Extension de l'espace d'acquisition pour les méthodes de Shape-from-silhouette

B. Michoud[1] E. Guillou[1] S. Bouakaz[1]

[1] LIRIS, Laboratoire d'InfoRmatique en Images et Systèmes d'information

Université Claude Bernard, Lyon 1
43, Boulevard du 11 novembre 1918
69622 Villeurbanne Cedex

{bmichoud, eguillou, sbouakaz}@liris.cnrs.fr

Résumé

L'acquisition de la forme tridimensionnelle d'un personnage est une étape indispensable pour un grand nombre d'applications de réalité virtuelle, augmentée et dans la conception de jeux vidéos. Celle-ci soit être complète et précise pour offrir le meilleur réalisme possible. Les méthodes dites "Shape From Silhouette" (SFS) permettent d'obtenir cette estimation en temps réel à partir de plusieurs caméras. L'une des limitations de ces méthodes est que le personnage doit être entièrement visible dans toutes les caméras pour être reconstruit entièrement. Dans cet article nous proposons une extension à SFS qui permet de reconstruire une estimation 3d de la forme d'un objet même s'il sort du champ de vision d'une ou plusieurs caméras.

Mots clefs

Réalité virtuelle, réalité augmentée, reconstruction géométrique

1 Introduction

Ce travail s'insère dans un projet de réalité augmentée dont l'un des objectifs est l'insertion en temps réel, d'un personne réelle filmée par plusieurs caméras, dans un décor virtuel. Pour assurer une insertion la plus réaliste possible, il est important de modéliser précisément les interactions géométriques et photométriques entre la personne et son environnement. Pour cela il est nécessaire de disposer d'une représentation tridimensionnelle de la personne.

La littérature propose un grand nombre de méthodes permettant l'acquisition de la forme d'une personne. L'une des approches les plus populaire est celle connue sous le nom de "Shape-from-silhouette" que nous noterons SFS dans la suite de cet article. Parmi les travaux consacrés à SFS, certains proposent l'acquisition d'une forme humaine en temps réel. On peut citer ceux de l'équipe de Kong Man Cheung [1] qui a été dans les premières à proposer en 2000 un algorithme temps réel. Depuis d'autres méthodes [2, 3, 4, 5] ont permis d'obtenir du temps réel en utilisant principalement les outils proposés par les cartes graphiques programmables [6, 7].

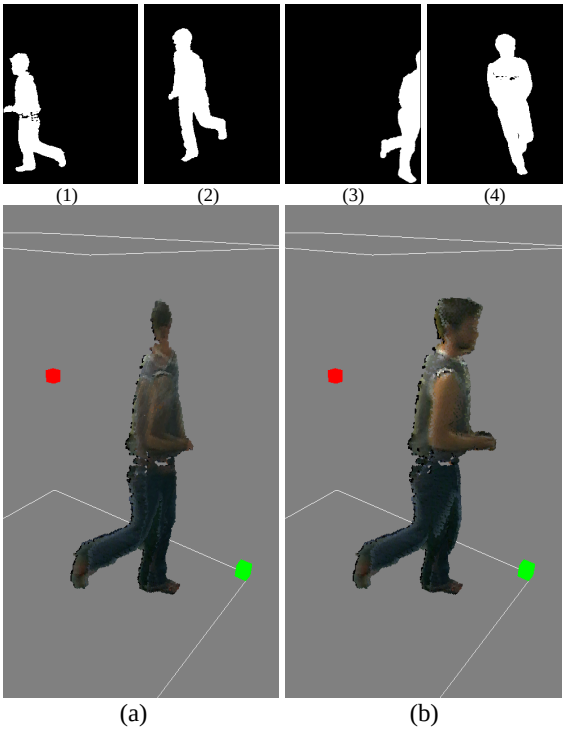

Figure 1 – Différences de reconstruction d'un objet lorsque celui-ci sort du champs de vision d'une caméra : (a) en utilisant les méthodes basées SFS actuelles ; (b) en utilisant notre algorithme. Le coloriage voxélique (réalisé par un lancer de rayon) n'est donné que pour faciliter la compréhension des images.

