Université Claude Bernard Lyon 1

N° ordre : 156 – 2009

École Doctorale Informatique et mathématiques — UFR d'Informatique

Reconstruction 3D à partir de séquences vidéo pour l'acquisition du mouvement de personnages en temps réel et sans marqueur

THÈSE

présentée et soutenue publiquement le 30 Septembre 2009

pour l'obtention du

Doctorat de l'Université de Lyon
délivré par l'Université Claude Bernard – Lyon 1

(spécialité informatique)

par

Brice Michoud

Composition du jury

Président :	M. Roger Mohr	Professeur ENSIMAG, Grenoble
Rapporteurs :	M. Michel Dhome	Directeur de recherche Université Blaise Pascal, Clermont - Ferrand
	M. Pascal Guitton	Professeur Université Bordeaux 1, Bordeaux
Examinateurs :	M. Kadi Bouatouch	Professeur Université de Rennes 1, Rennes
	M. Albert Dipanda	Professeur Université de Bourgogne, Dijon
	Mme. Saïda Bouakaz	Professeur Université Claude Bernard Lyon1, Villeurbanne
	M. Erwan Guillou	Maître de conférences Université Claude Bernard Lyon1, Villeurbanne

Laboratoire d'InfoRmatique en Image et Systèmes d'information — UMR 5205

LIRIS

Résumé

Dans cette thèse, nous nous intéressons à l'acquisition automatique de mouvements 3D de personnes. Cette opération doit être réalisée sans un équipement spécialisé (marqueurs ou habillage spécifique), pour rendre son utilisation générale, sous la contrainte du temps réel. Cette condition est nécessaire pour permettre des applications interactives. Pour répondre à ces questions, nous sommes amenés à traiter de la reconstruction et l'analyse de la forme 3D.

Concernant le problème de reconstruction 3D en temps réel d'entités en mouvement à partir de plusieurs vues, les approches existantes font souvent appel à des calculs complexes incompatibles avec la contrainte du temps réel. Les approches du type *Shape-From-Silhouette* (SFS) offrent un compromis intéressant entre efficacité algorithmique et précision. Ces dernières utilisent les silhouettes issues de chaque caméra pour proposer un volume englobant des objets. Cependant elles nécessitent un environnement particulièrement contraint, dont le placement minutieux des caméras. Les travaux présentés dans ce manuscrit généralisent l'utilisation des approches SFS à des environnements peu contrôlés. Ils s'appuient sur le domaine de visibilité de chaque caméra, et relaxent les contraintes de leur placement. L'introduction de critères géométriques réduit la quantité d'artefacts générés. Enfin une mesure de confiance sur l'existence de chaque point 3D compense les erreurs d'extraction de silhouette.

Les principales méthodes d'acquisition de mouvements sans marqueur s'appuient sur une modélisation paramétrique du corps. L'acquisition du mouvement revient à déterminer les paramètres offrant la meilleure corrélation entre le modèle et la reconstruction 3D. Ceci implique la résolution d'un système de grande dimension et donc des temps de calculs prohibitifs. Notre objectif étant le suivi temps réel, nous proposons des méthodes qui offrent la précision requise et le temps réel. En premier lieu nous présentons deux approches construites sur l'extraction de la structure topologique de la forme 3D. Ces premières approches temps réel, restent sensibles aux erreurs induites par l'utilisation d'un faible nombre de caméras. Pour augmenter la robustesse, nous nous appuyons sur un marquage naturel des extrémités du corps : la peau. Couplé à un suivi temporel par filtre de Kalman, à un recalage d'objets géométriques simples (ellipsoïdes, sphères, etc.), nous proposons un système temps réel, offrant une erreur de l'ordre de 6%. Celui-ci acquiert le mouvement d'une personne en temps réel à partir de deux vues. De par sa robustesse, il permet le suivi simultané de plusieurs personnes, même lors de contacts. Les résultats obtenus ouvrent des perspectives à un transfert vers des applications grand public.

Mots-clés: Acquisition de mouvements sans marqueur, Enveloppe Visuelle, temps réel, reconstruction multi-vues.

Title : 3D Video-based Reconstruction for Realtime and Markerless Motion Capture

Abstract

In this thesis, we aim at automatically capturing 3D motion of persons without markers. To make it flexible, and to consider interactive applications, we address real-time solution, without specialized instrumentation. Real-time body estimation and shape analyze lead to home motion capture application

We begin by addressing the problem of 3D real-time reconstruction of moving objects from multiple views. Existing approaches often involve complex computation methods, making them incompatible with real-time constraints. *Shape-From-Silhouette* (SFS) approaches provide interesting compromise between algorithm efficiency and accuracy. They estimate 3D objects from their silhouettes in each camera. However they require constrained environments and cameras placement. The works presented in this document generalize the use of SFS approaches to uncontrolled environments. They take into account each cameras field of view, which relaxes constraints on their placement. The introduction of a new geometric criterion reduces the amount of generated artifacts. Finally a measure of confidence on the truthfulness of each 3D point offset errors coming from silhouettes extraction.

The main methods of marker-less motion capture, are based on parametric modeling of the human body. The acquisition of movement goal is to determine the parameters that provide the best correlation between the model and the 3D reconstruction. This involves solving a problem of large scale and time calculations prohibitive. As our objective is the real-time monitoring, we have proposed simple and robust real-time methods, providing the accuracy required by the target applications. First we present two approaches built on the topological structure of the 3D shape extraction. This first approach, however, remains sensitive to errors induced by using a small number of cameras. The following approaches, more robust, use natural markings of the body extremities : the skin. Coupled with a temporal Kalman filter, a registration of simple geometric objects, or an ellipsoids' decomposition, we have proposed two real-time approaches, providing a mean error of 5%. The proposed system acquires the motion of a person in real time from two views. Thanks to the approach robustness, it allows the simultaneous monitoring of several people even in contacts. The results obtained open up prospects for a transfer to home applications.

Keywords: Markerless motion capture, Visual Hull, Real-time, Multi-Views Reconstruction.

Préparée

au Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS) UMR 5205, Bâtiment Nautibus, 43, bd du 11 novembre 1918, 69622 Villeurbanne cedex.

Remerciements

Mon jury de thèse a été pour moi un véritable honneur. Je remercie M. Roger MOHR d'avoir présidé ce jury avec enthousiasme, adresse et justesse. Je remercie M. Michel DHOME et M. Pascal GUITTON, dont le volume de mon manuscrit n'a pas entaché la rédaction de rapports particulièrement synthétiques et précis sur les différents travaux réalisés. Je remercie M. Kadi BOUATOUCH et M. Albert DIPANDA qui ont acceptés d'être examinateurs, leurs questionnements ont contribué à l'ouverture d'un débat scientifique passionnant. Je remercie Mme Saïda BOUAKAZ et M. Erwan GUILLOU qui ont su conseiller, diriger, et parfois canaliser mes ardeurs scientifiques, afin de rendre ce travail, dont je suis fier aujourd'hui.

Le travail présenté dans ce manuscrit clôt quatre années passées au laboratoire LIRIS. Ce document ne laisse apparaître que la partie visible de l'iceberg, dont les profondeurs sont soutenues et bâties par un ensemble de collègues que je ne saurai laisser dans l'ombre. Je pense notamment aux membres de l'équipe SAARA, et particulièrement à Erwan pour la confiance qu'il m'a accordé, à son dynamisme scientifique, à sa curiosité naturelle et à son courage pour les maintes relectures du manuscrit, à Saïda pour sa rigueur scientifique et son aide à la formalisation des procédés mis en jeux, à Hector pour son enseignement de la communication scientifique, à Mathieu pour nos discussions passionnées et ses nombreuses relectures, à Behzad pour avoir cru en moi dès mon Master 2, et encore à Martine d'avoir introduit et offert une dimension plastique et artistique à mon travail. Je tiens à remercier Saïd pour ses encouragements et sa confiance. Étant géographiquement proche (mais pas seulement) des membres de l'équipe R3AM, je tiens également à remercier Jean-Claude, Benjamin, Jean-Philippe et François pour nos discussions de passionnés d'image. Je tiens à remercier Amélie pour nos échanges et le partage de longs week-ends de rédaction, rythmés de menus sushis-café. Enfin je remercie Marie, Marianne, ainsi que tous ceux que j'oublie, pour l'environnement de travail dans lequel j'ai pu pleinement m'épanouir. Je tiens également à remercier certains membres de l'UFR d'informatique et particulièrement Samir pour sa confiance et son accompagnement, ainsi qu'Élodie, Florence, Éliane, Sylvain, Alex, Raph et tant d'autres qui on su m'intégrer au sein de leurs équipes pédagogiques et permettre des échanges fructueux.

Mon investissement fort dans ce travail n'aurait pu être couronné de succès sans le réconfort, le dévouement, le courage et encore l'amour de ma compagne, Fanny, qui a toujours su trouver les mots justes, et accepter mon absence quasi-totale de cette dernière année. Je pense bien sûr au soutient que j'ai reçu de toute ma famille (M.'s et C.'s) pour leurs encouragements et leurs relectures expresses et assidues. Je remercie également mes amis qui se sont armé de compréhension face à d'incessants "Bah, heu, non, désolé, demain j'travail ...".

Je suis sorti grandi de ces quatre années de travail intensif, dont la conclusion ne marque que le commencement de ma vie professionnelle ; commencement et fructification dont vous êtes tous, directement ou indirectement, artisans.

Je dédie cette thèse
à celles et ceux qui ont
cru et croient en moi,
à ceux qui ont permis
que je me construise
tel que je suis,
Merci.

Sommaire

Partie II Acquisition de mouvements en temps réel à partir de formes 3D

Chapitre 1

Introduction générale

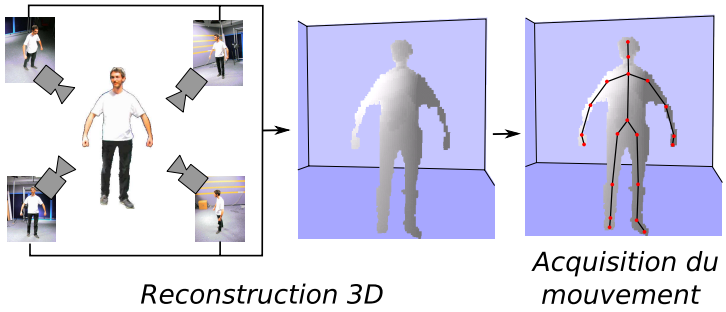


FIG. 1.1 – Présentation du processus d'acquisition de mouvements que nous avons adopté.

Le développement de nouveaux domaines de communication et de moyens d'interaction basés sur l'image, tels que la vidéo conférence, la réalité virtuelle ou la réalité augmentée comme technologie de communication, pour ne parler que des domaines les plus connus, nécessitent l'exploration et la maîtrise de nouveaux outils pour mieux appréhender l'image. Dans ce contexte, l'acquisition et la restitution des mouvements de personnages extraits de scènes réelles sont l'une des étapes clés. Souvent désigné par le terme "suivi de mouvements", selon l'application visée, ce processus peut correspondre à divers niveaux de précision allant de la simple détection de mouvements dans une scène à la "capture" tridimensionnelle du mouvement du personnage et son adaptation à un avatar.

Acquérir le mouvement de personnes revient à déterminer les paramètres qui régissent ce mouvement. Les solutions industrielles qui répondent à ce besoin ont recours à un matériel spécialisé. Les systèmes les plus répandus s'appuient sur des procédés optiques qui nécessitent l'utilisation de caméras spécifiques, associées à des marqueurs fixés aux articulations des personnes en mouvement. Placés sur des zones caractéristiques du corps, ces marqueurs permettent de les localiser facilement dans les images et de calculer leur position 3D. L'utilisation de matériel spécifique implique un coût important et un manque de souplesse, qui limitent le transfert de

ces technologies vers des applications grand public. Résoudre ce problème grâce à la vision par ordinateur et sans recours à une "instrumentation" apporte une souplesse inégalée à l'utilisateur. En l'absence de marqueurs, le problème devient plus difficile.

1.1 Contexte

Dans le cadre de nouvelles interactions homme-machine, nous nous intéressons à la mise en place d'un système temps réel qui doit permettre, à partir de quelques caméras, de fournir la position 3D des articulations de la ou des personnes filmées. Nous constatons le développement de nouvelles modalités d'interaction, entre autres, liées à des applications ludiques. Par exemple la manette $Wii-remote^{TM}$ (Nintendo) permet de récupérer la position 3D de la main du joueur pour interagir avec le jeu, la manette six-axis de la $Playstation3^{TM}$ (Sony) offre des interactions proches, ou encore la caméra $EyeToy^{TM}$ fournit une acquisition simple de mouvements 2D du corps en entier. Ce contexte ouvre de nouvelles perspectives au développement de systèmes d'acquisition de mouvements temps réel. Pour des interactions homme-machine orientées divertissement, et donc grand public, le système doit être léger (une seule machine) et fonctionner en temps réel (au moins trente postures acquises et analysées par seconde). A notre connaissance, il n'existe ni de travail de recherche abouti, ni de solution industrielle, capable de satisfaire ces deux contraintes fortes, sans marqueur ou capteur spécifique. Notre travail s'inscrit dans les demandes actuelles du marché du jeux vidéo, de la domotique, de l'analyse du geste sportif, ainsi que les systèmes de surveillance de la personne.

Les méthodes développés dans cette thèse s'intéressent à l'acquisition du mouvement de personnes sans l'utilisation de marqueurs. Afin de limiter la sensibilité aux auto-occultations, nous proposons de réaliser l'acquisition de mouvements à partir de plusieurs flux vidéos synchronisés. Compte tenu des éléments cités ci-dessus, ce processus doit être réalisé :

- sans marqueur, de façon à rendre son utilisation souple ;
- en temps réel, condition nécessaire pour envisager des applications interactives ;
- sans instrumentation, de façon à minimiser les coûts ;
- avec le minimum d'interventions manuelles.

Nous cherchons à obtenir une méthode stable et robuste à partir d'un faible nombre de caméras, même dans le cas où plusieurs personnes sont filmées simultanément. Pour atteindre ces objectifs, nous avons d'abord été amenés à traiter le problème de l'estimation de la géométrie 3D à partir de plusieurs vues en temps réel. Ensuite nous nous somme intéressés aux problèmes liés à l'analyse de la forme 3D, afin d'en extraire la pose des personnes filmées et d'en réaliser le

suivi au cours du temps. La figure 1.1 illustre les étapes essentielles de ce travail.

1.2 Reconstruction 3D temps réel à partir de plusieurs vues

Les méthodes que nous proposons traitent de la problématique de reconstruction 3D dense en temps réel de personnes en mouvement à partir de plusieurs vues. Les approches existantes font souvent appel à des méthodes de calcul assez complexes, les rendant incompatibles avec la contrainte du temps réel. Les approches du type *Shape-From-Silhouette* (que nous noterons dans la suite SFS) offrent un compromis intéressant entre efficacité algorithmique et bonnes propriétés. Ces méthodes reconstruisent en temps réel une forme 3D englobante des objets, à partir des silhouettes issues de chaque caméra. De par la simplicité de mise en œuvre, ces méthodes sont devenues populaires ; cependant elles souffrent de plusieurs limitations :

- Dans la formalisation actuelle de SFS, les objets ou parties d'objets de la scène sont reconstruits seulement s'ils sont totalement visibles par toutes les caméras. Ceci impose des contraintes rigides quant au placement des caméras. Ces contraintes sont d'autant plus difficiles à satisfaire que l'espace d'acquisition est réduit, et impliquent l'installation du système par un spécialiste.
- SFS génère des artefacts qui ne contiennent aucun objet réel. Leur existence peut induire en erreur l'acquisition et le suivi de mouvements, même s'il est possible d'en réduire le nombre en ajoutant des caméras.
- SFS est sensible à l'extraction erronée de silhouettes. En effet, si l'extraction de silhouettes fournit des résultats satisfaisants dans un environnement totalement contrôlé tel qu'un studio d'acquisition, ces résultats sont moins fiables dans le cas général.

Une partie des travaux menés dans cette thèse généralisent l'utilisation des approches SFS à des environnements peu contrôlés. Notre contribution porte sur une méthode qui se propose de pallier les trois limitations précédentes. Cette méthode s'appuie sur l'apport de chaque silhouette, qui relaxe la contrainte de visibilité totale et les contraintes de placement des caméras. L'introduction de nouveaux critères géométriques permet de réduire fortement la quantité d'artefacts générés. Enfin, nous proposons une mesure de confiance sur l'extraction des silhouettes, pour décider de la reconstruction d'un point 3D à partir de l'information issue de différentes caméras. Nous proposons une évaluation de ces travaux par un ensemble d'expérimentations réalisées à partir de données synthétiques et réelles. Une analyse comparative montre l'efficacité de notre méthode, face aux principales approches existantes.

1.3 Acquisition du mouvement 3D en temps réel

Estimer le mouvement d'une personne revient à retrouver à chaque instant la position 3D de ses articulations. Le second volet de nos travaux, porte sur l'analyse de la forme reconstruite afin d'extraire les positions 3D des articulations de ces personnes. La littérature propose plusieurs méthodes, dont la plupart s'appuient sur une modélisation paramétrique représentée par un squelette articulé du corps humain. Ces paramètres décrivent les positions, orientations ou les angles des articulations. La capture du mouvement revient à estimer les paramètres qui réalisent la meilleure corrélation entre les vues réelles et les vues de synthèses, obtenues par une instantiation du modèle. Cette représentation implique la résolution d'un système de grande dimension, donc des temps de calculs élevés, voire prohibitifs lorsque l'on cherche la précision. Cela amène souvent à faire un choix entre la précision et les temps de calculs de l'application. Comme l'un de nos objectifs est la capture du mouvement temps réel, nous faisons le choix de méthodes simples et robustes, qui respectent la précision requise par l'application visée.

Le processus d'acquisition de mouvements comporte souvent deux étapes : l'initialisation de la pose et le suivi du mouvement. Dans la plupart des méthodes publiées, l'étape d'initialisation est traitée manuellement ou s'appuie sur des poses prédéterminées (pose en T). Cette initialisation donne les paramètres anthropométriques (longueurs des membres), les positions initiales des articulations, etc.

Nous commençons par traiter de l'automatisation de la phase d'initialisation. L'approche que nous proposons s'appuie sur une analyse topologique de la forme 3D à l'aide d'un graphe de Reeb, construit selon une fonction de hauteur.

Une approche simple d'acquisition de mouvements est d'estimer la position 3D des articulations à chaque pas de temps. La méthode développée opère par recalage entre un squelette articulé et la forme 3D. Ce recalage, contraint par des longueurs anthropométriques, s'appuie sur une analyse topologique de la forme 3D. Celle-ci est représentée par un graphe de Reeb barycentrique, invariant par rotation. Cette approche se révèle efficace lorsque la topologie de la forme 3D est équivalente à la topologie du modèle utilisé. Typiquement lorsque les mains sont en contact avec d'autres parties du corps, cette hypothèse n'est plus valable. Afin de rendre cette approche plus robuste, nous utilisons l'information de couleur.

Partant du constat qu'une personne a rarement le visage et les mains couverts, il est alors possible de les localiser dans la forme 3D estimée. Cette identification basée sur la couleur de la peau, associée à des contraintes de longueurs des membres et à la notion de continuité temporelle, apporte une réponse à l'acquisition en temps réel des mouvements d'une personne filmée. Elle

relaxe la contrainte sur le nombre de caméras utilisées ; dans certains cas, deux caméras s'avèrent suffisantes. Cette approche étant principalement construite sur l'information des parties de peau, toute erreur liée à leur extraction influe sur la précision de la capture du mouvement. Afin de répondre à cette fragilité nous nous sommes ensuite orientés vers une approche qui profite à la fois de l'information de peau, et d'un mécanisme de détection des parties saillantes de la géométrie.

Cette dernière approche réalise en premier lieu l'extraction robuste des extrémité du corps, à partir d'une représentation de la forme par ellipsoïdes, et à l'aide des parties de peau. Ces extrémités, ainsi que la continuité temporelle du mouvement, permettent d'extraire les positions des articulations intermédiaires. Cette extraction est réalisée par la segmentation de la forme 3D en régions correspondantes à chaque partie rigide. Cette approche offre une acquisition du mouvement particulièrement robuste, que ce soit face à des mouvements complexes, ou face à la présence de bruit dans la reconstruction 3D, tout en offrant un processus temps réel. En associant l'efficacité de cette dernière méthode, aux mécanismes de suppression d'objets fantômes proposés partie 1, nous avons finalement proposé une approche d'acquisition du mouvement de plusieurs personnes simultanément, résistante aux contacts entre ces personnes.

Le système global proposé profite d'un mécanisme d'estimation de forme 3D temps réel robuste qui génère peu d'artefacts. Cela a pour conséquence de permettre l'acquisition du mouvement d'une personne en temps réel à partir de deux vues. Enfin, l'approche développée propose une généralisation pour l'acquisition en temps réel du mouvement de plusieurs personnes simultanément, même lors de contacts entre ces personnes. Les résultats obtenus montrent une acquisition robuste du mouvement dans un environnement peu contrôlé, et ouvrent des perspectives à un transfert vers des applications grand public.

Avant de présenter nos travaux, nous commençons par présenter les principales méthodes de capture de mouvements.

Mon jury de thèse a été pour moi un véritable honneur. Je remercie M. Roger MOHR d'avoir présidé ce jury avec enthousiasme, adresse et justesse. Je remercie M. Michel DHOME et M. Pascal GUITTON, dont le volume de mon manuscrit n'a pas entaché la rédaction de rapports particulièrement synthétiques et précis sur les différents travaux réalisés. Je remercie M. Kadi BOUATOUCH et M. Albert DIPANDA qui ont acceptés d'être examinateurs, leurs questionnements ont contribué à l'ouverture d'un débat scientifique passionnant. Je remercie Mme Saïda BOUAKAZ et M. Erwan GUILLOU qui ont su conseiller, diriger, et parfois canaliser mes ardeurs scientifiques, afin de rendre ce travail, dont je suis fier aujourd'hui.

Le travail présenté dans ce manuscrit clôt quatre années passées au laboratoire LIRIS. Ce document ne laisse apparaître que la partie visible de l'iceberg, dont les profondeurs sont soutenues et bâties par un ensemble de collègues que je ne saurai laisser dans l'ombre. Je pense

notamment aux membres de l'équipe SAARA, et particulièrement à Erwan pour la confiance qu'il m'a accordé, à son dynamisme scientifique, à sa curiosité naturelle et à son courage pour les maintes relectures du manuscrit, à Saïda pour sa rigueur scientifique et son aide à la formalisation des procédés mis en jeux, à Hector pour son enseignement de la communication scientifique, à Mathieu pour nos discussions passionnées et ses nombreuses relectures, à Behzad pour avoir cru en moi dès mon Master 2, et encore à Martine d'avoir introduit et offert une dimension plastique et artistique à mon travail. Je tiens à remercier Saïd pour ses encouragements et sa confiance. Étant géographiquement proche (mais pas seulement) des membres de l'équipe R3AM, je tiens également à remercier Jean-Claude, Benjamin, Jean-Philippe et François pour nos discussions de passionnés d'image. Je tiens à remercier Amélie pour nos échanges et le partage de longs week-ends de rédaction, rythmés de menus sushis-café. Enfin je remercie Marie, Marianne, ainsi que tous ceux que j'oublie, pour l'environnement de travail dans lequel j'ai pu pleinement m'épanouir. Je tiens également à remercier certains membres de l'UFR d'informatique et particulièrement Samir pour sa confiance et son accompagnement, ainsi qu'Élodie, Florence, Éliane, Sylvain, Alex, Raph et tant d'autres qui on su m'intégrer au sein de leurs équipes pédagogiques et permettre des échanges fructueux.

Mon investissement fort dans ce travail n'aurait pu être couronné de succès sans le réconfort, le dévouement, le courage et encore l'amour de ma compagne, Fanny, qui a toujours su trouver les mots justes, et accepter mon absence quasi-totale de cette dernière année. Je pense bien sûr au soutient que j'ai reçu de toute ma famille (M.'s et C.'s) pour leurs encouragements et leurs relectures expresses et assidues. Je remercie également mes amis qui se sont armé de compréhension face à d'incessants "Bah, heu, non, désolé, demain j'travail ...".

Je suis sorti grandi de ces quatre années de travail intensif, dont la conclusion ne marque que le commencement de ma vie professionnelle ; commencement et fructification dont vous êtes tous, directement ou indirectement, artisans.

Chapitre 2

Capture de mouvements : principales méthodes

Les premiers travaux de capture de mouvements ont été initiés par Muybridge [Muy87] et Marey [Mar87] à la fin du 19$^{\text{ème}}$ siècle. Ces études de la locomotion animale et humaine ont été réalisées en capturant leurs mouvements par photographies successives. En 1973, le psychologue Johansson [Joh73] a conduit l'expérimentation appelée "Moving Light Display" en attachant de petits marqueurs réflecteurs de lumière sur des sujets humains, afin d'étudier la trajectoire de certaines articulations. Ces expérimentations représentent les premiers pas d'un travail de recherche intensif.

Avec l'avènement de la photographie et de la vidéo numérique, l'analyse de mouvements s'est ouverte aux communautés de recherche des domaines de la vision par ordinateur et de l'informatique graphique, avec un intérêt particulier pour le mouvement humain. Ces études alimentent des domaines connexes tels que la création de personnages virtuels à l'apparence et aux mouvement réalistes, utilisés dans la production de jeux vidéo et de films. Pour rendre ces personnages virtuels convaincants, dans leurs apparences, mais aussi dans leurs mouvements, ils doivent avoir un comportement proche du monde réel. Les procédés d'acquisition de mouvements, tentent d'apporter une solution à ce problème, permettant ensuite le transfert de ces mouvements à des personnages virtuels.

Dans le domaine des interactions homme machine, l'un des buts recherchés est de créer des systèmes capables de réagir à la gestuelle de l'utilisateur [PSH97, SWP98, MAS02]. La réussite commerciale des consoles de salon Wii^{TM}, ou encore du périphérique $EyeToy^{TM}$ sont autant d'exemples de la capture du mouvement pour l'interaction homme-machine.

Dans notre travail, nous nous intéressons particulièrement au processus de capture du mou-

vement. Ce processus peut être découpé en deux parties :

- **Estimation du modèle de corps :** L'estimation du modèle de corps est le processus qui détermine automatiquement une représentation de la forme et de la structure cinématique de la personne filmée. Par forme nous désignons la structure géométrique de la personne et par structure cinématique nous désignons à la fois la précision de la structure articulaire ainsi que la longueur des parties rigides (les os).

- **L'acquisition de mouvements :** L'acquisition du mouvement est le processus qui estime les paramètres de cette représentation mathématique, décrivant les mouvements réalisés par une personne (souvent appelée sujet). Cette représentation mathématique contient deux composantes. La première décrit le modèle théorique de la structure cinématique du corps du sujet capturé. La seconde définit les paramètres qui décrivent le mouvement par rapport à la représentation choisie (voir la section 2.3.4). Le rôle des algorithmes d'acquisition de mouvements est d'estimer ces paramètres.

Il existe une grande variété d'approches décrites dans la littérature, dont le but est de répondre à ces deux problèmes. Elles diffèrent par les procédés physiques utilisés afin de collecter les données de mouvements. Certaines approches sont mécaniques, magnétiques ou inertielles, mais les plus répandues sont optiques, qui utilisent des images ou des vidéos. Certaines méthodes nécessitent l'utilisation de structures d'interaction physiques telles que des exo-squelettes, des capteurs de suivis, ou encore des marqueurs optiques [Men99].

Dans la suite nous présentons un état de l'art non exhaustif qui relate les principales méthodes de la littérature liées à l'acquisition du mouvement humain. Pour des études exhaustives, il est possible de se référer aux travaux réalisés par [Gav99, MG01, AC99, GF02, Pop07].

2.1 Systèmes non optiques

Au cours de ces dernières années, plusieurs types de techniques d'acquisition de mouvements ont été développés sur des approches non optiques, en utilisant certaines grandeurs physiques [Men99, MG01]. Certaines approches utilisent des systèmes fixés sur le corps des personnes à suivre et qui ne nécessitent pas d'appareillage extérieur, comme les approches à base d'exo-squelette, ou encore les combinaisons inertielles. Enfin d'autres méthodes nécessitent une mise en place et un environnement plus complexes comme les approches magnétiques.

2.1.1 Approches mécaniques

Les systèmes mécaniques d'acquisition de mouvements suivent directement les angles des articulations de l'acteur. Ce type d'approche est souvent appelé acquisition de mouvements par

exo-squelette. Un technicien attache une structure articulée au corps du sujet. Ainsi lorsque ce dernier bouge, les capteurs situés aux articulations mesurent les mouvements relatifs à chaque articulation de l'acteur. Ce type d'acquisition de mouvements fonctionne en temps réel, de coût raisonnable et sans fils, ce qui permet un espace d'acquisition peu limité. En général un exo-squelette est une structure rigide en métal ou en plastique, dont les articulations contiennent des potentiomètres. Ces structures se révèlent parfois fragiles et peuvent gêner à l'exécution de mouvements complexes ou en contacts avec le sol. De par son prix (de 25 000 euros à 75 000 euros) ce type de méthode s'est particulièrement répandu.

2.1.2 Approches inertielles

Les technologies inertielles sont basées sur l'utilisation de capteurs inertiels miniatures et utilisent des algorithmes de fusion de capteurs. Ces approches ne sont pas contraignantes à l'utilisation et proposent une approche d'acquisition efficace pour un faible coût. Les données du mouvement inertiel sont transmises sans fils à un ordinateur qui réalise la visualisation et le calcul du mouvement réalisé en temps réel. La plupart des approches inertielles utilisent des gyroscopes pour mesurer les rotations. Ces rotations sont ensuite transformées en mouvements articulaires. Ces approches ne nécessitent pas de caméras, d'émetteurs ou encore de marqueurs extérieurs. Les systèmes de capture de mouvements inertiels, estiment les 6 degrés de libertés de chaque articulation du corps en temps réel. Les principaux intérêts de ces approches sont qu'elles ne nécessitent pas de studio d'acquisition et que l'espace d'acquisition n'est limité que par la connexion sans fils entre la combinaison et l'ordinateur. Ces technologies sont particulièrement utilisées dans les industries du jeux vidéo, pour la génération d'animations.

2.1.3 Approches magnétiques

Les systèmes magnétiques calculent la position et orientation des parties du corps de l'acteur, en analysant les flux magnétiques relatifs de trois bobines orthogonales présentes dans les émetteurs et récepteurs. L'intensité relative (ou le courant) de chaque bobine permet de calculer à la fois la quantité et la direction du mouvement. Ces approches permettent d'extraire le mouvement de l'acteur en utilisant seulement les deux-tiers du nombre de marqueurs nécessaires pour les approches optiques à base de marqueur. Ces technologies ne sont pas sensibles à l'occultation des marqueurs magnétiques par des parties du corps de l'acteur. Cependant elles sont sensibles aux bruits électromagnétiques de l'environnement tels que les interférences électriques provenant d'objets métalliques dans la scène, des éclairages, des câbles, des ordinateurs, etc. De plus, les câbles connectés aux marqueurs peuvent gêner la liberté de mouvements du sujet. Enfin l'espace d'acquisition des mouvements est généralement limité, comparé à l'ensemble des approches mécaniques, optiques ou encore inertielles.

(a) (b)

FIG. 2.1 – Illustration d'une scène d'acquisition de mouvements par une approche optique à base de marqueurs, système notamment proposé par la société Vicon [Sys]

2.2 Systèmes optiques à base de marqueurs

Les systèmes commerciaux d'acquisition de mouvements les plus utilisés sont des systèmes optiques à base de marqueurs. Ces approches utilisent le principe des points éclairés en mouvement, initialement proposé par Johansson dans [Joh73]. Des marqueurs optiques, composés d'un matériau hautement réflecteur ou encore de Diodes Electro-Luminescentes, sont placés sur le sujet à suivre. Plusieurs caméras spécifiques à haute vitesse (parfois associées à une source de lumière caractéristique) sont utilisées pour filmer les personnes en mouvement. Le système suit la position des marqueurs dans les flux vidéos et reconstruit leurs trajectoires 3D par une triangulation [Gle00, HFP$^+$00]. Les principales difficultés algorithmiques sont de suivre les marqueurs optiques à travers le temps en évitant les ambiguïtés et de les mettre en correspondance dans toutes les vues [RL02]. Avant de commencer l'acquisition de mouvements, les personnes à capturer prennent une pose particulière appelée *T-pose*, c'est à dire leur corps est vertical avec les bras écartés de façon à former un T. A partir de cette posture, les correspondances entre marqueurs et articulations du modèle cinématique sont réalisées. Ces correspondances et la trajectoire des marqueurs, permettent d'estimer la position et l'orientation de chaque segment du modèle cinématique à chaque pas de temps du système. Le suivi des segments spécifiques peut être transformé en information de rotations pour chaque articulations du squelette d'animation. A cause d'auto-occultations, il arrive que certains marqueurs apparaissent puis disparaissent dans certaines caméras. Il est possible que la mise en correspondance échoue. Il est donc nécessaire de réaliser un post-traitement qui permet de lisser les données et de résoudre les problèmes de correspondance, généralement corrigés manuellement. La figure 2.1(a) présente un système d'acquisition de mouvements optique à base de marqueurs, proposé par la société Vicon [Sys]. La figure 2.1(b) représente l'une des caméras spécifiques, montée d'un anneau lumineux de DEL. Les marqueurs sur le corps reflètent la lumière émise depuis ces différentes sources.

Les approches optiques à base de marqueurs ne sont pas seulement utilisées pour la capture du mouvement humain, mais aussi pour des animaux, des objets, etc. La précision obtenue par ces systèmes atteint généralement une erreur inférieure au millimètre.

2.3 Systèmes optiques sans marqueur

L'estimation du mouvement humain à partir de systèmes optiques sans marqueur est un problème complexe. Inférer la pose d'un humain uniquement à partir d'images est équivalent à un problème de recherche dans l'espace des paramètres d'un modèle d'humanoïde choisi. Il existe certains facteurs qui font que ce problème est difficile.

Premièrement le corps humain est une structure complexe et très articulée. Même les modèles cinématiques simples du corps humain contiennent plus de trente degrés de liberté. Ce fait rend la recherche de solution difficile, surtout lorsque l'on souhaite acquérir les mouvements fins du corps humain. Il est donc nécessaire d'ajouter certaines contraintes au processus de recherche de solution. Les approches les plus populaires se basent sur l'extraction de primitives dans les images, telles que les contours [DC01,FAHF08], l'information de silhouette [CTMS03], qui correspondent à certaines caractéristiques du modèle d'humanoïde choisi. Il est possible d'utiliser des modèles qui décrivent la dynamique du mouvement humain et de réaliser une prédiction sur la posture suivante à partir de la pose courante. La robustesse est généralement augmentée lorsque le modèle dynamique est intégré dans un processus de suivi statistique [DBR00, SBF00, SBS02, FAHF08]. Les contraintes cinématiques et dynamiques du corps humain peuvent aussi permettre de réduire l'espace de recherche [HUF04].

Un second facteur impacte sur les difficultés algorithmiques des approches optiques sans marqueur. La mise en correspondance du modèle d'humanoïde dans l'espace 3D à partir d'informations 2D ne définit pas une relation bijective. En effet différentes configurations du corps peuvent théoriquement fournir le même résultat dans les images, surtout lorsque le nombre de caméras est faible.

Les approches de capture de mouvements sans marqueur présentées dans la littérature peuvent être classées en deux familles. Nous proposons de les différencier par la nature de l'information extraite. Ainsi certaines approches estiment le mouvement 2D dans le plan image et d'autres estiment le mouvement 3D. Dans la suite nous décrivons les stratégies utilisées afin de rendre ce problème moins difficile.

2.3.1 Estimation 2D

Une approche générale de l'analyse de mouvements humain contourne l'approche modèle du corps humain, pour se concentrer sur l'extraction de caractéristiques dans les images 2D. Ces méthodes sont algorithmiquement plus simples que les approches 3D. Elles ont été populaires dans la dernière décennie, principalement grâce à l'augmentation de la puissance de calcul des ordinateurs.

Les approches 2D ont été principalement développées pour le suivi de la main et l'analyse de la gestuelle. Les caractéristiques employées sont la forme, le mouvement et la position des objets d'intérêt dans les images. Des caractéristiques classiques de formes sont les formes x-y [FTOK96] ou encore les moments de Zernike [HSJ95]. Les paramètres de trajectoire du centre des régions ont aussi été utilisés dans [SP95, BD02]. Il est aussi possible de subdiviser les images en grilles régulières et d'utiliser les caractéristiques de chaque bloc pour estimer des caractéristiques spécifiques, par exemple la périodicité du mouvement. Des caractéristiques de ces blocs peuvent être la somme des flots optiques [PN94] ou encore la valeur des pixels [KK96]. Une autre approche se base une modélisation statistique de modèles de forme. Cootes *et al.* ont proposé dans [CTCG95] des modèles appelés *modèles de forme actifs*. Un apprentissage réalisé à partir de configurations connues, définit ainsi les modèles de forme actifs, qui sont ensuite utilisés pour suivre en 2D chaque partie du corps. Une analyse en composantes principales sur la position des primitives caractéristiques a été utilisée pour décrire des formes dans un espace de paramètre de faible dimension. Les modèles ainsi appris sont utilisés pour suivre des objets déformables tels que les mains. Baumberg et Hogg [BH94] utilisent des modèles de forme actifs basés sur des B-splines afin de suivre le mouvement de piétons.

D'autres approches utilisent une information structurelle, ou encore information cinématique du corps humain, afin de proposer des résultats de meilleure qualité. Généralement un modèle 2D explicite de la forme à suivre est recalé sur le mouvement observé dans les images par l'utilisation de différentes caractéristiques. Le type de modèle définit quelles seront les caractéristiques à extraire. Certaines approches utilisent les contours, les régions de l'image ou encore de simples points. Dans l'approche de Geurtz [Geu93], les poses sont inférées en recalant des ellipses 2D par une approche hiérarchique. D'autres approches mettent en correspondance un contour déformable avec l'information de silhouette dans une représentation temporelle [NA94]. Dans [GXT94] Guo *et al.* estiment le mouvement 2D d'humains en recalant un modèle filaire dans la silhouette des personnes filmées.

Les contours caractéristiques correspondants aux bras et aux pieds sont identifiés dans l'approche présentée dans [CH96]. Leung et Young [LY95] utilisent un modèle 2D décrit par un ensemble d'éléments de forme en U pour déterminer les extrémités, le tronc et les autres articulations. Les contours en mouvement sont détectés dans les silhouettes de personnes en mouvement et un modèle est ensuite recalé à l'aide de ces primitives.

Le système *Pfinder* proposé par Wren *et al.* [WADP97] se base sur une approche région. Chaque région de l'image aussi appelée *blob*, est décrite par une méthode statistique à base de distributions gaussiennes associant position et couleur. Chaque blob correspond à une partie

spécifique du corps humain, telles que les mains et la tête. Le suivi de la personne filmée est réalisé par, une prédiction de l'apparence des blobs dans la trame suivante, une association des pixels aux régions et une mise a jour des distributions de chaque blob. Cette approche est l'une des premières méthodes sans marqueur temps réel.

Dans le système W^4, Haritaoglu *et al.* [HHD98] utilisent une différence entre les frames successives, afin d'identifier les personnes en mouvement observées par une caméra. Plusieurs personnes peuvent être suivies en même temps. Le mouvement des différentes parties du corps peut ainsi être capté. Pour cela, une carte 2D du corps est utilisée, dont chaque partie est recalée dans les parties en mouvement correspondantes.

D'autres approches à base d'exemples estiment la pose de personnes en comparant l'image obtenue, avec un base de donnée d'images de poses de référence [SC02, Car00].
Enfin d'autres méthodes modélisent l'apparence des personnes observées et l'utilisent pour suivre chaque partie du corps indépendamment [RF03].

L'ensemble des approches que nous venons de voir focalisent uniquement sur la capture du mouvement en 2D.Ces approches sont souvent utilisées pour la reconnaissance de geste, la surveillance et d'autres applications qui ne nécessitent pas une estimation précise.

2.3.2 Estimation 3D

Une meilleure compréhension du mouvement humain devient possible lorsque les approches estiment les paramètres de mouvements d'un modèle 3D du corps humain. Le modèle cinématique le plus utilisé est constitué d'une chaîne d'os rigides connectés par des articulations. Le squelette ainsi formé, appelé squelette d'animation, est ensuite habillé par différentes primitives géométriques simples, de façon à modéliser les propriétés physiques du corps humain. Les primitives géométriques sont généralement des ellipsoïdes [CKBH00, MTHC03, CH04a, CH04b], des super-quadriques [ST05a, Gav98, KM95] en encore des cylindres [SBF00, GBUP95]. Certaines autres approches plus sophistiquées modélisent le corps humain par une surface implicite, générée à partir d'un ensemble de méta-balles [PF03].

A travers les travaux de la littérature, nous observons qu'une multitude d'approches différentes ont été proposées afin d'associer les paramètres d'un modèle d'humanoïde de façon optimale avec la pose du sujet, à partir d'une ou plusieurs vues.

L'approche *diviser pour régner* a souvent été utilisée à ces fins. Le mouvement de chaque partie rigide du modèle est estimé indépendamment, puis des contraintes mécaniques garan-

tissent les connexions entre ces parties. L'une des premières méthodes utilisant ce schéma a été proposée dans [Sha91]. Dans l'approche proposée par [MZD03], l'information de silhouette est calculée pour chaque vue de la personne à suivre. Chaque silhouette est ensuite automatiquement subdivisée en parties élémentaires identifiées, ainsi chaque partie du corps est suivie séparément.

En 1980, O'Rourke et Balder ont été pionniers en proposant dans [OB80] une approche par propagation de contraintes afin de diminuer l'espace de recherche. Un modèle 3D explicite de la cinématique humaine est utilisé. Une propagation des contraintes basées sur la longueur de chaque partie rigide du modèle cinématique, permet de déterminer la pose courante par un processus itératif. Les méthodes de propagation de contraintes ont aussi été employées dans d'autres travaux, tels que celui proposé par Chen et Lee [CL92].

Dans [OB80] O'Rourke et Balder proposent une architecture générale largement utilisée dans les méthodes actuelles d'acquisition de mouvements sans marqueur. Ce principe décrit l'acquisition de mouvements basée modèle, par la composition d'une étape de prédiction, d'une étape de synthèse, d'une analyse de l'image et d'une estimation de l'état du système. C'est à dire que pour chaque acquisition, le système fournit une prédiction de la posture courante, puis synthétise de nouvelles vues à partir du modèle afin de comparer les images synthétiques avec les image observées. Enfin les paramètres du modèle sont mis à jour en fonction de cette comparaison. Les systèmes d'acquisition diffèrent dans le choix des stratégies algorithmiques utilisées pour chaque étape.

Les méthodes d'*analyse par la synthèse* recherchent l'espace des configurations possibles en comparant des images synthétiques issues des paramètre actuels du modèle, afin de le comparer aux images observées à travers certaines caractéristiques comme le recouvrement, les distances de contours, etc. Les erreurs d'alignement sont ensuite analysées afin de générer les nouveaux paramètres du modèle [dlRG05, GGTS01, Koc93]. Dans la méthode [PF03] de Plankers et Fua, leur modèle à base de méta-balles est mis en correspondance avec le mouvement d'une personne filmée, à partir de l'alignement observé dans les silhouettes et dans les cartes de profondeurs issues d'un reconstruction géométrique de la scène.

Les approches à base de *modèles physiques* estiment les forces à appliquer à un modèle afin qu'il corresponde de façon optimale, à la pose courante de la personne filmée [KM95]. Dans l'algorithme proposé par Delamarre et Faugeras [DF99], les images de silhouette d'un personne filmée par plusieurs caméras sont calculées. Ensuite les forces à appliquer au modèle, sont proportionnelles au décalage observé entre les contours des silhouettes générées par le modèle et les contours des silhouette observées.

D'autres catégories d'approches proposent d'inverser les équations non-linéaires qui associent l'espace des états du modèle d'humanoïde avec l'espace image. Une façon de réaliser cette inversion est d'appliquer les principes de cinématique inverse [YAT00] Ce processus est largement utilisé dans les domaines de la robotique, principalement pour calculer la configuration du modèle de corps qui minimise les différences entre la projections du modèle et les images. La cinématique inverse permet d'approximer linéairement l'inversion des équations de mesures. Bregler et Malik [BM00b] proposent une méthode qui met en correspondance un squelette d'animation habillé de cylindre avec une ou plusieurs vues. La combinaison d'une modélisation statistique des régions dans les images, d'une paramétrisation par rotation et de contraintes sur le flot optique, amène une solution de mise en correspondance itérative. Une extension de cette approche est proposée dans [CRHD00] en ajoutant des contraintes supplémentaires sur la profondeur.

Certaines approches ont été proposées à partir de bibliothèques de mouvements connus. Dans [LESC04] Loy *et al.* proposent d'estimer la posture 3D d'une personne filmée par une caméra. Ainsi chaque image est comparée à une base de données de descripteurs images de mouvements connus. Le mouvement connu le plus proche de l'observable est utilisé pour initialiser un processus itératif de recalage. D'autres approches utilisent ce type de bibliothèque de mouvements que ce soit à partir d'une [OS08, SVD03] ou plusieurs caméras [OSIK06].

D'autres approches proposent l'utilisation de filtres statistiques dans le contexte de l'acquisition de mouvements humain. Toutes ces approches utilisent un modèle de traitement qui décrit la dynamique du corps humain, et d'un modèle de mesure qui décrit comment une image est formée par le corps dans une certaine pose. Le modèle de traitement prédit le vecteur d'état du corps synthétique pour le prochaine image, alors que le modèle de mesures permet d'affiner et de corriger la prédiction proposée, en fonction des nouvelles images. L'un des avantages de ces méthodes est qu'elles modélisent le bruit présent dans les mesures. Dans le cas d'un bruit supposé gaussien et si la dynamique peut être modélisée par un système linéaire, alors un filtre de Kalman permet de réaliser le suivi [MTHC03]. Cependant les caractéristiques dynamiques du corps humain ne sont pas nécessairement linéaires. Le filtrage particulaire, basé sur une approche Bayesienne, peut modéliser ces système non-linéaires [DBR00]. A chaque pas de temps, le filtre à particule propose plusieurs prédictions (de postures) avec les probabilités associées. Ces probabilités sont ensuite mises à jour en fonction des données observables (ressemblance). La performance de ce type d'approche a été démontré à travers plusieurs travaux [DBR00, DC01, MI00, SBF00, SBS02].

D'autres méthodes estiment la géométrie 3D de la scène à partir de plusieurs silhouettes, dans laquelle est recalé un modèle d'humanoïde. Gond *et al.* [GSCD08] proposent une approche construite sur la description du corps humain par un descripteur de forme. Ce descripteur est construit sur le partitionnement de la forme 3D par un cylindre composé de plusieurs cellules

d'intérêt. La mise en correspondance de la description de la reconstruction 3D du sujet, avec une base de mouvements appris, offre l'estimation de la posture. Cheung *et al.* [CKBH00], proposent une méthode temps réel qui estime par un ensemble d'ellipsoïdes la posture du personnage reconstruit par l'approche *Shape-From-Silhouette*. L'estimation de mouvements est cependant peu précise car chaque membre est modélisé par une seule quadrique. Dans [MTHC03], un système pour l'acquisition sans marqueur hors ligne est réalisé à l'aide d'un modèle d'humain plus détaillé. La méthode proposée par Theobalt *et al.* [TMSS02] offre une approche similaire basée sur un schéma multi-couches. L'information de couleur a été ajoutée par [CH04b] afin de rendre cette approche plus robuste. Cheung *et al.* [CBK03a] proposent une approche construite sur une estimation de la scène par *Shape-From-Silhouette*, qui permet à la fois d'estimer automatiquement des longueurs anthropométriques du modèle ainsi que le suivi des articulations à travers le temps.

2.3.3 Estimation de la forme du corps à partir d'information optique

La plupart des système d'acquisition de mouvements optiques à base de marqueurs, proposent une approche algorithmique simple qui permet de définir les paramètres anthropométriques du modèle, de façon à correspondre à la personne filmée [Men99,HFP+00]. Afin de réaliser ce dimensionnement, le sujet à acquérir est invité à prendre une pose caractéristique. La correspondance entre les marqueurs 3D et le modèle détermine les paramètres anthropométriques (longueurs des membres du modèle). Alors que ce processus permet de calculer précisément les longueurs des parties rigides, celui-ci ne fournit pas d'information liée à la forme du corps observé. Allen *et al.* [ACP02] proposent une approche qui capture la forme 3D du haut du corps par l'interpolation de différentes reconstructions issues de données 3D scannées. Un modèle des déformations du corps en fonction des paramètres de pose est proposé dans [SMP03]. Un squelette d'animation de la personne est connu a priori, dont le mouvement est capté par un système optique à base de marqueurs. La déformation de la géométrie du corps est estimée à partir de l'information de silhouette.

Si les marqueurs optiques sont proscrits, alors l'estimation totalement automatique de la forme 3D devient plus difficile. Dans la plupart des approches sans marqueur, le modèle est supposé connu. Ces modèles à priori sont ensuite adaptés aux caractéristique physiques de la personne à capter, parfois par un intervenant extérieur. Certaines approche proposent une méthode totalement automatique, sans utiliser d'information à priori.

Dans les travaux de Cheung *et al.* [CBK03a], le squelette d'animation est estimé à partir d'une séquence de reconstructions par *Shape-From-Silhouette*d'une personne en mouvement. Afin de fournir des résultats intéressants, il reste indispensable que des mouvements spécifiques soient réalisés pour chaque membre séparément. Kakadiaris et Metaxas [KM95] proposent d'estimer le modèle de forme à partir de plusieurs flux vidéos desquels sont extraits les silhouettes de la

personne filmée. Dans [dATM+04], Aguiar *et al.* proposent de déterminer automatiquement les paramètres du squelette d'animation à partir d'une reconstruction 3D de la géométrie de la personne filmée. Cette estimation volumique est calculée par l'approche *Shape-From-Silhouette*puis corrigée par des approches de stéréo-vision. Le schéma proposé utilise une approche qui à partir de l'information spatiale et temporelle, estime précisément les structures articulaires de la personne filmée. Cette méthode n'impose pas de mouvements particulier. Enfin récemment, Aguiar *et al.* estiment la forme 3D de l'individu filmé par plusieurs caméras en recalant un scanner 3D de ce même individu dans chaque frame [dAST+08].

2.3.4 Représentation du squelette d'animation

Une chaîne cinématique régit les mouvement physiologiquement possible des mouvements du corps humain. Cette chaîne cinématique peut être estimée selon plusieurs granulométries, en fonction des besoins applicatifs.

L'une des représentation les plus classique de la chaîne cinématique définit les mouvements humains par une représentation hiérarchique des différentes parties du corps rigides, connectées par des articulations (typiquement de type rotule). Les différences entre les diverses formalisations se situent dans la représentation des articulations. Chaque articulation du squelette, dit squelette d'animation, peut comporter 1, 2 ou 3 degrés de liberté en rotations :

- **Euler :** Il s'agit probablement de la représentation la plus courante pour laquelle un triplet de paramètres définit l'information de rotation. Cette approche est largement utilisée de par sa simplicité [Gav99, DF01, ST03, HUF05]. Cependant cette modélisation souffre de certains effets de bord tels que le blocage de cardan, aussi connu sous le nom de *Gimbal Lock* en anglais. Les angles d'euler peuvent s'apparenter à une généralisation des coordonnées sphériques. Le calcul de la rotation est réalisé dans une ordre bien précis. Ainsi si la rotation demandée transforme l'un des axes de rotation en un autre axe de rotation, le résultat de cette rotation retourne un vecteur nul.

- **Quaternions :** Pour palier à cette singularité de l'approche par angles d'Euler, d'autres approches utilisent les quaternions. Cette représentation est compacte et n'impose pas une forte charge de calcul. Cette modélisation a été utilisée dans [KM95]

- **Axe et angle :** Cette dernière représentation est très utilisée en robotique [MSZ94]. Elle s'appuie sur le fait que toute rotation peut être traduite à chaque instant par un angle de rotation autour d'un axe quelconque. Cette formalisation a été utilisée notamment dans [BMP04].

2.4 Discussion

Les approches monoculaires tentent de déterminer les paramètres du modèle afin de le faire correspondre aux observables de la scène, seulement à partir d'une image de la personne à capter. Les temps d'exécution de ce type de processus dépendent des primitives caractéristiques et du modèle utilisés. Lorsque la silhouette est utilisée, il est possible de capter le mouvement en temps interactif, mais une silhouette particulière peut correspondre à plusieurs poses. De plus il n'est pas rare que certains membres de cette personne soient occultés et ainsi ne soient pas suivis correctement. Enfin en présence de plusieurs personnes, les occultations deviennent encore plus importantes et ces méthodes ne sont en générale pas capables de fournir de résultats corrects.

Les approches multi-caméra permettent de compenser les occultations présentes dans certaines caméras, par l'information apportée par les autres caméras. Ayant préalablement calibré le système, il devient possible de déterminer la posture 3D précise des personnes capturées. Dans le cadre de cette thèse, nous faisons le choix de travailler sur un système multi-caméra, qui amène une robustesse supplémentaire face aux occultations.

La plupart des approches de capture de mouvements optiques sans marqueur ont pour objectifs de fournir un suivi précis des articulations, au détriment du temps de calcul. Ces méthodes tendent de plus en plus à recaler des modèles d'humanoïdes complexes avec les données observées. Ces recalages ou mises en correspondances sont généralement réalisés par l'utilisation de procédures de minimisation, consommatrices de ressources de calcul. A l'opposé, nous développons à travers cette thèse des méthodes dont les objectifs principaux peuvent être ordonnés de la façon suivante :

1. La vitesse d'exécution : Il existe actuellement un compromis entre la vitesse d'exécution et la précision avec laquelle l'estimation de mouvements est effectuée. Les approches temps réel ne peuvent avoir les mêmes résultats que les approches hors ligne. Dans notre contexte d'interaction homme-machine, il est indispensable que le suivi se fasse en temps réel.

2. La précision : La précision est une donnée importante lorsque les applications de la capture de mouvements le nécessitent, comme par exemple pour l'animation de personnages virtuels pour les jeux vidéos ou encore le cinéma d'animation. Dans notre contexte une précision de l'ordre de quelques centimètres s'avère suffisante.

3. Le degré d'automatisation : Nous souhaitons proposer un système en direction du grand public, ainsi il est important que le système soit le plus automatique possible. Certaines approches nécessitent une initialisation semi-automatique voir manuelle. Ainsi lorsque le suivi tombe en échec le système n'est pas capable de se réinitialiser. Dans notre contexte

il est important que le système puisse détecter l'échec du suivi, et soit capable de se réinitialiser automatiquement.

4. Le matériel employé : L'une des limitations majeures des systèmes commerciaux provient de leur prix. Dans notre contexte nous souhaitons proposer un système d'acquisition de mouvements accessible, c'est à dire qu'il ne doit pas nécessiter de matériel spécifique en dehors d'un ordinateur de bureau et de quelques webcams.

L'une des caractéristiques commune à l'ensemble des méthodes d'acquisition de mouvements optiques sans marqueur, réside dans la mise en place de mécanismes qui minimisent l'espace de recherche de solutions. En premier lieu, nous proposons de profiter de l'information géométrique provenant du calibrage des caméras afin de réduire cet espace. La première étape du système consiste à estimer en temps réel le volume 3D des personnes. Ensuite, capter leur mouvement revient à estimer pour chaque pas de temps, les points 3D de cette reconstruction volumique, qui correspondent aux articulations.

Première partie

Reconstruction 3D temps réel à partir de plusieurs vues

Introduction

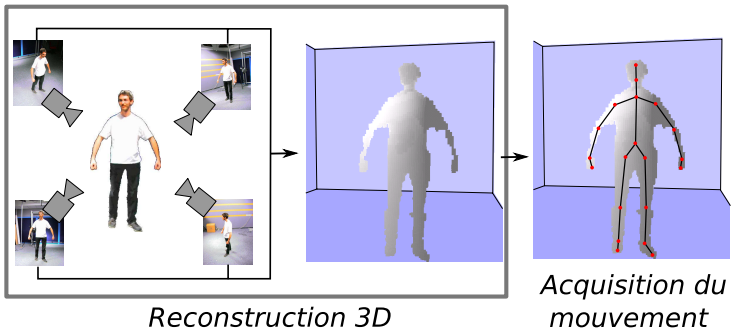


FIG. 1 – Retour sur le schéma du processus de capture de mouvements adopté : étape de reconstruction.

Notre objectif principal est de proposer une méthode d'acquisition sans marqueur et en temps réel, du mouvement de plusieurs personnes. Dans les système commerciaux les plus répandus, des marqueurs sont placés sur des parties caractéristiques du corps à acquérir afin de les repérer facilement dans les images et ainsi calculer leur position 3D précisément. En absence de marqueurs, le problème devient plus difficile. Acquérir le mouvement de personnes revient à déterminer l'information pertinente dans les images issues d'une ou plusieurs vues. Dans cette thèse nous faisons le choix d'extraire des primitives de bas niveaux dans les images 2D pour estimer la forme 3D des personnes filmées, afin d'en extraire la position des articulations directement en 3D (voir Figure 1). Dans cette première partie du document, nous traitons de la problématique de reconstruction 3D en temps réel d'objets articulés à partir de plusieurs vues.

La reconstruction de la forme 3D d'objets à partir de plusieurs images est l'une des problématiques les plus anciennes du domaine de la vision par ordinateur. A notre connaissance, la littérature propose peu de méthodes temps réel, qui estiment précisément la forme 3D d'objets filmés par une ou plusieurs caméras. L'approche *Shape-From-Silhouette* offre un compromis intéressant entre efficacité algorithmique et précision de la reconstruction. Cette méthode estime en temps réel une forme 3D englobant les objets à partir de leurs silhouettes issues de chaque caméras. De par sa simplicité de mise en œuvre, cette approche est devenue particulièrement populaire. Cependant elle souffre de plusieurs limitations. L'approche *Shape-From-Silhouette* classique est capable de reconstruire uniquement les objets entièrement visibles par toutes les caméras, ce qui

impose des contraintes de placement des caméras par rapport aux objets. Comme nous venons de le préciser, *Shape-From-Silhouette* construit une forme englobante des objets d'intérêt. Ainsi cette approche construit des artefacts qui ne contiennent aucun objet réel. La forme 3D calculée par *Shape-From-Silhouette* dépend des silhouettes en entrée. Ainsi toute erreur d'extraction de silhouette influe directement sur la qualité de reconstruction.

Dans la suite, proposons plusieurs contributions qui répondent à chacune de ces limitations, afin de rendre *Shape-From-Silhouette* apte au monde réel. Nous proposons une méthode d'estimation de forme 3D temps réel à partir des silhouettes des objets d'intérêt, qui construit moins d'artefacts que *Shape-From-Silhouette*. De plus cette nouvelle approche n'impose pas de contrainte forte sur le placement des caméras et se révèle robuste aux bruits dans les silhouettes.

Après un état de l'art des principales méthodes de reconstruction de la géométrie à partir d'images, nous présentons en détail l'approche *Shape-From-Silhouette* en identifiant les verrous les plus importants. Nous exposons ensuite nos différentes contributions liées à ces problématiques. Enfin les résultats obtenus soulignent de l'efficacité et la robustesse des méthodes proposées.

Chapitre 3

Principales méthodes de reconstruction 3D optiques

La problématique de la reconstruction géométrique est un domaine de recherche très vaste. Depuis le début des années 70, il a suscité beaucoup d'intérêt aussi bien dans les domaines de la recherche que pour des applications industrielles. Cette discipline est construite sur le développement d'analyse de données numériques provenant de diverses modalités : les approches par contacts physiques et les approches sans contact. Parmi les méthodes sans contact, les caméras numériques sont les capteurs les plus couramment utilisés. L'environnement est ainsi perçu à partir d'images numériques qu'il convient d'analyser. Plusieurs approches à partir de caméras sont apparues : les *approches monoculaires* qui estiment la géométrie de la scène à partir d'une seule vue ; les *approches multi-vues* construites sur l'utilisation de plusieurs vues prises simultanément ; ou encore les approches construites sur l'analyse d'une ou plusieurs vues prises à instants différents. Selon les applications visées, ces techniques œuvrent en lumière ambiante, nécessitent une lumière contrôlée qui peut parfois projeter un motif structuré.

Dans la suite nous revenons sur les principales méthodes d'estimation de la géométrie à partir d'images.

3.1 Méthodes monoculaires

Les approches de reconstruction qui utilisent une seule caméra, s'appuient sur l'extraction de certaines caractéristiques de l'image afin de déterminer l'information de profondeur. Cette information caractéristique peut être l'éclairement (*Shape From Shading*), la déformation de texture (*Shape From Texture*), la variation des paramètres de mise au point de l'objectif utilisé (*Shape from Focus/Defocus*), la détection de points de fuites [GMMB00] ou encore des approches par temps de vol de la lumière (*Caméras à temps de vol*). L'ensemble de ces approches utilisent certaines hypothèses fortes qui permettent de lever l'ambiguïté de profondeur, information en partie perdue lors du processus de formation de l'image.

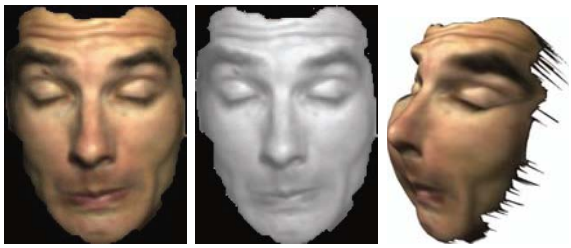


FIG. 3.1 – Résultat de l'approche *Shape From Shading* proposée par Meyer et al. dans [MBB07]. A gauche l'image source, transformée en niveaux de gris (au centre) et la reconstruction correspondante par *Shape From Shading*.

3.1.1 Shape From Shading

L'approche *Shape From Shading* estime la forme d'objets éclairés, à partir des variations graduelles de l'éclairement observé (voir figure 3.1). Ayant une image en niveau de gris, le but est de déterminer la position des sources de lumière, ainsi que la surface de l'objet pour chaque pixel de l'image. Même si le modèle d'éclairement est supposé lambertien, que la direction des sources de lumière est connue et que l'information de luminance peut être décrite en fonction des sources de lumière et de la surface, le problème n'en est pas moins difficile. Si la surface de la forme est décrite par l'information de normale, cela revient à résoudre pour chaque pixel une équation linéaire à trois inconnues. Si la surface est décrite en terme de gradient, cela revient a résoudre une équation non linéaire à deux inconnues. Ainsi trouver une unique solution pour *Shape From Shading* est difficile et il est souvent nécessaire de définir des contraintes supplémentaires [PF05, MBB07]. En dehors des problèmes de leur robustesse de ce type d'approche, les contraintes algorithmiques nécessaires à la résolution de *Shape From Shading* sont encore loin des contraintes temps réel.

3.1.2 Shape From Texture

En général les méthodes d'analyse de l'éclairement ne prennent pas en compte les variations de réflectance de la surface à estimer, c'est-à-dire les changements de couleur dus aux matériaux. Dans le monde réel, les changements de réflectance participent à l'apparence perçue des matériaux : le bois, la fourrure, l'herbe, les éraflures et les bosselures conduisent à des variations de l'éclairement perçu, mais ne relèvent pas du champ d'application des approches *Shape From Shading*. Ce groupe de phénomènes est généralement appelé *texture*. Les approches *Shape from texture* estiment la forme des objets en fonction des variations observées dans la texture [AS88, Gar92, LH05]. Ainsi les objets à construire doivent être texturés que ce soit naturellement ou artificiellement (voir figure 3.2).

FIG. 3.2 – Reconstruction de forme à partir de l'approche Shape From Texture proposée par Loh et Hartley [LH05]. L'image de gauche représente l'image source, la géométrie estimée est représentée au centre, enfin le modèle texturé est présenté sur l'image de droite.

3.1.3 Shape from Focus/Defocus

L'estimation de la profondeur *Depth From Defocus* ou de la forme *Shape From Defocus* consiste à reconstruire la forme 3D à partir de plusieurs images prises depuis une caméra fixe, avec différentes valeurs de mise au point. Avec certaines hypothèses de simplification sur la caméra, ce problème peut être posé comme étant l'inversion d'équations d'intégration qui décrivent le processus de création de l'image. Cette approche utilise le fait que, l'image de la scène sur le plan image, dépend de la radiance de la région, autant que de la forme de cette région. Cependant la valeur de chaque pixel provient de l'intégration d'une radiance inconnue par un noyau de convolution inconnu qui dépend aussi de la forme de la scène. Connaissant la valeur de l'intégrale en chaque pixel il faut alors estimer la valeur du noyau et de la radiance correspondante.

Plusieurs approches ont été proposées pour répondre à ce problème, en se basant sur l'hypothèse que la surface de la scène est localement plane et est parallèle au plan de la caméra (hypothèse de l'*equifocal*) [AFM98, DW90]. Cette hypothèse permet ainsi de décrire l'image observée comme étant une convolution linéaire ; il en revient un compromis fondamental entre la précision et la robustesse au bruit. Pour être résistant aux bruits, certaines approches proposent de calculer l'intégration sur des fenêtres de l'image de taille maximale, cependant pour être en accord avec l'hypothèse de l'*equifocal*, il est nécessaire que ces fenêtres soient les plus petites possibles. En particulier cette hypothèse est violée sur les contours d'occultation. Pour répondre à cette limitation d'autres approches proposent d'utiliser des noyaux de convolution différents en fonction du voisinage du pixel [CR98].

Même si certains travaux ont proposé une approche temps réel, ce type d'approche reste cependant peu robuste aux bruits ambiants et sensible au manque de texture des objets à reconstruire.

3.1.4 Caméras 2,5D à temps de vol

Les caméras 2,5D à temps de vol ont fait leur apparition au début des années 2000. Ce nouveau type de capteur offre un fonctionnement similaire à celui des radars. La différence provient du fait que l'onde envoyée n'est pas une onde acoustique, mais lumineuse.

FIG. 3.3 – Estimation de carte de profondeur par l'approche *Shape From Defocus* proposée par Jin et Favaro dans [JF02]. Les images de gauche et du centre représentent respectivement deux images de la même scène avec des paramètres de mise au point différents. L'image de gauche est produite avec une mise au point au loin, alors que l'image centrale est acquise avec une mise au point courte. L'image de droite représente l'estimation de profondeur correspondante. Notons que la scène est très texturée.

L'appareil se compose d'une source de lumière infrarouge d'amplitude modulée et d'un capteur qui mesure l'intensité rétrodiffusée de la lumière infrarouge. La source infrarouge émet continuellement une lumière dont l'intensité varie de façon sinusoïdale. Ainsi les différents objets de la scène reflètent une intensité de lumière infrarouge dépendante de leur distance à la caméra. Cela provient du fait que des objets à différentes distances de la caméra sont atteints par des parties différentes de l'onde sinusoïdale. La lumière réfléchie par les objets est ensuite comparée à la modulation d'amplitude de référence. Le déphasage entre l'onde émise et l'onde reçue est fonction du temps de vol de la lumière réfléchie par les objets de la scène. En mesurant l'intensité de lumière perçue en chaque pixel, il devient possible de calculer le déphasage correspondant et donc la distance entre les objets de la scène et la caméra.

Ces caméras estiment directement la carte de profondeur de la scène, mais se limitent pour le moment à de faibles résolutions, de l'ordre de 160×120 pixels pour les capteurs actuels, comme par exemple pour la caméra SwissRanger 4000 proposée par la société MESA. Une description détaillée de l'approche par temps de vol est décrite dans [GYB04].

Ces nouveaux capteurs amènent une solution alternative aux approches par stéréo-vision, mais leur coût reste important. De plus ceux-ci ne fournissent qu'une information 2,5D de la scène. Ainsi pour estimer le géométrie 3D, il est nécessaire d'utiliser plusieurs de ces périphériques, en utilisant par exemple la méthode proposée par [BG08]. Son son coût élevé fait que ce type de capteur, quoique prometteur, n'est pas conforme aux contexte que nous nous sommes fixé.

3.2 Approches multi-vues

Les approches monoculaires permettent l'estimation de la géométrie de la scène à partir de l'extraction de primitives depuis un seul point de vue. La présence de conditions environne-

mentales défavorables influe directement sur les résultats de ces approches. Afin de les rendre plus robustes et précises, une solution est de travailler sur l'analyse de plusieurs images prises simultanément depuis plusieurs points de vues.

3.2.1 Stéréo

Le problème de reconstruction de forme à partir de deux images appelée *stéréo-vision* a suscité beaucoup d'intérêt dans les domaines de la vision par ordinateur [KKB+93]. La difficulté principale provient du problème de la mise en correspondance, c'est à dire de trouver les points dans chaque image, qui correspondent à un même point dans la scène 3D [Fau96]. Une fois la correspondance trouvée pour tout couple de points issus des deux caméras, ces points peuvent être triangulés pour déterminer le point 3D de la scène correspondant. La mise en correspondance peut être facilitée en utilisant des contraintes géométriques et en faisant certaines hypothèses sur la scène [MP77, Bak81].

Supposant les paramètres de calibrage connus (voir annexe A), la *contrainte épipolaire* garantit qu'un point dans une image appartient à la droite épipolaire de ce point dans l'autre image.

Avec un objet opaque la *contrainte d'unicité* impose qu'un point dans une image a un unique correspondant dans l'autre image. Cependant en pratique il est possible que les correspondances soient multiples, par exemple lorsque l'objet a une seule couleur. Ainsi cette seconde contrainte n'est pas suffisante pour garantir un mise en correspondance correcte, mais elle peut être utilisée pour vérifier une correspondance calculée par une autre méthode.

Lorsqu'un objet de couleur uniforme doit être reconstruit par une approche stéréo, il peut être intéressant d'utiliser la *contrainte de continuité*, qui suppose que la surface de l'objet est lisse. Des correspondances erronées qui produisent des profondeurs inconsistantes peuvent ainsi être supprimées.

Enfin la *contrainte d'ordre* impose que les correspondances entre pixels des images sont dans le même ordre sur la droite épipolaire de chaque image (excepté lorsque les parties à traiter contiennent des occultations ou différents objets). Ainsi les correspondances pour une seule surface peuvent être vérifiées.

Les associations *denses* résolvent le problème de mise en correspondance, en recherchant plusieurs associations entre les images, de façon a obtenir une carte de profondeur dense. Il existe deux principales méthodes pour estimer une mise en correspondance dense. La première recherche les correspondances pour chaque pixel par des approches d'auto-correlation normalisée : en comparant une fenêtre autour de ce pixel dans l'autre image en utilisant une mesure de similarité pour décider de l'association [Bak81, MMHM02]. Cette première méthode s'avère sensible aux images bruitées et aux différences de conditions d'éclairage pour chaque vue. De plus elle calcule des estimations de profondeur de faible précision pour des régions d'intensité

FIG. 3.4 – Estimation de la carte de profondeur de la scène à partir des images de gauche et du centre. L'image de droite a été calculée en utilisant l'approche de Klaus et al. [KSK06]

similaire. Enfin cette approche se révèle lente à cause du nombre important de comparaisons a réaliser pour chaque paire d'images (une par pixel). Le second type d'approche pour la stéréo dense, tire avantage de caractéristiques dans les images, telles que les régions homogènes, les contours [ZM96], les droites ou encore les coins [JB02]. Ainsi pour que la carte de profondeur soit dense, il est nécessaire que les images contiennent beaucoup de primitives caractéristiques. Une analyse expérimentale des techniques de stéréo vision denses peut être trouvée dans [SZ00]. La figure 3.4 présente l'estimation d'une carte de profondeur par stéréo-vision dense.

Les méthodes que nous venons de citer sont généralement utilisées pour des scènes statiques, ou bien en approche image par image pour des objets dynamiques. La *stéréo-vision dynamique* utilise l'information de mouvements dans les images pour aider à la reconstruction des cartes de profondeur, que ce soit à partir d'une caméra en mouvement, ou encore pour une scène en mouvement. Le mouvement peut être estimé en utilisant le *flot optique* entre les images successives, puis combiné à l'approche stéréo afin d'obtenir une carte de profondeur relative [GT95]. D'autres approches estiment les mouvements de la caméra entre les différents points de vue et ainsi raffinent la géométrie pour chaque nouvelle image [TSJ92, ZM96]. La navigation visuelle est une application typique des approches de stéréo-vision dynamiques, utilisée par des robots afin d'identifier et éviter les obstacles.

En dehors des difficultés liées à la mise en œuvre de telles approches, les méthodes intrinsèques de minimisation et d'optimisation n'offrent pas aujourd'hui, de reconstruction suffisamment dense et robuste pour le temps réel.

3.2.2 Lumière structurée

Pour répondre aux difficultés des approches de stéréo-vision liées à la mise en correspondance, certaines méthodes dites "actives" projettent de l'énergie lumineuse dans la scène. En remplaçant l'une des caméras d'un système de stéréo-vision par un périphérique qui projette un motif connu sur la scène, la mise en correspondance revient à calculer les correspondances entre les pixels et les points du motif projeté [DW05]. La principale difficulté de ce type d'approche provient du choix du motif à projeter, ainsi que de la stratégie de codage de ce motif. En effet la mise en

correspondance dépend de la capacité à décoder les éléments du motif de façon à les localiser dans l'image.

Les méthodes de lumières structurées peuvent être classées en trois catégories [SPB04] : les approches par multiplexage dans le temps, les méthodes par codage direct et les approches intégrant le voisinage spatial :

- Les approches basée sur des stratégies de multiplexage dans le temps sont faciles à implémenter et peuvent produire une reconstruction de très grande précision et à haute résolution. Cependant ce type d'approche n'est pas adapté à le reconstruction d'objets dynamiques.
- D'autres méthodes codent directement l'information spatiale, c'est à dire que chaque sous-motif est caractérisé par exemple par son intensité, ou par sa forme. Ces méthodes offrent une bonne résolution spatiale. Mais leur application reste limitée à des environnements peu bruités et dont les conditions d'éclairage sont contrôlées.
- Les méthodes de la dernière catégorie sont basées sur la projection d'un motif dans lequel est inséré un codage spatial. Le code de chaque composant du motif dépend à la fois de sa valeur et de la valeur de ses voisins. En général cela permet de travailler avec des scènes dynamiques, mais avec une résolution moindre que les deux précédentes classes de méthode [AGD07].

L'information additionnelle du motif projeté permet en général une mise en correspondance plus robuste et rapide que les approches classiques de stéréo-vision. La projection d'un motif dans la scène revient à marquer la scène d'un ensemble de sous-motifs de couleur. Lorsque l'objet à reconstruire est un humain, cette projection peut être une gène pour l'utilisateur. Ainsi d'autres approches projettent une lumière infrarouge non visible par l'œil humain [MLDR07]. Si cette nouvelle solution permet d'éviter le problème d'artefacts, elle ajoute une contrainte de matériel difficile à satisfaire. Transformer un vidéo projecteur pour qu'il projette en infra-rouge est déjà une contrainte, de plus le prix d'achat de ce matériel interdit les applications à domicile pour le grand public.

3.2.3 *Shape-From-Silhouette*

Les approches par stéréo-vision sont construites sur l'extraction de primitives images, ensuite utilisées pour réaliser la mise en correspondance. Afin de fournir un estimation dense de la géométrie de la scène, il est souvent nécessaire d'extraire des caractéristiques de haut niveau, qui pénalisent les temps de reconstruction. De plus l'information obtenue s'apparente à une carte de profondeur 2,5D. Afin d'obtenir une représentation 3D, il est alors nécessaire de recourir à des approches de stéréo-vision à partir de plus de deux caméras, ce qui impacte à nouveaux les temps de calcul. Afin de diminuer la complexité algorithmique, d'autres approches sont construites sur l'extraction de primitives images plus bas niveau : les silhouettes. Celles-ci sont généralement représentées par un masque binaire qui indique pour tout pixel de chaque image, si celui-ci cor-

respond à l'un des objets à reconstruire.

Les méthodes d'estimation de forme 3D à partir de plusieurs silhouettes sont très utilisées dans les environnements multi-caméras, en particulier lorsque cette estimation doit être faite sous la contrainte du temps réel. Ces approches, appelées *Shape-From-Silhouette*, estiment l'enveloppe visuelle [Lau94](*Visual Hull*) des objets d'intérêt, qui est une estimation englobante de leur forme 3D (voir figure 3.5).

A partir d'un ensemble de n vues, l'information de silhouette correspondante à la projection des objets d'intérêt, est extraite des images capturées par une approche dite d'extraction de silhouette (voir la section 4.1). Le *cône de silhouette* d'une caméra est défini par l'ensemble des demi droites issues du centre optique de la caméra à travers les pixels appartenants à la silhouette. L'enveloppe visuelle (EV) est définie par la forme tridimensionnelle générée par l'intersection des *cônes de silhouette* de toutes les caméras [Lau94]. La forme produite fournit un volume englobant des objets d'intérêt. Les approches *Shape-From-Silhouette* ont été utilisées pour diverses applications, telles que la surveillance de foule [YHHGBG03], la modélisation 3D [APSK07] ainsi que l'acquisition de mouvements sans marqueur [CH04a, CBK03a, dATM$^+$04, CKBH00].

Il existe différentes implantations des approches *Shape-From-Silhouette* afin d'estimer l'enveloppe visuelle. Certaines offrent une estimation de l'enveloppe visuelle en temps réel, qu'elle soit volumique [MS03, LCO06, Gra03] ou surfacique [LMS04, MBR$^+$00].

Dans la mesure où l'enveloppe visuelle définit une forme 3D englobante, cette reconstruction s'avère peu précise lorsque le nombre de vues est réduit. D'autres approches proposent l'utilisation d'une information de couleur additionnelle afin de réduire la quantité d'artefacts construits.

Space Carving

Les approches *Shape-From-Silhouette* volumiques construisent un volume qui contient l'ensemble des points des objets d'intérêt. Les approches de *Space Carving* utilisent l'information de couleur afin d'estimer la forme plus précisément. Sous l'hypothèse que tout point de la surface des objets d'intérêt renvoie la même couleur dans toutes les images où il est observé, alors il devient possible d'affiner la reconstruction fournie par une approche *Shape-From-Silhouette*. Il suffit de supprimer les points reconstruits qui ne respectent pas cette hypothèse [SD97, CMS99].

Pour chaque voxel visible (non occulté) un test de validation compare la couleur des pixels sur lesquels ce voxel se projette. Si ces couleurs sont similaires, ce voxel est alors *photo-consistant* et est validé comme appartenant à la surface de l'objet [KS00]. Les voxels qui ne sont pas photo-consistants sont supprimés. Ainsi la liste des voxels visibles est mise à jour et le processus est réitéré jusqu'a ce que tous les voxels restants soient photo-consistants. L'ensemble de ces voxels décrit la *photohull* des objets d'intérêt (voir figure 3.5).

Les approches de *Space Carving* produisent en général une estimation de la forme plus précise

FIG. 3.5 – Comparaison de la reconstruction d'une même scène par les approches *Shape-From-Silhouette* (à gauche), *SpaceCarving* (au centre) et par l'approche temporelle proposée par Goldluecke et Magnor dans [GM04a] (à droite).

que celles proposées par *Shape-From-Silhouette*, au prix d'une forte sensibilité à l'apparence des objets d'intérêt et de calculs beaucoup plus lourds.

Modélisation temporelle

En sus de l'information de silhouette, d'autres approches utilisent la continuité temporelle afin de corriger les formes calculées par les approches *Shape-From-Silhouette*. En utilisant une seule caméra filmant l'objet d'intérêt disposé sur un plateau tournant, il devient possible de construire l'enveloppe visuelle de cet objet à partir d'un grand nombre de vues [Nie94, FCZ98, yKWMC01] Ces approches fournissent une méthode intéressante de reconstruction précise de forme 3D, mais imposent que les objets ne soient pas articulés. La continuité temporelle a été aussi utilisée dans un contexte multi-caméras, afin de réduire les artefacts produits par *Shape-From-Silhouette* [VBK02]. La forme d'objets d'intérêt dynamiques est calculée à chaque trame en utilisant une approche volumique. Ensuite le mouvement des objets est estimé en analysant les reconstructions successives (*sceneflow*). La géométrie est synthétisée à un instant particulier en interpolant l'information issue des trames précédentes, courantes et suivantes. La représentation englobante proposée par les méthodes *Shape-From-Silhouette* est souvent utilisée comme une initialisation des approches de reconstructions temporelles [CBK03b] afin de déterminer le mouvement rigide des objets de la scène et ainsi affiner leur reconstruction au cours du temps. D'autres approches proposent une représentation qui lie l'information spatiale et temporelle. Par exemple Goldluecke et Magnor [GM04b] proposent une méthode qui modélise la scène comme étant une surface dans un espace spatio-temporel, qui est ensuite corrigée en utilisant la photo-consistante à travers la séquence complète. Cette approche fournit des résultats intéressants (voir figure 3.5) mais impose des calculs particulièrement intensifs [GM04a].

3.3 Conclusion

Pour certaines applications, les approches monoculaires fournissent une solution pour estimer la géométrie de la scène lorsque l'on possède un modèle de la scène, ou lorsqu'une lumière

structurée est projetée sur la scène. Par contre lorsque la scène observée n'est pas connue ou lorsque les conditions d'éclairage ne peuvent être contraintes, il est souvent nécessaire d'utiliser plusieurs vues.

Parmi les méthodes multi-vues que nous venons d'exposer, les approches *Shape-From-Silhouette* se distinguent par leur orientation temps réel tout en produisant une estimation 3D de la forme des objets d'intérêt. De plus la propriété selon laquelle la forme reconstruite contient les objets d'intérêt offre la garantie que les positions des articulations sont elles-aussi incluses dans cette représentation. Ces avantages rendent les approches *Shape-From-Silhouette* volumiques aptes aux contraintes que nous nous sommes fixées.
Cependant celles-ci connaissent certaines limitations importantes souvent soulignées dans la littérature, qui peuvent nuire à son utilisation. Elles génèrent des artefacts géométriques, qui ne contiennent pas d'objets réels. De plus, la précision de la reconstruction s'avère dépendante à la qualité de l'extraction de l'information de silhouette.

Dans la suite nous détaillons les approches *Shape-From-Silhouette*. Nous revenons sur la notion d'enveloppe visuelle qui permet d'identifier les principaux verrous de ce type d'approches.

Chapitre 4

Les approches *Shape-From-Silhouette*

Parmi les méthodes qui estiment la géométrie d'objets à partir de plusieurs vues, les méthodes de la classe *Shape-From-Silhouette* (SFS) estiment la *forme* des objets filmés *à partir* de l'information de *silhouette*.

Baumgart a été le premier en 1974 à utiliser l'information de silhouette d'un objet pour en estimer la forme 3D [Bau74]. Dans sa thèse, Baumgart estime les surfaces polyédriques 3D de différents objets à partir de quatre vues. Depuis plusieurs variantes ont été proposées. Par exemple Cheung *et al.* [CKBH00] utilisent une représentation volumique pour décrire la forme 3D estimée. Potmesil [Pot87] et Noborio *et al.* [NFA88] suggèrent l'utilisation d'arbres octaux (octree) afin de diminuer les temps de calcul. Shanmukh et Pujari [SP91] proposent une méthode qui permet de déterminer les positions et orientations optimales des caméras pour *Shape-From-Silhouette*. Szeliski a proposé à partir d'un plateau tournant et une seule caméra, la modélisation 3D la forme d'un objet [Sze93]. En résumé, les approches SFS sont au fil du temps devenues des méthodes de reconstruction 3D populaires de par leur simplicité algorithmique et leur facilité d'implémentation.

Estimer des formes 3D à partir de silhouette présente plusieurs intérêts. Les silhouettes sont relativement simples à estimer, surtout dans des environnement intérieurs avec des caméras fixes et en présence de peu d'ombres. L'implémentation des méthodes SFS est simple, comparée aux approches de stéréo-vision classiques. De plus, il est garanti que la forme générée contient les objets d'intérêt (section 4.3). Cette propriété fondamentale a permis l'application des approches SFS en robotique pour éviter des obstacles et manipuler des objets. Ces avantages ont permis de résoudre divers problèmes issus du domaine de la vision par ordinateur. Ceci inclut les applications de modélisation, de suivi de mouvements, de surveillance, de la génération de vidéos 3D, etc.

Dans ce chapitre, nous présentons le principe général des approches SFS construites sur le

concept d'*Enveloppe Visuelle* (EV). L'EV est une formalisation de la forme reconstruite par les méthodes SFS. Cette caractéristique commune à l'ensemble des approches de la classe *Shape-From-Silhouette* explique l'abus de langage selon lequel *Shape-From-Silhouette* décrit l'approche qui estime l'enveloppe visuelle d'objets d'intérêt.

4.1 La silhouette et son extraction

Lorsque l'on souhaite reconstruire un ou des objets 3D (dit d'intérêt) appartenant à une scène, il est nécessaire de les distinguer dans les images. Dans chaque image nous souhaitons déterminer les pixels correspondants à ces objets d'intérêt du reste de la scène. L'ensemble de ces pixels, forment la **silhouette** de ces objets. Les pixels qui ne sont pas labelisés silhouettes sont appelés pixels de **fond**.

La silhouette des objets depuis un point de vue peut-être définie par la projection de tous les points 3D de ces objets sur le plan image 2D. Dans la suite, $\mathcal{O}$ désigne l'ensemble des points 3D qui appartiennent aux objets d'intérêt et $\text{Proj}_{\pi_i}(E)$ la projection de E, un point 3D ou un ensemble de points 3D, sur le plan image π_i de la caméra $\mathcal{C}_i$.

$\mathcal{S}_i$ la silhouette de $\mathcal{O}$ vu depuis la caméra $\mathcal{C}_i$ est définie par :

$$\mathcal{S}_i = \{\text{p} \in \pi_i : \exists \text{P} \in \mathcal{O} \text{ et } \text{p} = \text{Proj}_{\pi_i}(\text{P})\} \tag{4.1}$$

Nous rappelons que l'un de nos objectifs est d'estimer la forme 3D d'objets d'intérêt en temps réel. Ainsi l'extraction de silhouette nécessite d'être précise avec une forte contrainte temps réel.

Les travaux liés à cette problématique peuvent être classés en deux grandes familles : les approches monoculaires et les approches multi-caméras. Dans la suite nous nous focaliserons sur les approches monoculaires, plus enclines au respect de la contrainte du temps réel. En effet les approches multi-vues nécessitent généralement de lourds calculs [LWB07, TIKM07].

Les méthodes d'extraction de silhouettes monoculaires varient à la fois sur la nature des primitives extraites, ainsi que sur des hypothèses sous-jacentes. Les hypothèses les plus communes sont :

- le fond demeure statique au cours du temps,
- les caméras sont supposées statiques : leurs positions et paramètres ne seront pas modifiés au cours du temps,
- l'éclairage est supposé constant.

L'extraction de silhouette est un problème difficile qui continue à mobiliser d'importants travaux de recherche. Parmi les méthodes proposées, nous distinguons trois types d'approches : les méthodes basées différence de pixels, les méthodes basées régions et les méthodes de type contours.

4.1.1 Approches par différence de pixels

L'incrustation de personnages dans un fond hôte constitue l'une des premières applications d'extraction de silhouette pour les domaines de la télévision, par exemple l'incrustation d'un présentateur devant une carte météorologique, ou encore pour les effets spéciaux dans le cinéma. Les méthodes d'incrustation sont principalement construites sur une extraction de silhouette par clé chromatique (*Chroma-Keying*) : l'objet à segmenter est filmé sur un fond d'une couleur unie caractéristique, afin de pourvoir détecter dans les images les pixels correspondants soit au fond, soit à l'objet. Les couleurs clés les plus répandues sont le bleu et le vert [SB96] afin de limiter les ambiguïtés avec les couleurs présentes sur l'objet ou sur la personne à incruster. Ainsi l'image de silhouette est estimée par l'ensemble des pixels de l'image issue de la caméra, dont la différence avec la couleur clé est jugée importante. Ces approches proposent une extraction robuste de la silhouette. Elles sont particulièrement bien adaptées au temps réel dans la mesure où le nombre d'opérations élémentaires est faible. Néanmoins l'environnement nécessite d'être totalement contrôlé (fond contraint et conditions d'éclairage fixes).

Lorsque qu'un fond uniforme connu n'est pas envisageable, d'autres méthodes modélisent le fond de la scène. Ces méthodes nécessitent préalablement l'acquisition d'une séquence vidéo de la scène supposée statique, privée des objets d'intérêt. Cette première étape permet de modéliser pour chaque pixel une valeur de référence correspondant au fond. Cela suppose que les paramètres géométriques des caméras (voir annexe A) soient constants. Afin de réaliser l'extraction de silhouette, en chaque chaque pixel, la valeur observée est comparée à la valeur de référence. Si cette *différence* dépasse un seuil prédéfini, le pixel correspondant est alors classé en tant que silhouette. Sinon celui-ci est labellisé comme correspondant au fond. Nous obtenons ainsi une *carte binaire de silhouette* pour chaque image.

Les méthodes construites sur une modélisation du fond varient aussi bien sur la méthode d'apprentissage que sur la distance colorimétrique utilisée.

L'hypothèse d'un bruit gaussien est souvent formulée afin de modéliser les erreurs commises lors de l'observation. La couleur du fond observée est supposée soumise à un bruit capteur dont la distribution est gaussienne. Par exemple les caractéristiques de cette gaussienne peuvent être estimées à partir d'une séquence d'images observées d'une scène fixe. Plusieurs méthodes utilisent cette approche [WADP97, CKBH00]. Ces approches fournissent une caractérisation unimodale des pixels du fond. Ainsi ces approches ne modélisent pas les changement, même mineurs, qui peuvent intervenir dans la scène. De par leur approche probabiliste, ces méthodes peuvent fournir une *carte de probabilité* qui décrit l'appartenance de chaque pixel à la silhouette. Pour contourner certains défaut de la méthode précédente, des approches multimodales proposent de modéliser le fond par le mélange de gaussiennes [SG99, Ziv04], ou encore par noyaux non paramétriques [EDH+02]. Certaines approches proposent une mise à jour dynamique du modèle de fond au cours du temps [PSF08]. Ce type d'approche permet une extraction plus robuste aux petites variations de l'environnement, tout en restant accessible au temps réel.

