est relativement prévisible et cohérent, le contexte réel apporte son lot de difficultés supplémentaires, et l'analyse des performances est parfois bien plus complexe.

Le choix des indices visuels exploités et leur fusion doivent être soigneusement étudiés. Leur influence sur le comportement des stratégies de filtrage est capitale et le résultat du suivi semble être principalement conditionné par les mesures plus que par le choix de la stratégie elle-même. Ceci n'est pas ou peu abordée dans la littérature. Nous proposons un protocole permettant un couplage optimal des différentes mesures envisagées, et une configuration des paramètres $\sigma_{(.)}$ prenant en compte le compromis nécessaire entre erreur moyenne et dispersion des estimés fournis par le filtre. Cette méthode montre à quel point un filtre dont les fonctions de vraisemblance sont mal construites peut résulter en des performances catastrophiques. Nous constatons également qu'il ne suffit pas de mettre en œuvre un grand nombre d'indices visuels différents pour améliorer le résultat, et que leur sélection repose en grande partie sur le contexte applicatif envisagé. En outre, il convient d'exploiter des distances de similarité $D_{(.)}$ peu piquées et adaptées au suivi des membres.

La comparaison exhaustive des différentes stratégies de filtrage ne peut se faire que dans un cadre où ces indices visuels ont été soigneusement choisis. Ainsi, nous observons que dans un contexte multi-oculaire, les stratégies avancées surpassent la stratégie classique à tout point de vue. L'APF et le PARTITIONNÉ QRS semblent tout particulièrement se détacher du lot. Les propriétés des stratégies QMC en font également un choix intéressant assez peu étudié dans la littérature. Dans un contexte stéréoscopique où les indices visuels exploitables sont plus limités et moins informatifs, les stratégies classiques opèrent tout aussi bien que les autres, voire mieux. Ce comportement est explicable par le profil particulièrement complexe de la densité filtrée. Nous constatons alors que pour un jeu d'indices visuels donnés, le suivi n'est plus amélioré significativement au delà d'un certain nombre de particules. Le choix des mesures limite donc la précision atteignable par un filtre. Ce résultat est particulièrement important dans les contextes où les ressources calculatoires sont limitées.

De manière surprenante, nous constatons également que le comportement des stratégies à échantillonnage préférentiel n'affichent pas une meilleure performance que leurs homologues classiques, malgré la prise en compte d'une mesure 3D dans le placement des particules. Toutefois, nous avons établi que la propriété de ré-initialisation automatique est transposables aux stratégies classiques par la prise en compte de cette

mesure dans la fonction de vraisemblance, ce qui constitue une alternative aux stratégies avec échantillonnage préférentiel.

Ces différents enseignements vont nous guider dans l'intégration sur système réel. Les règles que nous avons tirées de ces expériences permettent de choisir de manière réfléchie une association entre un ensemble de mesures et une stratégie de filtrage pour un contexte donné.

Chapitre V

Intégration sur systèmes multi-caméras pour l'interaction homme-robot

Dans ce dernier chapitre, nous abordons le problème de l'interaction homme-robot par suivi visuel. Après une rapide présentation des problèmes et avancées récentes dans ce domaine, nous proposons un scénario envisagé dans le cadre d'une interaction homme-robot. Nous présentons quelques résultats qualitatifs pour nos deux systèmes (multi-caméras déportées et embarquées). Ce chapitre se conclut par des considérations sur l'intégration en cours sur la plate-forme robotique JIDO, puis par quelques perspectives.

1 Capture de mouvement par vision et interaction homme-robot

La capture de mouvement humain depuis des caméras embarquées ou déportées dans l'environnement constituent deux enjeux majeurs dans la communauté Robotique.

L'instrumentation de l'environnement par des caméras déportées autorise une perception plus globale de la scène et donc une interprétation à large échelle des activités de l'homme qui partage l'environnement avec le robot. Cette démarche s'inscrit dans une problématique émergente dans la communauté Robotique : la robotique ubiquiste. Au delà des capteurs visuels, cette problématique (scientifique et technologique) nous amène à considérer des réseaux de capteurs pouvant instrumenter une multitude de supports de l'environnement afin de percevoir les agents humains et robotisés. Dans ce contexte, les systèmes multi-caméras (très souvent synchrones) pour la capture de mouvement humain sont pléthores dans la communauté Vision [114]. Néanmoins, ces systèmes restent encore perfectibles et ne respectent pas toujours les contraintes robotiques (environnement encombré et évolutif, temps réel applicatif, ...). Pour étayer ces propos, le lecteur pourra se référer au chapitre I qui énumère quelques systèmes

existants dans ce contexte applicatif.

L'interaction physique homme-robot (donc proximale) depuis des caméras embarquées est également une problématique centrale dans la perspective de construire des robots compagnons, voire cognitifs. Davis [36] a démontré que la communication homme-homme repose à 65% sur les mouvements corporels. Les mouvements coordonnés homme-robot (donc le positionnement du robot à distance sociale [79]), l'exécution de tâches conjointes entre ces deux agents requièrent clairement la capture du mouvement humain à chaque instant. Enfin, la capture du mouvement est une étape intermédiaire vers l'interprétation de gestes ou de postures de l'homme depuis le robot dans son voisinage immédiat. Dans ce contexte, et comme évoqué au chapitre I, la capture, et *a fortiori* l'interprétation, de mouvement humain depuis un système robotisé autonome est une problématique ouverte dans la communauté Robotique. Certes, de nombreuses plates-formes robotiques intègrent des capacités de détection et suivi de l'homme mais limitées à une analyse fruste : (i) dans le plan du sol par signaux télémétriques laser (robots Maggie [65], Pearl [131], Biron [100], Minerva [156]) voire radio-fréquence [90], (ii) dans le plan image par vision monoculaire (robots Alpha [11], PeopleBot [26], Biron [100], RoboX [141]). Mentionnons ici nos travaux antérieurs sur le suivi 2D de personnes par le robot guide Rackham [18].

La capture de mouvement (3D) depuis un robot mobile reste marginale ; rappelons ici les travaux d'Azad *et al.* [8], prévus pour une plate-forme humanoïde mais avec de fortes restrictions sur l'apparence de l'homme, et ceux de Knoop *et al.* [94] basés sur une caméra active 3D et intégrés sur notre plate-forme JIDO. Classiquement, la stratégie ascendante visant à séquencer analyse spatio-temporelle puis interprétation aboutit logiquement à peu de travaux intégrés sur l'interprétation de gestes ou postures. Citons ici l'interprétation 2D (donc fronto-parallèle) de gestes statiques ou dynamiques puis l'intégration sur plates-formes robotiques dans [159, 165]. À notre connaissance, peu de travaux sur l'interprétation 3D de gestes ou postures ont abouti à une intégration robotique et à des évaluations dans des contextes variés : à l'heure actuelle, les systèmes Arma [154] et Horos [136] sont dédiés aux gestes déictiques et T-Rot [176] est consacré à la reconnaissance de postures. À ce titre, des travaux sur l'interprétation de commandes multimodales sont menés par ailleurs dans le groupe par B. Burger [19].

Fort de ces constats, le challenge visant à l'intégration de notre système de capture de mouvement humain, à terme de son interprétation, sur une plate-forme mobile du laboratoire prend alors pleinement son sens.

Les travaux présentés dans ce chapitre visent conjointement à intégrer nos algorithmes de capture de mouvement sur un système multi-caméras instrumentant l'environnement et un système stéréo embarqué sur notre robot assistant JIDO. La finalité est d'élaborer et dérouler, grâce aux deux systèmes perceptuels, un scénario réaliste mettant en jeu JIDO et un humain dans un environnement humain naturel type hall de laboratoire. La caractérisation, la paramétrisation des deux systèmes de HMC dédiés (systèmes multi-oculaires *vs.* stéréo) s'appuie logiquement sur l'expertise acquise durant les évaluations quantitatives précédentes (cf. chapitre IV). Les évaluations sur des séquences réalistes et variées acquises depuis les deux systèmes mais traitées hors-ligne

restent ici qualitatives car le déploiement de systèmes commerciaux de HMC est incompatible avec des environnements encombrés tels qu'un hall public (réflexions diverses, artefacts, ...). Suppléer à terme ces systèmes par des systèmes de vision standards est nécessaire dans un tel contexte et corrobore les nombreux travaux menés dans la littérature sur cette problématique. Dans les sections suivantes, nous exposons un scénario d'application, puis détaillons la caractérisation des deux systèmes HMC et les évaluations associées.

2 Scénario robotique envisagé

2.1 Étapes-clés du scénario

Nos fonctions de suivi tridimensionnel de posture sont testées dans un scénario réaliste d'interaction homme-robot. Celui-ci recouvre les deux contextes considérés dans ce manuscrit (caméras déportées et embarquées), et se déroule dans un environnement typique de la robotique, par essence non contrôlé, dynamique et évolutif. Les deux systèmes exploités sont similaires à ceux utilisés dans le chapitre précédent. Les participants sont le robot manipulateur mobile JIDO (cf. figure V.1) ainsi que son interlocuteur humain faisant l'objet de la capture de mouvement. D'autres usagers peuvent être présents dans l'environnement. Cependant, leur interaction avec le robot étant passive, ils sont seulement considérés comme des artefacts pouvant perturber la capture de mouvement du sujet d'intérêt.

Le scénario peut être scindé en les trois étapes consécutives suivantes.

1. Au départ, le système de vision trinoculaire (supposé synchrone) détecte et capture les mouvements d'un humain évoluant dans l'espace surveillé. La personne s'avance, dans le but d'interpeller le robot. Le suivi tridimensionnel de ses gestes est réalisé durant l'ensemble de son déplacement.
2. Par un geste distinctif (*e.g.* un balancement d'un/des bras au-dessus de la tête), voire une posture particulière, l'homme signifie au robot son souhait d'interagir avec lui. La reconnaissance de ce geste de commande sur la base de sa capture par le réseau de caméras déclenche un mouvement d'approche du robot.
3. Le robot se positionne en face du tuteur puis stoppe son mouvement. La fonction de capture de mouvement humain sur la base du capteur stéréoscopique embarqué est alors lancée. Le but est de capturer le mouvement humain durant l'exécution d'une tâche conjointe (*e.g.* la manipulation d'un objet), ou, à plus haut niveau, d'interpréter des gestes ou attitudes de l'homme.

Les exécutions de ce scénario élémentaire peuvent présenter des variantes. Du fait que les fonctions de suivi ne requièrent aucun apprentissage, quiconque peut jouer le rôle du tuteur. L'environnement peut être plus ou moins encombré, et les conditions d'éclairage de la scène sont *a priori* quelconques[1]. Les gestes de commande effectués par le tuteur dans la dernière partie du scénario peuvent être de complexité variable.

[1]Notons que ces conditions sont moins favorables que celles considérées jusqu'ici pour la capture de mouvement sur la base de systèmes multi-oculaires.

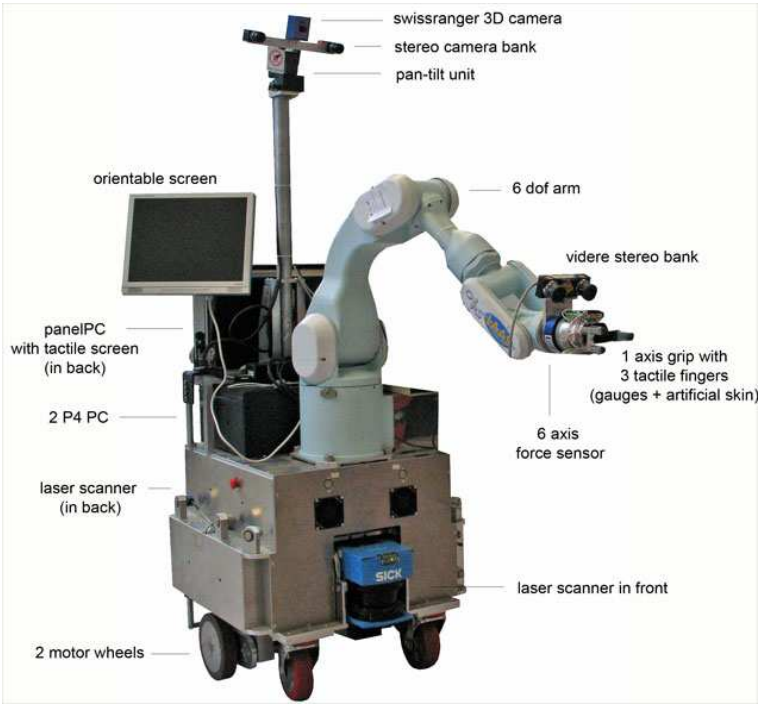


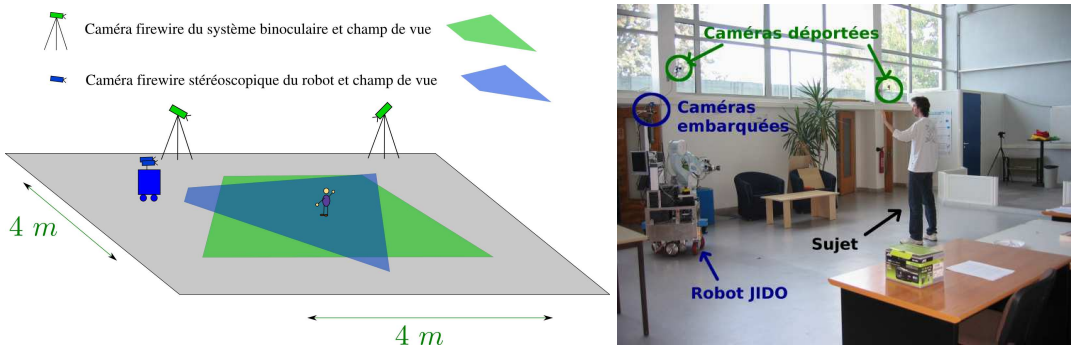
FIG. V.1 – Le robot JIDO.

2.2 Description de la plate-forme matérielle JIDO

Le robot JIDO est présenté en figure V.1. Il est constitué d'une base mobile non-holonome Neobotix MP-L655 et d'un bras manipulateur embarqué à 6 DDL Mitsubishi PA-10. Conçu pour l'interaction homme-robot, JIDO a été l'un des supports expérimentaux du projet européen intégré FP6-IP-COGNIRON (« The Cognitive Robot Companion », `www.cogniron.org`). Il embarque une paire de caméras stéréo montée sur une platine pan-tilt en haut d'un mât, une seconde paire de caméras stéréo fixée en bout de bras pour la manipulation référencée vision, deux capteurs laser rapides SICK, un panelPC proposant un écran tactile, ainsi que plusieurs écrans affichant des informations à l'attention de l'utilisateur. La paire stéréo que nous exploitons est celle située en haut du mât à $1.8\ m$. Elle présente une base de $27\ cm$.

2.3 Description du système multi-oculaire

À l'heure actuelle, notre plate-forme expérimentale ne comporte que deux caméras firewire IEEE1394b Flea 2 Color montées sur des fixations à $2,5\ m$ de hauteur afin de couvrir un large champ de vue. Le suivi est toutefois possible, même si les conditions sont sensiblement dégradées par rapport aux évaluations menées dans le cadre du chapitre IV. La figure V.2 schématise la situation et les différents acteurs impliqués.