A partir des silhouettes[1] d'un l'objet, les algorithmes SFS permettent d'estimer la forme 3d de cet objet. Si ces méthodes permettent de retrouver rapidement une forme globale du sujet filmé, l'une de ses limites réside dans la contrainte que seules les parties visibles depuis toutes les caméras à chaque instant, peuvent être reconstruites. Dans le cas où le sujet à reconstruire est une personne en

[1] images binaires associées aux images acquises d'un objet, où 1 représente l'objet et 0 représente le reste.

mouvement, il est difficile, à moins d'avoir des caméras haute définition placées loin de l'objet, d'assurer une visibilité totale dans chaque caméra. La figure 1 illustre cette contrainte, une grande partie de la personne n'est pas vue dans la silhouette 3 et n'est donc pas reconstruite (Figure 1.a).
Dans cet article, nous proposons une extension de l'algorithme SFS qui permet de pallier ce problème, en supposant que l'objet acquis soit partiellement visible depuis toutes les caméras. Après un court résumé des principes de SFS, le chapitre 3 présente notre extension de SFS qui permet de reconstruire une estimation 3d de la forme d'un objet O même s'il sort du champ de vision d'une ou plusieurs caméras. Nous discutons ensuite des résultats obtenus. Enfin dans le chapitre 5 nous concluons sur le travail effectué et proposons certaines perspectives de ce travail.

2 Méthodologie des algorithmes basés SFS

Les méthodes basées SFS sont fréquemment utilisées pour calculer une estimation 3d de la forme d'un objet. Le formalisme de construction de la VH d'un objet a été introduit par A. Laurentini [8].

Il peut être décrit comme suit :
Soit un objet 3d O filmé par n caméras cam_i. M_i est la matrice de projection associée à la caméra cam_i et I_i l'image acquise depuis cette caméra. Enfin, S_i est l'image de silhouette associée à I_i.
Soit un point 3d P. Si celui-ci est contenu dans le volume de O alors il se projette dans toutes les silhouettes :

$$\forall i = 1, ..., n \ , \ \exists p_i \in S_i \ , \ p_i = M_i.P.$$

où p_i est la projection de P sur la silhouette S_i.
La VH de O est alors définie comme le volume contenant l'ensemble des points 3d se projetant sur toutes les silhouettes S_i. Il y a principalement deux méthodes pour calculer la VH d'un objet O, que nous allons maintenant détailler.

L'approche basée surface

La VH d'un objet déduite d'un ensemble de n images de silhouette, est construite à partir de l'intersection des n cônes de silhouette. Ces cônes sont définis par la projection, dans l'espace 3d, des contours des silhouettes à travers le centre de projection de la caméra associée. Ainsi, la VH d'un objet O sera décrite par un ensemble de surfaces 2d, ces surfaces sont définies par l'intersection des surfaces des cônes de silhouette.
D'un coté, cette approche permet des calculs en temps réel [2, 7, 9]. De l'autre, les résultats obtenus ne sont pas utilisables pour calculer les informations volumiques nécessaires à la mise en correspondance avec des modèles génériques d'humanoïdes (utilisés pour l'estimation de la posture et l'interprétation de mouvement de la personne acquise) [10].

L'approche basée volume

Une approche équivalente définie la VH d'un objet O comme étant le volume maximum qui se projette exactement sur toutes les silhouettes de O [8]. Basé sur cette définition, l'approche le plus utilisée [1, 3, 6, 4, 5] calcule une estimation de la VH de O par un ensemble de voxels.
La zone d'acquisition est partitionnée en m voxels V_j où $j = 1, ..., m$. Soit v_{ij} l'ensemble des pixels de I_i sur lesquels se projetteV_j :

$$v_{ij} = (M_i.V_j) \cap I_i.$$

Le nombre nb_j de silhouettes sur lesquelles se projette V_j est défini par :