Les approches d'extraction de silhouette par différence de valeur présentent l'avantage de la simplicité. Cependant ces approches ne prennent pas en compte l'information spatiale, c'est à dire que la classification de chaque pixel ne prend pas en compte l'état des pixels voisins. Ces approches font souvent apparaître des classifications bruitées, spatialement incohérentes, ainsi que de petites régions aberrantes qui se traduisent généralement par la création de trous dans la silhouette ou la détection de parties de fond en tant que silhouette.

4.1.2 Approches région

Certaines approches d'extraction de silhouette tentent de combler ces déficiences à l'aide d'heuristiques de remplissage de régions dans les images, en propageant des critères locaux tels que la présence d'un faible gradient au sein d'une région et d'un fort gradient au frontières [TKBM99]. Si il s'agit de propager localement de l'information sur le gradient (opposé aux approches qui offrent une minimisation globale), cela peut aussi propager une erreur locale à toute une région, pouvant même aggraver l'erreur de segmentation dans certains cas. La mise en place d'un schéma de minimisation d'une fonction de coût sur une région, construite par exemple sur le gradient ou la différence de couleur, permet de palier à ce type de défauts locaux [RKB04]. Ces approches se révèlent en général coûteuses en temps de calcul, surtout si l'on souhaite acquérir plus de soixante images par seconde, condition nécessaire à l'acquisition et au suivi de mouvements rapides.

4.1.3 Approches contour

Certaines approches d'extraction de silhouette se focalisent sur la propriété selon laquelle les contours de la silhouette coïncident avec les contours d'occultation. Les approches contour s'intéressent d'avantage aux contours dans les images qu'aux régions qu'ils renferment. Ces méthodes consistent en général à formuler le problème d'extraction de silhouette comme un problème de minimisation d'une fonction d'énergie calculée sur le contour. Cette fonction peut aussi bien prendre en compte la présence de pics de gradient, que l'aspect des régions intérieures et extérieures à la silhouette. Elles se basent principalement sur des critères locaux tels que la couleur ou la texture. Cette formulation du problème appelle des solutions de type variationnelles, comme par exemple les contours actifs [KWT88] qui optimisent le placement d'un contour continu de topologie fixe, dans l'espace image. D'autre approches de type level-set représentent le contour comme un courbe de niveau d'une fonction définie dans le plan image. Ces approches amènent une plus grande flexibilité que les contours actifs, car elles permettent une variation de la topologie de la silhouette au cours du temps [RKPL07]. Le principal inconvénient des approches contour est lié à l'initialisation du système, généralement manuelle, par une solution proche de l'optimal. En effet dans le cas contraire ces approches risquent de converger ver une minimum local. Enfin même si certaines approches par contours actifs sont proches du temps réel, les approches variationnelles restent coûteuses et sont généralement réservées au traitement hors-ligne.

Parmi les différentes approches de soustraction de fond, les approches pixel construites sur une modélisation multi-modale du fond représentent un bon compromis entre qualité d'extraction et temps de calcul. Pour cette raison nous avons choisi l'approche proposée par Stauffer et Grimson dans [SG99]. Cette approche étant une méthode par raisonnement pixel, il n'y a pas de prise en compte explicite de l'information spatiale. Nous filtrons au préalable les images d'entrée par un filtre médian afin de réduire le bruit issu du capteur. Afin d'ajouter un critère de cohérence spatiale, l'extraction de silhouette est ensuite corrigée par l'application des opérateurs d'ouverture puis fermeture morphologiques.

4.2 Principe général de *Shape-From-Silhouette*

Nous considérerons la silhouette d'objets comme étant pour une caméra donnée, l'image de la projection de ces objets (voir équation 4.1). Si l'on considère le cône généralisé, défini par l'ensemble des demi-droites issue du centre optique de la caméra et qui intersectent la silhouette, alors ce cône contient par construction les objets d'intérêt (voir Figure 4.1). Dans la suite nous nommerons ce cône généralisé *Cône de silhouette*.

Si l'on considère les objets observés par n caméras, alors ceux-ci sont inclus dans l'intersection des n *cônes de silhouettes*. En s'appuyant sur cette propriété, l'approche *Shape-From-Silhouette* estime la forme 3D des objets d'intérêt par l'intersection des *cônes de silhouettes* relatifs aux n caméras.

Définition 1 *Soit* X *un point 3D et soit* $\mathcal{S}_i$ *la silhouette des objets d'intérêt issue de la caméra* $\mathcal{C}_i$*. On dit que le point* X *est consistant avec* $\mathcal{S}_i$ *si et seulement si* $\mathrm{Proj}_{\pi_i}(\mathrm{X}) \in \mathcal{S}_i$*.* □

Les points 3D qui sont consistants avec $\mathcal{S}_i$ sont par construction, présents dans le cône de silhouette de $\mathcal{C}_i$.

Définition 2 *Le point 3D* X *est silhouette-consistant si et seulement, pour tout* $i \in [1, \cdots, n]$ *avec* n *le nombre de caméras,* X *est consistant avec* $\mathcal{S}_i$*.* □

Si X est *silhouette-consistant*, alors X a une image dans chaque silhouette.

4.3 *Shape-From-Silhouette*, estimateur de l'enveloppe visuelle

Le terme d'*Enveloppe Visuelle* (EV) est la traduction de la notion de *Visual Hull* (VH), introduite en 1992 par Aldo Laurentini [Lau92]. L'EV désigne la forme calculée par *Shape-From-Silhouette* comme étant le volume qui résulte de l'intersection des cônes de silhouettes.

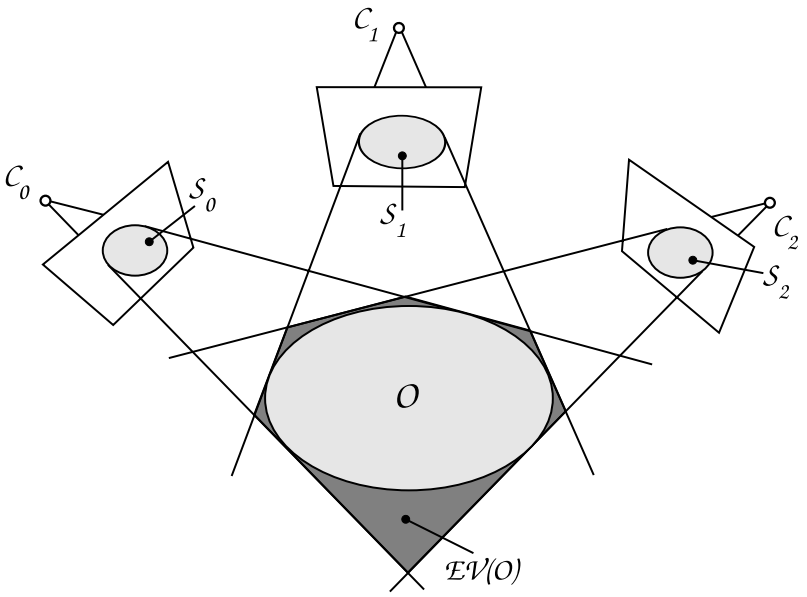


FIG. 4.1 – Représentation d'une coupe 2D de l'intersection des *cônes de silhouette* issus de plusieurs caméras. *Shape-From-Silhouette* calcule l'intersection des *cônes de silhouette* comme étant une estimation 3D de l'objet filmé. La forme estimée est appelée enveloppe visuelle de O .

.

Laurentini a étudié les aspects théoriques des *EV* d'objets 3D polyédriques [Lau94, Lau95] et d'objets courbes [Lau99, BL04]. Nous rappelons ci-dessous la définition de l'*EV* proposée par Laurentini dans [Lau94] :

Définition 3 *Soit* $\mathrm{EV}(\mathcal{O})$ *l'enveloppe visuelle d'un objet* $\mathcal{O}$ *relative à un ensemble de n points de vue. Pour tout point* $\mathrm{P} \in \mathrm{EV}(\mathcal{O})$ *et pour chaque caméra* $\mathcal{C}_i$ *de centre* C_i*, la demi-droite partant de* C_i *et passant par* P *contient au moins un point* Q *de* $\mathcal{O}$ *:*

$$\mathrm{EV}(\mathcal{O}) = \{\mathrm{P} : \forall i \in [1, \cdots, n], \exists \mathrm{Q} \in \mathcal{O}, \mathrm{Q} \in [\mathrm{C}_i, \mathrm{P})\}$$

□

Remarque 1 *L'enveloppe visuelle est le polyèdre englobant résultant de l'intersection des différents cônes de silhouettes associés aux caméras. Ce concept considère des caméras à champ de vue infini, ce qui implique que tous les objets d'intérêt sont toujours dans les champs de vision de toutes les caméras.* □

Alors l'ensemble des points O des objets d'intérêt, est contenu dans l'enveloppe visuelle de ces mêmes objets :

Propriété 1

$$\mathcal{O} \subseteq \mathrm{EV}(\mathcal{O})$$

□

Preuve Tout point $\mathrm{Q} \in \mathcal{O}$ satisfait à la définition 3 :

$$\forall \mathrm{Q} \in \mathcal{O} \text{ et } \forall i \in [1, \cdots, n], \mathrm{Q} \in [\mathrm{C}_i, \mathrm{Q}))$$

Avec n le nombre de caméras. Alors $\forall \mathrm{Q} \in \mathcal{O}, \mathrm{Q} \in \mathrm{EV}(\mathcal{O})$. □

La conséquence de cette propriété est particulièrement intéressante : un point $\mathrm{P} \notin \mathrm{EV}(\mathcal{O})$ ne peut appartenir à $\mathcal{O}$.

La notion d'*EV* peut également être définie à partir de la notion de *silhouette-equivalence*.

Définition 4 *Deux objets d'ensembles de points 3D respectifs $\mathcal{O}_1$ et $\mathcal{O}_2$ sont dits silhouette-équivalent par rapport aux caméras $\mathcal{C}_i$ avec $i \in [1, \cdots, n]$ si et seulement si*

$$\forall i \in [1, \cdots, n], \mathrm{Proj}_{\pi_i}(\mathcal{O}_1) = \mathrm{Proj}_{\pi_i}(\mathcal{O}_2)$$

□

Il en résulte la propriété suivante :

Propriété 2 $\mathrm{EV}(\mathcal{O})$ *est le volume maximal qui est silhouette-équivalent avec $\mathcal{O}$.* □

Preuve

- $\mathrm{EV}(\mathcal{O})$ est silhouette-équivalent avec $\mathcal{O}$:
  d'après la définition 3 la projection de tout point $\mathrm{P} \in \mathrm{EV}(\mathcal{O})$ dans chaque caméra $\mathcal{C}_i$ appartient à la silhouette $\mathcal{S}_i$. De plus la projection de tout point $\mathrm{Q} \in \mathcal{O}$ dans chaque caméra $\mathcal{C}_i$ appartient à la silhouette de $\mathrm{EV}(\mathcal{O})$ car $\mathcal{O} \subseteq \mathrm{EV}(\mathcal{O})$.
- $\mathrm{EV}(\mathcal{O})$ est le volume maximal silhouette-équivalent de $\mathcal{O}$:
  soit $\mathrm{P}' \notin \mathrm{EV}(\mathcal{O})$, il existe au moins une caméra $\mathcal{C}_j$ pour laquelle $[\mathrm{C}_j, \mathrm{P}')$ n'intersecte pas $\mathcal{O}$; donc la projection de P' sur le plan π_j dans la caméra $\mathcal{C}_j$ n'appartient pas à $\mathcal{S}_j$.

□

Les propriétés 1 et 2 sont considérées comme fondamentales. $\mathrm{EV}(\mathcal{O})$ délimite la portion de l'espace qui contient $\mathcal{O}$. De plus, ces propriétés offrent une méthode simple pour estimer l'EV.

L'EV calculée à partir d'un sous-ensemble de points de vue, est majorante de l'EV calculée sur l'ensemble de tous les points de vue.

Propriété 3 *Soit V un ensemble de n caméras, soit $V' \subset V$ un sous-ensemble de V et soient* $\mathrm{EV}(\mathcal{O})_V$ *et* $\mathrm{EV}(\mathcal{O})_{V'}$ *les enveloppes visuelles correspondant respectivement à V et V'. On a*

$$\mathrm{EV}(\mathcal{O})_V \subseteq \mathrm{EV}(\mathcal{O})_{V'}$$

□

Preuve Par définition de l'enveloppe visuelle, tout point $\mathrm{P} \in \mathrm{EV}(\mathcal{O})_V$ et pour toute caméra $\mathcal{C}_i \in V$, $\exists \mathrm{Q} \in \mathcal{O}, \mathrm{Q} \in [\mathrm{C}_i, \mathrm{P})$.
Soit $\mathrm{P} \in \mathrm{EV}(\mathcal{O})'_V$, ainsi $\forall \mathcal{C}_i \in V, \exists \mathrm{Q} \in \mathcal{O}, \mathrm{Q} \in [\mathrm{C}_i, \mathrm{P})$. Si $V' \subseteq V$, alors $\forall \mathcal{C}_i \in V', \mathcal{C}_i \in V$ et $\mathrm{P} \in \mathrm{EV}(\mathcal{O})'_V$.

On en déduit que $\forall \mathrm{P} \in \mathrm{EV}(\mathcal{O})_V, \mathrm{P} \in \mathrm{EV}(\mathcal{O})_{V'}$, ainsi $\mathrm{EV}(\mathcal{O})_V \subseteq \mathrm{EV}(\mathcal{O})_{V'}$. □

La propriété 3 offre un critère de précision de l'estimation de $\mathcal{O}$ par $\mathrm{EV}(\mathcal{O})$. Cette précision est fonction du nombre de caméras utilisées.

Comme nous venons de le voir, les silhouettes fournissent un moyen efficace pour estimer la forme des objets d'intérêt par leur enveloppe visuelle. La définition de l'enveloppe visuelle est construite sur un modèle de caméra dont le plan de projection est supposé infini. Or une caméra produit une image $\mathcal{I}_i$, partie fermée bornée du plan de projection. Dans la suite nous distinguerons deux types de silhouettes : les silhouettes "entières" $\mathcal{S}_i$ et les silhouettes réelles issues des caméras $\tilde{\mathcal{S}}_i$ avec

$$\tilde{\mathcal{S}}_i = \mathcal{S}_i \cap \mathcal{I}_i \tag{4.2}$$

C'est à dire que $\tilde{\mathcal{S}}_i$ est la silhouette de $\mathcal{O}$ visible par la caméra $\mathcal{C}_i$. Elle correspond à la portion de la silhouette $\mathcal{S}_i$ bornée par le champ de vision de $\mathcal{C}_i$.

4.4 Les approches surfaciques et volumiques

Les méthodes *Shape-From-Silhouette* calculent l'enveloppe visuelle des objets d'intérêt à partir de plusieurs images de silhouettes principalement selon deux axes. Les approches surfaciques estiment explicitement la surface de l'enveloppe visuelle, en calculant la surface de l'intersection des cônes de silhouettes. Les méthodes volumiques estiment le volume de l'EV. Elles sont principalement basées sur la propriété 2. L'EV est estimée en calculant l'ensemble de volume maximal qui est silhouette-équivalent avec les objets filmés.

4.4.1 Approches surfaciques

Les approches surfaciques se concentrent sur le calcul d'une représentation explicite de la surface de l'enveloppe visuelle. Des éléments de la surface de l'enveloppe visuelle, tels que des

points ou des facettes, sont estimés par intersection des surfaces des cônes de silhouette. La figure 4.2 représente le principe de l'estimation d'un objet par l'approche *Shape-From-Silhouette* surfacique.

Parmi les travaux relatifs aux approches surfaciques, Baumgart [Bau74] a été parmi les premiers à calculer une estimation polyédrique de la forme d'objets à partir d'une approximation polygonale du contour des silhouettes. [GW87, CB92, VF92] se sont intéressés à des points isolés de la surface, reconstruits en utilisant des approximations locales du second ordre de la surface. Pour cela, ils imposent une hypothèse supplémentaire, considérant que les surfaces sont de classe $\mathcal{C}^2$.

Les approches surfaciques sont construites sur une approximation polygonale des contours des silhouettes. En plus des difficultés liées à l'extraction de l'approximation polygonale des contours, la précision de cette approximation influe sur l'estimation de la surface de l'enveloppe visuelle. Ainsi ce type d'approche se révèle particulièrement sensible à l'extraction des silhouettes et aux imprécisions de calibrage des caméras. Cette sensibilité se traduit généralement par une estimation de la surface de l'enveloppe visuelle incomplète ou corrompue.

Cependant les approches surfaciques fournissent une représentation explicite de la surface particulièrement adaptée aux application du rendu de la forme 3D rapide et plus "réaliste" que celles proposées par les approches volumiques (voir figure 4.3). Ainsi les approches surfaciques sont particulièrement utilisées dans les contextes de reconstruction 3D de la géométrie pour le rendu depuis n'importe quel point de vue (*Freeviewpoint Rendering*) [MBM01, MBR+00, LMS04]. Boyer et Franco [BF03] proposent une technique qui calcule la surface de l'enveloppe visuelle à partir d'une estimation polyédrique des contours des silhouettes. Brand *et al.* [BKC04] décrivent une approche de géométrie différentielle afin de calculer une estimation précise de l'enveloppe visuelle à partir des contours des silhouettes. Li *et al.* présentent des méthodes implémentées sur carte graphique programmable, pour construire la surface de l'enveloppe visuelle en temps réel [LMS04]. Lazebnick *et al.* ont travaillé sur une méthode qui caractérise la surface de l'enveloppe visuelle comme étant un polyèdre généralisée. Ils utilisent la géométrie différentielle projective afin d'en calculer la forme [LFP07]. Matusik *et al.* indroduisent l'enveloppe visuelle basée image, qui reconstruit l'enveloppe visuelle en temps réel pour un point de vue spécifique [MBR+00]. D'autres travaux proposent aussi des approches de *Shape-From-Silhouette* surfaciques temps réel [LCO06, LMS04].

Ces approches estiment uniquement la surface de l'enveloppe visuelle des objets d'intérêt. Dans le contexte de l'acquisition de mouvements de personnes nous nous intéressons à l'extraction des points 3D correspondants aux positions des articulations. Or ces points n'appartiennent pas en général à la surface des objets, ni à la surface de l'enveloppe visuelle correspondante. Ainsi la représentation explicite de la surface offerte pas ce type d'approche s'avère moins adaptée à notre contexte que les approches volumiques de *Shape-From-Silhouette*.

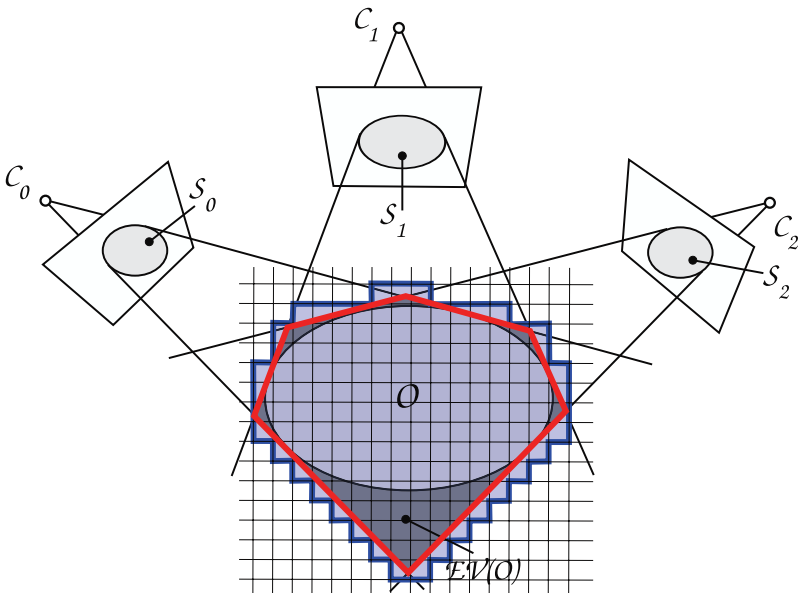


FIG. 4.2 – Illustrations d'une coupe de la géométrie estimée d'un objet O par l'approche *Shape-From-Silhouette* surfacique en rouge et par l'approche volumique à base de voxels en bleu.

.

4.4.2 Approches Volumiques

A l'opposé des approches de *Shape-From-Silhouette* surfaciques, les approches volumiques estiment le volume de l'enveloppe visuelle par un ensemble de primitive élémentaires : voxels, arbres octaux, etc. Ces approches sont principalement construites sur la propriété 2 selon laquelle l'enveloppe visuelle décrit le volume maximal silhouette-équivalent aux objets d'intérêt.

L'espace 3D d'intérêt supposé contenir l'enveloppe visuelle des objets d'intérêt est partitionné en un ensemble de cellules élémentaires. Ensuite chaque cellule est testée afin de vérifier son appartenance à l'enveloppe visuelle. Afin de valider la propriété de silhouette-équivalence, toute cellule dont la projection dans au moins une caméra n'intersecte pas de silhouette, est identifiée comme n'appartenant pas à l'enveloppe visuelle. L'ensemble des cellules validées forment une estimation englobante de l'enveloppe visuelle (voir figure 4.2). Chacune des cellule validée est dite *silhouette-consistante*.

Parmi les méthodes volumiques, certaines estiment l'enveloppe visuelle par un ensemble de voxels, alignés sur une grille régulière. Chaque voxel est testé en vérifiant s'il est silhouette-consistant. Cette étape est critique en terme de temps de calcul. Projeter la géométrie complète de chaque voxel dans chaque image peut s'avérer consommateur de temps machine. A l'opposé, il est possible de ne tester que la silhouette-consistance du centre de chaque voxel. L'algorithme 1 présente l'une méthode usuelle du calcul de *Shape-From-Silhouette* à base de voxels. Certaines approches intermédiaires [CKBH00] proposent de tester un échantillonnage des points de chaque voxel.

FIG. 4.3 – Comparaison des reconstructions d'un objet complexe par l'approche *Shape-From-Silhouette* surfacique (à gauche) et par l'approche volumique à base de voxels (à droite) à partir de 42 vues. Cette image est tirée des travaux de Boyer et Franco [BF03].

.

Diviser l'espace d'intérêt en $G \times G \times G$ voxels v_k avec $k \in [1, \cdots, G^3]$;
Pour k **de** 1 **à** G^3 **faire**
  $v_k \leftarrow$ *Dedans* ;
  Pour i **de** 1 **à** n **faire**
    Si $(\text{Proj}_{\pi_i}(v_k) \cap \tilde{\mathcal{S}}_i = \emptyset)$ **Alors**
      $v_k \leftarrow$ *Dehors* ;
    Fin Si
  Fin Pour
Fin Pour
L'EV est approximée par l'union des voxels v_k étiquetés *Dedans* ;

Algorithme 1: Algorithme standard de *Shape-From-Silhouette* à base de voxels

D'autre approches estiment l'*EV* sur GPU par un ensemble de voxels, en projetant les images de silhouettes dans l'espace des voxels en utilisant la technique de *projective-texture-mapping* [HLS04]. La grille de voxels peut être considérée comme étant une pile de plans d'images 2D. Pour chaque plan, si un pixel est intersecté par la projection de toutes les silhouettes, alors le voxel correspondant est à l'intérieur de l'enveloppe visuelle.

Certaines approches volumiques utilisent une structuration hiérarchique en arbre octal pour diminuer les temps de calculs [Pot87,NFA88,CH04b]. Ces méthodes testent la silhouette-consistance d'un pavé qui représente l'espace d'intérêt. Si celui-ci est indécis, c'est à dire que seul un sous-ensemble de points de ce pavé est silhouette-consistant alors ce pavé est partitionné en deux sur chaque axe et l'opération est répétée sur chacun des 8 sous-pavé. Le processus est conditionné à plusieurs critères d'arrêts. Le premier définit la profondeur maximale d'exploration, ou encore la

dimension minimale de chaque cellule. Un autre critère est qu'une décision totale soit prise pour chaque cellule de l'arbre octal. D'autres implémentations temps réel ont également été proposées en utilisant soit plusieurs machines [MS03], soit les capacités des cartes graphiques [LCO06], ou encore une faible résolution de grille volumique [Gra03].

Ces approches se révèlent en général plus rapides que les approches surfaciques. Elles ne nécessitent pas de polygonalisation des contours de silhouette et s'avèrent ainsi plus robustes aux silhouettes bruitées que les approches surfaciques. Un point de silhouette classé faux-positif sera compensé par certains points de silhouette vrai-négatifs issus des autres caméras. L'enveloppe visuelle définissant un volume englobant des objets d'intérêt, la reconstruction offerte par les approches volumiques peut être facilement affinée à l'aide de critères de cohérence basés par exemple sur la couleur dite cohérence photométrique.

Les approches volumiques offrent une estimation de l'enveloppe visuelle généralement plus grossière que les approches surfaciques, cela dépendant l'échantillonnage volumique (voir figure 4.3). Ainsi ces approches sont peu utilisées dans le contexte de la modélisation pour le rendu. Un échantillonnage de faible résolution implique que la forme estimée ne respecte pas exactement la propriété de silhouette-équivalence.

Les approches volumiques nécessitent la définition d'un espace d'intérêt englobant l'enveloppe visuelle des objets d'intérêt. Ainsi le choix de cet espace est une étape critique et peut amener à l'estimation que d'une sous-partie de l'EV.

4.5 Verrous de *Shape-From-Silhouette* et travaux associés

Les approches *Shape-From-Silhouette* permettent d'estimer en temps réel la forme 3D d'objets d'intérêt observés depuis plusieurs points de vue. Celles-ci sont construites sur l'information de silhouette dont l'extraction se révèle efficace en environnement contrôlé observé par un ensemble de caméras fixes. Comparées aux autres approches de reconstruction 3D à partir de plusieurs points de vues, les méthodes *Shape-From-Silhouette* s'avèrent simples à implémenter. Ces précédentes caractéristiques, couplées à la propriété que la forme obtenue contient les objets d'intérêt, ont rendu les méthodes *Shape-From-Silhouette* particulièrement populaires. Cependant elles présentent certains verrous.

Afin que la totalité des objets d'intérêt soient estimés par *Shape-From-Silhouette*, il est indispensable que ceux-ci soient totalement visibles dans toutes les caméras. Cela impose des contraintes fortes sur le placement des caméras. L'enveloppe visuelle fournissant un volume englobant des objets d'intérêt, certaines parties vides d'objets, dites entités fantômes, sont construites par les approches *Shape-From-Silhouette*. Enfin en présence d'un environnement d'acquisition faiblement contrôlé, l'extraction de silhouette est souvent bruitée. Cela a une répercussion sur la précision de la reconstruction de *Shape-From-Silhouette*.

4.5.1 Contraintes de placement des caméras

La propriété 2 selon laquelle l'enveloppe visuelle est le volume maximum silhouette-équivalent, définit la méthode de calcul de la plupart des approches volumiques de *Shape-From-Silhouette* (voir l'algorithme 1). Néanmoins, elle comporte un effet de bord indésirable car elle ne prend pas en compte le fait que certains objets peuvent être partiellement vus, ou totalement en dehors du champ de vision de certaines caméras (voir figure 4.4). Si certains objets ne respectent pas la contrainte de *visibilité totale* dans toutes les caméras, alors *Shape-From-Silhouette* ne peut pas estimer la totalité de l'EV correspondante.

Dans le cas de scène fixe, cette limitation est contournée en plaçant les objets de façon à ce qu'il soient tous totalement visibles depuis toutes les caméras. En présence d'objets dynamiques, comme par exemple plusieurs personnes en mouvement, il devient difficile de contraindre leurs déplacements. C'est à dire qu'il est difficile de garantir leur visibilité totale dans toutes les caméras. L'approche *Shape-From-Silhouette* classique souffre de contraintes sur le placement des caméras par rapport aux objets d'intérêt et inversement.

Cette limitation provient du fait que *Shape-From-Silhouette* ne peut reconstruire l'EV des objets seulement que s'ils sont dans l'intersection des cônes de vision des caméras, c'est à dire que :

$$\forall i \in [0, \cdots, n], \tilde{\mathcal{S}}_i = \mathcal{S}_i \tag{4.3}$$

Afin de contourner cette limitation, une solution naturelle serait d'augmenter le nombre de caméras dans la scène. Cependant l'espace d'acquisition est définit par l'intersection des cônes de vues des caméras. La conséquence de cette augmentation serait une réduction de l'espace d'acquisition.

A notre connaissance, peu de travaux proposent une solution pour relaxer les contraintes de placement des caméras pour les approches *Shape-From-Silhouette*. Franco et Boyer [BF03,FB08] sont parmi les premiers à identifier cette limitation et à proposer une nouvelle formalisation de *Shape-From-Silhouette*. Leur approche permet d'estimer l'enveloppe visuelle d'objets d'intérêt non vus par une ou plusieurs caméras. Ils proposent de prendre en compte les contributions de chaque silhouette uniquement dans le domaine de visibilité de cette silhouette. Dans [FB05, GFP07, GFP08] ils décrivent une modélisation probabiliste du volume d'intérêt afin d'estimer l'enveloppe visuelle sans contraintes sur le placement des caméras. L'ensemble des méthodes proposées sont cependant coûteuses en temps de calcul. Par exemple la méthode décrite dans [BF03] nécessite plusieurs secondes pour calculer l'EV. Les approches exposées dans [FB05, GFP07, GFP08] sont basées sur une résolution du problème par réseaux bayesiens et nécessitent au moins dix secondes de traitement par trame. Les résultats obtenus sont de bonne qualité, mais ces approches ne sont pas acceptables pour le temps réel. Relaxer les contraintes de placement des caméras peut rendre *Shape-From-Silhouette* plus apte à des conditions peu contrôlées. En

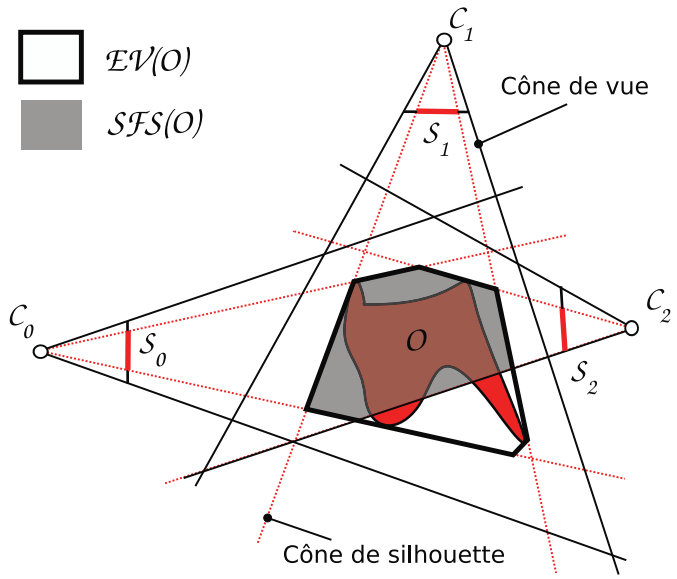


FIG. 4.4 – Configuration de caméra dans laquelle *Shape-From-Silhouette* ne peut estimer l'enveloppe visuelle des objets d'intérêt : L'objet n'est pas totalement visible depuis la caméra $\mathcal{C}_2$.

.

contrepartie leurs méthodes ajoutent des objets fantômes, ensembles connexes de point 3D qui ne correspondent à aucun objet d'intérêt.

4.5.2 Les entités fantômes

L'enveloppe visuelle d'objets d'intérêt est le volume maximal qui est silhouette-consistant avec ces mêmes objets filmés. Ce volume englobe la forme des objets d'intérêt (propriété 1). L'enveloppe visuelle des objets d'intérêt peut s'écrire de la façon suivante :

$$\mathrm{EV}(\mathcal{O}) = \mathcal{O} \cup \mathrm{E}_{fantôme} \tag{4.4}$$

avec $\mathrm{E}_{fantôme}$ les entités fantômes générées par l'enveloppe visuelle (voir figure 4.5). Ces entités fantômes sont des points 3D de l'espace qui sont silhouette-consistants, mais qui n'appartiennent pas aux objets d'intérêt.

Rappelons que l'enveloppe visuelle $\mathrm{EV}(\mathcal{O})$ d'un objet $\mathcal{O}$ est le volume maximal silhouette-équivalent avec $\mathcal{O}$. Une condition nécessaire pour qu'un point 3D P appartienne à $\mathcal{O}$ est que P soit silhouette-consistant avec $\mathcal{O}$. Cette condition est nécessaire mais pas suffisante. Le fait que la silhouette-consistance ne soit pas une relation d'équivalence implique que $\mathrm{EV}(\mathcal{O})$ contient à la fois $\mathcal{O}$ et des artefacts $\mathrm{E}_{fantôme}$.

Pour réduire les entités fantômes d'une EV, une solution est d'augmenter le nombre de vues (propriété 3). En réalité cela ne représente pas une solution ultime. En effet A. Laurentini a prouvé dans [Lau94] qu'avec une infinité de points de vue, il n'est pas possible de reconstruire

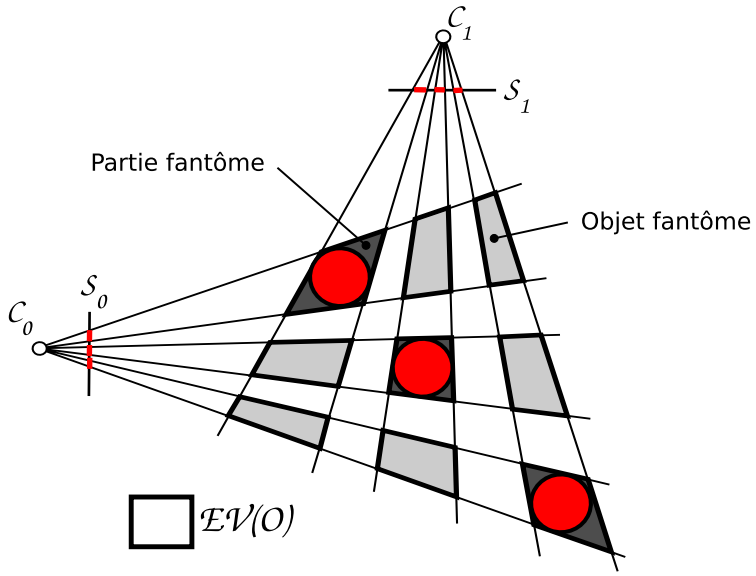


FIG. 4.5 – Dans cette configuration, l'enveloppe visuelle des objets d'intérêt est constituée de neuf composantes connexes. Six de ces composantes sont des *objets fantômes*représentée en gris clair. Les *parties fantômes*sont représentées en gris foncé. Ces entités fantômes de l'enveloppe visuelle ne contiennent pas d'objet réel.

.

les parties concaves de l'objet si elles ne définissent pas de contours dans au moins une des silhouette. Ces parties sont dites surfaces "silhouette-inactives". Il n'existe à priori pas de solution simple pour supprimer directement toutes les entités fantômes.

Nous proposons de séparer les entités fantômes en deux ensembles distincts :

$$\mathrm{EV}(\mathcal{O}) = \mathcal{O} \cup \mathrm{P}_{fantôme} \cup \mathrm{O}_{fantôme} \tag{4.5}$$

où $\mathrm{P}_{fantôme}$ représente les parties fantômes et $\mathrm{O}_{fantôme}$ représente les objets fantômes de l'enveloppe visuelle. Nous différencions ces deux types d'artefacts par le fait qu'un objet fantôme représente une partie connexe de l'enveloppe visuelle qui ne contient aucun objet réel, alors que les parties fantômes sont connectées à des objets réels.

Définition 5 *Soit* CC_j *l'une des composantes connexes de* $\mathrm{EV}(\mathcal{O})$. CC_j *est un* ***objet fantôme*** *de* $\mathrm{EV}(\mathcal{O})$ *si et seulement si* $\forall \mathrm{X} \in \mathrm{CC}_j : \mathrm{X} \notin \mathcal{O}$. □

Définition 6 *Soit* CC_j *l'une des composantes connexes de* $\mathrm{EV}(\mathcal{O})$. CC_j *contient une* ***partie fantôme*** *de* $\mathrm{EV}(\mathcal{O})$ *si et seulement si* $\exists(\mathrm{X}, \mathrm{Y}) \in \mathrm{CC}_j^2$ *tels que* $\mathrm{X} \in \mathcal{O}$ *et* $\mathrm{Y} \notin \mathcal{O}$. □

La figure 4.5 représente un cas typique où l'enveloppe visuelle contient des objets fantômes et des parties fantômes. Les entités fantômes apparaissent lorsqu'il existe des zones de l'espace consistantes avec l'ensemble des silhouettes, mais ne contenant aucun objet réel.

Remarque 2 *Il n'existe pas nécessairement d'objet fantôme dans une enveloppe visuelle. Cependant il existe toujours des parties fantômes dans une enveloppe visuelle, sauf lorsque* $\mathcal{O} = \mathrm{EV}(\mathcal{O})$.
□

Parmi l'ensemble des méthodes qui proposent un mécanisme de suppression d'entités fantômes, la plupart nécessitent de l'information supplémentaire. Dans [MH07], Miller et Hilton définissent le concept de *Safe Hull* (enveloppe sûre). Cette enveloppe définit les parties de l'enveloppe visuelle, dont on est sûr qu'elle contient des objets réels. Ils déterminent les portions de l'espace dits sûrs, en calculant l'intersection entre l'enveloppe visuelle et tous les rayons de silhouette [1] issus de toutes les caméras. Cette intersection fournit pour chaque rayon le nombre d'intervalles de l'enveloppe visuelle qu'il traverse. L'ensemble des points 3D de l'EV intersectés par des rayons qui traversent un seul intervalle de l'EV, constituent la *Safe Hull*. En présence de beaucoup de points de vue, cette approche supprime quelques entités fantômes. Cependant, leur approche souffre de temps de calculs particulièrement longs. De plus, rien ne garantit que la géométrie calculée par la *Safe Hull* contienne tous les objets de la scène. D'un côté cette approche supprime une partie des entités fantômes de l'EV, de l'autre côté leur approche viole une propriété fondamentale de l'enveloppe visuelle. En effet, en présence d'occultations partielles, des points 3D qui appartiennent aux objets réels seront supprimés.

La plupart des méthodes qui suppriment une partie des entités fantômes nécessitent une information supplémentaire comme par exemple les couleurs, les contours, ou encore une connaissance temporelle.

Bogomjakov et Gotsman [BG08] proposent une approche qui permet de supprimer certaines entités fantômes à l'aide de cartes de profondeurs, ou encore la mise en correspondance de silhouettes. Les cartes de profondeurs proviennent de caméras spécifiques 2.5D et la mise en correspondance peut-être mise en échec lorsque les objets ont une apparence proche. Bien que cette méthode offre le temps réel à l'aide d'informations supplémentaires, celle-ci ne garantit pas la suppression de tous les objets fantômes.

Pour calculer une géométrie plus fine, en plus de l'information de silhouettes, la *consistance-photométrique* a été prise en compte pour reconstruire la géométrie à partir de plusieurs vues. La plupart des approches qui ont été proposées sont volumiques. Chaque voxel est projeté dans toutes les images, pour vérifier si les régions dans les images où ce voxel se projette ont une couleur similaire. Si les régions ont une couleur similaire, alors le voxel correspondant est dit

[1] rayons limités aux points d'image classés silhouette

photo-consistant, c'est à dire qu'il semble être à la surface de l'objet réel. Cela revient à dire que si un voxel n'est pas photo-consistant, c'est qu'il n'appartient pas à la surface d'un objet réel. Pour diminuer les temps de calcul, les voxels précédemment calculés par une approche *Shape-From-Silhouette* sont testés couche par couche, en commençant par l'extérieur. De cette façon le modèle volumique est sculpté, pour obtenir au final une enveloppe photo-consistante (*PhotoHull*) [SD97,CMS99]. L'algorithme qui calcule la *PhotoHull* est généralement appelé *Space Carving*.

Les méthodes construites sur la photo-consistance supposent que les objets à reconstruire sont parfaitement diffus et texturés. De plus un voxel peut être classé comme photo-consistant si ses projections dans toutes les caméras sont des régions de couleur similaire. Cela ne garantit pas que ce voxel appartienne à la surface de l'objet. Tout comme le test de *silhouette-consistance*, la *photo-consistance* est une condition nécessaire, mais pas suffisante. D'un côté l'approche *Photo-Hull* permet de calculer un volume plus précis que l'EV, mais elle ne garantit pas de reconstruire les parties spéculaires de l'objet. De plus, par la détermination de la visibilité de chaque voxel, il n'existe à ce jour aucune implémentation en temps réel, mais seulement quelques une en temps interactif.

D'autres travaux [GFP08, KS00, MM04] proposent des reconstructions, qui sont en général, dépourvues d'objets fantômes en utilisant par exemple des mécanismes de mise en correspondance. Toutefois, l'ensemble de ces méthodes demandent des calculs intensifs pour l'évaluation d'inférences ou encore parfois par la mise en correspondance de pixels.

Les méthodes précédentes estiment la géométrie d'un objet à chaque trame, séparément, c'est a dire indépendamment de l'image précédente et de l'image suivante. Les propriétés physiques fondamentales de l'inertie, imposent que les objets en mouvement évoluent de façon continue à travers le temps. En utilisant le fait que le mouvement est continu, la géométrie calculée est plus précise. La prise en compte de la continuité temporelle a été intégrée à *Shape-From-Silhouette* de plusieurs façons. Certains travaux intègrent à la fois la cohérence temporelle avec l'approche de *PhotoHull* comme par exemple dans les travaux [VBK02,GM04a]. Dans [CBK03b] Cheung *et al.* ont proposé une méthode qui permet de fortement atténuer les entités fantômes en utilisant un recalage temporel basé sur des points caractéristiques dits *Colored Surface Points*. L'enveloppe visuelle est un englobant des objets réels. Donc parmi les points de surface de l'enveloppe visuelle, certains appartiennent aussi aux objets d'intérêt. Pour les détecter, Cheung *et al.* calculent la cohérence photométrique de chaque point de surface de l'enveloppe. Les points photo-consistants sont alors définis comme étant les Colored-Surface-Points (CSP). Ensuite ils calculent le recalage des points de l'enveloppe courante dans les images de la trame précédente et inversement. Le nombre de CSP augmente au cours du temps et ces points sont acceptés comme étant des points réels des objets filmés. Leur méthode permet de supprimer la plupart des entités fantômes lorsque les objets filmés sont lambertiens et contrastés.

Dans [SH07] Starck *et al.* proposent une méthode de reconstruction qui utilise l'ensemble des critères de cohérence cités ci-avant. Ils corrigent l'estimation volumique de l'enveloppe visuelle à l'aide de points caractéristique photo-consistants, ainsi qu'à travers un filtrage temporel ce qui produit des données en 4D. Enfin la forme 3D est déterminée à chaque pas de temps en utilisant une approche d'optimisation globale par *graphe cut*. Les résultats obtenus sont de très bonne qualité. Ces travaux représentent l'une des approches de référence.

L'ensemble de ces méthodes offrent des résultats intéressants, mais elles nécessitent de grosses capacités de calculs et ne peuvent aujourd'hui être implantées en temps réel.

Peu de méthodes de la littérature font la distinction entre parties fantômes et objets fantômes. Cependant il existe plusieurs approches qui ne traitent que des objets fantômes. Yang *et al.* ont proposé dans [YHHGBG03] une approche qui permet de supprimer la plupart des objets fantômes. Leur contexte de surveillance leur permet de supposer que les composantes connexes de l'EV qui contiennent un individu humain, ont un volume minimum. Enfin pour supprimer les objets fantômes restants, ils ont proposé un filtrage temporel. Ils supposent qu'une composante connexe de l'EV qui contient un objet réel en mouvement, décrit un mouvement lisse et continu. Malgré des résultats intéressants, cette approche ne peut être généralisée dans des scènes dynamiques avec des objets réels de tailles quelconques. L'approche proposée par Bogomjakov et Gotsman [BG08] offre une solution qui focalise sur la suppression des objets fantômes. Il est néanmoins nécessaire de déterminer les correspondances entre silhouettes. Les méthodes qui réalisent cette appariement sont cependant sensibles à l'apparence et aux occultations des objets présents dans la scène. Bien que cette approche fournisse le temps réel, elle ne garantit pas la suppression de tous les objets fantômes.

4.5.3 Sensibilité à la qualité de l'extraction des silhouettes

L'approche *Shape-From-Silhouette* estime l'enveloppe visuelle d'objets d'intérêt à partir de cartes binaires de silhouettes de ces objets, issues d'un ensemble de caméras calibrées. Si dans l'un des masques de silhouette, un pixel est détecté comme étant un faux pixel de fond, alors l'estimation de l'enveloppe visuelle sera incomplète, ce qui peut représenter une forte dégradation de l'information reconstruite et par la suite des difficultés à suivre les mouvements des personnes filmées. Ce type d'erreur provient du fait que l'étiquetage monoculaire est difficile à réaliser de façon fiable dans un environnement général et peu contrôlé. Diverses perturbations expliquent cette difficulté : le bruit des capteurs CCD, les ambiguïtés de couleur entre le fond et l'avant-plan ou encore les changements d'illumination de la scène (ce qui inclut les ombres d'objets d'intérêt). Une solution alternative est d'utiliser conjointement l'information offerte par l'ensemble des caméras. Cela revient à repousser la décision binaire initialement réalisée dans chaque image séparément, à l'étape du test de silhouette-consistance. En effet, si la projection d'un point 3D dans chaque image retourne une forte probabilité d'appartenir à la silhouette, sauf dans une

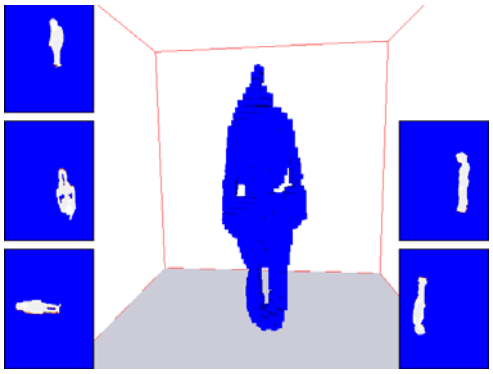


FIG. 4.6 – Sensibilité de *SFS* par rapport aux erreurs d'extraction de silhouette. La seconde image de silhouette (au centre à gauche) fournit une carte binaire de silhouette erronée. En conséquence la reconstruction par une approche classique de *Shape-From-Silhouette* est incomplète.

.

caméra, alors ce point 3D semble réellement appartenir à l'enveloppe visuelle. La littérature liée à cette problématique est limitée.

Une solution proposée dans [FB05] repose sur le concept de grille d'occupation, méthode populaire dans la communauté de la robotique. L'idée centrale de cette approche est de considérer que chaque pixel est un capteur apportant de l'information sur la scène observée et de le modéliser statistiquement comme tel. Les observations rapportées par les pixels de toutes les caméras peuvent alors être conjointement utilisées pour déduire où la matière se trouve dans la scène et avec quelle probabilité. Ce modèle présente certains avantages : la plupart des sources d'incertitude peuvent être prises en compte explicitement et aucune décision prématurée sur l'état des pixels et leur appartenance à une silhouette n'est nécessaire. Cette approche permet de s'affranchir des contraintes de visibilité communes aux méthodes classiques de reconstruction d'enveloppes visuelles. Le calcul de l'inférence est basé sur une résolution par réseaux bayesiens et nécessite au moins dix secondes pour traiter une seule trame. Cette approche offre une solution élégante à ce problème, mais n'est pas transposable dans notre contexte.

4.6 Contributions

Nous proposons dans les chapitres suivants nos contributions qui permettent d'estimer en temps réel la forme 3D de plusieurs objets, à partir de leurs silhouettes issues de plusieurs caméras. Le processus global mis en place n'impose plus de contraintes sur le positionnement des caméras. Les formes reconstruites ne contiennent pas d'objet fantômes. Enfin la reconstruction obtenue est robuste par rapport aux extractions de silhouette erronées.

Ces contributions répondent à trois des principaux verrous des approches *Shape-From-Silhouette*,

et ce sans autre information que les silhouettes des objets d'intérêt et le calibrage géométrique des caméras.

Nous commençons par présenter les différentes contributions que nous avons apporté pour relaxer les contraintes de placement des caméras. L'idée principale se base sur le fait que la décision de l'état d'un point de volume 3D doit être prise par rapport aux caméras qui voient ce point, et non l'ensemble de toutes les caméras. Nous définissons l'*Enveloppe Visuelle Étendue* qui présente une alternative aux *Enveloppes Visuelles*, dont la nouvelle formalisation ne suppose plus le champ de vision de chaque caméra infini. Cette nouvelle approche étend l'espace d'acquisition, mais ajoute cependant des objets fantômes supplémentaires. Dans la fin de ce chapitre, nous proposons plusieurs approches qui permettent de limiter l'ajout d'objets fantômes.

Nous présentons ensuite les mécanismes mis en place pour supprimer tous les objets fantômes générés par les approches qui estiment l'enveloppe visuelle. Nous exposons nos contributions en précisant le contexte de fonctionnement. Enfin nous justifions le fait qu'il n'est pas possible de trouver une solution optimale, dans la mesure où il existe une infinité de configurations d'objets réels qui fournissent la même enveloppe visuelle. Dans ce cas il n'est pas possible de les différencier simplement à partir des silhouettes, à moins de faire une hypothèse sur le placement des objets dans la scène.

Nous présentons dans la troisième partie une contribution simple qui rend les approches d'estimation de forme à partir de silhouettes plus robustes face aux problèmes d'extraction de silhouette. Cette méthode se repose sur le fait que l'information de silhouette dans une image, peut être remise en question par l'information de silhouette disponible dans les autres images

Enfin nous présentons les résultats de nos contributions. Nous les comparons aux méthodes de l'état de l'art. Nous discutons ensuite les apports réalisés et proposons certaines pistes de recherche.

Chapitre 5

Extension de l'espace d'acquisition

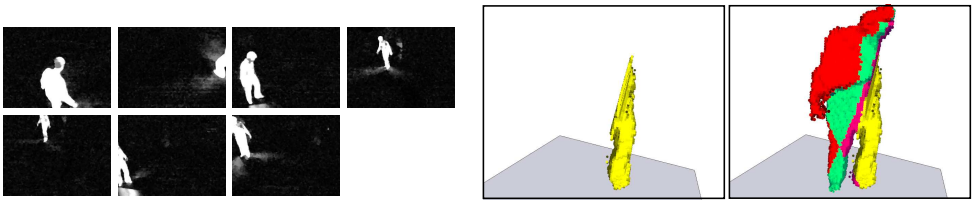


FIG. 5.1 – Lorsque les objets à construire ne sont pas totalement visibles depuis toute les caméras, les approches usuelles de SFS sont incapables de les estimer en entier (a). Ce chapitre focalise sur une extension permettant de relaxer cette contrainte (b).

Nous venons de présenter la définition et les propriétés fondamentales des enveloppes visuelles. La méthode *Shape-From-Silhouette* calcule l'enveloppe visuelle d'objets d'intérêt à partir des silhouettes de ces objets, issues de caméras calibrées. Lorsque les objets d'intérêt ne sont pas tous visibles dans toutes les caméras, *Shape-From-Silhouette* estime un sous-ensemble de l'enveloppe visuelle. Ainsi la forme calculée viole une propriété fondamentale des enveloppes visuelles : elle n'englobe pas les objets d'intérêt.

Notre objectif global est de réaliser l'acquisition du mouvement 3D de plusieurs personnages. Lors de mouvements de grande amplitude de ces personnes, elles ne sont pas toujours entièrement vues par toutes les caméras en même temps, surtout lorsque l'espace d'acquisition est réduit. Dans ce cas *Shape-From-Silhouette* estime une forme 3D tronquée qui risque de mettre en échec la méthode d'acquisition de mouvements (voir Fig. 5.1).

Shape-From-Silhouette estime la totalité de l'enveloppe visuelle lorsque les objets d'intérêt sont tous inclus dans l'intersection des cônes de vision des caméras. Cette condition impose certaines contraintes sur le placement et l'orientation de chaque caméra. Cette limitation provient de la définition des enveloppes visuelles, qui suppose que les champs de vue des caméras sont infinis, ou encore que tout point de l'espace 3D est vu par chaque caméra. Dans ce chapitre nous proposons une nouvelle formalisation de l'enveloppe visuelle dite *enveloppe visuelle étendue*,

qui tient compte du champ de vue réel des caméras. Cette nouvelle formalisation conserve les propriétés fondamentales des enveloppes visuelles, quelles que soient les positions et orientations des caméras.

Dans la prochaine section, nous définissons le concept d'*enveloppe visuelle étendue*. Cette nouvelle formalisation permet de construire une forme 3D qui contient les objets d'intérêt, qui est silhouette-équivalente, et ce sans contraintes sur le placement des caméras. Cependant le fait de relaxer ces contraintes, génère plus d'entités fantômes que l'enveloppe visuelle classique. Nous limitons ensuite l'ajout d'entités fantômes, en supposant que les objets d'intérêt sont vus par un nombre minimum de caméra, m. Nous présentons dans les sections suivantes différentes approches qui visent à ne pas ajouter d'objets fantômes par rapport à l'enveloppe visuelle classique.

5.1 Enveloppe visuelle étendue

L'approche *Shape-From-Silhouette* fournit un outil puissant pour estimer la forme 3D d'objets filmés par plusieurs caméras. Ces approches ont montré leur efficacité particulièrement dans des contextes de temps réel. L'une de ses limitations réside dans la contrainte de visibilité des objets à reconstruire. L'approche *Shape-From-Silhouette* calcule l'enveloppe visuelle des objets d'intérêt lorsqu'ils sont tous totalement visibles dans toutes les caméras. Violer cette condition implique que la forme calculé par *Shape-From-Silhouette* est un sous-ensemble de l'enveloppe visuelle des objets d'intérêt.

Comme nous l'avons souligné, le concept d'enveloppe visuelle est défini à partir d'un modèle théorique de caméra à champ de vue infini. C'est à dire que tout point 3D est vu par toutes les caméras en même temps. Cependant les caméras réelles ont une pyramide de vue limitée. Nous allons étendre la notion d'enveloppe visuelle, basée sur les observables réels : les silhouettes issues des caméras.

L'approche *Shape-From-Silhouette* estime la forme des objets contenus dans l'intersection des cônes de vision de chaque caméra. Il en découle que l'espace d'acquisition ne peut être étendu au delà de la zone d'acquisition. Cette dernière devient d'autant plus réduite que le nombre de caméras est important.

Nous notons $\tilde{\mathcal{S}}(\mathrm{X})$ l'ensemble des silhouettes et $\mathcal{I}(\mathrm{X})$ l'ensemble des images, dans lesquelles se projette le point 3D X :

$$\tilde{\mathcal{S}}(\mathrm{X}) = \{i \in [1, \cdots, n], \mathrm{Proj}_{\pi_i}(\mathrm{X}) \in \tilde{\mathcal{S}}_i\} \tag{5.1}$$

$$\mathcal{I}(\mathrm{X}) = \{i \in [1, \cdots, n], \mathrm{Proj}_{\pi_i}(\mathrm{X}) \in \mathcal{I}_i\}. \tag{5.2}$$

avec n le nombre de caméras, $\tilde{\mathcal{S}}_i$ la silhouette observable et $\mathcal{I}_i$ l'image issues de la caméra $\mathcal{C}_i$.

Notons que par construction

$$\tilde{\mathcal{S}}(\mathrm{X}) \subseteq \mathcal{I}(\mathrm{X}) \tag{5.3}$$

Shape-From-Silhouette estime la forme 3D d'un objet en calculant l'ensemble maximum de points 3D qui se projettent dans toutes les silhouettes. *Shape-From-Silhouette* à partir de n caméras peut s'écrire de la façon suivante :

$$SFS = \{\mathrm{X} \in R^3, \tilde{\mathcal{S}}(\mathrm{X}) = \mathcal{I}(\mathrm{X}) \text{ et } Card(\mathcal{I}(\mathrm{X})) = n\} \tag{5.4}$$

La reconstruction proposée par *Shape-From-Silhouette* est consistante avec toutes les silhouettes si tous les objets de la scènes sont situés dans la stricte intersection des cônes de vues de toutes les caméras. Si un point 3D est en dehors, alors *Shape-From-Silhouette* ne le reconstruira pas.

Nous proposons dans la suite d'estimer la forme des objets d'intérêt par l'ensemble maximal des points 3D qui sont consistants avec les caméras qui voient ce point.

Définition 7 *Soit* $\mathrm{EVE}(\mathcal{O})$ *l'enveloppe visuelle étendue d'objet* $\mathcal{O}$ *définie par l'ensemble des point 3D qui sont consistants pour les caméras qui voient ces points :*

$$\mathrm{EVE}(\mathcal{O}) = \{\mathrm{X} \in R^3 : \tilde{\mathcal{S}}(\mathrm{X}) = \mathcal{I}(\mathrm{X})\}$$

□

Remarque 3 *L'enveloppe visuelle étendue est un volume englobant défini par rapport aux centres des caméras et aux champs de vue associés. Cette nouvelle formalisation n'impose pas de contrainte de visibilité de* $\mathcal{O}$*.* □

La figure 5.2 représente un objet $\mathcal{O}$ observé par plusieurs caméras dont l'une voit $\mathcal{O}$ partiellement. L'objet est inclus dans $\mathrm{EVE}(\mathcal{O})$, alors que *Shape-From-Silhouette* ne le contient pas. En plus de supprimer les contraintes de placement des caméras, cette nouvelle formalisation construit une forme 3D qui contient tout point de $\mathcal{O}$.

Propriété 4

$$\mathrm{EV}(\mathcal{O}) \subseteq \mathrm{EVE}(\mathcal{O})$$

□

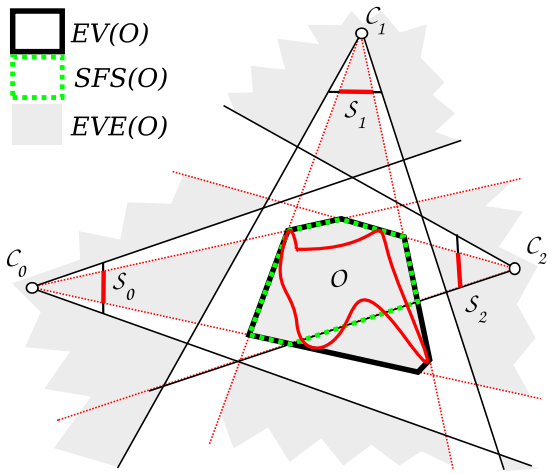


FIG. 5.2 – Représentation d'un coupe 2D d'objets $\mathcal{O}$, de EV($\mathcal{O}$), de *Shape-From-Silhouette* et de EVE($\mathcal{O}$).

.

Preuve Démontrons que tout point de EV($\mathcal{O}$) appartient à EVE($\mathcal{O}$).

Soit P $\in$ EV($\mathcal{O}$), alors $\forall i \in [1, \cdots, n], \exists$Q $\in \mathcal{O}$: Q $\in [C_i, \text{P})$.
Alors $\forall i \in [1, \cdots, n], \text{Proj}_{\pi_i}(\text{P}) \in \mathcal{S}_i$ or $\tilde{\mathcal{S}}_i = \mathcal{S}_i \cup \mathcal{I}_i$.
On en déduit que $\forall i \in \mathcal{I}(\text{P}), i \in \tilde{\mathcal{S}}(\text{P})$. De plus $\forall \text{X}, \tilde{\mathcal{S}}(\text{X}) \subseteq \mathcal{I}(\text{X})$. Ainsi $\tilde{\mathcal{S}}(\text{P}) = \mathcal{I}(\text{P})$.
Alors P $\in$ EVE($\mathcal{O}$), donc

$$\text{EV}(\mathcal{O}) \subseteq \text{EVE}(\mathcal{O})$$

□

On déduit de cette propriété que

Propriété 5

$$\mathcal{O} \subseteq \text{EV}(\mathcal{O}) \subseteq \text{EVE}(\mathcal{O})$$

□

Preuve D'après la propriété 1 $\mathcal{O} \subseteq$ EV($\mathcal{O}$) et d'après la propriété 4, EV($\mathcal{O}$) $\subseteq$ EVE($\mathcal{O}$). Ainsi

$$\mathcal{O} \subseteq \text{EV}(\mathcal{O}) \subseteq \text{EVE}(\mathcal{O})$$

□

Ces deux dernières propriétés prouvent que notre formalisation des *enveloppes visuelles étendues* respecte les propriétés fondamentales des enveloppes visuelles, sans imposer de contrainte de

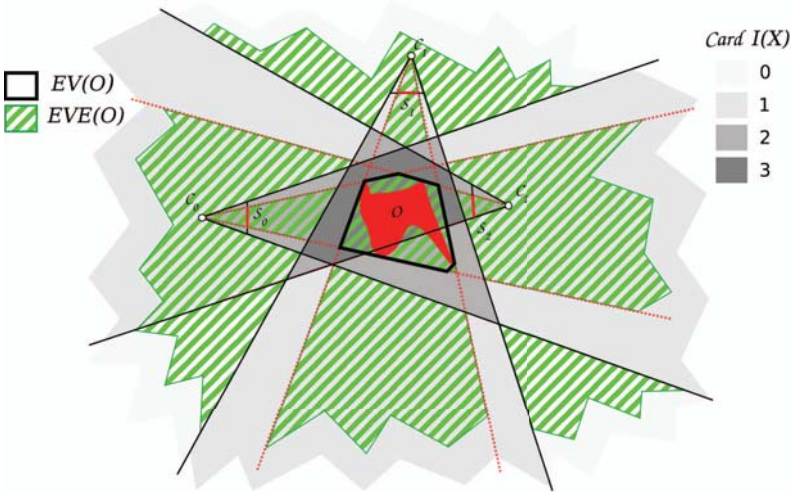


FIG. 5.3 – La densité des entités fantômes de EVE($\mathcal{O}$) est la plus forte dans les parties de l'espace vues par un faible nombre de caméras. Cette quantité diminue dans les zone couvertes par beaucoup de caméras

.

visibilité sur les objets à reconstruire. Tout point de $\mathcal{O}$ est reconstruit par *Shape-From-Silhouette* seulement lorsque il est vu par toutes les caméras. Notre approche est beaucoup moins restrictive dans la mesure où un point de $\mathcal{O}$ est reconstruit quelle que soit le configuration des caméras. L'une des conséquences de cette propriété est que si un point X n'appartient pas à EVE($\mathcal{O}$), alors X $\notin \mathcal{O}$. Une autre conséquence est que l'*enveloppe visuelle étendue* contient plus d'entités fantômes que l'enveloppe visuelle classique.

Nous venons de proposer une extension des enveloppes visuelles, qui conserve les mêmes propriétés fondamentales. Les propriétés des *enveloppes visuelles étendues* sont fondées sur les données observables du problème. Ainsi EVE($\mathcal{O}$) contient toujours $\mathcal{O}$, quelle que soit la configuration des caméras.

Le fait de relaxer les conditions de fonctionnement, amène à la construction de plus d'entités fantômes, par rapport aux enveloppes visuelles classiques. Même dans une configuration pour laquelle toutes les caméras voient tout point des objets à reconstruire EVE($\mathcal{O}$) $\neq$ EV($\mathcal{O}$). La quantité d'entités fantômes est importante dans les parties de l'espace vues par un faible nombre de caméras (voir aucune), alors que ce nombre diminue dans les parties de l'espace vues par beaucoup de caméras (voir figure 5.3).

En supposant que les objets d'intérêt sont vus par un nombre minimum de caméras que l'on notera m, il est possible de diminuer les entités fantômes. On note EVE$_m$ l'*enveloppe visuelle étendue* des objets d'intérêt vus par au moins m caméras :

Définition 8

$$\mathrm{EVE}_m = \{\mathrm{X} \in \mathrm{EVE}(\mathcal{O}), Card(\mathcal{I}(\mathrm{X})) \geq m\}$$

□

On en déduit la propriété suivante.

Propriété 6

$$Si\ \forall \mathrm{Q} \in \mathcal{O}, Card(\mathcal{I}(\mathrm{Q})) \geq m\ alors\ \mathcal{O} \subseteq \mathrm{EVE}_m$$

□

Preuve Tout point $\mathrm{Q} \in \mathcal{O}$ tel que $Card(\mathcal{I}(\mathrm{Q})) \geq m$ satisfait à la définition 8.

En effet $\forall \mathrm{Q} \in \mathcal{O}, \forall i \in [1, \cdots, n], \mathrm{Proj}_{\pi i}(\mathrm{Q}) \in \mathcal{S}_i$.

Or $\tilde{\mathcal{S}}_i = \mathcal{S}_i \cap \mathcal{I}_i$. On en déduit que $\forall i \in \mathcal{I}(\mathrm{Q}), i \in \tilde{\mathcal{S}}(\mathrm{Q})$.

De plus, d'après l'équation 5.3 $\forall \mathrm{X}, \tilde{\mathcal{S}}(\mathrm{X}) \subseteq \mathcal{I}(\mathrm{X})$.

Ainsi $\tilde{\mathcal{S}}(\mathrm{Q}) = \mathcal{I}(\mathrm{Q})$ et par hypothèse $Card(\mathcal{I}(\mathrm{Q})) \geq m$. On en déduit que

$$\forall \mathrm{Q} \in \mathcal{O}, Card(\mathcal{I}(\mathrm{Q})) \geq m \Rightarrow \mathcal{O} \subseteq \mathrm{EVE}_m$$

□

Tout point de $\mathcal{O}$ est reconstruit par *Shape-From-Silhouette* seulement lorsqu'il est vu par toutes les caméras. Notre approche est beaucoup moins restrictive dans la mesure où un point de $\mathcal{O}$ est reconstruit par EVE_m si il est visible par au moins m caméras.

Contrairement à $\mathrm{EVE}(\mathcal{O})$ qui contient $\mathrm{EV}(\mathcal{O})$, EVE_m ne contient pas toujours $\mathrm{EV}(\mathcal{O})$ (voir figure 5.4). L'effet observé selon lequel les entités fantômes sont plus denses dans les parties de l'espace observées par peu de caméras, est justifié par la propriété suivante.

Propriété 7

$$SFS = \mathrm{EVE}_n \subseteq \mathrm{EVE}_m \subseteq \mathrm{EVE}_0 = \mathrm{EVE}(\mathcal{O})$$

avec $m \in [1, \cdots, n]$. □

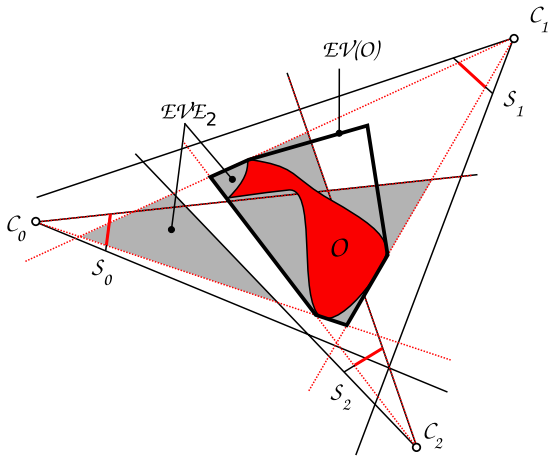


FIG. 5.4 – Représentation 2D de EVE_2 (en gris) d'un objet d'intérêt $\mathcal{O}$ et de son enveloppe visuelle $EV(\mathcal{O})$ (en noir). Tout point de $\mathcal{O}$ est contenu dans EVE_2, alors que certains points de $EV(\mathcal{O})$ n'appartiennent pas à EVE_2.

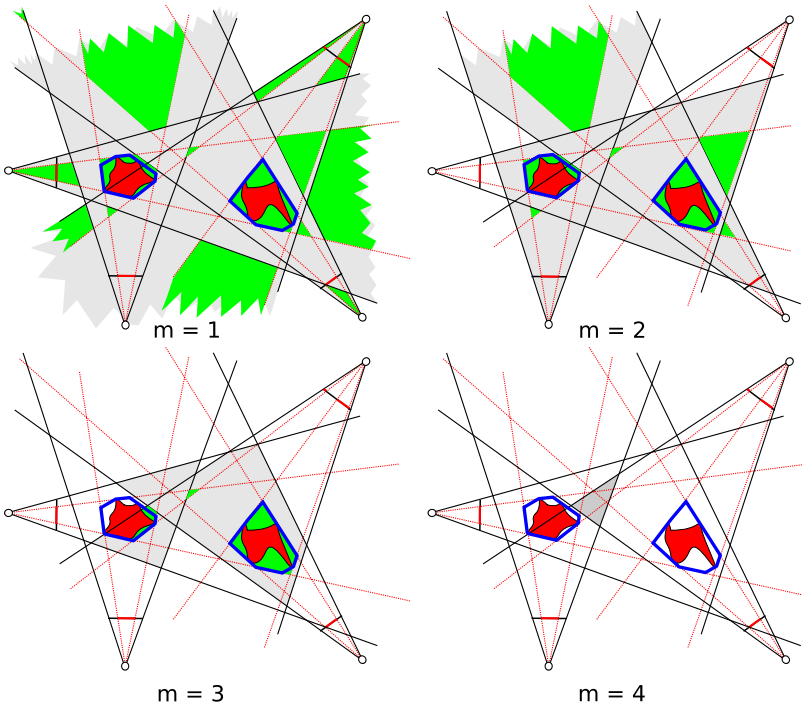


FIG. 5.5 – Représentation 2D de EVE_m de plusieurs objets d'intérêt et de l'enveloppe visuelle $EV(\mathcal{O})$ correspondante (en bleu). Pour chaque valeur de m les parties en gris représentent l'espace reconstructible et les parties en vert représentent EVE_m. Lorsque l'on augmente la valeur de m les entités fantômes diminuent. La reconstruction EVE_3 ne contient pas la totalité des objets d'intérêt car certaines parties des objets d'intérêt ne sont visible qu'à partir de deux caméras. EVE_4 est vide car aucune partie des objets d'intérêt n'est vu par toutes les caméras.

Il en résulte que EVE_m est une généralisation de *Shape-From-Silhouette* à m caméras. EVE_m estime *l'enveloppe visuelle étendue* des objets d'intérêt visibles dans aux moins m caméras. Le choix du paramètre m définit l'espace 3D reconstructible (figure 5.5). Ainsi tout point 3D X appartient à l'espace reconstructible de EVE_m si $Card(\mathcal{I}(\text{X})) \geq m$. De plus ce paramètre contrôle la quantité d'entités fantômes qui seront construites. Il est possible de diminuer la quantité d'entités fantômes, mais il n'est pas garanti des les supprimer.

Enfin le paramètre m définit la robustesse de EVE_m par rapport aux points de silhouettes faux-positifs. Pour un point 3D, la décision d'appartenance à EVE_m est prise en prenant en compte l'information d'au moins m caméras.

La valeur de m doit être choisie en fonction de l'application visée (ie, l'utilisation faite de l'estimation de la forme de $\mathcal{O}$). $m = 2$ est une valeur qui fournit un bon compromis entre la taille de l'espace de reconstruction, et l'ajout d'entités fantômes. En effet la plupart des entités fantômes sont issues de l'ensemble des points 3D consistants avec une ou aucune silhouette $\tilde{\mathcal{S}}_i$.

EVE_m estime la forme 3D de tout objet observé par m parmi n caméras. Ainsi le placement des caméras est moins contraint que pour l'approche *Shape-From-Silhouette*. Notre approche permet aussi d'étendre l'espace d'acquisition pour des configurations de caméras pensées pour *Shape-From-Silhouette*. Enfin l'algorithme qui estime EVE_m par un ensemble de voxels, est très proche de l'algorithme volumique de *Shape-From-Silhouette* :

Diviser l'espace d'intérêt en $G \times G \times G$ voxels v_k avec $k \in [1, \cdots, G^3]$;
$Card$: **entier** ;
Pour k **de** 1 **à** G^3 **faire**
  $v_k \leftarrow Dedans$;
  $Card \leftarrow 0$;
  Pour i **de** 1 **à** n **faire**
    Si ($\text{Proj}_{\pi_i}(v_k) \cap \mathcal{I}_i \neq \emptyset$) **Alors**
      $Card \leftarrow Card + 1$;
      Si ($\text{Proj}_{\pi_i}(v_k) \cap \mathcal{I}_i = \emptyset$) **Alors**
        $v_k \leftarrow Dehors$;
      Fin Si
    Fin Si
  Fin Pour
  Si ($Card < m$) **Alors**
    $v_k \leftarrow Dehors$;
  Fin Si
Fin Pour
EVE_m est approximée par l'union des voxels v_k étiquetés *Dedans* ;

Algorithme 2: Algorithme d'estimation de EVE_m à base de voxels.

Notons que Boyer et Franco ont proposé dans [BF03] une approche similaire à la notion d'enveloppe visuelle étendue, sans en étudier les propriétés. Leur solution se révèle équivalente à EVE_1. Notre contribution se distingue par son approche généralisée à toute valeur de m. De par notre formalisation, nous avons ainsi prouvé la propriété fondamentale : $\mathcal{O} \subseteq \text{EVE}_m$, tant que tout point de $\mathcal{O}$ est vu par au moins m caméras.

Dans la suite nous présentons certains résultats pratiques de l'approche EVE_m.

5.1.1 Résultats

Dans cette section nous présentons certains résultats expérimentaux de notre approche EVE_m par rapport à l'approche *Shape-From-Silhouette* classique. Une étude qualitative des résultats de EVE_m est développée dans la section 8.3.

Ayant diminué les contraintes de placement de caméras, le calibrage géométrique a été réalisé de façon spécifique. Les paramètres internes ont été déterminés en utilisant la méthode classique proposée par Zhang [Zha00]. Cependant cette même méthode n'a pu être utilisée pour le calibrage extrinsèque, car elle suppose que l'intersection des cônes de vision des caméras est non-vide. L'approche de calibrage par objets de calibrage virtuel présentée dans [SMP05] a permis de calibrer l'ensemble des caméras, par rapport à une caméra utilisée comme référence.

Chaque caméra délivre une image Bayer d'une résolution de 640×480 pixels à une cadence de 30 images par seconde. Chaque caméra est connectée à un ordinateur, équipé d'un processeur AMD ATHLON X2 3800+ disposant de 1Go de mémoire vive. Celui-ci réalise la transformation de l'image Bayer vers une image couleur. Il calcule une modélisation de fond par mixture de gaussienne [Ziv04], qui permet d'extraire de ces images les masques binaires de silhouette. Ces masques sont ensuite envoyés par réseau GigaByte vers la machine qui réalise le calcul de EVE_m. Cette dernière machine est construite autour d'un processeur AMD ATHLON X2 5600+ disposant de 1Go de mémoire vive, épaulé par une carte graphique NVIDIA Geforce 7900 GTX.

Nous avons porté l'algorithme 2 d'estimation de EVE_m à base de voxel, sur carte graphique programmable. Nous nous sommes basé sur l'approche proposée par Hazenfratz *et al.* [HLS04]. Les approches voxel sont généralement basées sur la projection des voxels dans le plan image de chaque caméra : un voxel appartient à l'*enveloppe visuelle étendue* s'il se projette dans toutes les silhouettes des caméras qui voient ce voxel. De façon à être mieux adaptée à une implémentation GPU, nous proposons d'utiliser la réciproque : Pour chaque caméra on projette à l'aide de la technique *Projective Texture Mapping* (voir [SKvW$^+$92]) le masque binaire de silhouette et le masque de vision de cette caméra dans la grille de voxels. Ainsi un voxel intersecté au minimum par m masques de visions et par autant de masques de silhouettes, appartient à EVE_m. La figure 5.6 illustre le fonctionnement de notre approche implémentée sur GPU.

Les résultats suivants représentent l'estimation de EVE_m par une grille de voxels d'une résolution de $128 \times 128 \times 128$ dans une boite de $2 \times 2 \times 2$ m, ce qui correspond à une précision de $1,56$ cm par coté de voxel. Dans la première expérimentation (voir Fig. 5.7) nous comparons

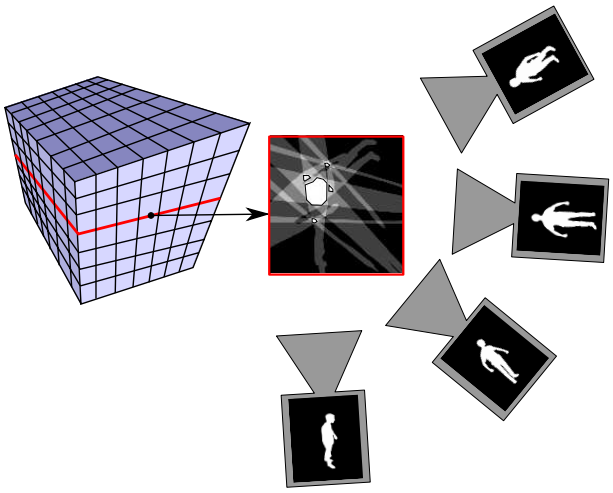


FIG. 5.6 – Le masque binaire de silhouette et le masque de vision de chaque caméra sont projetés sur chaque plan de la grille de voxel (seules les projections de masques de silhouette sont représentées). Chaque voxel intersecté par au moins m masques de vision et par le même nombre de masques de silhouettes appartient à EVE_m.

.

l'approche *Shape-From-Silhouette* classique à notre approche EVE_4. Sur les 6 caméras, 4 ont un vue partielle de l'objet à estimer. *Shape-From-Silhouette* est en échec car il ne peut rien reconstruire en dehors de l'intersection des cônes de vision des caméras. A l'opposé notre algorithme n'a aucun problème avec les vues partielles.

L'expérimentation suivante (voir Fig. 5.8) présente une configuration de caméras complexe. En effet l'intersection des cônes de vision est vide. Dans cette configuration *Shape-From-Silhouette* ne reconstruit rien. Notre approche reconstruit la totalité de l'objet d'intérêt.

Comparée à l'approche *Shape-From-Silhouette*, notre méthode fonctionne toujours en temps réel (plus de 65 estimations par seconde). Notre approche est plus flexible et permet d'estimer la forme 3D des objets d'intérêt, tant que tout point de ces objets est vu par au moins m caméras. Cependant la figure 5.9 montre une configuration dans laquelle *Shape-From-Silhouette* reconstruit la totalité des objets d'intérêt, mais où notre approche ajoute des objets fantômes.

5.1.2 Discussion

EVE_m calcule l'*enveloppe visuelle étendue* d'objets d'intérêt sans contraintes sur le placement des caméras. Cette approche ne nécessite aucun calcul supplémentaire par rapport à l'approche *Shape-From-Silhouette* et fonctionne en temps réel. Nous avons montré que la reconstruction 3D offerte par EVE_m contient tout point 3D des objets d'intérêt tant que toute partie de ceux-ci est visible par au moins m caméras.

La contrepartie de la suppression des contraintes de placement de caméras est la généra-

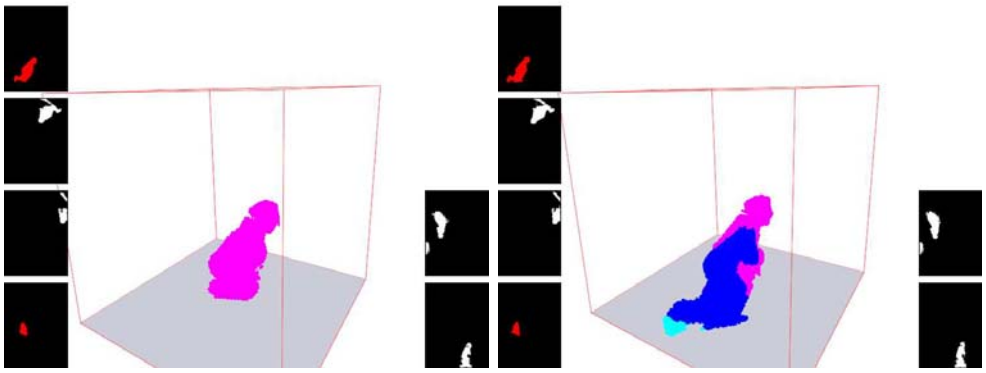


FIG. 5.7 – Une personnes à genoux est capturée par 6 caméras, 2 caméras voient entièrement la personne à reconstruire (silhouette en rouge), et 4 la voit partiellement (silhouette en blanc). A gauche certaines parties du corps ne sont pas reconstruites par *Shape-From-Silhouette*. A droite EVE_4 contient la totalité de l'objet d'intérêt. Les couleurs indiquent le nombre de caméras qui voient chaque point ; le magenta indique 6 caméras, le bleu indique 5 caméras, et le cyan indique 4 caméras.

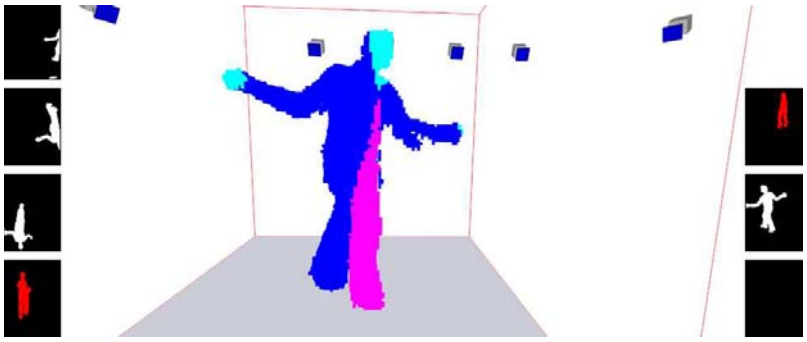


FIG. 5.8 – Reconstruction par EVE_4 d'une personne traversant différents champs de vue de caméras. La scène est filmée par 7 caméras, certaines voient la totalité de l'objet à estimer (en rouge), certaines le voient partiellement alors que certaines ne le voient pas du tout. La couleur indique le nombre de caméras qui voient chaque voxel. Notre système n'impose pas de contraintes sur le placement de caméras tant que tout point 3D a estimer est vu par au moins m caméras.

tion d'objets fantômes supplémentaires. Pour répondre à cette limitation nous avons proposé le paramètre m. Ce paramètre définit l'espace 3D reconstructible, c'est à dire que tout objet à l'intersection d'au moins m cônes de vision sera reconstruit par EVE_m. Il contrôle aussi l'ajout d'objets fantômes. Dans la suite nous proposons plusieurs méthodes, qui en fonction d'hypothèses particulières ou encore à l'aide d'informations supplémentaires comme la couleur, sont capables de supprimer les objets fantômes ajoutés par EVE_m.

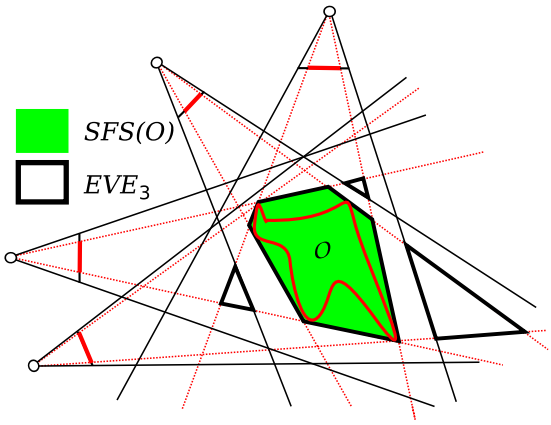


FIG. 5.9 – Configuration pour laquelle *Shape-From-Silhouette* reconstruit la totalité des objets d'intérêt. Cependant EVE_3 construit plus d'objets fantômes que l'approche classique *Shape-From-Silhouette*.

5.2 Hypothèse de visibilité partielle

Lors de nos premières expérimentations de *Shape-From-Silhouette* nous avons rapidement observé les contraintes de placement des caméras. C'est à dire qu'il est difficile de suivre les déplacements de personnes filmées lors de grands mouvements. En général celles-ci sont totalement visibles dans toutes les caméras, mais lors de mouvements rapides, comme par exemple la course, celles-ci deviennent partiellement visibles dans certaines caméras. Dans ce contexte il est possible d'étendre l'espace d'acquisition sans ajouter d'objet fantôme tant que tous les objets à estimer sont partiellement visibles dans toutes les caméras. C'est à dire que l'*on suppose que pour chaque objet de la scène, il existe un point de cet objet observé depuis toutes les caméras.* Chaque objet d'intérêt sera au moins reconstruit en partie par *Shape-From-Silhouette*. Nous déterminons ensuite les parties manquantes qui appartiennent à EVE et qui sont connectées aux formes reconstruites par *Shape-From-Silhouette*.

Par construction EVE est le volume maximal silhouette-équivalent avec les objets d'intérêt. Certaines composantes connexes CC_j de EVE contiennent donc des objets réels. Nous supposons que tout objet d'intérêt est partiellement reconstruit par $SFS = \text{EVE}_n$. Ainsi nous définissons l'approche *SFpS Shape-From-Partial-Silhouette*, union des composantes connexes de EVE qui intersectent la reconstruction offerte par EVE_n :

Définition 9

$$SFpS = \bigcup CC_j \subseteq \text{EVE} : \exists CC_i \subseteq \text{EVE}_n \ et \ CC_i \subseteq CC_j.$$

□

On en déduit la propriété suivante :

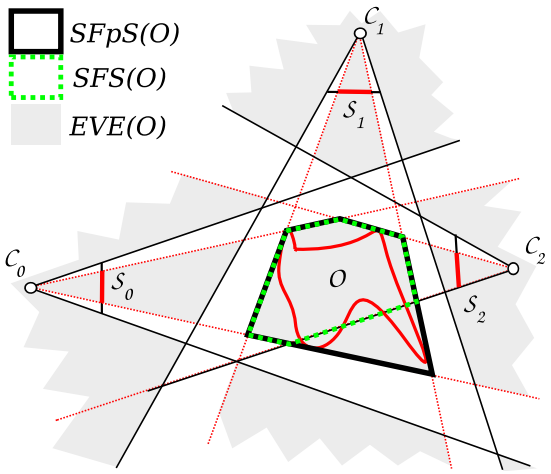


FIG. 5.10 – Illustration en 2D de la reconstruction de forme générée par $SFpS$. Les composantes connexes de EVE qui intersectent les composantes connexes de $\text{EVE}_n = SFS$ appartiennent à $SFpS$.

.

Propriété 8 *$SFpS$ n'ajoute pas d'objets fantômes par rapport à SFS.* □

Preuve $SFPpS$ est formé par les composantes connexes de EVE qui contiennent des composantes connexes de $\text{EVE}_n = SFS$, alors $SFpS$ contient au maximum le même nombre de composantes connexes que SFS et n'ajoute aucun objet fantôme. □

La figure 5.10 illustre la reconstruction de forme à l'aide de $SFpS$. Notre approche reconstruit la totalité des objets réels et n'ajoute pas d'objets fantômes.

5.2.1 Résultats

Notre algorithme a été testé sur différents jeux de données réelles, acquis depuis 4 caméras avec une résolution de 320×240 pixels.

La figure 5.11 montre les résultats obtenus sur un objet complexe. Il y a visibilité partielle dans les silhouettes 1, 3 et 4. L'approche classique SFS fourni une reconstruction partielle de l'objet (voir figure 5.11.a). Notre algorithme permet une reconstruction plus complète de la chaise du fait que les portions non visibles depuis une caméra sont visibles depuis les autres caméras. Le pied de la chaise n'est, quant à lui, pas reconstruit complètement, du fait qu'il n'est pas visible depuis la plupart des caméras.

La figure 5.12 montre les résultats obtenus lors de l'acquisition d'une personne en mouvement. Notre méthode ajoute de nouvelles informations valides (du point de vue des images de silhouette) à l'estimation de la forme de l'objet.

Les performances de notre approche sont très proches des temps de calculs de SFS. La différence provient du calcul des composantes connexes de EVE. Le temps de parcours étant négligeable par rapport au temps de calculs de la projection des voxels sur chaque silhouette. L'implémentation expérimentale de notre algorithme permet 60 estimations de forme par secondes (pour une résolution voxélique de 128^3), alors que notre implémentation de SFS atteint les 65 estimations de forme par seconde. Ces résultats illustrent le fait que notre algorithme est utilisable dans le cadre d'applications visant le temps réel.

5.2.2 Discussion

L'approche *Shape-From-Partial-Silhouette* permet de palier au problème de visibilité partielle des objets réels dans les caméras. Elle permet de reconstruire la totalité des objets, lorsque ceux-ci sont tous partiellement visibles dans toutes les caméras, et cette reconstruction est silhouette-équivalente.

Notre solution à l'avantage de ne pas ajouter d'objet fantômes par rapport à l'approche *Shape-From-Silhouette*. Elle ne nécessite aucune information supplémentaire, en dehors des paramètres de calibrage géométrique des caméras, ainsi que des masques de silhouettes. De plus, nous n'ajoutons qu'une étape de calcul par rapport *Shape-From-Silhouette*, ce qui permet de toujours fonctionner en temps réel. Les hypothèses de fonctionnement sont validées en présence d'un système réalisé pour l'approche *Shape-From-Silhouette* classique, et permet d'étendre l'espace d'acquisition. Cependant, dans des configurations plus complexes de caméras, comme par exemple l'intersection des cônes de vision de toutes les caméras vide, $SFpS$ est incapable de reconstruire les objets de la scène.

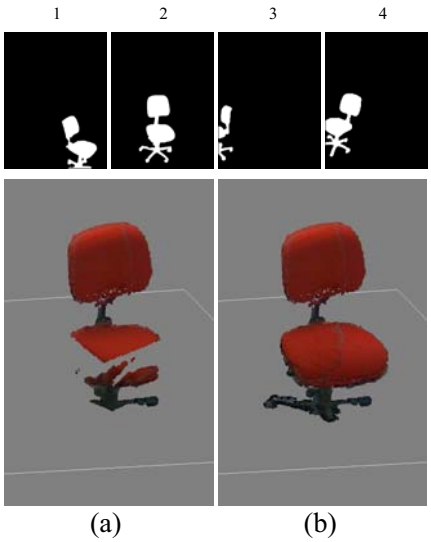


FIG. 5.11 – Estimation voxélique de la forme d'un objet complexe : (a) algorithme de base de SFS ; (b) notre algorithme $SFpS$.

.

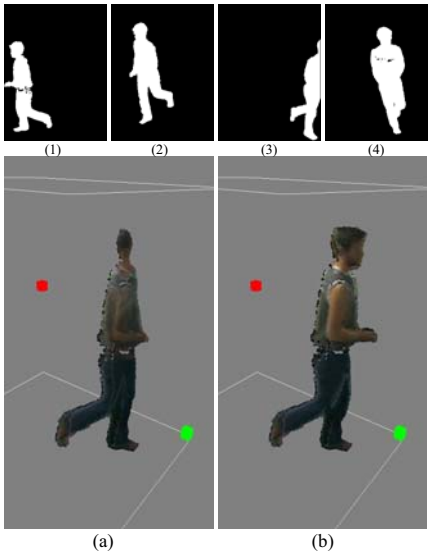


FIG. 5.12 – Différences de reconstruction d'un objet lorsque celui-ci sort du champs de vision d'une caméra : (a) en utilisant les méthodes basées SFS actuelles ; (b) en utilisant notre algorithme. Le coloriage des voxels (réalisé par un lancer de rayon) n'est donné que pour faciliter la compréhension des images.