(a) (b)

FIG. V.2 – (a) Localisation des différents capteurs et acteurs dans le hall d'étude ; (b) Situation réelle.

La section suivante présente plusieurs évaluations menées dans chacun des contextes multi-oculaire et stéréoscopique (embarqué) de manière indépendante.

3 Caractérisation et validation du suivi sur des séquences types

3.1 Choix des mesures et paramétrisation

Notre étude s'appuie logiquement sur les indices visuels décrits au chapitre II ainsi que sur les évaluations menées au chapitre IV. En contexte multi-oculaire, nous avons privilégié la segmentation de la silhouette sil, qui s'avère simple à mettre en œuvre, ainsi que la distance aux blobs de couleur peau $dist\_peau$ qui permet une meilleure localisation des mains. Nous utilisons occasionnellement la distance $sil2$ qui est relativement lourde à traiter en terme de temps de calcul, mais qui peut s'avérer utile pour certains mouvement complexes (marche notamment). De par les heuristiques décrites en chapitre II, et les évaluations du chapitre IV nous choisissons $\sigma_{sil} = 30$ et $\sigma_{dist\_peau} = 7$ et $\sigma_{sil2} = 0.1$.

En contexte stéréoscopique, nous exploitons la distance 3D aux blobs de couleur peau pour les contraintes fortes qu'elle apporte, ainsi que la distance 2D $dist\_peau$, apportant une information importante lorsque la triangulation n'est pas possible (impossibilité d'appariement, erreur de triangulation trop grande, ...). En complément nous utilisons la distance aux contours $dist$. Nous privilégions ici les indices visuels présentant des distances de similarité dont la courbure autour de l'optimum est relativement douce. En outre, l'extraction des contours et des régions de teinte chair est restreinte aux zones en mouvement. Ceci permet de s'affranchir en partie des arrière-plans encombrés et présentant des objets de couleur peau, problème que nous n'avions pas dans le cadre des évaluations du précédent chapitre. L'exploitation de ces mesures nous a menés à choisir $\sigma_{dist\_peau} = 7$, $\sigma_{blobs} = 0.08$ et $\sigma_{dist} = 5$.

3.2 Contexte multi-oculaire

Notre plate-forme expérimentale ne propose à ce jour que deux caméras de résolution 640×480. Nous utilisons un filtre APF avec 500 particules par couche. Nous présentons en figure V.3 un suivi à partir du système multi-oculaire sur une séquence de marche. Chaque paire d'images représente la vue courante de chacune des caméras. Bien que le suivi soit opérationnel, nous constatons qu'il est moins fonctionnel qu'avec trois caméras. En effet, sur plusieurs réalisations successives du suivi, il arrive que le système « décroche » partiellement notamment au niveau des pieds. La segmentation des jambes est de moins bonne qualité que celle du haut du corps du fait de l'ombre portée du sujet sur le sol. En outre, nous ne disposons pas de mesures spécifiques à la localisation des pieds. La multi-modalité potentielle de la fonction de vraisemblance qui en résulte perturbe sensiblement le suivi. Ces observations rejoignent celles de [10] et nos évaluations précédentes qui privilégient un système de vision trinoculaire. C'est pourquoi nous limitons par la suite le suivi au haut du corps. Le suivi du corps complet sera quant à lui subordonné à l'ajout d'une troisième caméra, difficile à positionner en pratique dans notre hall robotique.

La figure V.4 présente une séquence classique de mouvement de bras avec un léger déplacement du sujet. Le suivi est correct et l'estimé présente une dispersion moindre d'une exécution sur l'autre par rapport à la séquence précédente. En outre, sur une même séquence, le tressautement du modèle recalé est moins important. Ce mouvement plus lisse et plus constant est du à la réduction de l'espace de recherche (nous utilisons 14 DDL contre 22 précédemment). Actuellement, la vitesse de traitement du système est comprise entre 1 et 2 Hz.

3.3 Contexte stéréoscopique

Forts des évaluations précédemment menées, nous choisissons d'exploiter une stratégie de type PARTITIONNÉ QRS avec 500 particules par partition, pour des raisons de contraintes temps-réel. Nous avons évalué qualitativement le système sur plusieurs séquences de 320×240 pixels présentant des conditions variées (arrière-plan encombré, dégagé, différents vêtements, ...). La figure V.5 illustre une situation type d'interaction homme-robot dans un environnement humain (vue du robot et situation homme-robot filmée par une caméra extérieure au système). Seule une vue de la tête stéréoscopique est présentée. Nous constatons que l'initialisation se fait automatiquement quelques secondes après l'apparition du sujet dans le champ de vue. Le suivi est globalement satisfaisant, bien que la précision soit limitée par les mesures mises en place. Nous pouvons cependant remarquer sur la première moitié de la séquence que le suivi du bras droit est perturbé par la table de couleur chair située au second plan. Dans la deuxième moitié de la séquence, le sujet se rapproche de la caméra pour tendre la main vers le robot et la table est cachée. Le suivi est alors plus efficace. Enfin, lorsque le sujet sort (rapidement) du champ de vue, la dynamique n'est pas suffisante pour le suivre correctement, et le filtre décroche.

La figure V.6 présente un contexte plus encombré que le précédent. Le sujet rentre

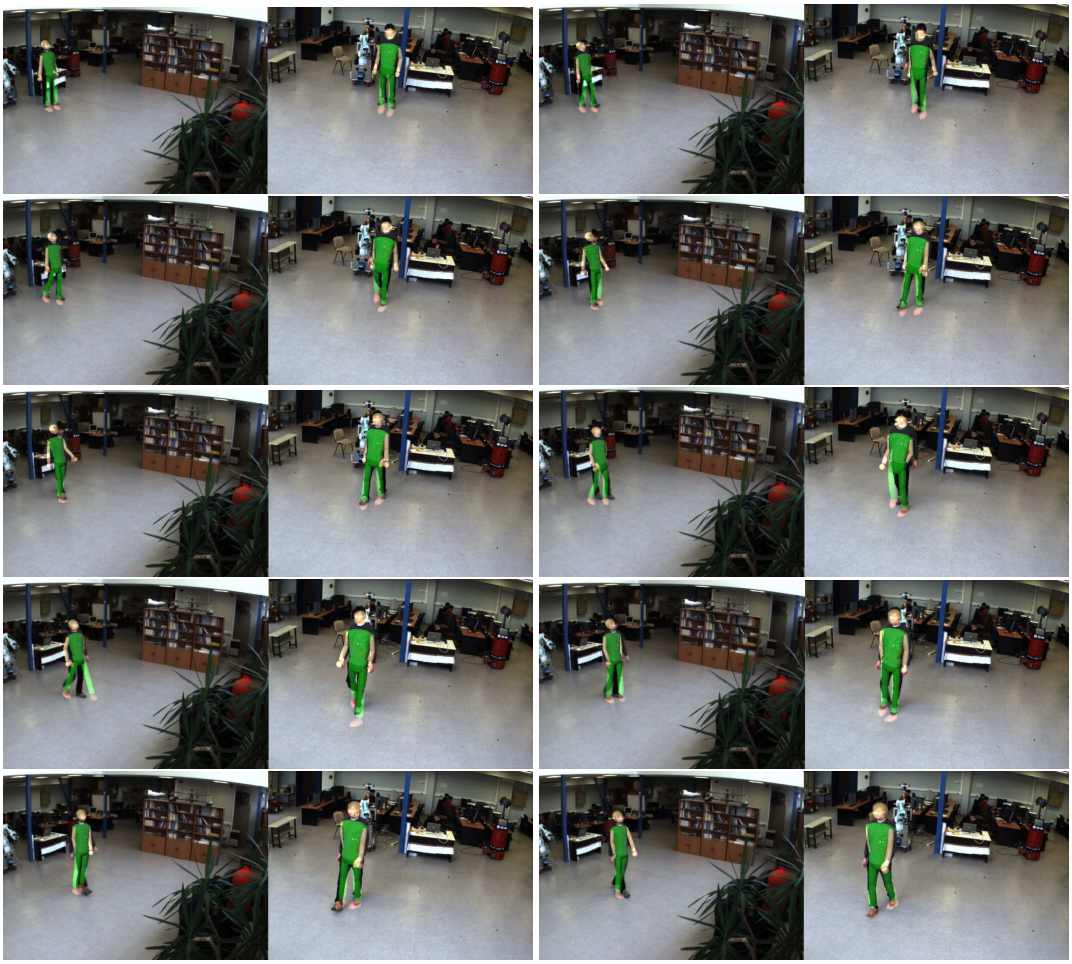


FIG. V.3 – Suivi par vision binoculaire déportée sur une séquence de marche.

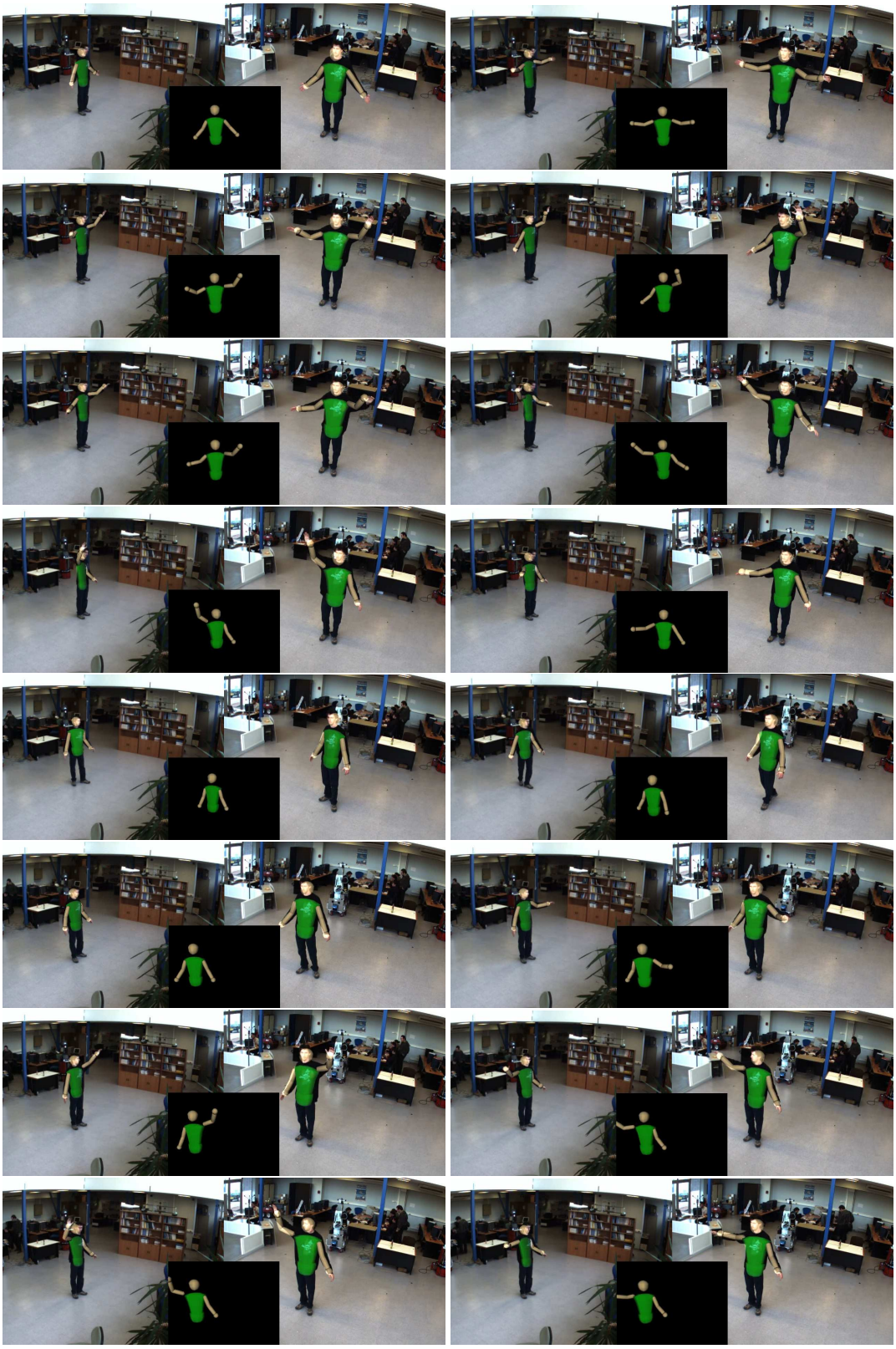


FIG. V.4 – Suivi par vision binoculaire déportée sur une séquence de mouvements simples.

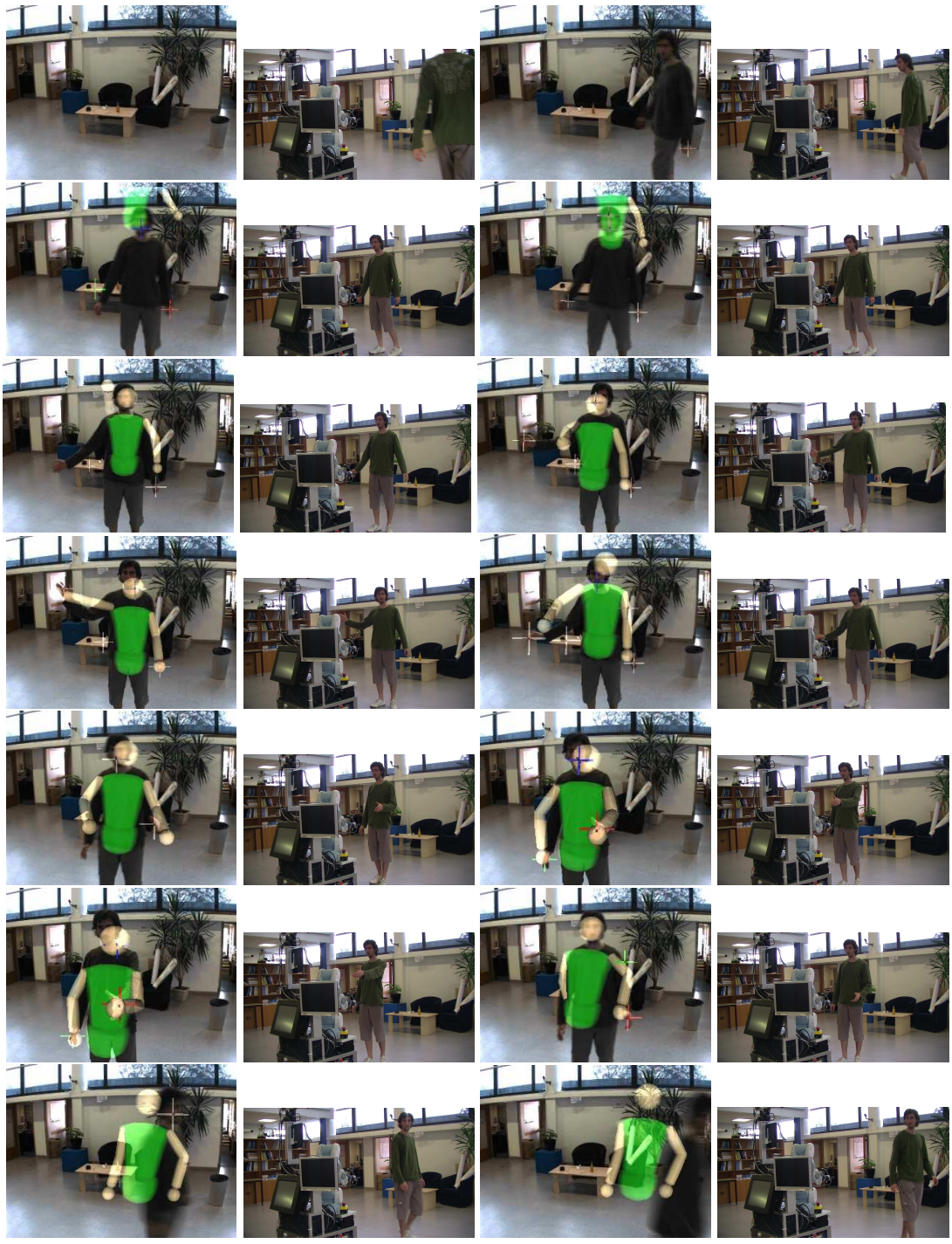


FIG. V.5 – Suivi par vision stéréoscopique sur une séquence avec fond quelconque. Les images issues de la caméra du robot avec la reprojection de l'avatar sont présentées en alternance avec la situation homme-robot.