$$nb_j = Card\{v_{ij} \ , \ v_{ij} \cap S_i \neq \emptyset\}.$$

Nous définissons $CS(O)$ la carte de silhouette de l'objet O comme étant l'ensemble de tous les couples (V_j, nb_j).
Si un voxel V_j se projette sur toutes les silhouettes de O alors il appartient à sa VH. Ainsi, $CS(O)$ peut être divisée en n sous-ensembles SFS_i :

$$SFS_i = \bigcup_{j=1}^{m} (V_j \ , \ nb_j = i).$$

L'estimation voxélique de la VH de O est alors définie par SFS_n où n est le nombre de caméras utilisées (voir Figure 2).

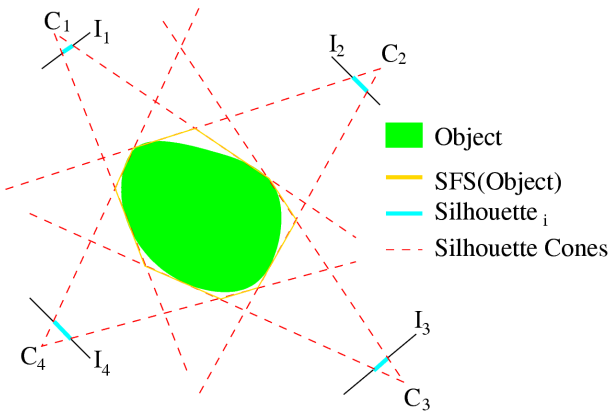

Figure 2 – Représentation 2d d'un objet O vu par 4 caméras, la VH correspondante et son estimation voxélique.

Si certaines parties du corps d'une personne en mouvement ne sont plus visibles depuis l'une des caméras, alors les informations correspondantes ne seront pas estimées dans sa VH. Pourtant, ces informations pourraient être obtenues à partir des autres caméras. Dans la suite, nous utiliserons ce postulat pour ajouter de nouvelles informations à la VH d'un objet.

3 Contributions

Dans la plupart des méthodes basées SFS, la zone d'acquisition, où un objet est reconstruit, correspond à l'intersection des cônes de vision des caméras. Les limitations de cette approche proviennent du fait que :

– la zone d'acquisition ne peut être étendue au delà de l'intersection des cônes de vision, notamment lorsqu'un grand nombre de caméras sont utilisées ;
– Il est difficile, pour une personne en mouvement, de rester visible à tout moment depuis toutes les caméras.

Pour éviter cela, nous allons prendre en compte le nombre potentiel de caméras qui voient un voxel. Si un voxel est visible depuis seulement $n-k$ caméras (où $k \in [n_{min}, ..., n-1]$ avec $n_{min} < n$) sur les n disponibles, alors il se projettera sur un maximum de $n-k$ silhouettes. Mais dans les méthodes actuelles basées SFS, il est nécessaire qu'un voxel soit vu par toutes les caméras ($nb_j = n$) pour qu'il soit contenu dans l'estimation voxélique de la VH d'un objet.

Dans un premier temps, nous utiliserons ce concept pour calculer une estimation de la forme d'un objet dans l'espace non visible depuis k caméras. Puis nous utiliserons les propriétés de connexité de l'objet acquis, afin de choisir les informations pertinentes pour étendre sa VH.

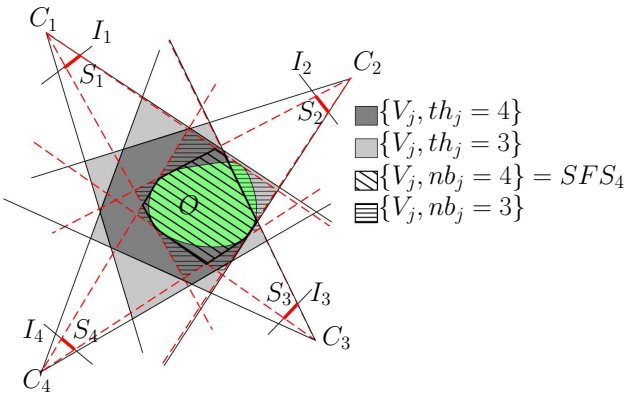

Figure 3 – Représentation 2d d'un objet O et son estimation en utilisant SFS_4. Les voxels qui étendent la forme de O sont ceux pour lesquels $th_j = nb_j = 3$.