.

5.3 Critère de qualité maximale

L'hypothèse de visibilité partielle présente une contrainte de placement de caméras, même si celle-ci est moins forte que la contrainte de placement de *Shape-From-Silhouette*. Supposons une configuration de caméra pour laquelle l'intersection des cônes de vues est vide. Alors l'approche $SFpS$ ne peut s'appliquer. La notion de qualité présentée propriété 3 permet d'étendre l'approche $SFpS$. Cela revient à supposer que les parties de l'espace de meilleur qualité (vue par le plus grand nombre de caméras) sont les parties dans lesquelles l'enveloppe visuelle étendue contient le moins d'objets fantômes.

Soit n_{max} le nombre maximum de caméras dans lesquelles un point 3D est silhouette-consistant. C'est à dire que

$$\mathrm{EVE}_{n_{max}} \neq \emptyset \text{ et } \forall m > n_{max}, \mathrm{EVE}_m = \emptyset \tag{5.5}$$

Ainsi nous définissons $SFpeS$ (*Shape-From-Partial-or-Empty-Silhouette*) une extension de $SFpS$ de la manière suivante :

Définition 10

$$SFpeS = \bigcup CC_j \subseteq \mathrm{EVE} : \exists CC_j \subseteq \mathrm{EVE}_{n_{max}}, CC_j \subseteq CC_j$$

□

$SFpS$ devient ainsi un cas particulier de $SFpeS$ avec $n_{max} = n$. La reconstruction de $SFpeS$ suppose que les objets fantômes sont reconstruits dans les parties qui ne sont pas de qualité maximale. Cependant $SFpeS$ ne garantit pas de contenir les objets réels. La figure 5.13 représente une configuration qui prouve que le critère de qualité ne permet pas de discriminer les objets fantômes.

Même si $SFpeS$ semble une méthode intéressante, cette approche peut supprimer certains objets réel. Cependant cela n'est pas le cas de l'approche $SFpS$ qui ne supprime pas d'objets réels, tant que ceux-ci sont partiellement visibles dans toutes les caméras.

5.4 Cohérence projective

Nous avons observé que EVE ajoute des objets fantômes par rapport à l'approche *Shape-From-Silhouette*. Cependant il est possible les supprimer lorsque l'on connaît les correspondances entre les composantes connexes de chaque silhouette. Pour présenter notre approche, nous ferons **dans un premier temps, l'hypothèse qu'il n'y a qu'un seul objet dans la scène et que celui-ci est inclus dans** EVE_m. C'est à dire que l'on cherche dans la reconstruction EVE_m l'unique composante connexe qui contient l'objet d'intérêt.

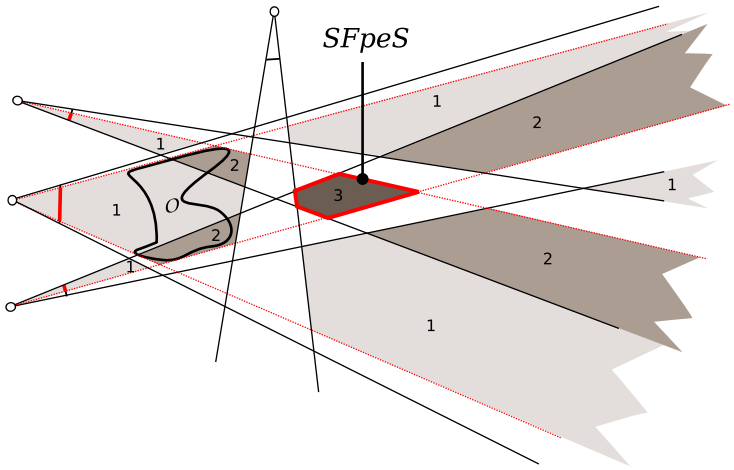


FIG. 5.13 – Représentation d'une coupe 2D de $SFpeS$ basée sur le critère de qualité maximum, c'est à dire les parties de EVE reconstruites à l'intersection du nombre maximal de cônes de vision. Ce critère n'est pas une condition suffisante pour distinguer les objets fantômes des objets réels. $SFpeS$ encadré en rouge, ne contient pas l'objet d'intérêt. Les parties grises représentent EVE_1.

Nous introduisons la notation cc_j qui représente une composante connexe 2D de l'une des silhouette observable $\tilde{\mathcal{S}}_i$.

Définition 11 *Soit* $\widetilde{\text{cc}}(\text{X})$ *l'ensemble des composantes connexes 2D des silhouettes* $\tilde{\mathcal{S}}(\text{X})$ *dans lesquelles se projette* X *:*

$$\widetilde{\text{cc}}(\text{X}) = \{j : \exists i \in [1, \cdots, n] \text{Proj}_{\pi_i}(\text{X}) \in \text{cc}_j\}$$

□

Notons que lorsqu'il n'y a qu'un objet d'intérêt dans la scène $\widetilde{\text{cc}}(\text{X}) = \tilde{\mathcal{S}}(\text{X})$.

Nous étendons la définition précédente aux ensembles de points 3D :

Définition 12 *Soit* $\widetilde{\text{cc}}(\text{E})$ *l'ensemble des composantes connexes 2D des silhouettes* $\tilde{\mathcal{S}}(\text{X})$ *dans lesquelles se projette tout point* X *de l'ensemble* E *:*

$$\widetilde{\text{cc}}(E) = \{j : \exists \text{X} \in E, j \in \widetilde{\text{cc}}(\text{X})\}$$

□

Propriété 9 *Soit* CC_j *l'une des composantes connexes de* EVE_m

$$\widetilde{\text{cc}}(CC_j) \neq \widetilde{\text{cc}}(\mathcal{O}) \Rightarrow CC_j \in \text{O}_{fantôme}$$

□

Preuve Démontrer ce théorème revient à démontrer la contraposée : c'est a dire $CC_j \notin \mathrm{O}_{fantôme} \Rightarrow \widetilde{\mathrm{cc}}(\mathcal{O}) = \widetilde{\mathrm{cc}}(CC_j)$. Nous commençons par démontrer $CC_j \notin \mathrm{O}_{fantôme} \Rightarrow \widetilde{\mathrm{cc}}(\mathcal{O}) \subseteq \widetilde{\mathrm{cc}}(CC_j)$.

Supposons que CC_j n'est pas un objet fantôme, alors par définition CC_j contient au moins un point appartenant à $\mathcal{O}$. Par hypothèse il n'y a qu'un objet dans la scène alors $\mathcal{O} \subseteq CC_j$.

Par définition $\forall i \in \widetilde{\mathrm{cc}}(\mathcal{O}), \exists \mathrm{X} \in \mathcal{O}, \mathrm{Proj}_{\pi_i}(\mathrm{x}) \in \widetilde{\mathrm{cc}}_i$. Or $\mathcal{O} \subseteq CC_j$ implique $\forall \mathrm{X} \in \mathcal{O}, \mathrm{X} \in CC_j$. Alors

$$\forall i \in \widetilde{\mathrm{cc}}(CC_j), \exists \mathrm{X} \in \mathcal{O}, \mathrm{Proj}_{\pi_i}(\mathrm{X}) \in \widetilde{\mathrm{cc}}(CC_j)$$

ainsi

$$CC_j \notin \mathrm{O}_{fantôme} \Rightarrow \widetilde{\mathrm{cc}}(\mathcal{O}) \subseteq \widetilde{\mathrm{cc}}(CC_j) \text{ est vrai.}$$

Nous souhaitons démontrer que $CC_j \notin \mathrm{O}_{fantôme} \Rightarrow \widetilde{\mathrm{cc}}(\mathcal{O}) = \widetilde{\mathrm{cc}}(CC_j)$. Supposons $i \in \widetilde{\mathrm{cc}}(CC_j), i \notin \widetilde{\mathrm{cc}}(\mathcal{O})$. Alors il existe un point 2D dans l'une des caméras qui ne résulte pas de la projection de $\mathcal{O}$. Donc il existe un autre objet dans la scène, ce qui est en contradiction avec l'hypothèse qu'il n'y à qu'un objet dans la scène.

On en déduit que

$$CC_j \notin \mathrm{O}_{fantôme} \Rightarrow \widetilde{\mathrm{cc}}(\mathcal{O}) = \widetilde{\mathrm{cc}}(CC_j) \text{ est vrai}$$

□

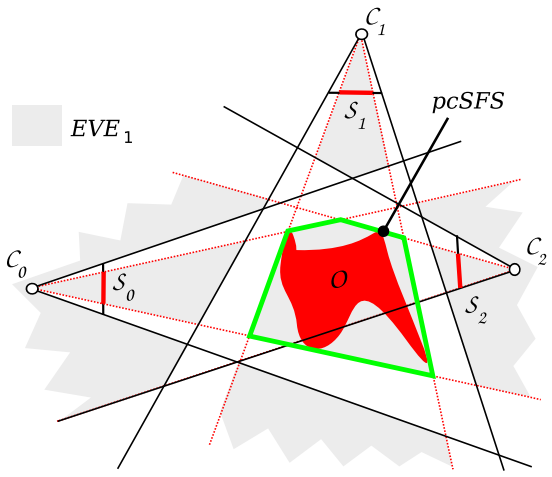


FIG. 5.14 – Représentation d'une coupe 2D de l'espace reconstruit par *pcSFS*. L'objet d'intérêt est contenu dans la reconstruction. Enfin cette configuration *pcSFS* ne contient aucun objet fantôme.

Cette propriété est particulièrement importante. En effet pour chaque CC_j de EVE_m, si $\widetilde{\mathrm{cc}}(\mathcal{O}) \neq \widetilde{\mathrm{cc}}(CC_j)$, alors CC_j est un objet fantôme et est ainsi supprimée de la reconstruction. De plus avec cette méthode nous garantissons que les parties supprimées ne contiennent pas

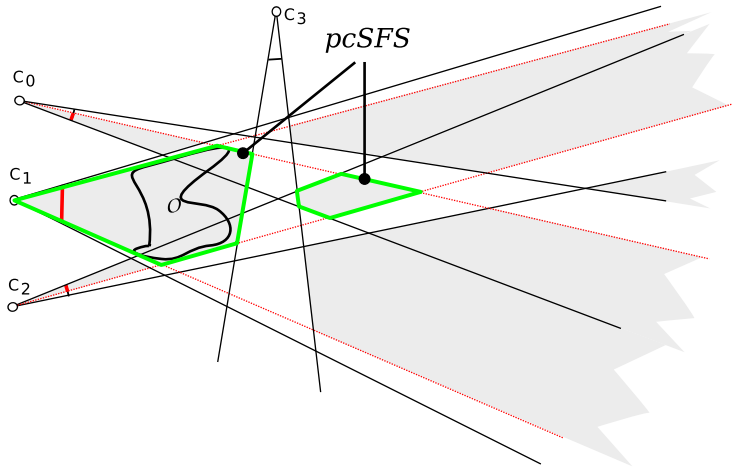


FIG. 5.15 – Représentation d'une coupe 2D de l'espace reconstruit par *pcSFS* pour la configuration proposée figure 5.13. *pcSFS* contient l'objet d'intérêt ainsi qu'un objet fantôme. L'approche garantit de reconstruire tous les objets d'intérêt. Cependant elle ne garantit pas de supprimer tous les objets fantômes.

l'objet réel. Il est important de noter que la réciproque de cette propriété n'est pas vrai. En effet la figure 5.15 présente une configuration où l'une des composantes connexes $CC_k \in \text{EVE}_m$ est un objet fantôme, alors que $\widetilde{\text{cc}}(\mathcal{O}) = \widetilde{\text{cc}}(CC_k)$.

Nous nommons par la suite *pcSFS* (*Projective-Coherent-Shape-From-Silhouette*) la méthode qui permet d'estimer la forme 3D d'un unique objet, à partir d'une configuration quelconque de caméras et ce sans ajouter d'objets fantômes :

Définition 13

$$pcSFS = \{CC_j \in \text{EVE}_m, \widetilde{\text{cc}}(CC_j) = \widetilde{\text{cc}}(\mathcal{O})\}$$

□

Nous avons fait l'hypothèse qu'il n'y ait qu'un seul objet dans la scène, alors $\widetilde{\text{cc}}(\mathcal{O})$ est l'ensemble des caméras qui ont une information de silhouette. Les figures 5.14 et 5.15 présentent la reconstruction de EVE_m dont les objets fantômes ont été discriminés par la propriété 9 et ainsi supprimés. De plus cette reconstruction est toujours silhouette-équivalente.

Nous venons de proposer *pcSFS* une approche qui permet de reconstruire un unique objet et ce sans ajouter d'objets fantômes, pour une configuration quelconque de caméras tant que toute partie de l'objet réel est totalement visible dans m caméras. De plus cette approche ajoute une seule étape de calcul supplémentaire qui correspond à l'étiquetage des parties connexes de EVE_m, dont le temps de calcul est négligeable par rapport aux calcul de la projection de la grille de voxel dans chaque caméra. Le fait que ce système soit limité à la reconstruction d'un

seul objet est une contrainte forte. Nous nous intéressons maintenant à généraliser le principe de *cohérence-projective* dans le cas de plusieurs objets.

5.4.1 Généralisation à plusieurs objets

Le théorème 9 est construit selon plusieurs hypothèses implicites. La première est de pouvoir calculer $\widetilde{\text{cc}}(\mathcal{O})$. Si il y a plusieurs objets dans la scène, il n'est pas possible, simplement à partir de critères géométriques, de calculer directement $\widetilde{\text{cc}}(\mathcal{O}_i)$ pour chaque objet $\mathcal{O}_i \subset \mathcal{O}$. C'est à dire qu'il faut pour chaque objet $\mathcal{O}_i$ déterminer dans quelles composantes connexes de silhouette il se projette. Cela revient à mettre en correspondances les composantes connexes de silhouettes dans toutes les caméras. Pour réaliser ces appariements, l'information de masques binaires de silhouettes et le calibrage géométrique des caméras ne suffisent pas. Nous avons utilisé la méthode proposée dans [Jos04], reconnue pour la qualité des appariements réalisés à partir de l'information de couleur.

Ayant les correspondances entre les composantes connexes de silhouettes, nous connaissons alors pour chaque objet réel les caméras qui le voient. Alors pour chaque objet nous pouvons nous ramener à l'hypothèse d'objet unique à reconstruire. C'est à dire que pour chaque objets $\mathcal{O}_i \subset \mathcal{O}$, on calcule séparément $pcSFS_i$, seulement à partir des composantes connexes des silhouettes qui sont associées à $\mathcal{O}_i$. Dans la suite de ce document $pcSFS$ décrit l'approche $pcSFS$ généralisée à plusieurs objets.

La figure 5.16 présente le résultat de la généralisation du critère de *cohérence-projective* à plusieurs objets. En présence d'objets de couleur similaire cette approche risque de fournir des résultats erronés. Lorsque qu'un objet d'intérêt occulte partiellement un autre objet depuis le point de vue d'une caméra, l'algorithme de mise en correspondance risque d'être mis en échec. Alors les conditions optimales pour le fonctionnement de cette approche sont que tout objet est visible depuis tout point de vue et qu'il n'y a pas d'occultations.

La généralisation du critère de cohérence projective est semblable à l'approche proposée par Bogomjakov et Gotsman dans [BG08]. Cependant dans leur approche, ils ne fournissent pas de justification formelle, concernant les propriétés de la forme reconstruite.

D'un côté cette généralisation permet d'estimer la forme 3D de plusieurs objets d'intérêt dans le cadre d'un positionnement libre de toutes les caméras, tout en supprimant une bonne partie des objets fantômes. De l'autre côté cette méthode nécessite le calcul des correspondances entre les composantes connexes de silhouettes. Or il n'existe pas de solution optimale pour résoudre ce problème et cela nécessite une information complémentaire de couleur. Enfin en présence d'objets d'intérêt d'apparence semblable, il faut prendre en compte la cohérence temporelle, de façon à mener à bien le processus de mise en correspondance. Enfin si le processus de mise en correspondance échoue, la reconstruction proposée risque de ne pas contenir les objets d'intérêt.

Nous présentons dans la suite certains résultats expérimentaux qui valident l'intérêt du critère de cohérence projective.

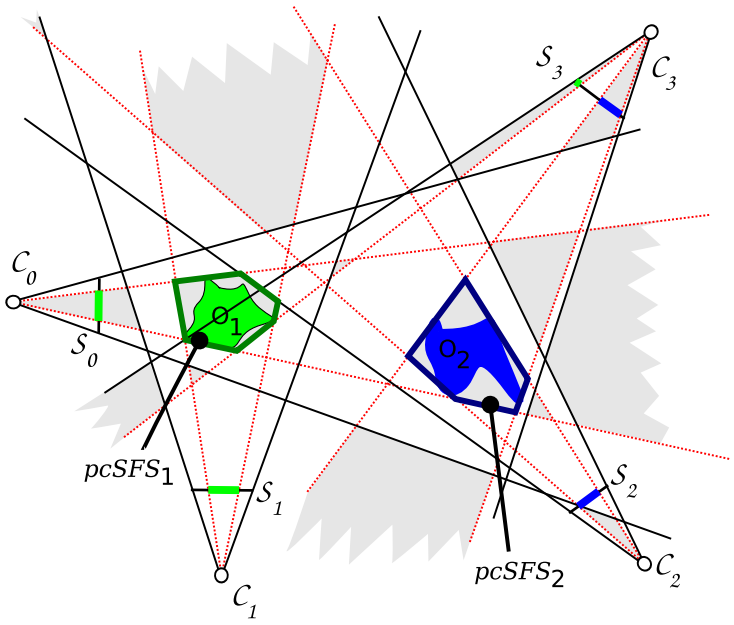


FIG. 5.16 – Illustration du fonctionnement de la généralisation du critère de *coherence-projective* en présence de plusieurs objets d'intérêt. La reconstruction est définie par l'union des reconstructions $pcSFS_i$.

5.4.2 Résultats

Dans cette section nous présentons les résultats expérimentaux du critère de cohérence projective. Une analyse quantitative et qualitative détaillée de toutes nos contributions est présentée dans le chapitre 8.

Les expérimentations suivantes ont été réalisées sur la plate forme présentée dans la section 5.1.1. La figure 5.17 présente les résultats du critère de cohérence projective en présence d'un seul objet d'intérêt. Ce critère a été appliqué sur la reconstruction de la scène proposée par EVE_1. Notons que la totalité des objets fantômes a été supprimée. La reconstruction présentée figure 5.17 illustre l'efficacité du critère de cohérence projective.

Il est important de noter la principale limitation de cette approche : le test de cohérence-projective est construit sur la supposition qu'il n'y a qu'un objet dans la scène. Ne pas respecter cette hypothèse revient à invalider la Propriété 9. Ce critère est basé sur la connaissance de la correspondance entre les composantes connexes de silhouettes (avec un seul objet, la correspondance est triviale). Pour plusieurs objets, nous calculons alors une mise en correspondance basée sur l'information de couleur. Ayant les correspondances, le test de cohérence projective est calculé séparément pour chaque objet en connaissant les composantes connexes de silhouette associées. Le critère est alors vérifié en prenant en compte l'information issue des caméras où il n'y a pas d'occultation partielle. La figure 5.19 présente le résultat de la généralisation du critère de cohérence projective en présence de plusieurs objets. Notons que si le processus de

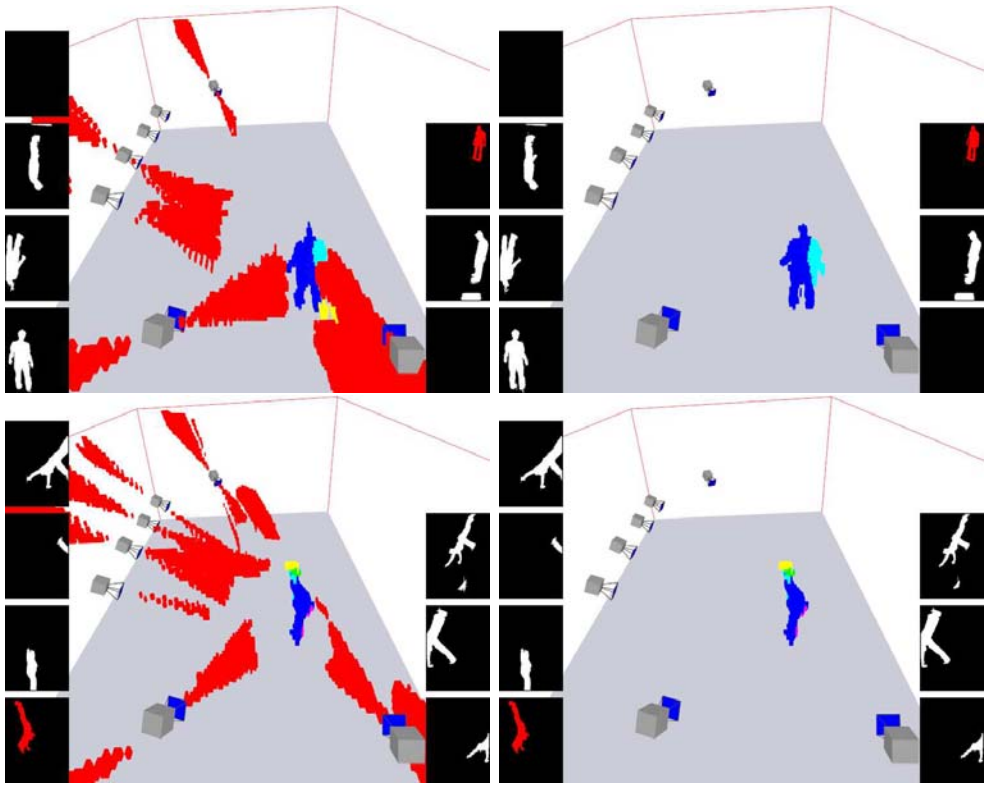


FIG. 5.17 – Deux trames différentes (haut et bas) capturée par notre plate forme. Le sujet à reconstruire est à la fois totalement visible (silhouettes rouges), partiellement visible (silhouettes blanches) ou encore non visible par chaque caméra. Les images de gauche présentent la reconstruction fournie par EVE_1. Le critère cohérence-projective supprime la totalité des objets fantômes (images de droite), qui sont des parties qui ne se projettent pas dans toutes les silhouettes de l'objet d'intérêt. Le code couleur indique le nombre de caméras qui voient chaque voxel : rouge = 1, jaune = 2, vert = 3, cyan = 4, bleu = 5, magenta = 6 et noir =7.

mise en correspondance est erroné, le test de cohérence projective peut supprimer des parties qui contiennent des objets d'intérêt.

Le test de cohérence projective nécessite un temps de calcul négligeable face au calcul de EVE_m. Ainsi le calcul de EVE_m et du test de cohérence-projective dans le cas d'un seul objet d'intérêt permet de calculer la forme de l'objet d'intérêt plus de 60 fois par secondes, pour une grille de voxel de résolution 128^3. En présence de plusieurs objets, l'étape de mise correspondance pénalise le système qui est alors capable d'estimer 30 reconstructions par seconde.

5.4.3 Discussion

Nous avons proposé un critère qui permet, en présence d'un seul objet d'intérêt, de conserver les parties de l'*enveloppe visuelle étendue* qui sont garanties de contenir cet objet d'intérêt. Dans sa première version, ce critère ne nécessite pas d'information supplémentaire que les cartes

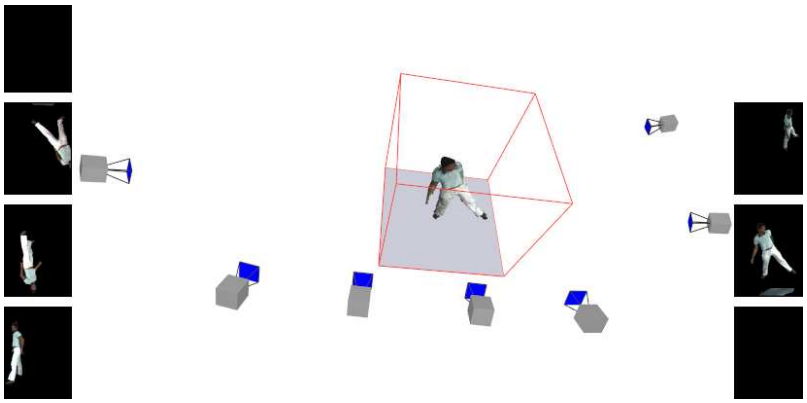


FIG. 5.18 – Résultat coloré de la reconstruction d'une personne se déplaçant dans la scène. Les images binaires de silhouettes sont représentées sur les extrémités gauches et droites. Deux caméras ne voient pas l'objet à reconstruire. EVE_1 contient l'objet d'intérêt, et le critère de *cohérence projective* permet de supprimer tous les objets fantômes.

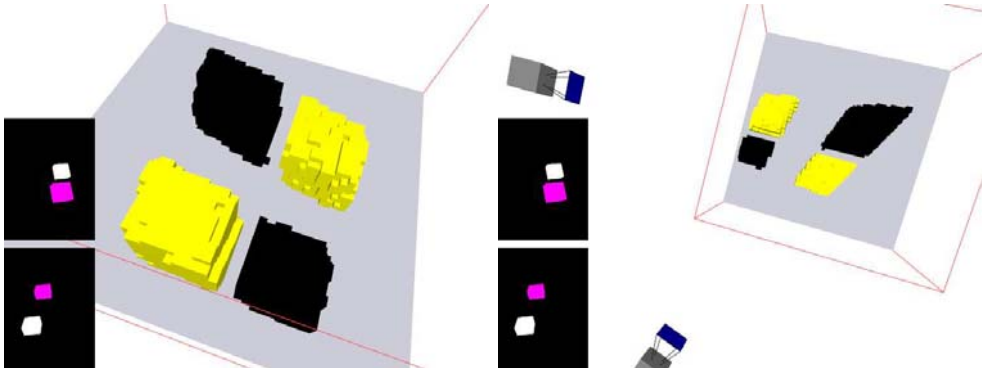


FIG. 5.19 – Deux points de vue d'une même reconstruction de deux objets d'intérêt par l'approche $\text{EVE}_2 = SFS$, filtrée par le critère de cohérence projective. Parmi les quatre parties estimées, les objets fantômes en noir sont identifiés par le critère de cohérence projective, alors que les parties jaunes contiennent les objets d'intérêt. Les couleurs observées dans les silhouettes indiquent la mise en correspondance des composantes connexes.

binaires de silhouettes, et le calibrage géométrique. Cependant la supposition qu'il n'y ait qu'un seul objet dans la scène est très restrictive. Nous avons proposé une généralisation de ce critère à plusieurs objets, lorsque la correspondance entre les composantes connexes de silhouette est disponible. Cette nouvelle approche est intéressante en théorie, mais le calcul pratique des correspondances est un problème difficile. Cette solution ne nous semble pas pertinente, dans la mesure ou le raisonnement est construit sur une information supplémentaire dont l'extraction n'est pas, à ce jour, suffisamment robuste.

5.5 Conclusion

A travers ce chapitre nous avons proposé une généralisation des enveloppes visuelles appelées *enveloppes visuelles étendues* (EVE) construites sur l'information de silhouette observable. L'enveloppe visuelle étendue d'objet d'intérêt à la propriété de contenir la totalité des objets d'intérêt, quelle que soit la configuration des caméras. Cependant cette formulation crée beaucoup d'entités fantômes. Nous avons alors proposé EVE_m , l'enveloppe visuelle étendue des objets d'intérêt, observés par au moins m caméras. Le paramètre m définit ainsi l'espace dans lequel tout objet d'intérêt est construit, et limite la construction d'entités fantômes. Ainsi EVE_m fournit un bon compromis entre une faible contrainte de placement des caméras et la génération de relativement peu d'entités fantômes.

Relaxer les conditions de fonctionnement de *Shape-From-Silhouette* par l'approche EVE_m a comme conséquence d'ajouter des entités fantômes. Nous avons proposée plusieurs critères qui limitent leurs apparitions.

Le premier critère dit de *visibilité partielle* n'ajoute pas d'objet fantômes par rapport aux enveloppes visuelles, mais suppose qu'une partie de tout objet à reconstruire est observée par chaque caméra. Avec cette approche les contraintes de placement de caméras ont été diminuées. Cette méthode s'applique particulièrement bien lorsque l'on souhaite étendre l'espace d'acquisition d'un configuration de caméras prévue pour *Shape-From-Silhouette*.

Le dernier critère de *cohérence-projective* permet de supprimer sans information supplémentaire, la plupart des objets fantômes issus de EVE, lorsqu'il n'y a qu'un objet d'intérêt dans la scène. Nous avons proposé une généralisation de ce critère en présence de plusieurs objets. Il est important de noter que la généralisation de la cohérence-projective pour plusieurs objets, apporte une solution générale pour la suppression des objets fantômes, que ce soit dans le cadre de EVE_m, mais aussi de toutes les approches qui estiment l'enveloppe visuelle dont bien sûr *Shape-From-Silhouette*. Cependant sa dépendance à l'étape de mise en correspondance entre les composantes connexes de silhouettes, fait que cette approche n'est pas efficace en présence d'occultations entre les objets.

Nous avons proposé plusieurs méthodes pour limiter la construction d'entités fantômes spécifiquement par les approches EVE ou encore EVE_m. Grâce à ces approches nous avons ensuite développé de nouvelles méthodes qui suppriment les objets fantômes générés par toute méthode d'estimation de forme à partir de silhouette : que ce soit EVE_m ou encore *Shape-From-Silhouette*. Nous présentons ces nouvelles approches dans le chapitre suivant. Les solutions que nous proposons sont uniquement construits sur des critères géométriques.

Chapitre 6

Suppression des objets fantômes

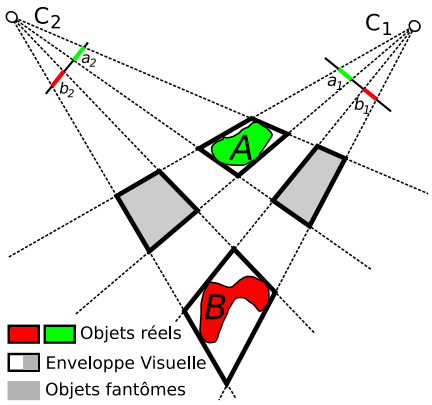


FIG. 6.1 – Configuration typique pour laquelle des objets fantômes apparaissent.

Les enveloppes visuelles et les enveloppe visuelles étendues estiment la forme 3D d'objets d'intérêt à partir des silhouettes de ces objets. Les volumes estimés produisent les mêmes silhouettes que les objets d'origine et ont la propriété de contenir tout point 3D des objets d'intérêt. Cependant ces approches ajoutent des points cohérents par rapport aux silhouettes qui n'appartiennent pas aux objets d'intérêt. Ces points sont nommés entités fantômes. Nous avons proposé dans le chapitre 4.5.2 de distinguer deux types d'entités fantômes : les parties fantômes et les objets fantômes. Les objets fantômes sont les composantes connexes de la reconstruction, qui ne contiennent aucun point 3D des objets d'intérêt. Les objets fantômes apparaissent lorsque la configuration des objets par rapport aux caméras est telle que les régions de l'espace occupées par des objets d'intérêt ne peuvent être distinguées des régions vides qui se projettent dans toutes les silhouettes. Ceux-ci sont générés par l'intersection de cônes de silhouettes qui appartiennent à des objets différents. Supposons deux objets A et B qui se projettent en des silhouettes a_1,b_1 et a_2, b_2 dans deux images (voir figure 6.1). Les objets corrects sont calculés par l'intersection des couples de cônes issus de a_1,a_2 et b_1,b_2. Deux objets fantômes sont générés par l'intersection des cônes issus de a_1,b_2 et a_2,b_1.

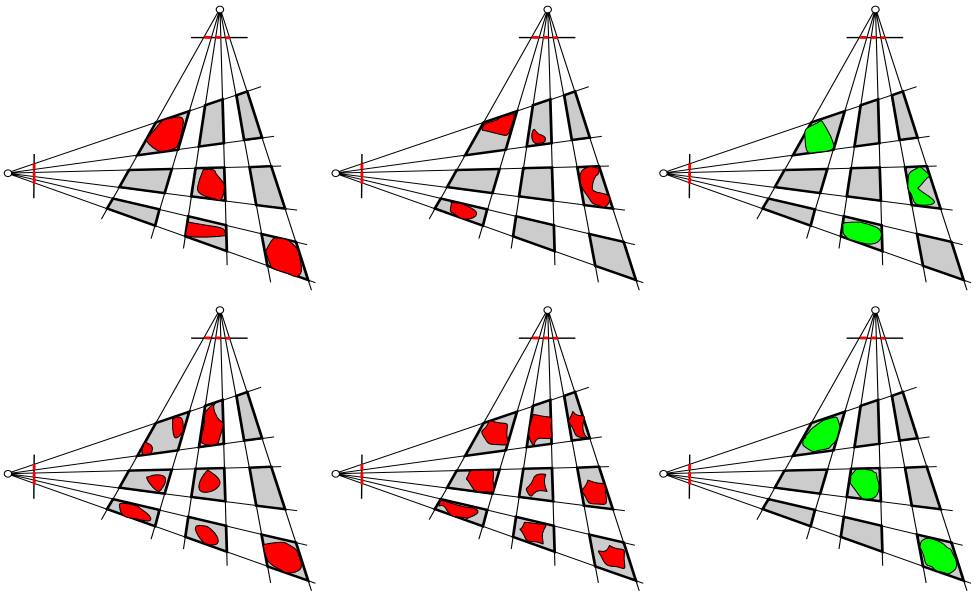


FIG. 6.2 – Représentation 2D de plusieurs configurations qui fournissent exactement la même EVE_2 (qui est dans cette configuration équivalente à EV). Cette illustration démontre que seulement à partir des silhouettes, il n'est pas possible de proposer une solution optimale qui supprime exactement tous les objets fantômes. Les objets représentés en vert appartiennent aux configurations silhouettes-équivalentes de EVE_2 dont le nombre d'objets est minimal.

Le nombre d'objets fantômes augmente lorsque le nombre d'objets d'intérêt augmente, surtout lorsque le nombre de caméras est faible. En effet il devient plus difficile d'avoir une caméra qui permet de séparer les objets dans la scène et de détecter les régions inoccupées de l'espace.

Nous rappelons que notre but est de réaliser le suivi et l'acquisition de mouvements de plusieurs individus. Dans ce contexte la présence d'objets fantômes peut nuire à la robustesse de l'acquisition de mouvements. Réaliser le suivi de plusieurs objets implique que le nombre d'objets reconstruits soit stable à travers le temps, or les objets fantômes apparaissent de façon non contrôlée.

L'une des propriétés fondamentales des approches EV et EVE est que tout point qui n'appartient pas à ces volumes, ne peut appartenir à un objet d'intérêt. Dans ce chapitre nous proposons deux approches dont la propriété est complémentaire : les composantes connexes conservées ne sont pas des objets fantômes. Parmi l'ensemble des travaux menés sur les enveloppes visuelles, il en existe peu qui permettent de supprimer les objets fantômes (voir le chapitre 4.5.2). Il faut cependant souligner le fait qu'il existe une infinité de configurations d'objets réels qui fournissent la même enveloppe visuelle (voir figure 6.2). Le problème de suppression des objets fantômes est sous-contraint et il ne peut exister de solution "optimale". Cependant, nous verrons dans la suite qu'il est possible de définir un critère qui garantit qu'une composante connexe de l'enveloppe visuelle contient au moins un point appartenant à un objet d'intérêt. Dans le chapitre

précédent, nous avons proposé plusieurs approches qui limitent l'apparition d'objets fantômes. La première solution proposée à la section 5.2 n'ajoute pas d'objets fantômes lorsque l'on étend l'espace d'acquisition. Mais elle ne supprime pas les objets fantômes générés par l'approche classique *Shape-From-Silhouette*. La seconde approche appelée *Cohérence projective* impose, dans le cas de plusieurs objets, la connaissance des correspondances entres les composantes connexes de silhouettes. Elle nécessite la prise en compte d'information supplémentaire de couleur et peut être mise en échec lorsque les objets ont une apparence très proche. De plus elle nécessite un calibrage photométrique des caméras.

Nous proposons deux contributions, qui permettent de supprimer la totalité des objets fantômes. La première est construite sur une propriété qui permet de déterminer les composantes connexes de l'enveloppe visuelle qui contiennent au moins un point appartenant à un objet d'intérêt. Ce critère permet de supprimer toutes les entités fantômes. Cependant certaines configurations pathologiques peuvent mettre cette première approche en échec. Nous proposons une seconde méthode, qui garantit de reconstruire une forme 3D sans objet fantôme, silhouette-équivalente. Dans la suite nous supposons que tout point 3D de $\mathcal{O}$ est observé par au moins m caméras, ainsi $\mathcal{O} \subseteq \mathrm{EVE}_m(\mathcal{O})$.

6.1 Enveloppe d'objets réels *EOR*

Nous souhaitons supprimer les objets fantômes présents dans les estimations $3D$ générées par EVE_m, qui est une généralisation de EVE ou de *Shape-From-Silhouette* (voir la Propriété 7). Nous proposons une caractérisation des objets fantômes, construite sur le fait qu'une composante connexe de EVE_m qui est la seule à se projeter sur un point de l'une des silhouettes $\tilde{\mathcal{S}}_i$ contient au moins un point des objets d'intérêt.

Rappelons la définition d'un objet fantôme :

Définition 14 *Soit* CC_j *l'une des composantes connexes de* EVE_m. *Si et seulement si* $\forall \mathrm{X} \in \mathrm{CC}_j : \mathrm{X} \notin \mathcal{O}$ *alors* CC_j *est un objet fantôme de* EVE_m. □

Propriété 10 *Soit* $\mathrm{p} \in \tilde{\mathcal{S}}_i$, *l'un des pixels de la silhouette observable* $\tilde{\mathcal{S}}_i$.

$$\exists!\ \mathrm{CC}_j \subseteq \mathrm{EVE}_m \text{ telle que } \mathrm{p} \in \mathrm{Proj}_{\pi_i}(\mathrm{CC}_j) \Rightarrow \mathrm{CC}_j \notin \mathrm{O}_{fantôme}$$

□

Preuve En accord avec la construction des silhouettes, $\forall \mathrm{p} \in \tilde{\mathcal{S}}_i$, il existe au moins un point $\mathrm{Q} \in \mathcal{O}$ tel que $\mathrm{p} = \mathrm{Proj}_{\pi_i}(\mathrm{Q})$.

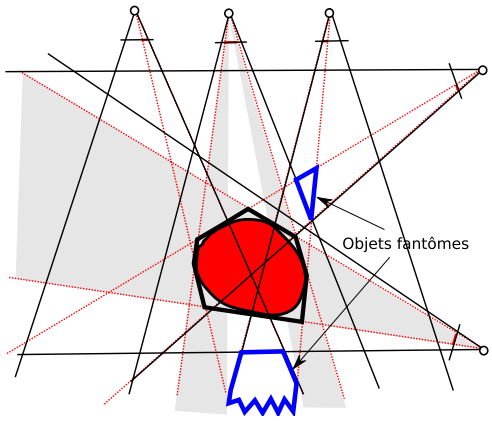


FIG. 6.3 – Représentation 2D des résultats de EOR : les objets réels sont représentés en rouge. Les composantes connexes noires sont conservées car elles satisfont la propriété 10. Les parties en gris indiquent les portions de l'espace où les composantes connexes de EVE_m sont conservées. Les objets fantômes (en bleu) ne sont pas conservés.

Si il existe une unique composante connexe $\text{CC}_j \subseteq \text{EVE}_m$ telle que $\text{p} \in \text{Proj}_{\pi_i}(\text{CC}_j)$, alors il existe au moins un point $\text{Q} \in \mathcal{O}$ tel que :

$$\text{Q} \in \text{CC}_j \text{ et } \text{Proj}_{\pi_i}(\text{Q}) = \text{p}$$

Parce que CC_j est la seule à se projeter sur le point p alors CC_j contient au moins un point $\text{Q} \in \mathcal{O}$ et n'est donc pas un objet fantôme. □

La propriété 10 caractérise les composantes connexes de EVE_m qui ne sont pas des objets fantômes. Nous introduisons la notion d'*enveloppe d'objets réels* notée EOR, l'union de toutes les composantes connexes qui satisfont la propriété 10 :

Définition 15

$$EOR = \bigcup \text{CC}_j, \exists \text{p} \in \tilde{\mathcal{S}}_i \ et \ \text{p} \in \text{Proj}_{\pi_i}(\text{CC}_j), \text{p} \notin \text{Proj}_{\pi_i}\text{CC}_k$$

avec CC_j *et* $\text{CC}_k \subset EV$ *et* $\text{CC}_k \neq \text{CC}_j$. □

Propriété 11 *EOR ne contient pas d'objet fantôme.* □

Preuve En effet EOR est l'union des composantes connexes de EVE_m qui contiennent au moins un point de $\mathcal{O}$. Ainsi EOR ne contient pas d'objet fantôme. □

Une méthode de calcul de EOR est présentée dans l'algorithme 3. Les figures 6.3 et 6.6 illustrent

le résultat de EOR en 2D. Les composantes connexes de EVE_2 qui sont les seules à se projeter sur l'un des pixels de l'une des silhouettes sont conservées.

$EOR \leftarrow \emptyset$;
[Calcul pour chaque point de silhouette, des composantes connexes de EVE_m *se projettent dessus.]*
Pour toute $\text{CC}_j \in \text{EVE}_m$ **faire**
  Pour $i \in [1, \cdots, n]$ **faire**
    Pour tout point $\text{p} \in \tilde{\mathcal{S}}_i$ **faire**
      Si $(\text{p} \in \text{Proj}_{\pi_i}(\text{CC}_j))$ **Alors**
        AjouterLien(p,CC_j) ;
      Fin Si
    Fin Pour
  Fin Pour
Fin Pour
[Les points de silhouette associés à une seule composante $\text{CC}_j \in \text{EVE}_m$*, indiquent que* $\text{CC}_j \in EOR$*.]*

Pour $i \in [1, \cdots, n]$ **faire**
  Pour tout point $\text{p} \in \tilde{\mathcal{S}}$ **faire**
    Si (NombreLiens(p) = 1) **Alors**
      $EOR \leftarrow EOR \cup$ DonneComposanteConnexeLiéeA(p) ;
    Fin Si
  Fin Pour
Fin Pour

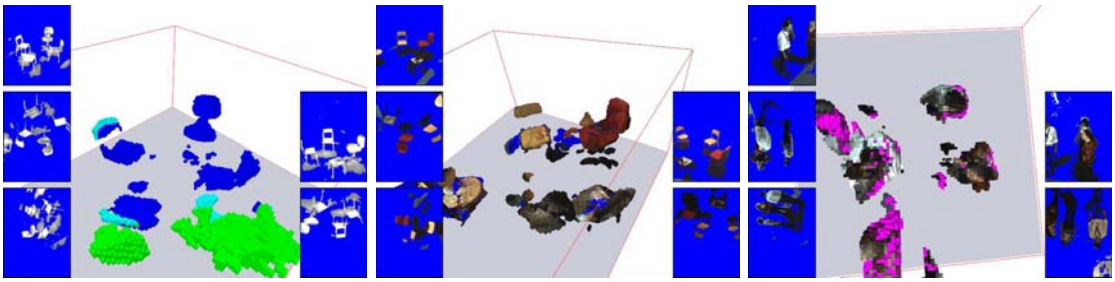
Algorithme 3: Algorithme de calcul de EOR

6.1.1 Résultats

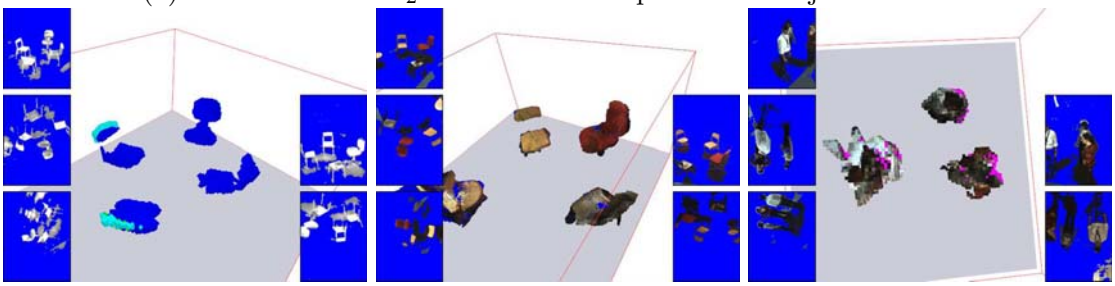
Dans cette section nous présentons des résultats expérimentaux qui démontrent l'efficacité de l'approche EOR. La plate-forme d'acquisition est composée de 5 caméras qui capturent 30 images par seconde, d'une résolution de 640×480 pixels. Chaque caméra est connectée à un ordinateur qui calcule l'extraction de silhouette utilisant l'approche proposée par [Ziv04] et l'envoi à un serveur à travers un réseau gigabit. Le calcul de EVE_m puis de EOR est réalisé sur une même machine, basée sur un processeur AMD ATHLON X2 5600+ avec une carte graphique Nvidia 7900 GTX.

L'approche EOR a été implémentée sur carte graphique programmable. Cela revient à calculer l'union des composantes connexes de EVE_m qui satisfont la propriété 10. Les composantes connexes de EVE_m qui sont les seules à se projeter sur un point de silhouette, font partie de EOR.

Il convient de calculer pour chaque point de silhouette la liste des composantes connexes dont



(a) Résultat de EVE_2. Nous notons la présence d'objet fantômes.



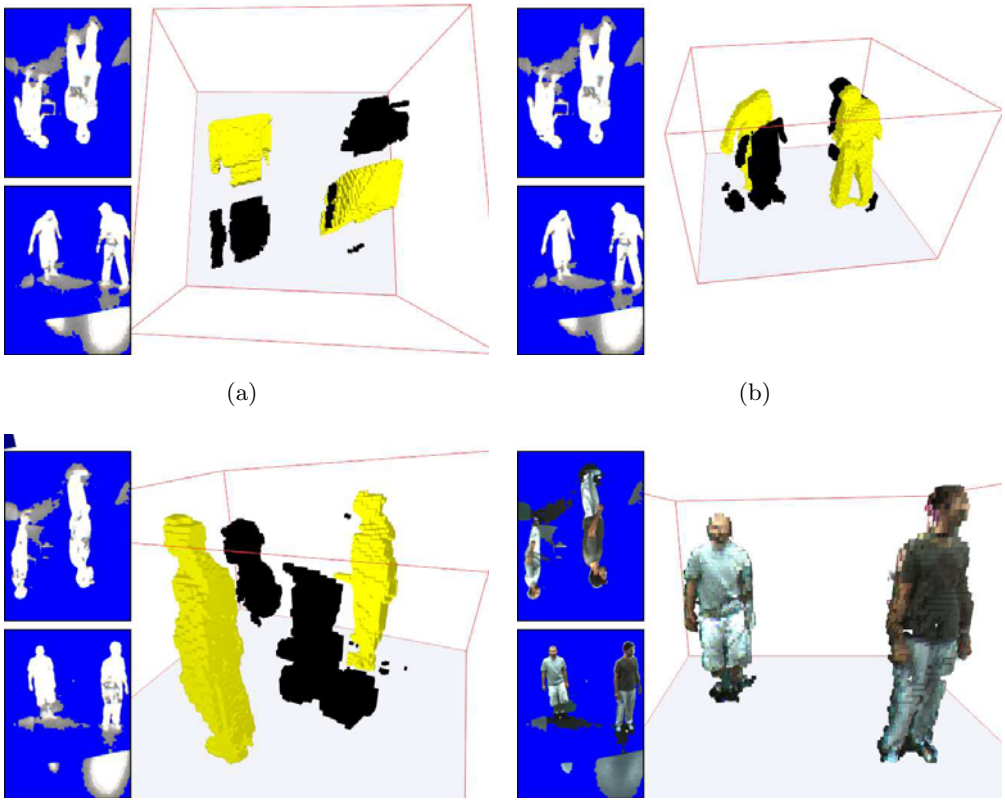
(b) Résultat de *EOR* à partir de EVE_2.

FIG. 6.4 – Résultats de *EOR* : la première ligne (a) représente l'estimation des objets d'intérêt par l'approche EVE_2. La seconde ligne (b) montre la reconstruction des objets d'intérêt par *EOR*. Chaque colonne représente une frame particulière. Dans la seconde et troisième colonne, la couleur est donnée pour une meilleur compréhension (les voxels dont la couleur ne peut être calculée sont représentés en pourpre).

la projection recouvre ce point. Nous réalisons le calcul tour à tour de la projection de chaque composante connexe de EVE_m dans toutes les caméras, ce qui revient au calcul de k rendu pour k composantes connexes. Ensuite pour chaque point de silhouette, si celui-ci est dans la projection d'une seule composante connexe, cette composante est marquée comme appartenant à *EOR*.

La figure 6.4 représente *EOR* calculée à partir de EVE_2 pour deux scènes complexes. Comme nous pouvons l'observer dans les images issues des caméras, les objets peuvent être occultés sans que cela ne mette *EOR* en échec. Nous soulignons que *EOR* ne contient pas d'objet fantômes et travaille à partir des reconstructions fournies aussi bien par l'approche EVE_m que par *Shape-From-Silhouette*. La figure 6.5 illustre l'efficacité de l'approche *EOR* pour une configuration difficile. Deux personnes sont filmées par deux caméras. Même en présence de bruit dans les silhouettes, *EOR* est capable de supprimer exactement tous les objets fantômes générés par EVE_2.

Notre implémentation de *EOR* fonctionne en temps réel, avec plus de 60 corrections par seconde. Elle s'avère moins rapide que celle proposée pour EVE_m. Cela provient du calcul des associations entre pixels de silhouette, et des composantes connexes. Pour une scène filmée par 5 caméras, le processus complet, c'est à dire de calcul de EVE_m puis de *EOR*, est effectué plus



(a) (b)

(c) (d)

FIG. 6.5 – Résultats de EOR : estimation de la forme 3D de deux personnes filmées par deux caméras. (a) et (b) représentent la même reconstruction de EOR en jaune depuis deux points de vue différents. Les objets fantômes sont correctement détectés (en noir). (c) et (d) soulignent l'efficacité de EOR pour la même scène. (d) montre la reconstruction EOR colorée.

de 40 fois par seconde. Les temps de calculs dépendent directement du nombre de caméras n.

6.1.2 Discussion

Avec notre approche, nous garantissons que tout objet réel contenu dans une des composantes connexes de EVE_m appartient à EOR tant que cette composante n'est pas **totalement occultée** dans toutes les caméras par une autre composante connexe de EVE_m.

Toutes les portions de l'espace qui ne sont pas dans EVE_m, ne peuvent contenir d'objet réel. A l'opposé, nous venons de proposer une méthode qui garantit que toutes les portions de l'espace qui appartiennent EOR contiennent au moins un point des objets d'intérêt. Cependant il n'est pas garanti que tout objet réel soit dans EOR. La figure 6.2 représente l'un des cas pathologiques dans lesquels EOR est vide. Cette limitation vient du fait qu'il existe un nombre infini de configurations d'objets qui peuvent produire les mêmes silhouettes, et ainsi la même enveloppe visuelle (voir la publication [Lau94] pour une étude plus précise).

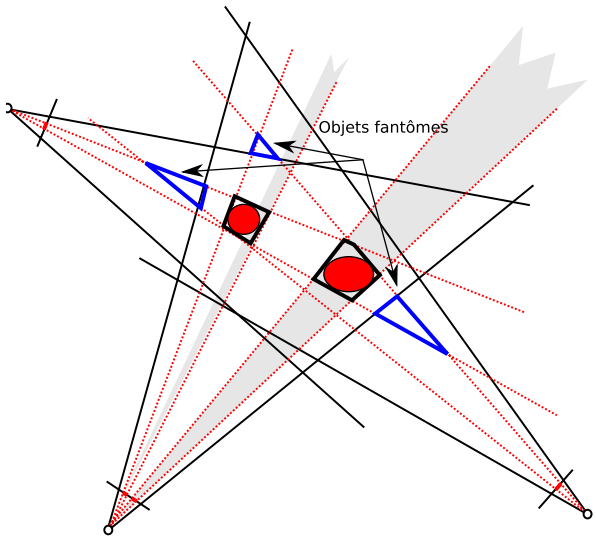


FIG. 6.6 – *EOR* est formée des composantes connexes représentées en noir ; elle conserve les composantes connexes de l'enveloppe visuelle qui contiennent des objets réel. Avec cette configuration, l'approche d'enveloppes sûres de Miller et Hilton conserve les parties de l'enveloppe visuelle, intersectées par les parties grises. Leur approche supprime alors certaines parties de l'enveloppe visuelle qui contiennent des objets réels.

L'*enveloppe d'objets réels EOR* est une approche formelle qui supprime tous les objets fantômes de l'enveloppe visuelle étendue. L'approche de Miller et Hilton [MH07] semble similaire. Notre approche est basée sur la notion de composantes connexes 3D, alors que leur travail est basé sur des composantes connexes 1D.

EOR ne contient pas certains objets réels, seulement lorsque ces objets sont **totalement occultés** par d'autres objets réels, et ce dans toutes les vues. Supposer qu'il existe au moins un pixel dans l'une des silhouettes qui correspond à une seule composante connexe n'est pas une supposition forte. La méthode de Miller et Hitlon supprime de l'enveloppe des parties d'objets réels, lorsqu'ils sont **patiellement occultés** par d'autres objets réels dans toutes les caméras. La supposition qu'aucun objet réel n'est partiellement occulté par un autre objet dans toutes les caméras est très contraignante. Notre formulation est moins restrictive que l'approche de Miller et Hilton. La figure 6.6 illustre une configuration pour laquelle notre approche conserve tous les objets réels, alors que l'approche de [MH07] supprime une partie des objets réels. Pour plus de détails, nous présentons dans le chapitre 8 une étude qualitative de nos contributions, dont *EOR*, comparées aux extensions de *Shape-From-Silhouette* qui font état de l'art, comme par exemple l'approche de Miller et Hilton.

Enfin nous soulignons que le calcul de *EOR* ne nécessite pas d'information supplémentaire en dehors des masques binaires de silhouettes et du calibrage géométrique des caméras.

6.2 Enveloppe d'objets réels silhouette-équivalente

Nous avons proposé l'approche *Enveloppe d'objets réels* qui supprime la totalité des objets fantômes générés par EVE_m ou encore par *Shape-From-Silhouette*. Cependant la forme reconstruite n'est pas nécessairement silhouette-équivalente. C'est à dire qu'il est possible que certains objets réels ne soient pas contenus dans EOR, ce qui risque de mettre en échec l'étape d'acquisition et de suivi de mouvements. Dans cette section nous proposons une approche qui fournit une forme silhouette-équivalente avec les objets d'intérêt et qui ne contient pas d'objet fantôme.

L'approche que nous proposons est construite sur l'idée clé qu'il existe au moins une combinaison de nombre minimal de composantes connexes de EV silhouette-équivalente avec les objets d'intérêt. Nous montrons dans la suite que si cette combinaison "minimale" est unique, alors celle-ci contient l'ensemble des points 3D des objets d'intérêt et ne contient pas d'objet fantôme. Nous nous intéressons en premier lieu à la suppression des objets fantômes de l'EV. Nous étendrons cette approche à EVE_m dans la section 6.2.5.

Rappelons la définition de la silhouette-équivalence de deux ensembles de points. Deux ensembles de points E et F sont silhouette-equivalent si et seulement si :

$$\forall i \in [1, \cdots, n], \text{Proj}_{\pi_i}(E) = \text{Proj}_{\pi_i}(F) \tag{6.1}$$

Nous observons une relation liant le nombre de composantes connexes de chaque silhouette, au nombre de composantes connexes de tout ensemble de points $3D$ silhouette-équivalent avec $\mathcal{O}$:

Proposition 1 *Soit U un sous-ensemble de $\mathbb{R}^3$ silhouette-équivalent avec $\mathcal{O}$, alors*

$$\forall i \in [1, \cdots, n], Card_{cc}(\mathcal{S}_i) \leq Card_{cc}(U) \tag{6.2}$$

où $Card_{cc}(U)$ le nombre de composantes connexes de l'ensemble U.

Preuve Comme U est silhouette-équivalent avec $\mathcal{O}$ alors :

$$\forall i \in [1, \cdots, n], \mathcal{S}_i = \text{Proj}_{\pi_i}(U) \tag{6.3}$$

soit k le nombre de composantes connexes de U identifiées cc_j, alors

$$\forall i \in [1, \cdots, n], \mathcal{S}_i = \text{Proj}_{\pi_i}(\bigcup_k cc_j \subseteq U) \tag{6.4}$$

$$\forall i \in [1, \cdots, n], \mathcal{S}_i = \bigcup_k \text{Proj}_{\pi_i}(cc_j \subseteq U) \tag{6.5}$$

L'image d'une composante connexe par une fonction continue est une composante connexe. Comme Proj_{π_i} est une fonction continue, $\text{Proj}_{\pi_i}(cc_j)$ est une composante connexe. L'union de k

composantes connexes contient au plus le même nombre de composantes connexes que k. Nous en déduisons :

$$\forall i \in [1, \cdots, n], Card_{cc}(\mathcal{S}_i) \leq Card_{cc}(U) \tag{6.6}$$

□

Notre objectif est de supprimer les objets fantômes créés par $EV(\mathcal{O})$. Les composantes connexes de $EV(\mathcal{O})$ peuvent être classées en deux ensembles : l'ensemble Tcc, union des composantes connexes de $EV(\mathcal{O})$ qui contiennent au moins un point de $\mathcal{O}$; $\mathrm{O}_{fantôme}$, union des objets fantômes de $EV(\mathcal{O})$ vide de tout point de $\mathcal{O}$.

$$Tcc = \bigcup cc_j \subseteq EV(\mathcal{O}), \exists \mathrm{P} \in cc_j, \mathrm{P} \in \mathcal{O} \tag{6.7}$$

$$Tcc = EV(\mathcal{O}) - \mathrm{O}_{fantôme} \tag{6.8}$$

Hypothèse : Dans la suite nous supposons que

$$\exists i \in [1, \cdots, n], Card_{cc}(\mathcal{S}_i) = Card_{cc}(Tcc) \tag{6.9}$$

Intuitivement, nous supposons qu'il existe au moins une vue qui distingue tous les objets visuellement discernables. Un objet visuellement discernable est l'union des objets réels, dont la projection tombe dans une même composante connexe de silhouette, pour chaque vue. Ces objets s'occultent mutuellement (partiellement ou totalement) pour chaque point de vue. En d'autres termes, nous supposons que nous avons k objets visuellement discernables, et que la silhouette qui contient le nombre maximum de composantes en contient k.

La figure 6.2 montre plusieurs scènes réelles qui offrent la même enveloppe visuelle en 2D. Les configurations qui valident cette hypothèse sont présentées en vert. Nous discuterons de la validité de cette hypothèse dans la section 6.2.4. Celle-ci représente une supposition plus réaliste et facile à valider que celles présentées dans les précédentes approches de la littérature visant à supprimer les objets fantômes. Nous montrons dans les prochaines sections, que même lorsque cette hypothèse n'est pas respectée, cette approche produit, au pire, la même reconstruction que *EOR*.

Notre but est d'estimer Tcc, l'union des composantes connexes de $EV(\mathcal{O})$ qui contiennent au moins un point de $\mathcal{O}$. A cet effet nous proposons le concept d'union minimale d'une enveloppe visuelle, qui offre une méthode pour calculer Tcc.

Définition 16 *Soit U une union de composantes connexes de $EV(\mathcal{O})$. U est une union minimale de $EV(\mathcal{O})$ si et seulement si, $\forall U'$ une union de composantes connexes de $EV(\mathcal{O})$ telle que $Card_{cc}(U') < Card_{cc}(U)$:*

$$\begin{cases} U \text{ est silhouette-équivalent avec } \mathcal{O} \\ U' \text{ n'est pas silhouette-équivalent avec } \mathcal{O} \end{cases} \tag{6.10}$$

A ce point, nous introduisons la proposition la plus importante, reliant Tcc au concept d'union minimale.

Proposition 2 *Si l'hypothèse 6.9 est validée, alors Tcc est une union minimale de $EV(\mathcal{O})$.*

Preuve Nous commençons par prouver que dans ce cas Tcc est silhouette-équivalent avec $\mathcal{O}$, ensuite nous montrons que Tcc est minimal.

1. Par construction, Tcc est un sous-ensemble de $EV(\mathcal{O})$ qui contient tout point de $\mathcal{O}$. Ainsi $\mathcal{O} \subseteq Tcc \subseteq EV(\mathcal{O})$. Dans [Lau94] Laurentini a prouvé que $EV(\mathcal{O})$ est le volume maximal silhouette-équivalent avec $\mathcal{O}$ (voir chapitre 4 propriété 2 pour la démonstration). Tcc contient $\mathcal{O}$ et est inclus dans $EV(\mathcal{O})$, alors Tcc est silhouette-équivalent avec $\mathcal{O}$.
2. Soit U' une union de composantes connexes de $EV(\mathcal{O})$ avec $Card_{cc}(U') < Card_{cc}(Tcc)$. L'hypothèse 6.9 donne

$$\exists i \in [1, \cdots, n], Card_{cc}(\mathcal{S}_i) = Card_{cc}(Tcc) \tag{6.11}$$

alors

$$\exists i \in [1, \cdots, n], Card_{cc}(\mathcal{S}_i) > Card_{cc}(U') \tag{6.12}$$

Nous déduisons de la proposition 1 que U' n'est pas silhouette-équivalent avec $\mathcal{O}$.

Finalement, $\exists i \in [1, \cdots, n], Card_{cc}(\mathcal{S}_i) = Card_{cc}(Tcc) \Rightarrow Tcc$ est une union minimale de $EV(\mathcal{O})$. □

Cette proposition offre une méthode qui permet de calculer Tcc. Il y a plusieurs cas possibles : soit il existe une unique union minimale de $EV(\mathcal{O})$, soit il en existe plusieurs.

6.2.1 Union minimale de $EV(\mathcal{O})$ unique

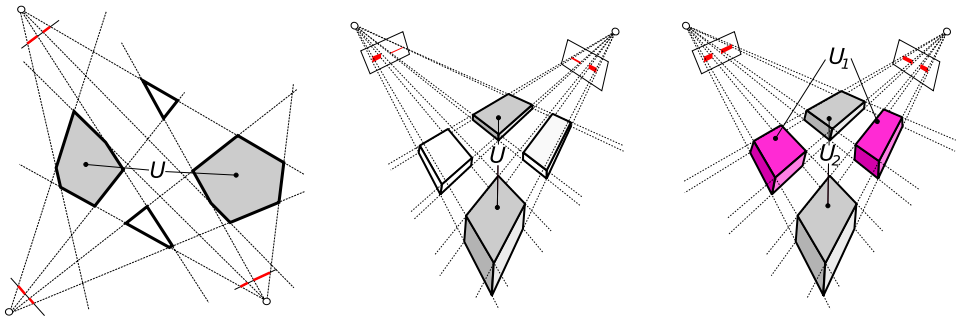
Comme Tcc existe toujours, alors :

Proposition 3 *Si il existe une union minimale unique de $EV(\mathcal{O})$, alors $U = Tcc$.*

Cette proposition offre une méthode robuste pour calculer Tcc en présence d'une unique union minimale de $EV(\mathcal{O})$ (voir Fig 6.7(a)). Avec une union minimale unique U, l'enveloppe visuelle de $\mathcal{O}$ privée d'objets fantômes, appelée Enveloppe d'Objets Réels Silhouette-Équivalente (EOR_{seq}) est définie par

$$EOR_{seq}(\mathcal{O}) = U \tag{6.13}$$

La proposition 3 est intéressante lorsqu'il existe une unique union minimale U de $EV(\mathcal{O})$ (voir Fig. 6.7(a) et Fig. 6.7(b)). Il est cependant possible que plusieurs unions-minimales existent. Ce cas se révèle rare dans des configurations réelles. Soient $U_1, \cdots, U_l$ l'ensemble des unions minimales de $EV(\mathcal{O})$. La Figure 6.7(c) montre un exemple 3D de plusieurs unions minimales.



(a) (b) (c)

FIG. 6.7 – Une enveloppe visuelle 2D (a), et deux EV 3D (b,c) et les unions minimales correspondantes. (a) et (b) présentent des exemples pour lesquels l'union minimale U (en gris) de $EV(\mathcal{O})$ (dont les bords sont en noir) est unique. (c) présente l'existence de deux unions minimales U_1 (en gris clair) et U_2 (en violet foncé) de $EV(\mathcal{O})$ (dont les bords sont en noir). La différence entre les configurations (b) et (c) provient du fait que les composantes connexes des silhouettes de (b) n'ont pas une taille similaire, en comparaison avec les silhouettes de (c).

6.2.2 Unions minimales de $EV(\mathcal{O})$ multiples

La plupart des précédentes approches de suppression d'objets fantômes, proposent des critères permettant d'identifier quelles parties de l'EV contiennent des points de $\mathcal{O}$. Avec plusieurs unions minimales, un critère analogue peut être : les composantes connexes générées par l'intersection de toutes les unions minimales, ne peuvent contenir d'objet fantômes. Ce critère définit une condition nécessaire mais non suffisante. Il est en effet possible de manquer certaines parties de $\mathcal{O}$. Cette effet de bord peut avoir encore plus d'effets négatifs sur la reconstruction que la présence même d'objets fantômes, spécialement pour des applications liées à l'analyse de scène, au suivi d'objets, etc.

Pour cette raison, nous préférons une solution qui impose que $\mathcal{O} \subseteq EOR_{seq}(\mathcal{O})$, au risque de conserver quelques objets fantômes. Sous l'hypothèse 6.9, Tcc est une union minimale de $EV(\mathcal{O})$ (voir Proposition 2). Ainsi, il existe une union parmi les unions minimales $U_1, \cdots, U_l$, telle que $U_j = Tcc$. En utilisant seulement l'information de silhouette et les données de calibrage géométrique des caméras, il n'est pas possible de déterminer U_j analytiquement. Nous proposons, dans le pire cas, de calculer une borne supérieure de Tcc qui s'avère cependant plus précise que $EV(\mathcal{O})$:

$$EOR_{seq}(\mathcal{O}) = \bigcup_{l} U_i \tag{6.14}$$

Proposition 4 *Si l'hypothèse 6.9 est valide, alors*

$$\mathcal{O} \subseteq EOR_{seq}(\mathcal{O}). \tag{6.15}$$

Preuve Par construction $\mathcal{O} \subseteq Tcc$. Comme l'hypothèse 6.9 est validée, alors Tcc est une union

minimale de $EV(\mathcal{O})$. Nous déduisons des équations 6.13 et 6.14 que $Tcc \subseteq EOR_{seq}(\mathcal{O})$. Ainsi $\mathcal{O} \subseteq EOR_{seq}(\mathcal{O})$. □

Ainsi, en présence d'une unique ou de plusieurs unions minimales de $EV(\mathcal{O})$, $\mathcal{O}$ est inclus dans $EOR_{seq}(\mathcal{O})$. Dans cette section, nous avons proposé le nouveau concept formel d'union minimale de $EV(\mathcal{O})$. Ce principe permet d'estimer exactement Tcc, les composantes connexes non fantômes de $EV(\mathcal{O})$, lorsqu'il existe une unique union minimale. Dans le cas contraire, la solution adoptée propose une reconstruction contenant $\mathcal{O}$, avec moins d'objets fantômes que ceux présents dans $EV(\mathcal{O})$. Dans la suite, nous exposons une solution efficace afin de calculer toutes les unions minimales de $EV(\mathcal{O})$.

6.2.3 Calcul de toutes les unions minimales

Le concept d'union minimale est construit sur le test de silhouette-équivalence d'unions de composantes connexes de $EV(\mathcal{O})$. Nous commençons par associer chaque pixel de chaque image d'entrée, aux composantes connexes qui se projettent en ce pixel. Pour une union de composantes connexes, nous calculons la différence entre les pixels de silhouettes, et ceux associés à cette union. Si la différence tombe en dessous d'un seuil, l'union est validées comme étant silhouette-équivalent avec $\mathcal{O}$.

En présence de ce test de silhouette-équivalence efficace, il faut ensuite générer toutes les unions des composantes connexes de $EV(\mathcal{O})$. Soit k le nombre de ses composantes. Il existe 2^k combinaisons de composantes connexes cc_j. Une solution peu efficace serait de toutes les énumérer dans l'ordre croissant du nombre de composantes utilisées, puis de tester la silhouette-équivalence de cette union. Dans le pire cas (en terme de calcul) $EV(\mathcal{O})$ ne contient pas d'objets fantômes. Alors l'algorithme aurait à réaliser 2^k tests de silhouette-équivalence.

Afin de réduire ces temps de calcul prohibitifs, nous observons que toutes les unions minimales de $EV(\mathcal{O})$ partagent une propriété intéressante : elles contiennent certaines composantes connexes qui sont absolument nécessaires afin de valider le critère de silhouette-équivalence. Ces composantes connexes sont aussi appelée enveloppes d'objets réels EOR, présenté chapitre 6.1. EOR est l'union des composantes connexes de $EV(\mathcal{O})$ qui vérifient la propriété 10, *i.e.* qui ne sont pas des objets fantômes.

Proposition 5 *EOR est un sous-ensemble de toute union minimale U de $EV(\mathcal{O})$:*

$$\forall U \textit{ une union minimale de } EV(\mathcal{O}), EOR \subseteq U \tag{6.16}$$

Preuve Par construction $\forall cc_j \subseteq EOR, EV(\mathcal{O}) - cc_j$ n'est pas silhouette-équivalent avec $\mathcal{O}$. Ceci implique que $\forall cc_j \subseteq EOR, cc_j \subset U$. □

En utilisant l'association réalisée entre chaque pixel et les composantes connexes cc_j de $EV(\mathcal{O})$, EOR est calculé par l'union des cc_j pour lesquelles un pixel s est uniquement associé à cc_j. Calculer EOR est équivalent en temps, à vérifier la silhouette-équivalence d'une union des composantes connexes de $EV(\mathcal{O})$. Soit $l \leq k$ le nombre de composantes connexes de $EV(\mathcal{O})$ validées par EOR. Dans le pire cas, nous devons vérifier la silhouette-équivalence de 2^{k-l} unions. Comme observé dans nos expérimentations, les valeurs usuelles de $k-l$ sont 3 ou 4, soit approximativement 16 unions. L'algorithme suivant décrit cette approche du calcul de EOR_{seq} :

```
l,q,k : entier ;
k ← nombre de composantes connexes de EV(O) ;
l ← nombre de composantes connexes de EOR ;
EOR_seq ← ∅ ;
Pour q de l à k faire
    Pour toute U, union de q composantes connexes de EV avec EOR ⊆ U faire
        Si (∀i ∈ [1, ··· , n], Proj_πi(U) = S_i) Alors
            k ← q ;
            EOR_seq ← EOR_seq ∪ U ;
        Fin Si
    Fin Pour
Fin Pour
```

Algorithme 4: Algorithme de calcul de EOR_{seq}.

En utilisant EOR, nous réduisons le nombre de tests de silhouette-équivalence par un facteur de 2^l. Cette approche propose une évaluation très efficace de l'ensemble de toutes les unions minimales. Ainsi $EOR_{seq}(\mathcal{O})$ est calculé en temps réel (voir la section résultats 6.2.6).

6.2.4 Discussion de l'hypothèse

Notre approche est construite sur l'hypothèse selon laquelle, ayant k objets visuellement discernables, l'une des silhouettes contient k composantes (voir hypothèse 6.9). Comparée avec les suppositions implicites proposées dans les approches précédentes telles que EOR, "Safe Hulls" (SH) [MH07] ou encore "Reduced Visual Hull" (REV) [BG08], notre approche s'avère moins contraignante.

Notre hypothèse s'avère valide même pour des scènes contenant des objets fantômes qui occultent totalement les composantes de Tcc (voir Fig.6.7(c)). EOR s'avère incapable de reconstruire les composantes de Tcc qui sont totalement occultées par d'autres composantes pour chaque vue. Dans le cas présenté figure 6.7(c), EOR offre une reconstruction vide, alors que EOR_{seq} contient tout $\mathcal{O}$. SH construit un point de Tcc lorsqu'il existe un rayon partant du centre d'une caméras vers ce points, qui traverse $EV(\mathcal{O})$ en une seule composante 1D. Ainsi

$SH(\mathcal{O}) \subseteq EOR(\mathcal{O})$. L'hypothèse implicite est même encore plus contraignante que celle de EOR. REV, tout comme la méthode $pcSFS$ présentée chapitre 5.4, est construit sur un processus de mise en correspondance des composantes connexes de silhouette entre chaque vue. Ceci impose une hypothèse encore plus forte, que chaque point de $\mathcal{O}$ n'est pas, même partiellement, occulté dans au moins deux vues.

Au final, notre hypothèse s'avère moins contraignante que celles proposées dans les précédentes approches de suppression de fantômes.

6.2.5 Généralisation à $\mathrm{EVE}_m(\mathcal{O})$

La formalisation de EOR_{seq} a été réalisée pour des reconstructions issues de l'enveloppe visuelle usuelle. Cependant nous avons proposé précédemment la définition des enveloppes visuelles étendues $\mathrm{EVE}_m(\mathcal{O})$ sur lesquelles nous souhaitons généraliser l'utilisation de EOR_{seq}. $\mathrm{EVE}_m(\mathcal{O})$ est construite sur une différentiation des silhouettes "théoriques" $\mathcal{S}_i$ et des silhouettes observables $\tilde{\mathcal{S}}_i$ dont la construction prend en compte le domaine de visibilité imposé par l'utilisation de caméras réelles :

$$\tilde{\mathcal{S}}_i = \mathcal{S}_i \cap \mathcal{I}_i \tag{6.17}$$

Nous observons que la proposition 1, liant le nombre de composantes connexes de chaque silhouette au nombre de composantes connexes d'un ensemble de point 3D silhouette-équivalent, n'est pas valide pour les silhouettes $\tilde{\mathcal{S}}_i$. Il est possible que la projection d'un ensemble connexe de points U présentant beaucoup de concavités, génère plusieurs composantes connexes dans $\tilde{\mathcal{S}}_i$. En effet l'intersection de deux connexes E et $\mathcal{I}_i$ ne décrit pas nécessairement un connexe. Afin de contourner ce problème, nous proposons de restreindre la proposition 1 aux composantes connexes de $\tilde{\mathcal{S}}_i$ qui ne partagent pas de frontières de $\mathcal{I}_i$. En effet, l'image d'une composante connexe cc_j de E fournit plus d'une composante connexe dans $\tilde{\mathcal{S}}_i$, seulement lorsque

$$\exists P \in \mathrm{Proj}_{\pi_i}(cc_j), P \notin \mathcal{I}_i.$$

Avec l'utilisation des silhouettes observées $\tilde{\mathcal{S}}_i$, la proposition 1 est validée, si l'on prend en compte uniquement les composantes connexes $\tilde{\mathcal{S}}_i$ qui ne partagent pas de frontières avec $\mathcal{I}_i$. Lors de l'utilisation de l'approche $\mathrm{EVE}_m(\mathcal{O})$, l'hypothèse formulée équation 6.9 nécessite l'existence d'au moins une caméra pour laquelle $\tilde{\mathcal{S}}_i = \mathcal{S}_i \cap \mathcal{I}_i = \mathcal{S}_i$. Avec cette adaptation de la proposition 1, nous venons d'étendre l'utilisation de EOR_{seq} à des reconstructions $3D$ provenant de $\mathrm{EVE}_m(\mathcal{O})$.

6.2.6 Résultats

EOR_{seq} a été utilisée dans plusieurs expérimentations particulièrement difficiles. La reconstruction à partir de données réelles permet de souligner la robustesse de notre approche même en présence des contraintes du monde réel. Des résultats à partir de données synthétiques sont

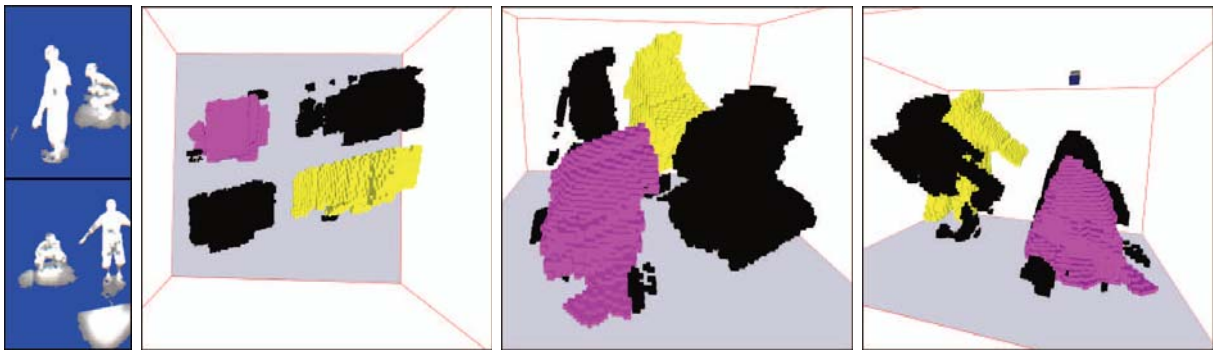


FIG. 6.8 – Une scène difficile reconstruite à partir de deux vues. EOR valide uniquement la partie jaune alors que l'union-minimale permet de valider en plus la partie violette. EOR_{seq} (en jaune et violet) contient tous les objets d'intérêt, sans les objets fantômes (en noir).

présentés afin de démontrer la qualité de reconstruction offerte par EOR_{seq} face aux approches précédentes de suppression d'objets fantômes purement géométriques.

L'estimation de l'enveloppe visuelle étendue $\text{EVE}_2(\mathcal{O})$ et l'étape de suppression d'objets fantômes sont calculés sur un seul ordinateur (Amd ATHLON X2 5600+) aidé par une carte graphique NVIDIA 9800GT. Notre implémentation utilise principalement le GPU pour calculer l'enveloppe visuelle étendue. Les composantes connexes de $\text{EVE}_2(\mathcal{O})$ sont extraites en utilisant le CPU. Enfin l'association entre pixels de silhouette et composantes connexes, sont estimées en utilisant le GPU. Avec une grille de voxels d'une résolution de $128 \times 128 \times 128$, le processus complet estime plus de 30 reconstructions par seconde.

L'espace d'acquisition est observé par 2 caméras capturant jusqu'à 30 images par seconde d'une résolution de 640×480 pixels. Chaque caméra est connectée à un ordinateur qui calcule l'extraction de silhouette, et l'envoie au serveur de reconstruction. Les expérimentations à partir de données réelles sont construites uniquement à partir de 2 vues dans le but de générer un nombre important d'objets fantômes.

Dans notre première expérimentation (voir figure 6.8) nous montrons l'efficacité de notre approche pour une scène particulièrement difficile observée par deux caméras. Les approches $\text{EVE}_2(\mathcal{O})$ et *Shape-From-Silhouette* construisent deux objets fantômes détectés par l'approche EOR_{seq} (en noir). L'approche précédente EOR valide partiellement les parties non objets fantôme (en jaune). L'approche EOR_{seq} se distingue par sa capacité à détecter exactement les objets fantômes (en noir) grâce au concept d'union-minimale.

La figure 6.9 présente une comparaison des reconstructions offertes par les approches EOR_{seq}, EVE_2, Safe Hulls [MH07] (SH), et EOR. L'ensemble de ces approches utilise uniquement l'information de calibrage des caméras et des masques de silhouette. L'expérimentation 1 est construite sur une scène composée de 3 personnes. Dans l'une des silhouettes, le nombre de composantes connexes est le même que le nombre d'objets visuellement discernables. Ainsi cette configuration suit notre hypothèse 6.9. $\text{EVE}_2(\mathcal{O})$ construit 6 objets fantômes, alors que SH et EOR

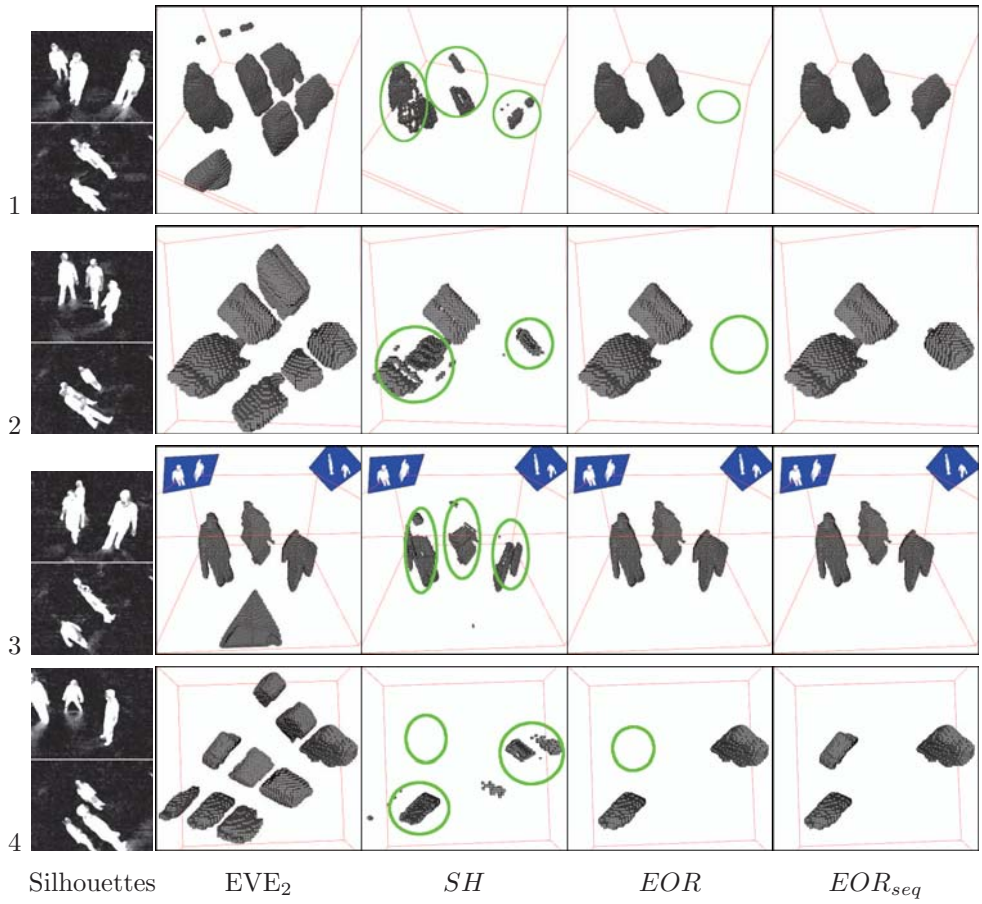


FIG. 6.9 – Comparaison visuelle des reconstructions offertes pour différents jeux de données réelles. Les silhouettes sont présentées colonne de gauche. Ces expérimentations génèrent beaucoup d'objets fantômes qui sont parfaitement supprimés par notre approche. L'utilisation de seulement deux vues en entrée, permet de souligner l'efficacité de EOR_{seq}. Les cercles verts indiquent les parties manquantes.

génèrent des reconstructions incomplètes. EOR_{seq} est capable de déterminer exactement toutes les composantes non fantômes. L'expérimentation 2 montre une configuration proche, pour laquelle EVE$_2(\mathcal{O})$ crée 3 objets fantômes. A nouveau, EOR_{seq} est la seule approche capable de supprimer exactement les objets fantômes. L'expérimentation 3 ne valide pas notre hypothèse de travail, car les deux silhouettes contiennent seulement deux composantes. Dans ce cas EOR_{seq} et EOR produisent la reconstruction attendue.

Lors de la dernière expérimentation à partir de données réelles, EVE$_2(\mathcal{O})$ génère 6 objets fantômes. EOR produit une forme moins corrompue que celle proposée par SH, mais cette approche n'est pas capable de reconstruite la troisième personne. Grace au concept d'union minimale, EOR_{seq} n'a aucun problème à détecter tous les objets fantômes.

Une évaluation visuelle offre une analyse intéressante des résultats. Cependant celle-ci ne

permet pas d'évaluer la qualité de la reconstruction offerte par EOR_{seq}. Dans la suite, nous proposons des expérimentations construites à partir de scènes synthétiques, dont les données de vérité terrain, permettent une évaluation quantitative et qualitative.

L'espace d'acquisition virtuel est composé de 8 caméras disposées en cercle, capturant des masques binaires de silhouette d'une résolution de 720×480 pixels. Nous avons choisi une expérimentation à partir de données synthétiques afin de comparer qualitativement les résultats apportés par EOR_{seq} aux approches géométriques précédentes qui ne nécessitent aucun information additionnelle en dehors des paramètres de calibrage des caméras et des masques binaires de silhouette : EVE_2, Safe Hulls [MH07], les Enveloppes d'Objets Réels EOR (voir section 6.1) et les Enveloppe d'Objets Réels Silhouette-équivalente EOR_{seq}.

Afin d'évaluer la qualité des différentes reconstructions, nous avons choisi trois critères de qualité : α correspondant au ratio de points des objets d'intérêt compris dans la reconstruction ; β étant le nombre d'objets fantômes construits par l'approche évaluée ; γ représentant la proportion de points faux-positifs estimés par la méthode à évaluer, par rapport à l'ensemble des points des objets d'intérêt. Les objets fantômes étant des composantes connexes de la reconstruction ne contenant aucun point des objets d'intérêt, β et γ sont liés. L'évaluation de la qualité dépend avant tout du contexte applicatif fixé. Notre but étant l'acquisition du mouvement de personnes, les critères précédents peuvent être ordonnés de la façon suivante : α est le critère le plus important car réaliser le suivi de parties manquantes se révèle particulièrement difficile ; β présente le nombre d'objets reconstruits inutilement ; γ souligne la précision de la reconstruction. Les jeux de données utilisés pour cette expérimentation et les reconstructions associées sont représentés figure 6.10. Les valeurs quantitatives correspondantes sont présentées tableau 6.1.

Nous observons que pour les différentes expérimentations, EOR_{seq} reconstruit un volume englobant les objets d'intérêt ($\alpha \approx 1.0$), et ne contenant pas ou peu d'objets fantômes. Ceux-ci apparaissent lorsqu'il existe plusieurs unions-minimales de $\text{EVE}_m(\mathcal{O})$. EOR_{seq} supprime le nombre maximum d'objets fantômes en respectant la propriété de silhouette-équivalence. Cette approche se révèle particulièrement robuste en présence de peu de caméras.

EOR supprime toujours tous les objets fantômes, mais supprime parfois certains points réels. EOR produit des reconstructions incomplètes, comparée à EOR_{seq}. Dans toutes ces configurations SH supprime parfois tous les objets fantômes et fournit souvent une reconstruction incomplète, spécialement en présence d'un faible nombre de caméras.

Dans [MH07], SH a été construit sur une approche *Shape-From-Silhouette* surfacique et s'avère peu robuste à l'échantillonnage d'une approche *Shape-From-Silhouette* volumique : dans le tableau 6.1 avec le jeu D, la reconstruction de SH contient 4 petits objets fantômes pour 3 ou 4 vues. Dans la mesure où EOR_{seq} est construite sur la propriété de silhouette-équivalence, EOR_{seq} est efficace aussi bien sur les approches *Shape-From-Silhouette* surfaciques que volumiques.

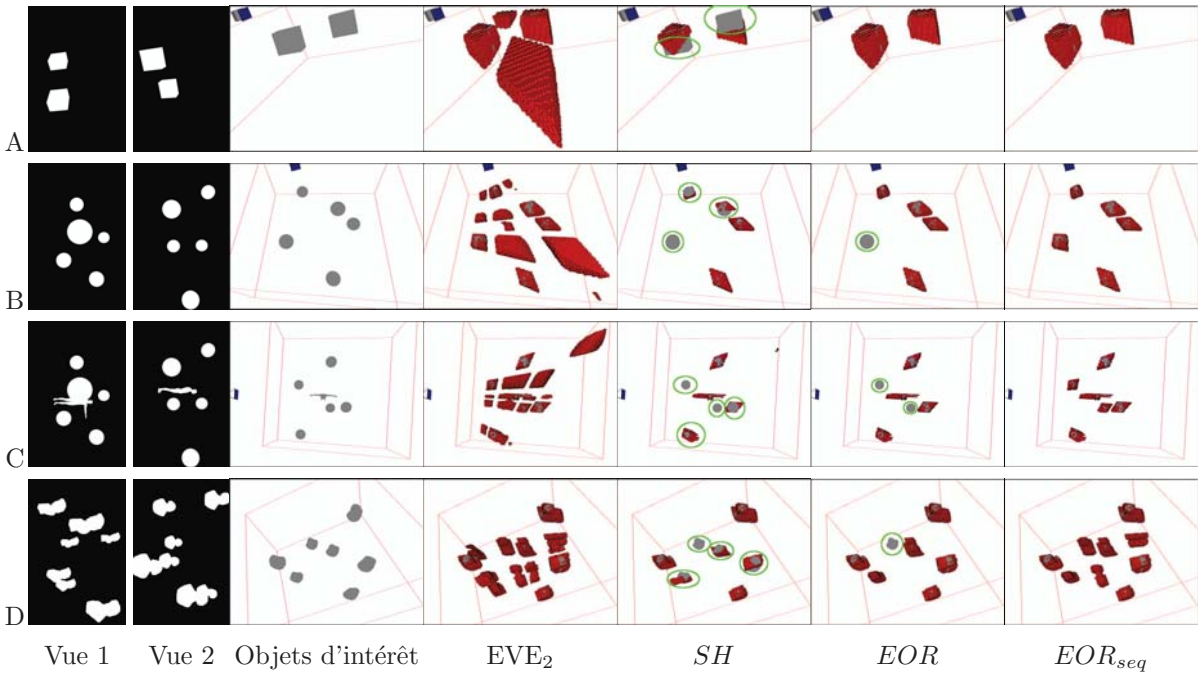


FIG. 6.10 – Reconstructions issues de différents jeux de données à partir de deux vues. Ces différentes configurations sont difficiles, ce qui a pour conséquence la création d'objets fantômes. Les deux colonnes de gauches présentent les masques de silhouette correspondants. Les configurations des objets d'intérêt observées depuis point de vue virtuel, sont présentées à la troisième colonne. Les colonnes suivantes décrivent successivement les reconstructions proposées par : EVE_2 volumique, Safe Hulls [MH07] (SH), EOR puis EOR_{seq}. Les cercles verts montrent les parties des objets d'intérêt qui ne sont pas reconstruites.