et sort plusieurs fois du champ de vue. Il effectue des mouvements non fronto-parallèles puis présente un objet au robot. D'une manière générale le système parvient à suivre les mouvements. Certains gestes complexes sont parfois sujets à décrochage temporaire, mais la localisation simultanée de la tête et des mains constitue une mesure suffisamment discriminante pour lui permettre de se réinitialiser. Son comportement global est là encore satisfaisant. Le suivi dans un contexte aussi encombré et présentant de nombreux objets de teinte chair autour du sujet est rendu possible par l'exploitation du mouvement dans les images. En effet, les images sont pondérées par un masque de mouvement, selon la méthode présentée chapitre II section 3.4. En terme de précision du suivi toutefois, nous constatons que le recalage du modèle n'est pas toujours parfait. La localisation des coudes notamment reste à améliorer. En terme de dispersion, l'estimé fourni par les filtres a tendance à tressauter d'une image à l'autre. Ceci est la contrepartie du faible nombre de particules que nous utilisons et reste cohérent avec les évaluations quantitatives menées dans le chapitre IV. De manière qualitative, nous observons que le suivi est répétable d'une exécution sur l'autre. Il arrive toutefois que le suivi du bras décroche temporairement (25% des cas) sur cette séquence comme sur la précédente, mais la localisation 3D des blobs peau permet une ré-initialisation régulière du système.

Nous présentons en figure V.7 et V.8 le suivi de sujets différents. La première séquence montre une personne en train de lire. Bien que le modèle que nous adoptons présente des dimensions fixes, le système parvient à suivre un sujet à la morphologie différente. Les approximations du modèle sont telles qu'elles peuvent être compensées par une estimation sensiblement différente sur la profondeur du sujet. L'activité est suivie par le système de manière satisfaisante en exploitant l'initialisation automatique, et ce en dépit des objets occultants. La deuxième séquence présente un suivi de geste simple mais qui est perturbé par le passage d'une personne en arrière-plan. Le système parvient à s'initialiser et à suivre le premier sujet. Le passage influence légèrement le comportement du filtre, mais l'estimé reste satisfaisant. Notons toutefois ici que la présence de la deuxième personne reste passagère et que celle-ci ne souhaite pas interagir avec le robot. Dans le cas où deux personnes se dirigent simultanément vers la plateforme mobile, le système est trop perturbé et le suivi n'est plus satisfaisant. Les deux séquences présentées ici offrent une répétabilité du suivi correcte. En outre, les mouvements relativement lents réalisés permettent une meilleure localisation des coudes grâce à l'utilisation du mouvement. Le temps de traitement du système oscille actuellement entre $5\ Hz$ et $8\ Hz$.

D'une manière générale, nous constatons que le succès du suivi est grandement conditionné par le choix et la paramétrisation correcte des mesures. Dans notre cas, la localisation des blobs de couleur peau est primordiale. Dès lors que certains artefacts sont présents (figure V.5), les performances sont significativement dégradées. Certes ces évaluations qualitatives montrent que le jeu d'indices visuels proposé est relativement robuste à la variabilité des environnements et des sujets observés, néanmoins il serait pertinent de pousser nos investigations sur les meilleures associations en termes d'indices visuels.

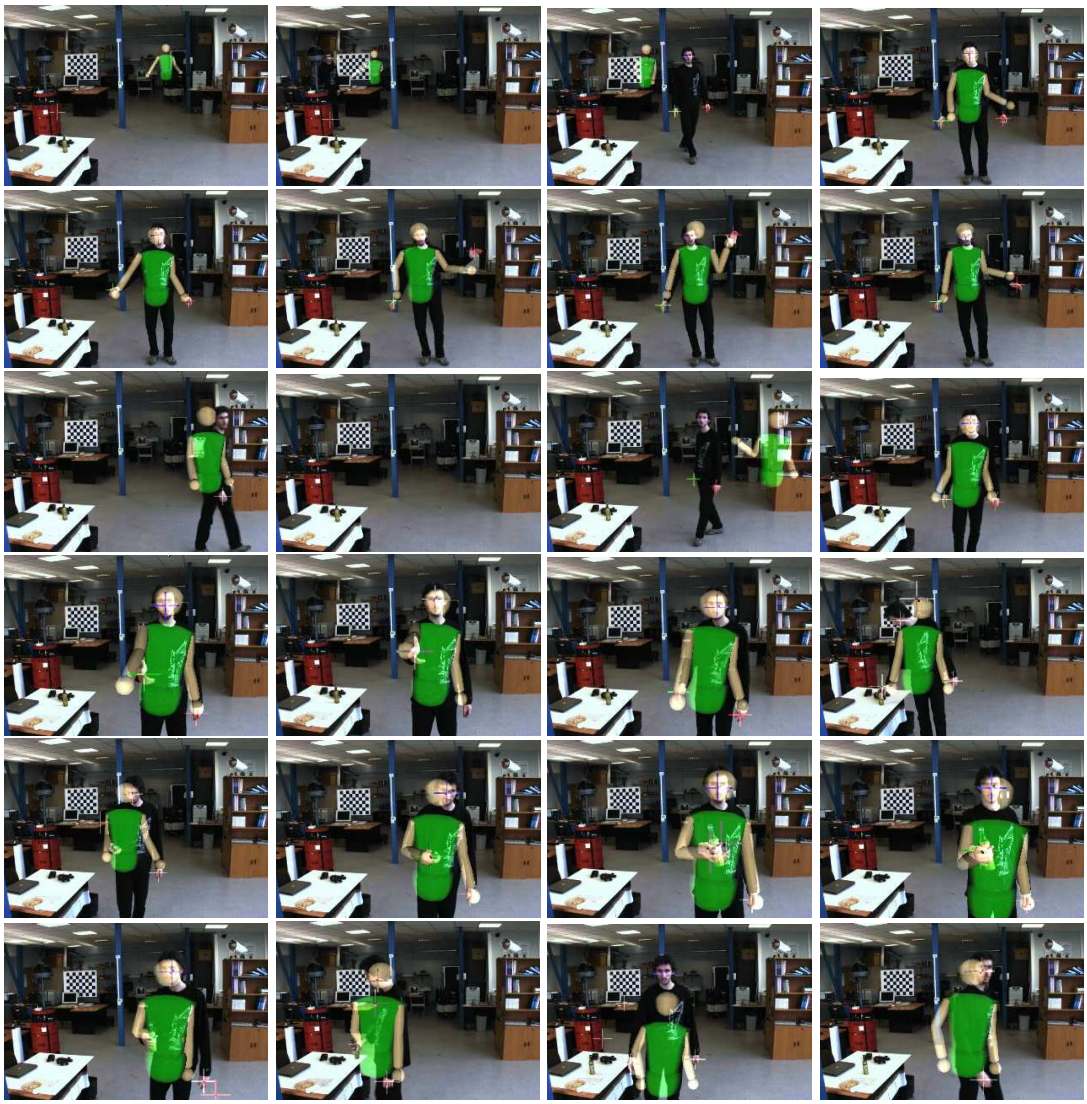


FIG. V.6 – Suivi par vision stéréoscopique sur une séquence complexe avec fond encombré.

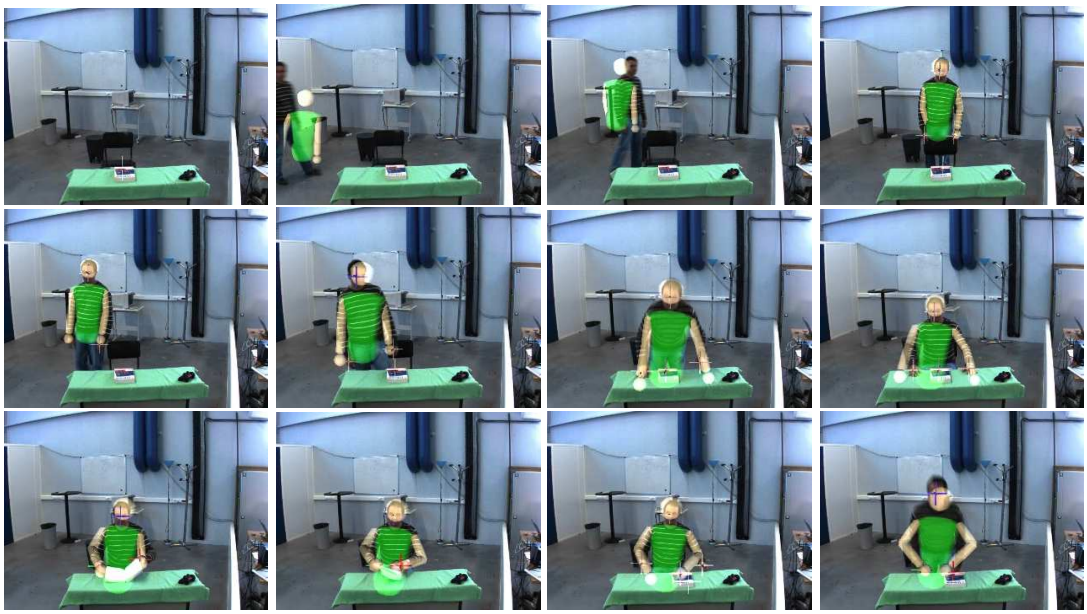


FIG. V.7 – Suivi par vision stéréoscopique sur un deuxième sujet lisant un livre. Le sujet est en partie occulté par la table.

FIG. V.8 – Suivi sur un troisième sujet perturbé par le passage d'une personne en arrière-plan.

Ainsi, à l'heure actuelle, notre système de suivi impose que le sujet porte des vêtements à manches longues et que l'arrière-plan présente peu d'objets de teinte chair situés derrière le sujet. En outre, la taille des images traitées limite l'efficacité du suivi à une distance d'environ $4,5\ m$ maximum. Les vidéos associées aux séquences présentées sont accessibles à l'URL : `www.laas.fr/~mfontmar`.

4 Vers l'exécution complète du scénario

Nos tests précédemment présentés sont menés sur des séquences prises depuis les différents systèmes et les algorithmes sont lancés « hors ligne ». Dans l'optique de la mise en place d'un système autonome, nous souhaitons cependant nous orienter vers une exécution « en ligne ».

Le robot JIDO embarque l'architecture logicielle « LAAS » (LAAS Architecture for Autonomous Systems) développée au laboratoire [4]. Celle-ci se divise en trois niveaux. En premier lieu, au dessus du matériel (capteurs et actionneurs), le niveau *fonctionnel* encapsule toutes les primitives d'action et de perception du robot dans des modules (Figure V.9). Chacun d'eux offre ses services et publie des données à l'attention de tout client potentiel : autre module fonctionnel, élément d'un niveau plus élevé dans l'architecture, opérateur. L'architecture garantit en outre que les activités codées dans les modules fonctionnels obéissent à des contraintes temporelles rigoureuses. Plus haut dans l'architecture, le niveau *exécutif* active ces modules, contrôle les fonctionnalités embarquées, et coordonne les services selon les besoins de la tâche globale. Le niveau *décisionnel* est en charge de la planification des tâches et de la supervision. Pour plus de détails, le lecteur peut consulter [28]. L'ensemble de l'architecture est codée en C/C++.

La capture de mouvement sur la base d'une paire de caméras stéréoscopiques complémente le sous-ensemble des modules fonctionnels dédiés à la perception de l'homme. À l'heure actuelle, JIDO dispose de plusieurs facultés de perception : le module ICU se focalise sur le suivi 2D des personnes et HUMREC sur la reconnaissance de visage. Le module GEST est en charge du suivi 3D du visage et des mains, le module HUMPOS localise l'homme en distance et en azimuth sur la base d'un capteur LASER. Notre objectif est de compléter cette panoplie de possibilités par une faculté de suivi 3D des gestes (module HMC). Les tests d'implémentation sont en cours.

Concernant le système binoculaire déporté, il convient de rajouter une troisième caméra au système actuel afin de le rendre plus fiable. Cependant, cet ajout risque de compromettre la vitesse de traitement déjà faible du système. Une alternative consiste à réduire la taille des images traitées, toutefois cela peut nuire à la surface de la zone utile du système. En outre, l'utilisation d'indices visuels tels que la silhouette duale $sil2$ est particulièrement consommatrice de temps. La mise en place d'un système temps-réel de capture de mouvement par caméras déportées demande ainsi de faire des compromis.

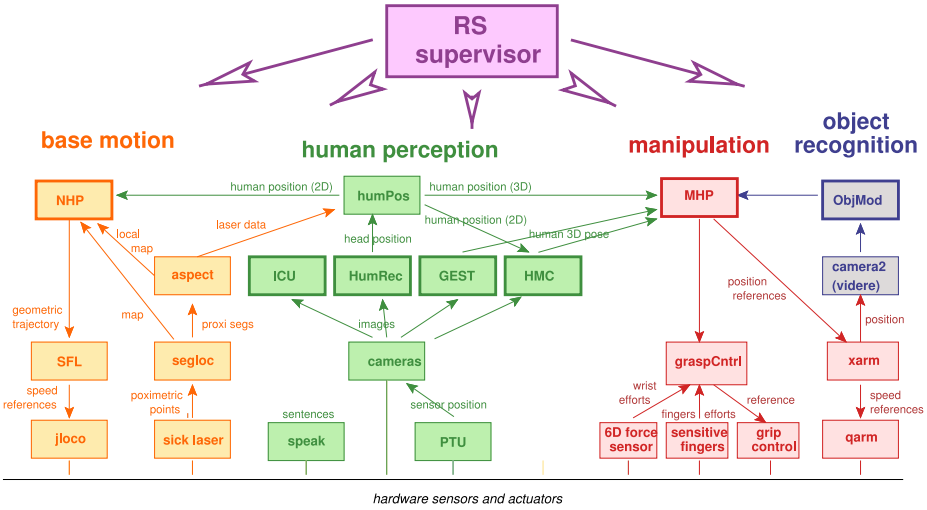


FIG. V.9 – Modules fonctionnels de l'architecture logicielle embarquée sur JIDO.