Soit th_j le nombre d'images sur lesquelles se projette le voxel V_j :

$$th_j = Card\{v_{ij}\ ,\ v_{ij} \cap I_i \neq \emptyset\}.$$

La carte de projection $CP(O)$ d'un objet O est alors définie comme étant l'ensemble de tous les couples (V_j, th_j).

Si un voxel V_j est contenu dans le volume de O, il se projette alors sur th_j images et nb_j silhouettes ; ainsi $th_j = nb_j$.

Nous cherchons l'ensemble des voxels V_j contenus dans le volume d'un objet O. Pour cela nous comparons $CS(O)$ et $CP(O)$:

– si $th_j \neq nb_j$ alors V_j n'est pas contenu dans le volume de O ;
– sinon V_j est potentiellement contenu dans le volume de O (Figure 3).

L'ensemble $R = \{V_j, nb_j = th_j\}$ de tous les voxels potentiels peut être séparé en n sous-ensembles R_i :

$$R_i = \{V_j\ ,\ nb_j = th_j = i\}.$$

Notons que

$$SFS_n = \bigcup_{V_j \in R_n} \{V_j\}.$$

Ainsi, pour étendre SFS_n, nous choisirons des voxels contenus dans les sous-ensembles R_k avec $k \in [n_{min}, ..., n-1]$ et $n_{min} < n$. Soit $\Re_{n_{min}}$ l'union de tous les R_k :

$$\Re_{n_{min}} = \bigcup_{k=n_{min}}^{n-1} R_k.$$

Un objet 3d est connexe ainsi l'estimation 3d de sa forme doit aussi être connexe. $\Re_{n_{min}}$ est l'union de L composantes connexes notées c_l où $l = 1, ..., L$. Afin de satisfaire la contrainte de connexité du volume reconstruit, nous choisissons les composantes connexes de $\Re_{n_{min}}$ qui sont connectées à SFS_n (comme décrit dans la figure 4).

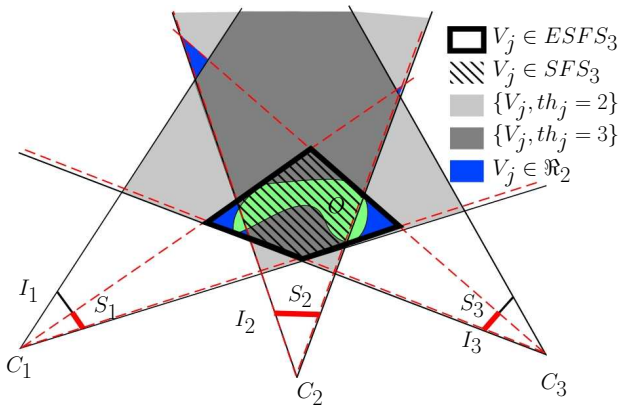

Figure 4 – Représentation 2d de O, son estimation volumique à partir de SFS_3 et de $ESFS_3$. Nous notons que $ESFS_3$ est plus précis que SFS_3.

Soit $C_{n_{min}}$ l'ensemble des composantes connexes de $\Re_{n_{min}}$ connectées à R_n :

$$C_{n_{min}} = \bigcup_{l=1}^{L} (c_l\ ,\ connexe(c_l \cup R_n)).$$

$ESFS_n$ défini le volume estimé à partir de notre algorithme, et étendant SFS_n :

$$ESFS_n = SFS_n \cup C_{n_{min}}.$$

L'estimation de forme définie par $ESFS_n$ dépend de la valeur de n_{min} : une partie de l'objet filmé sera reconstruite

si n_{min} caméras peuvent la voir. A. Laurentini [8] a montré que plus le nombre de caméras utilisé est grand, meilleure est l'estimation de la forme d'un objet.