6.2.7 Discussion

Dans cette section nous avons proposé une approche totalement automatique qui supprime en temps réel les objets fantômes construits par toute méthode reposant sur le concept d'enveloppe visuelle. Cette nouvelle approche ne nécessite pas d'information supplémentaire en dehors du calibrage des caméras et de l'information de silhouette. Celle-ci se révèle efficace, que ce soit pour des objets statiques ou dynamiques, estimés par une approche surfacique ou volumique. Nous avons prouvé qu'en respectant une hypothèse peu contraignante, notre approche est capable de supprimer la totalité des objets fantômes dans la majorité des cas. Une comparaison avec les précédentes approches déterministes, montre que EOR_{seq} fournit des reconstructions plus précises même en présence d'un grand nombre d'objets d'intérêt filmés par peu de caméras. Dans la mesure où EOR_{seq} ne nécessite pas d'information additionnelle, cette approche permet d'améliorer les applications actuelles construites sur l'estimation d'enveloppes visuelles.

	n	EVE_2			SH			EOR			EOR_{seq}		
		α	β	γ	α	β	γ	α	β	γ	α	β	γ
A	2	0.99	2	3.60	0.38	0	0.52	0.99	0	1.28	0.99	0	1.28
	3	0.99	0	0.80	0.99	0	0.80	0.99	0	0.80	0.99	0	0.80
	4	0.99	0	0.74	0.99	0	0.74	0.99	0	0.74	0.99	0	0.74
B	2	0.99	7	2.43	0.51	0	0.25	0.80	0	0.35	0.99	0	0.42
	3	0.99	1	0.31	0.99	0	0.26	0.99	0	0.26	0.99	0	0.26
	4	0.99	0	0.22	0.99	0	0.22	0.99	0	0.22	0.99	0	0.22
C	2	0.99	11	3.80	0.31	0	0.27	0.63	0	0.57	0.99	0	0.71
	3	0.99	7	0.56	0.94	0	0.30	0.99	0	0.31	0.99	0	0.31
	4	0.99	1	0.27	0.99	0	0.25	0.99	0	0.25	0.99	0	0.25
D	2	0.99	5	8.58	0.57	0	5.07	0.67	0	5.50	0.99	2	7.40
	3	0.99	5	4.20	0.69	4	2.90	0.99	0	3.82	0.99	0	3.82
	4	0.99	1	2.09	0.98	4	2.20	0.99	0	2.05	0.99	0	2.05

TAB. 6.1 – Comparaison des reconstructions issues des jeux de données synthétiques présentés Fig. 6.10. Les méthodes qui offrent les meilleurs résultats sont sur-lignées en vert alors que celles qui offrent les moins bons résultats sont sur-lignées en rouge clair. Pour toutes ces expérimentations, EOR_{seq} offre toujours les meilleurs résultats, comparés aux reconstructions offertes par EVE_2, SH, et EOR. n décrit le nombre de vues en entrée. Seules les reconstructions issues d'un faible nombre de caméras sont représentées afin de souligner les capacités de EOR_{seq}.

6.3 Conclusion

Dans ce chapitre nous avons proposé deux approches, EOR et EOR_{seq}, qui permettent de caractériser et supprimer les objets fantômes dans les reconstructions fournies par les approches EVE_m ou encore par *Shape-From-Silhouette*. EOR conserve tout objet d'intérêt lorsqu'il participe à la création des silhouettes, c'est à dire que sans cet objet, les silhouettes ne seraient pas celles observées. Cette supposition n'est pas très forte et suffit en général pour supprimer uniquement les objets fantômes. Dans le cas contraire l'approche EOR_{seq} garantit que la forme reconstruite est silhouette-équivalente. Lorsqu'il existe une unique *union-minimale*, EOR_{seq} contient tous les objets d'intérêt et ne contient aucun objet fantôme.

Les approches EOR et EOR_{seq} sont efficaces pour toute estimation générée par EVE_m. Cependant EVE_m et *Shape-From-Silhouette* sont basées sur l'exploitation des masques binaires de silhouette. Or les défauts de robustesse de l'extraction des masques de silhouettes sont directement responsables d'erreurs dans la reconstruction (généralement des trous). Dans le chapitre suivant nous proposons une méthode temps-réel qui répond à ce problème en retardant le plus possible la classification des pixels en tant que silhouette. L'idée est de profiter des paramètres de calibrage géométrique des caméras, pour fusionner l'information disponible pour chaque caméra et ainsi prendre une décision globale à toutes les caméras pour un point $3D$ donné.

Chapitre 7

Estimation de l'enveloppe visuelle par fusion de cartes de silhouettes

Les approches *Shape-From-Silhouette* estiment la forme 3D d'objets d'intérêt à partir de plusieurs images de silhouettes de ses objets, issues de plusieurs caméras calibrées. Les méthodes actuelles de calibrage offrent une souplesse de mise en œuvre et une précision suffisante pour les approches *Shape-From-Silhouette* [Zha00]. L'extraction de silhouettes en temps réel par les approches de kroma-keying est généralement de bonne qualité. Mais dans un contexte d'environnement peu contrôlé, il n'existe pas, à notre connaissance, de méthode d'extraction de silhouette temps réel suffisamment robuste.

Dans l'approche *Shape-From-Silhouette* classique, si l'un des pixels de silhouette est classé "fond" par erreur, alors l'ensemble des points 3D qui se projettent sur ce pixel sont classés comme n'appartenant pas aux objets d'intérêt. Ceci conduit à une reconstruction incomplète ou corrompue des objets d'intérêt.

Ce type d'erreur provient du fait que l'étiquetage monoculaire est difficile à réaliser de façon fiable dans un environnement général et peu contrôlé. Diverses perturbations expliquent cette difficulté : le bruit des capteurs CCD, les ambiguïtés de couleur entre le fond et les objets d'intérêt ou encore les changements d'illumination de la scène (ce qui inclut les ombres d'objets d'intérêt).

Dans [FB05], Franco et Boyer proposent une solution qui repose sur le concept de grille d'occupation, méthode populaire dans la communauté de la robotique. L'idée centrale est de considérer chaque pixel comme un capteur apportant de l'information sur la scène observée et de le modéliser statistiquement comme tel. Les observations rapportées par les pixels de toutes les caméras sont utilisées conjointement pour déduire où la matière se trouve dans la scène et avec quelle probabilité. Avec ce modèle, la plupart des sources d'incertitude peuvent être prises en compte explicitement et aucune décision prématurée sur l'état des pixels et leur appartenance à une silhouette n'est nécessaire. L'idée novatrice de leur approche est que pour chaque point

3D, la décision d'appartenance aux objet d'intérêt est prise en fusionnant l'information observée par toutes les caméras. Le choix d'une modélisation statistique offre une solution élégante à ce problème. Mais la résolution du système par inférence bayesienne interdit l'utilisation de cette approche dans un contexte de temps réel.

Notre contribution est construite sur la même idée clé. C'est à dire que la décision d'appartenance d'un point 3D à l'enveloppe visuelle étendue des objets d'intérêt, est prise en fusionnant l'information issue de toutes les caméras. L'information de silhouette issue d'une caméra peut-être remise en cause par l'information issue des autres caméras qui voient ce point 3D. Comparée à l'approche de Franco et Boyer, notre approche est algorithmiquement très simple et offre des résultats similaires, avec un facteur d'accélération d'environ 1000.

7.1 Les cartes probabilistes de silhouettes

Parmi les méthodes d'extractions de silhouette en temps réel, les approches les plus efficaces sont généralement construites sur une modélisation statistique du fond (voir la section 4.1). Cet apprentissage est effectué lorsque les objets d'intérêt ne sont pas encore dans la scène. Ensuite La probabilité pour que chaque pixel appartienne aux valeurs observées de fond est évaluée pour chaque nouvelle image.

Dans la suite nous nommons carte de probabilité de silhouette, l'image qui associe à chaque pixel y de l'image $\mathcal{I}_i$, sa probabilité $\tilde{P}_i(\mathrm{y})$ **estimée** de ne pas appartenir au fond.

La figure 7.1 présente la carte de probabilité de silhouette, déduite d'une modélisation statistique du fond par mélange de gaussiennes [Ziv04].

L'information de silhouette observable est déterminée à partir des cartes de silhouettes de la façon suivante :

$$\tilde{\mathcal{S}}_i = \{\mathrm{y} \in \mathcal{I}_i, \tilde{P}_i(\mathrm{y}) > T\} \tag{7.1}$$

avec T un seuil de confiance et $\tilde{P}_i(\mathrm{y})$ la probabilité **estimée** qu'un point 2D y de l'image $\mathcal{I}_i$ n'appartienne pas au fond, définie par

$$\tilde{P}_i(\mathrm{y}) = P_i(\mathrm{y}) + N \tag{7.2}$$

où $P_i(\mathrm{y})$ désigne la probabilité que le point y de l'image $\mathcal{I}_i$ n'appartienne pas au fond et N désigne un bruit additif.

La classification monoculaire de l'information de silhouette proposée dans l'équation 7.1 fournit des résultats peu robustes aux bruits du capteur de la caméra et aux ambiguïtés de couleur entre le fond et les objets d'intérêt. Cela provient du fait que le paramètre de bruit N n'est pas toujours suffisamment pris en compte.

(a) (b) (c) (d)

FIG. 7.1 – L'image (a) représente l'image d'une modélisation statistique de fond, observé par une caméra fixe. L'image (b) représente la trame courante observée par cette même caméra. L'image (c) représente la carte de probabilité correspondante : la valeur de chaque pixel représente sa probabilité de ne pas appartenir au fond (noir = 0, blanc = 1). L'image (d) indique l'estimation correspondante de $\tilde{\mathcal{S}}_i$ avec $T = 0,5$.

Pour rendre les approches *Shape-From-Silhouette* plus robustes face aux problèmes d'extraction de silhouettes, nous proposons de décider de l'appartenance d'un point 3D à l'enveloppe visuelle étendue des objets d'intérêt, en se basant cette fois sur les cartes probabilistes de silhouettes, plutôt que sur l'information peu robuste des silhouettes $\tilde{\mathcal{S}}_i$.

7.2 Enveloppe visuelle étendue à partir de cartes probabilistes de silhouettes

Rappelons la définition de EVE_m qui estime l'enveloppe visuelle étendue des objets d'intérêt observés par au moins m caméras :

Définition 17

$$\text{EVE}_m = \{\text{X} : \tilde{\mathcal{S}}(\text{X}) = \mathcal{I}(\text{X}) \ et \ Card(\mathcal{I}(\text{X})) \geq m\}$$

□

Supposons pour l'instant que le bruit N dans les cartes probabilistes de silhouettes est nul.

On en déduit ainsi

$$\forall \text{X} \in \text{EVE}_m, \forall i \in \mathcal{I}(\text{X}), P_i(\text{X}) > T$$

Soit $MoyP(\mathrm{X})$ la contribution moyenne empirique de la probabilité que la projection de X appartienne au fond pour l'ensemble des caméras :

$$MoyP(\mathrm{X}) = \frac{1}{Card(\mathcal{I}(\mathrm{X}))} \sum_{i \in \mathcal{I}(\mathrm{X})} P_i(\mathrm{Proj}_{\pi i}(\mathrm{Q}))$$

Ainsi on en déduit que :

$$\forall \mathrm{X} \in \mathrm{EVE}_m, MoyP(\mathrm{X}) > T$$

Supposons désormais que le bruit N suit la loi normale $\mathcal{N}(0;\, \sigma^2)$. Nous définissons de façon similaire $Moy\tilde{P}(\mathrm{X})$, la contribution moyenne **empirique** de la probabilité estimée que la projection de X appartienne au fond pour l'ensemble des caméras.

N suit une loi normale d'espérence 0, ainsi :

$$\forall \mathrm{X} \in \mathrm{EVE}_m, \lim_{Card(\mathcal{I}(\mathrm{X})) \to \infty} Moy\tilde{P}(\mathrm{X}) = MoyP(\mathrm{X})$$

et

$$\forall \mathrm{X} \in \mathrm{EVE}_m, \lim_{Card(\mathcal{I}(\mathrm{X})) \to \infty} Moy\tilde{P}(\mathrm{X}) > T \tag{7.3}$$

Dans la mesure où la variable aléatoire N suit la loi normale $\mathcal{N}(0;\, \sigma^2)$, on déduit la probabilité suivante :

$$Proba(N > -3\sigma) = Proba(N \leq 3\sigma) \approx 0,999$$

Ainsi

$$Proba(\tilde{P}_i(\mathrm{y}) > P_i(\mathrm{y}) - 3\sigma) \approx 0,999$$

Or $\forall \mathrm{X} \in \mathrm{EVE}_m, \forall i \in \mathcal{I}(\mathrm{X}), P_i(\mathrm{X}) > T$

On en déduit que

$$\forall \mathrm{X} \in \mathrm{EVE}_m, \forall i \in \mathcal{I}(\mathrm{X}), Proba(\tilde{P}_i(\mathrm{X}) > T - 3\sigma) > 0,999 \tag{7.4}$$

En présence d'un grand nombre de caméras, il provient des équations 7.3 et 7.4, que pour tout point $\mathrm{X} \in \mathrm{EVE}_m$, le bruit dans les cartes de silhouette est compensé et $Moy\tilde{P}(\mathrm{X}) > T$. Une très grande majorité des points $\mathrm{X} \in \mathrm{EVE}_m$ satisfont à la propriété $\tilde{P}_i(\mathrm{X}) > T - 3\sigma$. Ainsi nous venons de proposer plusieurs critères qui permettent d'estimer EVE_m à partir des cartes de probabilité dont le bruit est supposé additif et suivant une loi normale d'espérance nulle.

Définition 18 *Soit* $\widetilde{\mathrm{EVE}}_m$ *l'estimation de* EVE_m *à partir des cartes de probabilité de silhouette :*

$$\widetilde{\mathrm{EVE}}_m = \{\mathrm{X} : \tilde{\mathcal{S}}_E(\mathrm{X}) = \mathcal{I}(\mathrm{X}), Card(\mathcal{I}(\mathrm{X})) \geq m \text{ et } Moy\tilde{P}(\mathrm{X}) > T\}$$

avec $\tilde{\mathcal{S}}_E(\mathrm{X})$ *l'ensemble estimé des indices de silhouettes, dans lesquelles* X *possède une image.*

$$\tilde{\mathcal{S}}_E(\mathrm{X}) = \{i \in \mathcal{I}(\mathrm{X}) : \tilde{P}_i(\mathrm{X}) > T - 3\sigma\}$$

□

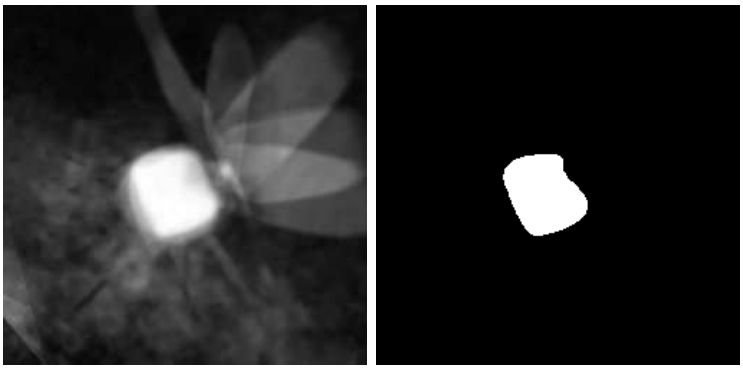


FIG. 7.2 – L'image de gauche représente en 2D les valeurs de $Moy\tilde{P}(\mathrm{X})$. Le noir correspond à une valeur de 0 et le blanc correspond à une valeur de 1. L'image de droite représente le résultat de $\widetilde{\mathrm{EVE}}_m$ avec $T = 0,7$ et $\sigma = 0,13$.

L'approche que nous venons de proposer permet pour chaque point 3D, de compenser l'information de probabilité erronée supposée minoritaire, par la contribution majoritairement observée. De plus le fait d'imposer un seuil minimum permet de prévenir les erreurs issues de valeurs extrêmes.

Afin de profiter du schéma de confrontation permettant de décider de l'appartenance d'un point 3D X à $\widetilde{\mathrm{EVE}}_m$, il est préférable qu'il y ait au moins deux informations à confronter ; c'est à dire que $m \geq 2$. Le paramètre T indique une valeur de confiance, définissant la précision moyenne de l'estimation de EVE_m. Le paramètre σ régit la dispersion maximale autorisée. Il participe à la définition du seuil bas, en dessous duquel la probabilité qu'un point 2D n'appartienne pas au fond est jugée trop faible. La figure 7.2 montre une représentation 2D des valeurs de $Moy\tilde{P}(\mathrm{X})$ (image de gauche) ainsi que la reconstruction $\widetilde{\mathrm{EVE}}_m$ correspondante. L'algorithme 5 présente une méthode pour estimer $\widetilde{\mathrm{EVE}}_m$ par un ensemble de voxels.

Il n'est à priori pas facile de déterminer le couple de paramètres (T, σ) qui offre la meilleure qualité de reconstruction. Nous reviendrons sur cette difficulté chapitre 8, dans lequel nous réalisons une analyse qualitative de cette approche à partir de données simulées.

Outre le fait que notre approche estime EVE_m et *Shape-From-Silhouette* (qui en est un cas particulier) de façon robuste en présence d'extraction de silhouettes bruitées, elle permet de calculer les silhouettes binaires corrigées des objets d'intérêt. En effet, le fait de confronter en 3D l'information des cartes de probabilité apporte une mécanisme de confrontation plus robuste que les approches monoculaires classiques. Soit $\tilde{\mathcal{S}}_{i_C}$ la silhouette binaire corrigée de la caméra $\mathcal{C}_i$:

$$\tilde{\mathcal{S}}_{i_C} = \{\mathrm{y} \in \mathcal{I}_i, \mathrm{y} \in \mathrm{Proj}_{\pi i}(\widetilde{\mathrm{EVE}}_m)\}. \tag{7.5}$$

Diviser l'espace d'intérêt en $G \times G \times G$ voxels v_k avec $k \in [1, \cdots, G^3]$;
Pour k **de** 1 **à** G^3 **faire**
  $v_k \leftarrow Dehors$;
  $\tilde{\mathcal{S}}_{Conf}(v_k) \leftarrow \emptyset$;
  $\mathcal{I}(v_k) \leftarrow \emptyset$;
  $Moy\tilde{P}(v_k) \leftarrow 0$;
  Pour i **de** 1 **à** n **faire**
    Si ($(\text{Proj}_{\pi_i}(v_k)) \cap \mathcal{I}_i \neq \emptyset$) **Alors**
      $\mathcal{I}(v_k) \leftarrow \mathcal{I}(v_k) \cup i$;
      $Moy\tilde{P}(v_k) \leftarrow Moy\tilde{P}(v_k) + \tilde{P}_i(\text{Proj}_{\pi_i}(v_k))$;
      Si ($\tilde{P}_i(\text{Proj}_{\pi_i}(v_k)) > T - 3\sigma$) **Alors**
        $\tilde{\mathcal{S}}_{Conf}(v_k) \leftarrow \tilde{\mathcal{S}}_{Conf}(v_k) \cup i$;
      Fin Si
    Fin Si
  Fin Pour
  $Moy\tilde{P}(v_k) \leftarrow Moy\tilde{P}(v_k)/Card(\mathcal{I}(v_k))$;
  Si ($\tilde{\mathcal{S}}_{Conf}(v_k) = \mathcal{I}(v_k)$ et $Card(\mathcal{I}(v_k)) \geq m$ et $Moy\tilde{P}(v_k) > T$) **Alors**
    $v_k \leftarrow Dedans$;
  Fin Si
Fin Pour
$\widetilde{\text{EVE}}_m$ est approximée par l'union des voxels v_k étiquetés *Dedans* ;

Algorithme 5: Algorithme d'estimation de $\widetilde{\text{EVE}}_m$ à base de voxels.

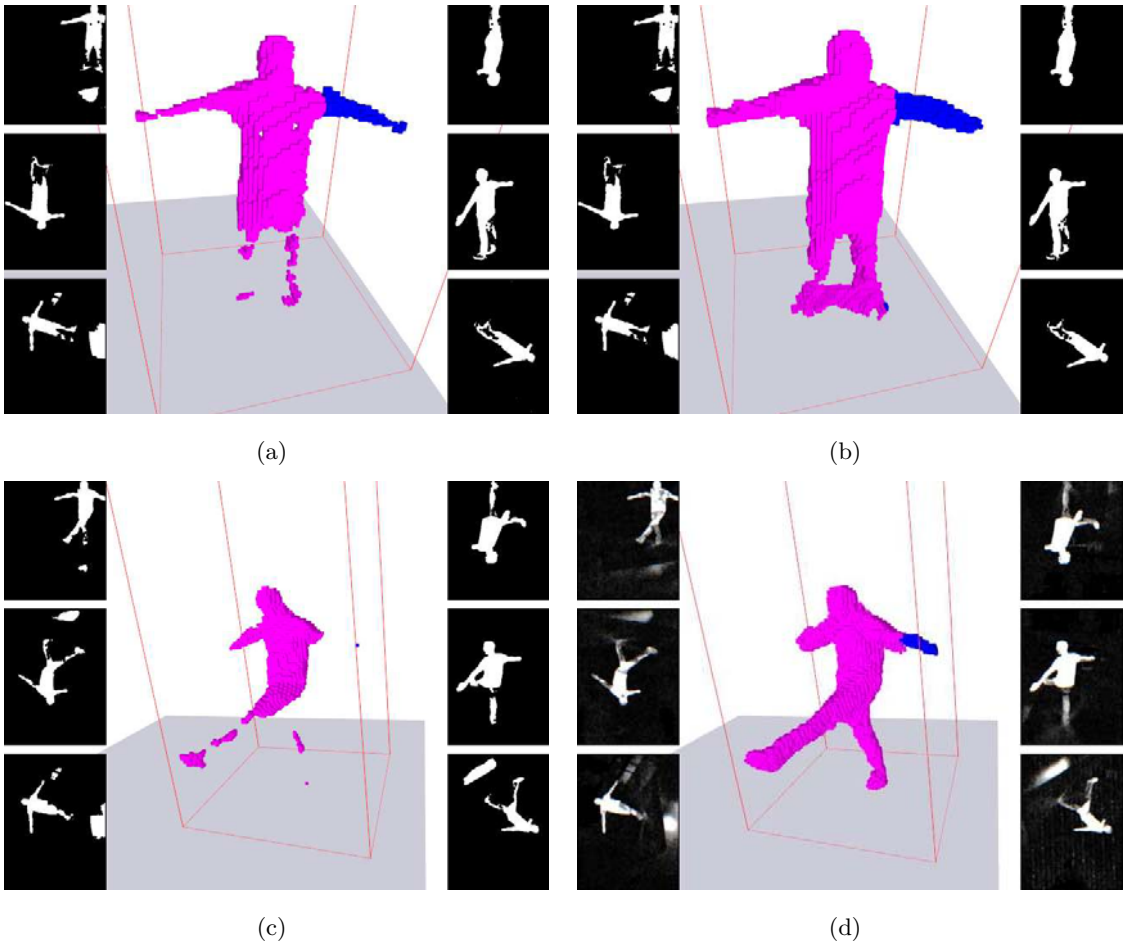
Dans la suite nous présentons les résultats de l'approche $\widetilde{\text{EVE}}_m$ à partir de données réelles, ainsi que la capacité de correction de silhouette proposée par cette approche.

7.3 Résultats

Nous allons illustrer à travers plusieurs expérimentations le fait que $\widetilde{\text{EVE}}_m$ est plus robuste aux extractions de silhouettes erronées que l'approche EVE_m.

Ces expérimentations ont été réalisées pour une scène filmée par six caméras qui fournissent 30 images par seconde d'une résolution de 640×480 pixels. Le calcul de $\widetilde{\text{EVE}}_m$ est réalisé sur un seul ordinateur avec un processeur AMD 5600+ X2 et une carte graphique NVIDIA GTX7900. L'espace 3D de $2m \times 2m \times 2m$ est estimé par une grille de voxels de résolution $128 \times 128 \times 128$.

La figure 7.3 confronte les résultats issus de l'approche $\widetilde{\text{EVE}}_m$ par rapport à l'approche EVE_m. La configuration présentée génère beaucoup de faux-négatifs dans les images binaires de silhouettes. Ainsi EVE_m n'est pas capable d'estimer correctement l'enveloppe visuelle étendue de l'objet d'intérêt (voir figures 7.3(a) et 7.3(c)).



(a) (b)

(c) (d)

FIG. 7.3 – Reconstructions $\widetilde{\text{EVE}}_4$ comparées aux reconstructions proposées par EVE_4. Les images (a) et (c) représentent le résultat de EVE_4 à partir des silhouettes observées $\tilde{\mathcal{S}}_i$, binarisées avec $T = 0,5$. Le bruit présent dans les silhouettes $\tilde{\mathcal{S}}_i$ implique que le reconstruction EVE_4 soit incomplète. Les images 7.3(b) et 7.3(d) représentent les reconstruction des mêmes scènes par l'approche $\widetilde{\text{EVE}}_4$, avec $T = 0,5$ et $\sigma = 0,13$. Cette reconstruction est robuste à l'information de silhouette bruitée.

Dans la figure 7.3(b), $\widetilde{\text{EVE}}_m$ reconstruit une grande partie de l'objet d'intérêt, mais la quantité de bruit est trop importante pour que la forme 3D complète soit estimée. Cependant dans la figure 7.3(d), $\widetilde{\text{EVE}}_m$ estime la totalité de l'objet d'intérêt.

Comme nous l'avons souligné précédemment l'approche $\widetilde{\text{EVE}}_m$ peut être utilisée pour corriger les cartes binaires de silhouettes. La figure 7.4 présente le résultat de ce processus de correction, pour la configuration présentée figures 7.3(c) et 7.3(d). Les silhouettes estimées par l'étiquetage monoculaire sont représentées sur la composante rouge de l'image. La correction est elle représentée sur la composante bleue. Nous observons que $\widetilde{\text{EVE}}_m$ permet de corriger le plupart des points détectés faux-négatifs ou faux-positifs par l'approche monoculaire. Les grandes zones rouges (au sol) ne sont pas cohérentes depuis tous les points de vue et n'appartiennent

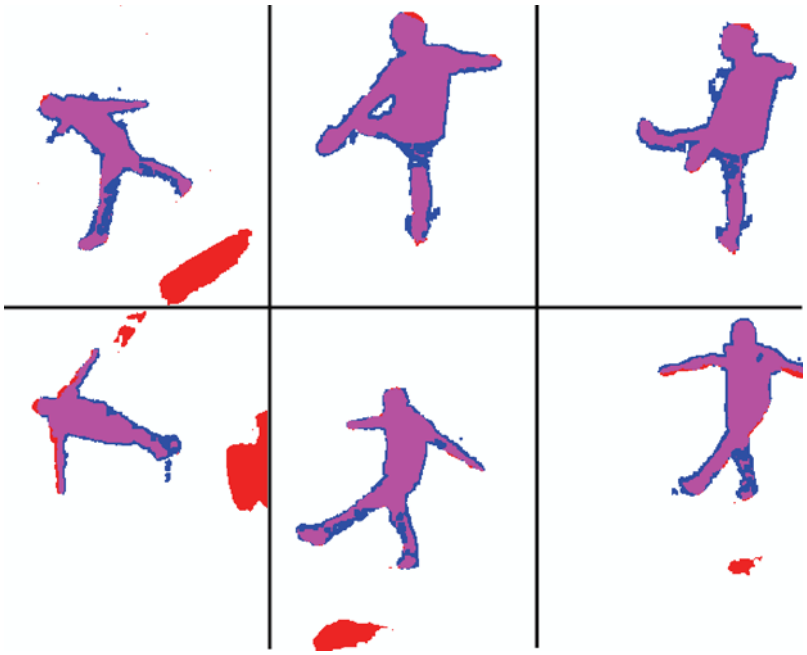


FIG. 7.4 – Illustration de la correction de silhouettes permise par l'approche $\widetilde{\text{EVE}}_m$. Les silhouettes observées proviennent de la scène présentée figure 7.3(d). La composante rouge représente les silhouettes $\tilde{\mathcal{S}}_i$ extraites de façon monoculaire. Les silhouettes $\tilde{\mathcal{S}}_{i_C}$ corrigées sont représentée sur la composante bleue de l'image. On observe la présence d'un liserai bleu autour de toutes les silhouettes. Celui-ci provient d'une résolution de la grille de voxel trop grossière. Enfin certaines zones de silhouettes "monoculaires" $\tilde{\mathcal{S}}_i$ ne sont pas recouvertes par les silhouettes corrigées $\tilde{\mathcal{S}}_{i_C}$. Les grandes zones rouges proviennent du reflet des éclairages de la scène sur le sol. N'étant pas cohérentes, ces parties ne sont pas représentées dans les silhouettes corrigées $\tilde{\mathcal{S}}_{i_C}$. Enfin certaines erreurs de calibrage sont à l'origine de petites zones des silhouettes $\tilde{\mathcal{S}}_i$ qui n'appartienne pas à $\tilde{\mathcal{S}}_{i_C}$.

donc pas aux silhouettes corrigées. Enfin l'apparition d'un liserai bleu autour des silhouettes $\tilde{\mathcal{S}}_i$ provient de l'estimation de l'EVE_m par un ensemble de voxels dont la résolution n'est pas suffisamment fine.

Comme pour la méthode EVE_m, nous avons implanté $\widetilde{\text{EVE}}_m$ sur carte graphique. L'ensemble des opérations de calcul de la contribution moyenne empirique est réalisé par un *fragment shader*. Nous obtenons des temps de calculs similaires à l'approche EVE_m. En effet la partie limitante de l'algorithme provient du transfert des voxels calculés sur la carte graphique vers la mémoire centrale de l'ordinateur. Pour une grille de dimension $2m \times 2m \times 2m$ de résolution 128^3, notre approche est capable de calculer plus de 100 reconstructions par seconde.

7.4 Discussion

Une très grande partie des méthodes qui estiment la forme 3D d'objets 3D à partir de silhouettes sont peu robustes aux silhouettes mal extraites. Dans ce chapitre noue avons proposé une approche pragmatique qui rend les méthodes EVE_m plus robustes. Notre contribution impose seulement quelques calculs supplémentaires par rapport à l'approche *Shape-From-Silhouette* ou encore sa généralisation EVE_m. Ainsi cette méthode fonctionne toujours en temps réel et permet, pour une grille de voxel de 128^3 éléments, d'estimer plus de 100 formes par seconde. Notre contribution peut-être comparée à l'approche proposée par Franco et Boyer dans [FB05]. Ces approchent tentent de répondent à une même problématique. Notre approche, propose en temps réel, des résultats semblables à ceux proposés par leur approche, nécessitant environ 10 secondes de calcul par trame. Nous ne pensons pas que le fait de porter leur algorithme sur carte graphique programmable permettra un gain de l'ordre de trois décades.

Grâce à cette contribution simple, nous rendons les méthodes *Shape-From-Silhouette* ou encore EVE_m aptes à des conditions expérimentales plus difficiles, toujours en temps réel.

Dans la section suivante nous présentons une évaluation du processus complet d'estimation 3D temps réel à partir de plusieurs vue sans contraintes de placement de caméras, robuste aux extractions erronées de silhouette et qui ne génère pas d'objets fantômes.

Chapitre 8

Résultats

Dans les précédents chapitres, nous avons proposé plusieurs approches ayant pour objectif de rendre *Shape-From-Silhouette* plus efficace dans le contexte de l'acquisition de mouvements de personnes en temps réel et sans marqueur. Pour chaque méthode proposée, nous avons principalement exposé les résultats issus de données réelles permettant de souligner "visuellement" leurs intérêts. Afin de fournir une évaluation qualitative, il est souhaitable que l'ensemble de nos contributions soient confrontées aux méthodes actuelles de *Shape-From-Silhouette*. Dans ce chapitre nous proposons de comparer ces méthodes à partir de données synthétiques, permettant de contrôler parfaitement le placement et la géométrie des objets d'intérêt, la génération de silhouettes exactes $\tilde{\mathcal{S}}_i$, tout en offrant une liberté totale du placement des caméras. Nous montrerons tout au long de nos expérimentations que les solutions apportées permettent la définition d'une méthode d'estimation temps réel de forme à partir de silhouette, sans contraintes sur le placement des caméras par rapport aux objets (et inversement), robuste aux silhouettes bruitées et qui ne crée pas d'objet fantôme.

8.1 Processus et critères d'évaluation

Notre objectif est de fournir une évaluation qualitative de nos approches, comparées aux approches actuelles de *Shape-From-Silhouette*. Les méthodes proposées réalisent une estimation volumique de la forme des objets d'intérêt par un ensemble de voxels. Ainsi les données de références dites de *vérité terrain* sont obtenues par voxélisation des objets d'intérêt. L'approximation de la forme 3D de ces objets est réalisée sur carte graphique programmable. De façon similaire à l'approche présentée dans [KPT99], nous utilisons une approche construite sur l'utilisation du buffer de profondeur (Z-buffer). La surface des objets d'intérêt virtuels, décrite par un maillage polyédrique, est supposée fermée.

L'intérieur des objets est coloré en blanc alors que le fond de la scène et l'extérieur des objets sont colorés en noir. Nous réalisons le rendu de ces objets sans changer la position d'une caméra virtuelle orthographique, mais en modifiant le plan avant de clipping. Ainsi pour chaque position

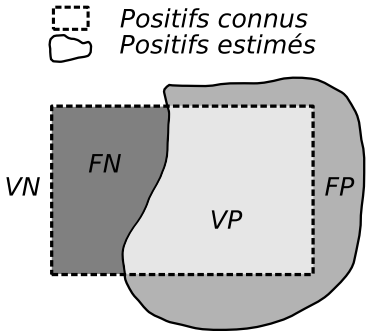


FIG. 8.1 – Représentation schématique des notions de Vrai-Positif (VP),Vrai-Négatifs (VN), Faux-Positif (FP) et Faux négatifs (FN). Ces états permettent de caractériser la labélisation de chaque élément, offerte par un classifieur binaire.

du plan de clipping de la caméra, nous calculons l'ensemble des voxels de ce plan, qui intersectent le volume intérieur des objets virtuels. Pour faciliter les comparaisons entre les données de référence et les données estimées, les différentes représentations volumiques sont calculées pour une même grille de voxels. Ainsi les comparaisons se font dans le même espace d'échantillonnage.

Les méthodes d'estimation de forme volumique d'objets d'intérêt de la famille *Shape-From-Silhouette* peuvent être considérées comme solutions à un problème de classification binaire des éléments d'une grille volumique. En adoptant ce point de vue, il devient possible d'utiliser certains des outils d'évaluation d'algorithme de classification binaire, généralement utilisés par la communauté de *machine learning* [DG06, LPD06], connaissant la vérité terrain associée à l'expérimentation.

Dans un problème de décision binaire, un classifieur décide qu'un élément testé soit labelisé négatif ou positif. Les décisions réalisées par le classifieur sont représentées par une structure souvent appelée matrice de confusion ou encore table de contingence. La matrice de confusion contient quatre informations :

- les Vrai-Positifs (VP) qui correspondent aux éléments que l'on sait positifs, évalués positifs par le classifieur,
- les Faux-Positifs (FP) qui correspondent aux éléments que l'on sait négatifs, estimés positifs,
- les Vrai-Négatifs associés aux éléments que l'on sait négatifs, évalués négatifs,
- les Faux-Négatifs, éléments que l'on sait positifs, évalués négatifs.

La figure 8.1 souligne les quatre états possibles associés à chaque élément classifié. Dans notre contexte de l'estimation de la forme volumique d'objets d'intérêt $\mathcal{O}$ connus, par une approche de type *Shape-From-Silhouette*, la matrice de confusion correspondante est présentée figure 8.2.

		Connaissance	
		$\in \mathcal{O}$	$\notin \mathcal{O}$
Estimation	$\in \mathcal{O}$	VP	FP
	$\notin \mathcal{O}$	FN	VN

FIG. 8.2 – Matrice de confusion adaptée aux contexte de l'estimation de forme.

L'information brute de la caractérisation de chaque élément n'est cependant pas suffisante pour comparer les différentes méthodes. Parmi les principaux outils d'évaluation utilisés dans la communauté de *Machine Learning*, le couple de paramètres *Précision-Rappel* (ou *Precision-Recall*) offre une caractérisation de la performance d'algorithmes de classification binaire, jugée plus discriminante que les autres outils classiques tels que la courbe *ROC* (*Receiver Operator Characteristic*) [DG06].

La précision (notée P dans la suite) est définie par le rapport du nombre d'éléments VP (c'est à dire le nombre de voxels correctement estimés comme appartenant à $\mathcal{O}$) sur le nombre total de voxels estimés comme appartenant à $\mathcal{O}$ (le nombre de voxels VP et FP) :

$$P = \frac{Card\ \{v_k \in VP\}}{Card\ \{v_k \in VP \cup FP\}}$$

La précision représente la proportion du nombre de voxels estimés comme appartenant à $\mathcal{O}$, qui appartiennent effectivement à $\mathcal{O}$. La précision peut-être considérée comme une mesure de l'exactitude ou encore de la fidélité de la reconstruction.

Le rappel (noté R dans la suite) est défini par la proportion des points de $\mathcal{O}$ contenus dans la reconstruction :

$$R = \frac{Card\ \{v_k \in VP\}}{Card\ \{v_k \in VP \cup FN\}}$$

Le rappel peut-être considéré comme une mesure de la complétude de la méthode de reconstruction à évaluer.

A travers nos contributions nous avons répondu à diverses limitations des approches *Shape-From-Silhouette*. Les plus importantes étant l'espace d'acquisition limité à l'intersection des cônes de vision des caméras, la génération d'objets fantômes et la sensibilité de la reconstruction à la qualité des silhouettes. Les conséquences sont que la totalité des objets d'intérêt ne sont pas toujours reconstruits, ou encore que certains objets reconstruits sont vides d'objet réel. Pour souligner les réponses que nos méthodes apportent nous proposons d'ajouter un critère d'évaluation F lié au nombre d'objets fantômes construits :

$$F = Card\{\mathrm{CC}_i \subset FP \cup VP, \forall v_k \in \mathrm{CC}_i, v_k \notin \mathcal{O}\}$$

Un objet fantôme désigne une composante connexe qui n'intersecte aucun objet d'intérêt. Nous rappelons que les objets fantômes sont des entités fantômes. Ainsi elles ont un impact sur la valeur de la précision.

En présence d'une méthode de reconstruction de la forme volumique "idéale", les critères d'évaluation choisis fournissent les valeurs suivantes : $R = 1,0$ indique que tout point de $\mathcal{O}$ est inclus dans la reconstruction. $P = 1,0$ souligne que la méthode ne reconstruit pas d'entités fantômes. Enfin $F = 0$ indique que cette méthode de reconstruction ne crée pas d'objet fantôme.

La "qualité" est une notion dépendante du contexte dans lequel on souhaite caractériser les méthodes. Dans notre contexte d'acquisition de mouvements temps réel de personnes à partir de plusieurs caméras, les critères peuvent être ordonnés de la façon suivante :

1. R représente le critère le plus important. Pour suivre les personnes filmée, il est indispensable qu'elles soient totalement reconstruites.
2. F indique que chaque objet 3D (composante connexe) reconstruit, contient au moins la géométrie d'une personne. Le fait que F soit proche de 0 permet de prévenir certaines erreurs pour le processus de suivi de mouvements. En effet le suivi d'objets vides peut rendre le processus d'acquisition de mouvements instable.
3. P représente l'exactitude de la reconstruction. Ainsi plus cette valeur est faible, plus la reconstruction contient des entités fantômes nuisibles à la précision du processus du suivi. C'est à dire que les données appropriées sont noyées dans des données non pertinentes.

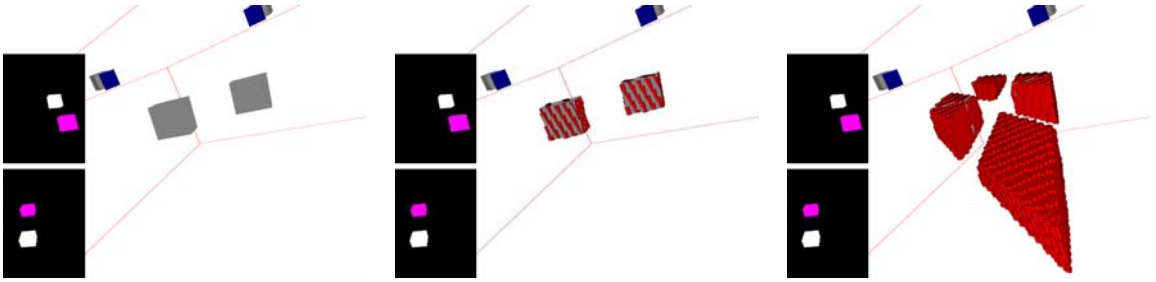
Dans la suite nous présentons les résultats de nos contributions par rapport à plusieurs axes d'expérimentation. Chaque configuration a été choisie pour la représentation de caractéristiques particulières, qui permettent d'illustrer les différences entre les approches. *Shape-From-Silhouette* souffre de plusieurs limitations, alors il est nécessaire d'étudier chaque limitation indépendamment, de façon a illustrer au mieu les intérêts et défauts de chaque méthode.

Nous commençons par présenter une étude qualitative des mécanismes mis en place pour la suppression des objets fantômes dans les approches *Shape-From-Silhouette*. Nous présentons ensuite les apports réalisés par notre approche d'*enveloppes visuelles étendues*. Ensuite nous soulignons la robustesse de l'approche $\widetilde{\text{EVE}}_n$ par rapport à une information de silhouette bruitée. Enfin nous présenterons une analyse comparative entre l'approche *Shape-From-Silhouette* classique face au pipeline de traitement qui profite de l'ensemble de nos contributions.

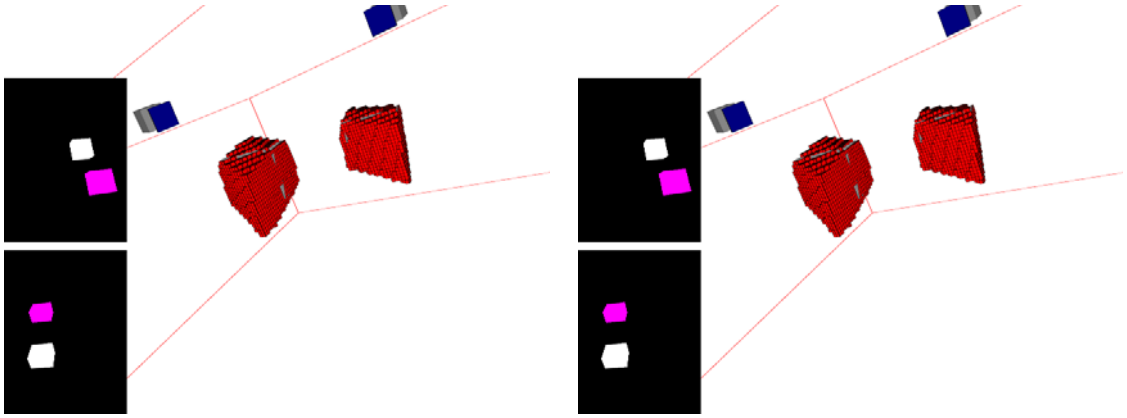
8.2 Mécanismes de suppression d'objets fantômes

Dans ce premier pan d'expérimentations, nous choisissons plusieurs configurations qui impliquent la génération d'objets fantômes. A travers ces trois configurations, nous allons illustrer

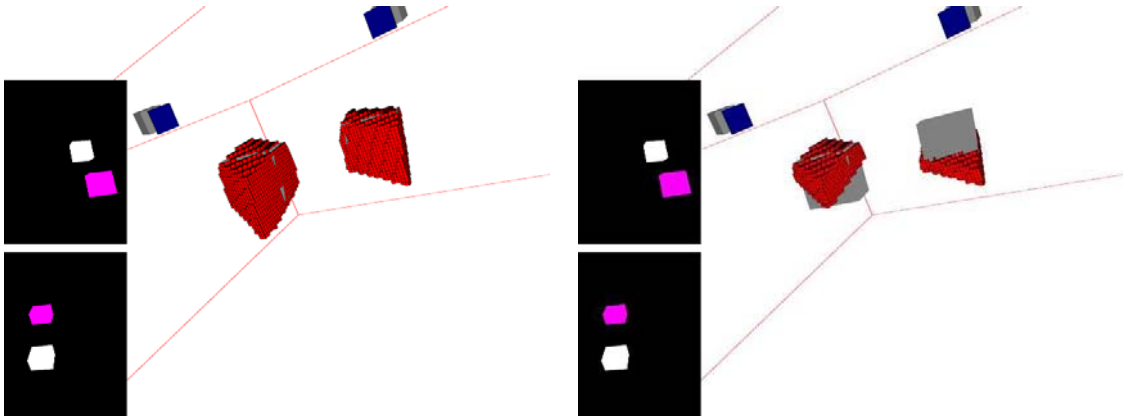
l'efficacité des approches EOR et EOR_{seq} décrites dans le chapitre 6, de l'approche $pcSFS$ présentée dans le chapitre 5, face aux approches classique de SFS et la méthode Safe Hulls [MH07]. Comme observé dans le chapitre 5 l'approche $pcSFS$ proposée, est équivalente à la méthode *Reduced Visual Hull* décrite par Bogomjakov et Gotsman dans [BG08]. Ainsi l'ensemble des évaluations réalisées pour $pcSFS$ s'appliquent aussi à l'approche *Reduced Visual Hull*. L'ensemble de ces approches permettent la suppression d'objets fantômes générés par une approche estimant l'enveloppe visuelle telle que EVE_m ou encore $\widetilde{\mathrm{EVE}_m}$ dans les sections suivantes.



(a) Objets d'intérêt (b) Objets d'intérêt voxelisés (c) SFS



(d) EOR (e) EOR_{seq}



(f) $pcSFS$ (g) $SafeHulls$

FIG. 8.3 – Reconstructions offertes par les méthodes SFS (c), EOR (d), EOR_{seq} (e), $pcSFS$ (f) et par l'approche $SafeHulls$ (g) proposée par Miller et Hilton dans [MH07]. La scène contient deux objets d'intérêt observés depuis deux caméras (a). L'image (b) représente la voxélisation des objets d'intérêt. Cette configuration génère deux objets fantômes correctement supprimés par les approches EOR , EOR_{seq} et $pcSFS$.

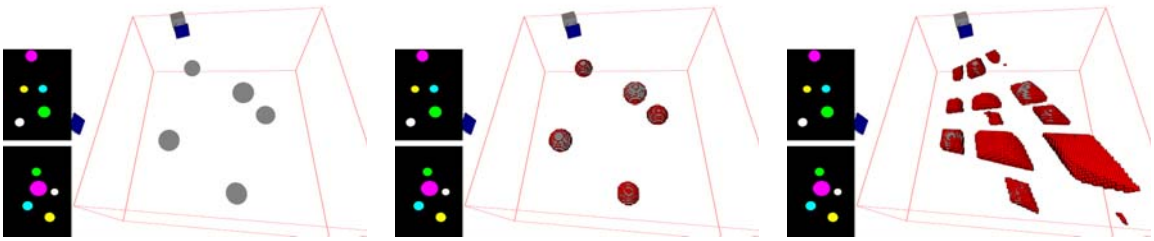
Notre première expérimentation met en avant les capacités de suppression d'objets fantômes dans une scène simple, composée de deux pavés disjoints. Le tableau 8.4 présente les valeurs des critères d'évaluation, en fonction du nombre de caméras utilisées n.

n	SFS	EOR	EOR_{seq}	$pcSFS$	$SafeHulls$	
2	1.000	1.000	1.000	1.000	0.381	R
	2	0	0	0	0	F
	0.217	0.438	0.438	0.438	0.424	P
3	1.000	1.000	1.000	1.000	1.000	R
	0	0	0	0	0	F
	0.572	0.572	0.572	0.572	0.572	P
4	0.999	0.999	0.999	0.999	0.999	R
	0	0	0	0	0	F
	0.556	0.556	0.556	0.556	0.556	P
5	0.999	0.999	0.999	0.999	0.999	R
	0	0	0	0	0	F
	0.595	0.595	0.595	0.595	0.595	P
6	0.999	0.999	0.999	0.999	0.999	R
	0	0	0	0	0	F
	0.808	0.808	0.808	0.808	0.808	P
7	0.999	0.999	0.999	0.999	0.999	R
	0	0	0	0	0	F
	0.810	0.810	0.810	0.810	0.810	P
8	0.999	0.999	0.999	0.999	0.999	R
	0	0	0	0	0	F
	0.812	0.812	0.812	0.812	0.812	P

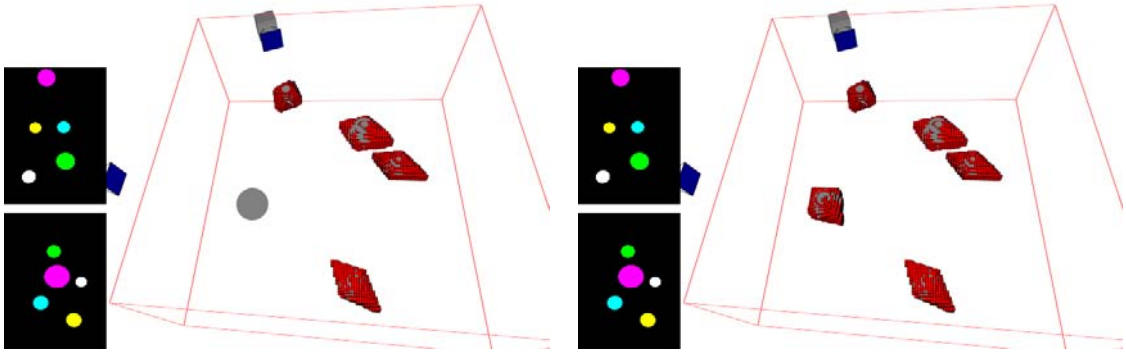
FIG. 8.4 – Valeurs des critères d'évaluation pour la configuration de scène présentée figure 8.3. Avec $n = 2$ caméras, les approches EOR, EOR_{seq}, $pcSFS$ offrent de meilleurs résultats que les approches SFS et $SafeHulls$. SFS offre une reconstruction complète, mais génère deux objets fantômes, alors que $SafeHulls$ fournit une géométrie corrompue. Pour $n > 2$, ces différentes méthodes offrent les mêmes reconstructions. Nous observons que la précision P augmente lorsque le nombre de caméras croit.

Lorsque $n = 2$ (voir figure 8.3), l'ensemble des approches que nous avons proposé suppriment les deux objets fantômes de SFS. L'approche $SafeHulls$ supprime les objets fantômes, mais ne contient pas la totalité des objets d'intérêt ($R = 0,381$). Lorsque le nombre de caméras augmente, l'ensemble des méthodes exposées offrent des résultats similaires.

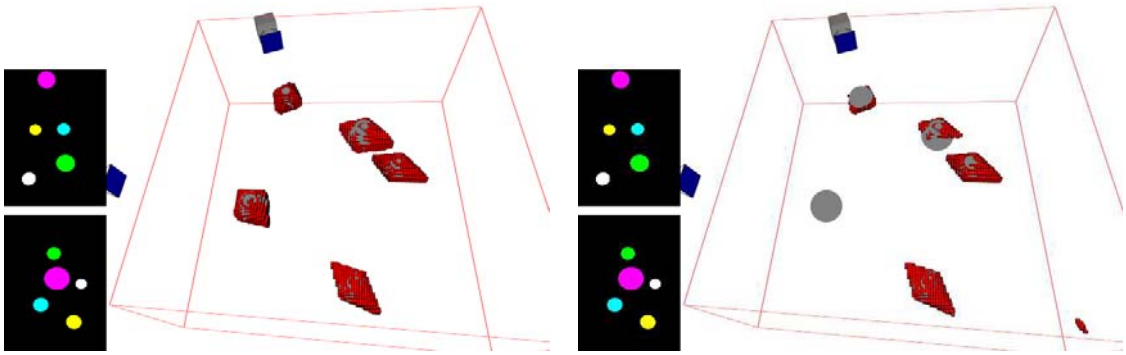
En présence d'un faible nombre de vues, EOR, EOR_{seq} et $pcSFS$ offrent la meilleure reconstruction, alors que les reconstructions proposées par l'approche de Miller et Hilton sont parfois incomplètes.



(a) Objets d'intérêt (b) Objets d'intérêt voxelisés (c) SFS



(d) EOR (e) EOR_{seq}



(f) $pcSFS$ (g) $SafeHulls$

FIG. 8.5 – Reconstruction à partir de deux points de vue des objets d'intérêt (a) par les méthodes SFS (c), EOR (d), EOR_{seq} (e), $pcSFS$ (f) et par l'approche $SafeHulls$ (g). Dans cette configuration les méthodes EOR_{seq} et $pcSFS$ sont les seules à supprimer exactement tous les objets fantômes de SFS. Par rapport à $pcSFS$, EOR_{seq} n'utilise pas l'information de couleur. Notons que EOR reconstruit une forme incomplète.

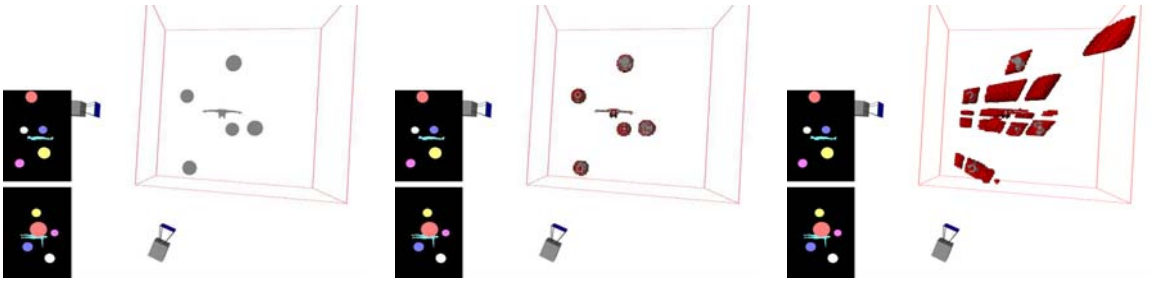
L'expérimentation suivante est construite sur une scène composée de 5 sphères (voir figure 8.5). Les valeurs correspondantes des critères d'évaluation sont présentées dans le tableau 8.6. Lorsque $n = 2$, SFS génère 8 objets fantômes. Les approches EOR_{seq} et $pcSFS$ suppriment

n	SFS	EOR	EOR_{seq}	$pcSFS$	$SafeHulls$	
2	0.997	0.797	0.997	0.997	0.511	R
	8	0	0	0	1	F
	0.291	0.355	0.701	0.701	0.671	P
3	0.997	0.997	0.997	0.997	0.997	R
	1	0	0	0	0	F
	0.765	0.791	0.791	0.791	0.791	P
4	0.996	0.996	0.996	0.996	0.996	R
	0	0	0	0	0	F
	0.817	0.817	0.817	0.817	0.817	P
5	0.996	0.996	0.996	0.996	0.996	R
	0	0	0	0	0	F
	0.889	0.889	0.889	0.889	0.889	P
6	0.995	0.995	0.995	0.995	0.995	R
	0	0	0	0	0	F
	0.940	0.940	0.940	0.940	0.940	P
7	0.994	0.994	0.994	0.994	0.994	R
	0	0	0	0	0	F
	0.963	0.963	0.963	0.963	0.963	P
8	0.993	0.993	0.993	0.993	0.993	R
	0	0	0	0	0	F
	0.967	0.967	0.967	0.967	0.967	P

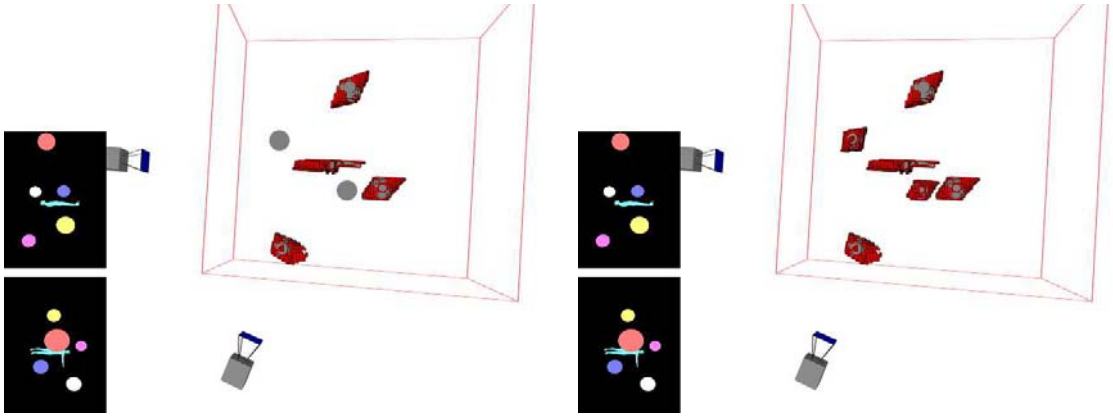
FIG. 8.6 – Valeurs des critères d'évaluation pour la configuration de scène présentée figure 8.5. Avec le nombre de caméras $n = 2$, EOR, et $SafeHulls$ offrent de moins bons résultats que les approches EOR_{seq}, $pcSFS$. Pour $n = 3$, en dehors de SFS, toutes les méthodes offrent la même reconstruction.

exactement les objets fantômes, ainsi ces approches se révèlent plus de deux fois plus précises que l'approche SFS. La méthode EOR supprime tous les objets fantômes mais ne contient pas l'une des sphères, car plusieurs composantes connexes de la reconstruction recouvrent exactement les composantes de silhouette associées à cette sphère. Grâce au critère de silhouette équivalence EOR_{seq} se montre plus efficace que EOR. Enfin l'approche $SafeHulls$ échoue en construisant un objets fantômes tout en fournissant une estimation incomplète : seul 51% des points des objets d'intérêt sont reconstruits. Lorsque $n = 3$, EOR, EOR_{seq}, $pcSFS$ et $SafeHulls$ suppriment correctement l'unique objet fantôme de SFS. Pour les autres valeurs de n les différentes méthodes offrent la même reconstruction.

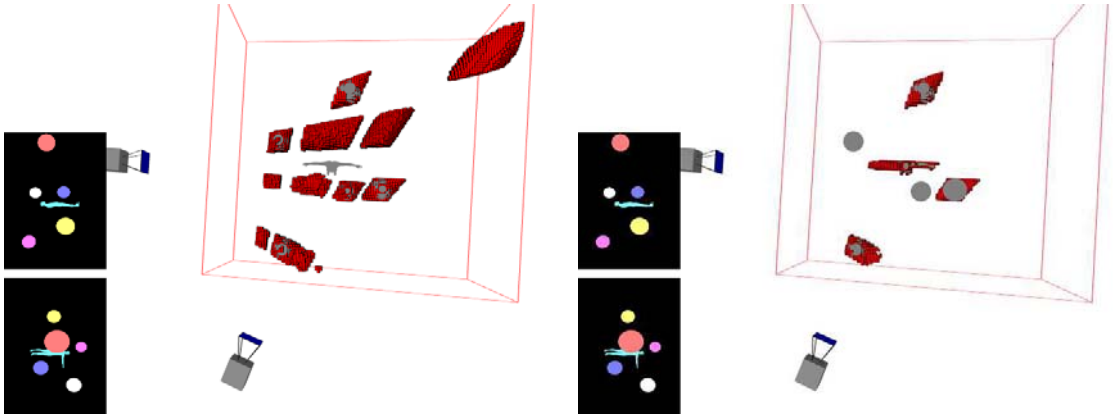
Dans cette configuration les méthodes les plus efficaces sont EOR_{seq} et $pcSFS$. Notons que $pcSFS$ nécessite cependant une information de couleur, afin de réaliser la mise en correspondance des composantes connexes de silhouette.



(a) Objets d'intérêt (b) Objets d'intérêt voxelisés (c) SFS



(d) EOR (e) EOR_{seq}



(f) $pcSFS$ (g) $SafeHulls$

FIG. 8.7 – Résultats de reconstruction des objets d'intérêt à partir de 2 caméras. Seule l'approche EOR_{seq} supprime exactement les objets fantômes.

La configuration présentée figure 8.7 est proche de celle utilisée lors de l'expérimentation précédente. Cependant l'ajout d'un objet génère des occultations dans les silhouettes, qui mettent en échec les mises en correspondance nécessaires à l'approche $pcSFS$. Ainsi pour l'ensemble des évaluations présentées Tableau 8.8, $pcSFS$ se trouve en échec. Dans cette configuration difficile, EOR_{seq} est la seule approche qui supprime exactement les objets fantômes de SFS et ce, quel que soit le nombre de caméras.

n	SFS	EOR	EOR_{seq}	$pcSFS$	$SafeHulls$	
2	0.998	0.634	0.998	0.911	0.305	R
	11	0	0	8	0	F
	0.207	0.526	0.584	0.213	0.534	P
3	0.996	0.996	0.996	0.907	0.940	R
	7	0	0	4	0	F
	0.638	0.760	0.760	0.656	0.758	P
4	0.998	0.997	0.997	0.906	0.997	R
	1	0	0	0	0	F
	0.786	0.798	0.798	0.804	0.798	P
5	0.997	0.997	0.997	0.912	0.997	R
	0	0	0	0	0	F
	0.869	0.869	0.869	0.885	0.869	P
6	0.997	0.997	0.997	0.910	0.997	R
	0	0	0	0	0	F
	0.922	0.922	0.922	0.935	0.922	P
7	0.995	0.995	0.995	0.912	0.995	R
	0	0	0	0	0	F
	0.938	0.938	0.938	0.951	0.938	P
8	0.993	0.993	0.993	0.906	0.993	R
	0	0	0	0	0	F
	0.942	0.942	0.942	0.953	0.942	P

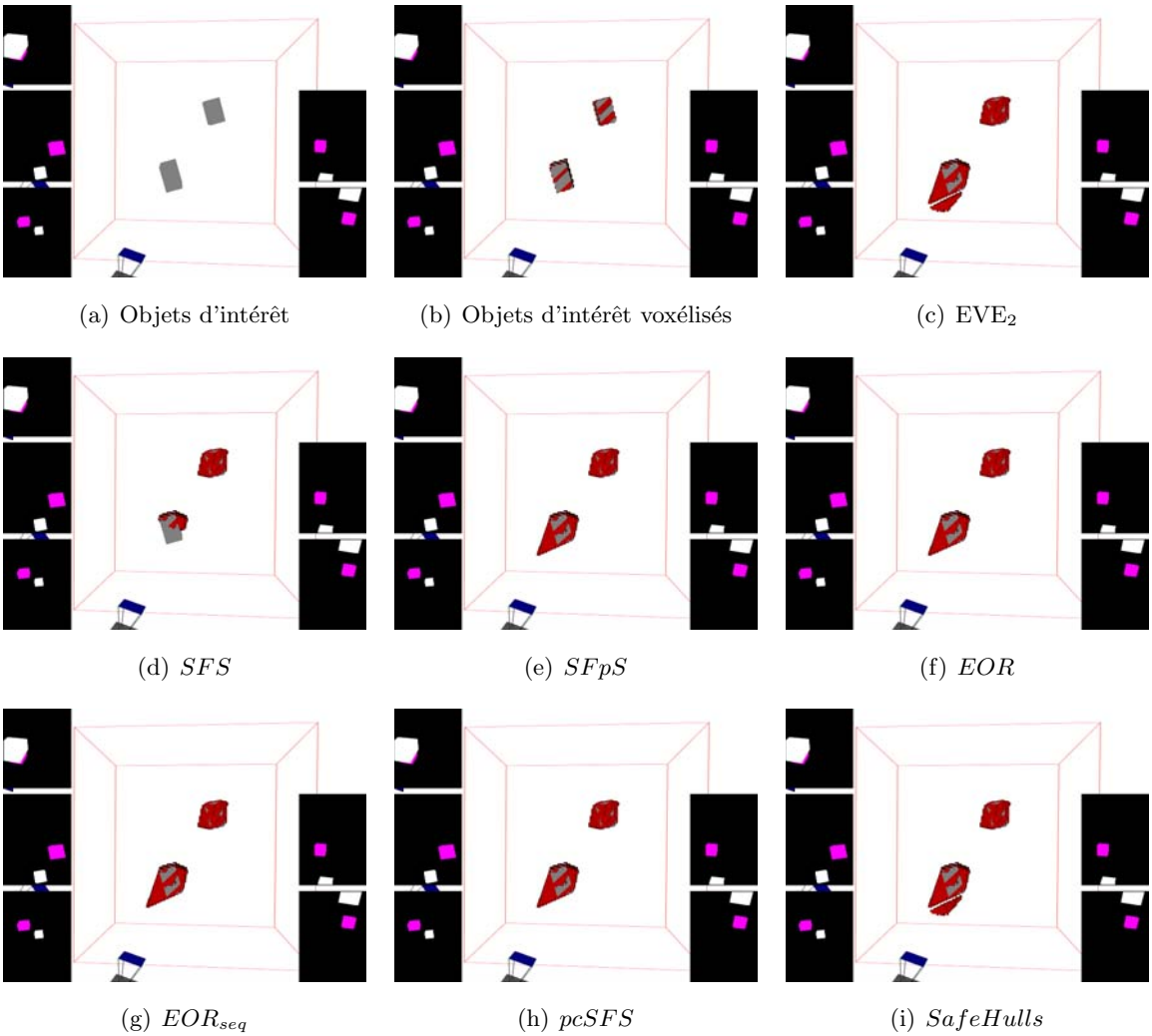
FIG. 8.8 – Valeurs des critères d'évaluation pour la configuration de scène présentée figure 8.7. Avec le nombre de caméras $n = 2$, seule l'approche EOR_{seq} supprime exactement les objets fantômes de SFS. Pour les valeurs de $n > 3$, EOR, $SafeHulls$ et EOR_{seq} fournissent la même reconstruction.

Pour ces trois expérimentations l'approche EOR_{seq} s'est montrée la plus efficace pour la suppression "exacte" des objets fantômes, pour toute valeur de $n \in [2, \cdots 8]$. Cette approche s'est même montrée particulièrement performante pour un faible nombre de caméras. Nous rappelons qu'il est cependant possible que EOR_{seq} ne supprime pas la totalité des objets fantômes (voir chapitre 6.2), mais nous n'avons pour le moment pas observé de configuration pour laquelle cela se produit.

Dans la suite nous nous intéressons à évaluer les intérêts de la formulation EVE_m, filtrée par les différentes approches de limitation d'objets fantômes telles que $SFpS$, ou encore de suppression : EOR, EOR_{seq}, $SafeHulls$ et $pcSFS$.

8.3 Extension du volume d'acquisition par l'approche des enveloppes visuelles étendues

Dans la suite nous présentons plusieurs expérimentations qui soulignent les intérêts des enveloppes visuelles étendues. Nous comparons les approches EVE_m et SFS ainsi que l'approche $SFpS$ qui profite de l'information issue de SFS afin de limiter la quantité d'objets fantômes reconstruits. Nous étudions ensuite l'application des approches EOR, EOR_{seq}, $pcSFS$ et $SafeHulls$ appliquées à la reconstruction offerte par EVE_m.



(a) Objets d'intérêt (b) Objets d'intérêt voxélisés (c) EVE_2

(d) SFS (e) $SFpS$ (f) EOR

(g) EOR_{seq} (h) $pcSFS$ (i) $SafeHulls$

FIG. 8.9 – Représentation d'une scène simple contenant deux objets d'intérêt observés par cinq caméras. L'un des deux pavés est partiellement visible pour 3 des 5 caméras. Ainsi SFS est incapable de le reconstruire totalement, alors que EVE_2 les contient tous. Toutes nos contributions suppriment exactement les objets fantômes. Cependant l'approche $SafeHulls$ en est incapable.

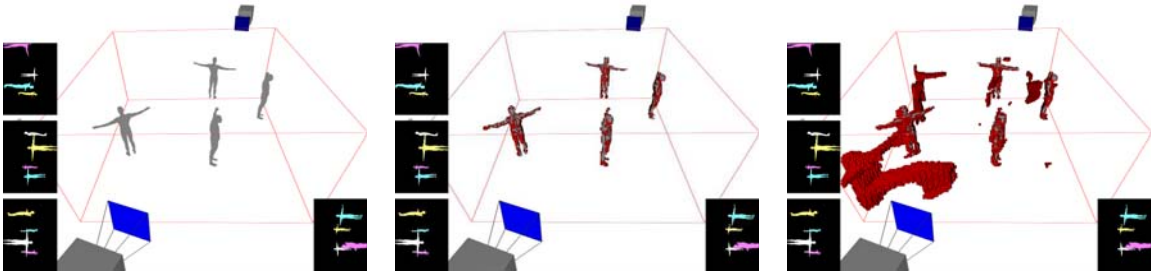
Cette nouvelle expérimentation est réalisée à partir de la configuration de scène présentée figure 8.9. La majorité des caméras sont disposées afin de respecter l'hypothèse de visibilité

n	EVE_2	SFS	$SFpS$	EOR	EOR_{seq}	$pcSFS$	$SafeHulls$	
2	0.999	0.999	0.999	0.999	0.999	0.999	0.999	R
	0	0	0	0	0	0	0	F
	0.436	0.436	0.436	0.436	0.436	0.436	0.436	P
3	0.999	0.954	0.999	0.999	0.999	0.999	0.999	R
	1	0	0	0	0	0	2	F
	0.442	0.603	0.452	0.452	0.452	0.452	0.449	P
4	0.999	0.810	0.999	0.999	0.999	0.999	0.999	R
	2	0	0	0	0	0	1	F
	0.472	0.564	0.490	0.490	0.490	0.490	0.483	P
5	0.999	0.719	0.999	0.999	0.999	0.999	0.999	R
	1	0	0	0	0	0	1	F
	0.503	0.610	0.516	0.516	0.516	0.516	0.512	P
6	0.999	0.719	0.999	0.999	0.999	0.999	0.999	R
	1	0	0	0	0	0	1	F
	0.589	0.809	0.608	0.608	0.608	0.608	0.594	P
7	0.999	0.719	0.999	0.999	0.999	0.999	0.999	R
	1	0	0	0	0	0	1	F
	0.591	0.513	0.610	0.610	0.610	0.610	0.595	P
8	0.999	0.673	0.999	0.999	0.999	0.999	0.999	R
	1	0	0	0	0	0	1	F
	0.417	0.826	0.612	0.612	0.612	0.612	0.603	P

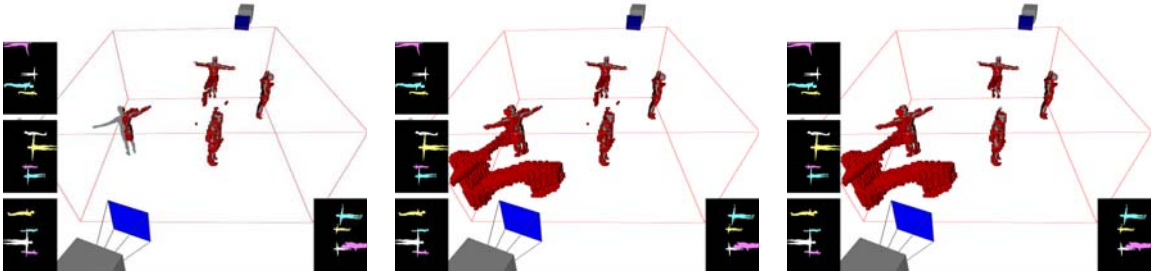
FIG. 8.10 – Valeurs des critères d'évaluation pour la scène présentée figure 8.9. Nous sommes dans une configurations ou la majorité des caméras ne voient pas totalement l'un des objets : l'hypothèse visibilité partielle est respectée. Pour $n = 2$ les deux objets d'intérêt sont totalement vus dans toutes les caméras. Ainsi EVE_2 et SFS fournissent la même reconstruction. Plus le nombre de caméras n augmente, moins SFS reconstruit les objets d'intérêt. Cela se traduit par le fait que les valeurs R de SFS diminuent lorsque le nombre de caméra augmente. Les approches $SFpS$, EOR, EOR_{seq} et $pcSFS$ suppriment exactement les objets fantômes générés par EVE_2. Cependant l'approche $SafeHulls$ en est incapable.

partielle. Cela revient à supposer que SFS est capable de reconstruire au moins une partie de chaque objet. Dans cette configuration l'extension $SFpS$ (*Shape From partial Silhouette*) supprime l'ensemble des objets fantômes générés par EVE_2.

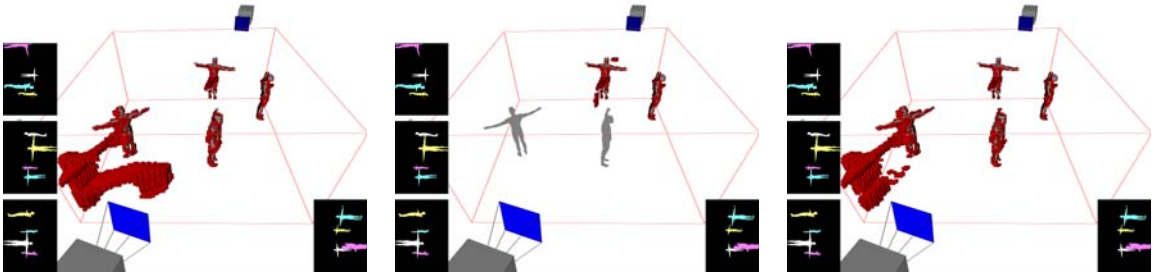
Les critères d'évaluation présentés au tableau 8.10 indiquent que les approches EVE_2, $SFpS$, EOR, EOR_{seq} et $pcSFS$ reconstruisent la totalité des objets de la scène. De plus $SFpS$, EOR, EOR_{seq} et $pcSFS$ suppriment l'ensemble des objets fantômes générés par EVE_2. Notons que $SFpS$ fournit des résultats satisfaisants car SFS ne contient pas d'objet fantômes. Ceci provient du faible nombre d'objets dans la scène. Nous verrons dans l'expérimentation suivante que $SFpS$ contient le même nombre d'objets fantômes que SFS.



(a) Objets d'intérêt (b) Objets d'intérêt voxélisés (c) EVE_2



(d) SFS (e) $SFpS$ (f) EOR



(g) EOR_{seq} (h) $pcSFS$ (i) $SafeHulls$

FIG. 8.11 – Estimation de 4 humanoïdes filmés par 4 caméras, dont un n'est que partiellement visible par deux caméras. Dans cette configuration SFS est incapable de reconstruire la totalité des objets d'intérêt et construit 3 objets fantômes. Pour cette configuration, $SFpS$ contient le même nombre d'objets fantômes que SFS et supprime une partie des objets fantômes calculés par EVE_2. Seules les approches EOR et EOR_{seq} permettent de supprimer exactement les objets fantômes de EVE_2. Nous observons que par rapport au critère de précision, $pcSFS$ se montre l'approche la plus performante.

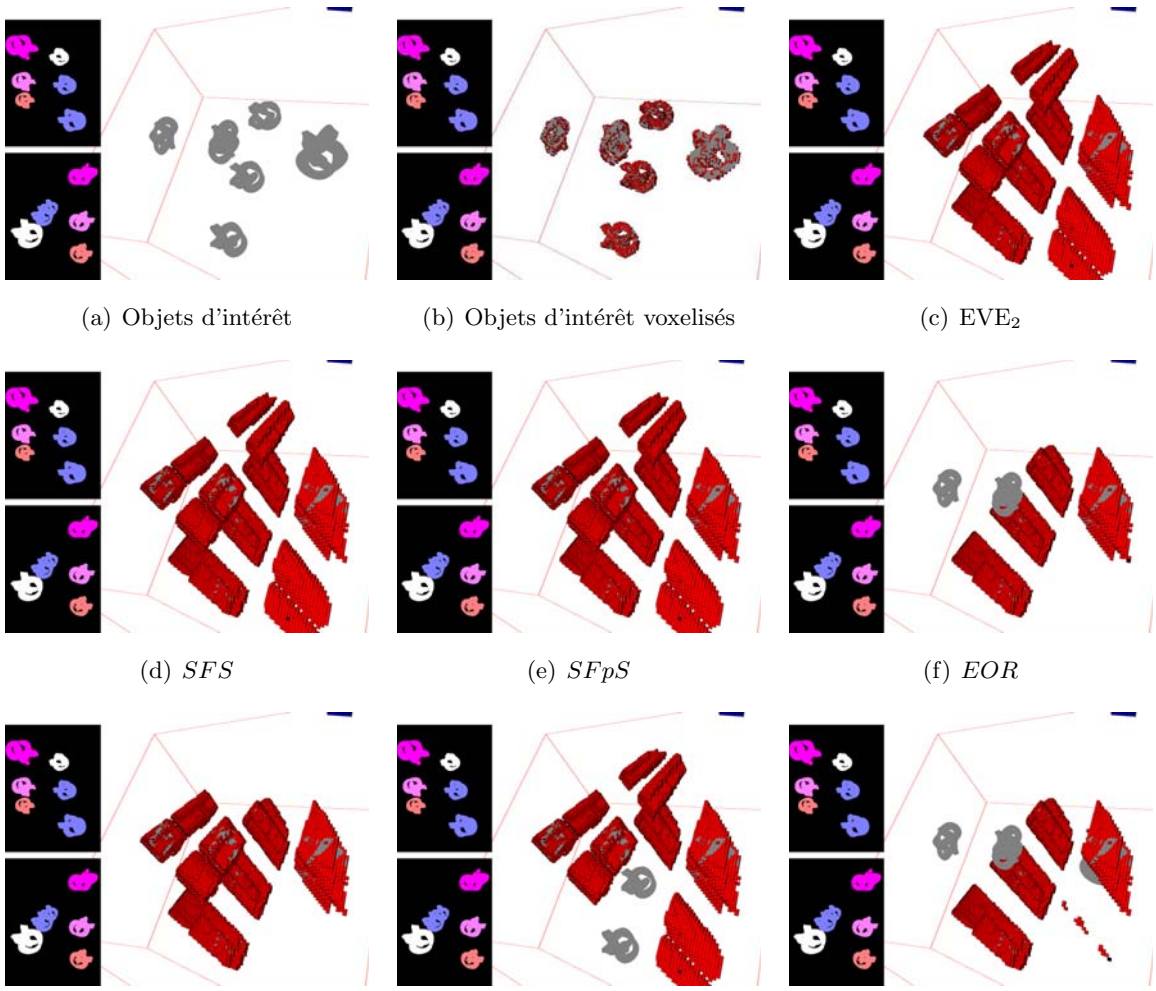
Cette nouvelle configuration de scène est composée de modèles d'humanoïdes (voir figure 8.11), disposés de telle sorte que l'hypothèse de visibilité partielle soit respectée. A nouveau SFS ne peut reconstruire la totalité des objets d'intérêt. Nous observons que EVE_2 ajoute beaucoup d'objets fantômes, par rapport à l'approche SFS. Pour toutes valeur de n, seules les approches EOR et EOR_{seq} suppriment exactement les objets fantômes (voir tableau 8.12). $SFpS$ contient le même nombre d'objets fantômes que SFS. Cette approche ne permet alors pas de supprimer les tous les objets fantômes générés par EVE_2. Les reconstructions offertes pas

n	EVE_2	SFS	$SFpS$	EOR	EOR_{seq}	$pcSFS$	$SafeHulls$	
2	0.995	0.995	0.995	0.995	0.995	0.747	0.480	R
	2	2	2	0	0	0	4	F
	0.111	0.111	0.111	0.115	0.115	0.124	0.192	P
3	0.997	0.865	0.997	0.997	0.997	0.499	0.924	R
	6	2	2	0	0	1	2	F
	0.192	0.508	0.207	0.210	0.210	0.484	0.301	P
4	0.995	0.845	0.995	0.995	0.995	0.498	0.981	R
	6	3	3	0	0	2	2	F
	0.192	0.624	0.226	0.229	0.229	0.625	0.350	P
5	0.993	0.843	0.993	0.993	0.993	0.497	0.993	R
	6	0	0	0	0	0	2	F
	0.192	0.756	0.698	0.698	0.698	0.762	0.667	P
6	0.992	0.839	0.992	0.992	0.992	0.496	0.991	R
	6	0	0	0	0	0	3	F
	0.190	0.799	0.758	0.758	0.758	0.804	0.721	P
7	0.992	0.839	0.992	0.992	0.992	0.496	0.992	R
	2	0	0	0	0	0	2	F
	0.245	0.821	0.798	0.798	0.798	0.804	0.757	P
8	0.991	0.839	0.991	0.991	0.991	0.496	0.991	R
	2	0	0	0	0	0	1	F
	0.467	0.879	0.826	0.826	0.826	0.844	0.804	P

FIG. 8.12 – Valeurs des critères d'évaluation pour la scène présentée figure 8.11. Nous sommes dans une configurations ou la majorité des caméra ne voient pas totalement deux des modèles humanoïdes. L'hypothèse de visibilité partielle est ainsi respectée. Pour $n = 2$ les objets d'intérêt sont totalement vus dans toutes les caméras. Ainsi EVE_2 et SFS fournissent la même reconstruction. Plus le nombre de caméras n augmente, moins SFS reconstruit les objets d'intérêt. Cela traduit par la diminution des valeurs R de SFS lorsque le nombre de caméra augmente.

$pcSFS$ ne sont pas correctes car l'étape de mise en correspondance des composantes connexes de silhouettes tombe en échec. En effet certains objets sont occultés par d'autres pour plusieurs caméras. Enfin l'approche $SafeHulls$ est incapable de construire la totalité des objets d'intérêt lorsque $n = 2$ et $n = 4$. De plus cette approche ne supprime pas toujours l'ensemble des objets fantômes de EVE_2 (voir figure 8.11).

Il ressort de cette expérimentation que seules les approches EOR et EOR_{seq} permettent de filtrer correctement la reconstruction proposée par EVE_2. En effet quel que soit le nombre de caméras compris entre 2 et 8, ces deux approches suppriment uniquement les objets fantômes de la reconstruction EVE_2. Les autres approches tombent en échec pour au moins l'une des configurations.



(a) Objets d'intérêt (b) Objets d'intérêt voxelisés (c) EVE_2

(d) SFS (e) $SFpS$ (f) EOR

(g) EOR_{seq} (h) $pcSFS$ (i) $SafeHulls$

FIG. 8.13 – Estimation de 6 noeuds gordiens filmés par 2 caméras. Dans la mesures où tous les objets sont totalement visibles dans toutes les caméras, les reconstructions EVE_2, SFS et donc $SFpS$ sont équivalentes. Seule l'approche EOR_{seq} reconstruit tous les objets d'intérêt. Cependant elle ne supprime pas tous les objets fantômes. Ce résultat provient du fait qu'il existe plusieurs union-minimales. L'approche $SafeHulls$ génère de nouveaux objets fantômes et ne contient pas la totalité des objets d'intérêt. L'approche $pcSFS$ est mise en échec car certains objets sont occultés par d'autres dans l'une des vues.

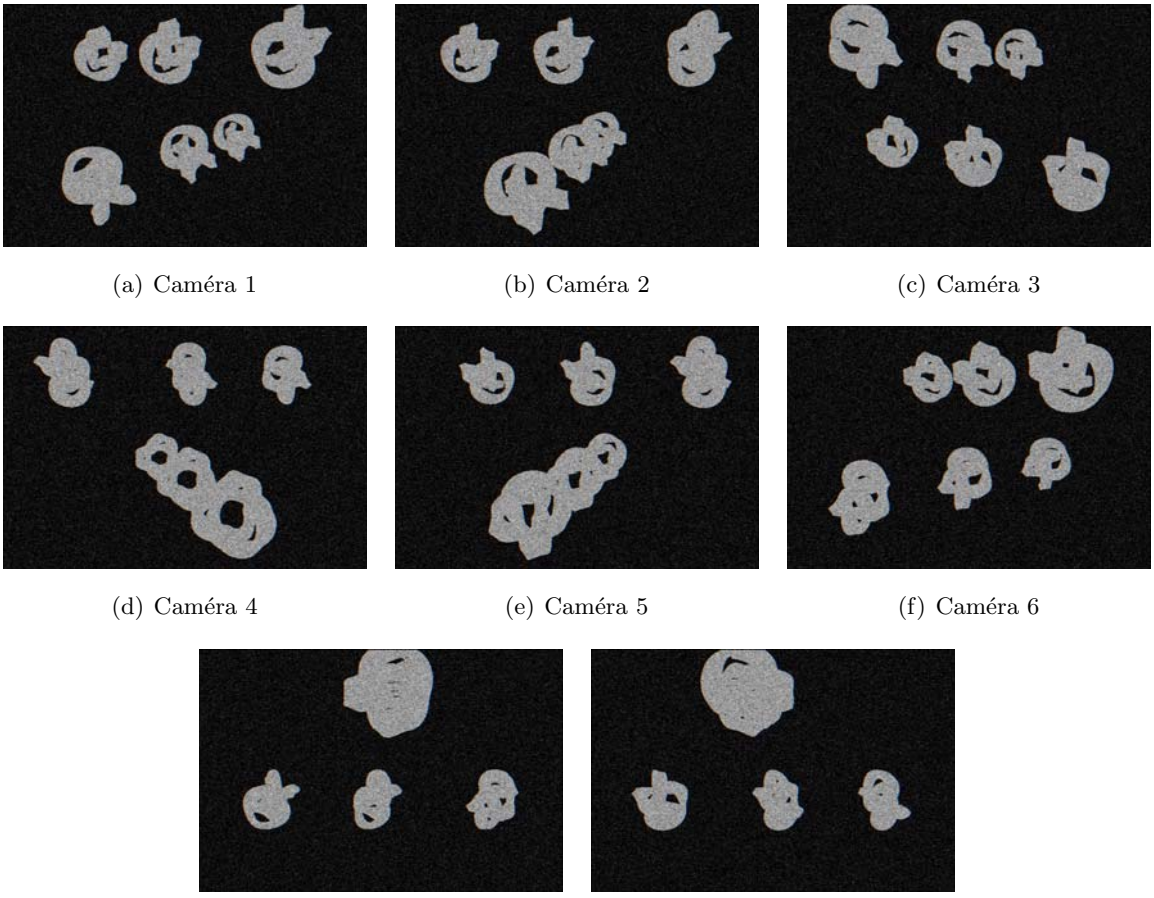
Cette dernière expérimentation est réalisée à partir de la scène présentée figure 8.13. Les critères d'évaluation indiqués tableau 8.14 soulignent les apports de l'approche EVE_m ainsi que de l'approche EOR_{seq}. Seules ces deux approches contiennent la totalité des objets d'intérêt. Nous observons aussi que le rappel R de l'approche SFS diminue lorsque le nombre de caméras augmente.

n	EVE_2	SFS	$SFpS$	EOR	EOR_{seq}	$pcSFS$	$SafeHulls$	
2	0.997	0.997	0.997	0.853	0.997	0.997	0.567	R
	11	11	11	0	4	11	3	F
	0.269	0.269	0.269	0.455	0.644	0.270	0.466	P
3	0.999	0.958	0.999	0.999	0.999	0.429	0.996	R
	9	4	4	0	0	2	1	F
	0.450	0.597	0.584	0.614	0.614	0.493	0.611	P
4	0.998	0.948	0.998	0.998	0.998	0.429	0.998	R
	5	0	0	0	0	0	1	F
	0.487	0.661	0.653	0.653	0.653	0.619	0.637	P
5	0.998	0.944	0.998	0.998	0.998	0.429	0.998	R
	4	0	0	0	0	0	1	F
	0.507	0.724	0.720	0.720	0.720	0.719	0.719	P
6	0.997	0.941	0.997	0.853	0.997	0.429	0.997	R
	3	0	0	0	0	0	0	F
	0.551	0.812	0.805	0.800	0.805	0.802	0.805	P
7	0.997	0.940	0.997	0.858	0.997	0.000	0.997	R
	3	0	0	0	0	0	0	F
	0.600	0.600	0.836	0.834	0.836	0.000	0.836	P
8	0.996	0.940	0.996	0.996	0.996	0.000	0.996	R
	1	0	0	0	0	0	0	F
	0.749	0.847	0.841	0.841	0.841	0.000	0.841	P

FIG. 8.14 – Valeurs des critères d'évaluation pour la scène présentée figure 8.13. Nous avons choisi une configuration pour laquelle l'hypothèse de visibilité partielle est respectée. Pour $n = 2$ les objets d'intérêt sont totalement visibles depuis toutes les caméras. Ainsi EVE_2, SFS et $SFpS$ fournissent la même reconstruction. Plus le nombre de caméras n augmente, moins SFS reconstruit les objets d'intérêt. Pour toutes valeurs de n seule l'approche EOR_{seq} respecte les critères de qualité fixés : elle reconstruit toujours l'ensemble des points qui appartiennent aux objets d'intérêt et supprime le plus souvent tous les objets fantômes générés par EVE_m.

A travers les expérimentations que nous avons mené au sein de cette section, les approches qui offrent les meilleurs résultats selon nos critères, sont EVE_m filtrée par EOR_{seq}. Dans la suite nous nous intéressons à la dernière contribution que nous avons proposée : rendre les approches SFS est EVE_m robustes aux silhouettes bruitées.

8.4 Estimation robuste de forme 3D à partir de cartes probabilistes de silhouette



(a) Caméra 1 (b) Caméra 2 (c) Caméra 3

(d) Caméra 4 (e) Caméra 5 (f) Caméra 6

(g) Caméra 7 (h) Caméra 8

FIG. 8.15 – Cartes de probabilité de silhouette issues de 8 caméras qui observent les 6 noeuds gordiens présentés figure 8.18. Le bruit synthétique additif suit une loi normale $\mathcal{N}(0;\, 0,15^2)$.

Dans le chapitre 7, nous avons proposé l'approche $\widetilde{\text{EVE}}_m$ qui estime EVE_m à partir de cartes de probabilité de silhouettes, en fusionnant l'information disponible. Nous présentons dans cette section une expérimentation qui souligne la robustesse de l'approche $\widetilde{\text{EVE}}_m$ aux bruit des silhouettes, comparée à l'approche EVE_m. Pour une même configuration de scène, nous avons évalué le critère de rappel R des méthodes $\widetilde{\text{EVE}}_2$ et EVE_2 pour différentes intensités de bruit injecté dans l'information de silhouettes.

