Reconstruction géométrique par estimation de posture

Geometrical reconstruction using pose estimation

Mathieu Barnachon[1,2,3] Brice Michoud[2,3] Erwan Guillou[2,3] Saïda Bouakaz[2,3]

[1] PlayAll, www.playall.fr
[2] Université de Lyon, CNRS
[3] Université Lyon1, LIRIS, UMR5205, F-69622, France

Résumé

La reconstruction de la géométrie à partir de caméras est un domaine qui pourrait concurrencer les reconstructions classiques. Parmi les méthodes de reconstruction 3D à partir de plusieurs vues, la méthode Shape From Silhouette (SFS)[4] permet d'estimer en temps réel un volume englobant d'un objet (un humain dans notre cas) à partir de ses silhouettes observées dans chaque vue. En dehors de sa simplicité d'implémentation et de sa robustesse, SFS produit des formes qui contiennent des entités fantômes : ensemble de points 3D qui n'appartiennent pas aux objets d'intérêts. Ces entités peuvent fortement perturber la précision et le réalisme de la forme 3D reconstruite.
Dans cet article nous proposons d'utiliser l'information additionnelle de posture de la personne obtenue par un procédé de capture de mouvement, afin de supprimer ces entités fantômes et d'affiner la géométrie. Par construction l'approche SFS fournit une forme 3D englobante de la personne filmée. Ainsi si l'on transforme la géométrie calculée d'une posture de référence vers la posture courante, l'intersection des reconstructions fournit une reconstruction plus précise que l'approche SFS classique. Dans ces travaux, nous présentons une approche temps réel incrémentale, qui affine automatiquement la reconstruction 3D de la personne filmée au cours du temps.

Mots Clef

Reconstruction basée image, capture de mouvement, *Shape From Silhouette*, reconstruction d'avatar.

Abstract

Geometrical reconstruction from multiple cameras can now concurrence classical reconstruction. Shape From Silhouette (SFS)[4] is a method that estimates the Visual Hull, in real time from observed silhouette in each view. Despite the fact that SFS is really simple and robust, it produces ghost parts : a set of 3D points that does not belong to real objects. These parts can significantly disturb the precision and the realism of the 3D shape.
In this paper, we suggest to use the motion estimation of people, getting from motion capture, in order to remove these ghost parts and refine geometry. By construction, SFS gives a 3D hull of the filmed person. If we transform the computed geometry from original pose to the current pose, the intersection of reconstruction and observed pose produces a more precise reconstruction than the classical SFS. In this work, we introduce a real time and incremental approach, which automatically refines the 3D reconstruction of the acquired object.

Keywords

Image-based reconstruction, motion capture, Shape From Silhouette, avatar reconstruction.

1 Introduction

Dans le domaine de la vision par ordinateur, la production d'une géométrie précise à partir d'images est un problème récurrent. Nous proposons dans cet article de résoudre l'acquisition de la géométrie d'un personnage, par l'intermédiaire de l'extension de l'approche *Shape From Silhouette*, construite sur l'information de posture de ce personnage. Cette extension fournit, en temps réel, la forme volumique de ce personnage plus précisement que les approches usuelles d'estimation de l'enveloppe visuelle (*Visual Hull*) et ce, à partir d'un faible nombre de vues.
Partant du volume fourni par *Shape From Silhouette* et d'une estimation de la posture du personnage, nous proposons une reconstruction de meilleure qualité, sans entité fantôme. De plus, nous rendons la reconstruction plus résistante aux trous dans les silhouettes en tenant compte de la fréquence d'apparition des voxels. Le principe de la reconstruction est illustré par le schéma 1.
Nous partons d'une version standard de *Shape From Silhouette* proposé par [4]. Il s'agit d'une version volumique du *Shape From Silhouette* qui travaille sur une grille de voxels. Par une technique de *skinning* nous affectons les voxels de la première posture aux positions des os à la pose courante par [10]. Ceci nous permet de déformer la recons-