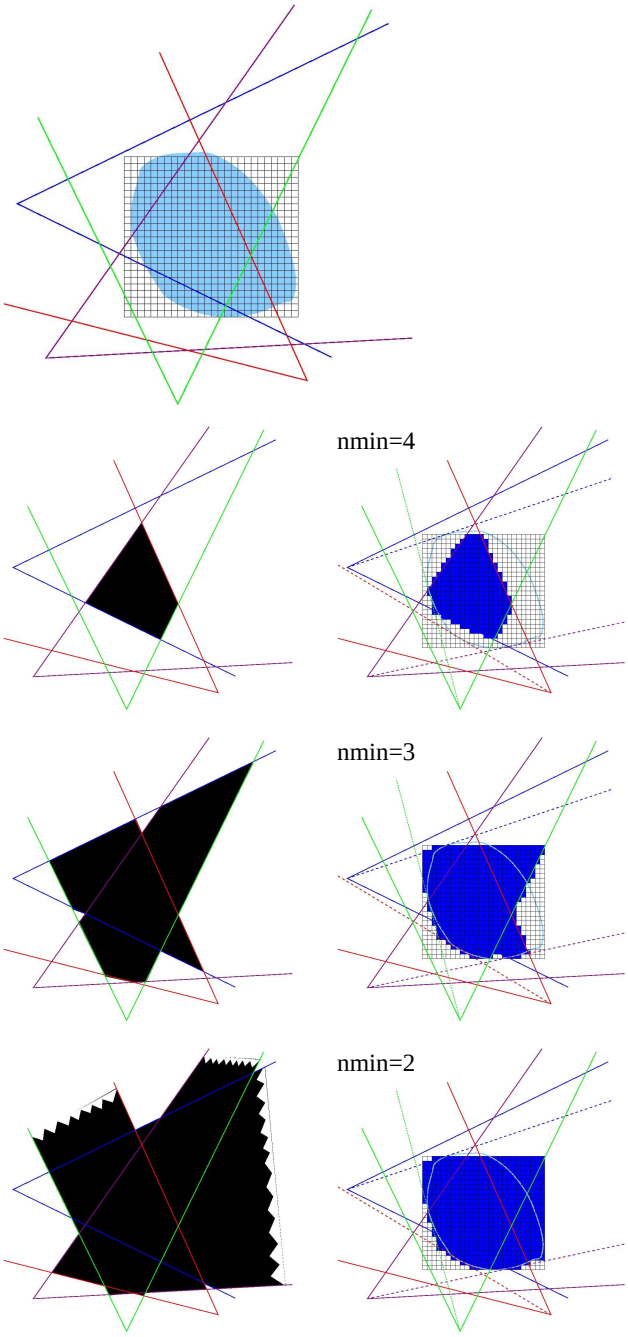

Figure 5 – Variation de l'estimation de la forme d'un objet en fonction de la valeur de nmin. L'image du haut représente la configuration caméras/objet/espace voxélique, à gauche l'espace d'acquisition disponible et à droite la forme obtenue.

Si $n_{min} = n - 1$, alors $ESFS_n$ complète l'estimation SFS_n, avec les informations visibles depuis $n-1$ caméras. Ceci est la meilleure estimation possible (en terme de précision) avec strictement moins de n caméras. De plus, la zone d'acquisition utilisée pour construire $ESFS_n$ est légèrement plus grande que celle utilisée pour SFS_n. Si n_{min} est proche de 1, alors la zone d'acquisition de $ESFS_n$ est nettement plus grande que celle de SFS_n. Mais la forme estimée sera moins précise pour les parties des O visibles depuis seulement n_{min} caméras. Ainsi, la valeur de n_{min} doit être choisie en fonction de l'application visée (ie, l'utilisation faite de l'estimation de la forme de O) comme montré dans la figure 5.