Dans le chapitre 7 nous avons supposé un bruit additif dans les silhouettes observée $\tilde{\mathcal{S}}_i$, suivant la loi normale de paramètres $\mathcal{N}(0;\, \sigma^2)$. Pour simuler ce bruit additif, nous utilisons la méthode de Box-Muller [BM58] qui génère des paires de nombres aléatoires à distribution normale centrée réduite, à partir d'une source de données aléatoires suivant une loi uniforme. Nous estimons cette dernière à partir du générateur pseudo-aléatoire *Mersenne Twister*, proposé par

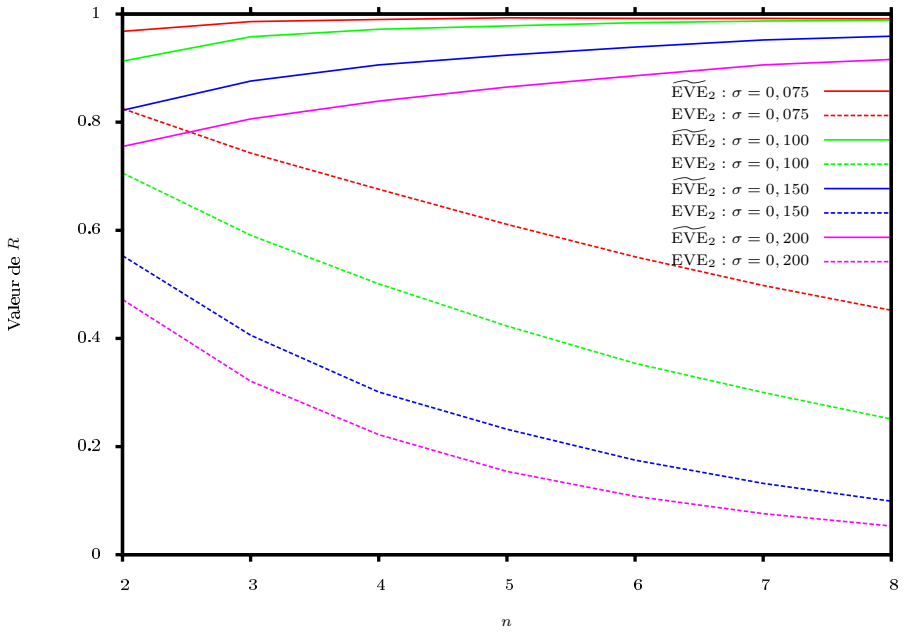


FIG. 8.16 – Rappel des approches $\widetilde{\text{EVE}}_2$ et de EVE_2 en fonction du paramètre de bruit σ. Lorsque le nombre de caméras augmente, le rappel de $\widetilde{\text{EVE}}_2$ converge vers 1 alors que le rappel de EVE_2 converge vers 0. Ces convergences sont cependant moins rapides lorsque le paramètre de bruit σ est important. Même en présence d'un bruit important ($\sigma = 0.200$) $\widetilde{\text{EVE}}_2$ construit plus de 75% des points d'intérêt.

Matsumoto et Nishimura [MN98]. Les données étant synthétiques, nous calculons les silhouettes "parfaites" des objets d'intérêt, en associant à chaque pixel de silhouette une valeur $T_M \in [0, 1]$ supérieure au seuil de binarisation T. Ensuite nous perturbons ces informations en ajoutant le bruit additif de loi normale de paramètres $\mathcal{N}(0; \sigma^2)$. La figure 8.15 présente les cartes de probabilité de silhouettes générées avec $T_M = 0,6$ et $\sigma = 0,15$ pour la scène présentée figures 8.13 et 8.18.

Les graphiques 8.16 (respectivement 8.17) comparent les valeurs de rappel (respectivement de précision) calculées pour les approches $\widetilde{\text{EVE}}_2$ et EVE_2 pour différentes intensités de bruit.

La figure 8.16 souligne les apports de l'approche $\widetilde{\text{EVE}}_2$ comparée à la méthode déterministe EVE_2. Nous observons que quel que soit le nombre de caméras et pour toute valeur de bruit, le rappel de $\widetilde{\text{EVE}}_2$ est supérieur au rappel de EVE_2. Comme indiqué dans le chapitre 7 décrivant l'approche $\widetilde{\text{EVE}}_2$, lorsque le nombre de caméras augmente, le rappel de l'approche $\widetilde{\text{EVE}}_2$ augmente, alors que celui de EVE_2 diminue. Plus le nombre de vues en entrée est important, plus l'approche $\widetilde{\text{EVE}}_2$ compense le bruit présent dans chaque silhouettes.

Cette expérimentation ne présente pas explicitement la comparaison entre $\widetilde{\text{EVE}}_2$ et SFS. En effet nous avons démontré au paragraphe 5.1, propriété 7 que $SFS(\mathcal{O}) \subseteq \text{EVE}_m(\mathcal{O})$. Nous

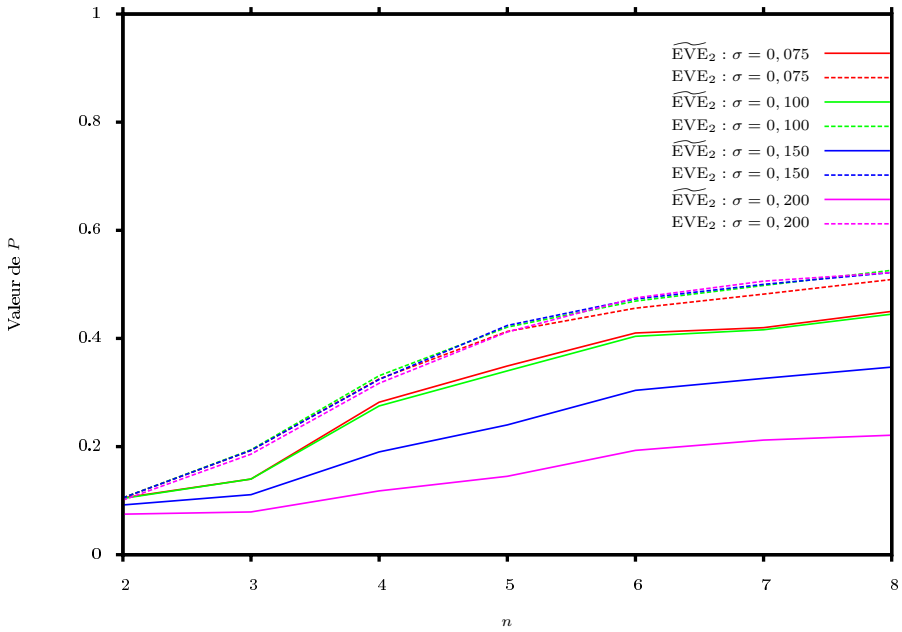
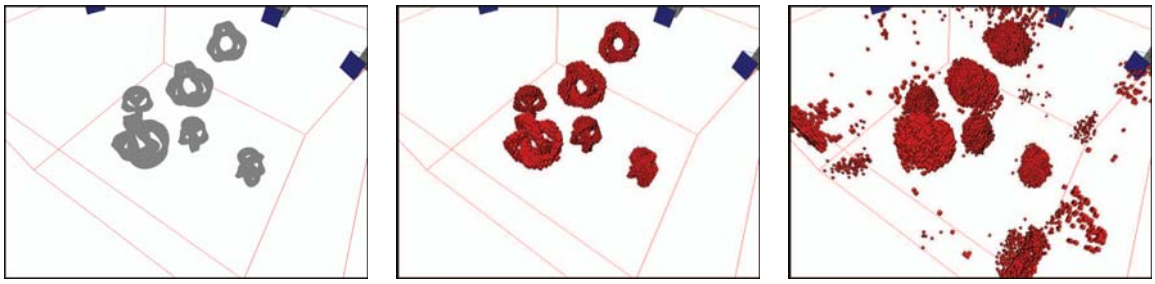


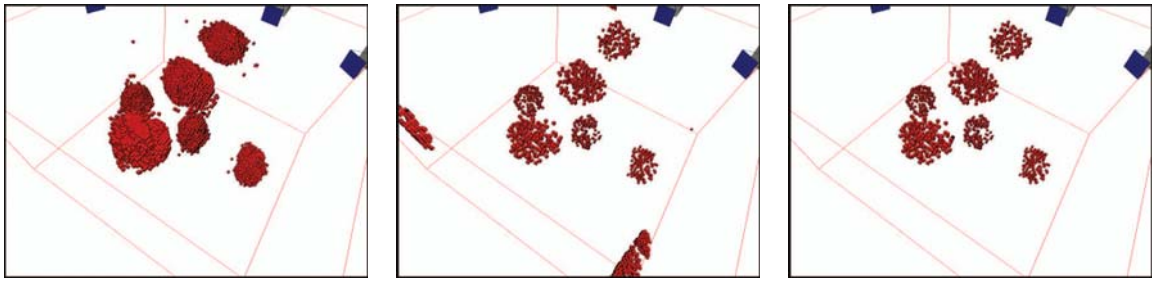
FIG. 8.17 – Précision des approches $\widetilde{\text{EVE}}_2$ et de EVE_2 en fonction du paramètre de bruit σ. Lorsque le nombre de caméras augmente, la précision de $\widetilde{\text{EVE}}_2$ et de EVE_2 croissent. Cependant nous observons que la précision $\widetilde{\text{EVE}}_2$ se révèle plus faible que celle associée à EVE_2. Cela provient du fait que $\widetilde{\text{EVE}}_m$ est construite afin de maximiser le rappel. Ceci a pour conséquence d'augmenter la quantité d'objets fantômes, ce qui réduit la précision.

en déduisons que SFS ne peut fournir un rappel plus important que EVE_2 et ainsi que $\widetilde{\text{EVE}}_2$.

La figure 8.17 présente la comparaison de la précision des approches $\widetilde{\text{EVE}}_2$ et EVE_2. A l'opposé du critère de rappel, la précision s'avère plus intéressante pour l'approche EVE_2. Cette observation provient du fait que l'approche $\widetilde{\text{EVE}}_2$ a été construite afin de maximiser la complétude de la reconstruction. L'une des conséquence est la création d'un nombre important d'objets fantômes de petite taille, ce qui influe sur le critère de précision. Nous verrons dans la section suivante que le fait d'associer $\widetilde{\text{EVE}}_2$ à EOR_{seq} permet de supprimer cet effet indésirable. Dans notre contexte pour lequel nous privilégions le critère de rappel, $\widetilde{\text{EVE}}_2$ s'avère plus performante que EVE_2 et SFS.



(a) Objets d'intérêt (b) Objets d'intérêt voxélisés (c) $\widetilde{\text{EVE}}_2$



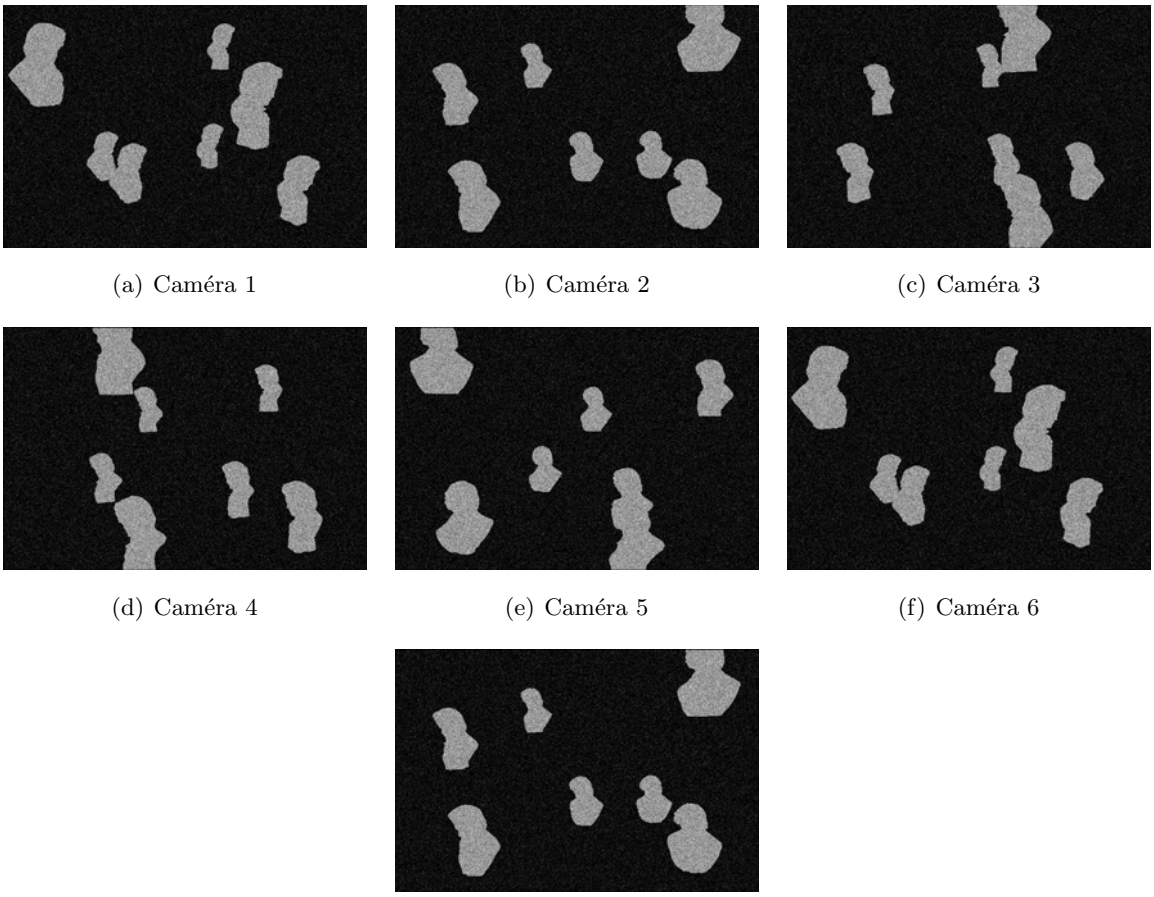
(d) $\widetilde{SFS}$ (e) EVE_2 (f) SFS

FIG. 8.18 – Scène d'expérimentation des approches $\widetilde{\text{EVE}}_2$ et $\widetilde{\text{EVE}}_n = \widetilde{SFS}$, composée de 6 objets d'intérêt, filmés par 8 caméras. Les cartes binaires de silhouettes sont bruitées de façon synthétique (voir figure 8.15). Les reconstructions $\widetilde{\text{EVE}}_2$ et $\widetilde{SFS}$ contiennent 95% des points des objets d'intérêt, alors que les approches EVE_2 et SFS ne contiennent qu'une faible partie de ces mêmes points. Les formes reconstruites par $\widetilde{\text{EVE}}_2$ et $\widetilde{SFS}$ contienne des points faux-positifs provenant du bruit dans l'information de silhouette.

La figure 8.18 représente les reconstructions proposées par $\widetilde{\text{EVE}}_2$, $\widetilde{\text{EVE}}_n$, EVE_2 et de SFS avec $n = 8$ et $\sigma = 0,150$. En présence des mêmes conditions de bruits, l'approche $\widetilde{\text{EVE}}_m$ produit des résultats supérieurs à ceux générés par l'approche EVE_m. Plus le nombre de caméras augmente, plus l'approche $\widetilde{\text{EVE}}_m$ est robuste aux bruits dans les silhouettes. Les approches EVE_m et SFS possèdent la propriété opposée ; plus le nombre de caméras augmente, moins la reconstruction contient de points de $\mathcal{O}$. Avec l'approche $\widetilde{\text{EVE}}_m$ nous proposons une méthode robuste au bruit dans les silhouettes, fonctionnant en temps réel.

La méthode $\widetilde{\text{EVE}}_m$ s'avère plus robuste que EVE_m. Cependant elle a tend à ajouter des points isolés faux-positifs. Ainsi le fait de rendre l'approche EVE_m robuste crée à nouveau des objets fantômes. Dans la suite nous présentons le pipeline de nos contributions, comparé à l'approche classique SFS.

8.5 Association des différentes contributions



(a) Caméra 1 (b) Caméra 2 (c) Caméra 3

(d) Caméra 4 (e) Caméra 5 (f) Caméra 6

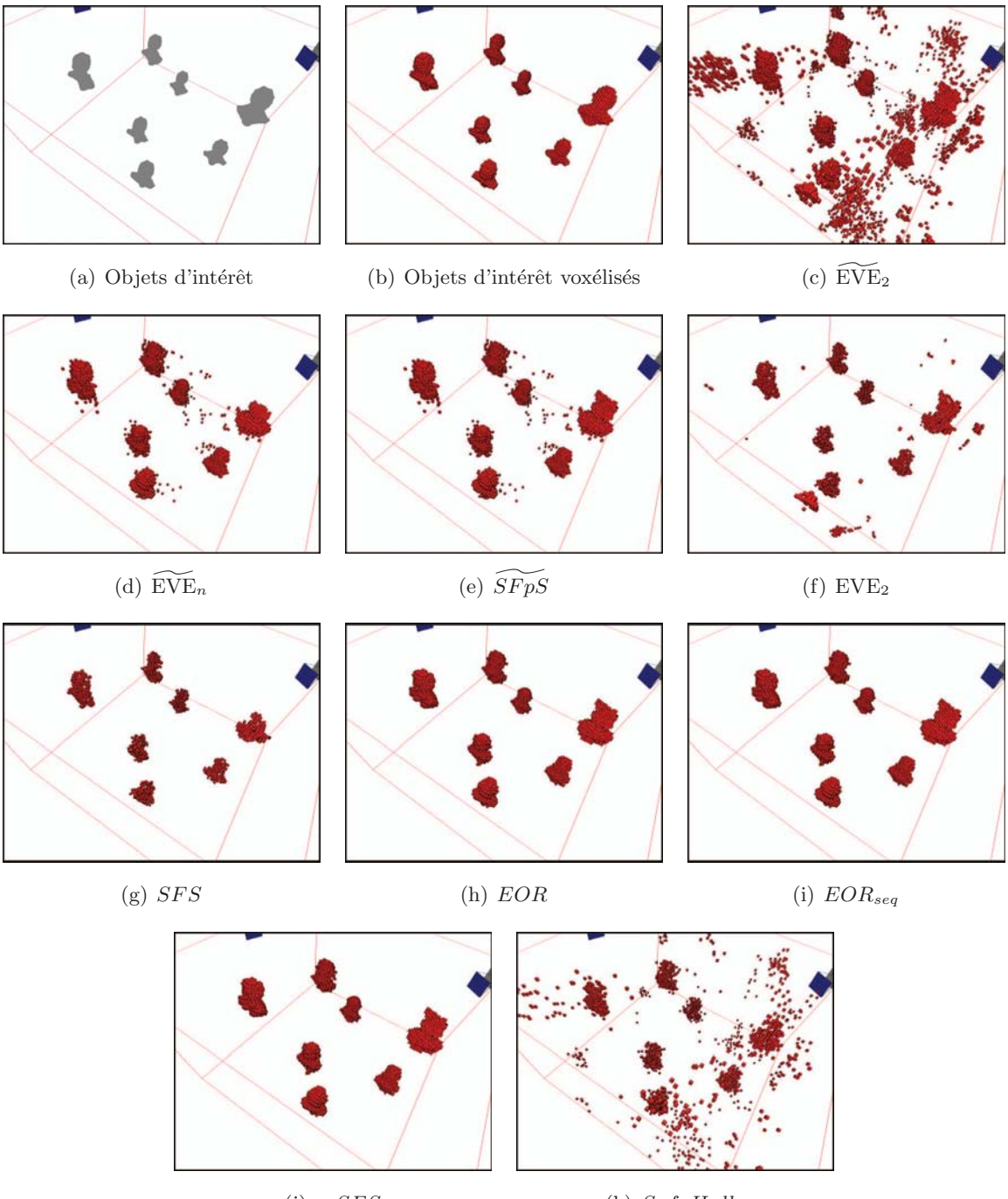
(g) Caméra 7

FIG. 8.19 – Cartes de probabilité de silhouette issues de 7 caméras qui observent 7 bustes de Ludwig van Beethoven présentés figure 8.19. Le bruit synthétique additif suit une loi normale $\mathcal{N}(0; 0,15^2)$. Notons que les objets sont partiellement visibles depuis 5 des caméras.

Dans la première partie de ce document, nous avons proposé plusieurs solutions dans le but de rendre les approches *Shape-From-Silhouette* plus robustes aux silhouettes bruitée, de ne pas générer d'objet fantômes, ou encore de relaxer les contraintes de visibilité totale. Les expérimentations menées dans les sections précédentes ont souligné séparément les apports de chaque méthode développée. Nous proposons dans la suite une expérimentation qui confronte simultanément les reconstructions issues de chacune de ces approches, à partir d'une scène critique (voir figure 8.20) : Les silhouettes sont bruitées artificiellement (voir figure 8.19), la contrainte de visibilité totale n'est pas respectée (voir figure 8.20) et enfin cette scène implique la génération d'objets fantômes. Dans la suite, l'ensemble des méthodes de suppression d'objets fantômes sont appliquées à la reconstruction offerte par $\widetilde{\text{EVE}}_2$.

La figure 8.20 décrit la scène d'expérimentation, ainsi que les reconstructions issues des différentes approches à partir de 7 caméras. Même en présence de ce nombre important de caméras seules les reconstructions offertes par EOR, EOR_{seq} et $pcSFS$ contiennent plus de 95% des points des objets d'intérêt, sans contenir d'objet fantôme. Le tableau 8.21 présente les différentes valeurs des critères d'évaluation pour cette scène, en fonction du nombre de caméras. Notons que quel que soit le nombre n de caméras, EOR supprime la totalité des objets fantômes, mais aussi parfois certaines composantes non fantômes. $SFpS$ contient le même nombre d'objets fantômes que $\widetilde{\mathrm{EVE}}_n$ et ne garanti donc pas de tous les supprimer de $\widetilde{\mathrm{EVE}}_2$. L'approche $pcSFS$ ne propose pas toujours un résultat correct. En effet l'étape de mise en correspondance des composantes connexes de silhouettes tombe en échec, car tous les objets ont une apparence similaire. Enfin l'approche $SafeHulls$ s'avère instable. Cela provient du fait que cette approche est basée sur une raisonnement 1D pour chaque pixel. Ainsi elle ne profite pas de la robustesse offerte par l'approche multi-caméras. L'association de $\widetilde{\mathrm{EVE}}_2$ filtrée par EOR_{seq} fournit la meilleure reconstruction. Pour la majorité des cas EOR_{seq} supprime exactement tous les objets fantômes.

Cette dernière configuration confirme le fait que les reconstructions les plus performantes sont réalisées par le couple $\widetilde{\mathrm{EVE}}_2$ filtrée par EOR_{seq}.



(a) Objets d'intérêt (b) Objets d'intérêt voxélisés (c) $\widetilde{\text{EVE}}_2$

(d) $\widetilde{\text{EVE}}_n$ (e) $\widetilde{SFpS}$ (f) EVE_2

(g) SFS (h) EOR (i) EOR_{seq}

(j) $pcSFS$ (k) $SafeHulls$

FIG. 8.20 – Reconstruction des objets d'intérêt (a) à partir des 7 cartes probabilistes de silhouettes altérées par un bruit additif suivant la loi normale $\mathcal{N}(0; 0.15^2)$ (voir figure 8.19). Les approches $\widetilde{\text{EVE}}_2$ et $\widetilde{SFpS}$ reconstruisent la totalité des points des objets d'intérêt $\mathcal{O}$, mais génèrent de petites entités fantômes. Pour cette configuration les approches EVE_2 et SFS ne contiennent qu'une fraction de $\mathcal{O}$. A partir de la reconstruction offerte par $\widetilde{\text{EVE}}_2$ la méthode $SafeHulls$ est incapable de supprimer tous les objets fantômes. Cependant EOR, EOR_{seq} se montrent particulièrement efficaces et suppriment exactement les objets fantômes.

n	$\widetilde{\text{EVE}_2}$	$\widetilde{SFS}$	$SFpS$	EVE_2	SFS	EOR	EOR_{seq}	$pcSFS$	$SafeHulls$	
2	0.799	0.799	0.799	0.525	0.525	0.229	0.799	0.799	0.024	R
	> 255	> 255	>255	71	71	0	1	15	205	F
	0.213	0.213	0.213	0.260	0.260	0.495	0.334	0.262	0.171	P
3	0.851	0.834	0.851	0.393	0.382	0.601	0.851	0.851	0.077	R
	> 255	> 255	> 255	63	22	0	0	7	172	F
	0.316	0.426	0.352	0.387	0.561	0.440	0.463	0.421	0.261	P
4	0.882	0.862	0.882	0.296	0.284	0.882	0.882	0.882	0.132	R
	> 255	220	220	55	18	0	0	8	191	F
	0.498	0.627	0.624	0.674	0.740	0.696	0.696	0.683	0.401	P
5	0.910	0.874	0.911	0.221	0.205	0.910	0.910	0.910	0.163	R
	> 255	89	89	61	15	0	0	1	201	F
	0.585	0.724	0.719	0.722	0.807	0.750	0.750	0.748	0.521	P
6	0.938	0.887	0.938	0.174	0.154	0.938	0.938	0.938	0.189	R
	> 255	44	44	45	21	0	0	0	205	F
	0.573	0.765	0.760	0.529	0.842	0.776	0.776	0.776	0.589	P
7	0.951	0.898	0.951	0.130	0.114	0.951	0.951	0.951	0.217	R
	>255	32	32	52	26	0	0	1	218	F
	0.529	0.775	0.769	0.310	0.838	0.783	0.783	0.782	0.638	P
8	0.962	0.908	0.962	0.096	0.086	0.962	0.962	0.962	0.247	R
	>255	29	29	33	26	0	0	1	221	F
	0.695	0.782	0.777	0.793	0.860	0.787	0.787	0.786	0.686	P

FIG. 8.21 – Valeurs des critères d'évaluation pour la scène représentée figure 8.20. Pour $n = 2$ les objets d'intérêt sont totalement vus dans toutes les caméras. Ainsi $\widetilde{\text{EVE}_2}$ et $\widetilde{\text{EVE}_n}$ fournissent la même reconstruction. Pour toute valeur de n, seule l'approche EOR supprime tous les objets fantômes. EOR_{seq} fournit les meilleurs résultats, car elle contient le maximum de points de $\mathcal{O}$ et supprime en général exactement les objets fantômes.

8.6 Discussion

Les expérimentations menées dans ce chapitre montrent que nos contributions répondent aux principales limitations des approches d'estimation de forme à partir de silhouettes.

Par rapport aux critères de qualité fixés, le couple de méthodes $\widetilde{\text{EVE}}_2$ filtrée par EOR_{seq} offre les meilleurs résultats de reconstruction pour toutes les expérimentations menées. Ce couple de méthode est résistant aux bruits présents dans les silhouettes. Cette association permet de relaxer les contraintes de placement des caméras : tout objet d'intérêt observé par au moins 2 caméras est contenu dans la reconstruction. Enfin cette approche se révèle très efficace pour supprimer les objets fantômes. Dans la majorité des configurations EOR_{seq} supprime exactement tous les objets fantômes. Dans le cas contraire la reconstruction contient tous les points des objets d'intérêt.

L'approche EOR fournit elle aussi des résultats intéressants et supprime toujours tous les objets fantômes. Cependant il arrive que cette méthode ne conserve pas tous les points appartenant aux objets d'intérêt. Cette dernière caractéristique peut s'avérer pénalisante lorsque tout objet d'intérêt doit être reconstruit.

L'approche $SFpS$ permet de supprimer les objets fantômes de $\widetilde{\text{EVE}}_2$ qui n'appartiennent pas à $\widetilde{\text{EVE}}_n = \widetilde{SFS}$. Cette méthode souffre de deux limitations : si $\widetilde{SFS}$ contient des objets fantômes, alors $SFpS$ les contient aussi. De plus, si un objet d'intérêt n'est pas partiellement reconstruit par $\widetilde{SFS}$, il ne peut appartenir à $SFpS$.

La méthode $pcSFS$ fournit une approche intéressante. Cependant elle est trop dépendante de la mise en correspondance des composantes connexes de silhouettes. Ainsi, même en présence de données synthétique, $pcSFS$ a été souvent incapable de supprimer la totalité des objets fantômes.

Enfin en comparant les approches EOR_{seq} et EOR à la méthode $SafeHulls$ proposée par Miller et Hilton, il ressort que nos contributions offrent de bien meilleurs résultats. En effet $SafeHulls$ est construite sur l'étude de composantes connexes 1D pour chaque caméra. Ainsi cette approche ne profite par du schéma multi-caméra et se révèle particulièrement instable en présence de bruit. Avec un faible nombre de points de vue, $SafeHulls$ supprime souvent des points qui appartiennent aux objets d'intérêt.

En conclusion, l'association des méthodes $\widetilde{\text{EVE}}_m$ et EOR_{seq}, répond aux principales limitations des approches *Shape-From-Silhouette*. Leur efficacité s'illustre particulièrement lorsque le nombre de caméra est faible (de l'ordre de 2 ou 3).

La majorité des applications de *Shape-From-Silhouette*, comme par exemple l'acquisition du mouvement de plusieurs personnes, ou encore la surveillance, peuvent désormais être réalisées à partir d'un faible nombre de caméras, même en environnement peu contrôlé.

Chapitre 9

Conclusion

L'approche *Shape-From-Silhouette* estime la forme 3D d'objets à partir de leurs silhouettes issues de caméras calibrées. Les silhouettes sont simples à estimer en environnement contrôlé et l'implémentation de *Shape-From-Silhouette* est relativement aisée. Cependant cette approche souffre de limitations fortes : le placement des caméras assez contraint, la génération d'objets fantômes et la sensibilité à la qualité d'extraction des silhouettes. A travers cette première partie, nous avons présenté des contributions qui apportent des réponses.

- Toute partie d'objet est reconstruite que si elle appartient à l'intersection des pyramides de vue de toutes les caméras. Ceci contraint le placement de ces dernières. Pour palier cette limitation, nous avons proposé une extension de l'enveloppe visuelle que nous avons appelé *enveloppe visuelle étendue* (EVE_m), qui intègre l'espace de visibilité de chaque caméra. La décision de reconstruction d'un point 3D se base sur la silhouette-consistance des caméras qui voient ce point, et non sur l'ensemble des caméras. La contrepartie de cette relaxation est la génération de plus d'objets fantômes. Nous avons exploré plusieurs voies en vue de limiter leurs ajouts.
- *Shape-From-Silhouette* construit des objets fantômes, composantes connexes 3D qui ne contiennent aucun objet réel. Nous avons proposé des solutions afin de supprimer ces objets. En introduisant une condition suffisante de l'existence d'un objet réel dans une composante connexe reconstruite, nous avons proposé la méthode EOR (*Enveloppe d'objets réels*), qui garantit la suppression de tout objet fantôme. Cependant elle n'assure pas de conserver tout objet réel. Une extension, *Enveloppe d'objets réels silhouette-équivalente* (EOR_{seq}), limite la suppression d'objets réels.
- *Shape-From-Silhouette* n'est pas robuste à une extraction erronée de silhouettes. En effet lorsqu'un point de silhouette est détecté faux-négatif, l'ensemble des points 3D des objets d'intérêt qui se projettent sur ce point ne sont pas reconstruits. Notre solution, $\widetilde{\text{EVE}}_m$ décide de l'appartenance d'un point 3D à la reconstruction 3D, en fusionnant l'information issue de cartes valuées de silhouettes de chaque caméra. L'information de silhouette issue

d'une caméra peut être remise en cause par l'information issue des autres caméras. Les expérimentations menées ont démontré que cette contribution apporte une alternative à la méthode proposée par Franco et Boyer [FB05], en offrant des résultats similaires, avec une accélération des calculs (de l'ordre de 1000 par rapport à [FB05]).

A travers ces trois axes de contributions, nous proposons une méthode d'estimation de forme 3D temps réel, robuste à une extraction de silhouettes bruitées, qui ne construit, en général, pas d'objets fantômes et qui n'impose pas de contrainte forte sur le placement des caméras.

L'étude menée chapitre 8, montre que l'association des approches $\widetilde{\text{EVE}}_m$ et du filtrage EOR_{seq}, offre les meilleures reconstructions. Quel que soit le nombre de caméras et quelle que soit la quantité de bruit dans les images, les résultats obtenus s'avèrent supérieurs à ceux proposés par les approches classiques de SFS, l'approche $SafeHulls$ [MH07] ou encore l'approche $ReducedVisualHull$ [BG08].

Les contributions apportées sont orientées vers des implémentations volumiques. La méthode $\widetilde{\text{EVE}}_m$ n'est pas directement transposable aux approches polyédriques. Cependant les formalisations de EVE_m et de EOR_{seq} sont adaptées à une implémentation surfacique.

L'association de $\widetilde{\text{EVE}}_m$ et de EOR_{seq} fournit une alternative intéressante à l'utilisation de *Shape-From-Silhouette*, offrant une reconstruction plus robuste, plus précise, moins contrainte, mais toujours en temps réel.

Deuxième partie

Acquisition de mouvements en temps réel à partir de formes 3D

Introduction

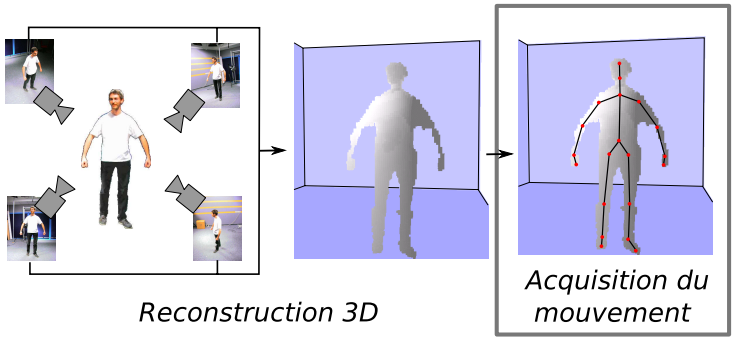


FIG. 1 – Retour sur le schéma du processus de capture de mouvements adopté : étape d'acquisition de mouvements.

L'objectif principal de cette thèse est de proposer des méthodes qui permettent l'acquisition du mouvement de personnes sans marqueur et en temps réel. Nous avons décrit dans la première partie de cette thèse, nos contributions réalisant une reconstruction 3D des personnes filmées à partir de plusieurs caméras. Le second volet de ces travaux porte sur l'analyse de cette forme, afin de localiser à chaque instant les positions tridimensionnelles des articulations de ces personnes (voir figure 1).

La plupart des méthodes de la littérature s'appuient sur une modélisation paramétrée du corps humain. Les paramètres décrivent, entre autres, soit les positions, les orientations ou encore les angles des articulations. La capture du mouvement revient à estimer les paramètres qui réalisent la meilleure correspondance entre les vues réelles et des vues de synthèses, obtenues par instantiation du modèle. Les temps calculs sont généralement élevés, voire prohibitifs lorsque l'on cherche la précision [GD96, SC05, ST05b]. Ceci limite l'application de ce type d'approches aux systèmes hors ligne.

A notre connaissance, peu de méthodes sans marqueurs fonctionnent en temps réel [KLPL02, CGH08, OS08]. L'acquisition du mouvement de corps articulés sans marqueur est un problème difficile, dont les représentations courantes impliquent la résolution d'un problème de grande dimension. A cette difficulté s'associe généralement une relation non-linéaire liant les paramètres du modèle aux observés dans l'espace image. Acquérir la posture revient à déterminer la réci-

proque de cette relation, permettant d'estimer les paramètres du modèle en fonction des observés. Celle-ci n'est en générale pas bijective, ce qui implique la mise en place de mécanismes complexes.

Nous cherchons à estimer les positions 3D des articulations de personnes à partir d'une reconstruction 3D de leur forme. Dans ce contexte la relation liant les paramètres du modèle (positions des articulations) et observés (la forme 3D) est linéaire.

Dans cette seconde partie, nous présentons plusieurs réponses au problème de la capture de mouvements, principalement orientées sur l'analyse de la forme 3D. Le travail présenté dans la suite décrit la mise en œuvre de plusieurs méthodes construites sur l'extraction de primitives pertinentes et robustes. Ces primitives permettent d'accélérer le processus de mise en correspondance entre modèle et observés. Notre objectif est la capture en temps réel du mouvement. Ainsi nous privilégions des approches rapides, amenant parfois une réduction de la précision par rapport aux systèmes à base de marqueurs.

Nos premières contributions sont construites sur une extraction de la topologie de la reconstruction 3D. Celle-ci fournit une représentation condensée de cette information 3D sous la forme d'une structure 1D. Nous verrons que dans le cadre de mouvements usuels, nos méthodes se révèlent particulièrement efficaces et respectent le temps réel. Néanmoins elles pré-supposent que la topologie de la forme 3D reflète la topologie de notre modèle. Pour cette raison, elles se montrent sensibles aux auto-contacts et à la présence d'artefacts dans la reconstruction.

Notre contribution suivante répond à ce problème par la détection des parties de peau, qui informent sur les positions des extrémités hautes du corps (tête,mains). Ces extrémités permettent de conduire un recalage d'objets géométriques simples, qui estime les différentes postures au cours du temps. Nous montrerons que cette approche construite sur un concept simple offre, en temps réel, la capture totalement automatique du mouvement d'un personnage, robuste aux auto-occultations.

Nous proposons ensuite une contribution originale qui utilise conjointement l'extraction des parties saillantes de la forme et la détection des parties de peau. La complémentarité de ces deux modalités offre une extraction robuste des extrémités du corps, ensuite utilisées pour segmenter chaque partie rigide du corps dans la forme 3D. Cette dernière approche offrant une précision et une robustesse intéressantes, nous l'étendons pour réaliser l'acquisition du mouvement de plusieurs personnes simultanément. Il en résulte une approche qui fournit la capture robuste du mouvement d'une personne en temps réel et de plusieurs personnages en temps interactif.

Nous terminons cette seconde partie en effectuant une analyse comparatives de nos méthodes, réalisée à partir de données terrain.

Dans chacun des chapitres suivants, nous présentons les notions et travaux de référence nécessaires à la présentation et à l'évaluation de nos contributions.

Chapitre 10

Acquisition du mouvement par recalage de structures 1D

Dans ce chapitre nous présentons plusieurs approches qui réalisent l'extraction de la pose d'un personnage à partir d'une estimation volumique de sa forme 3D. Nous choisissons de représenter la posture de ce personnage par une structure articulaire constituée de segments rigides connectés par des rotules. Cette représentation, généralement appelée squelette d'animation, définit une structure 1D. La majorité des méthodes à base de modèle articulaires recalent ce squelette d'animation habillé par un ensemble de primitives géométriques simples (cylindres, quadriques, etc) avec le volume 3D. Ces approches nécessitent une initialisation complexe des paramètres du modèle : les longueurs des membres, les paramètres de la primitive associée à chaque membre, etc. Il est possible d'utiliser une approche alternative : recaler le squelette d'animation sur squelette topologique de cette forme 3D. Avec cette approche, l'initialisation des paramètres du modèle se résume à la longueur des membre.

10.1 Squelettes topologiques

Le squelette topologique d'une forme 3D est une représentation "compacte" de celle-ci. Ce squelette présente ses caractéristiques topologiques principales : trous, cycles, nombre de branches, composantes connexes, etc. Parmi les méthodes publiées qui extraient le squelette topologique d'un objet représenté dans un espace discret, nous identifions plusieurs classes d'approches.

10.1.1 Approches par amincissement topologique

Les méthodes d'amincissement topologique calculent le squelette topologique d'un ensemble de voxels, en supprimant itérativement les voxels de la surface jusqu'à ce que l'amincissement souhaité soit obtenu. L'ensemble des algorithmes d'amincissement topologique opèrent dans un

espace discret (voxels) et s'appuient sur le concept de *point simple*, introduit pas Morgenthaler en 1981 [Mor81]. Un point simple (voir [KR89]) pour un état de l'art complet sur la topologie discrète) est un point de l'objet 3D qui peut être supprimé sans modifier la topologie initiale de l'objet. L'une des propriétés importantes des points simples est qu'ils peuvent être caractérisés localement, c'est à dire qu'un voxel peut être validé *point simple* seulement en inspectant son voisinage. Cela a pour conséquence de proposer des algorithmes efficaces.

Le processus d'amincissement commence par les voxels de surface et continue jusqu'à ce que tout voxel restant ne soit pas un point simple. A chaque itération, chaque voxel de surface est testé par rapport à certaines caractéristiques de conservation de topologie, généralement représentées par un masque volumique. Cependant supprimer tous les points simples peut produire un raccourcissement trop fort des branches du squelette topologique. Certaines approches proposent d'autres mécanismes qui permettent de conserver les propriétés géométriques de l'objet.

10.1.2 Approches par cartes de distance

La transformation en distances, ou cartes de distances, associe à chaque point P d'un objet, la distance la plus faible au contour de cet objet. Plusieurs types de distance peuvent être utilisées, comme par exemple la distance euclidienne, ou encore une approximation telle que la métrique de chanfrein <3,4,5> [Bor96]. Les crêtes décrites dans les cartes de distances 3D, sont composées des voxels qui sont localement centrés dans l'objet. Les méthodes de la littérature varient sur la façon de déterminer cet ensemble de voxels. Les approches par amincissements ordonnés sur la distance [CCZ07], utilisent une procédure d'amincissement topologique construite sur une fonction de priorité. Celle-ci ordonne le processus de suppression de points simples. D'autres méthodes utilisent l'information du gradient afin de déterminer les voxels de crête [BKS01]. Enfin certains travaux sont construits sur la propagation de fronts géodésiques [PFP04]. L'ensemble de ces approches génèrent souvent un ensemble de voxels candidats très important. Cet ensemble peut être réduit en utilisant certaines approches d'amincissement topologique.

Les approches par carte de distance permettent d'estimer finement la surface médiane. Elle ne peuvent fournir le squelette topologique de toute forme, sans l'utilisation de techniques supplémentaires de squelettisation de surface médiane, dépendantes de la nature d'objet à traiter.

10.1.3 Approches géométriques

Les méthodes géométriques s'appliquent généralement à des objets représentés par des maillages polygonaux ou encore par des nuages de points non organisés. Une approche classique est d'utiliser les diagrammes de Voronoi [OK95] générés à partir des arêtes de la représentation polygonale 3D, ou encore directement à partir du nuage de points [AL01]. Les arêtes internes du diagramme de Voronoi sont utilisées afin d'extraire une approximation de la surface médiane. Un squelette

topologique peut être estimé en simplifiant cette surface médiane en une structure 1D. Les approches géométriques offrant une surface médiane sont généralement très sensibles au bruit. De plus, elles ne peuvent fonctionner directement sur un objet voxélisé. Il est alors nécessaire de trianguler la surface de cet ensemble de voxels par exemple par l'approche *Marching Cube* [KKI99]. Cependant il existe certaines autres approches géométriques qui estiment directement une structure 1D.

Graphes de Reeb

Les approches topologiques basées sur les graphes de Reeb [Ree46] présentent l'avantage de préserver les propriétés topologiques du maillage [BMMP03], tout en offrant une structure 1D qui encode la topologie de cet objet. Ce graphe est basé sur la théorie de Morse [For98, Gra71] et est utilisé dans diverses applications de description de courbes et de surfaces. Celui-ci peut être formalisé de la façon suivante :

Définition 19 *Soit f une fonction réelle définie sur une variété compacte M, $f : M \to \mathbb{R}$. Le graphe de Reeb de f est l'espace quotient de f dans $M \times \mathbb{R}$, par la relation d'équivalence $(p_1, f(p_1)) \sim (p_2, f(p_2))$ vérifiée si et seulement si $f(p_1) = f(p_2)$, et p_1 et p_2 appartiennent à la même composante connexe de $f^{-1}(f(p_1))$ ou $f^{-1}(f(p_2))$.* □

Un graphe de Reeb est composé de nœuds représentant les points critiques de f, autrement dit les points de la variété M où les dérivées partielles de f s'annulent. Les arêtes du graphe représentent les composantes connexes de M reliant les points critiques de f. La figure 10.1 présente un graphe de Reeb calculé sur un bi-tore, selon la fonction *hauteur*. Les points critiques de cette fonction ont été marqués en rouge pour les minima, en vert pour les maxima et en noir pour les points selles.

En accord avec la théorie de Morse, la topologie d'un objet change uniquement lors de la connexion de points critiques de f, ainsi les arêtes du graphe représentent les composantes de l'objet, alors que les nœuds représentent les connexions entre les composantes de l'objet. Le graphe de Reeb n'est pas nécessairement inclus dans l'objet, car il n'est pas définit dans l'espace de l'objet. Cependant il est possible d'incorporer l'information géométrique en associant les arêtes aux centres des composantes connexes successives selon la fonction f. Cette approche définit ainsi un squelette topologique qui respecte la topologie de l'objet, ainsi que sa géométrie. Des extensions de graphe de Reeb ont été proposées pour des maillage polyédriques [TVD07a] et des structures discrètes régulières [XSW04] ou irrégulières [VCT06]. Ces approches varient principalement sur la fonction f utilisée. Certaines approches utilisent une fonction de hauteur [BS03], qui a une sensibilité à l'orientation de l'objet, d'autres utilisent la distance euclidienne par rapport à un point dans l'espace (généralement le barycentre de l'objet) [BMMP03] qui est invariante par rotation. Enfin d'autres approches utilisent les distances géodésiques plutôt que les distances euclidiennes,

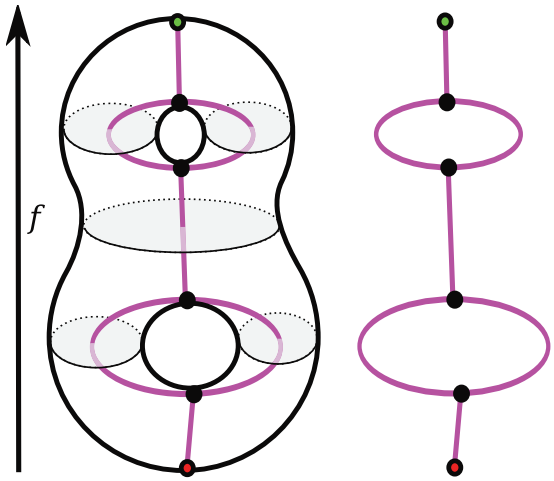


FIG. 10.1 – Graphe de Reeb d'un bi-tore, construit sur une fonction f de hauteur.

résistantes aux variations de pose de l'objet, mais consommatrices de temps de calculs [TVD07a].

D'autres approches géométriques fournissent directement le squelette topologique d'un objet maillé. Li *et al.* [LWTH01] construisent un squelette filaire en fusionnant les arêtes du maillage de l'objet dans un ordre dépendant de leur longueur. Cependant cette approche est sensible à la tesselation. Dans [Lie03] le squelette topologique est estimé par une décomposition du maillage d'origine.

Parmi les classes de méthodes évoquées ci-dessus, les approches géométriques et principalement les approches par graphe de Reeb, fournissent des méthodes d'extraction de squelette topologique parmi plus les rapides [CSM07]. De plus, les graphes de Reeb fournissent directement une structure 1D qui représente les principales caractéristiques topologiques de la forme 3D. Pour ces raisons, nous proposons deux approches d'acquisition de mouvements temps réel à partir des représentations compactes proposées par les graphes de Reeb, pour des objets représentés par des ensembles de voxels.

10.2 Acquisition de mouvements par Graphe de Reeb construit sur fonction de hauteur

Dans cette section nous présentons une approche simple qui permet à la fois de réaliser l'initialisation de façon automatique, ainsi que le suivi de mouvements d'une personne dont nous avons une estimation volumique 3D (notée V dans la suite). Nous commençons par présenter la méthode de calcul du graphe de Reeb, construit sur une fonction de hauteur, d'un objet voxélique. Nous présentons ensuite la méthode de recalage du squelette d'animation, dans le graphe de Reeb filtré, à l'aide de contraintes morphologiques simples. Nous montrons que malgré

sa construction simple, cette approche permet l'acquisition du mouvement d'une personne en temps réel.

10.2.1 Calcul du graphe de Reeb

Parmi les fonctions f utilisées pour le calcul de graphe de Reeb, les fonctions de *hauteur* ont été particulièrement utilisées de par leur simplicité de calcul. Ce type de fonction revient en général à associer à chaque point de l'objet, la valeur de son abscisse pour l'un des axes du repère lié à l'objet. Dans notre cas nous disposons d'un ensemble de voxels (notés $v_k \in V$) défini dans une grille régulière. Si le repère local de cette grille est aligné sur le repère global de la scène, alors l'altitude de chaque voxel par rapport au sol définit une fonction de hauteur sur l'axe vertical de la scène. Dans la suite nous choisissons de définir $f(v_k)$ par l'altitude de v_k. L'ensemble des voxels d'une même couche horizontale possèdent la même valeur de hauteur (voir figure 10.2(a)). Nous notons dans la suite V_j l'ensemble des voxels v_k d'une même couche j.

Nous calculons pour chaque couche j de la grille de voxels, les composantes connexes 2D notées dans la suite cc_i, en commençant par la couche de plus faible altitude (V_0). Ensuite nous définissons un graphe orienté $G(S, A)$ dont les sommets S représentent les barycentres de chaque composante connexe 2D. Les sommets $(s_l, s_m) \in S^2$ associés respectivement aux composantes connexes cc_l et cc_m, sont connectés par un arc si et seulement si les conditions suivantes sont validées :

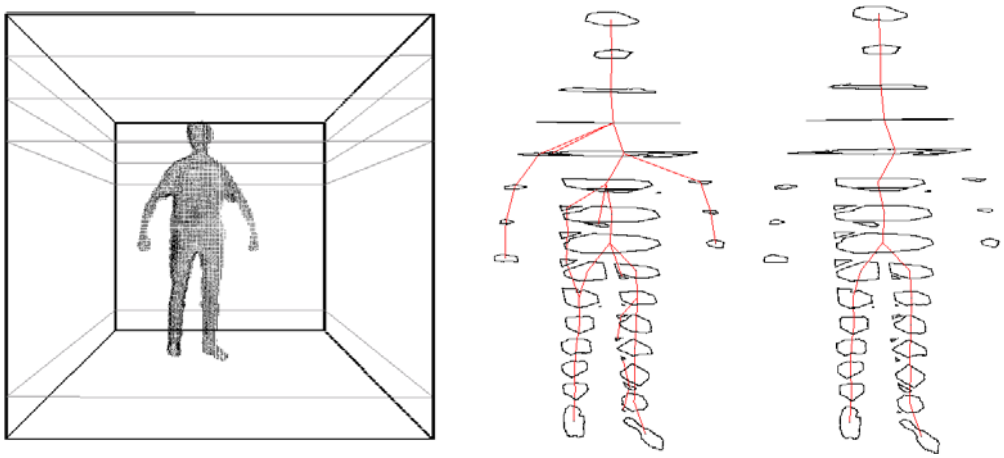
- $\text{cc}_l \subset V_i$ et $\text{cc}_m \subset V_{i+1}$, c'est à dire cc_m est sur la couche successive de cc_l dans le sens du parcours des couches ;

- $\exists v_n \in \text{cc}_l, \exists v_o \in \text{cc}_m : d(v_n, v_o) = 1$ avec d la distance d_6 ou d_{26}.

Ainsi les sommets de G sources[2], puits[3] et les sommets dont le degré entrant ou sortant est supérieur à 1, définissent respectivement les minima, maxima et selle, soit les points critiques de f. L'ensemble de ces sommets définissent les sommets du graphe de Reeb de V (voir figure 10.2(b)). Les arcs entre les sommets du graphe de Reeb, sont décrits par les chemins de G passant uniquement par des sommets dont le degré sortant et entrant est de 1.

Comme nous pouvons l'observer dans la figure 10.2(b), le graphe G de V ne fournit pas nécessairement la topologie du personnage filmé. Cela provient principalement des artefacts dus à la reconstruction du volume 3D par les approches *Shape-From-Silhouette*. De plus, en fonction de la résolution de la grille de voxels, le graphe G extrait peut être différent. Afin de réduire G aux sommets représentatifs, nous proposons de filtrer ce graphe afin de ne conserver que les

[2]sommet dont le degré entrant est de 0
[3]sommet dont le degré sortant est de 0



(a) (b) (c)

FIG. 10.2 – Représentation de couches horizontale de l'estimation volumique V d'un personnage (a). Graphe de Reeb de V calculé sur une fonction de hauteur (b). Représentation d'α-chemins de G avec $\alpha = 0,3$ (c).

arcs et sommets significatifs. Pour cela nous introduisons le concept d'α-représentativité d'un sommet de G

Définition 20 *Un sommet $s_i \in S$ est un sommet α-représentatif d'une couche j si et seulement si la composante connexe associée* cc_i *respecte la propriété suivante :*

$$\frac{Card(\mathrm{cc}_i)}{Card(\{v_k \in V_j\})} > \alpha \text{ où } \alpha \in [0,1] \tag{10.1}$$

□

On étend cette définition aux chemins de G :

Définition 21 *Un α-chemin est un chemin de G ne passant que par des sommets α-représentatifs.*
□

Dans la suite nous noterons G_α le sous-graphe de G contenant uniquement les α-chemins de G. Nous étendons cette notation à $G_{[\alpha_1;\alpha_2]}$ le sous-graphe de G contenant uniquement les α-chemins de Gn avec $\alpha_1 \leq \alpha \leq \alpha_2$.

La figure 10.2(c) illustre le filtrage du graphe présenté figure 10.2(b) avec $\alpha = 0,3$. Nous disposons désormais d'une représentation "compacte" de V par une structure 1D, dont la granularité peut être modulée par le paramètre α. Nous nous intéressons maintenant à réaliser l'initialisation de l'acquisition du mouvement.

10.2.2 Initialisation

Dans cette section, nous présentons une approche automatique qui estime à la longueurs des membres, ainsi que la posture initiale de la personne filmée. Les méthodes de la littérature peuvent être classées selon trois axes. Le premier concerne les approches qui nécessitent une intervention manuelle afin de définir les mesures anthropométriques et la posture [MBR06]. Les méthodes de la seconde classe nécessitent que la personne filmée prenne une pose type, comme par exemple la *T-pose* [FGDP02]. Ces méthodes fonctionnent généralement en temps réel. Enfin d'autres approches estiment automatiquement ces paramètres initiaux [MTHC03, dATM$^+$04] à partir d'une séquence vidéo analysée, au mieux, en temps interactif. Dans cette section nous proposons une approche intermédiaire qui offre en temps réel l'initialisation du suivi de mouvements, sans imposer de pose type, tant que la personne filmée est debout, bras et jambes un peu écartés.

L'anthropométrie définit les techniques qui concernent la mesure des particularités dimensionnelles d'un homme ou d'un animal. Elle est particulièrement utilisée en ergonomie. Pour l'homme elle concerne notamment :

- les dimensions : stature (la taille), la hauteur du buste, la longueur de chaque membre et chaque partie de membres (bras, avant-bras, etc) ;
- les masses : masse totale, masse de chaque partie du corps, le centre de gravité de chaque partie ;
- les circonférences : bassin, poitrine, membres, etc.

Dans notre contexte nous nous intéressons particulièrement aux données anthropométriques liées à la longueur de chaque partie rigide du corps en fonction de la stature. Dans l'homme de Vitruve, Léonard de Vinci propose l'utilisation de rapports anthropométriques afin d'illustrer certains grands principes de l'architecture. Depuis, les mesures anthropométriques [DT01] ont été utilisées à la fois dans les domaines de la santé [SYN$^+$00], pour les industries du vêtement [MC05] ou encore pour l'ergonomie [Nas78].

Soit $L_{stature}$ la stature de la personne filmée. Nous proposons les rapports anthropométriques suivants :

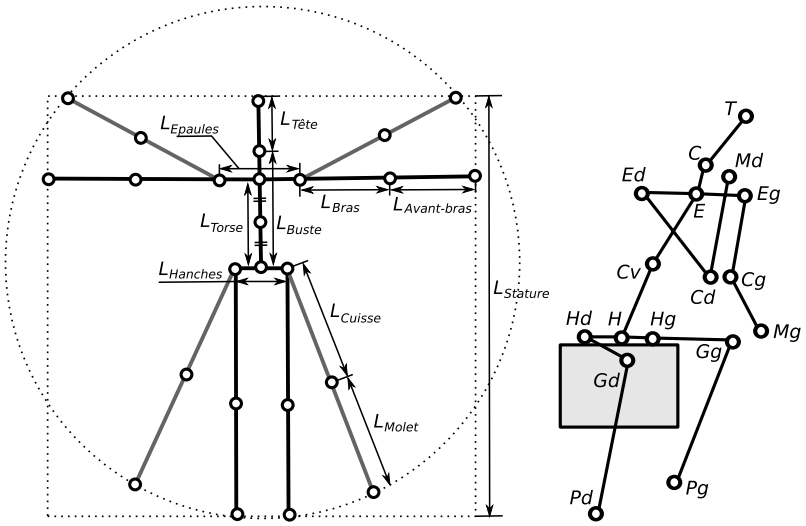


FIG. 10.3 – Représentation du squelette d'animation utilisé. L'image de gauche illustre les notations des longueurs anthropométriques. L'image de droite présente l'identification de chaque articulation.

L_{Buste}	$\approx$	$3L_{stature}/8$	la hauteur du buste
L_{Torse}	$\approx$	$5L_{stature}/16$	la longueur du torse
L_{Cuisse}	$\approx$	$L_{stature}/4$	la longueur des cuisses
L_{Mollet}	$\approx$	$L_{stature}/4$	la longueur des mollets
L_{Bras}	$\approx$	$L_{stature}/5$	la longueur des bras
$L_{AvantBras}$	$\approx$	$L_{stature}/5$	la longueur des avant-bras
$L_{Epaules}$	$\approx$	$L_{stature}/5$	la largeur des épaules
$L_{Hanches}$	$\approx$	$L_{stature}/6$	la largeur des hanches
L_{Tete}	$\approx$	$L_{stature}/8$	la hauteur de la tête

En dépit d'une caractérisation peu précise, ces rapports permettent de dimensionner le squelette d'animation choisi pour modéliser le personnage filmé, seulement en fonction de sa stature. La figure 10.3 représente le modèle articulaire utilisé.

Nous nous intéressons désormais à la procédure qui permet l'extraction de la posture. Le squelette d'animation contient 5 extrémités. Nous commençons par rechercher parmi les sommets maxima de G, ceux correspondants à ces extrémités.

Tête et pieds

Afin d'extraire les articulations de la tête T, et des pieds P_d et P_g nous commençons par filtrer le graphe G en calculant le graphe $G_{0,3}$ afin de n'extraire que les points critiques de G représentatifs (voir figure 10.4). La valeur $\alpha = 0,3$ se justifie par le fait que pour une couche donnée, la section de chaque bras est supposée représenter moins du tiers des parties du corps

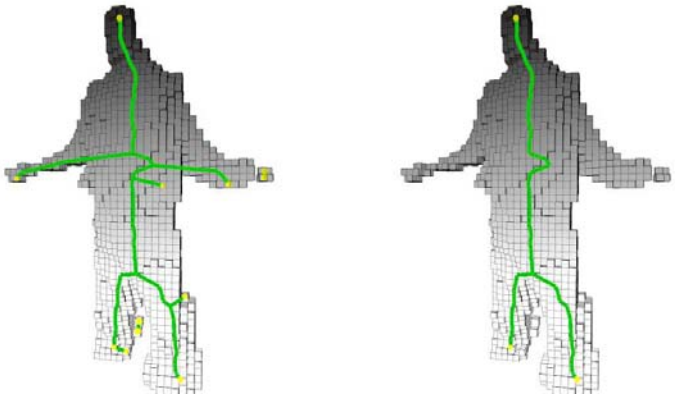


FIG. 10.4 – Représentation de G(à gauche) et de $G_{0,3}$ (à droite). $G_{0,3}$ définit les chemins permettant l'extraction des articulations de la tête et des jambes.

coupées par le plan de cette couche. Les sommets de G sont ensuite triés selon la fonction de hauteur f. Ainsi le sommet s_T maximal global de G et les sommets s_{P1} et s_{P2} de plus faible altitude définissent respectivement les positions 3D des articulations T, P_1 et P_2. Pour l'instant nous ne sommes pas capables de dissocier les articulations gauches des articulations droites. Cette identification sera réalisée dans la suite. La différence d'altitude entre T et le sol permet d'estimer la taille de l'individu, soit $L_{Stature}$.

Cou, épaules, centre de la colonne vertébrale et hanches

Disposant de rapports anthropométriques décrits précédemment, il devient possible d'estimer rapidement les positions du cou C, du centre des épaules E, le centre de la colonne vertébrale Cv ainsi que le centre des hanches H. Ayant préalablement déterminé le haut de la tête, nous parcourons le chemin qui prend source au sommet s_T, jusqu'à ce que le sommet courant s_C soit à une distance de L_{Tete} de s_T. Ainsi s_C décrit l'estimation de l'articulation C. Nous utilisons à nouveau cette approche afin de déterminer séquentiellement les sommets s_E à une distance $L_{Stature}/16$ de s_C, s_{Cv} à une distance $5L_{Stature}/32$ de s_E et enfin s_H à une distance $5L_{Stature}/32$ de s_{Cv}. Les articulations E, Cv et H sont respectivement associés aux sommets s_E, s_{Cv} et s_H.

Afin de déterminer l'axe des épaules, nous estimons l'axe principal d'inertie des voxels associés aux sommets du sous-chemin du graphe $G_{0,3}$ partant du sommet associé au cou s_C et s'arrêtant au sommet associé au centre de la colonne vertébrale s_{Cv}. Soit $\overrightarrow{\mathrm{u_E}}$ cet axe normalisé dont le sens est le même que le vecteur $\overrightarrow{\mathrm{P_1P_2}}$ partant du pieds 1 en direction du pieds 2. Ainsi les articulations E_1 et E_2 des épaules sont estimées par :

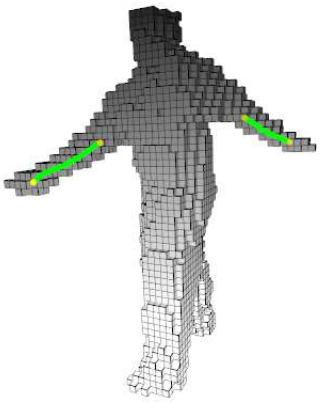


FIG. 10.5 – Représentation $G_{[0,1;0,3]}$. Les chemins décrits permettent d'extraire les positions des mains, des bras et des coudes.

$$E_2 = E + \overrightarrow{u_E}\frac{L_{Epaules}}{2} \tag{10.2}$$

$$E_1 = E - \overrightarrow{u_E}\frac{L_{Epaules}}{2} \tag{10.3}$$

Afin de déterminer l'axe des hanches, nous estimons l'axe principal d'inertie des voxels associés aux sommets du sous-chemin du graphe $G_{0,3}$ partant de s_{Cv} et s'arrêtant au sommet s_H. Soit $\overrightarrow{u_H}$ cet axe normalisé dont le sens est le même que le vecteur $\overrightarrow{P_1P_2}$. Ainsi les articulations H_1 et H_2 des hanches sont estimées par :

$$H_2 = H + \overrightarrow{u_H}\frac{L_{Hanches}}{2} \tag{10.4}$$

$$H_1 = H - \overrightarrow{u_H}\frac{L_{Hanches}}{2} \tag{10.5}$$

Mains

Dans les hypothèses de fonctionnement, nous supposons les bras et les mains écartés (ils ne touchent pas le corps). Ainsi les branches de chaque membre sont représentées dans G et les extrémités du corps correspondent aux points critiques extrema du graphe de Reeb. Le graphe $G_{0,3}$ permet de ne conserver que les branches et les point critiques du graphe qui correspondent à la tête, au buste, et au jambes. Afin de déterminer la position des mains, nous calculons le graphe $G_{[0,1;0,3]}$ qui supprime une partie des sommets de G issus des erreurs de reconstruction, et qui contient les branches associées aux bras (voir figure 10.5).

Comme nous l'avons précisé, les arcs du graphe de Reeb sont définis par les chemins de G qui relient les points critiques du graphe de Reeb. $G_{[0,1;0,3]}$ contient un sous-ensemble des arcs du graphe de Reeb, dont certains sont connectés aux points caractéristiques correspondants aux

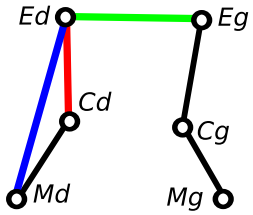


FIG. 10.6 – Trièdre de sommet E_d et de base les articulations C_d (rouge) , E_g (vert) et M_d (bleu) est direct.

mains. Les propriétés mécaniques élémentaires du corps imposent que les bras soient contenus dans une sphère Sph_1 centrée au centre des épaules E, de rayon $L_{Stature}/2$. Ainsi parmi les arcs de $G_{[0,1;0,3]}$ totalement contenus dans Sph_1, deux correspondent aux 2 bras. Nous supposons les bras suffisamment écartés, de façon à ce que les avant-bras ne touchent pas le buste. Les arcs qui représentent les bras ont ainsi une longueur d'au moins $L_{AvantBras}$. Si ces arcs sont au nombre de deux, ils sont associés à chaque épaule correspondante, par plus proche voisin. Enfin pour chaque arc validé, le point critique extremum associé correspond à l'extrémité de la main (voir figure 10.5).

Genoux et coudes

Pour chaque extrémité, l'arc connecté approxime le membre associé (tibia, avant-bras). De façon similaire à la détermination du centre du cou, nous parcourons pour chaque articulation, l'arc connecté jusqu'à obtenir la distance voulue. Si l'arc se révèle trop court, le dernier point de l'arc permet d'estimer la direction du membre parcouru. Ainsi connaissant la distance anthropométrique nous estimons la position correspondante. Pour extraire les coudes, nous travaillons sur les arcs du graphe de Reeb contenus dans $G_{[0,1;0,3]}$, alors que pour estimer la position des genoux, nous nous intéressons aux arcs appartenant à $G_{0,3}$.

Orientation gauche-droite

Nous avons déterminé l'ensemble des articulations du squelette d'animation choisi. Cependant les articulations gauches et droites n'ont pas été déterminées. Pour cela nous proposons d'utiliser les contraintes mécaniques liées aux rotations observées au niveau des coudes. En effet, dans une position confortable, les bras ne sont pas parfaitement tendus, ainsi l'angle du coude permet de déterminer l'orientation gauche-droite.

Considérons le trièdre de sommet E_1, l'épaule identifiée 1, et de base les articulations C_1, E_2 et M_1. Si ce trièdre est direct alors les articulations identifiées 1 (resp. 2) correspondent aux articulations de la partie droite (resp. gauche) du corps (voir figure 10.6).

L'ensemble des étapes que nous venons de proposer permettent une estimation de la posture

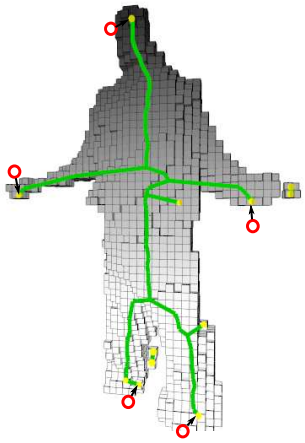


FIG. 10.7 – Illustration de la procédure de suivi des extrémités par plus proche voisin. La feuille (en jaune) la plus proche de chaque extrémités extraite lors de la précédente trame (en rouge), lui est associée.

initiale. Cependant ces étapes ne restent valables que lorsque l'individu respecte certaines conditions. En effet il est nécessaire que le point de plus haute altitude corresponde à la tête. Ainsi le personnage à suivre doit disposer ses mains en dessous du niveau de la tête, sans les coller au buste, et les tenir relativement tendues. Enfin il est aussi nécessaire que ses jambes soient un peu écartées. Ces critères, quoique peu limitant lors de l'initialisation, interdisent l'acquisition de mouvements qui ne respectent pas ses contraintes. Cependant, nous disposons de toute l'information nécessaire pour le réaliser : la structure topologique offre l'extraction des extrémités du corps, les chemins les reliant permettent d'estimer les articulations intermédiaires (genoux et coudes). Enfin les lois de la cinétique imposent une continuité temporelle liant la position de chaque articulation au temps.

10.2.3 Suivi des articulations

Nous proposons dans la suite une approche permettant d'estimer à chaque pas de temps la position courante de chaque articulation à partir de la posture extraite au pas de temps précédent et des observables courants. Cette approche peut être décrite selon deux blocs fonctionnels, le suivi des extrémités du corps, puis l'extraction des articulations intermédiaires.

Suivi des extrémités

Les extrémités du corps humain telles que les mains, les pieds et la tête définissent, lorsqu'elles ne sont pas en contact avec d'autre parties du corps, les feuilles du graphe de Reeb. Connaissant la position de chaque extrémité du corps à la trame précédente, il devient possible d'estimer leur position à partir des feuilles du graphe de Reeb G calculé à l'instant courant. Pour cela nous proposons une mise en correspondance construite sur plus proche voisin.

Nous calculons la distance euclidienne entre chaque extrémité extraite à la trame précédente, avec chaque feuille du graphe de Reeb courant. La feuille la plus proche de chaque extrémité est élue nouvelle position de cette extrémité (voir Fig. 10.7). Avant que cette mise à jour soit effective, nous vérifions qu'une feuille ne soit pas élue pour plusieurs extrémités. Dans ce cas l'extrémité qui lui est la plus proche est validée, alors que les autres ne seront pas mises à jour. En effet la fusion de plusieurs extrémités rendrait le système incapable de les dissocier par la suite. Notons que la vitesse de déplacement de chaque extrémité du corps dispose d'une borne supérieure, définie en fonction de la fréquence d'acquisition. Toute association dont la distance dépasse ce seuil est invalidée. Avec cette approche nous disposons désormais d'un mécanisme de suivi des extrémités simple, particulièrement rapide, offrant un processus temps réel. Notons que dans le cas ou une extrémité n'a pu être associée à l'une des feuilles de G, sa position à la trame précédente est conservée pour la trame courante. Disposant désormais d'une estimation de la position courante de chaque extrémité, nous nous intéressons désormais à l'extraction des articulation intermédiaires.

Extraction de articulations intermédiaires

Le graphe de Reeb G est décrit par un ensemble d'arcs sur lesquels peuvent être extrait des chemins reliant chaque couple de feuilles. Nous proposons d'utiliser cette information afin d'extraire les articulations intermédiaires. Parmi les feuilles validées comme extrémités, les articulations intermédiaires se situent à proximité de chemins se terminant en l'articulation extrême correspondante. La prise en compte des rapports anthropométriques proposés section 10.2.2 permet ensuite de déterminer le sommet de G correspondant à chaque articulation intermédiaire. La principale difficulté provient du choix de chaque chemin se terminant sur l'articulation considérée. En effet, si l'on souhaite rendre l'extraction de ces articulations résistante au bruit de reconstruction, il devient nécessaire de définir une stratégie dépendante de l'articulation traitée. Pour l'estimation de la position de chaque genou, nous proposons de réaliser la recherche sur le plus court chemin reliant le pied correspondant, à la tête. L'estimation de la position de chaque coude est elle aussi réalisée sur le chemin le plus court reliant la main au pied du même côté. Enfin l'estimation des positions du cou, du centre des épaules, du centre de la colonne vertébrale et du centre des hanches sont réalisées sur le chemin le plus court reliant la tête à l'un des pieds.

La figure 10.8 illustre l'extraction des articulations intermédiaire à partir des chemins précédemment décrits, ainsi qu'à l'aide des rapports anthropométrique que nous avons proposés. Notons que cette procédure est appliquée uniquement pour les extrémités dont l'extraction a été réalisée pour la trame courante. Dans le cas contraire, par exemple lorsque l'une des mains n'a pu être extraite à partir des feuilles de G, alors la position du coude au temps courant sera estimée par la position extraite au temps précédent.

Il reste désormais à extraire les positions des épaules et des hanches. Nous utilisons la même méthode que proposée lors de la section initialisation. Ainsi l'axe des épaules est estimé par

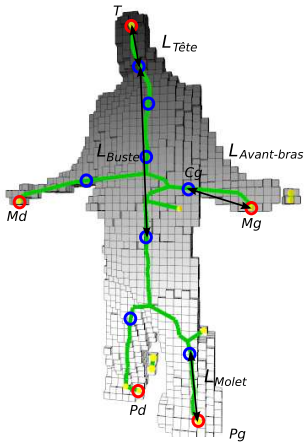


FIG. 10.8 – Illustration de la procédure d'extraction des articulations intermédiaires (en bleu) à partir des chemins entre les extrémités (en rouge) et des rapports anthropométriques.

le vecteur principal d'inertie des voxels associés aux sommets du chemin le plus court de G reliant le sommet s_E, centre des épaules, à s_{Cv}, centre de la colonne vertébrale. Une fois le sens gauche-droite de ce vecteur calculé, et en fonction du sens estimé au pas de temps précédent, les rapports anthropométriques permettent d'extraire les positions des épaules. Cette même approche est utilisée afin d'estimer l'axe et les positions des hanches, l'axe principal d'inertie étant calculé à partir des voxels associés au sommets du plus court chemin, reliant le centre de la colonne vertébrale s_{Cv} au centre des hanches s_H.

10.2.4 Résultats

Dans cette section nous présentons quelques résultats de cette approche à partir de données réelles. Pour une analyse qualitative et quantitative, le chapitre 13 présente plusieurs expérimentations réalisées à partir de données synthétiques.

Les expérimentations suivantes sont construites à partir d'une reconstruction de la forme 3D définie sur une grille de 64^3 voxels, échantillonnant un cube de $2m$ de côté. Cette reconstruction est réalisée à partir de 6 flux vidéos. La figure 10.9 présente plusieurs acquisitions de mouvements proposées par cette approche. Nous observons qu'elle fournit des résultats d'une qualité moyenne. Pour 4 des images présentées, les hanches sont inversées, ceci provient ses pertes dans le suivi de leur positions, lors de trames précédentes. Ces pertes sont la conséquence de l'absence de détection des extrémités correspondantes aux mains. L'image en bas à gauche souligne l'incapacité du graphe de Reeb sur fonction de hauteur à extraire les structures perpendiculaires pour lesquelles f offre la même valeur de potentiel. Ainsi les mouvements qui peuvent être correctement extraits par cette approche se limitent aux postures pour lesquelles aucun membre ou extrémité ne se trouve à l'horizontal. De plus cette approche ne s'avère pas résistante aux contacts. Elle s'avère incapable d'extraire les articulations des jambes pour la configuration présentée sur la première

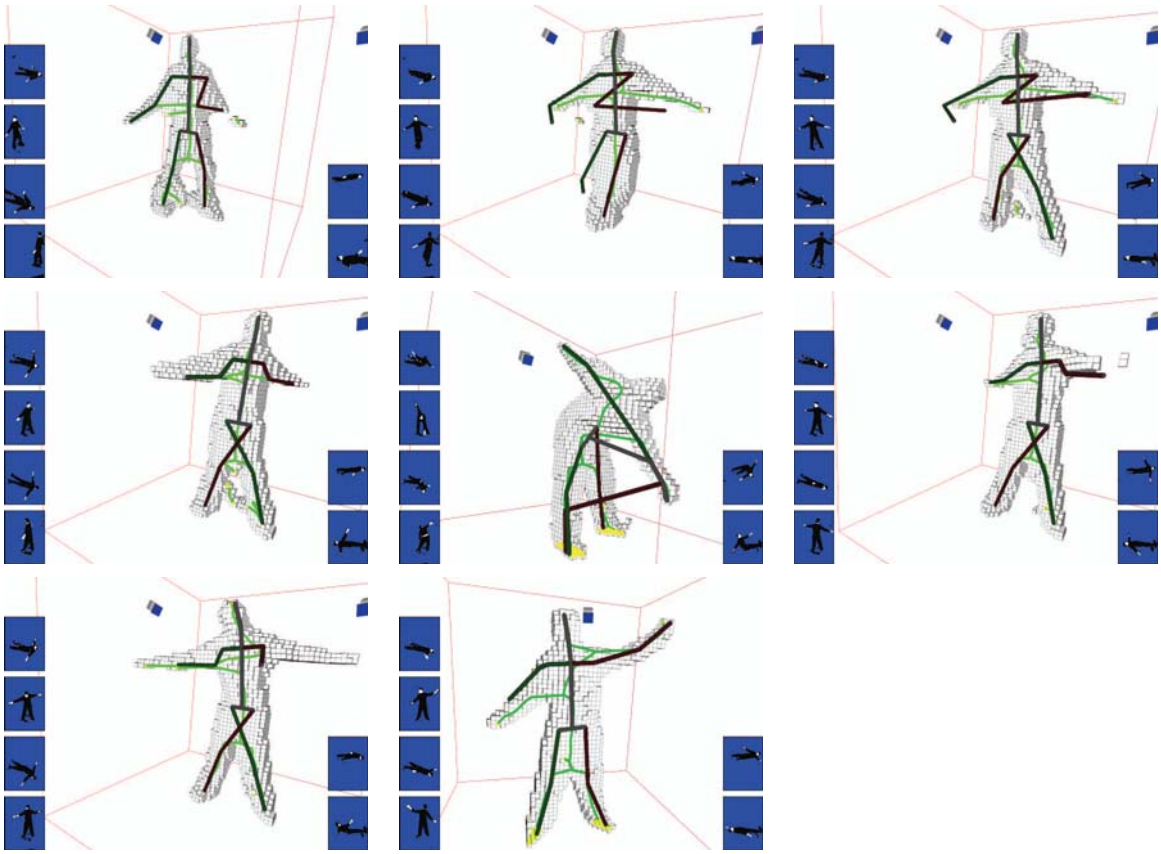


FIG. 10.9 – Résultats de l'approche sur un ensemble de mouvements simples.

ligne, au centre, de la figure 10.9.

Dans cette première section, nous avons proposé une approche simple et totalement automatique d'initialisation et d'acquisition de mouvements d'une personne. L'un de nos objectifs a été de réduire au maximum la complexité de l'approche, afin de déterminer les verrous principaux liés à l'initialisation et au suivi. Il en ressort une méthode permettant une initialisation efficace d'acquisition de mouvements. Cependant le manque de robustesse de l'extraction du graphe de Reeb sur fonction de hauteur, implique une acquisition de mouvements fragile. Cette représentation présente néanmoins l'avantage de la vitesse et de la simplicité. Cependant elle ne propose pas une extraction invariante par rotation. La topologie de la forme 3D extraite par graphe de Reeb construit sur fonction de hauteur n'est pas toujours complète. Si la personne filmée se met en pose T, les branches correspondantes aux bras ne sont pas détectées. Ainsi toute partie du corps horizontale ne peut être prise en compte par un fonction liée uniquement à l'altitude. Cette seconde limite provient du fait que l'extraction de ce type de graphe de Reeb, n'est pas invariante par posture. L'utilisation d'approches d'extraction de la topologie invariantes par rotation, permettrait un suivi de mouvements plus efficace.

Il existe principalement deux autres types d'approches de calcul de graphe de Reeb. La première s'appuie sur une fonction de distance barycentrique et la seconde sur les distances

géodésiques. L'approche géodésique offre généralement de bon résultats au sacrifice des temps de calcul, même si certains travaux récents proposent des approches rapides [PSBM07]. L'extraction de la topologie par graphe de Reeb barycentrique semble être une alternative intermédiaire aux approches construites sur fonction de hauteur et celles construites sur distances géodésiques.

Dans la section suivante nous décrivons une méthode d'extraction de graphe de Reeb barycentrique, qui sera ensuite utilisée pour réaliser le suivi de mouvements.

10.3 Acquisition du mouvement par Graphe de Reeb construit sur fonction de distance barycentrique

Nous commençons par proposer une méthode de calcul de graphe de Reeb barycentrique qui respecte le temps réel. Ensuite nous présentons le mécanisme de filtrage des points critiques du graphe de Reeb afin de réduire sa sensibilité aux bruits issus de la reconstruction. Nous terminons ce chapitre en présentant les résultats obtenus.

10.3.1 Calcul du graphe de Reeb barycentrique

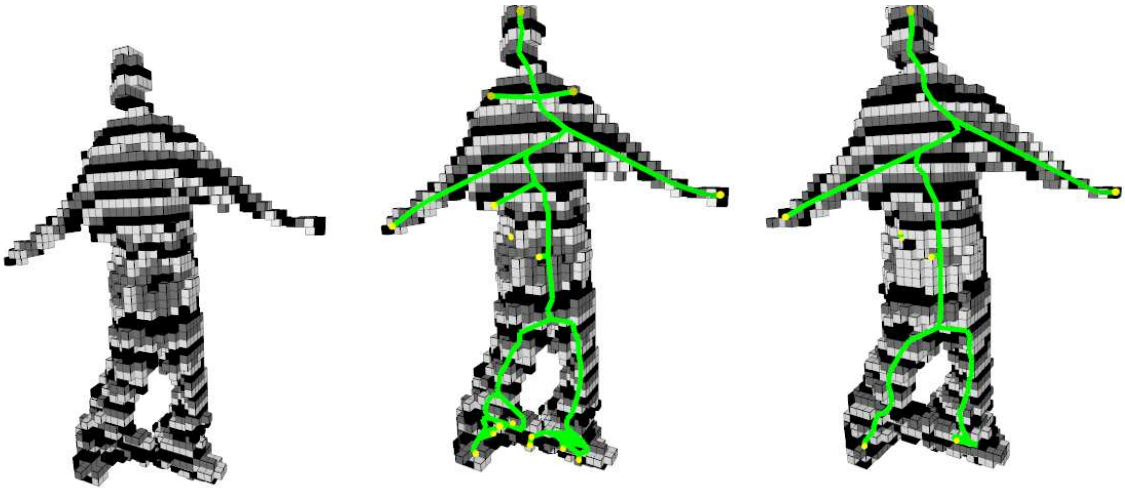
Parmi les méthodes de calcul de graphe de Reeb, certaines approches sont construites sur une fonction f, distance euclidienne de chaque point de la forme 3D au barycentre. Cette approche a pour avantage de proposer la construction d'un graphe invariant par rotation. Dans notre cas, nous disposons d'un ensemble de voxels (notés $v_k \in V$) défini dans une grille régulière. Les voxels intersectés par une sphère centrée sur le barycentre, fournissent la même valeur de f. Ainsi les voxels peuvent être représentés par couches "sphériques". Nous notons V_j l'ensemble des voxels appartenant à une même couche "sphérique" j.

Nous calculons pour chaque couche j les composantes connexes notées dans la suite cc_i, en commençant par la couche de plus faible distance (V_0) contenant le barycentre. Ensuite nous définissons un graphe orienté $G(S, A)$ dont les sommets S représentent les barycentres de chaque composante connexe cc_i. Les sommets $(s_l, s_m) \in S^2$ associés respectivement aux composantes connexes cc_l et cc_m, sont connectés par un arc si et seulement si les conditions suivantes sont respectées :

- $\text{cc}_l \subset V_i$ et $\text{cc}_m \subset V_{i+1}$, c'est à dire cc_m est sur la couche successive de cc_l dans le sens du parcours des couches ;

- $\exists v_n \in \text{cc}_l, \exists v_o \in \text{cc}_m : d(v_n, v_o) = 1$ avec d la distance d_{26}.

L'ensemble des points critiques de f définissent les sommets du graphe de Reeb de V (voir figure 10.10(b)). Remarquons qu'avec cette approche il n'existe qu'un seul sommet de G source, correspondant au barycentre de V. Les arcs liants les sommets du graphe de Reeb, sont décrits



(a) (b) (c)

FIG. 10.10 – Représentation des couches sphériques de l'estimation volumique V d'un personnage (a), le Graphe de Reeb barycentrique associé (b) et le graphe de Reeb barycentrique après filtrage médian (c).

par les chemins de G passant uniquement par des sommets dont le degré sortant et entrant est de 1.

Cette méthode du calcul du graphe barycentrique de V fournit une extraction du squelette topologique invariante par rotation. Cependant nous observons une forte sensibilité de cette approche aux bruits présents dans la reconstruction V. Nous proposons d'appliquer un filtre médian sur V, avant d'en extraire le squelette topologique. Filtrer des données binaires par un filtre médian sur un voisinage donné, revient à remplacer la valeur d'un voxel par l'information majoritaire observée sur ces voisins, lui compris. Afin de conserver la contrainte du temps réel, nous avons implanté ce filtre sur carte graphique programmable.

Nous observons que l'utilisation de ce filtre améliore considérablement l'extraction topologique (voir Fig. 10.10(c)). Les petites discontinuités de V sont filtrée, or ce sont les sommets du graphe de Reeb issus de petites discontinuités que nous ne considérons pas pertinents.

Nous avons souligné dans la section précédente l'efficacité de l'approche par graphe de Reeb sur fonction de hauteur pour initialiser un processus de capture de mouvements. Cependant son utilisation pour le suivi s'est avérée moins efficace. Ceci étant du à l'incomplétude de la topologie extraite. L'extraction du graphe de Reeb barycentrique offre une réponse à ce problème, dont la construction est invariante par rotation. Nous avons alors appliqué la procédure de suivi décrite section 10.2.3 en utilisant le graphe de Reeb barycentrique. La section suivante décrit les résultats obtenus.

10.3.2 Résultats

La figure 10.11 présente les acquisitions réalisées à partir du même jeu de données et de la même reconstruction 3D que les résultats présentés figure 10.9. Nous observons une nette amélioration de la capture de mouvements obtenue.

Même si l'approche barycentrique a tendance à offrir trop de feuilles par rapport au nombre d'extrémités suivies, le mécanisme mis en place s'avère suffisant pour compenser cet effet. La précision de cette approches aux petites variations topologiques de la forme 3D, lui permet de traiter des configurations plus complexes que précédemment (voir la dernière image de la figure 10.11). Enfin l'apport de l'approche barycentrique par rapport à la méthode précédente est souligné figure 10.12. La propriété d'invariance par rotation de cette nouvelle approche permet l'acquisition de mouvements, sans impliquer d'hypothèse sur l'orientation globale du corps suivi. Pour plus de détails, une analyse comparative entre ces deux approches est exposée au chapitre 13. Notons que pour la configuration présentée, ces deux approches sont capables de fournir plus de 40 estimations de mouvements par seconde, pour une grille de voxels d'une résolution de $64 \times 64 \times 64$.

10.4 Discussion

Nous avons présenté une méthode qui propose une initialisation totalement automatique et qui réalise la capture des mouvements d'une personne filmée par plusieurs caméras, en temps réel, et sans l'utilisation de marqueurs. Cette première approche, construite uniquement sur l'extraction et l'analyse de la topologie, offre une solution particulièrement rapide. Les résultats obtenus à partir de données réelles valident à la fois l'approche par recalage de structure 1D, et le respect de la contrainte du temps réel sur une seule machine.

En effet cette méthode dont l'efficacité dépend avant tout de la robustesse de la topologie extraite nous a poussé a comprendre et a étudier les limites des approches *Shape-From-Silhouette*, afin de les rendre plus robuste dans l'objectif de fournir le moins d'erreur possible en entrée du processus d'acquisition de mouvements.

Le processus de capture proposé est construit sur l'extraction et l'analyse de la topologie de la forme 3D. Cette approche se révèle efficace tant que cette topologie correspond à celle du squelette d'animation de notre modèle. Lors de contacts, ou encore la présence de parties fantômes dans la reconstruction, sont autant de vecteurs modifiant la topologie de la forme 3D. Ces éléments limitent l'ensemble des postures qui peuvent être captés aux mouvements pour lesquels les membres ne sont pas en contacts. La sensibilité de cette approche aux parties fantômes impose l'utilisation d'un nombre important de caméras. Lorsque ces conditions sont réunies, les résultats offerts se révèlent suffisantes pour certaines applications d'interactions

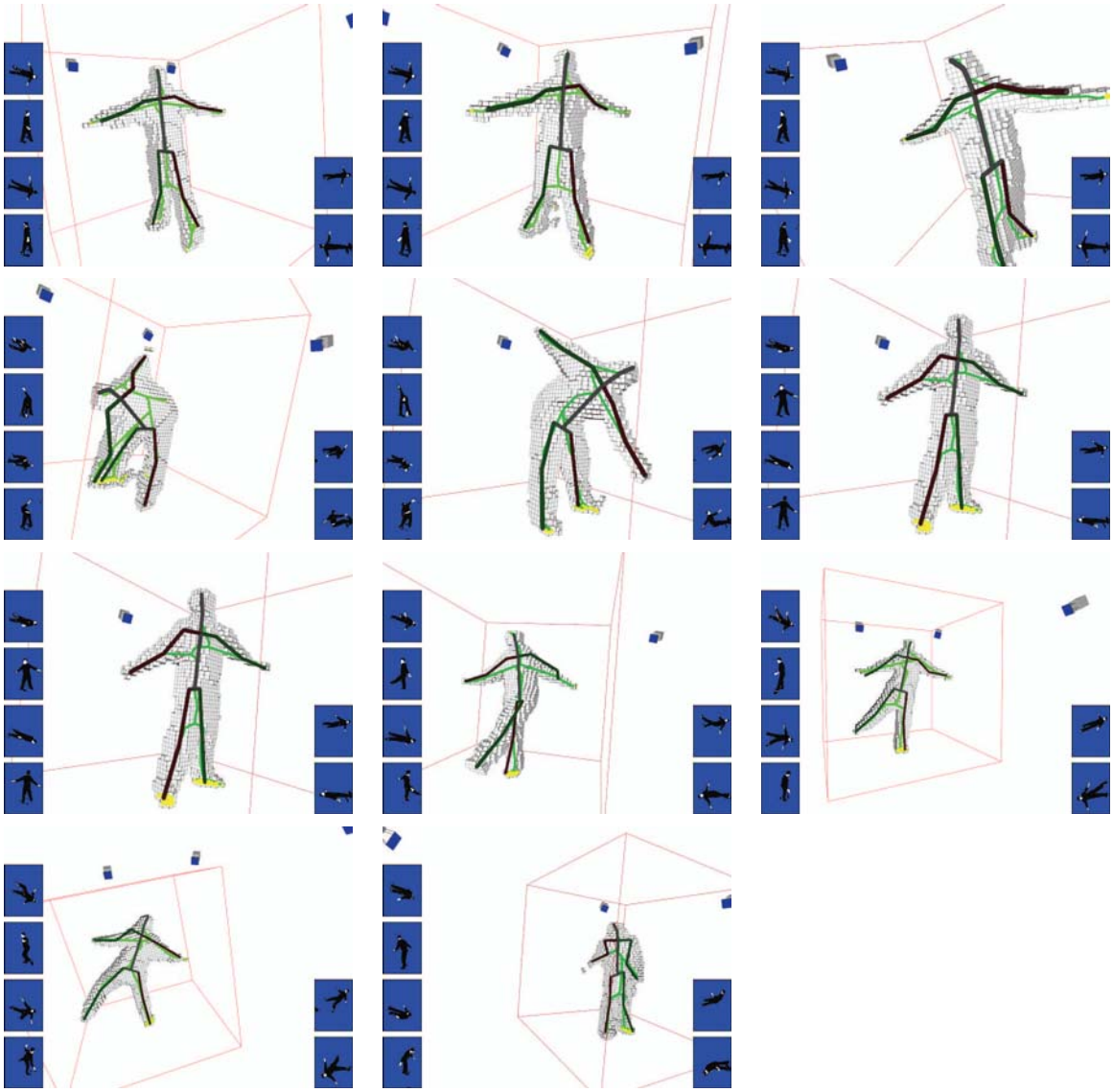


FIG. 10.11 – Résultats de l'approche construite sur l'extraction du graphe de Reeb barycentrique.