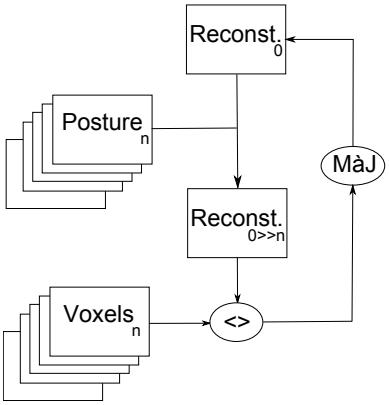


FIGURE 1 – Schéma de principe de la reconstruction.

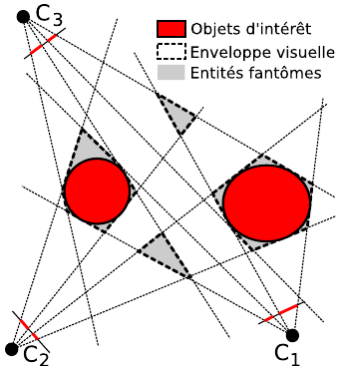


FIGURE 2 – Résultat de l'enveloppe visuelle lors de l'acquisition d'objets réel (en rouge) à l'aide de 3 caméras. Les entités fantômes apparaissent en gris.

truction de référence vers la reconstruction courante.
À l'aide de l'intersection entre la posture courante et la posture de référence, rapportée dans cette même posture, nous affinons la reconstuction. En effet, quelle que soit la posture considérée, SFS produit une forme englobante. Si l'on considère deux volumes englobants différents, alors leur intersection est aussi un volume englobant de la forme. Néanmoins, compte tenu de l'implémentation du *Shape From Silhouette* sur une grille de voxels, il est possible que ces volumes ne soient pas toujours des englobants exacts. C'est pour cela que nous allons utiliser la fréquence d'appartenance d'un point au volume englobant. Nous pourrons alors affiner le volume en fonction d'un seuil. Considérant qu'un voxel est d'autant plus sûr qu'il est observé un nombre de fois important, nous utilisons l'aspect incrémental pour produire un ensemble de confiance.
Une optimisation est proposée pour l'intersection en partant du principe que la grille de voxels peut être d'une résolution supérieure à la grille de voxels initiale. Nous montrerons qu'en subdivisant la grille par deux sur chacun des axes, soit une finesse huit fois plus importante, l'approche conserve néanmoins le temps réel.
La présentation de notre méthode se déroule selon le plan suivant : en section 2 nous présentons les méthodes de l'état de l'art ; section 3, nous développons la méthode de reconstruction géométrique ; la section 4 montre les résultats obtenus ; enfin, la section 5 conclue notre travail et évoque les perspectives associées.

2 État de l'art

Les méthodes d'estimation de l'enveloppe visuelle – concept introduit par Laurentini[6] – et notamment l'algorithme de *Shape From Silhouette*[4] tel qu'actuellement utilisé [11], permettent d'obtenir une représentation voxélique d'une personne filmée par n caméras en temps réel. Laurentini introduit le concept de volume maximal se projetant exactement dans les silhouettes issues des caméras. Cette méthode découpe les objets par la vision que l'on en a, à partir du cône de vision de chaque caméra, pour ensuite les projeter sur une grille régulière tridimensionnelle. Néanmoins, ces méthodes sont connues pour produire une grande quantité d'entités fantômes – entités qui ne représentent pas d'objet réel – bien qu'elles soient issues de la projection de silhouettes correctes (voir figure 2). Franco et Boyer[2] ont levé ces contraintes en introduisant une grille d'occupation en tant que représentation probabiliste spatiale et temporelle. Cette solution est malheureusement difficilement utilisable en temps réel, contrainte de ce travail.