4 Résultats

Notre algorithme a été testé sur différents jeux de données réelles, acquis depuis 4 caméras avec une résolution de 320x240 pixels. La meilleure précision de reconstruction, lorsque certaines parties de l'objets n'étaient pas visibles depuis une caméra, a été obtenue en utilisant $n_{min} = 1$.

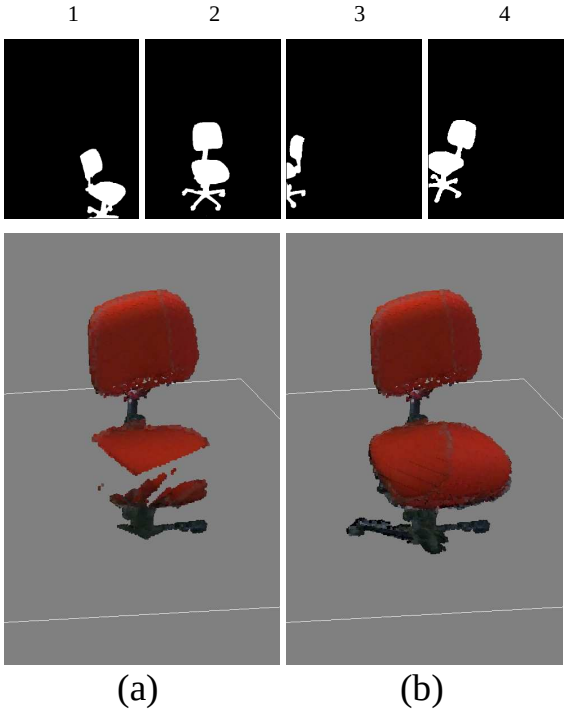

Figure 6 – Estimation voxélique de la forme d'un objet complexe : (a) algorithme de base de SFS ; (b) notre algorithme.

La figure 6 montre les résultats obtenus sur un objet complexe. Il y a visibilité partielle dans les silhouettes 1, 3 et 4. La méthode de base de SFS fourni une reconstruction partielle de l'objet (voir Figure 6.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 7 montre les résultats obtenus lors de l'acqui-

sition d'une personne en mouvement. Dans les deux cas (figures 6 et 7), notre méthode ajoute de nouvelles informations valides (du point de vue des images de silhouette) à l'estimation de la forme de l'objet.

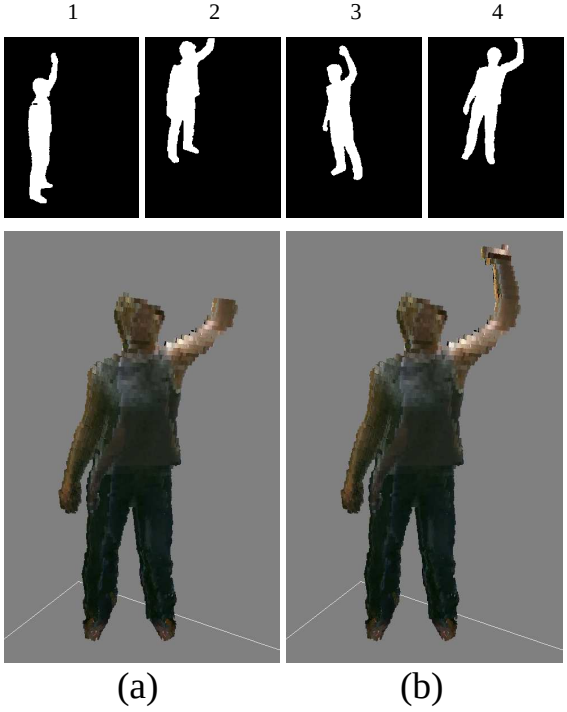

Figure 7 – Estimation voxélique de la forme d'une personne en mouvement : (a) algorithme de base de SFS ; (b) notre algorithme.