5 Conclusion et perspectives

Ce chapitre présente des évaluations sur des séquences acquises dans les deux contexte précités. Comparativement aux séquences du chapitre IV (capturées dans une salle spécifique au système de HMC commercial), ces séquences sont acquises depuis les deux systèmes multi-caméras qui viennent ici instrumenter un hall robotique et une plate-forme mobile. La caractérisation des deux fonctions HMC s'appuie logiquement sur les enseignements tirés des investigations précédentes. Le comportement satisfaisant des deux filtres dédiés corroborent les évaluations quantitatives même si les séquences analysées ont une complexité croissante. L'analyse, certes hors ligne, s'inscrit dans un scénario d'interaction homme-robot mettant en jeu les deux systèmes multi-caméras.

L'implémentation finale au sein de l'architecture logicielle de la plate-forme mobile est en cours et nous espérons prochainement coupler nos travaux avec les techniques de reconnaissance de geste actuellement investies par le groupe de recherche afin de proposer un robot autonome capable d'interactions poussées avec l'homme.

Conclusion générale

Le suivi visuel de mouvement humain est un problème très complexe. Des travaux de plus en plus nombreux sont menés depuis plus d'une vingtaine d'années, proposant un large panel de méthodes. Le nombre important d'applications envisageables constitue un enjeu majeur qui tire la recherche vers l'avant dans ce domaine : étude du mouvement, analyse médicale du geste, interaction homme-machine ou homme-robot, synthèse de mouvement, jeu vidéo, ...Toutefois, le suivi visuel demande encore un investissement important dans le but d'atteindre la mise en place de telles applications. Le problème est intrinsèquement mal posé (de par le lien ambigu qu'il existe entre la configuration de l'homme dans l'espace 3D et son apparence dans l'image), et même dans un contexte idéal, nous sommes confrontés à des cas particulièrement complexes. Les environnements réels ajoutent encore à la difficulté du problème, de par les conditions variées d'éclairage et d'encombrement de la scène étudiée. Les modèles que nous adoptons sont frustes et nuisent parfois au suivi.

Afin de traiter ce problème, de nombreuses approches sont proposées dans la littérature. Nous avons focalisé notre attention sur les filtres particulaires qui permettent de fusionner des données hétérogènes dans un cadre stochastiquement étayé. Dans l'optique d'intégrer un système de suivi sur une plate-forme robotique mobile autonome, nous avons mené nos travaux en gardant à l'esprit les contraintes de temps réel auxquelles nous sommes confrontés pour une telle application. Dans le but d'aborder les difficultés de manière graduelle, nous avons choisi d'étudier deux contextes de suivi visuel. Le premier repose sur un système multi-oculaire dans un environnement contrôlé où les caméras sont fixes et où le fond est connu. Le second, plus proche de l'application finale, exploite une caméra stéréoscopique dont on suppose ne pas connaître l'arrière-plan. Nous avons mis en place un système modulaire générique adapté au suivi visuel dans ces deux contextes. Plusieurs indices visuels 2D exploitant la forme, la couleur et le mouvement dans les images ont été présentés. En complément, nous faisons appel à une mesure 3D permettant la localisation de la tête et des mains du sujet dans l'espace.

Par la suite, nous avons rappelé le principe des stratégies de filtrage particulaire et présenté plusieurs variantes exploitant un partitionnement de l'espace de recherche, l'échantillonnage préférentiel, les approches par affinement, l'échantillonnage QMC. Afin d'évaluer l'ensemble des stratégies et des mesures proposées, nous avons mis en place diverses métriques permettant d'apprécier le comportement des filtres sous différents angles. Plusieurs évaluations préliminaires sur les stratégies de filtrage ont été

menées dans un environnement de synthèse.

Nous avons ensuite complété celles-ci par des évaluations quantitatives sur séquences réelles dans les deux contextes d'étude. La vérité de terrain a été acquise à partir d'un système de HMC commercial. Les évaluations en contexte multi-oculaire ont révélé que les stratégies de filtrage avancées ont un comportement plus satisfaisant que la stratégie classique à tout point de vue. De manière générale, les stratégies PARTITIONNÉ QRS et APF semblent sortir du lot. En contexte stéréoscopique, cette séparation est moins nette. Les fonctions de vraisemblance étant multimodales, des stratégies telles que l'APF présentent un comportement décevant par rapport à la stratégie classique.

Dans tous les cas, le choix des mesures et leur paramétrisation est primordial pour obtenir un comportement satisfaisant des filtres. Ceci est très peu mentionné dans la littérature. Nous avons établi que le choix optimal des paramètres $\sigma_{(.)}$ n'est pas identique selon le critère à minimiser. Les distances de similarités plus lissées sont plus facilement exploitables et ne compromettent pas la dispersion de l'estimateur. L'apport d'une mesure 3D est indéniable en terme de précision et de robustesse du suivi. En outre, l'introduction de cet attribut intermittent dans les fonctions de vraisemblance en complément des indices visuels 2D constitue une alternative aux stratégies avec échantillonnage préférentiel qui s'avèrent sujettes aux problèmes d'échantillonnage de l'historique. Enfin, il convient également de ne pas exploiter trop de mesures différentes afin de préserver un comportement « stable » des filtres. De manière plus générale, l'efficacité de l'approche envisagée est grandement conditionnée par le choix des indices visuels, plus que par celui de la stratégie de filtrage.

Forts des évaluations précédentes, nous avons caractérisé et validé deux systèmes de capture de mouvement depuis des caméras instrumentant l'environnement et une plate-forme robotique mobile. Leur performance respective est ici étudiée à partir de séquences (acquises puis traitées hors-ligne) mettant en jeu l'homme et le robot en environnement humain. L'intégration de notre système stéréo dans l'architecture logicielle du robot est actuellement en cours afin de permettre une analyse en ligne de séquences depuis JIDO. Cet objectif constitue un premier pas vers des applications « temps réel » d'interaction homme-robot proximale nécessitant la capture du mouvement humain.

Nous pouvons énumérer plusieurs extensions à moyen et long termes à ces travaux. Les évaluations du chapitre IV couvrent certes un large spectre mais les associations stratégies de filtrage *vs.* mesures considérées sont évidemment non exhaustives. Ainsi, il conviendrait de complèter la liste des stratégies évaluées par des stratégies plus récentes, citons ici les stratégies CSS [146] et APF avec partitionnement automatique [39]. Concernant les mesures, nous avons limité l'étude à des indices relativement classiques mais pertinents eu égard aux contraintes robotiques. Il pourrait être intéressant de mettre en œuvre des indices visuels plus évolués, notamment en exploitant plus finement le mouvement. Il serait également pertinent d'évaluer l'influence du modèle de dynamique et de sa paramétrisation dans nos filtres, voire exploiter des modèles plus complexes. Enfin, nous aimerions étudier le comportement des deux systèmes multi-caméras exhibés à la capture de mouvement sur des sujets divers et variés en terme de

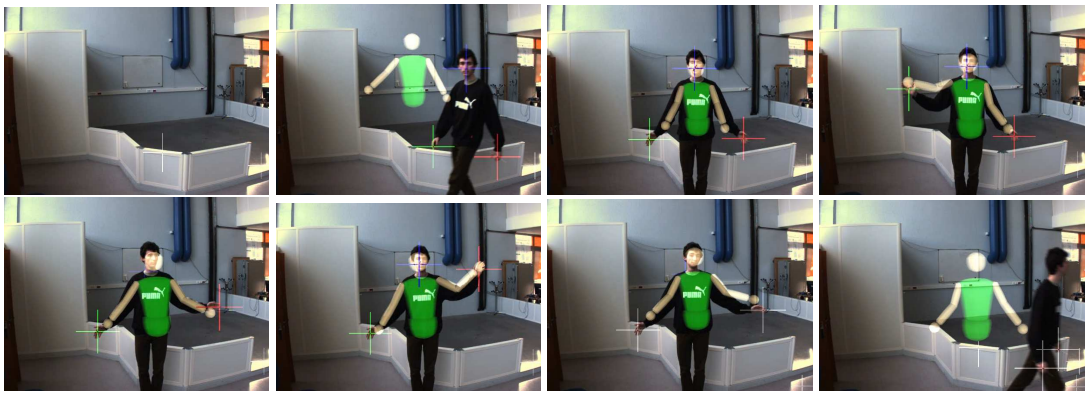


FIG. VI.1 – Au revoir...

morphologie. Au delà des quelques illustrations prometteuses du chapitre V, il semble opportun d'adapter en ligne les dimensions du modèle 3D, voire inclure sa paramétrisation dans l'espace des paramètres estimés à l'instar de [146].

A plus longue échéance, nous aimerions voir coopérer les deux systèmes de HMC déporté et embarqué. La fusion des deux flux vidéo, éventuellement asynchrones, permettrait de robustifier la capture du mouvement humain au voisinage immédiat du robot par une perception globale et locale. À notre connaissance, cette problématique n'a pas été encore étudiée. À plus haut niveau, nos systèmes de HMC, dans un fonctionnement couplé ou non, doivent permettre l'interprétation de gestes ou de postures à partir des travaux menés par ailleurs dans le groupe sur les modèles de Markov cachés et les réseaux bayésiens dynamiques [19]. Un challenge sera alors de coupler la perception de l'homme avec la perception de l'environnement *i.e.* l'espace et/ou les objets afin d'inférer une interprétation plus globale des actions et activités de l'homme dans son contexte environnemental, voire de plusieurs individus partageant l'espace avec le robot censé interagir avec ceux-ci. Dans cette perspective, les travaux réalisés sur la capture de mouvement humain depuis des caméras déportées ou embarquées constituent sans aucun doute un pré-requis indispensable.

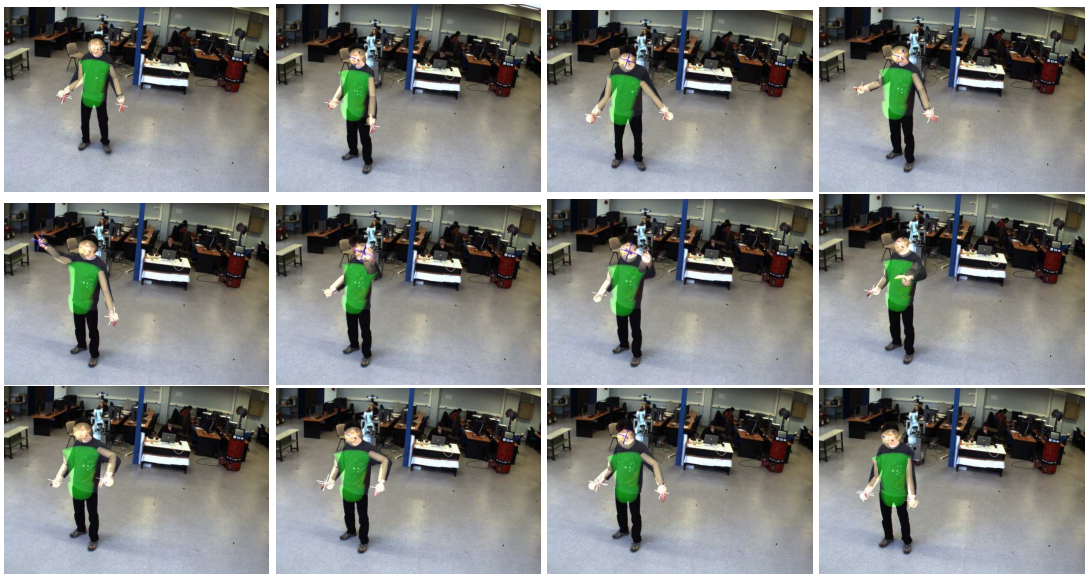


FIG. VI.2 – ... Et à bientôt !

FIG. VI.3 – Robot interaction from a different point of view. Dessin : Mathieu Warnier, scénario : Akin Sisbot, couleur et numérisation : Mathias Fontmarty.

Annexe A

Liste de publications

Cette annexe liste l'ensemble des publications qui ont ponctué le déroulement de cette thèse.

1 Congrès internationaux

- M. Fontmarty, F. Lerasle et P. Danès. Towards real-time markerless human motion capture from ambiance cameras using an hybrid particle filter. Dans *IEEE International Conference on Image Processing (ICIP'08), San Diego (CA), USA, octobre, 2008*
- M. Fontmarty, F. Lerasle, et P. and Danès. Data fusion within a modified annealed particle filter dedicated to human motion capture. Dans *International Conference on Intelligent Robots and Systems (IROS'07), San Diego, CA, USA novembre, 2007*

2 Symposiums internationaux

- M. Fontmarty, T. Germa, B. Burger, L.F. Marin et S. Knoop. Implementation of human perception algorithms on a mobile robot. Dans *6^{th} IFAC Symposium on Intelligent Autonomous Vehicles (IAV'07), Toulouse, FRANCE, septembre, 2007*
- E. A. Sisbot, A. Clodic, L.F. Marin, M. Fontmarty, L. Brèthes, et R. Alami. Implementing a human-aware robot system. Dans *IEEE International Symposium on Robot and Human Interactive Communication 2006 (RO-MAN'06), Hatfield, UK, septembre, 2006*

3 Congrès nationaux

- M. Fontmarty, F. Lerasle et P. Danès. Une stratégie hybride de filtrage particulaire pour le suivi de mouvement humain depuis un robot mobile. Dans *$16^{ème}$*

congrès francophone AFRIF-AFIA, Reconnaissance des Formes et Intelligence Artificielle (RFIA'08), Amiens, FRANCE, janvier, 2008

- M. Fontmarty, F. Lerasle, P. Danès et P. Menezes. Filtrage Particulaire pour la capture de mouvement dédiée à l'interaction Homme-Robot. Dans *11ème congrès francophone des jeunes chercheurs en vision par ordinateur (ORASIS'07), Obernai, FRANCE, juin, 2007*
- M. Fontmarty, F. Lerasle, et P. Danès. Suivi 3D de mouvements humains pour l'interaction Homme-Robot. Dans *8ème congrès des doctorants (EDSYS'07), Albi, FRANCE, mai, 2007*

4 Journaux internationaux

- M. Fontmarty, F. Lerasle et P. Danès. [Soumis] Evaluation of particle filter based human motion visual trackers for home environment surveillance. Dans *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), 2008*
- L. Brèthes, F. Lerasle, P. Danès, et M. Fontmarty. Particle filtering strategies for data fusion dedicated to visual tracking from a mobile robot. *Journal of Machine Vision and Applications (MVA), 2008*

5 Autres

- M. Fontmarty (couleurs), E. A. Sisbot (scénario) et M. Warnier (dessin). Human-Robot interaction from a different point of view (Comic). Dans *IEEE International Conference on Robotics and Automation (ICRA'08), Pasadena (CA), USA, mai, 2008*

Annexe B

Prétraitement des images

L'acquisition des images est faite à l'aide de caméras couleur Point Grey Flea grand angle *via* le bus *Firewire* (IEEE 1394). Les images fournies ont des résolutions allant de 160×120 à 1024×768, mais nous travaillons au maximum sur des images 640×480, essentiellement pour des raisons de temps de traitement et de bande passante lors de l'acquisition et du transfert du flot d'images. Suivant le contexte dans lequel nous travaillons et le type de caméra utilisé, les images produites peuvent être sensiblement différentes (cf figure B.1). Les caméras embarquées sur le robot par exemple sont de type mono CCD, proposant ainsi des images mosaïquées. De même, l'objectif de la caméra et l'éclairage ambiant donnent lieu à un certain nombre d'aberrations qu'il faut corriger afin d'exploiter au mieux ces images. Ces pré-traitements, présentés figure B.2, sont détaillés par la suite.