Dans notre contexte, l'acquisition du mouvement est aussi importante que l'estimation de la géométrie. En effet, sans une bonne restitution des mouvements, il n'est pas possible d'effectuer une déformation de la posture initiale correcte, ce qui conduit inévitablement à des erreurs dans l'intersection des grilles de voxels. Bien qu'il existe des méthodes monoculaires [1] ou [18], elles souffrent d'erreurs et de manques de précision. En mode multi-images, des problèmes continuent de se poser quant à l'occultation de certaines parties du corps par une autre (auto-occultations). En effet, lorsque la topologie du sujet n'est plus clairement identifiable, des méthodes comme [12] tombent en défaut.

Partie charnière de notre méthode, le *skinning* est la solution adoptée quant à la déformation de la géométrie associée à la posture de référence vers la posture courante. Il se doit d'être le plus précis possible, tout en conservant son aspect temps réel. Le *skinning* est une technique très connue dans l'animation de personnages consistant à déformer – un maillage généralement – suivant un squelette d'animation. Néanmoins, il n'est pas possible ici de connaître à priori l'influence de chaque os sur la géométrie. Il est pourtant nécessaire de connaître l'influence de chaque membre – os – sur les parties du corps de façon automatique. Dans notre cas, il s'agit d'apparier les voxels à un os de manière unique. Ensuite, nous appliquons la déformation nécessaire à la reconstruction de la posture initiale. Il existe un très grand nombre de méthodes pour réaliser ces déformations, comme les méthodes utilisant des déformations linéaires, *Skeletal-Subspace Deformation (SSD)*[7], *Animation Space*[9], *Multi-Weight Enveloping*[16]. Les méthodes linéaires présentent comme avan-

tages leur simplicité et leur rapidité, néanmoins, elles ne permettent pas toujours d'éviter des désagréments comme les contractions de maillage aux articulations.
Une autre famille de *skinning* est celle des déformations non linéaires, par courbes [17], [15], ou *dual quaternions*[3]. Ces méthodes sont beaucoup plus longues et coûteuses à mettre en œuvre.

3 Reconstruction géométrique

Nous présentons en premier lieu la méthode de *skinning* permettant de transformer la géométrie initiale vers la forme courante. La première étape consiste à déterminer pour chaque point les os qui l'influence. Il existe peu de méthodes de calcul automatique de cette influence. La plupart des approches utilisent une « peinture » de l'influence par un artiste. Cette version n'est pas souhaitable ici à cause du temps réel. De plus, il faudrait réaliser ce processus pour chaque trame réalisée. Diverses techniques sont envisageables, comme la création d'aires d'influence par des ellipsoïdes, affectation à l'os le plus proche, ou un découpage par plans selon les articulations.
Une technique simple de *skinning* linéaire a été développée. S'appuyant sur la topologie fixe du squelette et sur une affectation contrainte par les articulations, le *skinning* linéaire par plans produit des résultats corrects en un temps court. Le principe de cette affectation est de séparer la géométrie suivant les plans bissecteurs des os. Considérant la distance d'un point à un segment (équation 1), nous affectons chaque voxel au segment – os rigide – dont il est le plus proche. Pour l'efficacité des calculs, nous faisons le choix de travailler uniquement sur le centre des voxels.

$$Dist_{C\rightarrow[AB]} \quad = \quad Min\left(Dist_{X\rightarrow C}\right), \forall X \in [AB] \ , \qquad (1)$$

Avec $Dist_{X\rightarrow C}$ la distance euclidienne, $[AB]$ l'os, et C le centre du voxel considéré.
En conservant la distance minimale, calculée par l'équation 1, entre un voxel et un os, nous effectuons une découpe des influences selon les plans bissecteurs aux os. La figure 3 représente la découpe effectuée par notre approche.
Il est facile, connaissant l'os auquel un point se rapporte, de le déformer de façon identique à celui-ci. Pour cela, le mouvement entre les extrémités des articulations dans la trame courante et celle dans la trame de référence est établi. Nous obtenons alors la matrice de passage de la posture initiale à la posture courante. En appliquant cette matrice aux voxels de la posture initiale, nous obtenons leurs positions dans la posture courante.
Ayant une déformation de la posture de référence dans la posture observée, il est désormais possible d'affiner la géométrie de la personne. L'enveloppe visuelle produit un nombre plus ou moins important de voxels qui ne sont pas des objets réels. Lors de la première reconstruction, nous savons, par définition, que tous les points de l'objet réel sont présents. Il n'est dès lors plus utile d'en ajouter. Pour obtenir une forme plus précise, il est nécessaire de retirer