Les performances de notre algorithme sont très proches de celles de l'algorithme de base, seules deux étapes ont été ajoutées :

1. Le calcul de la carte de projection $CP(O)$, qui est réalisé une fois en pré-traitement de la phase d'acquisition. En effet, $CP(O)$ dépend uniquement des paramètres intrinsèques et extrinsèques des caméras qui sont constants durant l'acquisition ;

2. Le calcul de la connexité 3d : à chaque pas de temps, nous parcourons les ensembles R_n et R_k. Le temps de parcours étant négligeable par rapport au temps de calcul 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 et $n_{min} = 1$, alors que notre implémentation de la méthode de base de SFS atteint les 65 estimations de forme par seconde. Ces résultats nous montrent que notre algorithme est utilisable dans le cadre d'applications visant le temps réel.

5 Conclusions et perspectives

Dans cet article, nous avons proposé une extension de l'algorithme de "Shape-from-silhouette". Cette méthode permet d'avoir une zone d'acquisition étendue par rapport à celle disponible avec les algorithme habituels de SFS. Hormis le problème du calibrage des caméras, notre seule hypothèse porte sur le fait que l'objet doit être majoritairement visible dans toutes les caméras.
L'estimation de forme obtenue à partir de notre méthode contient celle qui peut être obtenue par SFS. Notre extension permet, de plus, d'estimer la forme des parties de l'objet qui ne sont pas visibles depuis une ou plusieurs caméras, du moment qu'elles le soient depuis les autres. De plus cette méthode est applicable même avec une contrainte temps réel.
Nous travaillons actuellement sur la formalisation de l'erreur de reconstruction en fonction du nombre de caméras utilisées. Cette méthode est utilisée dans le cadre d'une application de suivi de mouvement en temps réel et devrait être implémentée sur GPU afin de pouvoir travailler avec des caméras haute fréquence.

Références

[1] Kong Man Cheung, Takeo Kanade, J.-Y. Bouguet, et M. Holler. A real time system for robust 3d voxel reconstruction of human motions. Dans *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, volume 2, pages 714 – 720, Juin 2000.

[2] Wojciech Matusik, Chris Buehler, et Leonard McMillan. Polyhedral visual hulls for Real-Time rendering. Dans *12th Eurographics Workshop on Rendering Techniques*, pages 115–126, 2001.

[3] Jean-Marc Hasenfratz, Marc Lapierre, et François Sillion. A real-time system for full body interaction with virtual worlds. *Eurographics Symposium on Virtual Environments*, pages 147–156, 2004.

[4] Fabrice Caillette, Aphrodite Galata, et Toby Howard. Real-Time 3-D Human Body Tracking using Variable Length Markov Models. Dans *Proceedings of British Machine Vision Conference (BMVC)*, volume 1, pages 469–478, 2005.

[5] Bastian Goldluecke et Marcus Magnor. Real-time free-viewpoint video rendering from volumetric geometry. *Visual Communications and Image Processing 2003*, 5150(1) :1152–1158, 2003.

[6] Jean-Marc Hasenfratz, Marc Lapierre, Jean-Dominique Gascuel, et Edmond Boyer. Real-time capture, reconstruction and insertion into virtual world of human actors. Dans *Vision, Video and Graphics*, pages 49–56. Eurographics, Elsevier, 2003.

[7] M. Li, M. Magnor, et H. Seidel. Hardware accelerated visual hull reconstruction and rendering. Dans *Graphics Interface*, 2003.

[8] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(2) :150–162, 1994.

[9] Jean-Sébastien Franco et Edmond Boyer. Exact polyhedral visual hulls. Dans *Fourteenth British Machine Vision Conference (BMVC)*, pages 329–338, Septembre 2003. Norwich, UK.

[10] Ivana Mikic, Mohan Trivedi, Edward Hunter, et Pamela Cosman. Human body model acquisition and tracking using voxel data. *Int. J. Comput. Vision*, 53(3) :199–223, 2003.