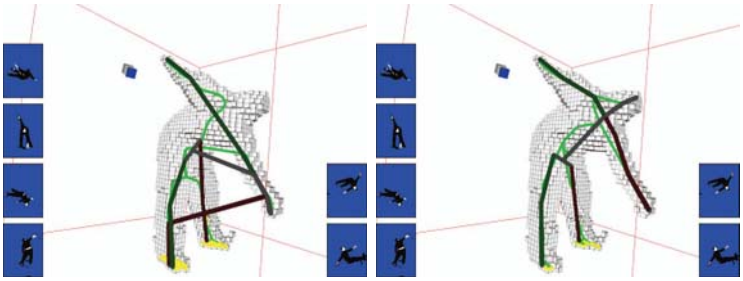


FIG. 10.12 – Comparaison de l'estimation d'une même posture par l'approche construite sur le graphe de Reeb construit sur fonction de hauteur (a) et barycentrique (b). Parce qu'elle est invariante par rotation, l'approche barycentrique offre une extraction précise.

homme-machines pour lesquelles la précision n'est pas un élément déterminant.

Il ressort de cette première solution que la principale difficulté est liée à l'extraction des extrémités du corps, sensible par essence, aux contacts. Pour cette raison, nous nous intéressons dans la suite au développement de méthodes robustes permettant leur extraction. Un premier élément de réponse est construit sur un marquage naturel de certaines extrémités du corps : la peau.

Chapitre 11

Recalage d'objets géométriques simples, dirigé par la couleur de peau

Les méthodes d'acquisition de mouvements décrites précédemment extraient certaines caractéristiques de la forme 3D, telles que la structure topologique. Ces méthodes nécessitent une estimation relativement précise de la forme 3D et imposent l'utilisation d'un nombre important de caméras (de l'ordre de 6). Nous avons formulé nos objectifs en direction d'approches qui imposent le moins de contraintes possibles. Un nombre de caméras important peut devenir une contrainte forte.

Lorsque l'on diminue le nombre de caméras, la quantité d'entité fantôme construite augmente (voir le chapitre 4) et la forme 3D estimée ne reflète plus suffisamment les caractéristiques topologiques de la personne filmée. Afin de répondre à cette problématique, nous proposons d'utiliser une information supplémentaire qui compense cette augmentation des entités fantômes : les parties de peau.

Dans ce chapitre, nous proposons une approche qui ne nécessite que quelques caméras et réalise l'acquisition du mouvement d'un personnage en temps réel. Cette méthode est construite sur l'utilisation d'un modèle d'humanoïde composé d'objets géométriques simples, dont le recalage dans la forme 3D est dirigé par l'information de peau. Partant du constat qu'une personne a rarement le visage et les mains couverts, il est possible de les localiser dans la forme 3D estimée. Cette identification est basée sur la couleur de la peau. Associée à des contraintes de longueurs des membres et à la notion de continuité temporelle, elle permet d'apporter une réponse à l'acquisition en temps réel des mouvements d'une personne filmée ; en relaxant la contrainte sur le nombre de caméras utilisées.

Nous commençons par extraire les zones de peau dans chaque image couleur. Ensuite nous calculons la reconstruction 3D de ces parties de peau, en utilisant les approches proposées dans la première partir de cette thèse. Le mécanisme de suppression des objets fantômes EOR_{seq} (voir chapitre 6) est utilisé afin de supprimer les zones 3D de peau fantômes. Les zones de

peau conservées décrivent généralement le visage et les mains. A partir de ces extrémités du corps, un modèle géométrique est recalé dans la forme 3D, offrant l'extraction des articulations principales du corps. Le processus global est capable de capturer plus de 30 postures par seconde sur un ordinateur usuel. La fusion de plusieurs modalités telles que la peau, la forme 3D et les contraintes liées aux longueurs des membres, rend cette approche robuste, même en présence d'un faible nombre de caméras.

Nous commençons par présenter le processus d'estimation conjointe de la forme 3D et des parties de peau à partir de plusieurs caméras calibrées. Nous détaillons ensuite la méthode d'acquisition de mouvements.

11.1 Estimation conjointe de la forme 3D et des parties de peau

Nous avons proposé dans la première partie de cette thèse nos contributions liées à la reconstruction 3D de personnes à partir des silhouettes issues de plusieurs caméras calibrées. Soit $\mathcal{V}_{\text{Tous}}$ l'ensemble des voxels qui représentent la personne filmée.

Nous souhaitons extraire les parties de $\mathcal{V}_{\text{Tous}}$ qui contiennent les mains et le visage. Cette extraction peut être réalisée en calculant la couleur de chaque voxel $\mathcal{V}_{\text{Tous}}$, puis en sélectionnant les voxels qui ont une couleur de peau. En présence d'images couleur, la couleur de chaque voxel de $\mathcal{V}_{\text{Tous}}$ est calculée en interpolant l'information de couleur issue des caméras pour lesquelles ce voxel n'est pas occulté. Cette approche, en plus d'être coûteuse (calcul de la visibilité de chaque voxel depuis chaque point de vue) n'est pas efficace en présence de peu de caméras. Les parties fantômes construites sont généralement nombreuse et peuvent ainsi être sources d'erreur.

Les masques binaires des parties de peau dans chaque image représentent la projection des mains et du visage dans les caméras. Ces masques sont considérés comme étant une information de silhouette, permettant d'estimer leur forme 3D par une approche *Shape-From-Silhouette*. Nous avons montré que la reconstruction de plusieurs objets à partir de quelques caméras, génère des objets fantômes que nous pouvons supprimer grâce à notre approche EOR_{seq}. Dans la suite nous noterons $\mathcal{V}_{\text{Peau}}$ l'ensemble des voxels issus des masques de peau, filtrés par EOR_{seq}.

11.1.1 Extraction de l'information de peau dans des images couleur

La caractérisation de la couleur de peau amène deux problèmes : le choix de l'espace colorimétrique et la méthode de modélisation de la couleur de peau [KMB07].

La colorimétrie, l'informatique graphique ou encore les standards de transmission vidéo ont donné naissance à diverses représentations de la couleur. L'espace RGB est très utilisé pour les applications informatiques et la sauvegarde d'images numériques. La forte corrélation entre les canaux, la non-uniformité perçue ou encore le mélange des informations de luminance et de chrominance, n'en font pas un choix intéressant. Cependant certaines approches de détection de

peau travaillent sur cet espace [JR02, BM00a]. L'espace RGB normalisé est une représentation où chaque composante de RGB est normalisée par la somme des trois valeurs. Pour des surfaces lambertiennes, RGB normalisé est invariant (sous certaines conditions) à l'orientation de la surface par rapport aux sources de lumière. Ce type d'espace a notamment été utilisé dans les travaux de [BCL01, ZSQ99]. HSV, HSL et HSI sont des espaces colorimétriques qui séparent l'information de teinte, de saturation et de luminance. Ces espaces décrivent la couleur avec des éléments intuitifs, basés sur la perception de la couleur. Ceux-ci ont été utilisés pour la caractérisation de la couleur peau [MGR98, SSA00], mais nécessitent en général une quantité de calcul importante. L'espace $YCrCb$ provient d'un codage non linéaire de l'espace RGB, souvent utilisé dans certains schémas de compression d'image. Cet espace propose une alternative intéressante aux espaces HSx, tout en consommant peu de temps machine [CB00]. Le terme de couleur de peau ne définit pas un propriété physique, mais plutôt la perception que l'on peut en avoir. Les espaces de couleur perceptuellement uniformes, tels que $CIELAB$ et $CIELUV$, permettent une bonne caractérisation de la couleur de la peau [TFAS00, ZSQ99]. Passer de RGB ou $YCrCb$ à ce type de représentation nécessite une charge intensive de calcul.

Les caméras utilisées lors des expérimentations fournissent un signal $YCrCb$. Nous travaillerons directement dans cet espace, afin de ne pas surcharger la machine calcul et limiter les erreurs dues aux conversions entre espaces. Cet espace fournit la séparation de l'information de luminance de l'information de chrominance. Afin de rendre l'extraction de la peau robuste aux conditions d'éclairage, nous ne travaillerons que sur le plan de chrominance $CrCb$.

Parmi les modélisations de la peau proposées dans la littérature, les approches peuvent être classées selon plusieurs axes.

Caractérisation explicite des régions de peau

Une méthode pour construire un classifieur de peau est de définir explicitement à travers un ensemble de critères simples, les limites des clusters de peau dans certains espaces de couleur. Le principal avantage de cette approche est la simplicité des règles de détection de peau, qui permettent la construction d'une classifieur rapide [KS03, FFB96]. La plus grande difficulté pour réaliser un bon taux de reconnaissance avec cette méthode provient du choix de l'espace de couleur et des règles de décision empiriques associées. Certaines approches sont construites sur des algorithmes d'apprentissage qui permettent de déterminer automatiquement l'espace colorimétrique ainsi que les règles de détection [GM02]. Ce type d'approche reste de manière générale, peu robuste aux variations de l'éclairement et fournit des taux de reconnaissance relativement faibles.

Modélisation non-paramétrique de la distribution de peau

L'idée clé des méthodes de modélisation non-paramétrique de la peau est d'estimer la distribution de la couleur de la peau à partir de données d'apprentissage, sans en extraire un modèle explicite. Les résultats de ces approches sont parfois appelés carte de probabilité de peau [BM00a, Gom02] en associant une probabilité à chaque point de l'espace couleur discrétisé. Les approches de cette classe s'appuient en général sur une estimation de la densité de probabilité par histogramme, par apprentissage et classification bayesienne [ZSQ99, SCHG04, CB00] ou encore par des approches par réseaux de neurones [BCL01]. Nous nous intéressons particulièrement aux approches par histogrammes, particulièrement rapides, ainsi en accord avec la contrainte du temps réel.

Plusieurs approches de détection de visage et de suivi s'appuient sur un histogramme normalisé afin de segmenter les pixels de couleur de peau [SSA00,SSA04,SH00]. L'espace colorimétrique, généralement limité au plan de chrominance, est sous-échantillonné en éléments auxquels correspondent plusieurs chrominances. Chaque élément contient la fréquence d'apparition de chaque chrominance durant l'étape d'apprentissage. L'histogramme est ensuite normalisé, de façon à le convertir en une distribution discrète de probabilité :

$$P_{c|peau} = \frac{peau[c]}{Norme} \tag{11.1}$$

où $peau[c]$ est la fréquence de la couleur c observée, sachant qu'elle est de la couleur de peau et $Norme$ est le coefficient de normalisation correspondant, à la somme des valeurs de l'histogramme [JR02], ou encore à la valeur maximale de l'histogramme [ZSQ99]. Ainsi $P_{c|peau}$ constitue une mesure de confiance pour une couleur c d'être une couleur de peau. Cette approche peut ensuite être étendue en calculant l'histogramme $\neg peau[c]$ puis $P_{c|\neg peau}$ de la même manière. Ainsi caractériser si une couleur c est couleur de peau peut être vérifié par le critère suivant :

$$\frac{P_{c|peau}}{P_{c|\neg peau}} > K \tag{11.2}$$

avec $K \in [0, 1]$ un seuil.

Les deux principaux avantages de ce type d'approche, résident à la fois dans la vitesse de l'apprentissage et d'utilisation, ainsi que sur leur indépendance théorique à la distribution de la couleur de peau, ce qui n'est pas le cas pour la majorité des modélisations paramétriques.

Modélisation paramétrique de la distribution de peau

Les approches les plus répandues, construites sur des histogrammes, nécessitent que le jeu de données d'apprentissage soit suffisamment représentatif des différents types de peau à reconnaître. Il est alors nécessaire de déterminer une représentation de la peau qui puisse généraliser et interpoler les données d'apprentissage. Plusieurs approches paramétriques ont été

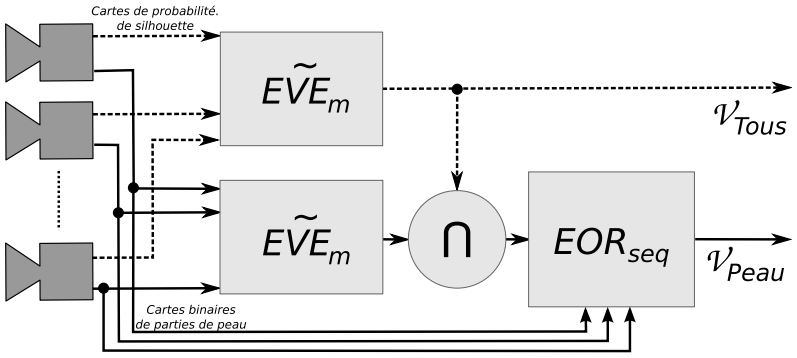


FIG. 11.1 – Représentation fonctionnelle du processus de reconstruction de forme 3D et des parties de peau.

proposées. Certaines méthodes estiment la distribution de la couleur de peau par une gaussienne [TFAS00, HAMJ02]. D'autres modèles plus précis, capables de décrire des distributions complexes, sont basés sur l'utilisation de mélange de gaussiennes [TFAS00, MGR98, YW07]. En observant les distributions de peau et de non peau dans divers espaces de couleur, Lee et Yoo [LY98] ont proposé un modèle à base d'ellipses, qui offre de meilleurs résultats qu'une modélisation par une ou plusieurs gaussiennes, avec des temps d'apprentissage et de détection inférieurs.

Afin de respecter la contrainte du temps réel, nous utilisons l'approche construite sur un histogramme normalisé du plan chromatique $CrCb$, qui offre un compromis intéressant entre performance de la caractérisation et temps de calculs. L'apprentissage de la couleur de peau a été réalisé pour un échantillon de quelques personnes et fournit dans notre contexte expérimental des résultats satisfaisants. Dans le cas de l'utilisation étendue à un ensemble de personnes, il devient nécessaire de prévoir une procédure de mise à jour du modèle de peau, afin de l'adapter aux personnages filmés.

11.1.2 Reconstruction 3D de la forme et des parties de peau

En utilisant plusieurs canaux de couleur du *frame-buffer* de la carte graphique programmable, nous calculons simultanément $\widetilde{\mathrm{EVE}}_m(Silh)$ à partir des cartes de silhouettes et $\widetilde{\mathrm{EVE}}_m(Peau)$ à partir des cartes de parties de peau.

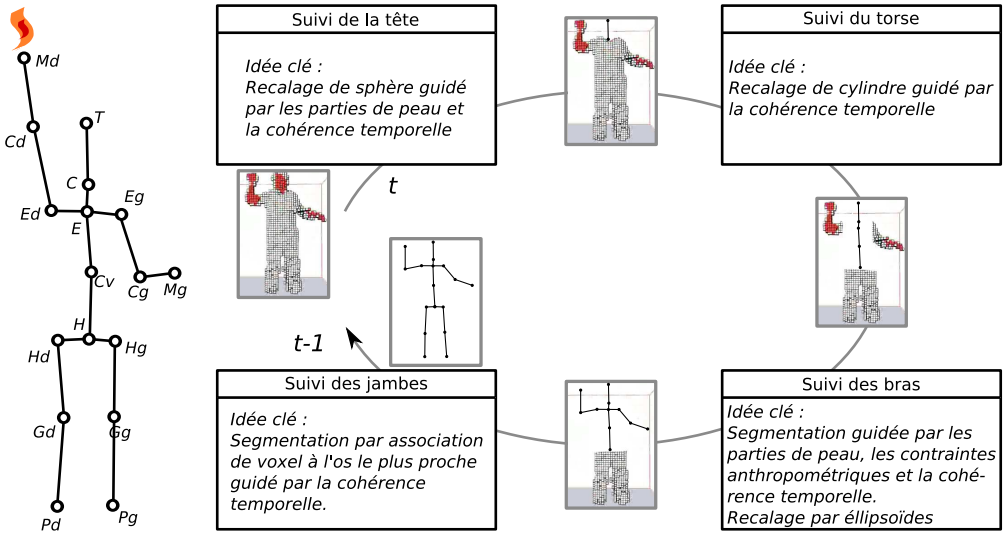
L'approche d'extraction des parties de peau proposée offre un bon taux de reconnaissance, mais aussi un taux important de faux positifs dus au fond de la scène. Afin de supprimer les artefacts de reconstruction issus de ces faux-positifs, nous filtrons $\widetilde{\mathrm{EVE}}_m(Peau)$ en calculant l'intersection avec $\widetilde{\mathrm{EVE}}_m(Silh)$. En effet les parties de peau de la personne sont incluses dans sa forme.

Nous devons prendre garde au fait que les parties de peau de la personne filmée ne sont pas toujours visibles dans toutes les caméras en même temps. Nous relaxons cette contrainte en calculant $\widetilde{\text{EVE}}_m(Peau)$ avec une valeur du seuil d'acceptation T bas (voir le chapitre 7). Ceci permet de compenser les parties de peau occultées.

En présence de plusieurs parties de peau 3D observées par un faible nombre caméras, la quantité d'objets fantômes construits est importante. Afin de supprimer les objets de peau fantômes, nous appliquons ensuite l'approche EOR_{seq} sur la forme 3D construite par $\widetilde{\text{EVE}}_m(Peau)$.

La figure 11.1 présente notre approche de calcul de $\mathcal{V}_{\text{Tous}}$ et de $\mathcal{V}_{\text{Peau}}$. Nous analysons ensuite ces estimations volumiques afin de réaliser la capture du mouvement en temps réel de la personne filmée.

11.2 Vue d'ensemble



(a) (b)

FIG. 11.2 – (a) Modèle d'animation utilisé et identification de chaque articulation. (b) Vue d'ensemble du suivi de mouvements (b)

L'objectif principal de l'acquisition du mouvement est de déterminer, à chaque instant, la posture du corps du personnage filmé. Connaissant une représentation 3D de sa forme, cette acquisition peut être réalisée par l'association de chaque voxel à l'un des os du modèle d'humanoïde utilisé (rappelé figure 11.2(a)). Dans ce chapitre nous proposons une approche construite sur ce principe, en segmentant les différentes parties du corps de la personne filmée dans sa forme estimée. Le système proposé est construit sur un ensemble d'heuristiques simples et rapides à évaluer. Cette approche, moins précise que les approches de recalage par minimisation,

permet de conserver le temps réel. Nous proposons un processus utilisant l'analyse conjointe de la forme 3D et des parties de peau, la cohérence temporelle et des contraintes liées aux distances anthropométriques du corps humain.

Notre approche procède en deux étapes : l'initialisation, puis le suivi. L'étape d'initialisation (voir la section 11.4) estime la posture initiale du sujet à partir des rapports anthropométriques présentés chapitre 10.2.2. A partir de cette information, nous profitons de la continuité temporelle des mouvements humains afin de réaliser le suivi des articulations du sujet au cours du temps (voir la section 11.3). Nous supposons que seules les mains et le visage de la personne sont découverts et que les vêtements n'ont pas une couleur similaire à la peau du sujet.

Nous introduisons les notations utilisées dans la suite de ce chapitre. Pour chaque partie rigide du corps x (voir la figure 11.2(a)), L_{x} dénote sa longueur, $\overrightarrow{D_{\mathrm{x}}}$ son orientation, R_{x} son rayon (pour une modélisation par boule ou par cylindre) et $\mathcal{V}_x$ l'ensemble des voxels qui la représente. $\mathcal{E}_{\mathcal{V}_x}$ est l'ellispoïde d'inertie de $\mathcal{V}_x$ (voir annexe B) et $CdG(\mathcal{V}_x)$ son barycentre. L'indice g (resp. d) décrit chaque partie gauche (resp. droite). Pour une quantité Q (position d'une articulation, ensemble de voxels, ...) Q^t dénote sa valeur au temps t, et $Q^t(i)$ désigne cette quantité à l'étape i pour des d'algorithmes itératifs.

Notons que nous utilisons les mêmes approches pour résoudre à la fois les étapes d'initialisation et de suivi. La différence provient de conditions initiales différentes. Nous commençons par présenter les approches génériques qui réalisent le suivi de chaque articulation au cours du temps. Nous traitons ensuite de l'adaptation de ces méthodes spécifiquement au processus d'initialisation.

Notre approche travaille à partir d'un ensemble de voxels "actifs" $\mathcal{V}^t_{\mathrm{Actif}}$. A chaque temps t, cet ensemble est initialisé par $\mathcal{V}^t_{\mathrm{Tous}}$. Les articulations sont extraites séquentiellement, permettant au fur et à mesure de diminuer l'espace de recherche des articulations suivantes. Après avoir traité un groupe d'articulations, les voxels associés sont supprimés de $\mathcal{V}^t_{\mathrm{Actif}}$.

11.3 Suivi des articulations

Afin de suivre les mouvements du personnage filmé à un instant t, nous supposons que les longueurs de ses membres (fixes à travers le temps) et sa posture au temps précédent $t-1$ sont connues. Comme nous l'avons précisé, nous proposons une extraction séquentielle des articulations (voir figure 11.2(b)). Nous commençons par estimer les articulations de la tête. En effet son déplacement moins rapide que les extrémités du corps ainsi que l'information de peau, permettent d'extraire sa position facilement, et de façon robuste. Nous estimons ensuite les articulations du buste, connecté à la tête. Enfin les articulations des membres sont extraites.

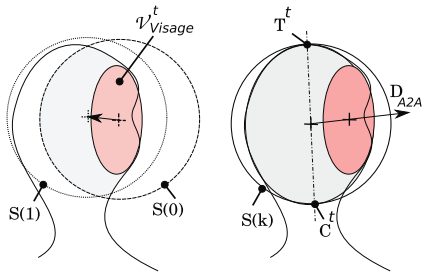


FIG. 11.3 – (a) Estimation des articulations de la tête par recalage de boule (les parties rosées décrivent $\mathcal{V}^t_{\text{Visage}}$, le gris décrit $\mathcal{V}^t_{\text{Tete}}(i)$), (b) estimation des articulations de la tête.

11.3.1 Suivi de la tête

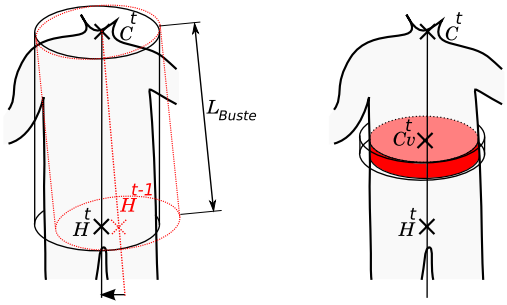
Cette étape a pour objectif de déterminer les articulations T^t et C^t, respectivement la position du haut de la tête et la connexion entre la tête et le cou à l'instant t. Pour cela, il nous faut identifier dans $\mathcal{V}^t_{\text{Actif}}$, les voxels qui correspondent à la tête ($\mathcal{V}^t_{\text{Tete}}$). Nous utilisons une méthode de recalage d'une boule $\mathcal{B}oule$ de centre B et de diamètre L_{Tete}, inspirée de [MTHC03]. L'initialisation du processus se fait en centrant la boule sur le visage (voir figure 11.3(a)). $\mathcal{V}^t_{\text{Peau}}$ contient à la fois les voxels du visage et les voxels correspondants aux mains. En utilisant un critère de cohérence temporelle, $\mathcal{V}^t_{\text{Visage}}$ est estimé par la composante connexe de $\mathcal{V}^t_{\text{Peau}}$ la plus proche[4] de la position précédente du visage $\mathcal{V}^{t-1}_{\text{Visage}}$. Le centre de la tête est estimé par l'algorithme suivant :

$i \leftarrow 0$;
$k \leftarrow 0$;
$B(0) \leftarrow CdG(\mathcal{V}^t_{\text{Visage}})$;
Répéter
  $\mathcal{V}^t_{\text{Tete}}(i) \leftarrow \mathcal{V}^t_{\text{Actif}} \cap \mathcal{B}oule(i)$;
  $B(i+1) \leftarrow CdG(\mathcal{V}^t_{\text{Tete}}(i))$;
  $i++$;
jusqu'à ce que ($||B(i) - B(i-1)|| < \epsilon_{Tete}$)
$k \leftarrow i - 1$;
$\mathcal{V}^t_{\text{Tete}} \leftarrow \mathcal{V}^t_{\text{Tete}}(k)$;
$\mathcal{V}^t_{\text{Actif}} \leftarrow \mathcal{V}^t_{\text{Actif}} - \mathcal{V}^t_{\text{Tete}}$;
$\mathcal{V}^t_{\text{Peau}} \leftarrow \mathcal{V}^t_{\text{Peau}} - \mathcal{V}^t_{\text{Visage}}$;

Algorithme 6: Algorithme de recalage de boule

La direction Haut-Bas $\overrightarrow{D^t_{\text{HautBas}}}$ de la personne est déterminée par $\overrightarrow{\text{B(k) CdG}(\mathcal{V}^t_{\text{Actif}})}$. Les

[4] en utilisant une distance euclidienne d'un point à un ellipsoïde



(a) (b)

FIG. 11.4 – (a) Segmentation du buste par recalage de cylindre. (b) estimation de l'articulation Cv^t.

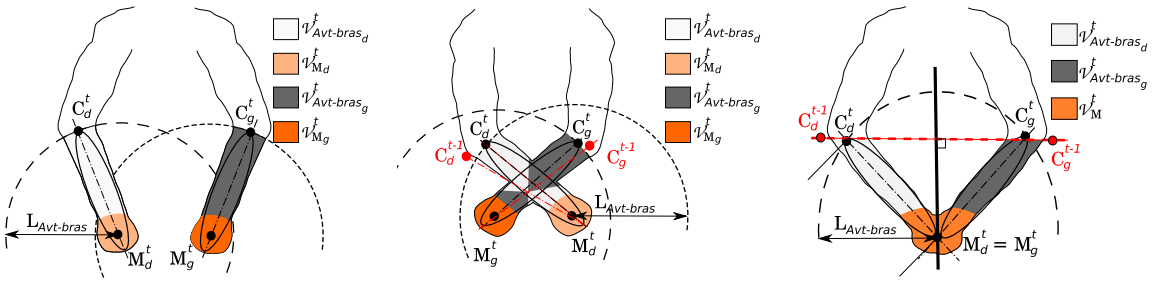
points d'intersection entre la sphère associée à $\mathcal{B}oule(\mathrm{k})$ et l'axe principal de $\mathcal{E}_{\mathcal{V}^t_{\mathrm{Tete}}}$ définissent les articulations T^t et C^t de la tête selon le sens $\overrightarrow{D^t_{\mathrm{HautBas}}}$ (voir figure 11.3(b)).

11.3.2 Suivi du buste

Cette étape a pour objectif de déterminer les positions du centre des hanches H^t et du centre de la colonne vertébrale Cv^t à l'instant t. Pour cela, il nous faut identifier dans $\mathcal{V}^t_{\mathrm{Actif}}$, les voxels qui correspondent au buste ($\mathcal{V}^t_{\mathrm{Buste}}$). Nous proposons d'estimer la forme globale du buste par un cylindre $\mathcal{C}yl$ de diamètre L_{Epaule}, de hauteur L_{Buste} et dont l'une de ses bases est ancrée sur le cou C^t. Nous cherchons à recaler ce cylindre dans $\mathcal{V}^t_{\mathrm{Actif}}$, afin de déterminer l'orientation générale du torse. Cette méthode de recalage est inspirée de la méthode proposée pour l'extraction de la tête. La direction $\overrightarrow{D^t_{\mathrm{Buste}}}(0)$ de l'axe de $\mathcal{C}yl$ est initialisée par le vecteur $\overrightarrow{\mathrm{C^tH^{t-1}}}$ (voir figure 11.4(a)). L'estimation de l'orientation globale du buste est réalisée par l'algorithme suivant :

$i \leftarrow 0$;
$k \leftarrow 0$;
$\overrightarrow{D^t_{\mathrm{Buste}}}(0) \leftarrow \mathrm{C}^t - \mathrm{H}^{t-1}$;
Répéter
  $\mathcal{V}^t_{\mathrm{Buste}}(i) \leftarrow \mathcal{V}^t_{\mathrm{Actif}} \cap \mathcal{C}yl(i)$;
  $\overrightarrow{D^t_{\mathrm{Buste}}}(i+1) \leftarrow \mathrm{C}^t - CdG(\mathcal{V}^t_{\mathrm{Buste}}(i))$;
  $i++$;
jusqu'à ce que ($||CdG(\mathcal{V}^t_{\mathrm{Buste}}(i)) - CdG(\mathcal{V}^t_{\mathrm{Buste}}(i-1))|| < \epsilon_{Buste}$)
$k \leftarrow i - 1$;
$\mathcal{V}^t_{\mathrm{Buste}} \leftarrow \mathcal{V}^t_{\mathrm{Buste}}(k)$;
$\mathcal{V}^t_{\mathrm{Actif}} \leftarrow \mathcal{V}^t_{\mathrm{Actif}} - \mathcal{V}^t_{\mathrm{Buste}}$;

Algorithme 7: Algorithme de recalage de cylindre



(a) (b) (c)

FIG. 11.5 – Suivi des avant-bras dans différentes configurations possibles : (a) représente une configurations ou les avant-bras ne se touchent pas,(b) présente une configuration pour laquelle ils se touchent et (c) représente une configuration ou les mains sont en contact.

La position du centre des hanches H^t est estimé par le centre de la seconde base de $\mathcal{C}yl(k)$ (voir la figure 11.4(a)). Le centre de la colonne vertébrale est estimé par le barycentre des voxels intersectés par un cylindre d'axe $\overrightarrow{D^t_{\text{Buste}}}$, centré sur le barycentre de $\mathcal{V}^t_{\text{Buste}}$(k), de hauteur $L_{\text{Buste}}/4$ et de diamètre L_{Epaule} (voir la figure 11.4(b)).

11.3.3 Suivi des mains et des avant-bras

Nous commençons par estimer la position des mains à partir de l'ensemble des voxels de peau. A l'aide de la longueur des avant-bras, nous déterminons ensuite la position des coudes. La cohérence temporelle est ensuite utilisée afin d'en déterminer le côté.

Soit $\mathcal{V}^t_{\text{M}}$ l'ensemble des voxels potentiels des mains. $L_{\text{Stature}}/2$ est un majorant de la longueur d'un bras. $\mathcal{V}^t_{\text{Peau}}$ contient les voxels des mains ainsi $\mathcal{V}^t_{\text{M}}$ est estimé par l'ensemble des voxels de $\mathcal{V}^t_{\text{Peau}}$ qui sont intersectés par une boule centrée en C^t de rayon $L_{\text{Stature}}/2$. Plusieurs cas se dégagent, en fonction du nombre de composantes connexes contenues dans $\mathcal{V}^t_{\text{M}}$ (voir figure 11.5).

Deux mains distinctes

Lorsque l'ensemble $\mathcal{V}^t_{\text{M}}$ contient au moins deux composantes connexes, il est possible d'extraire les positions des deux mains. Les deux composantes connexes $\mathcal{V}^t_{\text{M1}}$ et $\mathcal{V}^t_{\text{M2}}$ de volume maximum définissent leurs positions :

$$\text{M}^t_x = CdG(\mathcal{V}^t_{\text{M}_x}) \text{ avec } x \in [1,2]. \tag{11.3}$$

Connaissant une estimation de la longueur des avant-bras $L_{\text{AvantBras}}$, il est possible de déterminer les espaces où peuvent se situer chaque avant-bras. Pour un côté $x \in [1,2]$, nous déterminons l'ensemble de voxels de $\mathcal{V}^t_{\text{Actif}}$ qui peuvent représenter l'avant-bras du côté x :

$$\mathcal{V}^t_{\text{AvantBras}-pot_x} = \mathcal{V}^t_{\text{Actif}} \cap \mathcal{B}oule(\text{M}^t_x, L_{\text{AvantBras}}) \tag{11.4}$$

L'avant-bras du côté x est connecté à la main M_x^t. Ainsi cet avant-bras et cette main appartiennent à une même composante connexe de $\mathcal{V}^t_{\mathrm{AvantBras}-pot_x}$:

$$\mathcal{V}^t_{\mathrm{AvantBras}x} = cc_j \in \mathcal{V}^t_{\mathrm{AvantBras}-pot_x}, \mathcal{V}^t_{\mathrm{M}x} \subset cc_j. \tag{11.5}$$

Deux cas sont possibles :

- $\mathcal{V}^t_{\mathrm{AvantBras1}} \cap \mathcal{V}^t_{\mathrm{AvantBras2}} = \emptyset$, alors les avant-bras ne sont pas en contact entre eux. Nous calculons la position de chaque coude C_x^t à une distance $L_{\mathrm{AvantBras}}$ de la main M_x^t, porté par l'axe principal de $\mathcal{E}_{\mathcal{V}^t_{\mathrm{AvantBras}x}}$. Les côtés gauches et droits sont calculés en utilisant le critère de cohérence temporelle : le côté de l'avant-bras x est le même que l'avant-bras le plus proche calculé au temps $t-1$. Cette configuration est présentée figure 11.5(a).
- Dans le cas contraire les avant-bras sont en contact. Dans cette configuration, nous commençons par identifier le côté de chaque main en utilisant la propriété de longueur constante des avant-bras. M_x^t est du côté droit si

$$|d(\mathrm{M}_x^t, \mathrm{C}_d^{t-1}) - L_{\mathrm{AvantBras}}| < |d(\mathrm{M}_x^t, \mathrm{C}_g^{t-1}) - L_{\mathrm{AvantBras}}|, \tag{11.6}$$

  sinon M_x^t est du côté gauche.
  Les voxels de $\mathcal{V}^t_{\mathrm{AvantBras1}} \cap \mathcal{V}^t_{\mathrm{AvantBras2}}$ sont segmentés en deux sous-ensembles $\mathcal{V}^t_{\mathrm{AvantBras}d}$ et $\mathcal{V}^t_{\mathrm{AvantBras}g}$. Pour chaque voxel v_i, s'il est plus proche $[\mathrm{M}_d^t\mathrm{C}_d^{t-1}]$ que de $[\mathrm{M}_g^t\mathrm{C}_g^{t-1}]$ alors v_i est ajouté à $\mathcal{V}^t_{\mathrm{AvantBras}d}$, sinon v_i est ajouté à $\mathcal{V}^t_{\mathrm{AvantBras}g}$. L'axe principal de $\mathcal{E}_{\mathcal{V}^t_{\mathrm{AvantBras}d}}$, de $\mathcal{E}_{\mathcal{V}^t_{\mathrm{AvantBras}g}}$ et $L_{\mathrm{AvantBras}}$ sont utilisés pour calculer les positions des coudes C_d^t et C_g^t. Une configuration de ce type est présentée figure 11.5(b).

Une main, ou deux mains jointes

Lorsque $\mathcal{V}^t_{\mathrm{M}}$ ne contient qu'une seule composante connexe, alors celle-ci correspond soit aux deux mains en contact, soit à une seule main (c'est à dire que l'autre main n'est pas visible). Nous utilisons la cohérence temporelle afin de déterminer dans quel cas nous nous trouvons.

- Si ${\mathrm{M}^{t-1}}_d$ et ${\mathrm{M}^{t-1}}_g$ sont proches de $\mathcal{V}^t_{\mathrm{M}}$, alors les mains sont jointes (voir figure 11.5(c)). Ainsi ${\mathrm{M}^t}_d = {\mathrm{M}^t}_g = CdG(\mathcal{V}^t_{\mathrm{M}})$. $\mathcal{V}^t_{\mathrm{AvantBras}}$ est calculé comme proposé précédemment. Nous segmentons $\mathcal{V}^t_{\mathrm{AvantBras}}$ en deux parties $\mathcal{V}^t_{\mathrm{AvantBras}d}$ et $\mathcal{V}^t_{\mathrm{AvantBras}g}$ selon le plan orthogonal à $[\mathrm{C}_d^{t-1}\mathrm{C}_g^{t-1}]$ contenant ${\mathrm{M}^t}_d$. $L_{\mathrm{AvantBras}}$ et les axes principaux de $\mathcal{E}_{\mathcal{V}^t_{\mathrm{AvantBras}d}}$ et $\mathcal{E}_{\mathcal{V}^t_{\mathrm{AvantBras}g}}$ sont utilisés afin de calculer les coudes C_d^t et C_g^t.
- Dans le cas contraire, la main la plus proche ${\mathrm{M}^{t-1}}_x$ de $\mathcal{V}^t_{\mathrm{M}}$ est utilisée afin de déterminer le côté de ${\mathrm{M}^t}_x$, et ${\mathrm{M}^t}_x = CdG(\mathcal{V}^t_{\mathrm{M}})$. Nous calculons $\mathcal{V}^t_{\mathrm{AvantBras}}$ comme proposé précédemment et son axe principal d'inertie est utilisé pour calculer ${\mathrm{C}^t}_x$. La main la plus distante de $\mathcal{V}^t_{\mathrm{M}}$ ne peut être estimée au temps t. Dans ce cas nous conservons la position calculée au temps $t-1$.

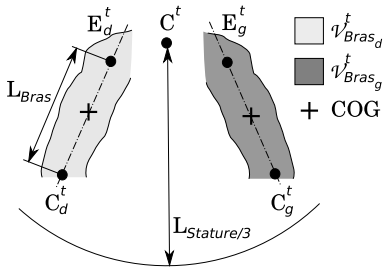


FIG. 11.6 – Illustration de l'algorithme de suivi des épaules et des bras.

Aucune main visible

Si $\mathcal{V}^t_{\mathrm{M}}$ est vide, alors aucune main n'est visible. Dans ce cas nous conservons les positions calculées au temps $t-1$.

Dans tous les cas, $\mathcal{V}^t_{\mathrm{Actif}}$ est mis à jour en supprimant les voxels appartenant aux avant-bras et aux mains.

11.3.4 Suivi des épaules

Nous avons déterminé les positions des articulations de la tête, du torse et des coudes. Nous finalisons le suivi des articulations hautes du corps en déterminant la position des épaules. Les contraintes de longueurs du corps humain imposent que les bras soient contenus dans la boule de rayon $L_{\mathrm{Stature}}/3$ centrée sur le cou. L'intersection de $\mathcal{V}^t_{\mathrm{Actif}}$ par cette boule, contient les voxels appartenant aux bras et certains voxels de bruit. Soit $\mathcal{V}^t_{\mathrm{Bras}}$ cet ensemble de voxels.

Le coude est l'une des extrémités des bras, la seconde correspond à l'épaule. Nous connaissons la position courante des coudes C^t_d et C^t_g. Il est donc possible de déterminer les voxels appartenant aux bras. Pour un côté x, $\mathcal{V}^t_{\mathrm{Bras}x}$ est la composante connexe de $\mathcal{V}^t_{\mathrm{Bras}}$ la plus proche[5] de ${\mathrm{C}^t}_x$ (voir figure 11.6). L_{Bras} étant connue, la position courante de l'épaule ${\mathrm{E}^t}_x$ est donnée par :

$${\mathrm{E}^t}_x = {\mathrm{C}^t}_x + \frac{CdG(\mathcal{V}^t_{\mathrm{Bras}x}) - {\mathrm{C}^t}_x}{||CdG(\mathcal{V}^t_{\mathrm{Bras}x}) - {\mathrm{C}^t}_x||} L_{\mathrm{Bras}}. \tag{11.7}$$

$\mathcal{V}^t_{\mathrm{Actif}}$ est mis à jour en supprimant les éléments appartenant aux bras.

11.3.5 Suivi des jambes

Nous venons de déterminer toutes les articulations hautes du corps. $\mathcal{V}^t_{\mathrm{Actif}}$ étant mis à jour à chaque étape en supprimant les voxels utilisés, celui-ci ne contient plus que ceux associés aux jambes. Notre approche d'extraction des jambes est inspirée de la méthode *point to line mapping* souvent utilisée pour l'association d'un squelette d'animation à un maillage 3D [SHSI99]. Les

[5] en terme de distance euclidienne

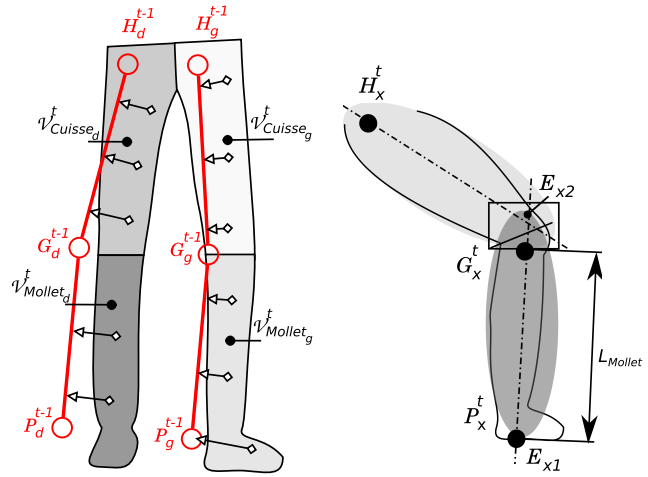


FIG. 11.7 – (a) étape du *binding* du suivi des jambes et illustration de l'estimation des articulations des jambes (b).

éléments de $\mathcal{V}^t_{\text{Actif}}$ sont séparés en quatre ensembles $\mathcal{V}^t_{\text{Cuisse}g}$, $\mathcal{V}^t_{\text{Mollet}g}$, $\mathcal{V}^t_{\text{Cuisse}d}$ et $\mathcal{V}^t_{\text{Mollet}d}$ en fonction de la distance de chaque voxel aux segments $[\text{H}_g^{t-1}, \text{G}_g^{t-1}]$, $[\text{G}_g^{t-1}, \text{P}_g^{t-1}]$, $[\text{H}_d^{t-1}, \text{G}_d^{t-1}]$, et $[\text{G}_d^{t-1}, \text{P}_d^{t-1}]$ (voir figure 11.7(a)). Pour chaque côté x, nous calculons l'ellipsoïde d'inertie $\mathcal{E}_{\mathcal{V}^t_{\text{Mollet}x}}$ (soient Ex_1 et Ex_2 ses points extrêmes) et l'ellipsoïde d'inertie $\mathcal{E}_{\mathcal{V}^t_{\text{Cuisse}x}}$.

Le genou est le point d'intersection entre la cuisse et le mollet (figure 11.7(b)), ainsi la position du pied $\text{P}^t{}_x$ est estimée par le point extrême de $\mathcal{E}_{\mathcal{V}^t_{\text{Mollet}x}}$ le plus distant de $\mathcal{V}^t_{\text{Cuisse}x}$. Le genou est aligné sur $[Ex_1Ex_2]$ du côté opposé à $\text{P}^t{}_x$ et à une distance de L_{Mollet} de ce point. La hanche correspondante $\text{H}^t{}_x$ est estimée par le point extrême $\mathcal{E}_{\mathcal{V}^t_{\text{Cuisse}x}}$ le plus distant de $\mathcal{V}^t_{\text{Mollet}x}$, corrigée pour être à une distance L_{Cuisse} du genou $\text{G}^t{}_x$.

11.4 Initialisation

Nous proposons une approche temps réel et automatique d'initialisation de suivi de mouvements, pour tout type de mouvements tant que la personne filmée est debout, que ses mains, non jointes, sont en dessous du niveau de la tête et que ses pieds ne sont pas joints. Après avoir estimé la taille de l'individu, notre approche calcule les paramètres de chaque partie du corps séquentiellement, dans le même ordre que la méthode de suivi. Comme précisé précédemment, les mêmes approches sont utilisées à la fois pour réaliser les étapes d'initialisation et de suivi. Les différences sont liées aux conditions utilisées. Dans la suite, nous précisons ces conditions pour chaque groupe d'articulations. Comme lors du suivi de mouvements, nous initialisons l'ensemble de voxels $\mathcal{V}_{\text{Actif}}$ par $\mathcal{V}_{\text{Tous}}$. Notons que si l'une des étapes suivantes échoue, alors le système tente de s'initialiser à la construction suivante.

Paramètres anthropométriques

Nous avons proposé dans le chapitre 10.2.2 un ensemble de rapports qui lient les mesures anthropométriques de certaines parties du corps, à la stature d'un individu. Ainsi il nous suffit de déterminer $L_{stature}$ la stature de la personne filmée, pour estimer l'ensemble des mesures anthropométriques. Celles-ci sont nécessaires afin d'utiliser l'approche l'extraction des articulations présentée dans la section précédente. Nous estimons le paramètre $L_{stature}$, par le voxel de $\mathcal{V}_{\text{Tous}}$ de plus haute altitude, lorsque $\mathcal{V}_{\text{Tous}}$ comporte un nombre d'éléments suffisant.

Initialisation de la tête

Nous avons pour objectif de déterminer les articulations T^0 et C^0 de la tête. En accord avec les hypothèses d'initialisation précisées ci-dessus, les voxels du visage du personnage filmé appartiennent à la composante connexe de $\mathcal{V}^0_{\text{Peau}}$ de plus haute altitude. Cette composante connexe est utilisée pour "simuler" la position précédente du visage $\mathcal{V}^{-1}_{\text{Visage}}$. L'approche de suivi de la tête présentée section 11.3.1 est utilisée afin de calculer T^0 et C^0.

Initialisation du torse

L'algorithme du suivi du torse présenté section 11.3.2 est appliqué en simulant la position précédente des hanches C^{-1} par $CdG(\mathcal{V}^0_{\text{Actif}})$. Les positions des hanches H^0 et du centre de la colonne vertébrale Cv^0 sont ensuite calculés en utilisant l'algorithme 7.

Initialisation des bras

Nous initialisons les positions des bras et des avant-bras en utilisant l'algorithme de suivi présenté section 11.3.3. Dans la mesure où nous n'avons pas de connaissance de la position précédente des bras, nous ne pouvons estimer les bras que lorsque les avant-bras sont distincts (ne sont pas en contact). Après avoir vérifié cette propriété, nous calculons ${\text{M}^0}_d$, ${\text{M}^0}_g$,${\text{C}^0}_d$ et ${\text{C}^0}_g$.

Initialisation des épaules

L'algorithme de suivi des épaules proposé section 11.3.4 ne nécessite pas de connaître la posture de la position précédente. Ainsi nous l'utilisons directement afin de calculer la position des épaules ${\text{E}^0}_d$ et ${\text{E}^0}_g$.

Initialisation des jambes

L'algorithme de suivi des jambes présenté section 11.3.5 nécessite la connaissance de leur position précédente. Nous proposons de simuler cette précédente information, en estimant de façon

grossière la position des genoux, des pieds et des hanches. Ensuite nous utilisons l'algorithme de suivi afin de calculer leur position précise.

$\mathcal{V}^0_{\text{Actif}}$ contient les voxels qui n'ont pas été associés aux autres parties du corps. Nous avons supposé le sujet filmé en appui sur ces pieds, qui se trouvent en contact avec le sol. Pour cette raison nous commençons par calculer l'ensemble de composantes connexes des voxels de $\mathcal{V}^0_{\text{Actif}}$ ayant une altitude inférieure à $L_{\text{Stature}}/8$. Si il y a moins de deux composantes connexes, cela provient du fait que les pieds sont en contact et qu'ils ne peuvent être distingués. Dans le cas contraire nous utilisons les deux composantes principales $\mathcal{V}^0_{\text{P}d}$ et $\mathcal{V}^0_{\text{P}g}$. Les genoux, les pieds et les hanches sont estimés par les équations suivantes :

$$\text{G}^{-1}{}_x = \text{H}^0 + v_x \frac{L_{\text{Cuisse}}}{||CdG(\mathcal{V}^0_{\text{P}x}) - \text{H}^0||}, \tag{11.8}$$

$$\text{P}^{-1}{}_x = \text{H}^0 + v_x \frac{L_{\text{Cuisse}} + L_{\text{Mollet}}}{||CdG(\mathcal{V}^0_{\text{P}x}) - \text{H}^0||}, \tag{11.9}$$

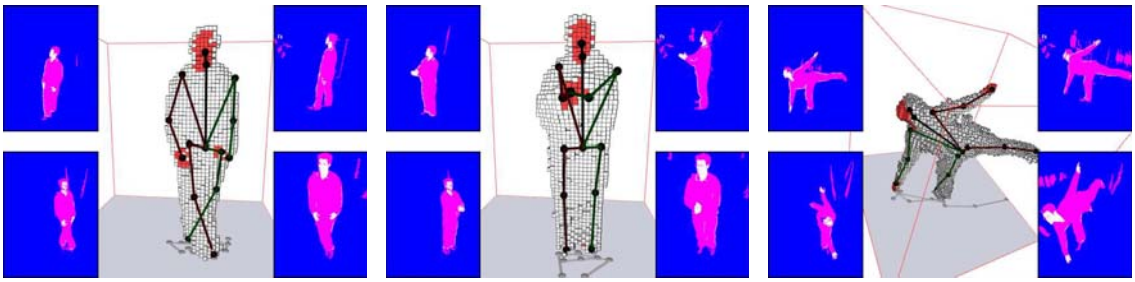
$$\text{H}^{-1}{}_x = \text{H}^0. \tag{11.10}$$

Finalement nous calculons $\text{P}^0{}_d$, $\text{G}^0{}_d$, $\text{P}^0{}_g$ et $\text{G}^0{}_g$ en utilisant l'algorithme de suivi des jambes présenté section 11.3.5.

11.5 Résultats

Dans cette section, nous présentons les résultats expérimentaux de cette approche réalisés à partir de données réelles. Pour une analyse quantitative ce cette approche, le chapitre 13 propose une étude comparative des approches développées dans la seconde partie de la thèse. L'infrastructure d'acquisition est composée de 4 webcams Phillips (SPC900NC) toutes connectées à un seul ordinateur usuel (CPU : P4 3.2Ghz,GPU : NVIDIA Quadro 3450) qui réalise la soustraction de fond, l'extraction des parties de peau, la reconstruction et l'acquisition de mouvements. Les webcams utilisées produisent des images couleur d'une résolution de 320×240 pixels à une cadence de 30 images par seconde.

Notre approche a été appliquée à la capture de mouvements rapides et complexes. Par une analyse conjointe de la forme et des parties de peau, notre algorithme est capable d'extraire correctement des postures généralement jugées difficiles, telles que celle présentée figure 11.8(a). Ce type de posture est difficile car la topologie de la forme 3D n'est pas cohérente avec la topologie d'un humain. La cohérence temporelle est une des clés du succès de l'acquisition de la posture présentée figure 11.8(b). Cette configuration souligne la réponse de notre approche face à un cas de mains jointes (section 11.3.3). Une posture difficile est montrée figure 11.8(c) et est parfaitement extraite par notre système. Les images présentées figure 11.9 montrent les résultats obtenus pour une grande variété de mouvements.



(a) (b) (c)

FIG. 11.8 – Les images (a) (b) et (c) soulignent les résultats de l'approche pour des postures difficiles. La posture extraite est illustrée par un squelette d'animation dont le côté droit est représenté en rouge, et le côté gauche en vert. Les voxels correspondants aux parties de peau sont repérés en rouge.

TAB. 11.1 – Comparaison des temps d'exécution de notre approche, avec les méthodes actuelles. Les données indiquées correspondent à celles publiées dans les articles originaux.

Réference	Nbr. d'articulations	Initialisation temps réel	Cadence	Nbr. de processeurs
Nous	17	Oui	$\approx$ 30 fps	1
[OSIK06]	21	Non	$\approx$ 25 fps	> 8
[CH04a]	15	Non	$\approx$ 15 fps	1
[TS06]	19	Non	$\approx$ 10 fps	1
[CGH08]	15	Non	$\approx$ 10 fps	1
[MBR06]	15	Non	$\approx$ 1 fps	1

Notre implantation actuelle de l'algorithme est capable d'acquérir plus de 35 postures par seconde sur le matériel décrit ci-dessus, soit une cadence supérieure à celle des webcams. Le tableau 11.1 montre quelques comparaisons des vitesses d'exécution avec les principales approches actuelles de suivi de mouvements sans marqueur rapides. Notre approche offre la meilleure cadence en utilisant un ordinateur mono-processeur. De plus cette approche se révèle être l'une des seules à offrir un système totalement automatique.

11.6 Discussion

Dans ce chapitre, nous avons décrit un système d'acquisition temps-réel et sans marqueur de mouvements, fonctionnant à partir d'un faible nombre de caméras sur un seul ordinateur usuel. Ce système est basé à la fois sur une analyse 3D de la forme, sur des contraintes morphologiques, ainsi que sur la connaissance des parties 3D de peau. Celle-ci s'avère totalement automatique. En combinant différents types d'informations, cette approche se révèle robuste face aux occulta-

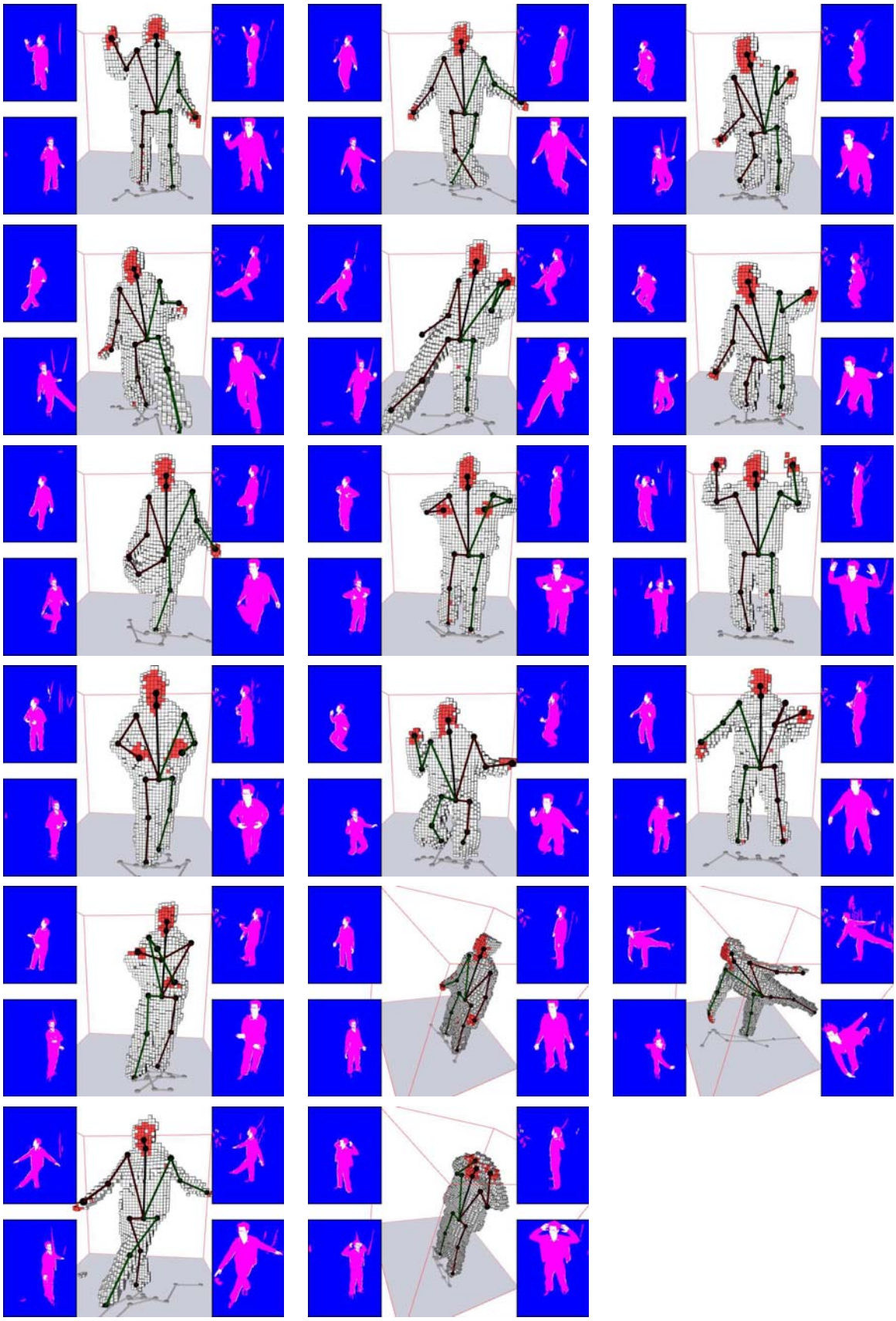


FIG. 11.9 – Résultats pour une grande variété de mouvements.

tions et aux auto-contacts. Elle estime les 17 principales articulation du corps humain à plus de 30 postures par seconde. Pour cette raison elle se révèle particulièrement adaptée à la mise en place de nouvelles interactions homme-machine. Même si elle est construite sur des heuristiques simples, les résultats obtenus montrent que l'approche s'avère robuste.

Grâce à l'utilisation des parties de peau, la problématique de l'extraction des extrémités hautes du corps a été partiellement résolue, sous réserve que l'apprentissage de la peau de la personne filmée ait été réalisé. Dans le cas contraire, le système n'est pas capable de suivre fidèlement les mouvements réalisés. De plus, nous observons que les extrémités basses du corps sont extraites de façon moins précise que les extrémités hautes. Afin de contourner ces deux limites, nous proposons dans la suite une méthode qui permet l'extraction des extrémités à partir des parties de peau, complétées des zones saillantes de la forme 3D. Nous montrerons que cette fusion d'information, couplée à un modèle simple du mouvement de chaque extrémité, permet une acquisition plus précise, avec un faible impact sur les temps de calcul.

Chapitre 12

Suivi conjoint d'ellipsoïdes saillantes et des parties de peau

Dans les chapitre précédents, nous avons proposé deux approches d'acquisition et de suivi du mouvement du corps d'un personnage, en utilisant l'information topologique de la forme 3D disponible ou l'information de couleur de peau. Ces deux approches amènent des outils complémentaires, qui une fois associés, peuvent permettre une capture efficace de la posture courante des personnages reconstruits.

Les approches construites sur l'extraction temps-réel de la topologie, souffrent en général d'une extraction non invariante par transformations rigides. Elles se révèlent souvent sensibles à l'échantillonnage des données et peu robustes au bruit. Pour cette raison nous nous intéressons désormais à l'extraction des parties saillantes de la forme plutôt qu'à toute l'information topologique disponible. Certaines postures, par exemple lorsque les bras sont alignés au corps, amènent à la disparition des parties saillantes de la forme 3D correspondantes à ces extrémités. Pour cette raison nous utiliserons l'information des parties 3D de peau, dont l'extraction est peu sensible aux contacts et aux occultations.
Dans ce chapitre nous proposons une approche d'acquisition de mouvements qui tire avantage de la complémentarité des parties saillantes de la forme, et des parties de peau. Cette méthode est construite en deux étapes (voir figure 12.1).

La première étape réalise la capture des extrémités du corps au cours du temps. En premier lieu, celle-ci détermine un ensemble de points caractéristiques des extrémités du corps à partir d'une estimation des zones saillantes dans le géométrie et des parties de peau. Une modélisation des mouvements, associée à un filtre de kalman, permettent de déterminer les zones correspondantes aux extrémités du corps.

La seconde étape, construite sur une généralisation de la méthode *Point To Line Mapping*, permet de segmenter la forme 3D en régions, correspondantes à chaque partie rigide du corps. L'extraction des articulations intermédiaires est réalisée par la recherche de l'ellipsoïde représen-

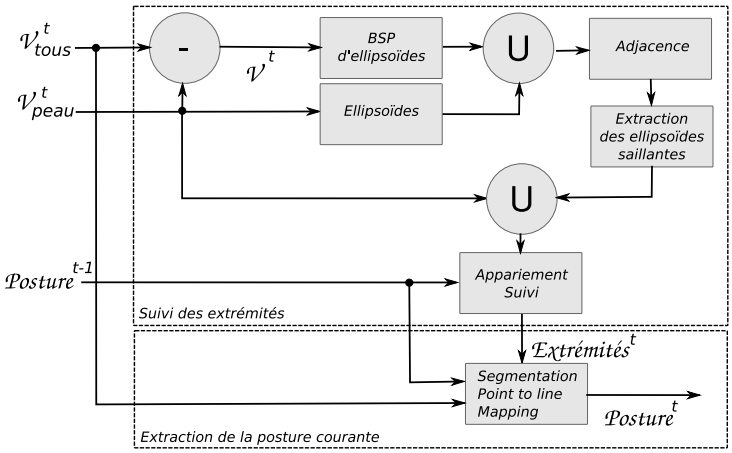


FIG. 12.1 – Vue d'ensemble du processus d'acquisition de mouvements exposé dans ce chapitre.

tant chaque région. Nous montrerons les mécanismes mis en place pour pallier aux cas critiques d'auto-occulation par exemple.

Nous présentons ensuite une généralisation de ces travaux pour l'acquisition simultanée du mouvement de plusieurs personnes en temps réel, même en présence de contacts entre ces personnes. Grâce aux contributions présentées dans la partie 1 de ce document, la forme estimée est vide d'objets fantômes. En présence de cette géométrie pertinente, l'utilisation d'un filtre de Kalman permet le suivi robuste du déplacement global de chaque personnage. Lors de contacts entre participants, une procédure de coupe permet la séparation de la géométrie afin de fournir des ensembles distincts associés à chaque personnage. Ceux-ci sont ensuite utilisés pour le suivi du mouvement respectif de chaque sujet.

Dans ce chapitre, nous nous focalisons sur la composante suivi du processus d'acquisition de mouvements. En effet nous considérons les procédures d'initialisation présentées aux chapitres 10.2.2 et 11.4 suffisamment précises, pour fournir une première estimation de la posture du sujet.

12.1 Suivi des extrémités du corps

Cette première partie réalise l'acquisition des positions des extrémités du corps au cours du temps (voir figure 12.1). Nous commençons par estimer la géométrie de la personne filmée par un ensemble d'ellipsoïdes, qui offre une représentation compacte. Parmi celles-ci nous extrayons les ellipsoïdes saillantes. Complétées par les ellipsoïdes des parties de peau, l'ensemble de ces ellipsoïdes décrit les zones d'intérêt qui semblent contenir les extrémité du corps. Nous traitons séparément des données relatives à la forme et aux parties de peau. En effet chaque composante connexe de peau définit une ellipsoïde particulière, qui par essence est d'intérêt *i.e.* elle représente certainement une extrémité du corps. Une modélisation des mouvements, associée à la continuité du mouvement humain, permettent de déterminer les zones d'intérêt correspondantes aux extrémités du corps. Dans la suite la reconstruction 3D de la forme au temps courant t est

notée $\mathcal{V}^t_{\text{Tous}}$. La reconstruction 3D des parties de peau est réalisée par la méthode proposée dans le chapitre précédent. Celles-ci sont représentées par l'ensemble $\mathcal{V}^t_{\text{Peau}}$.

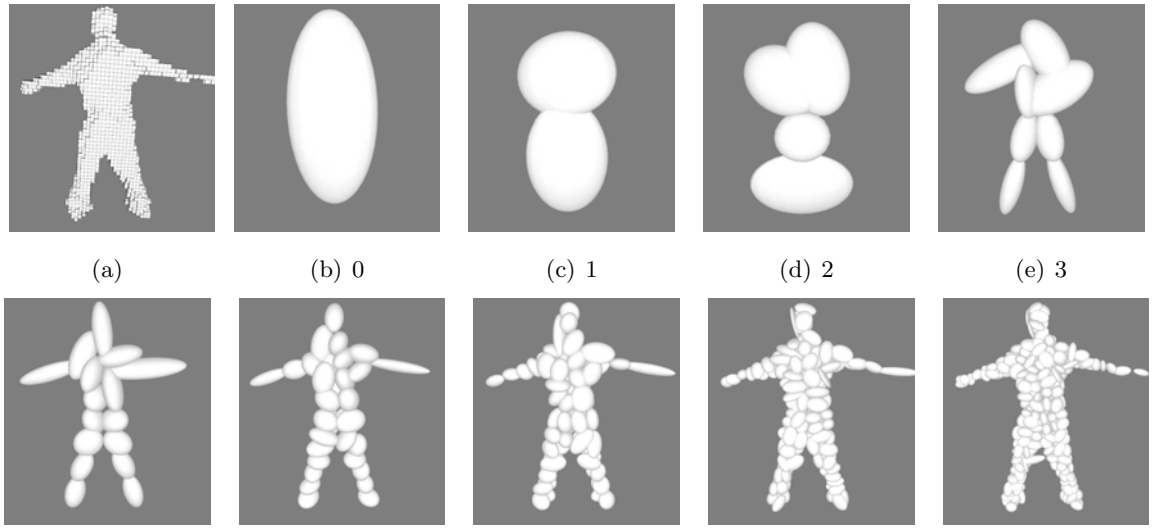
12.1.1 Segmentation de forme par d'ellipsoïdes

Dans les précédentes approches, nous nous sommes intéressé à l'extraction de la posture d'un personnage à partir de sa forme estimée par un ensemble de voxels. Il en résulte des approches offrant des résultats intéressants, mais dont la précision et les temps de calcul dépendent de la résolution de la grille d'échantillonnage. De plus, l'information par voxels peut se trouver bruitée à la fois par un extraction de silhouette imparfaite, mais aussi par la présence de parties fantômes connectées à la géométrie réelle, dont la quantité est inhérente au nombre de vues utilisées pour la reconstruction. Nous avons opté pour une représentation de la géométrie par un ensemble d'ellipsoïdes. Elles définissent une représentation compacte de l'information spatiale, tout en minimisant le nombre de primitives nécessaires. Cette structure représente uniquement les caractéristiques topologiques globales de cette forme, plutôt que les caractéristiques locales généralement issues des bruits de la reconstruction. Enfin cette représentation est indépendante de la résolution d'origine.

Certains travaux, notamment ceux présentés par de Aguiar *et al.* dans [dATM+04] et par Theobalt *et al.* dans [TdAM+04] sont construits sur une idée semblable. Leur travaux focalisent sur une méthode décomposition/fusion de la forme 3D par une ensemble de quadriques ou d'ellipsoïdes afin de déterminer la structure articulaire de la forme. Ces approches calculent ensuite la mise en correspondance de chaque primitive pour la séquence entière d'acquisition. Notre approche n'utilise que l'approche de décomposition par ellipsoïdes, car nous connaissons la structure articulaire de la personne filmée. La méthode proposée par Cheung *et al.* dans [CKBH00] réalise la segmentation de la forme 3D par un ensemble fini d'ellipsoïdes, dont chacune représente le buste, les membres et la tête du personnage reconstruit. Leur approche ne tient cependant pas compte des contraintes de tailles liées au corps humain. Le segmentation de chaque partie du corps tombe en défaut lorsque les membres dont trop proches les uns des autres. Cette méthode ne fournit pas de bons résultats en présence d'auto-contacts. Néanmoins, elle offre les avantages de la simplicité et du temps réel, mais le choix d'un faible nombre d'ellipsoïdes impose un suivi grossier de chaque partie du corps.

Nous proposons de segmenter la forme 3D par une structure hiérarchique d'ellipsoïdes. En premier lieu, nous évaluons l'ellipsoïde d'inertie de la forme complète (voir annexe B). Chaque ellipsoïde est ensuite divisée récursivement suivant le plan perpendiculaire à leur axe principal. Cette subdivision s'arrête lorsque le niveau de profondeur souhaité est obtenu, générant ainsi un Bsp (*Binary Space Partition*) de la forme initiale par ellipsoïdes.

Comme nous l'avons précisé, les parties de peau définissent directement des ellipsoïdes d'intérêt. Après avoir calculé les composantes connexes de peau, chacune d'entre elles est estimée par une ellipsoïde qui, par défaut, est déclarée ellipsoïde d'intérêt. De façon à ne pas réaliser de



(a) (b) 0 (c) 1 (d) 2 (e) 3

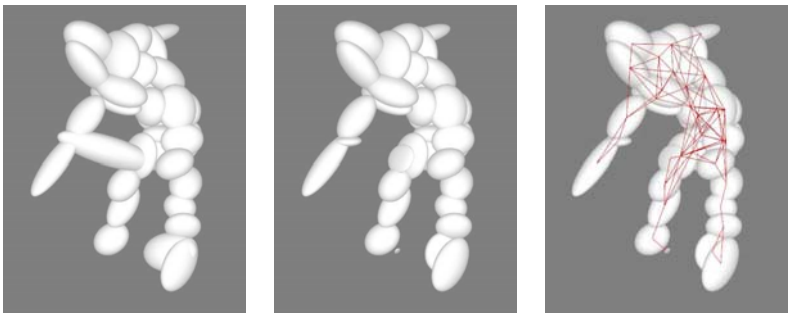
(f) 4 (g) 5 (h) 6 (i) 7 (j) 8

FIG. 12.2 – Segmentation de la forme d'origine (a) par ellipsoïdes pour plusieurs niveaux de récursion. Afin de faciliter la lecture, cette segmentation a été réalisée avec $\mathcal{V}^t = \mathcal{V}^t_{\text{Tous}}$.

doublons d'ellipsoïdes d'intérêt, la décomposition par BSP est réalisée uniquement sur le sous-ensemble des voxels non étiquetés peau : $\mathcal{V}^t = \mathcal{V}^t_{\text{Tous}} - \mathcal{V}^t_{\text{Peau}}$ Cette décomposition est réalisée indépendamment pour chaque forme $\mathcal{V}^t$ obtenue au temps t. La méthode de décomposition peut être exprimée en opérations élémentaires de la façon suivante :

1. La forme $\mathcal{V}^t$ est approximée par une ellipsoïde $\mathcal{E}_{\mathcal{V}^t}$.
2. Si la profondeur désirée est obtenue, la décomposition s'arrête. Sinon on continue à l'étape 3.
3. L'ensemble des voxels est partagé en deux sous-ensembles S_1 et S_2 par le plan orthogonal à l'axe principal de $\mathcal{E}_{\mathcal{V}^t}$, passant son centre.
4. S_1 et S_2 sont ensuite individuellement approximés chacun par une ellipsoïde. Pour chaque sous-ensemble, la procédure est répétée à partir de l'étape 2.

De cette façon nous obtenons un ensemble d'ellipsoïdes (Fig. 12.2) et les sous-ensembles de $\mathcal{V}^t$ correspondants. Après un nombre suffisant de subdivisions (typiquement 7), les points d'un même sous-ensemble de voxels appartiennent à une même partie rigide du corps du sujet suivi. Il est cependant possible que des sous-ensembles ne soient pas connexes, indiquant que les ellipsoïdes correspondantes puissent appartenir à plusieurs parties rigides non connectées (voir Fig. 12.3(a)). Nous finalisons cette segmentation en veillant à ce qu'une ellipsoïde soit associée à un sous-ensemble connexe de $\mathcal{V}^t$. Dans le cas contraire, chaque composante connexe est à son tour estimée par une nouvelle ellipsoïde (voir Fig. 12.3(b)).



(a) (b) (c)

FIG. 12.3 – Le mécanisme de découpe peut créer des sous-ensembles de voxels non connexes (a), qui donnent alors naissance à de nouvelles ellipsoïdes (b). (c) Notre représentation est étendue par la prise en compte des adjacences (représentées par des segments).

Nous associons désormais les ellipsoïdes issues de la segmentation de $\mathcal{V}^t$ et celles provenant des parties de peau 3D $\mathcal{V}^t_{\text{Peau}}$. Avec cette approche, la forme 3D est représentée par un ensemble de primitives (ellipsoïdes) beaucoup plus restreint que la représentation initiale (grille de voxels). L'espace de recherche pour l'extraction de la posture se trouve fortement réduit.

Nous souhaitons que cette représentation code aussi la topologie de la forme représentée. Pour cela nous calculons l'ensemble des relations d'adjacence entre les ellipsoïdes. Deux ellipsoïdes $\mathcal{E}_{\mathcal{V}^t_1}$ et $\mathcal{E}_{\mathcal{V}^t_2}$ sont liées par une relation d'adjacence, si et seulement si

$$\exists v_1 \in \mathcal{V}^t_1 \text{ et } \exists v_2 \in \mathcal{V}^t_2, d_{26}(v_1, v_2) = 1 \tag{12.1}$$

La figure 12.3(c) représente les adjacences entre les ellipsoïdes d'un même niveau de récursion. Finalement nous obtenons un ensemble d'ellipsoïdes (Fig. 12.3(b)), la liste d'adjacences entre ces ellipsoïdes (Fig. 12.3(c)) et les sous-ensembles de voxels correspondants. Cet ensemble segmente $\mathcal{V}^t$ afin de réduire l'espace de recherche des zones d'intérêt.

12.1.2 Extraction des zones d'intérêt

Nous focalisons sur l'extraction des extrémités du corps de la personne reconstruite. En général, ces extrémités coïncident avec les parties saillantes de la forme reconstruite. De plus, parmi les cinq extrémités du corps, la tête et les mains sont généralement découvertes, laissant apparaître des parties couleur chair. Ces deux types de caractéristiques définissent les zones d'intérêt à extraire, qui permettront par la suite le suivi des extrémités du corps. Nous commençons par détailler l'approche d'extraction des parties saillantes à partir de la représentation par ellipsoïdes de la forme 3D.

Les parties saillantes[6] d'un objet sont intuitivement identifiées comme étant les parties qui forment le relief de cet objet. Dans notre cas nous nous intéressons aux ellipsoïdes proéminentes

[6]Dépassant, étant en dehors, en formant un relief. *Définition provenant de l'encyclopédie Hachette*

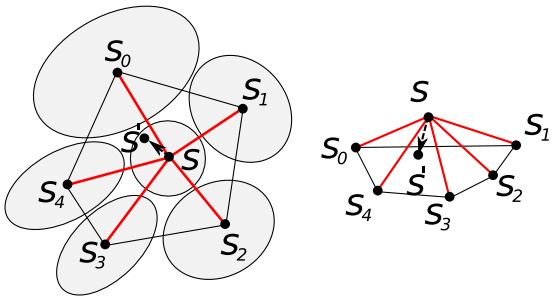


FIG. 12.4 – Principe de l'opérateur de contraction proposé, appliqué à un exemple simple, observé depuis deux points de vue différents. Par soucis de clarté, les ellipsoïdes ne sont pas représentées sur le schéma de droite.

de la forme. Dans la littérature, peu d'approches se sont intéressées à l'estimation des parties saillantes à partir d'une représentation géométrique volumique. Un ensemble important de travaux ont été proposé pour la détection de points d'intérêt dans des maillages surfaciques, en vue de morphing de maillages [Ale99], d'indexation d'objets [TVD08, ZTS02], de paramétrisation de surface [KS04, PSS01] ou encore de segmentation de maillage [KLT05, TVD07b].

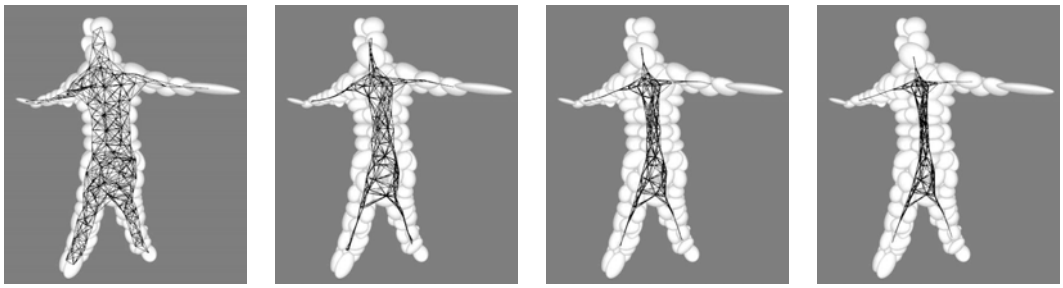
Dans [ZMT05, ZH04], les extrema locaux d'une fonction de distance géodésique définissent les points critiques. Cependant il est nécessaire de filtrer l'extraction des points provenant de variations très locale, nécessitant la définition d'un seuil par l'utilisateur. D'autres approches sont construites sur la contraction de maillage, permettant en général l'extraction du squelette topologique ainsi que de ses points d'intérêt proéminents [ATC+08].

Nous proposons une approche simplifiée des travaux présentés par Kin-Chung *et al.* [ATC+08].

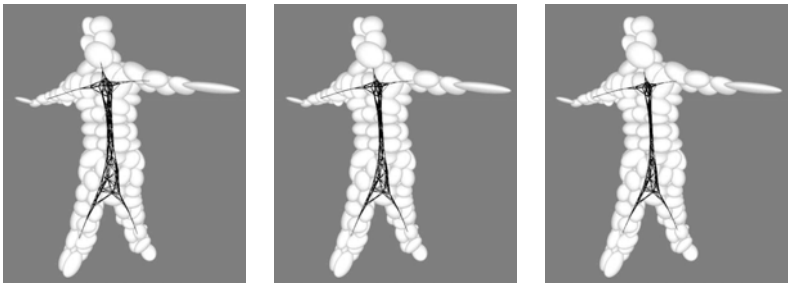
La représentation de la forme par ellipsoïdes calculée précédemment, définit un maillage volumique dont les sommets sont les centres des ellipsoïdes et les arêtes sont décrites par les relations d'adjacence entre ces ellipsoïdes. Nous réalisons une contraction de ce maillage. Un estimateur de rugosité locale permet ensuite d'extraire les groupes d'ellipsoïdes proéminentes.

Contraction de maillage

Notre approche est construite sur un processus de contraction géométrique dont l'objectif est de filtrer la position de chaque sommet de notre maillage volumique. Ce processus itératif définit la position s' filtrée d'un sommet s par une combinaison pondérée de ses sommets adjacents s_i. Cette opération revient généralement à définir la position de s' comme étant le barycentre pondéré, ou non, de ses sommets adjacents [KG00, Fie88]. Dans notre cas, la représentation choisie contient l'information de volume local de chaque ellipsoïde. Nous considérons que, plus le volume m_i de l'un des voisins s_i d'un sommet s est important, plus il influe sur la position filtrée s'. De plus cette influence est modulée par la distance euclidienne d_i de s à s_i. Nous définissons un opérateur de filtrage spatial d'un sommet s vers le sommet filtré s' en fonction de



(a) géométrie initiale (b) 1 itération (c) 2 itérations (d) 3 itérations



(e) 4 itérations (f) 5 itérations (g) 6 itérations

FIG. 12.5 – Filtrage d'une forme représentée un ensemble d'ellipsoïdes, par contraction géométrique ($\alpha = 0,5$).

son voisinage par :

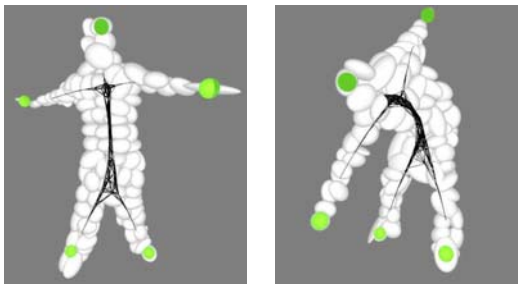
$$s' = \frac{\sum_i (\frac{\alpha}{d_i} + (1-\alpha)m_i)s_i}{\sum_i \frac{\alpha}{d_i} + (1-\alpha)m_i)} \tag{12.2}$$

avec $\alpha \in [0,1]$ un coefficient de pondération (voir figure 12.4).

La figure 12.5 présente le résultat de cet opérateur de contraction géométrique pour différentes valeurs d'itérations. Ce processus de contraction géométrique n'altère par la connectivité du maillage original. De plus, la contraction géométrique offre un filtrage implicite de la géométrie initiale. Ainsi ce type d'approche se révèle peu sensible au bruit. Celle-ci travaille directement à partir de la géométrie initiale, offrant de surcroît une méthode insensible à la translation, la rotation et à la posture.

Estimation locale de proéminence

Après avoir appliqué l'opérateur défini équation 12.2, afin de filtrer les hautes fréquences la forme initiale, nous cherchons à identifier les sommets saillants *i.e.* proéminents, de la forme filtrée. La méthode que nous proposons utilise à nouveau l'opérateur de contraction géométrique. En effet, celui-ci agit en supprimant la rugosité locale de chaque sommet à partir de l'information offerte par son voisinage. Si cette rugosité offre une amplitude suffisante, celle-ci subsistera même après lissage. Ainsi nous proposons de valider un sommet s comme étant saillant s'il satisfait la



(a) (b)

FIG. 12.6 – Représentation des ellipsoïdes saillantes détectées pour une valeur de $k = 0,8$.

condition suivante :

$$||s' - s|| > k\frac{\sum_i d_i}{\sum_i 1} \tag{12.3}$$

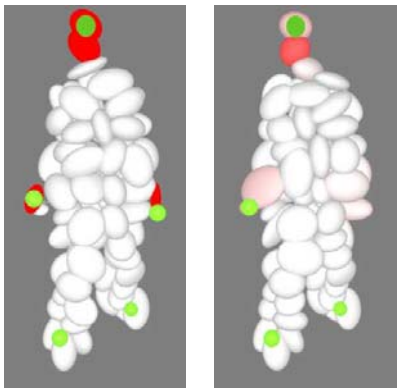
avec $k \in [0,1]$ un seuil. Notons que cette formulation est construite sur une variation prenant en compte la grandeur locale du déplacement. Le choix de k définit la quantité de saillance demandée. Une valeur typique de k est $0,8$.

La figure 12.6 illustre le mécanisme de détection des parties saillantes. En fonction de la profondeur choisie lors de la construction des ellipsoïdes, il est possible que plusieurs ellipsoïdes adjacentes soient validées par ce détecteur. Nous proposons de les fusionner de façon à ne conserver que les contributions globales. Cette fusion est réalisée de façon à conserver les propriétés de connexité de la reconstruction. Ainsi les ellipsoïdes saillantes connexes sont fusionnées en une seule ellipsoïde équivalente.

Enfin les ellipsoïdes jugées proéminentes forment l'ensemble des ellipsoïdes d'intérêt, utilisées par la suite pour le suivi des extrémités du corps. Les ellipsoïdes de peau, qui au vu des caractéristiques géométriques n'ont pas été détectées comme saillantes, sont ajoutées à la liste des ellipsoïdes d'intérêt.

Lorsque la topologie de la forme reconstruite est conforme à la topologie d'un humain, le critère géométrique présenté équation 12.3 est suffisant pour extraire toutes les extrémités (voir figure 12.6). Cependant, en cas de contacts entre les extrémités et d'autres parties le corps, celles-ci ne coïncident plus avec des parties saillantes de la forme. Afin de répondre à cette limitation, l'intégration des parties de peau – correspondantes à trois des cinq extrémités du corps – complète et rend plus robuste l'extraction d'ellipsoïdes d'intérêt. La figure 12.7 présente l'avantage procuré par l'utilisation de l'information de couleur de peau pour l'extraction des ellipsoïdes d'intérêt, par rapport à l'approche uniquement géométrique.

Dans cette section nous avons proposé plusieurs outils permettant l'extraction des régions d'intérêt de la forme qui correspondent aux extrémités du corps. Dans la suite nous détaillons les méthodes mises en place de façon à réaliser le suivi de ces extrémités de façon robuste et en temps réel.



(a) (b)

FIG. 12.7 – Représentation des ellipsoïdes saillantes détectées avec parties de peau (a) pour une valeur de $k = 0,8$. Observons que dans cette configuration, l'approche uniquement géométrique (b) ne parvient pas à déterminer toutes les ellipsoïdes d'intérêt.

12.1.3 Suivi des zones d'intérêt

Notre objectif est de réaliser le suivi des cinq extrémités du corps à partir des ellipsoïdes d'intérêt, pour ensuite extraire toute la chaîne articulaire de la personne filmée. Le nombre d'ellipsoïdes d'intérêt extraites n'est pas nécessairement constant. De plus les positions estimées ne se révèlent toujours très précises. Ceci est principalement du aux imprécisions provenant des caméras qui influent la précision de la forme, à des occultations, des ombres, ou encore à la modification d'apparence du sujet lors de mouvements. Quelle qu'en soit la source, les mesures varient de façon "aléatoire" par rapport aux données qui pourraient être offertes par un système idéal. Dans la suite, nous modéliserons l'ensemble des imprécisions obtenues comme étant un bruit additionnel au processus de suivi.

Nous souhaitons estimer le mouvement de chacune des extrémités de cette personne en utilisant toutes les mesures préalablement réalisées. L'utilisation conjointe de ces mesures doit permettre de déterminer de façon robuste la trajectoire de chaque extrémité. L'ingrédient clé est la définition d'un modèle de mouvements connu à priori (déplacement à accélération constante par exemple). En présence de ce modèle, nous pouvons déterminer ses paramètres correspondant aux observations, en sus de la position de chaque extrémité. Cette étape est usuellement divisée en deux phases (voir figure 12.8). La première phase, généralement appelée phase de *prédiction*, utilise l'information apprise précédemment pour affiner le modèle, afin de prévoir la prochaine position de chaque extrémité. La seconde phase, dite phase de *correction* a pour objectif de réconcilier la mesure effective, avec la prédiction construite à partir des mesures précédentes.

Les méthodes permettant ce type de processus en deux étapes sont généralement appelées estimateurs. Les filtres de Kalman et les méthodes de Monte-Carlo séquentielles sont les familles les plus répandues.

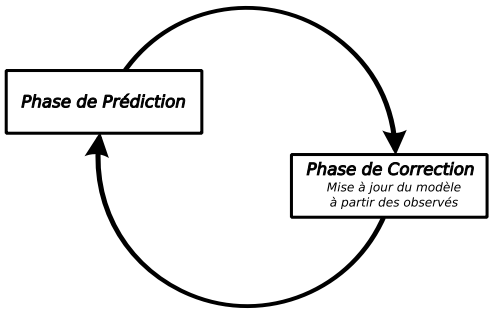


FIG. 12.8 – Cycle de Prédiction/Correction : La prédiction construite à partir des observations déjà réalisées suivie par la mise à jour du modèle à partir d'une nouvelle mesure.

- Le filtre de Kalman est un filtre récursif qui estime les états d'un système dynamique à partir d'une série de mesures altérées par un bruit supposé Gaussien. Il existe plusieurs filtres de Kalman permettant d'estimer différents types de systèmes dynamique. Le filtre de Kalman (KF) gère des systèmes linéaires [Kal60] alors que le filtre de Kalman étendu (Extended Kalman Filter, EKF) [May79] et le filtre de Kalman "sans parfum" (Unscented Kalman Filter, UKF) [JU97] permettent l'estimation de systèmes non-linéaires. Ces différents filtres modélisent cependant une seule hypothèse. En effet le modèle de la distribution de probabilité de l'hypothèse est une gaussienne, il n'est donc pas possible de représenter simultanément plusieurs hypothèses.

- Les méthodes de Monte-Carlo séquentielles offrent une solution optimale aux problèmes d'estimation des états d'une système non-linéaire à partir d'une série de mesures altérées par un bruit non nécessairement Gaussien. Ces méthodes sont généralement connues sous le nom de filtre à particules (particle filter) [GSS93] ou encore d'algorithme de Condensation [IB98]. Elles offrent une alternative aux filtres de Kalman étendu et "non parfumé", avec l'avantage, sous réserve d'un échantillonnage suffisant, d'approcher l'estimation bayesienne optimale. Elles peuvent se révéler plus précises que EKF ou UKF. Cette précision amène néanmoins un modèle calculatoire qui se révèle particulièrement coûteux. Avec l'augmentation de la puissance de calcul, ces méthodes ont été récemment utilisées pour des applications temps réel tels que le suivi des mains [STW07].

Notre objectif est l'acquisition de mouvements en temps réel. Le suivi des ellipsoïdes d'intérêt ne doit donc pas représenter l'un des verrous calculatoires de l'approche complète du suivi de mouvements. Nous faisons le choix d'utiliser un filtre de Kalman, plutôt qu'une approche de Monte-Carlo séquentielle qui pourrait offrir un suivi plus précis au détriment des temps de calcul. Le choix entre le filtre de kalman usuel (KF) et les filtres EKF et UKF dépend avant tout de la linéarité du modèle choisi pour le suivi. Commençons par expliciter de modèle le

mouvement utilisé.

Modèle de mouvements

Nous souhaitons déterminer la position $p(t + \triangle t)$ des extrémités du corps au temps $t + \triangle t$ à partir des observations précédentes. En addition, l'état associé à chaque extrémité inclut sa vitesse $v = (v_x, v_y, v_z)$ et son accélération $a = (a_x, a_y, a_z)$. L'état de chaque ellipsoïde est alors représenté par le vecteur $[p_x, p_y, p_z, v_x, v_y, v_z, a_x, a_y, a_z]^T$. La matrice de transition d'état, indiquant la transformation réalisée du vecteur d'état au temps $t + \triangle t$ à partir du vecteur d'état au temps t, provient du champ d'étude de la cinématique du point. Sous l'hypothèse d'un mouvement à accélération constante, le modèle de mouvements est décrit par :

$$p(t + \triangle t) = p(t) + v(t)\triangle t + a(t)\frac{\triangle t^2}{2}, \tag{12.4}$$

$$v(t + \triangle t) = v(t) + a(t)\triangle t, \tag{12.5}$$

$$a(t + \triangle t) = a(t). \tag{12.6}$$

avec p, v, a et $\triangle t$ respectivement la position, la vitesse, l'accélération et le temps écoulé entre deux mesures. Ces équations proviennent d'un développement de Taylor contenant seulement les termes du second ordre. Nous utilisons ce modèle pour décrire le mouvement de chaque extrémité du corps. Il définit une fonction linéaire, ce qui implique l'utilisation d'un filtre de Kalman. Pour une description plus détaillée, Welch et Bishop proposent une introduction complète au filtre de Kalman, en soulignant certaines de ses applications usuelles [WB01].

Pour le filtre de Kalman, la transition du vecteur d'état x de t à $t+1$ peut être exprimée par

$$x_{t+1} = Ax_t + w_t, \tag{12.7}$$

avec A la matrice de transition d'état et w_t un terme de bruit. Ce terme de bruit est une variable aléatoire gaussienne d'espérance nulle et de matrice de covariance Q. Sa distribution de probabilité est :

$$P(w) \sim N(0, Q). \tag{12.8}$$

Dans la suite la matrice de covariance Q sera appelée matrice de covariance du bruit du processus. Celle-ci modélise les changements possibles du vecteur d'état entre t et $t + 1$ qui ne sont pas pris en compte par la matrice de transition d'état. Il est supposé que w_t et x_t sont des variables indépendantes.