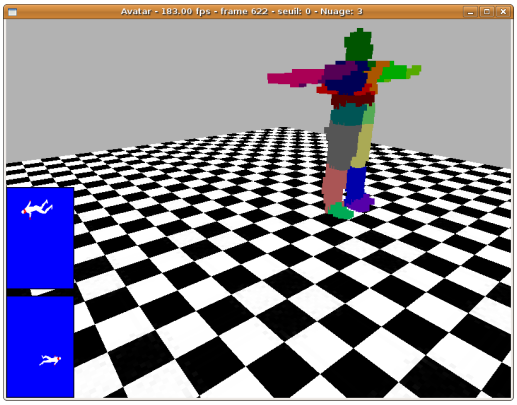


FIGURE 3 – *skinning* automatique de la posture de référence avec une couleur par os.

des points de cette enveloppe. Ce retranchement est réalisé par l'intersection entre la géométrie de référence transformé dans la posture courante et la géométrie courante. Les points hors de l'intersection seront éliminés, car ils ne peuvent être des points du personnage.
L'approche n'étant pas basée sur un modèle générique que l'on déforme, comme dans [13], mais sur une reconstruction incrémentale du sujet en fonction du temps, il est donc nécessaire de définir une posture initiale. Ceci nous permet d'avoir une méthode générique, mais nous contraint à privilégier une reconstruction par rapport aux autres. La première reconstruction sera la référence. Nous imposons également, pour des raisons d'efficacité du *skinning* que le personnage soit en *T-pose* (voir figure 3). Cette posture est courante en animation, et les performances du *skinning* sont ainsi meilleures. Le personnage n'est contraint de conserver cette posture que le temps d'une estimation.
La géométrie de référence contient, par essence, l'ensemble de l'information pertinente. Pour l'extraire, nous profitons du mouvement du sujet.
L'estimation de la posture permet de reporter la posture initiale vers la posture actuelle du personnage. Les deux géométries contiennent l'information pertinente plus des artéfacts dû à la reconstruction. Le fait de rapporter ses deux géométries dans le même référentiel permet d'extraire cette information : l'intersection produit alors un volume supérieur ou égal au volume des objets réels (voir équation 2).

$$\left.\begin{array}{l} R \subseteq A \\ R \subseteq B \end{array}\right\} \quad \Rightarrow \quad R \subseteq (A \cap B) \qquad (2)$$

Où :
– R est le volume des objets réels ;
– A est le volume issu de l'enveloppe visuelle de la trame de référence ;
– B est le volume issu de l'enveloppe visuelle de la trame courante.

Chaque nouvelle trame permet d'affiner la géométrie de référence. Si le *skinning* et l'estimation de la posture étaient

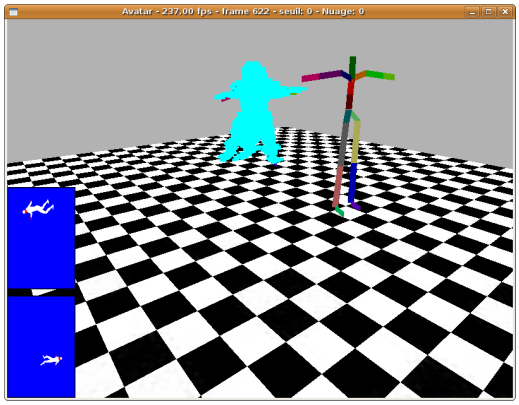


FIGURE 4 – Données en entrée : voxels (bleus), estimation de la posture courante (sous les voxels), et posture de référence.

parfaits, nous n'aurions plus qu'à retirer de la trame initiale les points qui n'ont pas de correspondant dans la trame courante. Cependant, les imprécisions inhérentes à tout *skinning* conduisent à calculer des fréquences pour chaque point de référence. Nous proposons une mesure de confiance pour chacun des voxels. Nous filtrons la reconstruction en fonction de la fréquence de validation de chaque voxel. La figure 7 souligne différentes fréquences de validation en fonction du temps.