1 « Démosaïquage »

Bien que les technologies récentes dans le domaine de la vision permettent l'utilisation de caméras couleur tri-CCD (une rétine CCD par couleur primaire), les caméras mono-CCD sont très usitées, de par leur coût plus abordable. Afin de restituer une image couleur, ces capteurs sont recouverts d'un filtre permettant de ne laisser passer qu'une seule des trois composantes chromatiques par pixel. Ce procédé donne lieu à des images dites « mosaïquées » ou « Bayerisées », du nom du type de filtre utilisé (figure B.2 (a)). L'image telle qu'elle est fournie par la caméra peut être vue comme une simple image d'intensité (1 seule information par pixel), mais son interprétation dans le domaine des couleurs est possible à l'aide d'une opération de démosaïquage, consistant en une interpolation chromatique entre chaque pixel (figure B.3). Les caméras embarquées sur la plate-forme mobile que nous manipulons nécessitent ce type de pré-traitement, contrairement aux caméras de surveillance. La méthode que nous utilisons est basée sur une interpolation bilinéaire. Une étude relativement complète des techniques d'interpolation proposées dans la littérature est présentée dans [6].

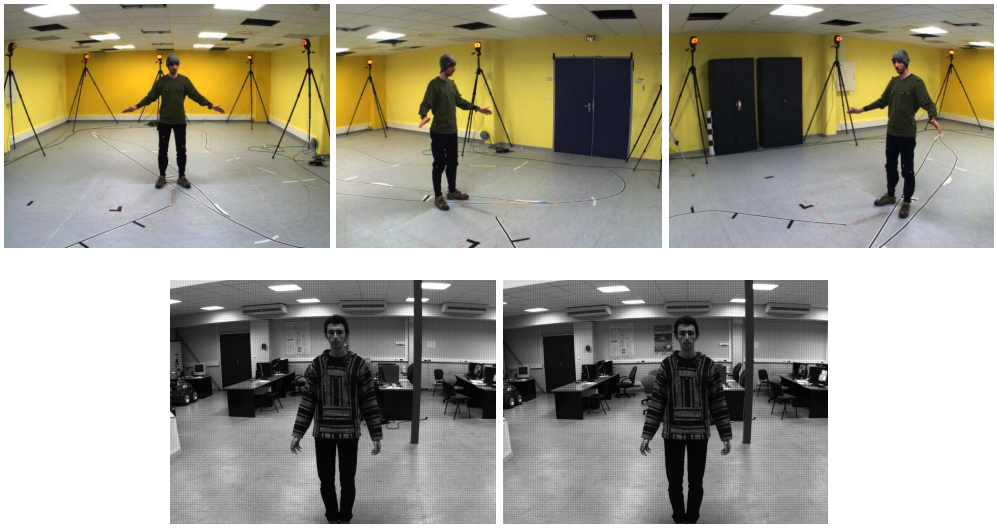


FIG. B.1 – Les images “source” provenant de trois caméras de surveillance (haut) et de caméras stéréoscopiques embarquées sur le robot (bas).

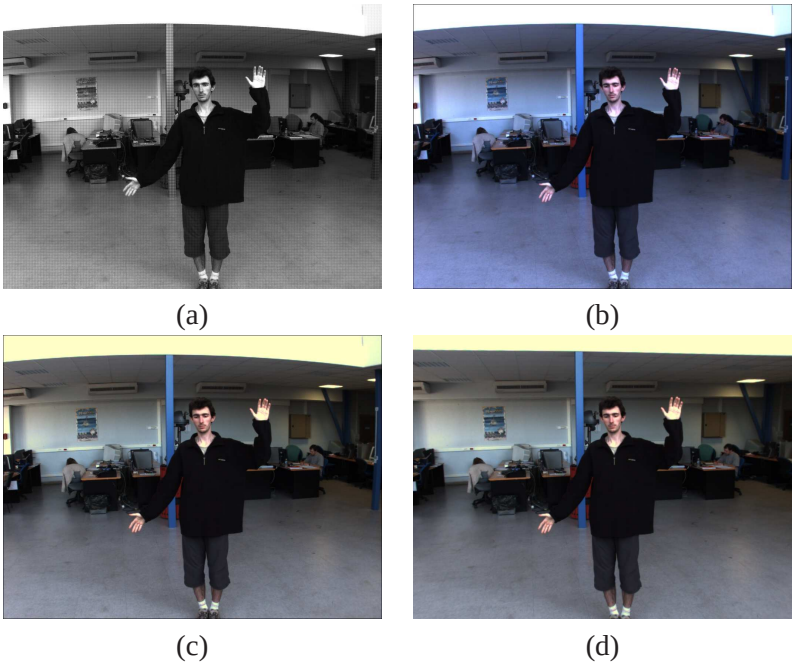


FIG. B.2 – Les quatre étapes du pré-traitement des images : (a) image originale brute (issue des caméras) ; (b) image couleur démosaïquée ; (c) image après application de la balance des blancs ; (d) image finale corrigée en distorsion.

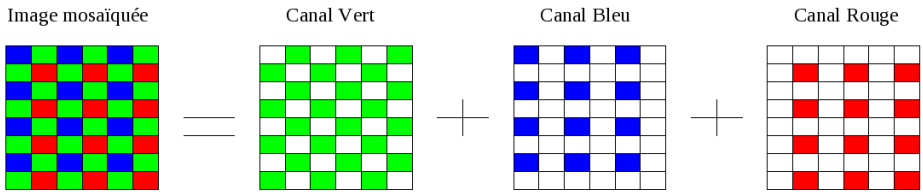


FIG. B.3 – L'image mosaïquée (gauche) permet de reconstituer les trois composantes chromatiques de chaque pixel. La valeur des cases blanches est calculée par interpolation des valeurs voisines.

2 Balance des blancs

Dans certains environnements, les caméras fournissent des images aux couleurs altérées par l'éclairage ambiant. Les tonalités présentent ainsi une couleur dominante (cf. image B.2 (b)). Il convient alors de procéder à un équilibrage des couleurs, aussi appelé balance des blancs, afin de retrouver le panachage des couleurs, idéalement tel qu'il est perçu par l'oeil humain (cf. image B.2 (c)). De nombreux algorithmes sont proposés dans la littérature (hypothèse du monde blanc [6], rétinex [27], correction automatique [57]). Notre approche est basée sur l'hypothèse du « monde gris ». Elle a l'avantage d'être simple à mettre en oeuvre, et peu gourmande en temps de calcul. Cette hypothèse suppose que la moyenne des couleurs d'une image est proche du gris dans le cas général. La moyenne chromatique réelle (R_m, G_m, B_m) de chaque image est calculée et, afin de la ramener à une teinte grise, nous calculons les nouvelles couleurs (R', G', B') de chaque pixel à partir des valeurs originales (R, G, B) *via*

$$R' = R \times \frac{G_m}{R_m}, \qquad G' = G, \qquad B' = B \times \frac{G_m}{B_m} \tag{B.1}$$

La moyenne est ainsi corrigée en (G_m, G_m, G_m). Notons ici que nous avons choisi comme référence la couleur verte car elle est deux fois plus présente que les autres au sein de la matrice de Bayer. De nombreuses variantes existent, dont la plus connue consiste à prendre comme valeur de référence l'intensité moyenne d'une zone prédéfinie de l'image. Le lecteur curieux pourra ici encore se référer à [6] pour plus de détails.

3 Correction de distorsion

Après avoir corrigé les aberrations chromatiques de l'image, la dernière étape rectifie quant à elle les aberrations géométriques. De part l'utilisation d'objectifs grand angle, les images affichent parfois une distorsion importante (figure B.2 (c)), qu'il convient de corriger. Ceci permet une meilleure cohérence vis-à-vis de la localisation du sujet par la suite. La figure B.2 (d) présente le résultat d'une correction en distorsion

radiale à l'ordre 3. Nous utilisons l'algorithme classique consistant en une interpolation bilinéaire selon les paramètres intrinsèques de la caméra obtenus après une calibration [43].

L'ensemble des pré-traitements réalisés sur l'image constitue la première étape de notre système et permet donc d'obtenir une information plus exploitable afin de suivre au mieux les mouvements de sujet.

Annexe C

Détails des algorithmes de filtrage

Nous exposons dans cette annexe les algorithmes de filtrage que nous n'avons pas détaillés dans le chapitre III pour ne pas le surcharger. Nous présentons donc une version mathématiquement correcte de l'APF en table C.1 qui ne diffère de la version classique que par son étape de rééchantillonnage, et dont il serait par ailleurs intéressant d'étudier le comportement, l'IAPF en table C.2, et la version QRS du PARTITIONNÉ dénommée PARTITIONNÉ QRS en table C.3.

$\{(\boldsymbol{x}_k^{(i)}, w_k^{(i)})\}_{i=1}^N = \mathrm{APF2}(\{\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1}^N, \boldsymbol{z}_k)$

1: **SI** $k = 0$**, ALORS** Échantillonner $\boldsymbol{x}_0^{(1)}, \ldots, \boldsymbol{x}_0^{(N)}$ i.i.d. selon $p_0(\boldsymbol{x}_0)$, et poser $w_0^{(i)} = \frac{1}{N}$ **FIN SI**
2: **SI** $k \geq 1$ **ALORS** $\{$ —$\{(\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)})\}_{i=1}^N$ représente $p(\boldsymbol{x}_{k-1}|\boldsymbol{z}_{1:k-1})$—$\}$
3: Poser $\{(\boldsymbol{x}_{k,0}^{(i)}, w_{k,0}^{(i)})\}_{i=1}^N = \{(\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)})\}_{i=1}^N$, $\lambda_{k,0}^{(i)} = w_{k,0}^{(i)}$, et choisir $1 < \alpha_1 < \cdots < \alpha_L$ et $\beta_1 < \cdots < \beta_L < 1$
4: **POUR** $l = 1, \ldots, L-1$, **FAIRE**
5: **POUR** $i = 1, \ldots, N$, **FAIRE**
6: Échantillonner indépendamment $\boldsymbol{x}_{k,l}^{(i)} \sim p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1}^{(i)})$
7: Définir le poids intermédiaire $\lambda_{k,l}^{(i)} \propto \lambda_{k,l-1}^{(i)} p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)})$
8: **FIN POUR**
9: Normaliser les poids intermédiaires de sorte que $\sum_{i=1}^N \lambda_{k,l}^{(i)} = 1$
10: Échantillonner dans $(1, \ldots, N)$ l'ensemble $\{s^{(i)}\}_{i=1}^N$ selon $P(s^{(i)} = i) = \lambda_{k,l}^{(i)}$. Poser $w_{k,l}^{(s^{(i)})} = \frac{w_{k,l-1}^{(i)}}{\lambda_{k,l}^{(i)}}$, normaliser l'ensemble $\{w_{k,l}^{(s^{(i)})}\}_{i=1}^N$, puis poser $\lambda_{k,l}^{(s^{(i)})} = w_{k,l}^{(s^{(i)})}$. Renommer $\{(\boldsymbol{x}_{k,l}^{(s^{(i)})}, w_{k,l}^{(s^{(i)})}, \lambda_{k,l}^{(s^{(i)})})\}_{i=1}^N$ en $\{(\boldsymbol{x}_{k,l}^{(i)}, w_{k,l}^{(i)}, \lambda_{k,l}^{(i)})\}_{i=1}^N$. —$\{(\boldsymbol{x}_{k,l}^{(i)}, w_{k,l}^{(i)})\}_{i=1}^N$ représente $\int \cdots \int p_l(x_{k,l}|x_{k,l-1}) p_{l-1}(x_{k,l-1}|x_{k,l-2}) \ldots p_1(x_{k,1}|x_{k,0}) p_0(x_{k,0}|z_{1:k-1}) \mathrm{d}x_{k,0} \mathrm{d}x_{k,1} \ldots \mathrm{d}x_{k,l-1}$—
11: **FIN POUR**
12: **POUR** $i = 1, \ldots, N$, **FAIRE**
13: Échantillonner indépendamment $\boldsymbol{x}_k^{(i)} = \boldsymbol{x}_{k,L}^{(i)} \sim p(\boldsymbol{x}_k|\boldsymbol{x}_{k,L-1}^{(i)})$
14: Associer le poids $w_k^{(i)} \propto w_{k-1}^{(i)} p(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)})$
15: **FIN POUR**
16: Calculer l'estimé du MMSE $\mathrm{E}_{p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})}[\boldsymbol{x}_k] = \sum_{i=1}^N w_k^{(i)} \boldsymbol{x}_k^{(i)}$
17: **FIN SI**

TAB. C.1 – Version mathématiquement correcte de l'algorithme de l'« Annealed Particle Filter ». L'étape 10 fait intervenir des rééchantillonnages pondérés, tels que définis dans [101].

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

SI $k = 0$**, ALORS** Échantillonner $\boldsymbol{x}_0^{(1)}, \ldots, \boldsymbol{x}_0^{(N)}$ i.i.d. selon $p_0(\boldsymbol{x}_0)$, et poser $w_0^{(i)} = \frac{1}{N}$ **FIN SI**
SI $k \geq 1$ **ALORS** $\{$ —$\{(\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)})\}_{i=1}^N$ représente $p(\boldsymbol{x}_{k-1}|\boldsymbol{z}_{1:k-1})$—$\}$
  Poser $\{(\boldsymbol{x}_{k,0}^{(i)}, w_{k,0}^{(i)})\}_{i=1}^N = \{(\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)})\}_{i=1}^N$. Choisir $1 \leq \alpha_1 < \cdots < \alpha_L$ et $\beta_1 < \cdots < \beta_L \leq 1$
  Poser $l = 1$ — Gestion de la première itération —
  POUR $i = 1, \ldots, N$, **FAIRE**
    Échantillonner indépendamment $\boldsymbol{x}_{k,l}^{(i)} \sim q(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1}^{(i)}, \boldsymbol{z}_k)$
    Associer le poids $w_{k,l}^{(i)} \propto w_{k,l-1}^{(i)} \frac{p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)}) p_l(\boldsymbol{x}_{k,l}^{(i)}|\boldsymbol{x}_{k,l-1}^{(i)})}{q(\boldsymbol{x}_{k,l}^{(i)}|\boldsymbol{x}_{k,l-1}^{(i)}, \boldsymbol{z}_k)}$
  FIN POUR
  Normaliser les poids de sorte que $\sum_{i=1}^N w_{k,l}^{(i)} = 1$
  Rééchantillonner $\{(\boldsymbol{x}_{k,l}^{(i)}, w_{k,l}^{(i)})\}_{i=1}^N$
  — Gestion des itérations suivantes —
  POUR $l = 2, \ldots, L$, **FAIRE**
    POUR $i = 1, \ldots, N$, **FAIRE**
      Échantillonner indépendamment $\boldsymbol{x}_{k,l}^{(i)} \sim p_l(\boldsymbol{x}_{k,l}|\boldsymbol{x}_{k,l-1}^{(i)})$
      Associer le poids $w_{k,l}^{(i)} \propto w_{k,l-1}^{(i)} p_l(\boldsymbol{z}_k|\boldsymbol{x}_{k,l}^{(i)})$
    FIN POUR
    Normaliser les poids de sorte que $\sum_{i=1}^N w_{k,l}^{(i)} = 1$
    SI $l < L$ **ALORS**
      Rééchantillonner $\{(\boldsymbol{x}_{k,l}^{(i)}, w_{k,l}^{(i)})\}_{i=1}^N$
    FIN SI
  FIN POUR
  Renommer $\{(\boldsymbol{x}_{k,L}^{(i)}, w_{k,L}^{(i)})\}_{i=1}^N$ en $\{(\boldsymbol{x}_k^{(i)}, w_k^{(i)})\}_{i=1}^N$
  Calculer l'estimé du MMSE $\mathrm{E}_{p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})}[\boldsymbol{x}_k] = \sum_{i=1}^N w_k^{(i)} \boldsymbol{x}_k^{(i)}$
FIN SI

TAB. C.2 – Algorithme de l'« I-Annealed Particle Filter ».