Dans la mesure où les reconstructions sont acquises à fréquence constante et où nous ne nous intéressons pas aux valeurs physiques de la vitesse et de l'accélération, nous supposons que

$\triangle t = 1$. Ainsi la matrice de transition A s'écrit :

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{12.9}$$

Pour le filtre usuel de Kalman, la relation entre le phénomène que l'on observe et son état est supposée linéaire. Ainsi l'observable y_t peut être définie à partir de l'état x_t :

$$y_t = Cx_t + u_t, \tag{12.10}$$

où C est une matrice $n \times m$ qui associe le vecteur d'état de dimension m au vecteur d'observation de dimension n. De la même façon que w_t l'est pour le processus, u_t décrit le bruit de mesure. Celui-ci est lui aussi supposé suivre une loi normale :

$$P(v) \sim N(0, R), \tag{12.11}$$

avec R la matrice de covariance appelée matrice de covariance du bruit de mesure.

Dans notre configuration, nous souhaitons suivre les extrémités du corps humain décrites par leurs coordonnées dans l'espace. C est défini par :

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{12.12}$$

Nous disposons désormais des données nécessaires à la prédiction du vecteur d'état associé à chaque extrémité du corps au temps $t + 1$. Une fois cette prédiction réalisée, il faut déterminer parmi les observations réalisées au temps $t+1$, les correspondances avec les extrémités modélisées, afin de réaliser la procédure de correction. Cette étape revient à réaliser la mise en correspondance entre les éléments suivis et les observables.

Mise en correspondance

Nous cherchons à déterminer parmi les observables *i.e.* les ellipsoïdes d'intérêt, lesquelles correspondent aux extrémités du corps. Le filtre de Kalman offre une prédiction de la position courante des extrémités notées $\tilde{y_i}$ avec $i \in [0, \cdots, n-1]$. Soient p_j avec $j \in [0, \cdots, m-1]$ les positions des m ellipsoïdes d'intérêt. Le choix des ellipsoïdes jugées correctes est réalisé sur

une mesure de similarité entre observable et prédiction construite. Nous pouvons utiliser la distance euclidienne ou encore la distance de Mahanalobis. Cette distance une mesure introduite par Mahalanobis [Mah36]. Elle est basée sur la corrélation entre des variables par lesquelles différents modèles peuvent être identifiés et analysés. Elle diffère de la distance euclidienne par le fait qu'elle prend en compte la corrélation de la série de données. Ainsi, à la différence de la distance euclidienne où toutes les composantes des vecteurs sont traitées de la même façon, la distance de Mahalanobis accorde un poids moins important aux composantes les plus bruitées (en supposant que chaque composante est une variable aléatoire de type gaussien). Nous utiliserons la distance de Mahalanobis (notée D_M) afin de prendre en compte les distributions estimées pour chaque point prédit. La distance de Mahalanobis d'une série de valeurs de moyenne $\mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_p)$ et de matrice de covariance Σ pour un vecteur à plusieurs variables $x = (x_1, x_2, x_3, \ldots, x_p)$ est définie par

$$D_M(x, \mu, \Sigma) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}. \tag{12.13}$$

Une approche intuitive pour réaliser la mise en correspondance est d'associer à chaque prédiction $\tilde{y_i}$ l'ellipsoïde d'intérêt de position p_j la plus proche, selon la distance de Mahalanobis. Cependant ce type d'approche directe favorise la fusion de plusieurs extrémités lors de contacts, qui ne pourront être séparées par la suite. Pour cette raison, nous proposons de réaliser un appariement injectif, c'est à dire qu'une ellipsoïde d'intérêt est associée à au plus une extrémité. La solution doit respecter ce critère tout en minimisant la somme des distances de Mahalanobis associées à chaque couple. Cette somme sera dans la suite appelée, par abus de langage, *Energie globale d'appariement* (E) définie par

$$E = \sum_{i=0}^{n-1} D_M(p_j, \mu_{\tilde{y_i}}, \Sigma_{\tilde{y_i}}), \tag{12.14}$$

avec $\tilde{y_i}$ apparié à Σ_{p_j}.

La réalisation d'un appariement injectif suppose que le nombre d'ellipsoïdes d'intérêt soit supérieur ou égal au nombre d'extrémités à suivre : $m \geq n$. Dans le cas contraire nous réaliserons les $m < n$ appariements minimisant E. L'état associé à chaque extrémité non appariée ne sera pas corrigé.

Il existe plusieurs types d'approches permettant de réaliser cette appariement en tenant compte des deux critères cités précédemment. La première consiste à générer l'ensemble de tous les appariements possibles, afin de déterminer ceux qui respectent le critère d'injection et l'énergie minimale. Supposons le nombre d'extrémités à suivre $n = 8$ et $m = 8$ observations. Une recherche exhaustive revient à calculer le nombre d'arrangement de 8 objets dans une grille de $8 \times 8 = 64$. Ainsi le nombre d'arrangements possibles est $A_8^{8^2} = \frac{64!}{56!} = 178462987637760$. Cette première solution brutale permet d'estimer exactement la ou les solutions au problème. Cependant le nombre d'opérations à exécuter se révèle trop important dans le contexte d'une résolution en temps réel.

$$Dist[5][6] = \begin{bmatrix} 0,840188 & 0,394383 & 0,783099 & 0,798440 & 0,911647 & \mathbf{0,197551} \\ 0,335223 & 0,768230 & 0,277775 & 0,553970 & \mathbf{0,477397} & 0,628871 \\ \mathbf{0,364784} & 0,513401 & 0,952230 & 0,916195 & 0,635712 & 0,717297 \\ 0,141603 & 0,606969 & \mathbf{0,016301} & 0,242887 & 0,137232 & 0,804177 \\ 0,156679 & 0,400944 & 0,129790 & \mathbf{0,108809} & 0,998924 & 0,218257 \end{bmatrix}$$

FIG. 12.9 – Matrice des distances entre $n = 5$ prédits avec $m = 6$ observés. Pour cet exemple, la solution injective d'énergie minimale $E_{min} \approx 1,09$ est indiquée en gras.

A l'opposé, la programmation par contraintes se révèle bien plus rapide pour ce problème. Un algorithme de "réparation itérative" nécessite un premier appariement, par exemple en associant chaque prédiction à l'observation la plus proche. Cette méthode estime ensuite le nombre de conflits et utilise une heuristique afin de les supprimer. La méthode de moindre conflit, supprime l'appariement le plus conflictuel (liant la prédiction $\tilde{y_i}$ à l'observé p_j) en le remplaçant par l'appariement (liant $\tilde{y_i}$ à l'observé p_k) le moins conflictuel. Cette approche permet de résoudre ce problème en un faible nombre d'étapes. Il permet par exemple de résoudre le *problème des 8 reines* généralement en moins de 50 étapes pour un échiquier d'un million de cases par côté. Malgré sa rapidité de calcul, cette approche ne garantit cependant pas de déterminer une solution. Cette dernière caractéristique nous interdit l'utilisation de ce type d'approche.

Si l'on prend en compte la connaissance que l'on a sur l'appariement recherché, il devient possible de construite de façon incrémentale un arbre de recherche dont chaque branche de solution peut-être invalidée relativement tôt dans la procédure de recherche, permettant de déterminer exactement toute solution au problème. Cet algorithme récursif de recherche avec retour arrière, calcule une solution au problème d'appariement sur une grille de $n \times m$ à partir d'une solution quelconque du problème de $(n-1) \times (m-1)$ appariements, par l'adjonction d'un nouveau couple de possibles. La récurrence commence avec la solution du problème de 1×1 appariements. L'algorithme réalisant cet appariement est présenté algorithme 8. La fonction *CréeConflit ?* vérifie si l'ajout d'un lien d'appariement entre la position prédite i et l'ellipsoïde d'intérêt j ne génère pas un conflit, *i.e.* un appariement non injectif. La procédure *ParcoursArbreRecherche* parcourt l'arbre de recherche, en coupant toute branche générant un conflit ou dont l'énergie supérieure est à l'énergie minimale de l'appariement d'une solution déjà évaluée E_{min}. Notons que dans le pire des cas, cette approche vérifiera $\frac{m!}{(m-n)!}$ solutions. Ainsi lorsque $n \leq m$, l'appel de la fonction *Apparie*(n, m) retourne la solution au problème d'appariement. La figure 12.9 représente un appariement réalisé par cette approche.

A cause de la contrainte d'injectivité, cet algorithme ne peut fournir de solution lorsque le nombre d'observables est trop faible, *i.e.* $n > m$. Alors, au lieu d'apparier les prédits aux observés, nous cherchons à apparier les observés aux prédits. Il suffit de remplacer la matrice des distances de chaque prédit vers chaque observée $Dist[N][M]$ par $Dist[M][N] = Dist[N][M]^T$.

Ensuite l'appel à la fonction Apparie(m, n) fournit le tableau contenant les appariement pour chaque observé. Notons que certains prédits ne sont pas appariés, et ne seront alors pas corrigés.

```
E_min ← +∞ ; [On initialise l'énergie d'appariement minimale]
Dist[N][M] ; [Matrice des distances entre N prédits et M observés]
Fonction CréeConflit ?(Lignes[M], n, m) : booléen
    i : entier ;
    Pour (i ← 1 ; i ≤ n ; i++) faire
        Si (Lignes[n − i] = m) Alors
            Retourner vrai ;
        Fin Si
    Fin Pour
    Retourner faux ;
Fin
Procédure ParcoursArbreRecherche(E_cour,Lignes[M], n)
    m : entier ;
    E_loc : réel ;
    Pour (m ← 0 ; m ≤ M ; m++) faire
        Si (CréeConflit ?(Lignes, n, m) = faux ET E_loc < E_min) Alors
            E_loc ← E_cour + Dist[n][m]
            Lignes[n] ← m ;
            Si (n = N − 1) Alors
                E_min ← E_loc ;
            Sinon
                ParcoursArbreRecherche(E_loc,Lignes, n + 1) ;
            Fin Si
        Fin Si
    Fin Pour
Fin
Fonction Apparie(n, m) : Tableau d'entiers
    Lignes[M] : Tableau d'entiers ;
    ParcoursArbreRecherche(0, 0,Lignes, 0) ;
    Retourner Lignes[M] ;
Fin
```

Algorithme 8: Algorithme du calcul de l'appariement optimal de N prédictions avec M observés : Injectif et Minimisant E. Il est supposé que $N \leq M$.

Procédure de correction

Nous venons de proposer une méthode d'appariement efficace offrant une mise en correspondance injective de chaque prédit avec un observé, tout en respectant le critère d'énergie minimale. Cette approche est construite sur l'hypothèse selon laquelle chaque extrémité réelle

de la forme est représentée par une ellipsoïde d'intérêt. Cependant il est possible que plusieurs extrémités soient très proches et soient ainsi représentées par une seule ellipsoïde d'intérêt. Il est de plus possible que l'une des extrémités de la personne ne soit pas détectée. Dans ces deux cas, il devient nécessaire de pondérer les mesures appariées afin de corriger l'état de la prédiction du filtre de Kalman.

Nous choisissons d'utiliser la valeur de distance de Mahalanobis pour réaliser cette pondération. En effet pour une prédiction de paramètres μ, Σ et un point x, $D_M(x, \mu, \Sigma) \leq 1$ indique que le point x est situé à l'intérieur de l'ellipsoïde d'inertie équivalente. Pour x en dehors de cette ellipsoïde, $D_M(x, \mu, \Sigma) > 1$. Cette mesure fournit alors directement un facteur de pondération permettant le mélange de la prédiction courante $\tilde{y_i}$ et de l'observé courant apparié p_j, définissant le vecteur de correction.

$$y_i = \begin{cases} p_j & \text{si } D_M(p_j, \mu_{\tilde{y_i}}, \Sigma_{\tilde{y_i}}) \leq 1 \\ \tilde{y_i} + \dfrac{(p_j - \tilde{y_i})}{D_M(p_j, \mu_{\tilde{y_i}}, \Sigma_{\tilde{y_i}})} & \text{sinon} \end{cases} \tag{12.15}$$

Notons que les paramètres de bruit du processus Q et du bruit de mesure R influent sur le comportement et les paramètres de chaque $\mu_{\tilde{y_i}}, \Sigma_{\tilde{y_i}}$.

Ainsi nous disposons de toutes les étapes nécessaires offrant le suivi de chaque extrémité du corps de la personne filmée, en temps réel. Nous nous intéressons désormais à l'extraction des articulations intermédiaires du corps.

12.2 Extraction des articulations

Ayant une estimation des positions des extrémités du corps, nous nous intéressons à l'extraction de l'ensemble des articulations principales de la personne reconstruite. Après l'extraction des articulations relatives aux épaules et aux hanches, nous étendons l'approche de "point to segment mapping" (voir section 11.3.5) afin de segmenter chaque partie rigide du corps dans la forme 3D. L'estimation des ellipsoïdes d'inerties correspondantes permettent d'extraire précisément et de façon robuste, la position de chaque articulation.

12.2.1 Estimation de la position des épaules et du bassin

Nous proposons d'estimer la position du centre des épaules et des hanches, à partir de l'extrémité correspondante au haut de la tête et du centre de gravité du corps.

Nous commençons par estimer la position du cou par la méthode proposée section 11.3.1, construite sur le recalage d'une sphère, dont le diamètre est défini par la longueur anthropométrique de la tête. L'objectif est de déterminer le centre de la tête afin de déterminer la seconde extrémité de la tête ; le cou. Le haut de la tête sert d'initialisation du centre de la sphère. Ensuite l'intersection de cette sphère avec l'ensemble des voxels indique le recouvrement réalisé. Le

centre de la sphère est ensuite décalé sur le centre de gravité de l'intersection entre la sphère et l'ensemble des voxels. Cette procédure est réitérée jusqu'à stabilisation. Cette première procédure offre une estimation de la position du cou, à son tour nécessaire à l'estimation du centre des épaules. Nous observons que le recalage par sphère offre une solution permettant d'approcher le centre d'une forme à tendance sphérique. Cette même procédure peut cependant définir un espace local de recherche à partir duquel le centrage est réalisé. C'est sur cette propriété du recalage par sphère que nous construisons la recherche du centre des épaules et du bassin.

Le vecteur cou - centre de gravité offre une estimation de la direction du cou vers le centre des épaules. La distance entre le cou et le centre des épaules est contrainte par une mesure anthropométrique permettant de réduire l'ensemble des solutions, comme appartenant à la surface d'une sphère centrée sur le cou, de rayon $L_{stature}/8$. Une première estimation du centre des épaules E est définie par le point du segment cou - centre de gravité, de distance $L_{stature}/8$ du cou. Nous appliquons ensuite l'algorithme de recalage d'une sphère de diamètre $L_{Epaules} = L_{stature}/5$ (majorant la largeur du buste), centrée sur E. A chaque étape, nous imposons que la distance séparant le cou du centre de cette sphère soit constante et égale à $L_{stature}/8$.

La procédure d'estimation du centre des hanches H est construite sur la même approche. Une première estimation de H est définie par le point appartenant au segment [CCdg] à une distance $L_{stature}/6$ de E. Le recalage de sphère utilisé pour estimer E est utilisé pour affiner l'estimation du centre des hanches H.

Cette étape vise à fournir une estimation des épaules et des hanches plutôt qu'une extraction précise de leur positions, qui sera réalisée par l'étape présentée dans la prochaine section. Pour cette raison nous proposons une approche qui présuppose que l'axe des épaule et des hanches appartiennent à des plans orthogonaux à la colonne vertébrale décrite par le segment [EH]. Nous connaissons une estimation anthropométrique de la largeur des hanches ($L_{Hanches} \approx L_{stature}/6$) et des épaules $L_{Epaules}$, les épaules (respectivement les hanches) se situent donc dans le cylindre centré en E (resp. H), d'axe [EH] et de hauteur inférieure à $L_{Buste}/4$. Soient $\mathcal{V}_{\mathrm{E}}$ et $\mathcal{V}_{\mathrm{H}}$ les ensembles de voxels de la reconstruction intersectés respectivement par ces deux cylindres. Nous calculons l'axe principal d'inertie des ensembles $\mathcal{V}_{\mathrm{E}}$ et $\mathcal{V}_{\mathrm{H}}$ afin d'estimer l'axe des épaules et des hanches. La direction de chaque axe est calculée en utilisant la direction gauche-droite connue pour ce même axe lors de la frame précédente. Nous utilisons le produit scalaire entre la direction associée à la frame précédente, et celle associée à la frame courante. Soient $Dir(\mathrm{E}_g, \mathrm{E}_d)$ et $Dir(h_g, h_d)$ ces vecteurs. Une estimation des épaules E_g et E_d ainsi que des hanches gauche

h_g et droite h_d est offerte par :

$$\mathrm{E}_g = \mathrm{E} + \frac{L_{Epaules} Dir(\mathrm{E}_g, \mathrm{E}_d)}{2||Dir(\mathrm{E}_g, \mathrm{E}_d)||} \tag{12.16}$$

$$\mathrm{E}_d = \mathrm{E} - \frac{L_{Epaules} Dir(\mathrm{E}_g, \mathrm{E}_d)}{2||Dir(\mathrm{E}_g, \mathrm{E}_d)||} \tag{12.17}$$

$$\mathrm{H}_d = \mathrm{H} + \frac{L_{Hanches} Dir(\mathrm{H}_g, \mathrm{H}_d)}{2||Dir(\mathrm{H}_g, \mathrm{H}_d)||} \tag{12.18}$$

$$\mathrm{H}_g = \mathrm{H} - \frac{L_{Hanches} Dir(\mathrm{H}_g, \mathrm{H}_d)}{2||Dir(\mathrm{H}_g, \mathrm{H}_d)||} \tag{12.19}$$

Nous avons calculé les positions des extrémités, ainsi qu'une estimation des articulations du buste. Nous proposons dans la suite une approche robuste permettant l'affinage des positions de ces extrémités, ainsi que l'extraction robuste de la position des articulations intermédiaires *i.e.* les coudes et les genoux.

12.2.2 Extraction robuste de l'ensemble des articulations

Notre objectif est de déterminer de façon robuste la position de chaque articulation principale à partir de la reconstruction 3D de la personne. Les approches usuelles sont généralement construites sur la minimisation d'une distance entre un modèle paramétrique du corps de la personne et les observables, qui dans notre cas s'apparente à la forme 3D estimée. Notre approche est construite sur une approche opposée qui vise à segmenter dans la forme 3D les zones correspondantes à chaque partie rigide du corps.

Nous avons proposé section 11.3.5 une approche permettant de segmenter les parties de la forme 3D n'ayant pas été associées aux articulations hautes du corps, ni au buste, afin de déterminer la position des jambes. Cette méthode dite de "point to segment mapping" réalise la segmentation de l'ensemble des voxels par plus proche partie rigide. C'est cette même approche que nous allons généraliser à un ensemble quelconque d'articulations.

Nous représentons chaque partie rigide du corps par un segment. Nous commençons par segmenter la forme 3D en n ensembles de voxels $\mathcal{V}_i$ associés à n parties rigides. Chaque ensemble $\mathcal{V}_i$ est constitué des voxels dont la distance au $i^{\text{ème}}$ segment S_i est inférieure à la distance à tout autre segment $S_j \in S$. Dans la suite, nous considérons la distance d'un point p à un segment, comme étant la distance euclidienne minimale de ce point à un point q de ce segment :

$$D(p, S_i) = \underset{q \in S_i}{Min} |p - q|. \tag{12.20}$$

Ainsi chaque ensemble $\mathcal{V}_i$ est formé des voxels v tels que :

$$\mathcal{V}_i = \{v \in \mathcal{V}^t | \forall S_j \in S, D(p_v, S_i) \leq D(p_v, S_j)\}. \tag{12.21}$$

Nous cherchons à segmenter le corps en régions représentant chaque partie rigide suivie. Pour cette raison il est nécessaire de représenter toutes les parties rigides, afin d'éviter des associations erronées telles que, par exemple l'association de voxels du buste à l'un des bras. Ainsi les

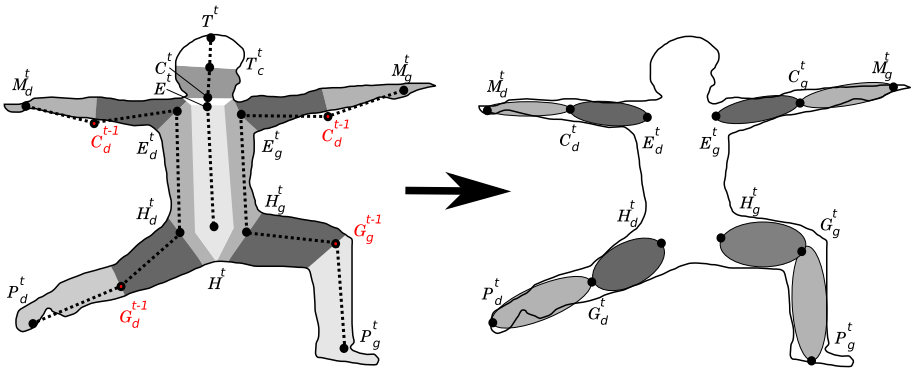


FIG. 12.10 – Determination des articulations intermédiaires : (à gauche) segmentation de la forme par l'approche *Point To Line Mapping* généralisée, (à droite) l'ellipsoïde correspondante à chaque région permet l'estimation des articulations intermédiaires et l'affinage des extrémités.

segments utilisés sont au nombre de 13 :

- 2 segments par bras : [M_x C_x] et [C_x E_x] avec $x \in \{g, d\}$ le côté considéré
- 2 segments par jambe : [P_x G_x] et [G_x H_x],
- 2 segments pour la tête : [C T_c] et [T_c T],
- 3 segments pour le buste : [E H], [E_g H_g] et [E_d H_d].

L'utilisation de 3 segments dans le buste permet de prévenir de l'association de voxels du buste à l'un des membres (voir figure 12.10). La distance étant uniforme pour tout segment, représenter le buste uniquement par le segment associé à la colonne vertébrale peut s'avérer insuffisant, surtout en présence d'une forme issue d'un faible nombre de caméras. Remarquons l'utilisation des positions des coudes et des genoux que nous n'avons pas encore estimées. Plusieurs alternatives sont possibles. La première consiste à utiliser leur position calculée lors au pas de temps précédent, au risque de ne pouvoir récupérer une erreur commise précédemment. A l'inverse, une estimation grossière de la position courante des articulations intermédiaires est donnée par le barycentre des articulations extrêmes de chaque membre. Au vu de nos expérimentations, cette seconde solution offre au système une robustesse aux erreurs de suivi.

Le correspondant d'une partie rigide dans la forme 3D représente un connexe. Cependant la méthode de reconstruction à partir de silhouette souffre de l'ajout de parties fantômes, qui peut se traduire par la présence de plusieurs composantes connexes dans certains ensembles $\mathcal{V}_i$. Afin de contourner cette difficulté, nous proposons de calculer pour chaque $\mathcal{V}_i$ l'ensemble des composantes connexes et les ellipsoïdes d'inertie associées. Ces dernières permettent de conserver uniquement la composante connexe correspondante à chaque articulation connue (pieds, mains, épaules, etc.). Ainsi seule la composante connexe de $\mathcal{V}_i$ dont l'ellipsoïde d'inertie est la plus proche

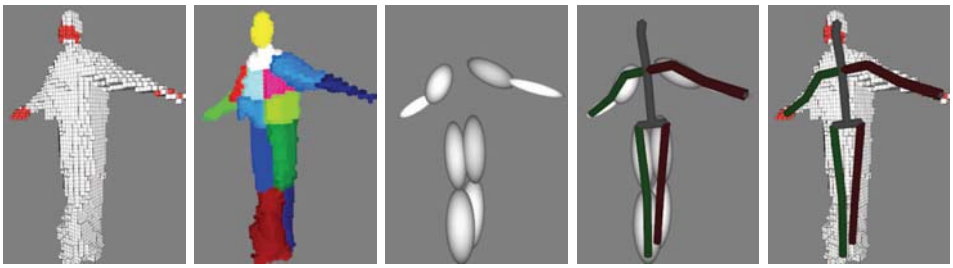


FIG. 12.11 – Les différentes étapes du processus d'extraction des articulations intermédiaires de gauche à droite : la forme 3D d'origine, la segmentation par plus proche segment en ensembles de voxels $\mathcal{V}_i$, la sélection de l'ellipsoïde représentative offrant l'estimation de chaque articulation intermédiaire puis raffinement de toutes les articulations.

(en terme de distance de Mahanalobis) de l'articulation connue du segment S_i est conservée. Les autres sont supprimées, étant considérées comme un bruit de reconstruction.Ces même ellipsoïdes fournissent une autre information essentielle : l'axe principal d'inertie des voxels associés à chaque partie rigide S_i (voir figure 12.10).

Chaque articulation intermédiaire est associée à deux parties rigides, dont on connaît désormais les ellipsoïdes d'inertie. Ainsi cette articulation se situe à proximité des extrémités les plus proches de ces deux ellipsoïdes. Les extrémités symétriques de ces ellipsoïdes sont utilisées afin d'affiner la position des extrémités de chaque membre. Les longueurs anthropométriques sont utilisées pour corriger la position de chaque articulation intermédiaire.

La figure 12.11 illustre les différentes étapes de l'extraction de la position des articulations principales.

Cette approche d'extraction et d'affinage de la position des articulations principales est robuste aux défauts provenant du calcul de la forme 3D de la personne filmée. En effet la segmentation de la forme est réalisée en prenant en compte la topologie de la structure articulaire, ainsi que la connaissance de la position des extrémités. Ceci permet de s'affranchir des artefacts fantômes, dont des branches topologiques ne correspondent pas aux extrémités suivies. La méthode de segmentation volumique permet de plus, d'offrir une approche résistante aux auto-contacts. En effet, cette approche minimise le recouvrement du volume par un ensemble d'ellipsoïdes contraintes par leur extrémités. Les parties du corps en contact sont correctement segmentées, sans utiliser de critères de cohérence liés à la courbure ou aux contours, incapables de détecter les auto-contacts, et souvent sensibles au bruit. Enfin l'extraction de l'axe principal de chaque segment rigide offre un mécanisme intrinsèque de filtrage de la géométrie calculée.

12.3 Résultats

Dans cette section nous présentons les résultats de cette approche issus d'expérimentations construites à partir de données réelles. Pour une étude quantitative le chapitre 13 présente une

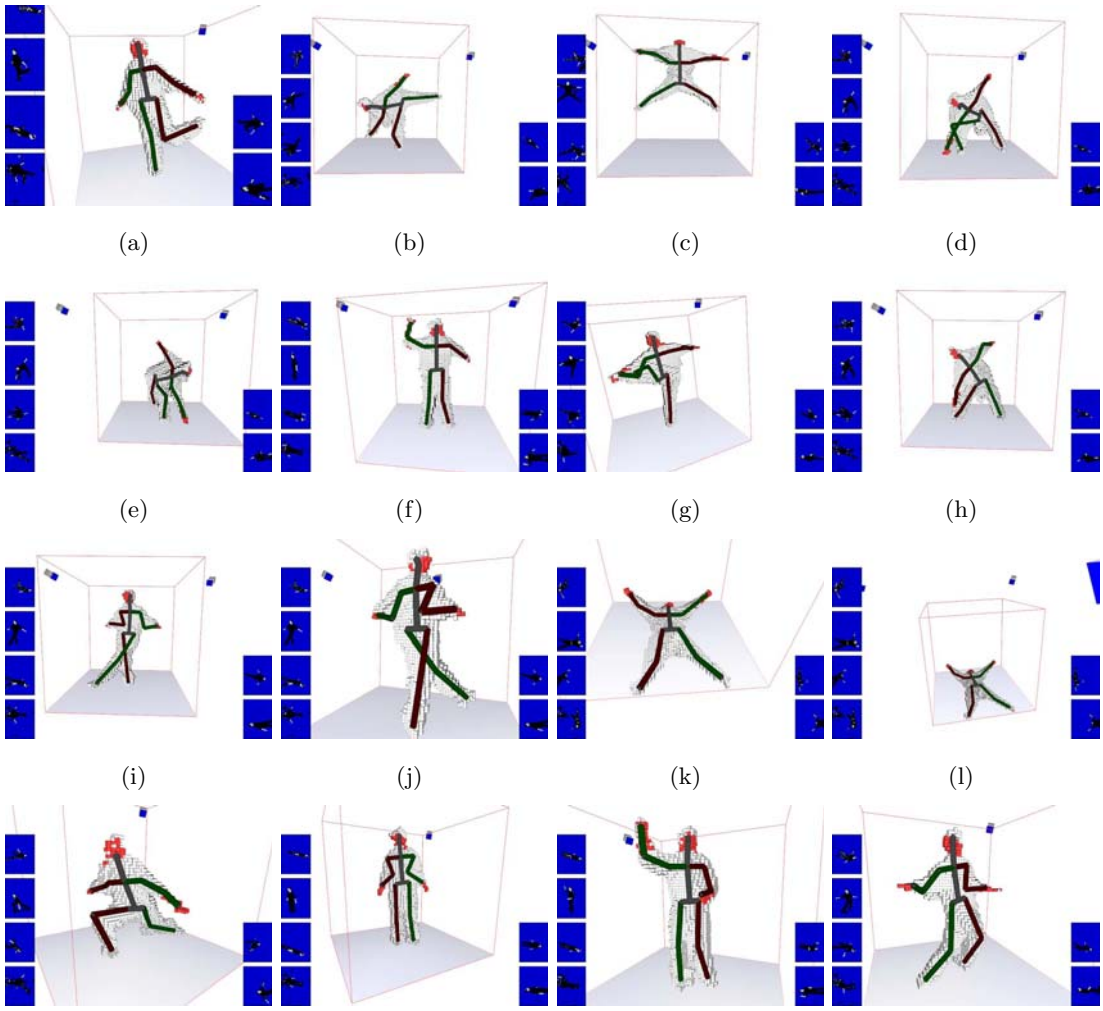


FIG. 12.12 – Résultats à partir d'une varité de mouvements complexes.

analyse comparative des différentes méthodes proposées.

Les expérimentations suivantes ont étés réalisées à partir d'une reconstruction de la forme 3D calculée à partir de 6 flux vidéos, définie sur une grille de 64^3 voxels échantillonnant un cube de $2m$ de côté. Chaque caméra fournit 30 images par seconde d'une résolution de 640×480 pixels. Chaque caméra est connectée à un ordinateur qui réalise la soustraction de fond et l'extraction des parties de peu. Ensuite ces informations sont envoyées par réseau à un unique ordinateur (CPU : Intel Core2quad @ 3Ghz, GPU Nvidia gtx260) qui réalise le calcul de la forme 3D et des parties de peau 3D, ainsi que l'acquisition du mouvement.

Les figures 12.12 et 12.13 présentent plusieurs acquisitions de mouvements complexes. L'ensemble des ces postures sont correctement extraites, même en présence de configurations extrêmes.

Par exemple, la figure 12.12(c) montre la capacité de cette approche à gérer les mouvements

rapides tels que des sauts. Notons que dans cette configuration, le fait que l'acteur sorte de l'espace reconstruit ne parvient pas à mettre cette approche en échec.

Les figures 12.12(d,g,h) soulignent l'efficacité de cette approche, même en présence de contacts francs entre bras et jambes. Enfin les figures 12.12(k,l) montrent que cette approche est insensible à l'orientation du corps de l'acteur, qui se trouve allongé sur le sol.

Sur l'ensemble des expérimentations présentées figures 12.12, les parties de peau sont correctement extraites. Ainsi seul le suivi des parties de peau semblerait suffire à l'extraction de la tête et des mains. Cependant le fait d'utiliser conjointement les parties de peaux et les zones saillantes offre une approche robuste aux auto-occultations, permettant l'acquisition robuste de mouvements tels que ceux présentés figures 12.13(g,h,i). L'extraction des articulations intermédiaires par segmentation de la forme 3D par ellipsoïdes, permet d'extraire des poses sans être perturbée par les contacts entre les différentes parties du corps (Figures 12.13(b,i,j,k,l,n,o,p)).

Nous observons que cette approche fournit des résultats d'une qualité supérieure aux approches précédentes. Elle se révèle particulièrement robuste face aux contacts, aux mouvements rapides, aux parties de peau mal extraites ou encore à l'orientation du corps. La mise en œuvre technique valide à nouveau la contrainte du temps réel, réalisant l'acquisition de plus de 40 mouvements par seconde, soit une cadence supérieure à celle des caméras.

12.4 Vers une généralisation à plusieurs personnages

Les approches d'acquisition de mouvements sans marqueur proposées précédemment permettent le suivi d'une seule personne. Nous ne nous étions alors pas orientés vers une approche muti-personnages, par appréhension des difficultés liées aux contacts entre ces personnes.

Les méthodes présentées chapitre 10 sont construites sur l'analyse de la topologie de la forme calculée. Lors de contacts entre les personnes reconstruites, la topologie change, mettant ce type d'approche en défaut. De plus les parties fantôme de la reconstruction 3D génère des branches topologiques fantômes, qui rendent ces méthodes instables.

La méthode présentée chapitre 11 utilise l'information de couleur de peau afin de guider un recalage d'objets géométriques simples dans la forme 3D calculée. Le recalage présenté est construit sur le centrage de ces objets simples dans l'ensemble de voxels calculé, prenant intrinsèquement en compte les frontières avec les parties vides de l'espace. Ainsi en présence de contacts entre les personnages, le recalage des parties de corps en contacts se trouve en difficulté. De plus le recalage au temps t dépend des positions estimées au temps $t-1$, au risque de répercuter une erreur aux frames suivantes, mettant en échec l'acquisition et le suivi de mouvements.

A l'opposé, l'approche présentée dans ce chapitre se révèle naturellement robuste aux contacts et aux parties fantômes tant que l'estimation des extrémités est correcte. Il est cependant nécessaire de pouvoir estimer la position du centre de gravité de chaque personnage, indirectement nécessaire à la segmentation de la forme 3D et à l'extraction des positions des articulations. Dans

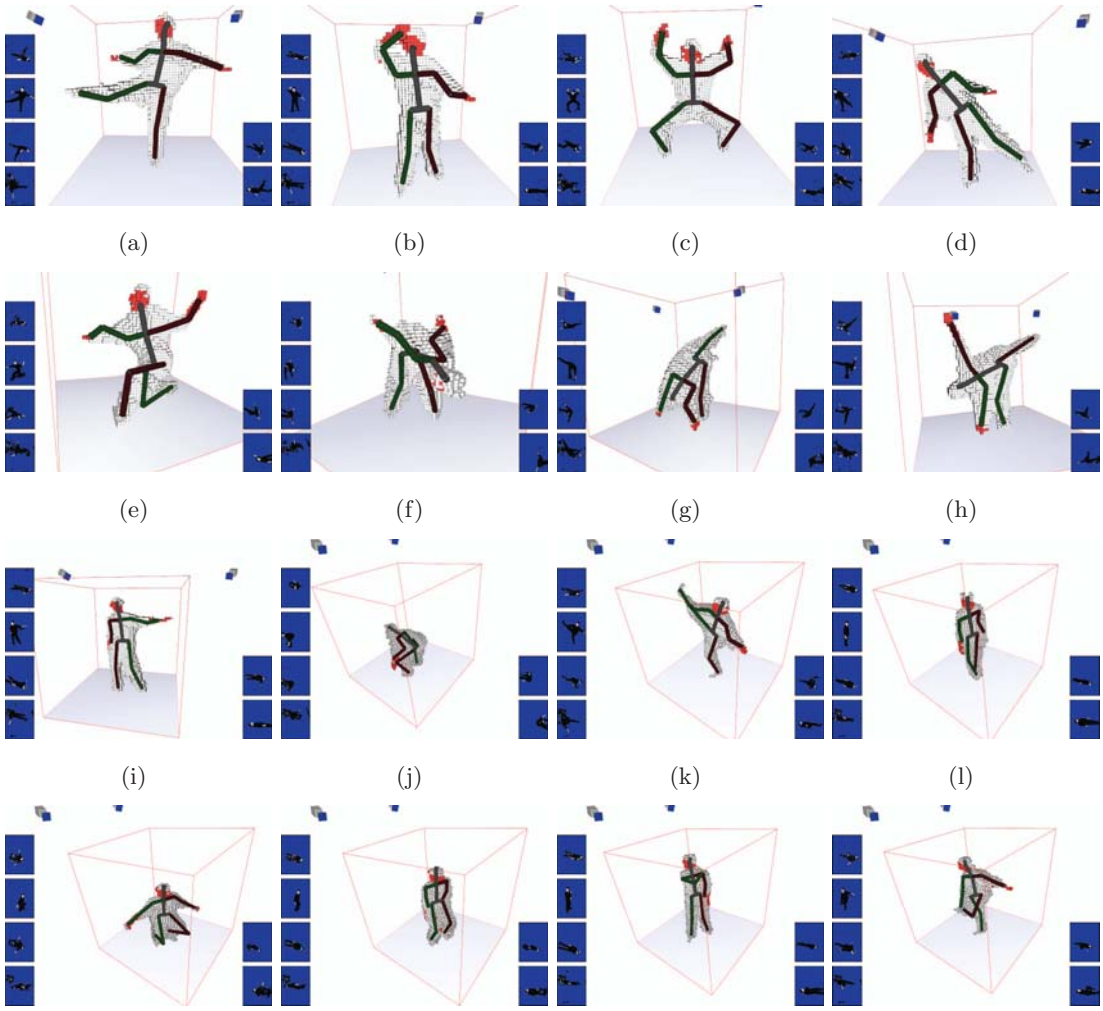


FIG. 12.13 – Acquisition de mouvements offerte par notre approche à partir de mouvements usuels.

la suite nous proposons une généralisation de cette approche en vue de réaliser l'acquisition du mouvement de plusieurs personnages, simultanément.

La littérature liée à la problématique d'acquisition de mouvements sans marqueur, concerne principalement le scénario mono-personnage. A notre connaissance, peu de travaux de la communauté sont relatifs à l'acquisition sans marqueur du mouvement de plusieurs personnages [GD96, LN09]. Ce faible nombre de contributions est principalement lié aux difficultés engendrées par les occultations entre personnages ou encore lors de contacts entre ceux-ci. Cependant une réponse partielle à ce problème peut-être trouvée dans les travaux issus du domaine de la surveillance [MZFN04, ZNL01, HHD00, BC09]. Ceux-ci, principalement orientés sur des approches monoculaires, visent à réaliser le suivi global de plusieurs personnes en 2D. Le principal

problème à résoudre correspond à la gestion des occultations entre personnes. Il est généralement résolu par une approche de prédiction/correction du déplacement de chaque personnage.

Les problèmes liés à l'occultation peuvent cependant être réduits lors de l'utilisation d'un système multi-caméras. En effet les formes 3D de plusieurs personnages ne peuvent fusionner en une seule composante que lorsque ces personnages s'occultent (*i.e.* se projettent dans une même composante connexe de silhouette) dans chaque vue. Cependant, la contrepartie du passage de l'information 2D de plusieurs caméras en 3D, est la génération d'entités fantômes. Nous avons proposé dans la première partie de cette thèse (voir chapitre 6) une solution très efficace de suppression d'objets fantômes. De plus, la méthode de suivi de mouvements présentée dans le chapitre précédent, se révèle robuste face aux fantômes topologiques de la forme 3D. Ainsi la génération de fantômes ne constitue pas, dans notre cas, un frein à l'acquisition du mouvement de personnages.

Pour acquérir le mouvement de chaque personne, il est alors nécessaire de suivre le déplacement de chaque individu dans la scène, afin de déterminer l'ensemble des voxels associés à chacun d'eux. Grâce aux travaux présentés jusqu'ici, les seules difficultés rencontrées pour la généralisation de l'acquisition à plusieurs personnages, sont liées à la présence de contacts entre certaines de ces personnes. Dans notre contexte, celles-ci ne peuvent être distingués d'occultations. Dans la suite nous proposons une méthode construite sur une approche de prédiction/correction permettant de déterminer à chaque instant, la position du centre de gravité et l'ensemble de voxels associés à chaque individu, afin d'en réaliser le suivi simultané.

12.4.1 Suivi de chaque personnage

Notre objectif est de réaliser le suivi du centre de gravité de chaque personne présente dans l'espace de reconstruction. Nous proposons d'utiliser le mécanisme de suivi proposé section 12.1.3, construit sur un modèle de mouvements supposé à accélération quasi-constante (équation 12.7). Notons que ce suivi est différent dans la mesure où le nombre de points à suivre (centre de gravité de chaque personnage) peut varier au cours du temps. Cette variation peut être détectée par l'étude du nombre de composantes connexes présentes dans la reconstruction. En effet, l'utilisation de l'approche de suppression des objets fantômes EOR_{seq} décrite chapitre 6.2, offre la garantie que chaque composante "complète" contient au moins un personnage.

L'augmentation du nombre de composantes connexes non-frontières de la reconstruction, indique l'entrée d'une nouvelle personne dans la scène. A l'initialisation du système, nous supposons qu'aucune personne n'est présente dans l'espace d'acquisition. Lors de l'entrée d'un nouveau personnage dans cet espace, une instance de suivi de ce personnage est initialisée. La détection de l'entrée de ce personnage offre cependant une difficulté. Lorsque celui-ci entre dans la boite de reconstruction, seule une partie de sa géométrie est reconstruite, n'impliquant pas nécessairement que la forme calculée soit connexe, par exemple lorsque seuls une jambe et un bras sont

reconstruit. Nous observons que ce problème apparaît uniquement lorsque que la forme calculée contient des voxels frontières de la grille. Cette difficulté est alors contournée en prenant uniquement en compte les nouvelles composantes connexes de la forme qui ne possèdent pas de voxels frontière.

Lorsque le nombre de composantes connexes non-frontières de reconstructions successives est le même que le nombre de personnes suivies, alors le mécanisme d'appariement présenté précédemment résout le problème. La nouvelle position de chaque personne est prédite. Ensuite l'appariement , dans ce cas bijectif, est réalisé entre les centres de gravité des composantes de la reconstruction pour chaque personne et les paramètres du modèle de déplacement de chaque personnage sont mis à jour. Nous considérons peu probable qu'une personne entre et qu'une autre sorte de la scène simultanément.

La diminution du nombre de composantes de la reconstruction entre deux trames successives peut provenir de deux cas différents :

- Une ou plusieurs personnes sortent de la scène, impliquant que les composantes connexes correspondantes passent à l'état frontière. Le mécanisme de prédiction permet d'estimer la position courante de chaque personne. L'appariement est réalisé en prenant aussi en comptes les composantes connexes frontières. L'instance offrant le suivi de la personne appariée à une composante frontière est alors supprimé. Le nombre de personnes présentes dans la scène est décrémenté.

- Plusieurs personnages sont "visuellement" en contact, c'est à dire qu'aucune vue en entrée ne permet de les distinguer. Ce second cas nécessite une attention particulière. En première étape l'appariement est réalisé de façon à déterminer pour toute composante connexe non frontière l'association au personnage correspondant. Ainsi au moins l'un des personnages n'est pas apparié. Chaque personnage est alors associé à la composante connexe la plus proche en terme de distance de Mahalanobis. De cette façon nous obtenons les personnes qui sont contenues dans chaque composante . L'étape suivante revient à mettre à jour le modèle de déplacement de chaque personne. Cette mise à jour est généralement réalisée en utilisant la position du centre de gravité de la composante connexe à laquelle chaque personnage est nouvellement associé. Ceci n'est valable que pour les composantes auxquelles sont appariées une seule personne. Pour les autres, nous proposons de segmenter chaque composante connexe incriminée en autant de région que le nombre de personnes qu'elles contiennent : pour une composante, chaque point est associé à la personne dont la distance de Mahanalobis est minimisée. Le modèle de mouvements de ces personnes est mis à jour à partir du centre de gravité de l'ensemble des points qui lui sont associés.

Nous venons de présenter une mécanisme simple permettant le suivi de la trajectoire de plusieurs personnages présents simultanément dans l'espace reconstruit. Nous avons proposé un mécanisme permettant de rendre l'approche robuste aux contacts. L'approche d'acquisition de mouvements présentée dans ce chapitre peut être réalisée à partir de la composante associée à chaque personnage. En effet nous avons observée sa robustesse face aux auto-occultations et contacts. Cependant nous proposons, pour chaque personnage, de contraindre la recherche des points saillants en ne prenant en compte que les voxels présents dans une sphère centrée sur le centre de gravité dont le diamètre correspond à sa stature estimée.

12.4.2 Résultats de la généralisation à plusieurs personnes

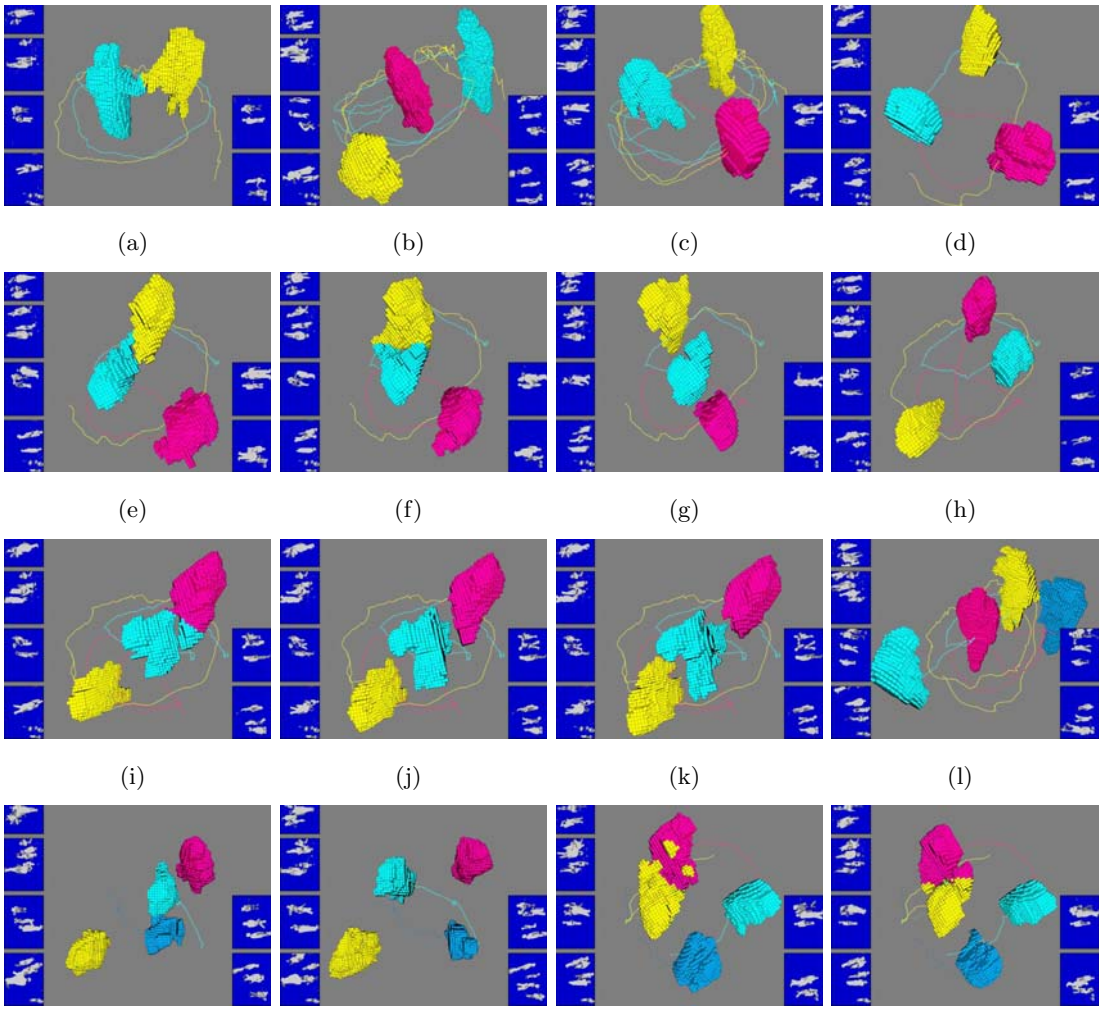
Dans la suite nous présentons les résultats liées à l'acquisition simultanée du mouvement de plusieurs personnes. Ces expérimentations ont été réalisées à partir de données réelles. La reconstruction de forme 3D est définie sur une grille de $128 \times 128 \times 64$ voxels échantillonnant un pavé de $5m \times 5m \times 2,5m$ de côtés. La reconstruction est réalisée par l'approche $\widetilde{\text{EVE}}_m$ présentée chapitre 7, à partir de 6 flux vidéos. Chaque caméra fournit 30 images par seconde d'une résolution de 640×480 pixels. Chacune est connectée à un ordinateur qui réalise la soustraction de fond et l'extraction des parties de peau. Ces informations sont envoyées par réseau à un unique ordinateur (CPU : Intel Core2quad @ 3ghz, GPU Nvidia gtx260) qui réalise le calcul de la forme 3D, dont les objets fantômes sont supprimés par l'approche EOR_{seq}, le suivi de chaque personnage et l'acquisition du mouvement de chacun d'entre eux.

La figure 12.14 montre l'efficacité de l'approche proposée réalisant le suivi de personnages. Les données utilisées en entrée ont été choisies pour leurs conditions d'acquisition difficiles. En effet nous observons la présence d'une grande quantité de bruit dans l'extraction des silhouettes. Même en présence de reconstruction bruitée, cette approche s'avère robuste.

La scène présentée figure 12.14(a) est constituée de deux personnes qui, après avoir parcouru l'espace en rond, se serrent la main. Les lignes fines indiquent la trajectoire de chaque personne. Leur couleur indique à quelle personne elle appartient. Même en présence de contacts entre ces personnes, l'approche se révèle capable d'estimer efficacement leur position.

Les images présentées figures 12.14(e,f,g) montrent le déroulement temporel du suivi de trois personnes, dont deux (en bleu et en jaune) ne sont pas séparables du point de vue des silhouettes. A nouveau notre approche détermine correctement la trajectoire de chacune des personnes. Ce système s'avère capable de traiter des scènes encore plus complexes comprenant 4 personnes souvent en contact (Fig. 12.14(l,m,n,o,p)).

Le suivi réalisé n'entrave pas les temps de calcul de la reconstruction, offrant un suivi de plusieurs personnes simultanément à une cadence de plus de 60 reconstructions par seconde. Les expérimentations menées permettent de valider l'efficacité de notre procédure de suivi de personnes proposée. Celui-ci s'avère robuste aux bruits issus de la reconstruction ainsi qu'aux



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

FIG. 12.14 – Représentation des trajectoires estimées des personnages présents simultanément dans une même scène d'acquisition.

contacts entre les sujets. Enfin celle-ci respecte la contrainte du temps réel.

12.4.3 Acquisition du mouvement de plusieurs personnages

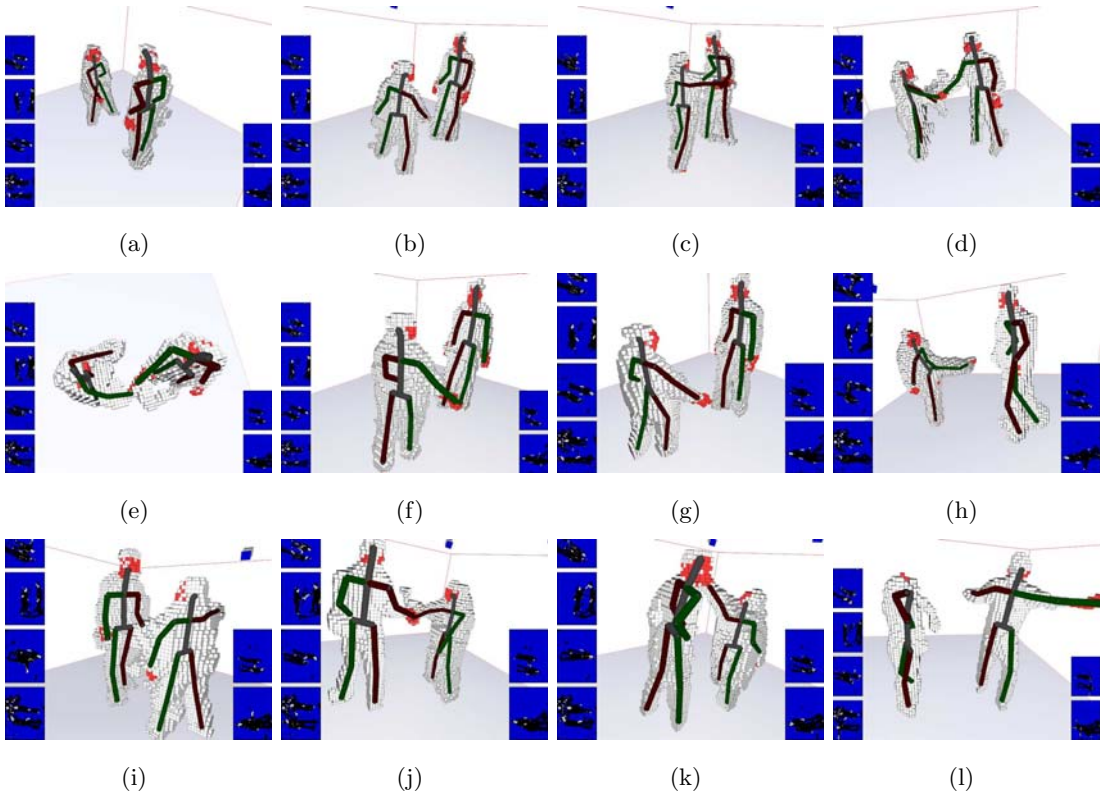


FIG. 12.15 – Capture simultanée du mouvement de plusieurs personnages. Cette méthode s'avère robuste aux contacts entre les sujets.

Disposant de l'estimation du centre de gravité et de la géométrie de chaque personnage présents dans la scène, nous appliquons la méthode d'acquisition de mouvements par ellipsoïdes à chaque personne détectée. La figure 12.15 présente le résultat de cette expérimentation. En présence de mouvements simples tels que ceux présentés figures 12.15(a,b,c) l'acquisition du mouvement offre une estimation correcte de la posture des deux sujets. Lors de contacts entre ces personnes (figures 12.15(d,e,j,k)) cette approche se révèle capable d'extraire la posture de chaque personnage sans difficulté. Cela provient principalement de la combinaison d'une approche efficace du suivi de chaque personnage, associée à la robustesse de l'approche par ellipsoïdes. Cette robustesse est particulièrement soulignée figures 12.15(e,j) où l'on distingue un membre fantôme généré par la méthode de reconstruction.

Cette extension de nos travaux permet de traiter de la présence de plusieurs personnes. Elle se révèle particulièrement robuste face aux contacts, aux parties fantômes générées par la reconstruction, ou encore face aux parties de peau mal extraites (voir figures 12.15(b,i,j,k,l)). L'implantation de cette approche offre des temps de calcul permettant une application en temps interactifs sur un seul pc. Cette approche se révèle capable de capturer plus de 10 mouvements par seconde de deux personnes simultanément. Cependant cette mise en œuvre étant construite

comme une validation de la méthode plutôt qu'un travail d'optimisation, nous avons l'espoir d'accélérer le processus en parallélisant le processus de capture de mouvements, maillon de la chaîne de traitement le plus consommateur de temps de calcul.

12.5 Discussion

Dans ce chapitre nous avons proposé une approche originale permettant l'acquisition totalement automatique du mouvement d'une personne en temps réel ou de plusieurs personnages en temps interactif.

En premier lieu, cette nouvelle approche recherche les extrémités du sujet au temps courant, en fonction de leurs positions calculées au temps précédent. Les extrémités sont extraites à partir d'une représentation hiérarchique de la forme par des ellipsoïdes. Cette représentation offre un filtrage des petites discontinuités de la forme 3D initiale, permettant une extraction robuste des zones saillantes, auxquelles sont ajoutées les parties de peau. Un procédé de suivi construit sur un filtre de Kalman permet de déterminer les zones saillantes correspondantes aux extrémités.

Les articulations intermédiaires sont ensuite estimées en segmentant la forme 3D au temps courant. Nous commençons par estimer la position de chaque segment rigide, en utilisant les positions des articulation intermédiaires estimées au temps précédent et les positions des extrémités extraites au temps courant. La segmentation est réalisée en associant chaque point de la reconstruction au segment rigide estimé le plus proche, définissant une région par segment rigide. Enfin l'extraction des ellipsoïdes représentatives de chaque région permet une estimation précise de la position des articulations.

A travers une expérimentation construite à partir de données réelles représentant des postures difficiles, cette approche s'est illustrée par sa robustesse face aux bruits issus de la reconstruction, et par sa précision. Ce sont ces caractéristiques qui ont permis d'envisager une extension afin de réaliser l'acquisition du mouvement de plusieurs personnes simultanément.

Après avoir identifié les cas de contacts entre les sujets, comme étant la partie critique, nous avons proposé une approche de suivi de la trajectoire de chaque personne. Nous avons observé que la présence de plusieurs personnes au sein d'une même scène implique un augmentation importante des parties fantômes générées. Cependant cette approche s'est révélée peu sensible à cette augmentation importante d'artefacts, comportement souligné par les résultats obtenus.

Il en ressort une approche de capture du mouvement de plusieurs personnes, robuste aux contacts et aux occultations. De plus, à notre connaissance, cette méthode fournit l'un des premiers systèmes d'acquisition de mouvements de plusieurs personnes sans marqueur, fonctionnant en temps interactif.

Dans le chapitre suivant nous proposons une évaluation comparative des différentes méthodes

de capture de mouvements que nous avons proposées.

Chapitre 13

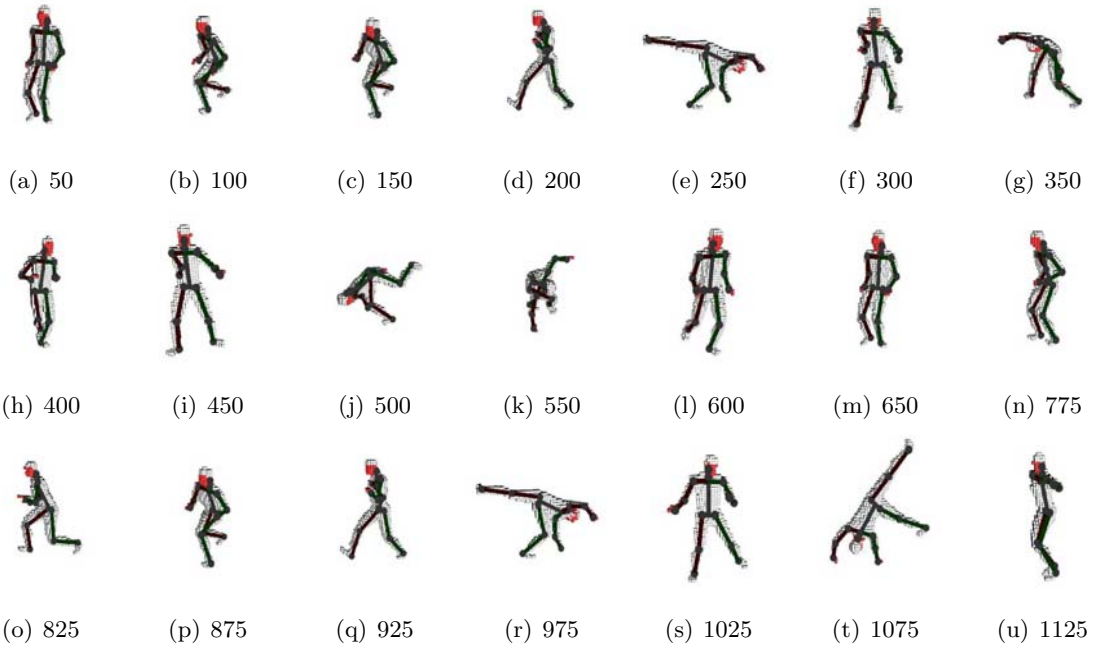
Analyse comparative

Dans les précédents chapitres, nous avons présenté différentes approches ayant pour objectif l'acquisition du mouvement 3D de personnages sans marqueur à partir de plusieurs caméras. Pour chaque approche, nous avons présenté les résultats issus de données réelles. Ces expérimentations n'offrant qu'une analyse visuelle de l'efficacité de chaque méthode, nous proposons dans ce chapitre de confronter ces différentes approches à l'aide de données terrain.

L'expérimentation suivante est articulée autour de cinq jeux de données sélectionnés pour l'hétérogénéité et la complexité des mouvements réalisés. Les données terrain ont été générées à partir de données de motion capture de type "Vicon" [Sys] appliquées à un avatar humain. Pour chaque posture, la géométrie de cet avatar animé a été rendue pour plusieurs points de vue, permettant la génération des silhouettes et des masques de peau correspondants. Ces silhouettes et les paramètres géométriques des caméras sont utilisés comme entrée pour chaque approche. Les expérimentations réalisées prennent en compte à la fois l'estimation de la forme 3D, ainsi que son analyse, permettant une évaluation de la solution complète d'acquisition de mouvements.

Nous avons proposé quatre approches d'acquisition différentes. Les deux premières présentées chapitres 10.2 et 10.3 sont construites sur l'extraction de la topologie de la forme 3D reconstruite. La première réalise cette extraction par la construction du graphe de Reeb construit sur fonction de hauteur, noté *Reeb H* dans la suite. La seconde est construite sur l'extraction de la structure topologique par graphe de Reeb barycentrique, notée *Reeb B*. La troisième approche présentée chapitre 11 est construite sur le recalage d'objets géométriques simples, guidé par les parties de peau. Nous nous référerons à cette approche par la notation *Recal*. Enfin la dernière méthode proposée chapitre 12, construite sur l'extraction et le suivi de points caractéristiques par une représentation de la forme par ellipsoïdes, est notée *Elli* dans la suite. Dans cette expérimentation, nous nous focalisons sur la capture du mouvement d'un seul personnage, n'ayant pas à disposition de données terrain provenant de la capture du mouvement de plusieurs personnes en contact. Les résultats concernant l'approche *Elli* mono-personnage, sont également valables pour sa généralisation, qui ajoute seulement un mécanisme du suivi global de chaque personnage.

13.1 Expérimentation 1



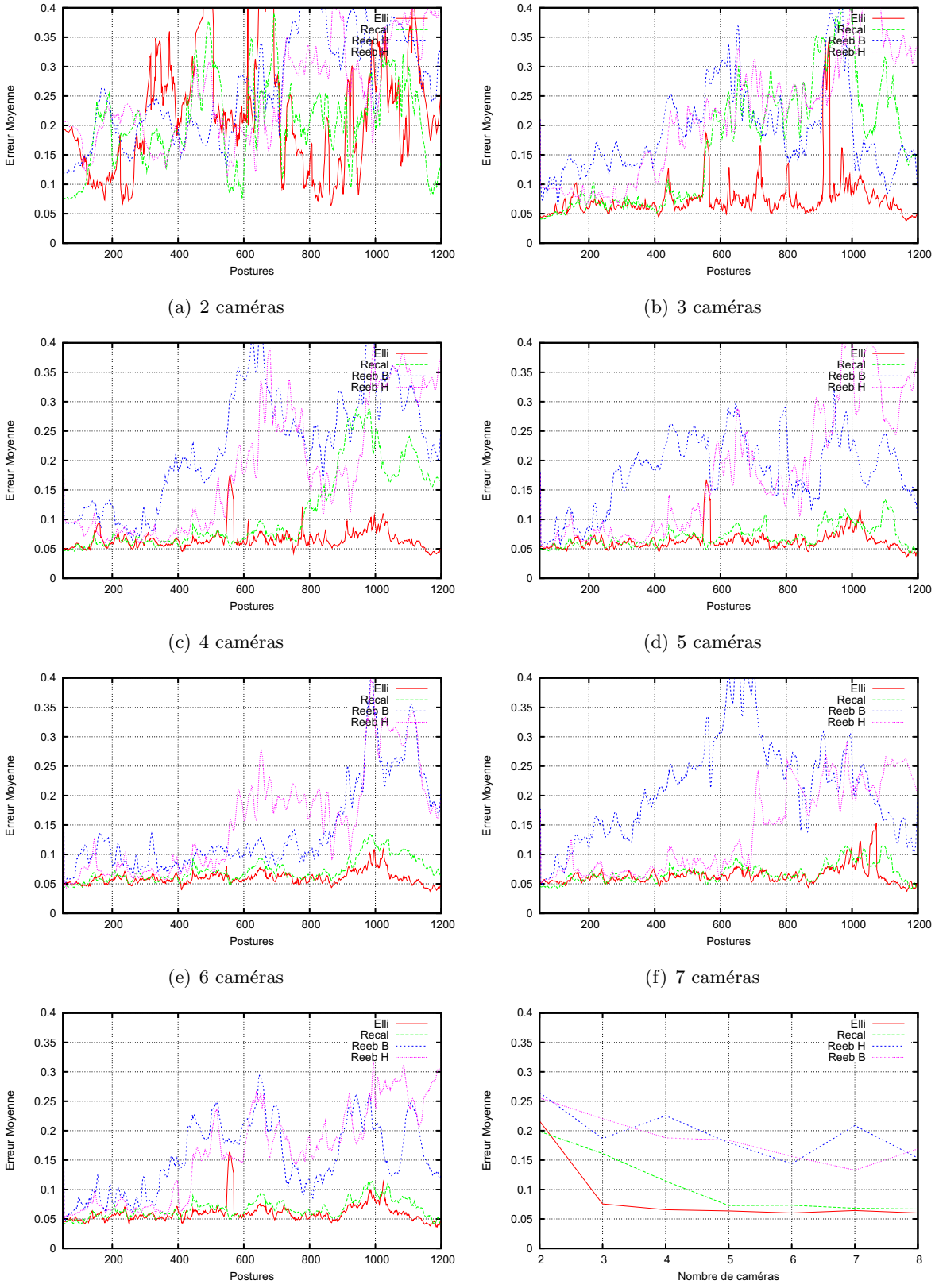
(a) 50 (b) 100 (c) 150 (d) 200 (e) 250 (f) 300 (g) 350

(h) 400 (i) 450 (j) 500 (k) 550 (l) 600 (m) 650 (n) 775

(o) 825 (p) 875 (q) 925 (r) 975 (s) 1025 (t) 1075 (u) 1125

FIG. 13.1 – Données de références de la première séquence. La valeur présente sous chaque image, indique le numéro de posture correspondant.

Cette première expérimentation est construite sur une séquence complexe (voir figure 13.1) dont les mouvement offrent une amplitude et une vitesse élevées. Connaissant la posture pour chaque frame, nous présentons figure 13.2 l'erreur moyenne quadratique entre la position estimée et la position connue de chaque articulation au cours de la séquence (figure 13.2(a-g)), en fonction du nombre de vues utilisées. Pour l'ensemble des expérimentations menées dans ce chapitre, cette mesure est définie relativement à la stature (taille) de l'individu. Ainsi une erreur de 1.0 indique que la distance moyenne entre la position estimée et connue de chaque articulation, est de l'ordre de la stature de la personne suivie. L'ensemble des expérimentations ont été réalisées à partir d'une grille de voxels d'une résolution de 64^3 pour une boite de 2 mètres de côté, définissant des voxels de $3,125$ cm de côté. En présence de deux vues (figure 13.2(a)), l'ensemble des méthodes offrent une estimation grossière du mouvement. Ceci s'explique par la conjugaison d'une estimation très bruitée de la forme 3D, à un mouvement de grande amplitude. Pour 3 et 4 points de vue, seule l'approche *Elli* offre une estimation raisonnable du mouvement. Quoique parfois en échecs, représentés par des pics, celle-ci est capable de reprendre rapidement le suivi, alors que les autres approches en sont incapables. A partir de 5 vues, les approches *Elli* et *Recal* offrent un suivi d'une erreur de l'ordre de 6% très inférieure aux approches construites sur le graphe de Reeb, sensibles au bruits de reconstruction. Nous observons figure 13.2(h), la rapidité de convergence de l'approche *Elli*, qui offre des erreurs similaires au delà de 3 vues.



(a) 2 caméras

(b) 3 caméras

(c) 4 caméras

(d) 5 caméras

(e) 6 caméras

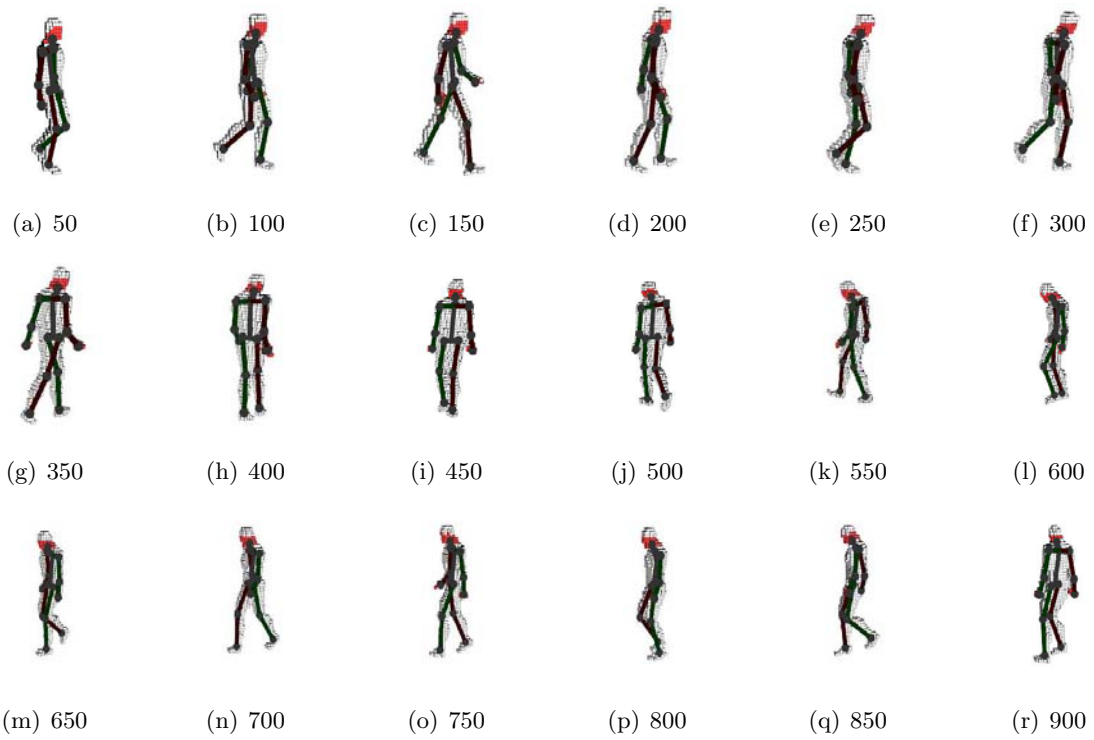
(f) 7 caméras

(g) 8 caméras

(h) Global

FIG. 13.2 – Erreur quadratique moyenne entre la position estimée et connue de chaque articulation. Les graphiques (a) à (g) soulignent cette erreur pour chaque posture, en fonction du nombre de points de vue. (h) présente l'erreur quadratique moyenne globale en fonction du nombre de vues.

13.2 Expérimentation 2



(a) 50 (b) 100 (c) 150 (d) 200 (e) 250 (f) 300

(g) 350 (h) 400 (i) 450 (j) 500 (k) 550 (l) 600

(m) 650 (n) 700 (o) 750 (p) 800 (q) 850 (r) 900

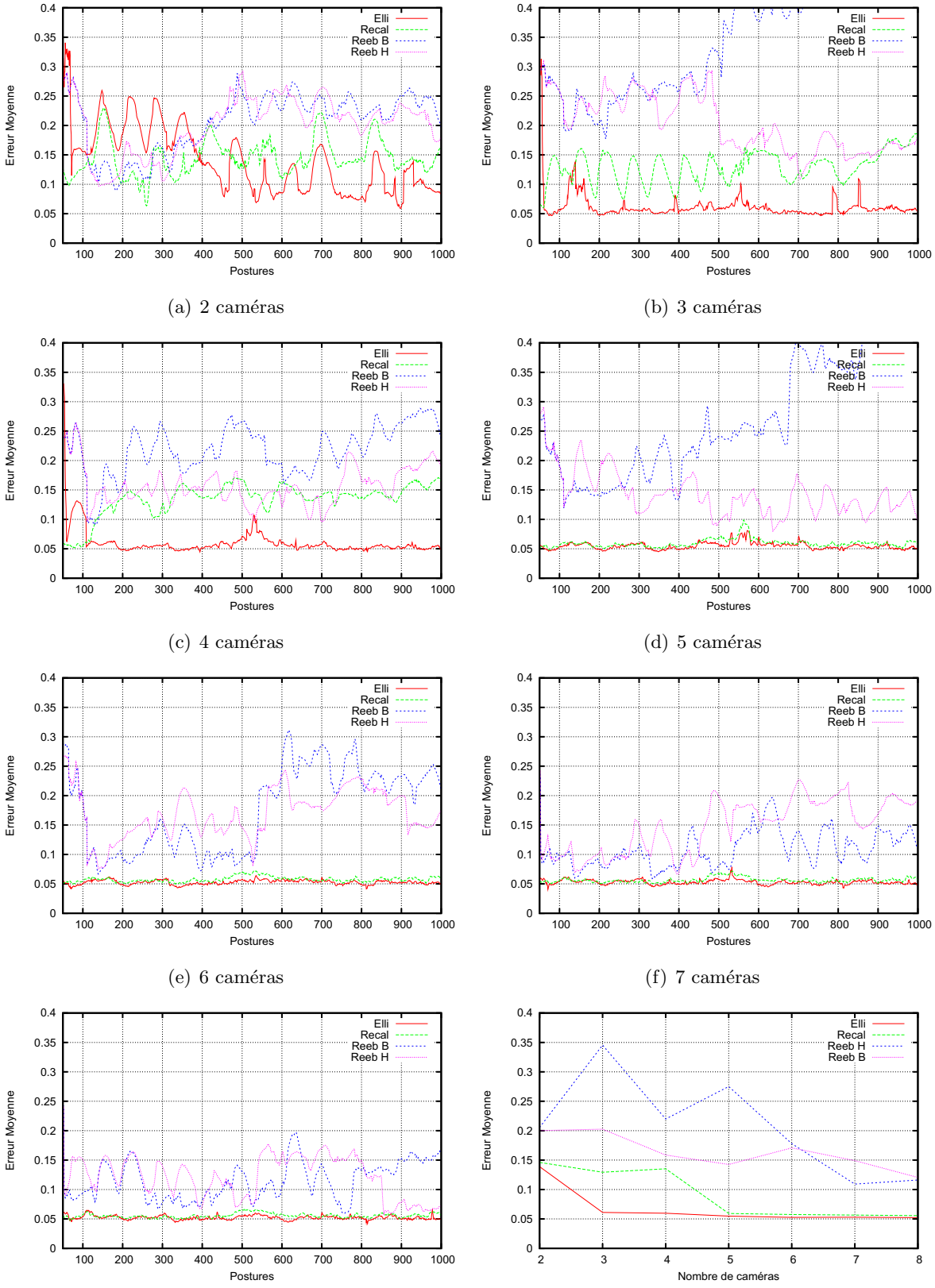
FIG. 13.3 – Mouvements de référence du second jeu de données. Ceux-ci représentent un mouvement de marche en ligne droite, suivi d'un virage sur la gauche, et enfin de quelques pas en arrière.

Cette seconde expérimentation est construite sur un mouvement de marche impliquant des contacts entre les membres supérieurs et le buste, ainsi qu'entre les membres inférieurs (voir figure 13.3).

Ce mouvement étant d'amplitude et de vitesse inférieures à l'expérimentation précédente, les résultats obtenus à partir de deux vues s'avèrent de meilleure qualité, avec une erreur comprise entre 15% et 20%.

A nouveau, seule l'approche *Elli* offre un suivi précis des articulations en présence de 3 et 4 vues (figure 13.4(b,c)) avec une erreur de l'ordre de 7%. Les pics présents correspondent à l'inversion des pieds et genoux droits et gauches, rapidement compensée et corrigée.

A partir de 5 vues, les approches *Elli* et *Recal* offrent une précision de l'ordre de 5%, soit une erreur équivalente la largeur de la paume de la main. Ce résultat proviennent principalement de l'utilisation des parties de peau, qui permettent de prévenir de la perte des mains lors de contacts avec le buste. La méthode *Elli* offre un suivi dont l'erreur converge vers 5% dès l'utilisation de 3 points de vue (figure 13.4(h)).



(a) 2 caméras

(b) 3 caméras

(c) 4 caméras

(d) 5 caméras

(e) 6 caméras

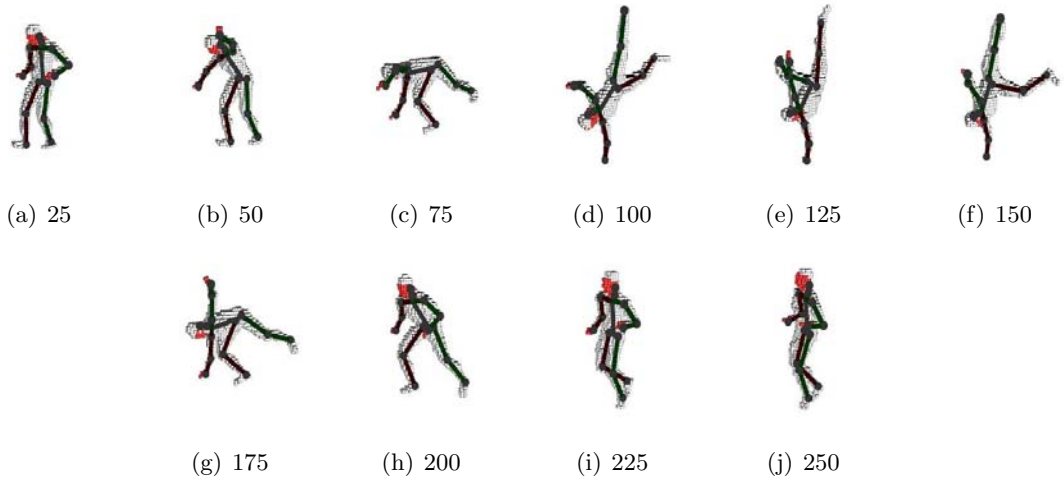
(f) 7 caméras

(g) 8 caméras

(h) Global

FIG. 13.4 – Erreur quadratique moyenne entre la position estimée et celle connue de chaque articulation, en fonction du nombre de caméras.

13.3 Expérimentation 3



(a) 25 (b) 50 (c) 75 (d) 100 (e) 125 (f) 150

(g) 175 (h) 200 (i) 225 (j) 250

FIG. 13.5 – Vérité terrain, troisième jeu de données. Cette séquence représente un personnage se mettant en équilibre sur un main, avec la seconde tenant la jambe opposée, puis revenant à la position d'origine.

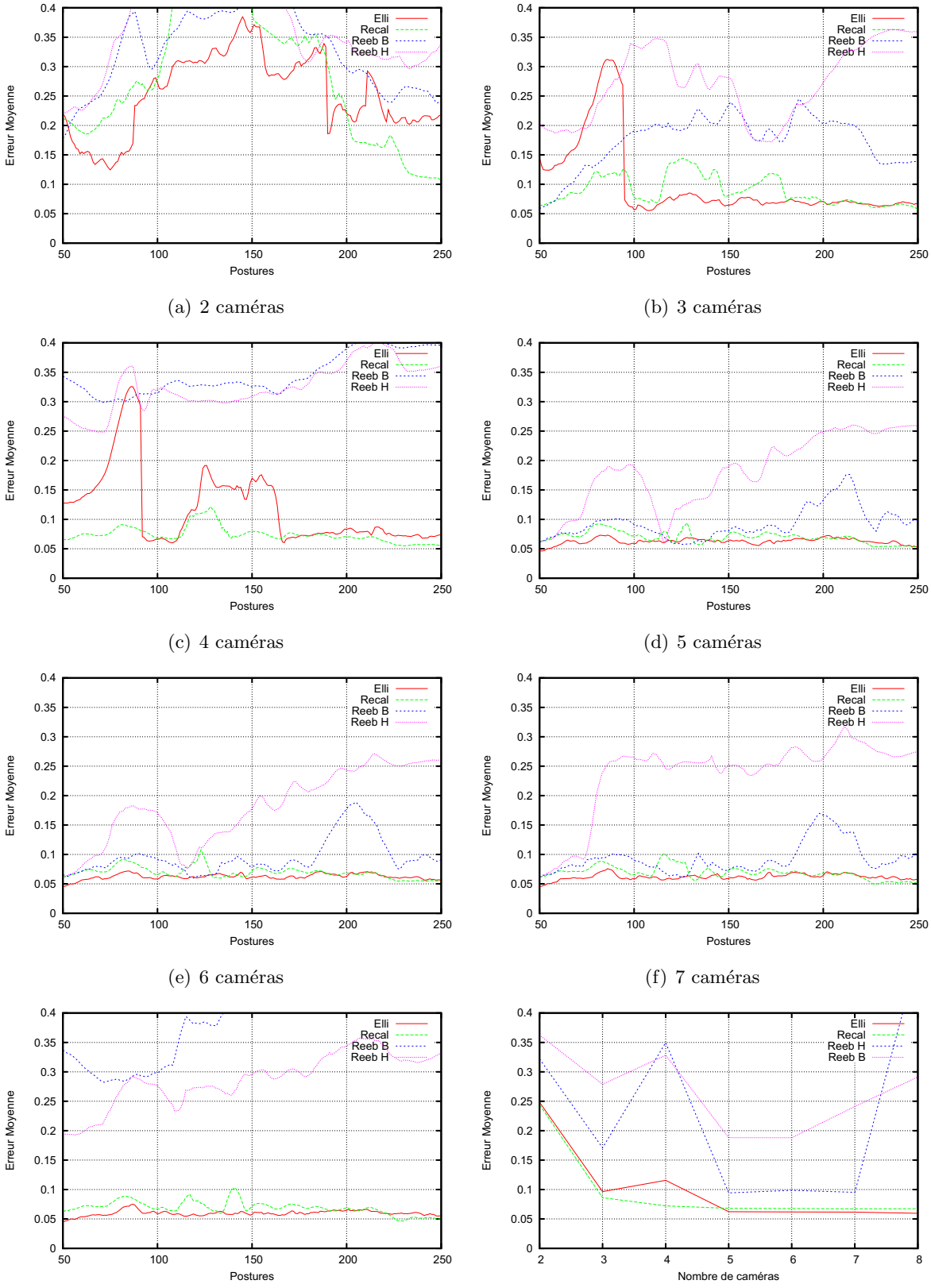
Cette expérimentation est réalisée à partir d'une séquence de mouvements impliquant un contact entre l'une des mains et de la jambe opposée (voir figure 13.5). Pour cette raison, cette séquence risque de mettre en difficulté les approches *Reeb H*, *Reeb B* ainsi que *Elli*. En effet celles-ci sont construites sur l'hypothèse selon laquelle un point saillant de la forme 3D, correspond à au plus une extrémité du corps.

Pour la reconstruction à partir de 2 vues, l'ensemble des approches offrent des résultats similaires. Cependant l'approche *Elli* offre une estimation dont l'erreur moyenne est la plus faible.

Contrairement aux expérimentations précédentes, pour l'utilisation de 2 ou 3 caméras, l'approche *Recal* surpasse les approches construites sur graphe de Reeb, ainsi que l'approche *Elli*. Cela provient d'une inversion des côtés gauche-droite des articulations. Cependant *Elli* est capable de compenser ce phénomène assez rapidement.

Cette dernière méthode offre à nouveau le meilleur suivi en présence de plus de 3 vues en entrée, suivie de près par l'approche *Recal*, cependant plus sensible aux contacts entre les bras et les jambes. Cette différence provient de la méthode de réduction du nombre de voxels utilisée dans l'approche *Recal*, qui supprime une partie des voxels des jambes, alors que celles-ci n'ont pas encore été déterminées. Ce phénomène s'illustre particulièrement pour les postures voisines de la posture 125 présentée figure 13.5(e).

Nous observons que pour cette expérimentation, les approches *Recal* et *Elli* offrent toutes deux une convergence rapide vers une erreur de 6% (figure 13.6(h)).



(a) 2 caméras

(b) 3 caméras

(c) 4 caméras

(d) 5 caméras

(e) 6 caméras

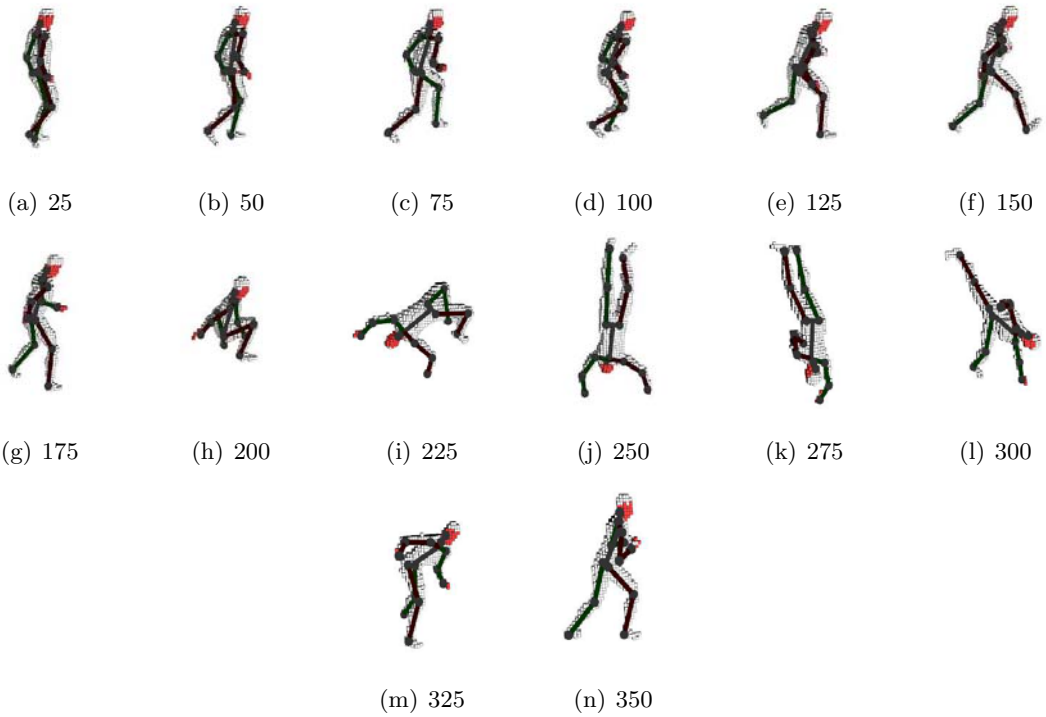
(f) 7 caméras

(g) 8 caméras

(h) Global

FIG. 13.6 – Erreur quadratique moyenne entre la position estimée et celle connue de chaque articulation, en fonction du nombre de caméras.

13.4 Expérimentation 4



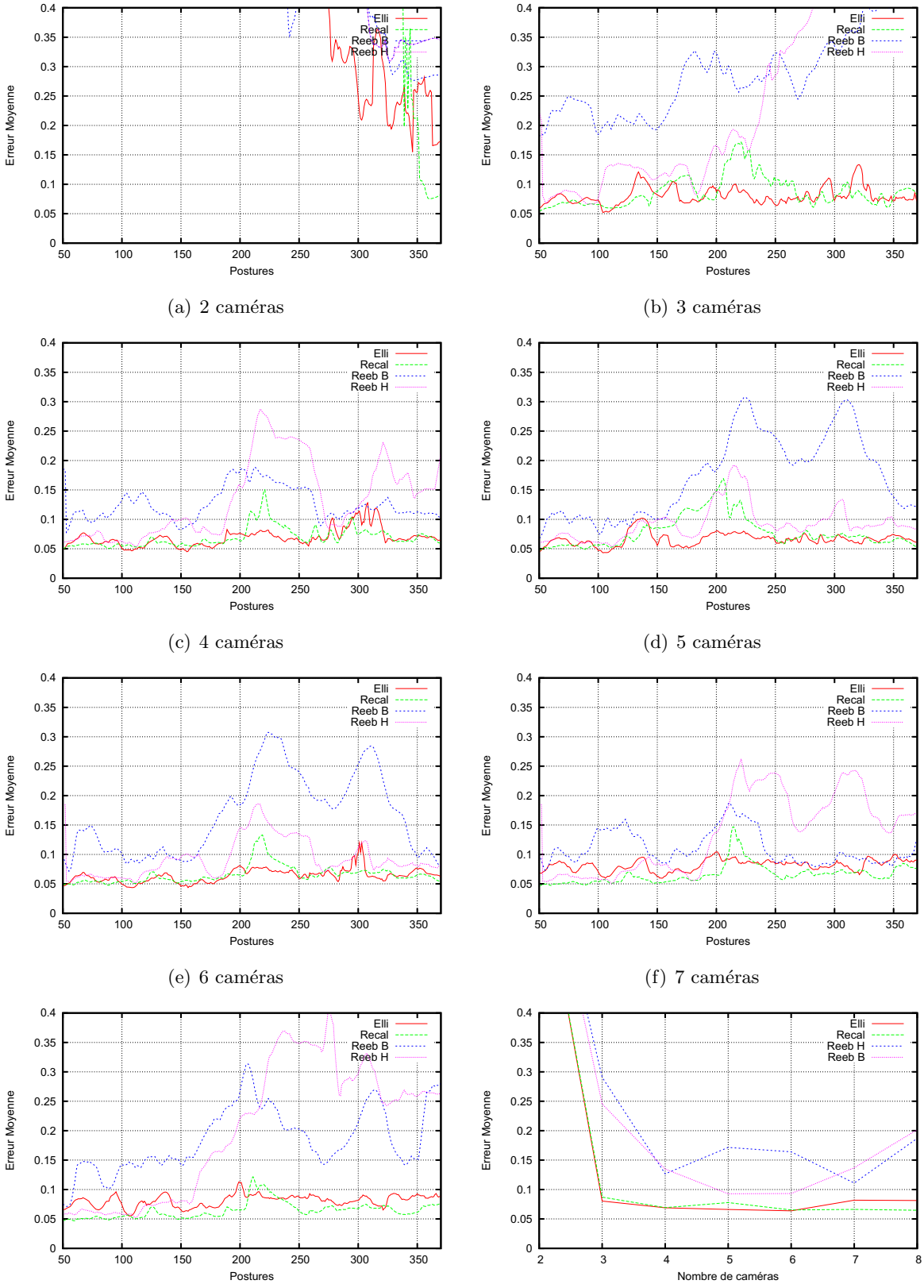
(a) 25 (b) 50 (c) 75 (d) 100 (e) 125 (f) 150

(g) 175 (h) 200 (i) 225 (j) 250 (k) 275 (l) 300

(m) 325 (n) 350

FIG. 13.7 – Mouvements de référence du quatrième jeu de données. La séquence montre un personnage passant successivement de la position debout, à accroupi, puis réalisant un saut main inversé, passant les jambe au dessus de la tête, pour revenir enfin à une pose d'équilibre.