Les propriétés de l'enveloppe visuelle ne dépendent pas de l'échantillonnage sur lequel on travail. Néanmoins, cela conditionne la qualité du résultat final. Passer d'une grille de voxels de 64 de côté, à une grille de 128 de côté, conduit à augmenter les calculs par $128^3 - 64^3 = 1835008$, soit près de 2 millions de plus par reconstruction. Cela se traduit inévitablement par un surcoût trop fort actuellement pour le temps réel. Toutefois, comme la posture initiale ne comprend qu'une portion de la grille de voxels, il est possible de la sur-échantillonner, au moins par deux sur chacun des axes. On obtient alors une grille plus fine.

Lors de la transformation de la trame de référence vers la trame courante, le sur-échantillonnage des voxels permet une précision plus importante quant à la correspondance des points entres eux.

4 Résultats

Les calculs sont réalisés sur un Core 2 Duo à 2,40 GHz, équipé d'une carte graphique Nvidia GeForce 8700M GT et de 4Go de RAM. La grille de voxels est une grille cubique de 64^3 voxels, correspondant à la taille d'une boîte réelle de 2,5m de côté. On obtient ainsi une définition de 4cm environ pour un voxel. L'évaluation de la méthode est réalisé à partir de données synthètique, où l'erreur due à la soustraction de fond, ainsi que de l'estimation du mouvement est minimisée.

L'ensemble des résultats est fourni pour une configuration de 2 caméras. Dans cette configuration, l'apport de notre méthode est le plus prononcé.

L'analyse du nombre de voxels au cours du temps (que nous appellerons volume par abus de langage) nous montre de fortes variations. En effet, ce volume est dépendant de l'action du personnage devant les caméras et de la position des dites caméras. Nous parvenons à résoudre ce problème en filtrant les voxels sur leur correspondance avec les voxels déformés de la posture d'origine. Ce volume est alors quasi constant, comme on peut le voir sur la figure 4(a). La qualité des données en entrées peut être jugée à partir de figure 4. On remarquera notamment que SFS produit trois jambes pour le personnage, néanmoins, nous n'avons pas de problème au niveau de l'estimation du mouvement. La posture de référence est une posture en T, ceci permet d'avoir un bon *skinning* initial, comme le montre la figure 5(a).

La figure 5(b) illustre la déformation de la posture de référence à l'aide des informations de *skinning* et de l'estimation de la posture. Le résultat est une géométrie de meilleur qualité que les estimations offertes par l'approche SFS.

La figure 5(c) montre la superposition de la déformation à l'acquisition. De nombreux voxels sont en trop, mais la posture déformée est cohérente avec la valeur observée, ce qui nous permet de faire l'intersection des deux.

Néanmoins, rien ne garantit que ce volume soit correct. Nous avons pour cela comparé le volume filtré avec un volume de référence. Le volume de référence est obtenu à partir de la reconstruction avec 8 caméras. Les figures 4(b) et 4(c) montrent qu'il est possible d'encadrer le volume de référence par deux seuils, ici respectivement 60% et 30%. Ceci permet de mettre en valeur un seuil pouvant fournir le volume réel.