$\{(\boldsymbol{x}_k^{(i)}, w_k^{(i)})\}_{i=1}^N = \text{PARTITIONNÉ QRS}(\{(\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)})\}_{i=1}^N, \boldsymbol{z}_k)$

SI $k = 0$**, ALORS** Échantillonner une séquence QMC randomisée de Sobol $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(N)}$ selon $\mathcal{U}_{[0,1)^d}(\boldsymbol{u})$, la convertir en $\boldsymbol{x}_0^{(1)}, \ldots, \boldsymbol{x}_0^{(N)} \sim p_0(\boldsymbol{x}_0)$, et poser $w_0^{(i)} = \frac{1}{N}$. **FIN SI**
SI $k \geq 1$ **ALORS** $\{$—$\{(\boldsymbol{x}_{k-1}^{(i)}, w_{k-1}^{(i)})\}_{i=1}^N$ représente $p(\boldsymbol{x}_{k-1}|\boldsymbol{z}_{1:k-1})$—$\}$
  Poser $\tau_0^{(i)} = w_{k-1}^{(i)}$ et $\boldsymbol{x}_k^{0,(i)} = \boldsymbol{x}_{k-1}^{(i)}$ pour $i = 1, \ldots, N$
  POUR $m = 1, \ldots, M$**, FAIRE**
    Sélectionner avec remise $s^{(1)}, \ldots, s^{(N)}$ dans $\{1, \ldots, N\}$ tels que $P(s^{(i)} = j) = \tau_{m-1}^{(j)}$
    Poser $C_j = card(\{i|s^{(i)} = j\})$
    POUR $j = 1, \ldots, N$**, FAIRE**
      Échantillonner une séquence QMC randomisée de Sobol $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(C_j)}$ selon $\mathcal{U}_{[0,1)^d}(\boldsymbol{u})$ et la convertir en $\boldsymbol{x}_k^{m,(\sum_{l=1}^{j-1} C_l+1)}, \ldots, \boldsymbol{x}_k^{m,(\sum_{l=1}^{j-1} C_l+C_j)} \sim p_m(\boldsymbol{x}_k^m|\boldsymbol{x}_k^{m-1,(j)})$
    FIN POUR
    Mettre à jour les poids *via* $\tau_m^{(i)} \propto l_m(\boldsymbol{z}_k|\boldsymbol{x}_k^{m,(i)})$
    Normaliser les poids de sorte que $\sum_i \tau_m^{(i)} = 1$
  FIN POUR
  Poser $w_k^{(i)} = \tau_m^{(i)}$ et $\boldsymbol{x}_k^{(i)} = \boldsymbol{x}_k^{m,(i)}$ pour $i = 1, \ldots, N$
  Calculer le MMSE $\mathrm{E}_{p(\boldsymbol{x}_k|\boldsymbol{z}_{1:k})}[\boldsymbol{x}_k] = \sum_{i=1}^N w_k^{(i)} \boldsymbol{x}_k^{(i)}$
FIN SI

TAB. C.3 – Algorithme PARTITIONNÉ QRS, ou filtrage particulaire PARTITIONNÉ exploitant les techniques QMC. Par souci de simplification d'écriture, $\boldsymbol{x}_k^m$ désigne ici le vecteur complet à l'étape m du filtre PARTITIONNÉ, et non plus la seule sous-partition m.

Annexe D

Analyse du comportement des filtres

Nous présentons ci-après les résultats obtenus sur les séquences de test dans le chapitre IV selon :

Multi-oculaire – séquence 3	Figure D.1
Multi-oculaire – séquence 4	Figure D.2
Stéréoscopique – séquence 1	Figure D.3
Stéréoscopique – séquence 3	Figure D.4

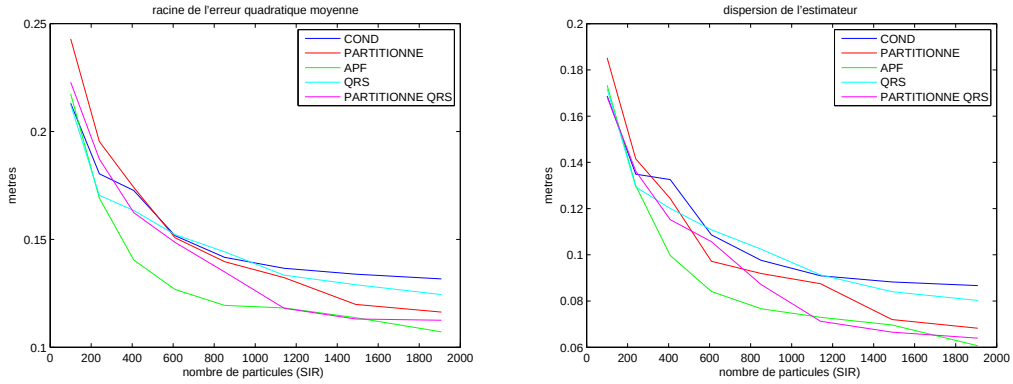


FIG. D.1 – Influence du nombre de particules N et de la stratégie de filtrage en contexte **multi-oculaire** sur la **séquence 3** pour $R = 30$ réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais.

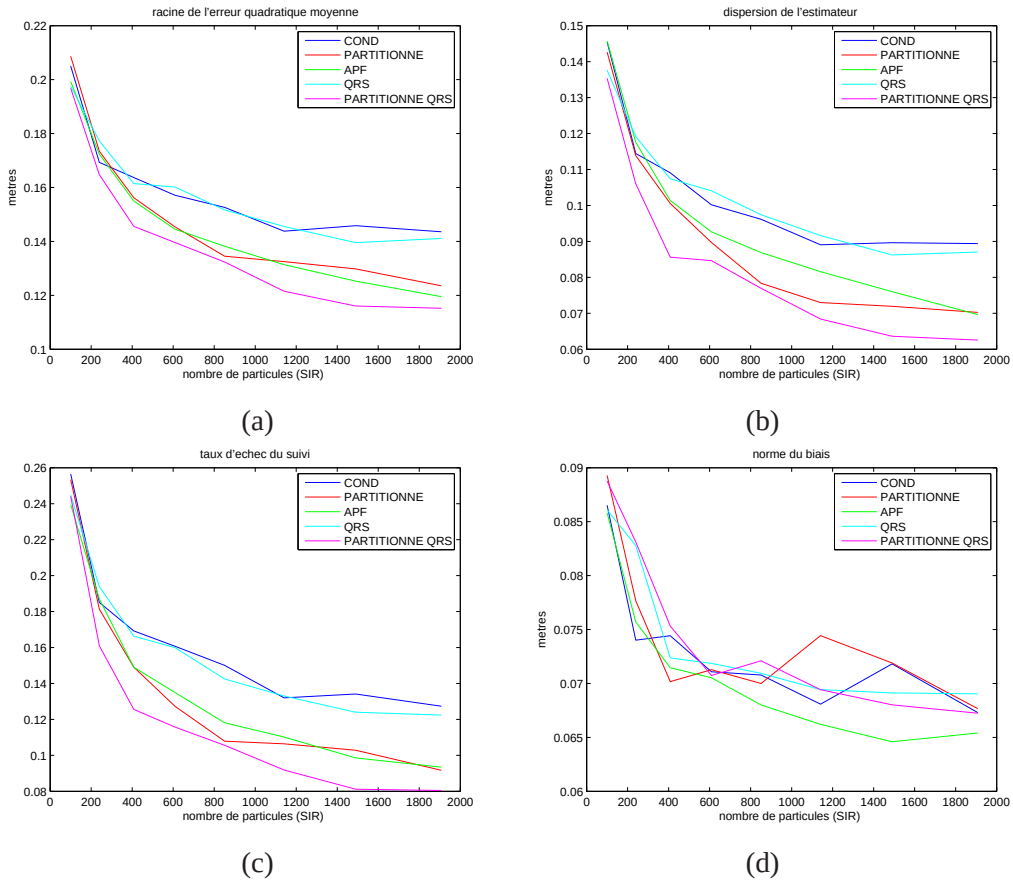


FIG. D.2 – Influence du nombre de particules N et de la stratégie de filtrage en contexte **multi-oculaire** sur la **séquence 4** pour $R = 30$ réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais.

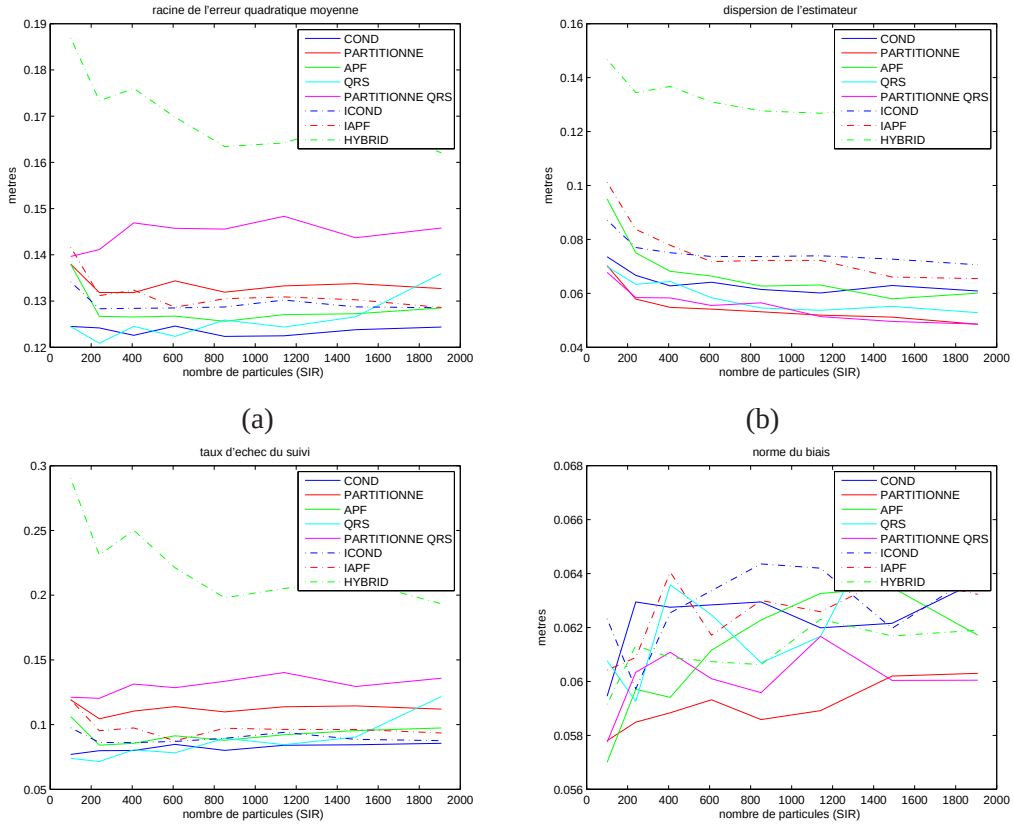


FIG. D.3 – Influence du nombre de particules N et de la stratégie de filtrage en contexte **stéréoscopique** sur la **séquence 1** pour $R = 30$ réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais.

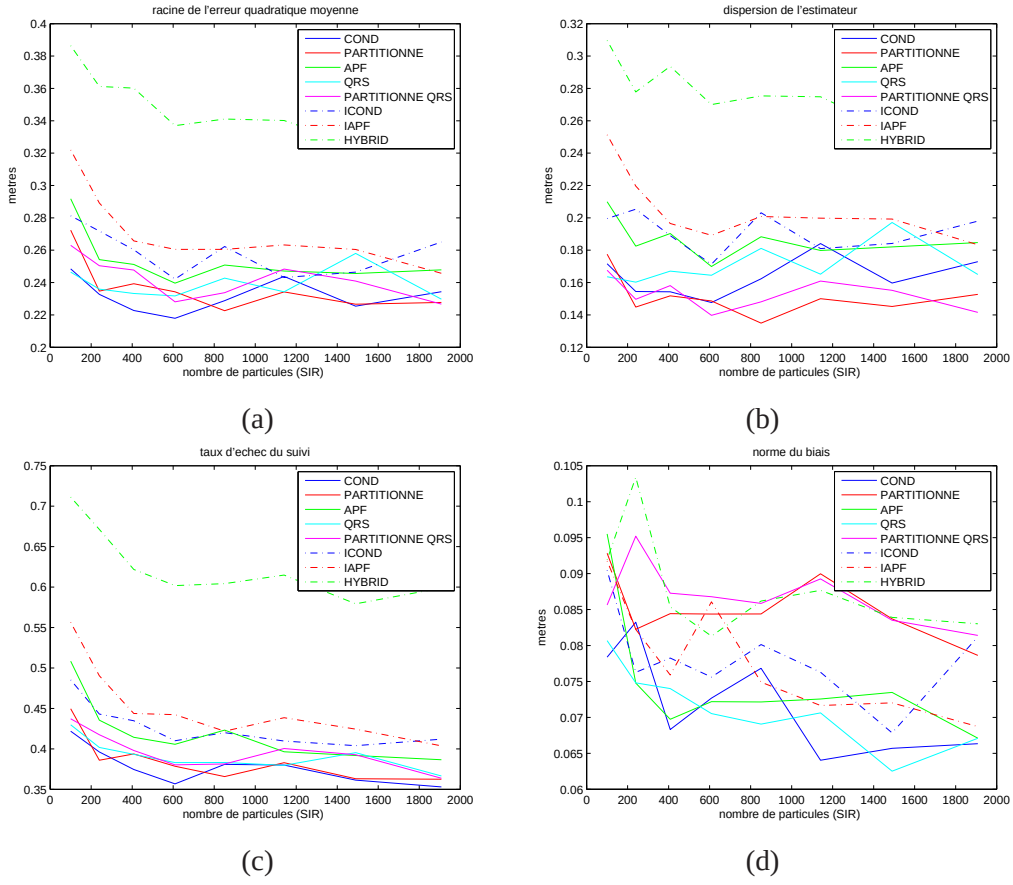


FIG. D.4 – Influence du nombre de particules N et de la stratégie de filtrage en contexte **stéréoscopique** sur la **séquence 3** pour $R = 30$ réalisations du filtre : (a) RMSE, (b) dispersion de l'estimateur, (c) taux d'échec, (d) norme du biais.