A travers ce nouveau jeu de données, nous confrontons les différentes approches développées face à un mouvement rapide, présentant des contacts importants de tous les membres avec le buste lors de l'impulsion du saut de main (voir figure 13.7). Il en ressort qu'aucune méthode n'est capable de s'initialiser lors d'une reconstruction issue de deux points de vue. En effet la configuration des caméras implique la génération d'un nombre élevé de partie fantômes, mettant en échec l'initialisation du suivi. En présence de 3 caméras, seule l'approche *Reeb B* tombe en échec. Cet effet provient principalement l'extraction du graphe de Reeb barycentrique qui se révèle particulièrement sensible aux bruits de reconstruction. Notons que l'approche *Reeb H* offre un comportement correcte jusqu'à la frame 250, initiatrice d'un décrochage de l'approche. Pour toutes les expériences menées pour une reconstruction issue de 4 à 8 caméras, les approches *Elli* et *Recal* offrent un suivi comparable. Cependant l'approche *Elli* se révèle plus robuste que l'approche *Recal*, dont nous observons un pic d'erreur présent sur l'ensemble des graphes 13.8(c-g) au voisinage de la posture 210. A cet instant l'acteur réalise le mouvement le plus rapide de la séquence, offrant aussi des contacts importants entre le buste et les membres. La figure 13.8(h) souligne à nouveau la forte convergence de l'erreur de l'approche *Elli*. Notons que pour ce jeu de données, l'approche *Recal* le même type de comportement.



(a) 2 caméras

(b) 3 caméras

(c) 4 caméras

(d) 5 caméras

(e) 6 caméras

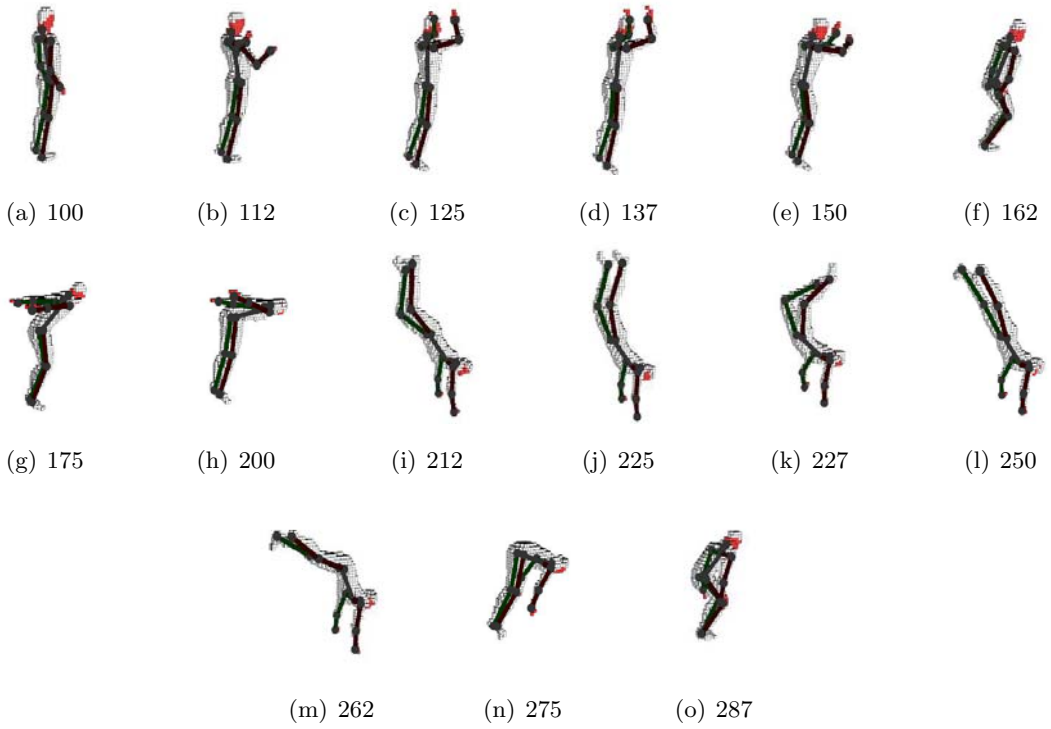
(f) 7 caméras

(g) 8 caméras

(h) Global

FIG. 13.8 – Erreur quadratique moyenne entre la position estimée et celle connue de chaque articulation, en fonction du nombre de caméras.

13.5 Expérimentation 5



(a) 100 (b) 112 (c) 125 (d) 137 (e) 150 (f) 162

(g) 175 (h) 200 (i) 212 (j) 225 (k) 227 (l) 250

(m) 262 (n) 275 (o) 287

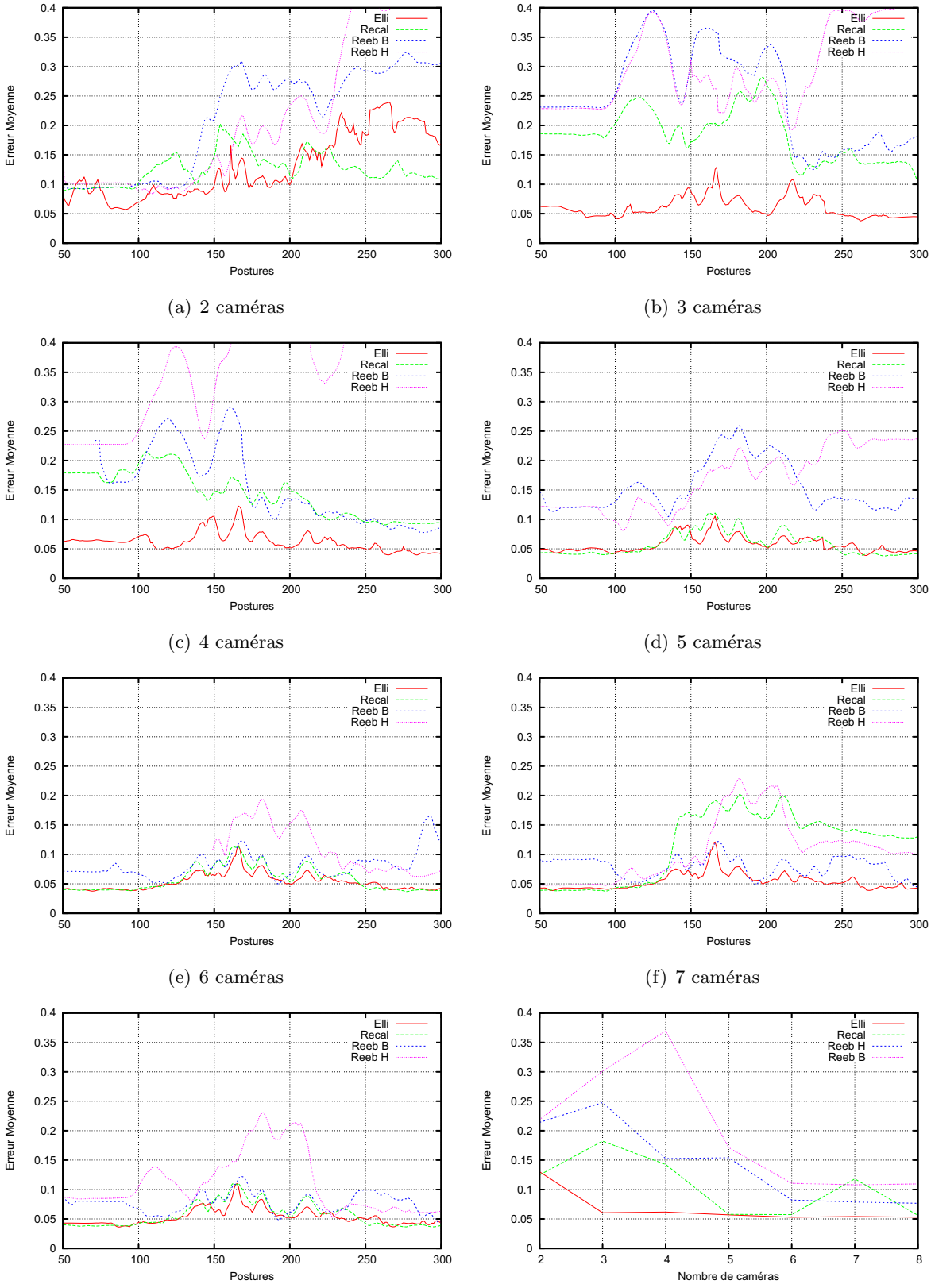
FIG. 13.9 – Vérité terrain des mouvements du dernier jeu de données. La séquence représente une gymnaste se mettant en équilibre sauté, revenant ensuite à sa position d'origine.

Cette dernière expérimentation est réalisée à partir d'une séquence de mouvements rapides, contenant de fortes accélérations et décélérations pour des extrémités du corps (voir figure 13.9). De plus les jambes se retrouvent en contact lors de l'amortissement de la figure réalisée.

En présence de deux points de vue, l'ensemble des approches fournissent une estimation semblable jusqu'à la posture 130 associée à la prise d'élan de l'acteur, pour laquelle l'accélération associée à chaque articulation se révèle forte. Ensuite seules les approches *Elli* et *Recal* offrent une estimation acceptable du mouvement. Cependant une inversion des côtés gauche et droit, pénalise l'erreur le suivi de l'approche *Elli* pour les 100 dernières postures.

Seule l'approche *Elli* se révèle capable d'estimer correctement cette séquence pour une reconstruction provenant de 3 ou 4 caméras, grâce au processus de suivi des extrémités, construit sur un filtre de Kalman.

Enfin les approches *Elli* et *Recal* offrent un bon suivi des articulations pour l'ensemble des autres configurations, avec une erreur moyenne de l'ordre de 6%. Le graphe présenté figure 13.10(h) illustre la supériorité de l'approche *Elli* comparée aux autres méthodes proposées.



(a) 2 caméras

(b) 3 caméras

(c) 4 caméras

(d) 5 caméras

(e) 6 caméras

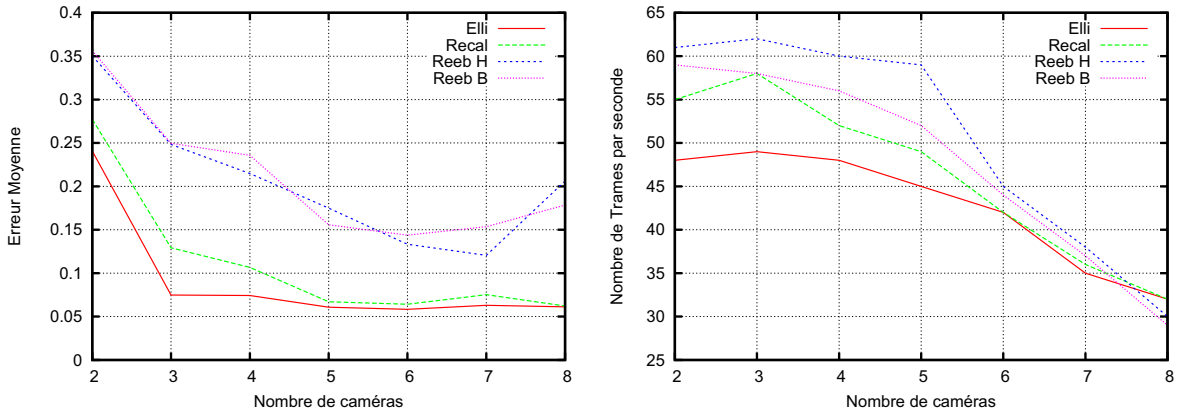
(f) 7 caméras

(g) 8 caméras

(h) Global

FIG. 13.10 – Erreur quadratique moyenne entre la position estimée et celle connue de chaque articulation, en fonction du nombre de caméras.

13.6 Synthèse



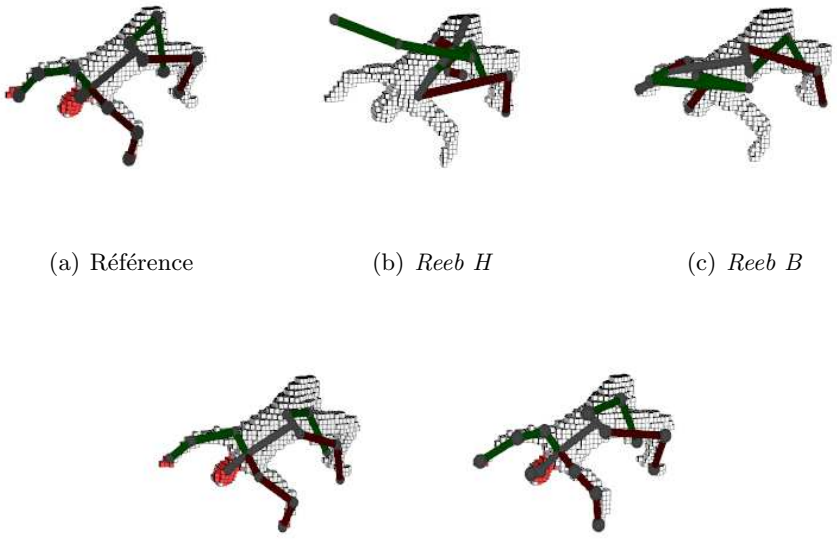
(a) (b)

FIG. 13.11 – (a) Synthèse de l'erreur quadratique de chaque méthode, représentée en fonction du nombre de vues utilisées. (b) Les temps de calculs correspondants soulignent le respect du temps réel.

A travers ces cinq expérimentations construites sur des jeux de données variés, nous avons confronté les méthodes de suivi de mouvements sans marqueur proposées. Le graphe présenté dans la figure 13.11(a) souligne les erreurs quadratiques moyennes entre les positions connues des articulations et l'estimation de ces positions par chacune des méthodes. Il en ressort que les approches construites uniquement sur l'extraction de la topologie (*Reeb H* et *Reeb B*), se révèlent sensibles aux imprécisions de reconstruction qui sont directement liées au nombre de vues utilisées. Cette sensibilité peut être associée à deux facteurs. Le premier concerne l'extraction de la topologie par construction de graphes de Reeb, qui ne fournit pas nécessairement l'estimation la plus robuste. Le second facteur correspond à l'invalidité de l'hypothèse sous-jacente selon laquelle la topologie de la forme reconstruite correspond à la topologie du corps humain. Cette hypothèse est principalement invalidée lors de l'utilisation d'un faible nombre de points de vue ou lors de contacts entre différentes parties du corps. Dans ce cas, l'erreur peut être supérieure à 15%. Cependant ces méthodes fournissent les temps de calculs les plus faibles, avec 50 à 60 acquisitions de mouvements par seconde (voir figure 13.11(b)).

Bien que construite sur des mécanismes de recalage simples, dirigés par l'information de peau, l'approche *Recal* offre un suivi d'une erreur de moins de 10% lors de l'utilisation de plus de deux vues. De plus l'erreur tend vers 7% lors de l'utilisation de 5 caméras et plus. Enfin les temps de calculs présentés (voir figure 13.11(b)) soulignent l'efficacité algorithmique de cette approche, qui respecte la contrainte du temps réel, en offrant une cadence de plus de 30 captures par seconde.

La dernière méthode, *Elli*, exposée chapitre 12, offre le suivi de meilleur précision (figure



(a) Référence (b) *Reeb H* (c) *Reeb B*

(d) *Recal* (e) *Elli*

FIG. 13.12 – Comparaison de l'acquisition du mouvement réalisé par les différentes méthodes proposées dans la seconde partie de cette thèse. La forme 3D est reconstruite à partir de 6 points de vue. Le mouvement réalisé s'avère particulièrement difficile en raison de sa vitesse d'exécution.

13.11(a)). En effet le suivi par filtre de Kalman, des extrémités complétées par les parties de peau, offre un mécanisme particulièrement efficace face aux auto-occultations, auto-contacts et parties fantômes. Cela provient principalement du filtrage intrinsèque de la forme par la représentation de la forme 3D par d'ellipsoïdes, qui implique une extraction robuste des extrémités. Les parties de peau offrent une insensibilité par rapport aux contacts des bras avec toute autre partie du corps. Enfin le mécanisme de segmentation de la forme 3D par plus proche segment offre un recalage rapide et efficace de chaque partie rigide du corps suivi. Cette efficacité est soulignée par la rapidité de convergence de l'erreur de la capture des mouvements dès l'utilisation de 3 points de vue (voir figure 13.11(a)).

Remarquons que la valeur asymptotique de 5% d'erreur provient du fait que les rapports anthropométriques utilisés ne correspondent pas à ceux de l'acteur, initiateur des mouvements de données terrain. Nous n'avons pas modifié les rapports anthropométriques utilisés, afin de proposer une analyse objective des méthodes proposées. Cependant le fait de les adapter aux personnages suivis aurait comme effet de diminuer l'erreur d'acquisition. La figure 13.11 indique que l'approche *Elli* s'avère la moins rapide. Notons cependant que cette approche fournit plus de 45 acquisitions de mouvements par seconde pour moins de 5 vues. De plus la diminution de la cadence en fonction du nombre de caméras, se révèle plus faible que pour les autres approches.

Méthode	Vitesse	Précision	Reprise
Reeb H	++	−	−
Reeb B	+	−	−
Recal	−	+	+
Elli	−−	++	++

FIG. 13.13 – Tableau comparatif des différentes approches proposées.

La figure 13.12 fournit une comparaison visuelle de l'acquisition de mouvements offert par chaque méthode proposée pour une posture complexe. La méthode *Elli* s'illustre à nouveau en offrant la meilleur estimation. Le tableau présenté figure 13.13 fournit une synthèse comparative des différentes méthodes de capture de mouvements proposées. Même si l'approche *Elli* s'avère en moyenne un peu moins rapide, ses facultés de reprise, ainsi que la précision obtenue en font la méthode la plus efficace. En sus des mesures appuyant cette position, nous rappelons que, par ces facultés, cette approche permet l'acquisition du mouvement de plusieurs personnages simultanément.

Chapitre 14

Conclusion

Nous nous sommes intéressés à la problématique de la capture du mouvement de personnes en temps réel et sans marqueur à partir de plusieurs caméras. Nous avons fait le choix de travailler sur une description tridimensionnelle de ces personnes. Ainsi nous avons exploré les axes suivants.

– L'extraction de la structure topologique de la forme 3D permet de localiser les extrémités du corps humain, puis les articulations intermédiaires. La représentation de cette structure par graphe de Reeb sur fonction de hauteur a donné une solution pour l'initialisation de la posture. L'extension de cette structure par un graphe de Reeb barycentrique, invariant par rotation, a donné une première solution pour l'acquisition automatique de mouvements. Ce type d'approche suppose que la topologie extraite corresponde à la topologie du corps humain. Cette hypothèse est mise en défaut lors d'auto-contacts, ce qui restreint la capture à des mouvements simples.

– Nous avons répondu à ce problème en exploitant la couleur de peau, qui permet d'identifier dans la forme 3D les mains et le visage. Cette information, conjuguée à un recalage rapide contraint par les rapports anthropométriques, permet la capture de la posture en temps réel. Le processus de suivi des extrémités peut être mis en difficulté lors de mouvements rapides.

– Nous apportons une réponse par un mécanisme d'extraction et de suivi robuste des extrémités du corps, par l'extraction les zones saillantes de la forme tridimensionnelle, auxquelles sont associées les parties 3D de peau. Un filtre de Kalman permet de suivre de façon robuste chaque extrémité. Une généralisation de la méthode *Point To Line Mapping* segmente la forme 3D en régions associées à chaque partie rigide du corps. Les ellipsoïdes représentatives de chaque région, conduisent à l'extraction de la posture du corps en entier. Cette approche est robuste aux contacts entre les membres, aux bruits issus de la reconstruction et aux mouvements rapides.

La généralisation de cette approche permet l'acquisition des mouvements de plusieurs personnes simultanément. Les extensions proposées dans la première partie de thèse permettent de supprimer tout objet fantôme, ainsi chaque composante reconstruite contient au moins un personnage. Grâce à cette propriété, nous avons proposé un processus qui réalise l'acquisition en temps interactif du mouvement de chaque personne, même lors de contacts entre elles.

Nous avons réalisé une analyse comparative de l'ensemble des méthodes, construite à partir de données terrain. Les résultats obtenus présentent une bonne performance, allant jusqu'à une erreur moyenne de moins de 6% dès l'utilisation de trois caméras. Notons que toutes les approches proposées satisfont à la contrainte du temps réel.
Nous sommes conscients que nos méthodes ne fournissent pas encore une précision comparable aux approches à base de marqueurs. Néanmoins, elles ouvrent la voie à des applications interactives grand public.

Conclusion générale

Nous nous sommes intéressé à la problématique de l'acquisition du mouvement de personnages en temps réel et sans marqueur. Nous avons été amenés à traiter du problème de la reconstruction 3D en temps réel et de l'analyse de cette forme afin d'en déduire le mouvement. Nos contributions concernent les approches *Shape-From-Silhouette* ainsi que l'acquisition de mouvements à partir de formes 3D.

Nos contributions apportées aux approches *Shape-From-Silhouette* sont les suivantes :

- Les contraintes de placement des caméras ont été relaxées, grâce à la prise en compte d'un critère de consistance portant sur le domaine de visibilité de chaque caméra. Il en résulte que lors de l'ajout d'une caméra, en plus d'améliorer la précision, l'espace d'acquisition est aussi augmenté, contrairement aux approches usuelles de *Shape-From-Silhouette*.
- Les objets fantômes sont supprimés. Nous avons proposé des critères formels construits sur des propriétés géométriques qui permettent d'assurer la suppression des composantes connexes fantômes. Cela a pour effet de diminuer fortement la quantité d'artefacts reconstruits. Comme elles ne nécessitent pas d'information supplémentaire en dehors des silhouettes et des paramètres géométriques des caméras, elles sont génériques et peuvent être appliquées à toute approche estimant l'enveloppe visuelle.
- La reconstruction 3D est robuste aux silhouettes bruitées. Nous avons proposé une approche qui fusionne dans l'espace 3D l'information des cartes valuées de silhouettes. L'existence d'un élément 3D est validée en fonction de l'information majoritaire issue de l'ensemble des caméras. Ainsi la reconstruction obtenue est plus robuste au bruit présent dans les silhouettes, qui peut provenir des conditions d'illumination de la scène, peu contrôlées.

L'ensemble de nos contributions liées à *Shape-From-Silhouette*, apporte une réponse dont l'implantation respecte la contrainte du temps réel. Les méthodes relatives à l'extension de l'espace d'acquisition et à la suppression des objets fantômes, s'appliquent à toute estimation de l'enveloppe visuelle, qu'elle soit volumique ou surfacique. L'utilisation conjointe de ces méthodes offre une estimation de la forme plus robuste et précise que les approches purement géométriques de la littérature.

Même si nous avons traité des principales limitations des approches *Shape-From-Silhouette*,

nous n'avons pas proposé de solution concernant la suppression des parties fantômes. Celles-ci sont notamment responsables, dans notre contexte, de la création de membres (bras, jambes) fantômes, qui peuvent fortement nuire au processus de capture de mouvements. Ces membres fantômes se caractérisent en tant que branches topologiques fantômes. Une réponse peut être apportée par l'analyse topologique de la forme reconstruite. Dans ce but, nous travaillons actuellement sur une extension de notre méthode *EOR*.

Concernant l'acquisition du mouvement en temps réel et sans marqueur, le problème est traité en s'appuyant sur la forme 3D obtenue. Nos contributions concernant cette problématiques sont les suivantes :

- L'initialisation totalement automatique du processus. Nous avons présenté des approches permettant de réaliser la capture d'une première posture de l'individu filmé, sans lui imposer de pose caractéristique telle que la pose en T. La seule contrainte est que le sujet arrive en marchant et que ses membres ne soient pas totalement collés aux corps.
- La capture du mouvement à partir de graphes de Reeb. Nous avons été parmi les premiers à proposer l'utilisation de graphes de Reeb pour extraire la topologie de la forme calculée. Ceci permet d'extraire les extrémités et les articulations intermédiaires d'un corps humain.
- Le respect de la contrainte du temps réel sur une seule machine. Le respect de cette contrainte, n'est pas simplement le résultat de l'augmentation des puissances de calcul. L'ensemble des méthodes développées contribuent largement à cette performance, par la recherche de primitives 3D pertinentes (couleur de peau, zones saillantes) et discriminantes, ainsi que la proposition de méthodes adaptées.
- L'acquisition multi-personnages de mouvements en temps interactif. Une généralisation de ces méthodes, permet l'acquisition sans marqueur du mouvement de plusieurs personnes, présentes simultanément, et robuste aux contacts entre elles. Nous sommes à notre connaissance, parmi les premiers à offrir ce processus de capture en temps interactif.

Les solutions apportées pourraient fournir une capture du mouvement encore plus précise en intégrant des contraintes articulaires. Actuellement, en dehors des longueurs anthropométriques, nous n'utilisons pas de contraintes bio-mécaniques. Celles-ci permettraient d'imposer que la posture capturée soit physiquement réaliste, et ainsi réduire l'espace de recherche.

L'ensemble des contributions proposées, offre une réponse qui valide les objectifs que nous nous sommes fixés. Les solutions apportées ouvrent des perspectives à un transfert vers des applications grand public. Elles n'ont pas pour objectif de remplacer les solutions construites à base de marqueurs, qui offrent actuellement une précision supérieure. Elles ouvrent la voie à une utilisation dans le cadre d'applications interactives. Le principal frein de ce transfert

concerne la sensibilité de l'extraction des silhouettes en présence d'un milieu non contrôlé, tel que celui que l'on peut rencontrer à domicile. Une solution matérielle dédiée permettrait de limiter, voire supprimer cette sensibilité. L'une des pistes serait l'utilisation d'une bande de fréquence lumineuse peu sensible aux contributions des éclairages artificiels et naturels.

Les caméras à temps de vol sont peu sensibles aux conditions d'éclairage artificiel. Ce type de solution existe dans le commerce, leur coût reste prohibitif (de l'ordre de 6000 euros) pour une utilisation grand public. Récemment, la société Microsoft a dévoilé le projet Natal [Nat] qui vise à utiliser ce type de caméras en tant qu'interface de jeux vidéo. Ceci augure d'une généralisation de ce type de produit en masse, amenant une baisse du coût de ces périphériques.

Si l'objectif de notre travail a été l'acquisition du mouvement, l'une des prolongations naturelles serait l'analyse et l'interprétation de ces mouvements. Cette problématique s'avère très étendue. Elle nécessite la prise en compte du type d'IHM visée et du contexte de l'interaction. Certains contextes imposent une analyse précise du mouvement, principalement lorsque la performance est recherchée, par exemple des applications d'entraînement au geste sportif, etc. Prenons l'exemple du jeu vidéo, il s'agit principalement de mettre en correspondance le mouvement réalisé avec une grammaire réduite de mouvements connus.

L'ensemble de ces perspectives ne font pas seulement appel au domaine de la vision par ordinateur, mais nécessitent en plus la prise en compte des aspects cognitifs ou ergonomiques permettant la mise en place et l'adaptation rapide des utilisateurs à de telles interactions. Même si ces questions ne semblent pas directement toucher nos domaines de recherche, leur prise en compte est cependant fondamentale quant à l'intégration et à la valorisation de ces travaux de recherche au sein de notre société.

Bibliographie

[AC99] J. K. Aggarwal and Q. Cai. Human motion analysis : a review. *Comput. Vis. Image Underst.*, 73(3) :428–440, 1999.

[ACP02] Brett Allen, Brian Curless, and Zoran Popović. Articulated body deformation from range scan data. *ACM Trans. Graph.*, 21(3) :612–619, 2002.

[AFM98] Naoki Asada, Hisanaga Fujiwara, and Takashi Matsuyama. Edge and depth from focus. *Int. J. Comput. Vision*, 26(2) :153–163, 1998.

[AGD07] C. ALBITAR, P. GRAEBLING, and C. DOIGNON. Robust structured light coding for 3d reconstruction. In *Eleventh IEEE International Conference on Computer Vision ICCV'07*, Rio de Janeiro, Brazil, October 2007.

[AL01] D. Attali and J.-O. Lachaud. Delaunay conforming iso-surface ; skeleton extraction and noise removal. *Computational Geometry : Theory and Applications*, 19(2-3) :175–189, 2001.

[Ale99] Marc Alexa. Merging polyhedral shapes with scattered features. In *SMI '99 : Proceedings of the International Conference on Shape Modeling and Applications*, page 202, Washington, DC, USA, 1999. IEEE Computer Society.

[APSK07] Ehsan Aganj, Jean-Philippe Pons, Florent Ségonne, and Renaud Keriven. Spatio-temporal shape from silhouette using four-dimensional delaunay meshing. In *ICCV*, pages 1–8. IEEE, 2007.

[AS88] Y. Aloimonos and M.J. Swain. Shape from texture. *BioCyber*, 58(5) :345–360, 1988.

[ATC+08] Oscar Kin-Chung Au, Chiew-Lan Tai, Hung-Kuo Chu, Daniel Cohen-Or, and Tong-Yee Lee. Skeleton extraction by mesh contraction. In *SIGGRAPH '08 : ACM SIGGRAPH 2008 papers*, pages 1–10, New York, NY, USA, 2008. ACM.

[Bak81] Henry Harlyn Baker. *Depth from edge and intensity based stereo.* PhD thesis, Champaign, IL, USA, 1981.

[Bau74] Bruce Guenther Baumgart. *Geometric modeling for computer vision.* PhD thesis, Stanford, CA, USA, 1974.

[BC09] F. Bardet and T. Chateau. Real time multi-object tracking with few particles. In *Visapp, International Conference on Vision Theory and Applications*, Lisboa, Portugal, Fevrier 2009.

[BCL01] David A Brown, Ian Craw, and Julian Lewthwaite. A som based approach to skin detection with application in real time systems. In *in Proc. of the British Machine Vision Conference*, 2001.

[BD02] Gary R. Bradski and James W. Davis. Motion segmentation and pose recognition with motion history gradients. *Mach. Vision Appl.*, 13(3) :174–184, 2002.

[BF03] Edmond Boyer and Franco Franco. A hybrid approach for computing visual hulls of complex objects. *cvpr*, 01 :695, 2003.

[BG08] Alexander Bogomjakov and Craig Gotsman. Reduced depth and visual hulls of complex 3d scenes. *Computer Graphics Forum*, 27(2) :175–182, April 2008.

[BH94] A M Baumberg and D C Hogg. An efficient method for contour tracking using active shape models. In *In Proceeding of the Workshop on Motion of Nonrigid and Articulated Objects. IEEE Computer Society*, pages 194–199, 1994.

[BKC04] Matthew Brand, Kongbin Kang, and David B. Cooper. Algebraic solution for the visual hull. *cvpr*, 01 :30–35, 2004.

[BKS01] Ingmar Bitter, Arie E. Kaufman, and Mie Sato. Penalized-distance volumetric skeleton algorithm. *IEEE Transactions on Visualization and Computer Graphics*, 7(3) :195–206, 2001.

[BL04] Andrea Bottino and Aldo Laurentini. The visual hull of smooth curved objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12) :1622–1632, 2004.

[BM58] G. E. P. Box and Mervin E. Muller. A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, 29(2) :610–611, 1958.

[BM00a] Jason Brand and John S. Mason. A comparative assessment of three approaches to pixel-level human skin-detection. *Pattern Recognition, International Conference on*, 1 :5056, 2000.

[BM00b] Christoph Bregler and Jitendra Malik. Tracking people with twists and exponential maps. Technical report, Berkeley, CA, USA, 2000.

[BMMP03] Silvia Biasotti, Simone Marini, Michela Mortara, and Giuseppe Patané. An overview on properties and efficacy of topological skeletons in shape modelling. In *SMI '03 : Proceedings of the Shape Modeling International 2003*, page 245, Washington, DC, USA, 2003. IEEE Computer Society.

[BMP04] Christoph Bregler, Jitendra Malik, and Katherine Pullen. Twist based acquisition and tracking of animal and human kinematics. *Int. J. Comput. Vision*, 56(3) :179–194, 2004.

[Bor96] Gunilla Borgefors. On digital distance transforms in three dimensions. *Comput. Vis. Image Underst.*, 64(3) :368–376, 1996.

[BS03] S. Biasotti and M. Spagnuolo. Shape understanding by contour driven retiling. *The Visual Computer*, 19 :2–3, 2003.

[Car00] Stefan Carlsson. Recognizing walking people. In *ECCV '00 : Proceedings of the 6th European Conference on Computer Vision-Part I*, pages 472–486, London, UK, 2000. Springer-Verlag.

[CB92] Roberto Cipolla and Andrew Blake. Surface shape from the deformation of apparent contours. *Int. J. Comput. Vision*, 9(2) :83–112, 1992.

[CB00] D. Chai and A. Bouzerdoum. A bayesian approach to skin color classification in ycbcr colorspace. In *TENCON 2000. Proceedings*, volume 2, pages 421–424, 2000.

[CBK03a] German K.M. Cheung, Simon Baker, and Takeo Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. *cvpr*, 01 :77, 2003.

[CBK03b] German K.M. Cheung, Simon Baker, and Takeo Kanade. Visual hull alignment and refinement across time : A 3d reconstruction algorithm combining shape-from-silhouette with stereo. *cvpr*, 02 :375, 2003.

[CCZ07] Michel Couprie, David Coeurjolly, and Rita Zrour. Discrete bisector function and euclidean skeleton in 2d and 3d. *Image and Vision Computing*, 2007.

[CGH08] Fabrice Caillette, Aphrodite Galata, and Toby Howard. Real-time 3-d human body tracking using learnt models of behaviour. *Comput. Vis. Image Underst.*, 109(2) :112–125, 2008.

[CH96] I-C. Chang and C-L. Huang. Ribbon-based motion analysis of human body movements. *icpr*, 03 :436, 1996.

[CH04a] F. Caillette and T. Howard. Real-time markerless human body tracking using colored voxels and 3d blobs. *Mixed and Augmented Reality, 2004. ISMAR 2004. Third IEEE and ACM International Symposium on*, pages 266–267, Nov. 2004.

[CH04b] Fabrice Caillette and Toby Howard. Real-Time Markerless Human Body Tracking with Multi-View 3-D Voxel Reconstruction. In *Proceedings of British Machine Vision Conference (BMVC)*, volume 2, pages 597–606, September 2004.

[CKBH00] Kong Man Cheung, Takeo Kanade, J.-Y. Bouguet, and M. Holler. A real time system for robust 3d voxel reconstruction of human motions. In *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, volume 2, pages 714 – 720, June 2000.

[CL92] Z. Chen and H.J. Lee. Knowledge-guided visual perception of 3-d human gait from a single image sequence. 22 :336–342, 1992.

[CMS99] W. Bruce Culbertson, Thomas Malzbender, and Greg Slabaugh. Generalized voxel coloring. pages 100–115. Springer-Verlag, 1999.

[CR98] S. Chaudhuri and A.N. Rajagopalan. *Depth from Defocus : A Real Aperture Imaging Approach*. Springer-Verlag, 1998.

[CRHD00] Michele Covell, Ali Rahimi, Michael Harville, and Trevor Darrell. Articulated-pose estimation using brightness and depth-constancy constraints. *cvpr*, 02 :2438, 2000.

[CSM07] Nicu D. Cornea, Deborah Silver, and Patrick Min. Curve-skeleton properties, applications, and algorithms. *IEEE Transactions on Visualization and Computer Graphics*, 13(3) :530–548, 2007.

[CTCG95] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Comput. Vis. Image Underst.*, 61(1) :38–59, 1995.

[CTMS03] Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 22(3) :569–577, 2003.

[dAST+08] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *SIGGRAPH '08 : ACM SIGGRAPH 2008 papers*, pages 1–10, New York, NY, USA, 2008. ACM.

[dATM+04] Edilson de Aguiar, Christian Theobalt, Marcus Magnor, Holger Theisel, and Hans-Peter Seidel. m^3 : Marker-free model reconstruction and motion tracking from 3d voxel data. *pg*, 00 :101–110, 2004.

[DBR00] J. Duetscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. *cvpr*, 02 :2126, 2000.

[DC01] T.W. Drummond and R. Cipolla. Real-time tracking of highly articulated structures in the presence of noisy measurements. In *IEEE Internation Conference on Computer Vision*, pages II : 315–320, 2001.

[DF99] Quentin Delamarre and Olivier Faugeras. 3d articulated models and multi-view tracking with silhouettes. *iccv*, 02 :716, 1999.

[DF01] Quentin Delamarre and Olivier Faugeras. 3d articulated models and multiview tracking with physical forces. *Comput. Vis. Image Underst.*, 81(3) :328–357, 2001.

[DG06] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *ICML '06 : Proceedings of the 23rd international conference on Machine learning*, pages 233–240, New York, NY, USA, 2006. ACM.

[dlRG05] Jean-Baptiste de la Rivière and Pascal Guitton. Model-based video tracking for gestural interaction. *Virtual Reality*, 8(4) :213–221, 2005.

[DT01] H. Dreyfuss and A. R. Tilley. *The Measure of Man and Woman : Human Factors in Design*. John Wiley & Sons, 2001.

[DW90] Trevor Darell and K. Wohn. Depth from focus using pyramid architecture. *Pattern Recogn. Lett.*, 11(12) :787–796, 1990.

[DW05] Albert Dipanda and Sanghyuk Woo. Towards a real-time 3d shape reconstruction using a structured light system. *Pattern Recognition*, 38(10) :1632–1650, 2005.

[EDH+02] Ahmed Elgammal, Ramani Duraiswami, David Harwood, Larry S. Davis, R. Duraiswami, and D. Harwood. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. In *Proceedings of the IEEE*, pages 1151–1163, 2002.

[FAHF08] A. Fossati, E. Arnaud, R. Horaud, and P. Fua. Tracking articulated bodies using generalized expectation maximization. In *CVPR Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment*, Anchorage, AK, USA, June 2008.

[Fau96] Olivier Faugeras. *Three-Dimensional Computer Vision : A Geometric Viewpoint*. The MIT press, Cambridge, Massachusetts, 1996.

[FB05] Jean-Sébastien Franco and Edmond Boyer. Fusion of multi-view silhouette cues using a space occupancy grid. In *Proc. of the 10th International Conference on Computer Vision*, oct 2005.

[FB08] Jean-Sebastien Franco and Edmond Boyer. Efficient polyhedral modeling from silhouettes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[FCZ98] Andrew W. Fitzgibbon, Geoff Cross, and Andrew Zisserman. Automatic 3d model construction for turn-table sequences. pages 155–170. Springer-Verlag, 1998.

[FFB96] Margaret M. Fleck, David A. Forsyth, and Chris Bregler. Finding naked people. In *ECCV '96 : Proceedings of the 4th European Conference on Computer Vision-Volume II*, pages 593–602, London, UK, 1996. Springer-Verlag.

[FGDP02] P. Fua, A. Gruen, N. D'Apuzzo, and R. Plankers. Markerless Full Body Shape and Motion Capture from Video Sequences. In *Symposium on Close Range Imaging, International Society for Photogrammetry and Remote Sensing, Corfu, Greece*, 2002.

[Fie88] David A. Field. Laplacian smoothing and delaunay triangulations. *Communications in Applied Numerical Methods*, 4(6) :709–712, 1988.

[For98] R. Forman. Morse theory for cell complexes. *Advances in Mathematics*, 134(1) :90–145, 1998.

[FTOK96] W. T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma. Computer vision for computer games. *fg*, 00 :100, 1996.

[Gar92] J. Garding. Shape from texture for smooth curved surfaces in perspective projection. *JMIV*, 2 :329–352, 1992.

[Gav98] Dariu Mihai Gavrila. *Vision-based 3-D tracking of humans in action.* PhD thesis, College Park, MD, USA, 1998.

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

[GBUP95] L. Goncalves, E. Di Bernardo, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3d. *iccv*, 00 :764, 1995.

[GD96] D. M. Gavrila and L. S. Davis. 3-d model-based tracking of humans in action : a multi-view approach. In *CVPR '96 : Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96)*, page 73, Washington, DC, USA, 1996. IEEE Computer Society.

[Geu93] Alexander Geurtz. *Model based shape estimation.* Thèse, Ecole polytechnique fédérale de Lauzanne, 1993.

[GF02] Michael Gleicher and Nicola Ferrier. Evaluating video-based motion capture. In *CA '02 : Proceedings of the Computer Animation*, page 75, Washington, DC, USA, 2002. IEEE Computer Society.

[GFP07] Li Guan, Jean-Sébastien Franco, and Marc Pollefeys. 3d occlusion inference from silhouette cues. In *CVPR*. IEEE Computer Society, 2007.

[GFP08] Li Guan, Jean-Sebastien Franco, and Marc Pollefeys. Multi-object shape estimation and tracking from silhouette cues. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, (USA)*, jun 2008.

[GGTS01] N. Grammalidis, G. Goussis, G. Troufakos, and M.G. Strintzis. Estimating body animation parameters from depth images using analysis by synthesis. *dcv*, 00 :93, 2001.

[Gle00] Michael Gleicher. Animation from observation : Motion capture and motion editing. *SIGGRAPH Comput. Graph.*, 33(4) :51–54, 2000.

[GM02] Giovani Gomez and Eduardo F. Morales. Automatic feature construction and a simple rule induction algorithm for skin detection. In *In Proc. of the ICML Workshop on Machine Learning in Computer Vision*, pages 31–38, 2002.

[GM04a] Bastian Goldluecke and Marcus Magnor. Space-time isosurface evolution for temporally coherent 3d reconstruction. *cvpr*, 01 :350–355, 2004.

[GM04b] Bastian Goldluecke and Marcus Magnor. Space-time isosurface evolution for temporally coherent 3d reconstruction. In *In International Conference on Computer Vision and Pattern Recognition*, pages 350–355, 2004.

[GMMB00] E. Guillou, Daniel Meneveaux, Eric Maisel, and Kadi Bouatouch. Using vanishing points for camera calibration and coarse 3d reconstruction from a single image. *The Visual Computer*, 16(7) :396–410, 2000.

[Gom02] Giovani Gomez. On selecting colour components for skin detection. *Pattern Recognition, International Conference on*, 2 :20961, 2002.

[Gra71] A. Gramain. *Topologie des surfaces*. Presses Universitaires Françaises, 1971.

[Gra03] O. Grau. Studio production system for dynamic 3D content. In T. Ebrahimi and T. Sikora, editors, *Visual Communications and Image Processing 2003. Edited by Ebrahimi, Touradj ; Sikora, Thomas. Proceedings of the SPIE, Volume 5150, pp. 80-89 (2003).*, volume 5150 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pages 80–89, June 2003.

[GSCD08] Laetitia Gond, Patrick Sayd, Thierry Chateau, and Michel Dhome. A 3d shape descriptor for human pose recovery. In Francisco J. Perales López and Robert B. Fisher, editors, *AMDO*, volume 5098 of *Lecture Notes in Computer Science*, pages 370–379. Springer, 2008.

[GSS93] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2) :107–113, 1993.

[GT95] Enrico Grosso and Massimo Tistarelli. Active/dynamic stereo vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(9) :868–879, 1995.

[GW87] Peter Giblin and Richard Weiss. Reconstruction of surfaces from profiles. Technical report, Amherst, MA, USA, 1987.

[GXT94] Y. Guo, G. Xu, and S. Tsuji. Tracking human body motion based on a stick figure model. 5 :1–9, 1994.

[GYB04] S. Burak Gokturk, Hakan Yalcin, and Cyrus Bamji. A time-of-flight depth sensor - system description, issues and solutions. In *CVPRW '04 : Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 3*, page 35, Washington, DC, USA, 2004. IEEE Computer Society.

[HAMJ02] R.L. Hsu, M. Abdel-Mottaleb, and A.K. Jain. Face detection in color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5) :696–706, 2002.

[HFP+00] L. Herda, P. Fua, R. Plänkers, R. Boulic, and D. Thalmann. Skeleton-based motion capture for robust reconstruction of human motion. In *CA '00 : Proceedings of the Computer Animation*, page 77, Washington, DC, USA, 2000. IEEE Computer Society.

[HHD98] I. Haritaoglu, D. Harwood, and L.S. Davis. W4 : A real time system for detecting and tracking people. *cvpr*, 00 :962, 1998.

[HHD00] Ismail Haritaoglu, David Harwood, and Larry S. Davis. W4 : Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :809–830, 2000.

[HLS04] Jean-Marc Hasenfratz, Marc Lapierre, and François Sillion. A real-time system for full body interaction with virtual worlds. *Eurographics Symposium on Virtual Environments*, pages 147–156, 2004.

[HSJ95] E. Hunter, J. Schlenzig, and R. Jain. Posture estimation in reduced-model gesture input systems. *IEEE International Workshop on Automatic Face and Gesture Recognition*, pages 290–295, june 1995.

[HUF04] Lorna Herda, Raquel Urtasun, and Pascal Fua. Hierarchical implicit surface joint limits to constrain video-based motion capture. In *In Proc. ECCV*, pages 405–418. Springer, 2004.

[HUF05] L. Herda, R. Urtasun, and P. Fua. Hierarchical implicit surface joint limits for human body tracking. *Comput. Vis. Image Underst.*, 99(2) :189–209, 2005.

[IB98] Michael Isard and Andrew Blake. Condensation - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29 :5–28, 1998.

[JB02] K. Jin and B. Boufama. Towards a fast and reliable dense matching algorithm. In *VI02*, page 178, 2002.

[JF02] Hailin Jin and Paolo Favaro. A variational approach to shape from defocus. In *In Proc. European Conference on Computer Vision*, pages 18–30, 2002.

[Joh73] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14 :201–211, 1973.

[Jos04] Neel Joshi. Color calibration for arrays of inexpensive image sensors. Technical report, Stanford University, 2004.

[JR02] Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. *Int. J. Comput. Vision*, 46(1) :81–96, 2002.

[JU97] S. Julier and J. Uhlmann. A new extension of the kalman filter to nonlinear systems. 1997.

[Kal60] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering*, (82 (Series D)) :35–45, 1960.

[KG00] Zachi Karni and Craig Gotsman. Spectral compression of mesh geometry. In *SIGGRAPH '00 : Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 279–286, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[KK96] R. Kjeldsen and J. Kender. Toward the use of gesture in traditional user interfaces. *fg*, 00 :151, 1996.

[KKB+93] Andreas Koschan, Andreas Koschan, Technische Universität Berlin, Technische Universität Berlin, and Technischer Bericht. What is new in computational stereo since 1989 : A survey on current stereo papers. Technical report, 1993.

[KKI99] Y. Kenmochi, K. Kotani, and A. Imiya. Marching cubes method with connectivity. pages IV :361–365, 1999.

[KLPL02] Sung-Eun Kim, Ran-Hee Lee, Chang-Joon Park, and In-Ho Lee. Mimic : Real-time marker-free motion capture system to create an agent in the virtual space. In *ICCE '02 : Proceedings of the International Conference on Computers in Education*, page 48, Washington, DC, USA, 2002. IEEE Computer Society.

[KLT05] Sagi Katz, George Leifman, and Ayellet Tal. Mesh segmentation using feature point and core extraction. *The Visual Computer*, 21(8) :649–658, 2005.

[KM95] I.A. Kakadiaris and D. Metaxas. 3d human body model acquisition from multiple views. *iccv*, 00 :618, 1995.

[KMB07] P. Kakumanu, S. Makrogiannis, and N. Bourbakis. A survey of skin-color modeling and detection methods. *Pattern Recogn.*, 40(3) :1106–1122, 2007.

[Koc93] R. Koch. Dynamic 3-d scene analysis through synthesis feedback control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6) :556–568, 1993.

[KPT99] Evaggelia-Aggeliki Karabassi, Georgios Papaioannou, and Theoharis Theoharis. A fast depth-buffer-based voxelization algorithm. *J. Graph. Tools*, 4(4) :5–10, 1999.

[KR89] T. Y. Kong and A. Rosenfeld. Digital topology : introduction and survey. *Comput. Vision Graph. Image Process.*, 48(3) :357–393, 1989.

[KS00] Kiriakos N. Kutulakos and Steven M. Seitz. A theory of shape by space carving. *Int. J. Comput. Vision*, 38(3) :199–218, 2000.

[KS03] J. Kovac and P. Peerand F. Solina. Human skin color clustering for face detection. In *EUROCON 2003. Computer as a Tool. The IEEE Region 8*, volume 2, pages 144–148, september 2003.

[KS04] Vladislav Kraevoy and Alla Sheffer. Cross-parameterization and compatible remeshing of 3d models. *ACM Trans. Graph.*, 23(3) :861–869, 2004.

[KSK06] A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *ICPR06*, pages III : 15–18, 2006.

[KWT88] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes : Active contour models. *International Journal of Computer Vision*, V1(4) :321–331, January 1988.

[Lau92] Aldo Laurentini. The visual hull of solids of revolution. *11th IAPR International Conference on Pattern Recognition*, I :720–724, 1992.

[Lau94] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(2) :150–162, 1994.

[Lau95] Aldo Laurentini. How far 3d shapes can be understood from 2d silhouettes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(2) :188–195, 1995.

[Lau99] Aldo Laurentini. The visual hull of curved objects. *iccv*, 01 :356, 1999.

[LCO06] Chulhan Lee, Junho Cho, and Kyoungsu Oh. Hardware-accelerated jaggy-free visual hulls with silhouette maps. In *VRST '06 : Proceedings of the ACM symposium on Virtual reality software and technology*, pages 87–90, New York, NY, USA, 2006. ACM.

[LESC04] G. Loy, M. Eriksson, J. Sullivan, and S. Carlsson. Monocular 3d reconstruction of human motion in long action sequences. In *ECCV04*, pages Vol IV : 442–455, 2004.

[LFP07] Svetlana Lazebnik, Yasutaka Furukawa, and Jean Ponce. Projective visual hulls. *Int. J. Comput. Vision*, 74(2) :137–165, 2007.

[LH05] A.M. Loh and R. Hartley. Shape from non-homogeneous, non-stationary, anisotropic, perspective texture. *Proc. British Machine Vision Conf*, pages 69–78, 2005.

[Lie03] André Lieutier. Any open bounded subset of rn has the same homotopy type than its medial axis. In *SM '03 : Proceedings of the eighth ACM symposium on Solid modeling and applications*, pages 65–75, New York, NY, USA, 2003. ACM.

[LMS04] Ming Li, Marcus Magnor, and Hans-Peter Seidel. A hybrid hardware-accelerated algorithm for high quality rendering of visual hulls. In *GI '04 : Proceedings of Graphics Interface 2004*, pages 41–48, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2004. Canadian Human-Computer Communications Society.

[LN09] Mun Wai Lee and Ramakant Nevatia. Human pose tracking in monocular sequence using multilevel structured models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(1) :27–38, 2009.

[LPD06] Thomas C. W. Landgrebe, Pavel Paclik, and Robert P. W. Duin. Precision-recall operating characteristic (p-roc) curves in imprecise environments. In *ICPR '06 : Proceedings of the 18th International Conference on Pattern Recognition*, pages 123–127, Washington, DC, USA, 2006. IEEE Computer Society.

[LWB07] Wonwoo Lee, Woo Wontack, and Edmond Boyer. Identifying foreground from multiple images. In *In Proceedings of the Eighth Asian Conference on Computer Vision*, December 2007.

[LWTH01] Xuetao Li, Tong Wing Woon, Tiow Seng Tan, and Zhiyong Huang. Decomposing polygon meshes for interactive applications. In *I3D '01 : Proceedings of the 2001 symposium on Interactive 3D graphics*, pages 35–42, New York, NY, USA, 2001. ACM.

[LY95] Maylor K. Leung and Yee-Hong Yang. First sight : A human body outline labeling system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(4) :359–377, 1995.

[LY98] Jae Y. Lee and Suk I. Yoo. An elliptical boundary model for skin color detection. In *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

[Mah36] P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings National Institute of Science, India*, volume 2, pages 49–55, April 1936.

[Mar87] Etienne-Jules Marey. Le mécanisme du vol des oiseaux éclairé par la chronophotographie. *La nature : revue des sciences et de leurs applications aux arts et à l'industrie*, pages 8–14, 1887.

[MAS02] S. Malassiotis, N. Aifanti, and M. G. Strintzis. A gesture recognition system using 3d data. *3dpvt*, 0 :190, 2002.

[May79] Peter S. Maybeck. *Stochastic models, estimation, and control*, volume 141 of *Mathematics in Science and Engineering*. 1979.

[MBB07] Alexandre Meyer, Hector Briceno Pulido, and Saida Bouakaz. User-guided Shape from Shading to Reconstruct Fine Details from a Single Photograph. In *ACCV'2007 : 8th Asian Conference on Computer Vision*, October 2007.

[MBM01] Wojciech Matusik, Chris Buehler, and Leonard McMillan. Polyhedral visual hulls for real-time rendering. In *Proceedings of the 12th Eurographics Workshop on Rendering Techniques*, pages 115–126, London, UK, 2001. Springer-Verlag.

[MBR+00] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J. Gortler, and Leonard McMillan. Image-based visual hulls. In *SIGGRAPH '00 : Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 369–374, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[MBR06] Clément Ménier, Edmond Boyer, and Bruno Raffin. 3d skeleton-based body pose recovery. In *Proceedings of the 3rd International Symposium on 3D Data Processing, Visualization and Transmission, Chapel Hill (USA)*, june 2006.

[MC05] R. Motmans and E. Ceriez. Tables d'anthropométrie. Ergonomie RC, Leuven, Belgium, 2005. http ://www.dinbelg.be/anthropometrie.htm.

[Men99] Alberto Menache. *Understanding Motion Capture for Computer Animation and Video Games*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

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

[MGR98] S.J. McKenna, S.G. Gong, and Y. Raja. Modelling facial colour and identity with gaussian mixtures. 31(12) :1883–1892, December 1998.

[MH07] Gregor Miller and Adrian Hilton. Safe hulls. In *Proc. 4th European Conference on Visual Media Production*. IET, November 2007.

[MI00] John MacCormick and Michael Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *ECCV '00 : Proceedings of the 6th European Conference on Computer Vision-Part II*, pages 3–19, London, UK, 2000. Springer-Verlag.

[MLDR07] Daniel Modrow, Claudio Laloni, Guenter Doemens, and Gerhard Rigoll. 3d face scanning systems based on invisible infrared coded light. In George Bebis, Richard D. Boyle, Bahram Parvin, Darko Koracin, Nikos Paragios, Tanveer Fathima Syeda-Mahmood, Tao Ju, Zicheng Liu, Sabine Coquillart, Carolina Cruz-Neira, Torsten Müller, and Thomas Malzbender, editors, *ISVC (1)*, volume 4841 of *Lecture Notes in Computer Science*, pages 521–530. Springer, 2007.

[MM04] Philippos Mordohai and Gerard Medioni. Dense multiple view stereo with general camera placement using tensor voting. In *3DPVT '04 : Proceedings of the 3D Data Processing, Visualization, and Transmission, 2nd International Symposium on (3DPVT'04)*, pages 725–732, Washington, DC, USA, 2004. IEEE Computer Society.

[MMHM02] Karsten Muhlmann, Dennis Maier, Jurgen Hesser, and Reinhard Manner. Calculating dense disparity maps from color stereo images, an efficient implementation. *International Journal of Computer Vision*, 47 :79–88, 2002.

[MN98] Makoto Matsumoto and Takuji Nishimura. Mersenne twister : a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.*, 8(1) :3–30, 1998.

[Mor81] D.G. Morgenthaler. Three-dimensional simple points : serial erosion, parallel thinning and skeletonization - tr-1005. Technical report, Univ. of Maryland, 1981.

[MP77] D. Marr and T. Poggio. A theory of human stereo vision. Technical report, Cambridge, MA, USA, 1977.

[MS03] Marcus Magnor and Hans-Peter Seidel. Capturing the shape of a dynamic world - fast ! In *SMI '03 : Proceedings of the Shape Modeling International 2003*, page 3, Washington, DC, USA, 2003. IEEE Computer Society.

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

[MTHC03] Ivana Mikić, Mohan Trivedi, Edward Hunter, and Pamela Cosman. Human body model acquisition and tracking using voxel data. *Int. J. Comput. Vision*, 53(3) :199–223, 2003.

[Muy87] Eadweard Muybridge. *Animal Locomotion*. 1887.

[MZD03] Anurag Mittal, Liang Zhao, and Larry S. Davis. Human body pose estimation using silhouette shape analysis. In *AVSS '03 : Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, page 263, Washington, DC, USA, 2003. IEEE Computer Society.

[MZFN04] Tao Member-Zhao and Ram Fellow-Nevatia. Tracking multiple humans in complex situations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9) :1208–1221, 2004.

[NA94] Sourabh A. Niyogi and Edward H. Adelson. Analyzing and recognizing walking figures in xyt. pages 469–474, 1994.

[Nas78] Nasa. In *Anthropometric source book. Volume 1 Anthropometry for designers*. NASA Report Number : NASA-RP-1024, S-479-VOL-1, 1978.

[Nat] Microsoft Project Natal. `http ://www.xbox.com/en-US/live/projectnatal/`.

[NFA88] Hiroshi Noborio, Shozo Fukuda, and Suguru Arimoto. Construction of the octree approximating a three-dimensional object by using multiple views. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10(6) :769–782, 1988.

[Nie94] Wolfgang Niem. Robust and fast modelling of 3d natural objects from multiple views. In *SPIE Proceedings : Image and Video Processing II*, pages 388–397, 1994.

[OB80] J. O'Rourke and N.I. Badler. Model-based image analysis of human motion using constraint propagation. 2(6) :522–536, November 1980.

[OK95] R. L. Ogniewicz and O. Kübler. Hierarchic voronoi skeletons. *Pattern Recognition*, 28(3) :343–359, 1995.

[OS08] Ryuzo Okada and Björn Stenger. A single camera motion capture system for human-computer interaction. In *IEICE Transactions on Information and Systems 2008 - Special Section on Machine Vision and its Applications*, pages 1855–1862, 2008.

[OSIK06] R. Okada, B. Stenger, T. Ike, and N. Kondoh. Virtual fashion show using real-time markerless motion capture. pages II :801–810, 2006.

[PF03] Ralf Plankers and Pascal Fua. Articulated soft objects for multiview shape and motion capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9) :1182–1187, 2003.

[PF05] E. Prados and O. Faugeras. Shape From Shading : a well-posed problem ? In *Proceedings of CVPR'05*. IEEE, june 2005.

[PFP04] D. Perchet, C. I. Fetita, and F. J. Preteux. Advanced navigation tools for virtual bronchoscopy. In E. R. Dougherty, J. T. Astola, and K. O. Egiazarian, editors, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 5298 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pages 147–158, May 2004.

[PN94] Ramprasad Polana and Al Nelson. Low level recognition of human motion (or how to get your man without finding his body parts. In *In Proc. of IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects*, pages 77–82, 1994.

[Pop07] Ronald Poppe. Vision-based human motion analysis : An overview. *Comput. Vis. Image Underst.*, 108(1-2) :4–18, 2007.

[Pot87] Michael Potmesil. Generating octree models of 3d objects from their silhouettes in a sequence of images. *Comput. Vision Graph. Image Process.*, 40(1) :1–29, 1987.

[PSBM07] Valerio Pascucci, Giorgio Scorzelli, Peer-Timo Bremer, and Ajith Mascarenhas. Robust on-line computation of reeb graphs : simplicity and speed. In *SIGGRAPH '07 : ACM SIGGRAPH 2007 papers*, page 58, New York, NY, USA, 2007. ACM.

[PSF08] J. Pilet, C. Strecha, and P. Fua. Making background subtraction robust to sudden illumination changes. In *European Conference on Computer Vision*, Marseille, France, October 2008.

[PSH97] Vladimir I. Pavlovic, Rajeev Sharma, and Thomas S. Huang. Visual interpretation of hand gestures for human-computer interaction : A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7) :677–695, 1997.

[PSS01] Emil Praun, Wim Sweldens, and Peter Schröder. Consistent mesh parameterizations. In *SIGGRAPH '01 : Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 179–184, New York, NY, USA, 2001. ACM.

[Ree46] G. Reeb. Sur les points singuliers d'une forme de pfaff completement integrable ou d'une fonction numerique. *Comptes Rendus Acad. Sciences Park*, 222 :847–849, 1946.

[RF03] Deva Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. *cvpr*, 02 :467, 2003.

[RKB04] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut" : interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3) :309–314, August 2004.

[RKPL07] Myung-Cheol Roh, Tae-Yong Kim, Jihun Park, and Seong-Whan Lee. Accurate object contour tracking based on boundary edge selection. *Pattern Recogn.*, 40(3) :931–943, 2007.

[RL02] Maurice Ringer and Joan Lasenby. Multiple hypothesis tracking for automatic optical motion capture. In Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, editors, *ECCV (1)*, volume 2350 of *Lecture Notes in Computer Science*, pages 524–536. Springer, 2002.

[SB96] Alvy Ray Smith and James F. Blinn. Blue screen matting. In *SIGGRAPH '96 : Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 259–268, New York, NY, USA, 1996. ACM.

[SBF00] Hedvig Sidenbladh, Michael J. Black, and David J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV '00 : Proceedings of the 6th European Conference on Computer Vision-Part II*, pages 702–718, London, UK, 2000. Springer-Verlag.

[SBS02] Hedvig Sidenbladh, Michael J. Black, and Leonid Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *ECCV '02 : Proceedings of the 7th European Conference on Computer Vision-Part I*, pages 784–800, London, UK, 2002. Springer-Verlag.

[SC02] Josephine Sullivan and Stefan Carlsson. Recognizing and tracking human action. In *ECCV '02 : Proceedings of the 7th European Conference on Computer Vision-Part I*, pages 629–644, London, UK, 2002. Springer-Verlag.

[SC05] Aravind Sundaresan and Rama Chellappa. Markerless motion capture using multiple cameras. In *CVIIE '05 : Proceedings of the Computer Vision for Interactive and Intelligent Environment*, pages 15–26, Washington, DC, USA, 2005. IEEE Computer Society.

[SCHG04] Nicu Sebe, Ira Cohen, Thomas S. Huang, and Theo Gevers. Skin detection : A bayesian network approach. *Pattern Recognition, International Conference on*, 2 :903–906, 2004.

[SD97] Steven M. Seitz and Charles R. Dyer. Photorealistic scene reconstruction by voxel coloring. *cvpr*, 00 :1067, 1997.

[SG99] Chris Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. *cvpr*, 02 :2246, 1999.

[SH00] Maricor Soriano and Sami Huovinen. Skin detection in video under changing illumination conditions. In *In Proc. 15th International Conference on Pattern Recognition*, pages 839–842, 2000.

[SH07] J. Starck and A. Hilton. Surface capture for performance based animation. *IEEE Computer Graphics and Applications*, 27 :21–31, 2007.

[Sha91] T. Shakunaga. Pose estimation of jointed structures. pages 566–572, 1991.

[SHSI99] Wei Sun, Adrian Hilton, Raymond Smith, and John Illingworth. Layered animation of captured data. In *Animation and Simulation '99 (10th Eurographics Workshop Proceedings*, pages 145–154, 1999.

[SKvW+92] Mark Segal, Carl Korobkin, Rolf van Widenfelt, Jim Foran, and Paul Haeberli. Fast shadows and lighting effects using texture mapping. In *Proceedings of SIGGRAPH*, 1992.

[SMP03] Peter Sand, Leonard McMillan, and Jovan Popović. Continuous capture of skin deformation. *ACM Trans. Graph.*, 22(3) :578–586, 2003.

[SMP05] Tomás Svoboda, Daniel Martinec, and Tomás Pajdla. A convenient multicamera self-calibration for virtual environments. *Presence : Teleoper. Virtual Environ.*, 14(4) :407–422, 2005.

[SP91] K. Shanmukh and Arun K. Pujari. Volume intersection with optimal set of directions. *Pattern Recogn. Lett.*, 12(3) :165–170, March 1991.

[SP95] Thad Starner and Alex Pentland. Real-time american sign language recognition from video using hidden markov models. *iscv*, 00 :265, 1995.

[SPB04] Joaquim Salvi, Jordi Pagès, and Joan Batlle. Pattern codification strategies in structured light systems. *Pattern Recognition*, 37 :827–849, 2004.

[SSA00] Leonid Sigal, Stan Sclaroff, and Vassilis Athitsos. Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2 :2152, 2000.

[SSA04] Leonid Sigal, Stan Sclaroff, and Vassilis Athitsos. Skin color-based video segmentation under time-varying illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7) :862–877, 2004.

[ST03] Cristian Sminchisescu and Bill Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotics Research*, 22 :371–392, 2003.

[ST05a] Cristian Sminchisescu and Bill Triggs. Building roadmaps of minima and transitions in visual models. *Int. J. Comput. Vision*, 61(1) :81–101, 2005.

[ST05b] Cristian Sminchisescu and Bill Triggs. Building roadmaps of minima and transitions in visual models. *Int. J. Comput. Vision*, 61(1) :81–101, 2005.

[STW07] Caifeng Shan, Tieniu Tan, and Yucheng Wei. Real-time hand tracking using a mean shift embedded particle filter. *Pattern Recogn.*, 40(7) :1958–1970, 2007.

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

[SWP98] Thad Starner, Joshua Weaver, and Alex Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12) :1371–1375, 1998.

[SYN+00] A. SEPOU, MC. YANZA, E. NGUEMBI, R. NGBALE, G. KOURIAH, A. KOUABOSSO, A. PENGUELE, F. NADJI-ADJIM, and K. KALAMBAY. Importance des mesures anthropométriques dans la pratique centrafricaine des soins de santé primaire. *Médecine d'afrique noire*, 7(47) :328–332, 2000.

[Sys] Vicon Motion Systems. `http ://www.vicon.com`.

[SZ00] Richard Szeliski and Ramin Zabih. An experimental comparison of stereo algorithms. In *ICCV '99 : Proceedings of the International Workshop on Vision Algorithms*, pages 1–19, London, UK, 2000. Springer-Verlag.

[Sze93] Richard Szeliski. Rapid octree construction from image sequences. *CVGIP : Image Underst.*, 58(1) :23–32, 1993.

[TdAM+04] Christian Theobalt, Edilson de Aguiar, Marcus A. Magnor, Holger Theisel, and Hans-Peter Seidel. Marker-free kinematic skeleton estimation from sequences of volume data. In *VRST '04 : Proceedings of the ACM symposium on Virtual reality software and technology*, pages 57–64, New York, NY, USA, 2004. ACM.

[TFAS00] Jean-Christophe Terrillon, Hideo Fukamachi, Shigeru Akamatsu, and Mahdad Nouri Shirazi. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *FG*, pages 54–63. IEEE Computer Society, 2000.

[TIKM07] Masahiro Toyoura, Masaaki Iiyama, Koh Kakusho, and Michihiko Minoh. Silhouette extraction with random pattern backgrounds for the volume intersection method. In *3DIM '07 : Proceedings of the Sixth International Conference on 3-D Digital Imaging and Modeling*, pages 225–232, Washington, DC, USA, 2007. IEEE Computer Society.

[TKBM99] Kentaro Toyama, John Krumm, Barry Brumitt, and Brian Meyers. Wallflower : Principles and practice of background maintenance. In *ICCV (1)*, pages 255–261, 1999.

[TM00] Ben Tordoff and David W Murray. Violating rotating camera geometry : The effect of radial distortion on self-calibration. *icpr*, 01 :1423, 2000.

[TMSS02] Christian Theobalt, Marcus Magnor, Pascal Schüler, and Hans-Peter Seidel. Combining 2d feature tracking and volume reconstruction for online video-based human motion capture. In *PG '02 : Proceedings of the 10th Pacific Conference on Computer Graphics and Applications*, page 96, Washington, DC, USA, 2002. IEEE Computer Society.

[TS06] T. Tangkuampien and D. Suter. Real-time human pose inference using kernel principal component pre-image approximations. In *In British Machine Vision Conference*, 2006.

[TSJ92] Arun P. Tirumalai, Brian G. Schunck, and Ramesh C. Jain. Dynamic stereo with self-calibration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(12) :1184–1189, 1992.

[TVD07a] Julien Tierny, Jean-Philippe Vandeborre, and Mohamed Daoudi. Enhancing 3d mesh topological skeletons with discrete contour constrictions. *The Visual Computer*, 2007.

[TVD07b] Julien Tierny, Jean-Philippe Vandeborre, and Mohamed Daoudi. Topology driven 3D mesh hierarchical segmentation. In *IEEE International Conference on Shape Modeling and Applications (SMI 2007)*, pages 215–220, Lyon, France, 2007.

[TVD08] Julien Tierny, Jean-Philippe Vandeborre, and Mohamed Daoudi. Partial 3d shape retrieval by reeb pattern unfolding. *Computer Graphics Forum (to appear)*, 2008.

[VBK02] Sundar Vedula, Simon Baker, and Takeo Kanade. Spatio-temporal view interpolation. In *EGRW '02 : Proceedings of the 13th Eurographics workshop on Rendering*, pages 65–76, Aire-la-Ville, Switzerland, Switzerland, 2002. Eurographics Association.

[VCT06] Antoine Vacavant, David Coeurjolly, and Laure Tougne. Topological and Geometrical Reconstruction of Complex Objects on Irregular Isothetic Grids. In Springer Verlag, editor, *13th International Conference on Discrete Geometry for Computer Imagery*, Lecture Notes in Compter Science, pages 470–481, October 2006.

[VF92] Régis Vaillant and Olivier D. Faugeras. Using extremal boundaries for 3-d object modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2) :157–173, 1992.

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

[WB01] Greg Welch and Gary Bishop. An introduction to the kalman filter. Technical report, 2001.

[WCH92] J. Weng, P. Cohen, and M. Herniou. Camera calibration with distortion models and accuracy evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(10) :965–980, 1992.

[WSZL08] Jianhua Wang, Fanhuai Shi, Jing Zhang, and Yuncai Liu. A new calibration model of camera lens distortion. *Pattern Recogn.*, 41(2) :607–615, 2008.

[XSW04] Yijun Xiao, Paul Siebert, and Naoufel Werghi. Topological segmentation of discrete human body shapes in various postures based on geodesic distance. In *ICPR '04 : Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3*, pages 131–135, Washington, DC, USA, 2004. IEEE Computer Society.

[YAT00] S. Yonemoto, D. Arita, and R. Taniguchi. Real-time human motion analysis and ik-based human figure control. In *HUMO '00 : Proceedings of the Workshop on Human Motion (HUMO'00)*, page 149, Washington, DC, USA, 2000. IEEE Computer Society.

[YHHGBG03] Danny B. Yang, nos Héctor H. González-Ba and Leonidas J. Guibas. Counting people in crowds with a real-time network of simple image sensors. In *ICCV '03 : Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 122, Washington, DC, USA, 2003. IEEE Computer Society.

[yKWMC01] Kwan yee K. Wong, Paulo R. S. Mendonça, and Roberto Cipolla. Head model acquisition from silhouettes, May 2001.

[YW07] Z.W. Yu and H.S. Wong. Fast gaussian mixture clustering for skin detection. pages IV : 341–344, 2007.

[ZH04] Yinan Zhou and Zhiyong Huang. Decomposing polygon meshes by means of critical points. In *MMM '04 : Proceedings of the 10th International Multimedia Modelling Conference*, page 187, Washington, DC, USA, 2004. IEEE Computer Society.

[Zha00] Z. Zhang. A flexible new technique for camera calibration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(11) :1330–1334, 2000.

[Ziv04] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *ICPR 2004. Proceedings of the 17th International Conference on Pattern Recognition*, volume 2, pages 28–31 Vol.2, 2004.

[ZM96] ChangSheng Zhao and Roger Mohr. Global three-dimensional surface reconstruction from occluding contours. *Comput. Vis. Image Underst.*, 64(1) :62–96, 1996.

[ZMT05] Eugene Zhang, Konstantin Mischaikow, and Greg Turk. Feature-based surface parameterization and texture mapping. *ACM Trans. Graph.*, 24(1) :1–27, 2005.

[ZNL01] Tao Zhao, Ramakant Nevatia, and Fengjun Lv. Segmentation and tracking of multiple humans in complex situations. In *CVPR (2)*, pages 194–201. IEEE Computer Society, 2001.

[ZSQ99] Benjamin D. Zarit, Boaz J. Super, and Francis K. H. Quek. Comparison of five color models in skin pixel classification. In *RATFG-RTS '99 : Proceedings of the International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, page 58, Washington, DC, USA, 1999. IEEE Computer Society.

[ZTS02] Emanuel Zuckerberger, Ayellet Tal, and Shymon Shlafman. Polyhedral surface decomposition with applications. *Computers & Graphics*, 26(5) :733–743, 2002.

Publications dans le cadre de la thèse

[BGM08] Mathieu Barnachon, Erwan Guillou, and Brice Michoud. Reconstruction Géométrique et Photométrique Incrémentale d'un Personnage en Mouvement à partir de Séquences Vidéo. In *AFIG 2008*, November 2008.

[BMGB09] Mathieu Barnachon, Brice Michoud, Erwan Guillou, and Saida Bouakaz. Reconstruction géométrique par estimation de posture. In *ORASIS*, June 2009.

[FFM08] François Fouquet, Jean-Philippe Farrugia, and Brice Michoud. Extraction d'un Environnement Photométrique et Géométrique pour la Réalité Augmentée. In Loïc Barthe Mathias Paulin, editor, *AFIG 2008 - 21èmes Journées de l'Association Française d'Informatique Graphique*, pages 227–236. IRIT Presse, November 2008.

[MBGP08] Brice Michoud, Saida Bouakaz, Erwan Guillou, and Hector Briceno Pulido. Largest Silhouette-Equivalent Volume for 3D Shapes Modeling without Ghost Object. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications, In conjunction with ECCV 2008*, October 2008.

[MGB05a] Brice Michoud, Erwan Guillou, and Saida Bouakaz. Human model and pose Reconstruction from Multi-views. In *International Conference on Machine Intelligence (ACIDCA-ICMI)*, November 2005.

[MGB05b] Brice Michoud, Erwan Guillou, and Saida Bouakaz. Reconstruction d'humanoïde à partir de plusieurs vues. In *Journées AFIG 2005*, pages 171–181, November 2005.

[MGB06a] Brice Michoud, Erwan Guillou, and Saida Bouakaz. Extension de l'espace d'acsuisition pour les méthodes de Shape From Silhouette. In *COmpression et REprésentation de Signaux Audiovisuels (CORESA) 2006*, November 2006.

[MGB06b] Brice Michoud, Erwan Guillou, and Saida Bouakaz. Shape From Silhouette : Towards a Solution for Partial Visibility Problem. In D.Fellner C.Hansen, editor, *Eurographics 2006*, Eurographics 2006 Short Papers Preceedings, pages 13–16, September 2006.

[MGB07a] Brice Michoud, Erwan Guillou, and Saida Bouakaz. Real-time and Markerless 3D Human Motion Capture using Multiple Views. In *2nd Human Motion Workshop*, October 2007. Taux d'acceptation : (29%).

[MGB07b] Brice Michoud, Erwan Guillou, and Saida Bouakaz. Real-time and Markerless Full-Body Human Motion Capture. In *Groupe de Travail Animation et Simulation 2007 (GTAS'07)*, June 2007.

[MGB+09] Brice Michoud, Erwan Guillou, Saida Bouakaz, Mathieu Barnachon, and Alexandre Meyer. Towards Removing Ghost-Components from Visual-Hull Estimations. In *ICIG*, September 2009.

[MGPB07a] Brice Michoud, Erwan Guillou, Hector Briceno Pulido, and Saida Bouakaz. Real-time and Marker-free 3D Motion capture for Home Entertainment Oriented Applications. In *ACCV'2007 : 8th Asian Conference on Computer Vision*, November 2007. Taux d'acceptation : (32%).

[MGPB07b] Brice Michoud, Erwan Guillou, Hector Briceno Pulido, and Saida Bouakaz. Real-Time Marker-free Motion Capture from multiple cameras. In *ICCV'2007 : Eleventh IEEE International Conference on Computer Vision*, October 2007. Taux d'acceptation : (23.5%).

[MGPB08a] Brice Michoud, Erwan Guillou, Hector Briceno Pulido, and Saida Bouakaz. Removing Camera Placement Constraints in Shape from Silhouette on Large Acquisition Volumes. Technical Report RR-LIRIS-2008-016, LIRIS UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumière Lyon 2/Ecole Centrale de Lyon, July 2008.

[MGPB08b] Brice Michoud, Erwan Guillou, Hector Briceno Pulido, and Saida Bouakaz. Silhouettes Fusion for 3D Shapes Modeling with Ghost Object Removal. Technical Report RR-LIRIS-2008-015, LIRIS UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumière Lyon 2/Ecole Centrale de Lyon, July 2008.

Annexe A

Modélisation de la caméra

Lorsqu'une caméra est utilisée pour extraire des information géométriques contenues dans une scène, l'une des tâches les plus importantes est de lier les coordonnées image des éléments de la scène avec leur coordonnées spatiales. Le processus de calibrage consiste à déterminer pour une caméra, les éléments qui gouvernent la transformation de passage existant entre les coordonnées spatiales de la scène et les coordonnées image. Ainsi le calibrage est un processus important principalement pour les raisons suivantes :

- il permet de mesurer précisément les positions spatiales des objets de la scène
- il estime la position de la caméra par rapport à la scène

La formation de l'image correspond à une transformation Proj_π, qui à partir du point de la scène $\text{P}(X, Y, Z)$ fournit un point image $\text{p}(u, v)$. L'étape de calibrage consiste de manière générale à estimer la transformation Proj_π telle que $\text{p} = \text{Proj}_\pi(\text{P})$. La littérature propose de nombreuses méthodes permettant d'approximer cette fonction. Elles possèdent en commun :

- Une modélisation de la caméra : le comportement de la caméra est modélisé en utilisant des éléments de géométrie projective et des caractéristiques propres aux optiques. Ces informations propres à la caméra sont dits *paramètres de calibrage*.
- L'utilisation d'un objet de calibrage : un ensemble de points 3D de référence (positions relatives ou absolues connues précisément) sont présentés à la caméra. Ils sont liés aux points image correspondants perçus par la caméra. Cette association permet de former un ensemble d'équations faisant intervenir les paramètres de calibrage.
- L'approximation de la fonction de transfert : la fonction Proj_π est approximée par des techniques de résolutions dépendantes de la complexité du modèle de caméra défini.

Parmi les paramètres de calibrage, il existe typiquement deux classes de paramètres qui définissent la transformation de passage des points de l'espace aux points de l'image.

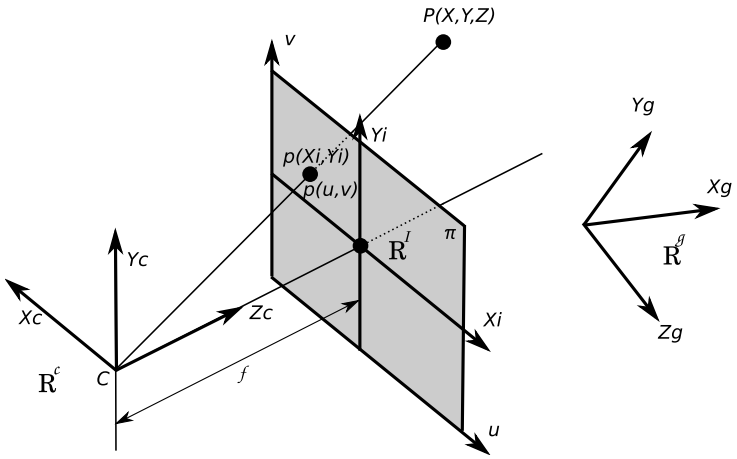


FIG. A.1 – Représentation du modèle de caméra à sténopé ou encore appelé caméra à projection perspective pure.

Les paramètres intrinsèques

Les paramètres intrinsèques caractérisent les propriétés inhérentes à la caméra et aux optiques :

- la position du centre image ;
- les facteurs d'échelle suivant les deux axes image ;
- la distance focale effective ;
- les coefficients de distortion des lentilles.

Les paramètres extrinsèques

Les paramètres extrinsèques donnent la position et l'orientation de la caméra par rapport au système de coordonnées de la scène. Ceux-ci peuvent être exprimés par :

- une translation
- une rotation suivant chacun des trois axes du repère de la scène.

Modèle interne de la caméra

Le modèle de caméra le plus communément utilisé est le modèle de caméra dit à sténopé. Ce modèle caractérise la transformation permettant le passage du repère $\mathcal{R}^{\mathcal{C}}$ lié à la caméra au repère image $\mathcal{R}^{\mathcal{I}}$ (voir figure A.1).

Ceci revient à modéliser chaque caméra $\mathcal{C}$ par son plan optique π et son centre optique C. Les caractéristiques de ce modèle sont les suivantes :

– un axe optique $[CZ_c)$ perpendiculaire au plan image π

– un centre C localisé sur l'axe optique à une distance focale f du plan image.

On suppose que la caméra réalise une transformation perspective pure de centre C à une distance f du plan image. Par conséquent la relation entre un point $\mathrm{P}(X_c, Y_c, Z_c)$ dans le repère $\mathcal{R}^{\mathcal{C}}$ caméra et sa projection $\mathrm{p}(X_i, Y_i)$ sur le plan image π est donnée par la relation suivante :

$$\frac{X_c}{X_i} = \frac{Y_c}{Y_i} = \frac{Z_c}{f} \tag{A.1}$$

La relation A.1 peut ainsi s'écrire sous forme matricielle en utilisant les coordonnées homogènes :

$$\begin{bmatrix} sX_i \\ sY_i \\ s \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \tag{A.2}$$

Cette relation fournit notamment $s = Z_c$.

Pour une caméra numérique, le plan caméra sur lequel vient se former l'image est en fait divisé en trames rectangulaires possédant des pas d'échantillonnage suivant les directions verticale et horizontale (voir figure A.2). La relation entre les coordonnées pixels (u, v) et les coordonnées du plan image (X_i, Y_i) est donnée par :

$$\begin{cases} X_i = \frac{u-u_0}{k_u} \\ Y_i = \frac{v-v_0}{k_v} \end{cases} \tag{A.3}$$

où u_0 et v_0 sont les coordonnées pixels de l'origine du repère image, k_u et k_v sont respectivement les pas d'échantillonnage horizontal et vertical.

A partir des équations A.2 et A.3, on peut établir une relation entre les coordonnées pixels et les coordonnées liées au repère caméra.

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \begin{bmatrix} k_u f & 0 & u_0 & 0 \\ 0 & k_v f & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \tag{A.4}$$

où f, u_0, v_0, k_u et k_v représentent les paramètres intrinsèques de la caméra.

Les paramètre intrinsèques permettent de relier les points image en pixels, aux points de la scène perçus dans le repère liée à la caméra. Leur détermination constitue la première tâche du processus de calibrage

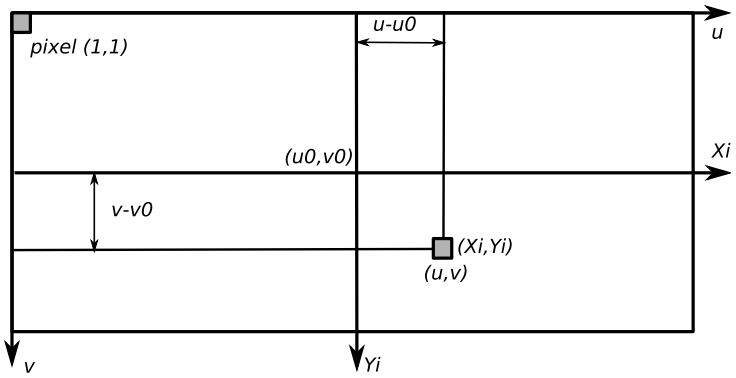


FIG. A.2 – Relation entre les coordonnées pixel (u, v) et le coordonnées du plan image (X_i, Y_i).

Modèle externe de la caméra

La modélisation externe permet de relier un repère absolu 3D de la scène au repère 3D lié à la caméra dans le but de connaître la position de cette dernière dans la scène. Pour cela, on doit établir préalablement la relation entretenue entre le repère $\mathcal{R}^{\mathcal{C}}$ et le repère absolu $\mathcal{R}^{\mathcal{G}}$:

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = R \begin{bmatrix} X_g \\ Y_g \\ Z_g \end{bmatrix} + T \tag{A.5}$$

où R et T sont les composantes rotation et translation du déplacement rigide qui font coïncider les deux repères $\mathcal{R}^{\mathcal{G}}$ et $\mathcal{R}^{\mathcal{C}}$.

En coordonnées homogènes, cette relation s'écrit :

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = \begin{bmatrix} R & T \\ 0\ 0\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_g \\ Y_g \\ Z_g \\ 1 \end{bmatrix} \tag{A.6}$$

où R et T définissent les paramètres extrinsèques de la caméra.

Le modèle global de la caméra

Le modèle global d'une caméra permet de déterminer la projection d'un point connu dans le repère absolu $\mathcal{R}^{\mathcal{G}}$ sur le plan image π de la caméra $\mathcal{C}$. La transformation entre les coordonnées des points dans le repère absolu et les coordonnées pixels est donnée par :

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \begin{bmatrix} K_u f & 0 & u_0 & 0 \\ 0 & K_v f & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ 0\ 0\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_g \\ Y_g \\ Z_g \\ 1 \end{bmatrix} \tag{A.7}$$

ou de manière plus concise :

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = M \begin{bmatrix} X_g \\ Y_g \\ Z_g \\ 1 \end{bmatrix} \tag{A.8}$$

où M est une matrice 3×4 représentant le modèle global de la caméra. La matrice M est dite matrice de calibrage de la caméra. Calibrer une caméra revient à déterminer la matrice M pour une caméra et une configuration donnée.

Correction des erreurs de distortion

Lors du processus de projection sur le plan π, les optiques de caméras induisent des déformations. De nombreuses méthodes se sont attaché à corriger les distortions optiques [WSZL08, TM00]. Zhang propose dans [Zha00] une méthode de correction des distortions optiques radiales, dues à la symétrie des lentilles, et tangentielles issues d'un mauvais alignement de la lentille. La modélisation standard [WCH92] des distortions radiales et tangentielles des optiques de caméra définit l'association entre les coordonnées pixel distordues (u, v) observables, aux coordonnées corrigées (u_{corri}, v_{corri}) qui ne sont pas physiquement observable, par la transformation suivante :

$$\begin{cases} u_{corri} = u + (u - u_0)(K_1 r^2 + K2r^4) + 2P_1(u - u_0)(v - v_0) + P_2(r^2 + 2(u - u_0)^2) \\ v_{corri} = v + (v - v_0)(K_1 r^2 + K2r^4) + 2P_2(u - u_0)(v - v_0) + P_1(r^2 + 2(v - v_0)^2) \end{cases} \tag{A.9}$$

où $r^2 = (u - u_0)^2 + (v - v_0)^2$.

Notons que les paramètres P_1 et P_2 correspondent aux distorsions tangentielles, alors que les paramètres K_1 et K_2 définissent les distorsions radiales. r est le rayon d'un point image par rapport au centre de distortion supposé en (u_0, v_0). Seuls quelques paramètre de distortions sont recherchés (K_1, K_2, P_1 et P_2) car les paramètres d'ordre supérieur ont en général une influence négligeable [WCH92].

Nous avons fait le choix d'utiliser les travaux de Zhang présentés dans [Zha00]. Cette approche fournit une flexibilité de calibrage intéressante et une précision suffisante.

La notation $\text{Proj}_{\pi_i}(\text{P})$ représente la transformation qui, à partir du point de la scène $\text{P}(X, Y, Z)$ fournit un point image $\text{p}(u_{corri}, v_{corri})$ pour la caméra $\mathcal{C}_i$. Ainsi lorsque nous utilisons cette notion par rapport à un modèle de caméra théorique, nous supposons $u_{corri} = u$ et $v_{corri} = v$.

Annexe B

Ellipsoïde d'inertie d'une ensemble de points 3D

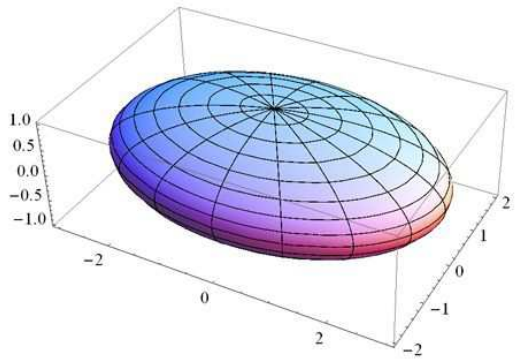


FIG. B.1 – Représentation d'une ellipsoïde.

Rappelons la définition de l'ellipsoïde d'inertie d'une ensemble de points $\mathcal{V}$ (dans notre cas l'ensemble des voxels représentés par leur centres) avec $v = (x_v, y_v, z_v) \in \mathcal{V}$. Soient (x_E, y_E, z_E), R_E et $(\alpha_E, \beta_E, \gamma_E)$ respectivement le centre de l'ellipsoïde $\mathcal{E}_\mathcal{V}$, la matrice de rotation passant du repère global vers le repère de l'ellipsoïde et $(\alpha_E, \beta_E, \gamma_E)$ la longueur des trois axes de l'ellipsoïde.

Ces paramètres sont estimés à partir des moments de $\mathcal{V}$ d'ordre zéro, un et deux :

$$M_0 = \sum_{v \in E} 1 \tag{B.1}$$

$$M_u = \frac{1}{M_0} \sum_{v \in E} u_v \tag{B.2}$$

$$M_{uw} = \frac{1}{M_0} \sum_{v \in E} u_v w_v \tag{B.3}$$

avec $(u, w) \in (x, y, z)^2$.

A partir de ces équations, les paramètres de l'ellipsoïde sont donnés par :

$$(x_E, y_E, z_E) = (M_x, M_y, M_z) \tag{B.4}$$

$$M = 4 \begin{bmatrix} M_{xx} & M_{xy} & M_{xz} \\ M_{yx} & M_{yy} & M_{yz} \\ M_{zx} & M_{zy} & M_{zz} \end{bmatrix} \tag{B.5}$$

$$= R_E \begin{bmatrix} \alpha_E^2 & 0 & 0 \\ 0 & \beta_E^2 & 0 \\ 0 & 0 & \gamma_E^2 \end{bmatrix} R_E^T \tag{B.6}$$

avec M la matrice de moments, (x_E, y_E, z_E) le centre de l'ellipsoïde, R_E sa matrice de rotation, et $(\alpha_E, \beta_E, \gamma_E)$ son vecteur d'échelle. R_E et $(\alpha_E, \beta_E, \gamma_E)$ se calculent par la décomposition de M en vecteurs propres et valeurs propres.