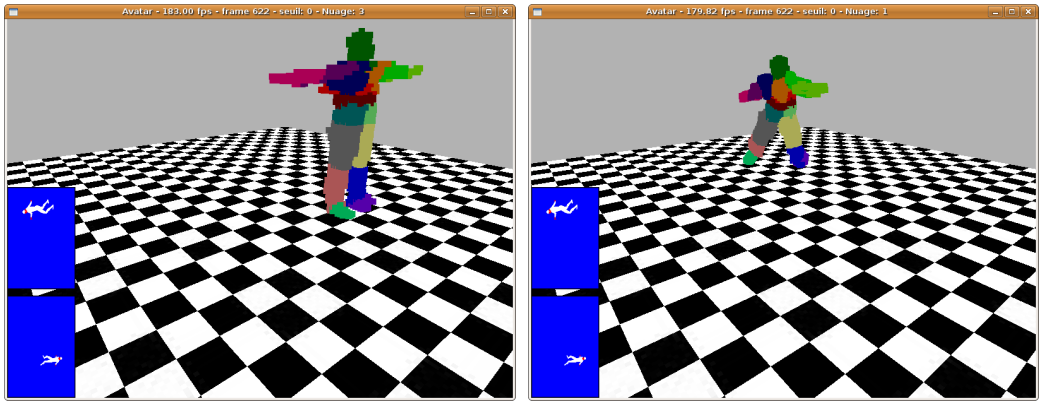
Il demeure que le volume filtré ne représente pas forcément le personnage tel qu'il est. Ceci s'explique par le fait que nous approchons l'enveloppe visuelle à l'aide de SFS. SFS peut donner des artéfacts dus à des ambiguitées (connus sous le nom de géométrie fantôme). Ceci est inhérent à toute méthode estimant l'enveloppe visuelle, telle SFS.

La reconstruction montre un aspect plus proche de la réalité avec un seuil bas. Il serait plus intéressant d'avoir un seuil haut, cela prouverait que notre méthode se comporte généralement bien. Nous avons choisi des cas difficiles et une limitation volontaire à deux caméras.

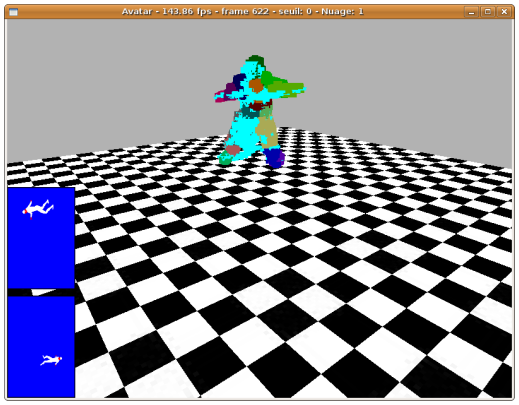
La reconstruction faite est, par essence, plus précise que SFS et ce, quelque soit le nombre de trames considérées. Par cette approche, nous sommes en mesure d'affiner la forme à chaque nouvelle trame, en fonction des précedentes, donc de façon itérative. De plus cette méthode conserve le temps réel, soit plus de 32 estimations calculées par seconde.

5 Conclusion

Dans ce travail, nous avons décrit et expérimenté une amélioration de SFS par l'utilisation de la connaissance de posture du personnage. La méthode est totalement automatique. Le système est basé sur des transformations géométriques élémentaires, ce qui le rend particulièrement effi-



(a) posture initiale *skinnée* (b) posture de référence rapportée dans la posture courante



(c) Voxels courants sur voxels *skinnés*

FIGURE 5 – Ensembles calculés par notre méthode.

cace, voir facilement implémentable sur GPU.

Nous avons démontré qu'il permettait de supprimer les parties fantôme inhérentes à SFS, et ce même avec un nombre très restreint de caméras. On remarquera que le système peut facilement s'étendre à plusieurs personnages – humains ou non – présents en même temps dans la scène, dès lors qu'on est capable de fournir une estimation de leur posture.