Glossaire

ACP : Analyse de Composantes Principales. Technique bayésienne permettant le calcul des axes principaux d'une distribution.

BP : « Belief Propagation » ou propagation de croyance.

CONDENSATION : « CONditional DENSity propagATION », premier algorithme de filtrage particulaire proposé.

DDL : Degré de liberté.

DEL : Diode Électro-Luminescente.

EKF : « Extended Kalman Filter », extension du filtre de Kalman aux systèmes non linéaires.

FP : Filtre Particulaire.

HMC : Human Motion Capture, ou capture de mouvement humain en français.

ICP : « Iterative Closest Point », algorithme d'appariement de 2 ensembles de points.

i.i.d. : indépendants et identiquement distribués.

KF : « Kalman Filter » ou filtre de Kalman. Technique de filtrage pour les systèmes dynamiques stochastiques linéaires gaussiens.

PARTITIONNÉ : Évolution de la CONDENSATION dans les cas où l'espace d'état est divisible en plusieurs partitions.

MC : Monte Carlo.

MMSE : « Minimum Mean Square Error », ou minimum de l'erreur quadratique moyenne. C'est un estimateur de l'espérance d'une variable aléatoire.

OUPF : « Optimized Unscented Particle Filter », filtre particulaire exploitant la transformation « Unscented », à l'instar du filtre UKF.

QMC : Quasi Monte Carlo.

RMSE : « Root Mean Square Error », ou racine de l'erreur quadratique moyenne.

SCP : Somme Cumulée des Poids. Elle est utilisée dans l'algorithme de rééchantillonnage systématique.

SMD : « Stochastic Meta-Descent », algorithme d'optimisation mixant des techniques déterministes et probabilistes.

UKF : « Unscented Kalman Filter », extension du filtre de Kalman aux systèmes non linéaires, reconnue comme plus efficace que l'EKF.

Bibliographie

[1] J. Y. Bouguet. http ://www.vision.caltech.edu/bouguetj/calib_doc/ — camera calibration toolbox for matlab.

[2] J. K. Aggarwal and Q. Cai. Human motion analysis : A review. *Computer Vision and Image Understanding*, 73(3) :428–440, 1999.

[3] 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.

[4] R. Alami, R. Chatila, S. Fleury, and F. Ingrand. An architecture for autonomy. *International Journal of Robotics Research (IJRR'98)*, 17(4) :315–337, 1998.

[5] 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.

[6] G. Avina. *Navigation visuelle d'un robot mobile dans un environnement extérieur semi-structuré*. PhD thesis, Institut National Polytechnique de Toulouse, 2005.

[7] P. Azad, A. Ude, T. Asfour, G. Cheng, and R. Dillmann. Image-based markerless 3D human motion capture using multiple cues. In *International Workshop on Vision Based Human-Robot Interaction*, Palermo, Italy, March 2006.

[8] 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.

[9] 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 (HUMANOIDS'O4)*, volume 2, pages 941–959, California, USA, November 2004.

[10] A. Balan, L. Sigal, and M. Black. A quantitative evaluation of video-based 3D person tracking. In *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS'05)*, pages 349–356, Washington, USA, October 2005.

[11] M. Benewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke. Towards a humanoid museum guide robot that interacts with multiple persons. In *International Conference on Humanoid Robots (HUMANOID'05)*, pages 418–423, Tsukuba, JAPAN, 2005.

[12] 0. Bernier and P. Cheung-Mon-Chang. Real-time 3D articulated pose tracking using particle filtering and belief propagation on factor graphs. In *British Machine Vision Conference (BMVC'06)*, volume 1, pages 5–8, 2006.

[13] A. Blake, M. Isard, and J. MacCormick. *Sequential Monte Carlo Methods in Practice*, chapter Statistical Models of Visual Shape and Motion, pages 339–358. Springer-Verlag, 2001.

[14] M. Bray, E. Koller-Meier, and L. Van Gool. 3D hand tracking by rapid stochastic gradient descent using a skinning model. In *Automatic Face and Gesture Recognition, (FGR'04)*, pages 675–680, May 2004.

[15] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'98)*, Santa Barbara (CA), USA, 1998.

[16] L. Bretzner, I. Laptev, and T. Lindeberg. Hand gesture recognition using multi-scale color features, hierarchical models and particle filtering. In *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 405 – 410, May 2002.

[17] L. Brèthes. *Suivi visuel par filtrage particulaire. Application à l'interaction homme-robot*. PhD thesis, Université de Toulouse, UPS, 2005.

[18] L. Brèthes, F. Lerasle, P. Danès, and M. Fontmarty. Particle filtering strategies for data fusion dedicated to visual tracking from a mobile robot. *Journal of Machine Vision and Applications (MVA)*, 2008.

[19] B. Burger, F. Lerasle, and I. Ferrané. Mutual assistance between speech and vision for human-robot interaction. In *International Conference on Intelligent Robots and Systems (IROS'08)*, Nice, FRANCE, September 2008.

[20] S. R. Buss. Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods. available at http ://euclid.ucsd.edu/ sbuss/ResearchWeb/ikmethods/iksurvey.pdf, 2004.

[21] F. Caillette, A. Galata, and T. Howard. Real-time 3D human body tracking using learnt models of behaviour. *Computer Vision and Image Understanding*, 2007.

[22] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6) :679–698, 1986.

[23] G. Casella and Robert C. P. Rao-blackwellisation of sampling schemes. *Biometrika*, 83 :81–94, 1996.

[24] Zhe Chen. Bayesian filtering : from Kalman filters to particles filters, and beyond. available on http ://www.math.u-bordeaux.fr/∼delmoral/chen_bayesian.pdf, 2003.

[25] K.M. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler. A real time system for robust 3D voxel reconstruction of human position. In *IEEE Conference on Conference on Vision and Pattern Recognition (CVPR'00)*, volume 2, pages 714–720, Hilton Head Island, South Carolina, USA, June 2000.

[26] G. Cielniak, A. Treptow, and T. Duckett. Quantitative performance evaluation of a people tracking system on a mobile robot. In *European Conference on Mobile Robots (ECMR'05)*, Ancona, Italy, September 2005.

[27] F. Ciurea and B. Funt. Tunin retinex parameters. *Journal of eletronic imaging*, 13(1) :58–64, January 2004.

[28] A. Clodic, V. Montreuil, R. Alami, and R. Chatila. A decisional framework for autonomous robots interacting with humans. In *IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN'05)*, 2005.

[29] S. Corazza, L. Mündermann, and T. Andriacchi. Markerless motion capture methods for the estimation of human body kinematics. In *9th international symposium on the 3D analysis of human movement*, June 2006.

[30] D. Crisan and A. Doucet. A survey of convergence results on particle filter methods for practitioners. *IEEE Transactions on Signal Processing*, 50(3) :736–746, March 2002.

[31] F. Daum and J. Huang. Dynamic Quasi-Monte Carlo for nonlinear filters. In *Signal Processing, Sensor Fusion, and Target Recognition XII*, volume 5096 of *Proceedings of SPIE*, pages 267–278, August 2003.

[32] F. Daum and J. Huang. Mysterious computational complexity of particle filters. In *Signal and Data Processing of Small Targets*, volume 4728 of *Proceedings of SPIE*, Bellingham, MA, USA, August 2003.

[33] F. Daum and J. Huang. Nonlinear filtering with Quasi-Monte Carlo methods. In *Signal and Data Processing of Small Targets*, volume 5204 of *Proceedings of SPIE*, pages 458–479, Bellingham, MA, USA, December 2003.

[34] F. Daum and J. Huang. Physics based computational complexity of nonlinear filters. In *Signal and Data Processing of Small Targets*, volume 5428 of *Proceedings of SPIE*, pages 145–153, Bellingham, MA, USA, August 2004.

[35] F. Daum and J. Huang. Quasi-Monte Carlo hybrid particle filters. In *Signal and Data Processing of Small Targets*, volume 5428 of *Proceedings of SPIE*, pages 497–508, Bellingham, MA, USA, August 2004.

[36] F. Davis. *Inside Intuition-What we know about non-verbal communication*. McGraw-Hill Book Co., 1971.

[37] Q. Delamarre and O. Faugeras. 3D articulated models and multi-view tracking with physical forces. *Computer Vision and Image Understanding*, 81(3) :328–357, 2001.

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

[39] J. Deutscher, A. Davison, and I. Reid. Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. In *IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR'01), pages 669–676, Kauaii Marriott, Hawaii, USA, 2001.

[40] J. Deutscher, B. North, B. Bascle, and A. Blake. Tracking trough singularities and discontinuities by random sampling. In *International Conference on Computer Vision (ICCV'99)*, page 1144, Washington, DC, USA, 1999. IEEE Computer Society.

[41] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *International Journal of Computer Vision (IJCV'05)*, 21(3) :185–205, 2005.

[42] L. Devroye. *Non uniform random variate generation*. Springer-Verlag, 1986.

[43] M. Devy, V. Garric, and J. J. Orteu. Camera calibration from multiple views of a 2d object using a global non linear minimization method. In *International Conference on Intelligent Robots and Systems (IROS'97)*, pages 1583–1589, 97.

[44] D. Doermann and D. Mihalcik. Tools and techniques for video performance evaluation. In *15th International Conference on Pattern Recognition (ICPR'00)*, volume 4, pages 167–170, Barcelona, Spain, September 2000.

[45] A. Doucet, N. De Freitas, and N. J. Gordon. *Sequential Monte Carlo Methods in Practice*. Series Statistics For Engineering and Information Science. Springer-Verlag, New York, 2001.

[46] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte-Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3) :197–208, 2000.

[47] A. Erol, G. Bebis, M. Nicolescu, R. Boyle, and X. Twombly. Vision-based hand pose estimation : a review. *Computer Vision and Image Understanding (CVIU'07)*, 108 :52–73, 2007.

[48] K.-T. Fang, Y. Wang, and P. M. Bentler. Some applications of number-theoretic methods in statistics. *Statistical Science*, 9(3) :416–428, 1994.

[49] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini. Learning about objects through action-initial steps towards artificial cognition. In *International Conference on Robotics and Automation (ICRA'03)*, pages 3140–3145, Taipei, TAIWAN, 2003.

[50] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems (RAS'03)*, 42 :143–166, 2003.

[51] M. Fontmarty, T. Germa, B. Burger, L.F. Marin, and S. Knoop. Implementation of human perception algorithms on a mobile robot. In *6^{th} IFAC Symposium on Intelligent Autonomous Vehicles (IAV'07)*, Toulouse, FRANCE, September 2007.

[52] M. Fontmarty, F. Lerasle, and P. Danès. Data fusion within a modified annealed particle filter dedicated to human motion capture. In *International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, November 2007.

[53] M. Fontmarty, F. Lerasle, and P. Danès. [article en révision] Evaluation of particle filter based human motion visual trackers for home environment sur-

veillance. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 2008.

[54] M. Fontmarty, F. Lerasle, and P. Danès. Towards real-time markerless human motion capture from ambiance cameras using an hybrid particle filter. In *IEEE International Conference on Image Processing (ICIP'08)*, San Diego (CA), USA, October 2008.

[55] M. Fontmarty, F. Lerasle, and P. Danès. Une stratégie hybride de filtrage particulaire pour le suivi de mouvement humain depuis un robot mobile. In *16ème congrès francophone AFRIF-AFIA, Reconnaissance des Formes et Intelligence Artificielle (RFIA'08)*, Amiens, FRANCE, January 2008.

[56] 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,FRANCE, June 2007.

[57] F. Gasparini and R. Schettini. Color balancing of digital photos using simple image statistics. *Journal of Pattern Recognition*, 37(6) :1201–1217, June 2004.

[58] D. M. Gavrila. 3D model-based tracking of human in actions : A multi-view approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'96)*, pages 73–80, 1996.

[59] D. M. Gavrila. Multi-feature hierarchical template matching using distance transforms, 1998.

[60] D. M. Gavrila. The visual analysis of human movement : a survey. *Computer Vision and Image Understanding*, 73(1) :82–98, 1999.

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

[62] M. Gleicher and N. Ferrier. Evaluating video-based motion capture. In *Proceedings of Computer Animation*, pages 75–80, Genève, SUISSE, June 2002.

[63] L. Goncalves, E. Di Bernardo, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3D. In *International Conference on Computer Vision (ICCV'95)*, 1995.

[64] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F Radar and Signal Processing*, 140(2) :107–113, April 1993.

[65] J. Gorostiza, R. Barber, A. Khamis, and M. Malfaz. Multimodal human-robot interaction framework for a personal robot. In *Int. Symp. on Robot and Human Interactive Communication (RO-MAN'06)*, pages 39–44, Hatfield, UK, September 2006.

[66] U. Grenander, Y. Chow, and D. M. Keenan. *Hands : A Pattern Theoretical Study of Bilogical Shapes*. Springer-Verlag, New York, 1991.

[67] D. Guo and X. Wang. Quasi-Monte Carlo filtering in nonlinear dynamic systems. *IEEE transactions on signal processing*, 54(6) :2087–2098, June 2006.

[68] A. Gupta, A Mittal, and L.S. Davis. Constraint integration for efficient multiview pose estimation of humans with self-occlusions. *Transactions on Pattern Analysis Machine Intelligence (PAMI'08)*, 30(3) :493–506, March 2008.

[69] I. Gupta, A. Mittal, and L. S. Davis. Constraint integration for multiview pose estimation of humans with self-occlusions. In *International Symposium on 3D data processing visualisation and transmission (3DPVT'06)*, pages 900–907, Chapel Hill, USA, June 2006.

[70] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P.-J. Nordlund. Particle filters for positioning, navigation and tracking. *IEEE Transactions on Signal Processing*, 50(2) :425–437, 2002.

[71] J. Hadamar. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 1902.

[72] B. Han, Y. Zhu, D. Comaniciu, and L. Davis. Kernel-based bayesian filtering for object tracking. In *IEEE Conference on Vision and Pattern Recognition (CVPR'05)*, 2005.

[73] A. Hilton and P. Fua. Modeling people toward vision based understanding of a person shape, appearance and movement. *Computer Vision and Image Understanding*, 81(3) :227–230, 2001.

[74] A. Hilton, P. Fua, and R. Ronfard. Modeling people : vision based understanding of a person shape, appearance, movement and behaviour. *Computer Vision and Image Understanding*, 103(2–3) :87–89, November 2006.

[75] E. Hjelmas. Face detection : a survey. *Computer Vision and Image Understanding*, 83(3) :236–274, 2001.

[76] D. Hogg. A program to see a walking person. *Image Vision Computer*, 1(1) :5–19, 1983.

[77] X.-L. Hu, T. B. Schön, and L. Ljung. A basic convergence result for particle filtering. *IEEE Transactions on Signal Processing*, 2008.

[78] G. Hua, Yang M.-H., and Y. Wu. Learning to estimate human pose with data driven belief propagation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 747–754, June 2005.

[79] H. Huttenrauch, K. Severinson, A. Green, and E. Topp. Investigating spatial relationships in human-robot interaction. In *International Conference on Intelligent Robots and Systems (IROS'06)*, Beijing, China, October 2006.

[80] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *European Conference on Computer Vision (ECCV'96)*, pages 343–356, Cambridge, UK, April 1996.

[81] M. Isard and A. Blake. CONDENSATION – Conditional density propagation for visual tracking. *International Journal on Computer Vision (IJCV'98)*, 29(1) :5–28, 1998.

[82] 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.

[83] Y. Iwashita, R. Kurazume, T. Hasegawa, and K. Hara. Robust motion capture system against target occlusion using fast level set method. In *International Conference on Robotics and Automation (ICRA'06)*, pages 168–174, Orlando (FL), USA, May 2006.

[84] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'98)*, pages 274–280, Fort Collins (CO), USA, June 1999.

[85] 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.

[86] S. Julier and J. Uhlmann. A new extension of the kalman filter to nonlinear systems. In *International Symposium on Aerospace/Defense Sensing, Simulation and Controls*, Orlando, FL, 1997.

[87] I. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In *IEEE Conference on Computer Vision and Pattern Vision (CVPR'96)*, pages 81–87, 1996.

[88] I. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion. *Transactions on Pattern Analysis and Machine Intelligence*, 22(12) :1453–1459, 2000.

[89] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering*, 82 :35–45, 1960.

[90] T. Kanda, H. Ishiguro, M. Imai, and T. Ono. Development and evaluation of interactive humanoid robots. *Proceedings of IEEE*, 92(11) :1839–1850, 2004.

[91] R. Kehl, M. Bray, and L. J. Van Gool. Full body tracking from multiple views using stochastic sampling. In *IEEE Conference on Conputer Vision and Pattern Recognition*, pages 129–136, San Diego (CA), USA, June 2005.

[92] R. Kehl and L. Van Gool. Real-time pointing gesture recognition for an immersive environment. In *Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 577–582, May 2004.

[93] G. Kitagawa. Monte-carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1) :1–25, 1996.

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

[95] D. Knossow. *Analyse et Capture multi-caméras du mouvement humain*. PhD thesis, Institut National Polytechnique de Grenoble, 2007.

[96] A. Kong, J. Liu, and W. Wong. Sequential imputations and bayesian missing data problems. *Journal of the American Statistical Association*, 89(425) :278–288, 1994.

[97] V. Kulyukin, C. Gharpure, J. Nicholson, and S. Pavithran. RFID in robot-assisted indoor navigation for the visually impaired. In *International Conference on Intelligent Robots and Systems (IROS'04)*, pages 1979–1984, Sendai, JAPAN, 2004.

[98] F. Lerasle, G. Rives, and M. Dhome. Tracking of human limbs by multiocular vision. *Computer Vision and Image Understanding*, 75(3) :229–246, 1999.

[99] J. Lichtenauer, Reinders M. J. T., and Hendriks E. A. Influence of the observation likelihood function on particle filtering performance in tracking applications. In *Automatic Face and Gesture Recognition (FGR'04)*, pages 767–772, Seoul, KOREA, May 2004.

[100] J.F. Maas, T. Spexard, J. Fritsch, B. Wrede, and G. Sagerer. BIRON, what's the topic ? a multi-modal topic tracker for improved human-robot interaction. In *International Symposium on Robot and Human Interactive Communication (RO-MAN'06)*, Hatfield, UK, September 2006.

[101] J. MacCormick. *Probabilistic models and stochastic algorithms for visual tracking*. PhD thesis, University of Oxford, 2000.

[102] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *International Conference on Comuter Vision (ICCV'99)*, pages 572–578, 1999.

[103] 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.

[104] S. J. MacKenna and H. Nait-Charif. Tracking human motion using auxiliary particle filters and iterated likelihood weighting. *Image and Vision Computing*, 25(6) :852–862, June 2007.

[105] P. S. Maybeck. *Stochastic models, estimation and control*, volume 1 of *Mathematics in Science and Engineering*. Academic Press, 1979.

[106] P. Menezes. *Multi-cue visual tracking for human robot interaction*. PhD thesis, University of Coimbra, January 2007.

[107] 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.

[108] P. Menezes, F. Lerasle, J. Dias, and R. Chatila. Appearance-based tracking of 3D articulated structures. In *International Symposium on Robotics (ISR'05)*, Tokyo, JAPAN, 2005.

[109] P. Menezes, F. Lerasle, J. Dias, and R. Chatila. A single camera motion capture system dedicated to gestures imitation. In *International Conference on Humanoid Robots (HUMANOID'05)*, pages 430–435, Tsukuba, 2005.

[110] P. Menezes, F. Lerasle, J. Dias, and R. Chatila. Tracking of human limbs by monocular vision. Technical Report 05447, LAAS-CNRS, September 2005.

[111] D. Metaxas, D. Samaras, and J. Oliensis. Using multiple cues for hand tracking and model refinement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03)*, 2003.

[112] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman. Human body model acquisition and tracking using voxel data. *International Journal on Computer Vision*, 53(3) :199–223, July 2003.

[113] T. Moeslund and E. Granum. A survey on computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81 :231–268, 2001.

[114] T. Moeslund, A. Hilton, and V. Krüger. A survey of advanced vision-based human motion capture and analysis. *Computer Vision and Image Understanding (CVIU'06)*, 104 :174–192, December 2006.

[115] J. Mulligan. Upper body pose estimation from stereo and hand face tracking. In *Computer and Robot Vision*, pages 413–420, May 2005.

[116] L. Mündermann, S. Corazza, and T. P. Andriacchi. Accurately measuring human movement using articulated ICP with soft-joint constraints and a repository of articulated models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–6, June 2007.

[117] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision : A survey. In *Trans. on Pattern Analysis Machine Intelligence (PAMI'08)*, 2008.

[118] E Muybridge. *Muybridge's complete human and animal locomotion : all 781 plates from the 1887 animal locomotion*, volume 1. Dover Publications, 1979.

[119] E Muybridge. *Muybridge's complete human and animal locomotion : all 781 plates from the 1887 animal locomotion*, volume 2. Dover Publications, 1979.

[120] P. Noriega. *Modèle du corps humain pour le suivi de geste en monoculaire*. PhD thesis, 2007, Université Pierre et Marie Curie — Paris 6.

[121] P. Noriega and O Bernier. Suivi 3D monoculaire du haut du corps par une propagation des croyances sous contraintes articulaires. In *Reconnaissance de Formes et Intelligence Artificielle*, Amiens, FRANCE, January 2008.

[122] K. Ogawara, X. Li, and K. Ikeuchi. Markerless human motion estimation using articulated deformable model. In *International Conference on Robotics and Automation (ICRA'07)*, pages 46–51, Roma, Italy, 2007.

[123] D. Ormoneit, C. Lemieux, and D.J. Fleet. Lattice particle filters. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (UAI'01)*, pages 395–402, San Francisco, CA, USA, 2001.

[124] J. O'Rourke and N. I. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2(6) :522–536, November 1980.

[125] H. Ouhaddi and P. Horain. 3D hand gesture tracking by model registration. In *Proceedings of International Workshop on Synthetic - Natural Hybrid Coding and Three Dimensional Imaging (IWSNHC3DI'99)*, Santorini, GRÈCE, September 1999.

[126] J. Park, S. Park, and J. K. Aggarwal. Human motion tracking by combining view-based and model-based methods for monocular vision. In *International Conference on Computational Science and its Applications (ICCSA'03)*, pages 650–659, 2003.

[127] V. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction : A review. *Transactions On Pattern Analysis and Machine Intelligence*, 19(7) :677–695, 1997.

[128] P. Pérez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proceedings of IEEE*, 92(3) :495–513, 2004.

[129] P. Pérez, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *European Conference on Computer Vision (ECCV'02)*, pages 661–675, Berlin, 2002.

[130] V. Philomin, R. Duraiswami, and L. S. Davis. Quasi-random sampling for CONDENSATION. In *European Conference on Computer Vision (ECCV'00)*, pages 134–149, Dublin, Ireland, 2000.

[131] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun. Towards robotic assistants in nursing homes : challenges and results. *Robotics and Autonomous Systems (RAS'03)*, 42 :271–281, 2003.

[132] M.K. Pitt and N. Shephard. Filtering via simulation : Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446), 1999.

[133] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose : Tracking people by finding stylized poses. In *IEEE Conference on Vision and Pattern Recognition (CVPR'05)*, pages 271–278, San Diego, USA, June 2005.

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

[135] J. M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *International Conference on Computer Vision (ICCV'95)*, pages 612–617, 1995.

[136] J. Richarz, C. Martin, A. Scheidig, and H.M. Gross. There you go ! - estimating pointing gestures in monocular images for mobile robot instruction. In *International Symposium on Robot and Human Interactive Communication (ROMAN'06)*, pages 546–551, Hartfield, UK, September 2006.

[137] K. Rohr. Incremental recognition of pedestrians from image sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'93)*, pages 8–13, New York City (NY), USA, June 1993.

[138] K. Rohr. Towards model-based recognition of human movements in image sequences. In *Computer Graphics, Vision, Image Processing : Image Understanding*, volume 1, pages 94–115, 1994.

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

[140] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *European Conference on Computer Vision (ECCV'00)*, pages 702–718, Dublin, Ireland, 2000.

[141] R. Siegwart, O. Arras, S. Bouabdallah, D. Burnier, G. Froidevaux, X. Greppin, B. Jensen, A. Lorotte, L. Mayor, M. Meisser, R. Philippsen, R. Piguet, G. Ramel, G. Terrien, and N. Tomatis. Robox at expo 0.2 : a large scale installation of personal robots. *Robotics and Autonomous Systems (RAS'03)*, 42 :203–222, 2003.

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

[143] L. Sigal and M. J. Black. Humaneva : Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical report, Department of Computer Science Brown University, 2006.

[144] E.A. Sisbot, A. Clodic, L.F. Marin, M. Fontmarty, L. Brèthes, and R. Alami. Implementing a human-aware robot system. In *IEEE International Symposium on Robot and Human Interactive Communication 2006 (RO-MAN'06)*, Hatfield, UK, September 2006.

[145] C. Sminchisescu. *Estimation algorithms for ambiguous visual models - Three Dimensional Human Modeling and Motion Reconstruction in Monocular Video Sequences*. PhD thesis, Institut National Politechnique de Grenoble (INRIA), July 2002.

[146] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3D body tracking. In *IEEE Conference on Pattern Vision Recognition, (CVPR'01)*, pages 447–454, Kauaii Marriott, Hawaii, USA, December 2001.

[147] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal on Robotic Research (IJRR'03)*, 6(22) :371–393, May 2003.

[148] A. W. B. Smith and B. C. Lovell. Measurement function design for visual tracking applications. In *International Conference on Pattern Recognition (ICPR'06)*, pages 789–792, 2006.

[149] J. M. Soares, P. Horain, and A. Bideau. Communication gestuelle et télévirtualité : Interaction autour d une application partagée et acquisition des gestes par vision artificielle en temps réel. In *Actes du colloque Compression et REprésentation des Signaux Audiovisuels(CORESA'04)*, pages 187–190, Lille, FRANCE, May 2004.

[150] H. W. Sorenson. *Kalman filtering : theory and application*. IEEE Press, 1985.

[151] M. Splong and M. Vidyasagar. *Robot Dynamics and Control*. John Wiley and Sons, 1989.

[152] B. Stenger, P. R. S. Mendonça, and R. Cipolla. Model-based hand tracking using an unscented kalman filter. In *British Machine Vision Conference (BMVC'01)*, volume 1, pages 63–72, Manchester, UK, September 2001.

[153] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla. Filtering using a tree-based estimator. In *International Conference on Computer Vision (ICCV'03)*, pages 1063–1070, 2003.

[154] R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. Natural human-robot interaction using speech head pose and gestures. In *International Conference on Intelligent Robots and Systems (IROS'04)*, Sendal, Japan, October 2004.

[155] A. Sundaresan and R. Chellappa. Markerless motion capture using multiple cameras. In *Computer Vision for Interactive and Intelligent Environment*, pages 15–26, November 2005.

[156] S. Thrun, M. Beetz, M. Bennewitz, W. Burgard, A.B. Cremers, F. Dellaert, D. Fox, D. Halnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz. Probabilistic algorithms and the interactive museum tour-guide robot MINERVA. *International Journal of Robotics Research (IJRR'00)*, July 2000.

[157] D. Tolani, A. Goswami, and N. I. Badler. Real-time inverse kinematics techniques for anthropomorphic limbs. *Graphical models*, 62(5) :353–388, 2000.

[158] P. Torma and C. Szepesvári. Sequential importance sampling for visual tracking reconsidered. In *AI and Statistics*, pages 198–205, 2003.

[159] J. Triesch and C. Von der Malsburg. A system for person-independent hand posture recognition against complex backgrounds. *Transactions on Pattern Analysis Machine Intelligence (PAMI'01)*, 23(12) :1449–1453, 2001.

[160] R. Urtasun and P. Fua. 3D human body tracking using deterministic temporal motion models. In *European Conference on Computer Vision (ECCV'04)*, pages 92–106, 2004.

[161] R. Van Der Merwe, N. De Freitas, A. Doucet, and E. Wan. The unscented particle filter. In *Advances in Neural Information Processing Systems 13*, November 2001.

[162] R. Van Der Merwe and E. Wan. The square-root unscented kalman filter for state and parameter estimation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, Salt Lake City, Utah, May 2001. IEEE.

[163] 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.

[164] S. Wachter and H. Nagel. Tracking persons in monocular image sequences. *Computer Vision and Image Understanding*, 74(3) :174–192, 1999.

[165] S. Waldherr, S. Thrun, and R. Romero. A gesture-based interface for human-robot interaction. *Autonomous Robots (AR'00)*, 9(2) :151–173, 2000.

[166] P. Wang and J. Rehg. A modular approach to the analysis and evaluation of particle filters for figure tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 790–797, New York, USA, 2006.

[167] http ://www.cs.brown.edu/research/vision/motioncapture — muybridge 3D.

[168] http ://www.ptgrey.com/products/flea2/index.asp — point grey research inc.

[169] http ://vision.cs.brown.edu/humaneva/ — humaneva.

[170] http ://mocap.cs.cmu.edu — CMU Graphics Lab.

[171] http ://www.motionanalysis.com — Motion Analysis Corporation.

[172] http ://www.vicon.com — Motion Capture Systems from Vicon.

[173] http ://viper-toolkit.sourceforge.net/ — VIPER : The video performance evaluation resource.

[174] M. Woo, J. Neider, T. Davis, and D. Shreiner. *OpenGL 1.2 programming guide : the official guide to learning openGL*. Addison-Wesley, 1999.

[175] Y. Yacoob and L. Davis. Learned temporal models of image motion. In *International Conference of Computer Vision*, pages 446–453, 1998.

[176] J. Yang, A. Park, and S.W. Lee. Gesture spotting and recognition for human-robot interaction. *Transactions on Robotics*, 23(2) :256–270, 2007.

[177] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Transactions on information theory*, 51(7) :2282–2312, July 2005.

[178] Xu Zhao and Yuncai Liu. Generative tracking of 3D human motion by hierarchical annealed geneteic algorithm. *International Journal of Pattern Recognition*, 2008.

[179] 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.