En conclusion, il s'agit d'une amélioration de SFS exploitable en temps-réel, qui permet le rendu d'avatar de qualité. La rapidité d'acquisition d'un personnage, ainsi que son exploitation et son post-traitement, sont très rapides. Elles sont utilisables dans de nombreuses conditions, la méthode étant fonctionnelle avec des estimations de posture issues d'autres méthodes, plus précises, telles que des données Vicon.

L'association des voxels aux os, ainsi que leurs déformations restent le point faible de notre méthode. C'est pour cela que l'utilisation d'un *skinning* plus performant, tant en terme de peinture de la zone d'influence, que de déformation, peut-être non linéaire, est la première piste à investiguer pour augmenter l'efficacité de la méthode.

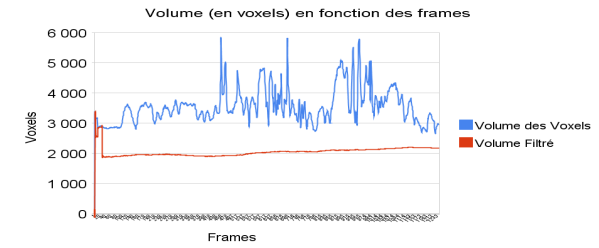
Pour parfaire la reconstruction géométrique, il serait intéressant de mettre en œuvre la méthode de classification des voxels décrites dans [8]. Ceci afin d'obtenir des informations pertinentes sur la photométrie du personnage. Nous pourrions alors obtenir les sources lumineuses, ainsi que les propriétés de réflexion des matériaux. Le modèle serait d'autant plus réutilisable dans un environnement différent, avec d'autres sources et/ou types d'éclairage.

L'investigation d'une méthode combinant géométrie et photométrie est envisageable, telle que *Voxel Coloring*[14] ou *Space Carving*[5]. La moyenne et l'écart type des couleurs par voxels nous conduit à ajouter une information sur la cohérence de chaque voxel. En effet, un voxel ayant un trop grand écart type au niveau de ses couleurs peut apparaître comme un point non photo-consistant. Ajouter cette information à la reconstruction géométrique pourrait augmenter le degré de fiabilité des informations, et ainsi produire une reconstruction plus précise.

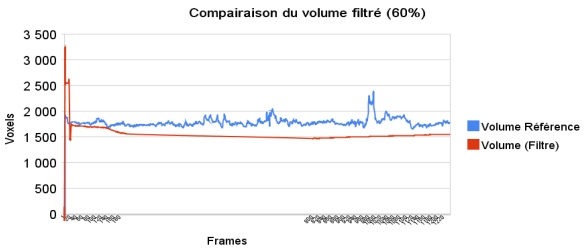
Références

[1] Ankur Agarwal and Bill Triggs. Monocular human motion capture with a mixture of regressors. In *CVPR '05 : Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, 2005.

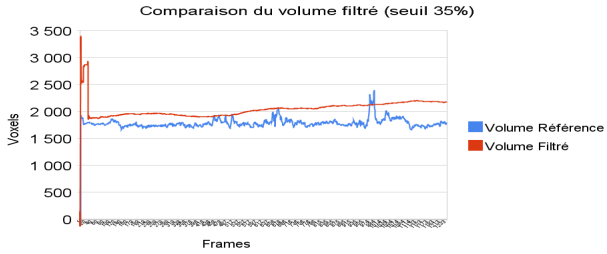
[2] Li Guan, Jean-Sebastien Franco, and Marc Pollefeys. Multi-object shape estimation and tracking from sil-



(a) Variations des volumes avec 2 caméras



(b) Comparaison des volumes (seuil 60%)



(c) Comparaison des volumes (seuil 30%)

FIGURE 6 – Variations de volume.

houette cues. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[3] Ladislav Kavan, Steven Collins, Jin Zara, and Carol O'Sullivan. Skinning with dual quaternions. In *I3D '07 : Proceedings of the 2007 symposium on Interactive 3D graphics and games*, 2007.

[4] Simon Baker Kong Man Cheung and Takeo Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[5] Kiriakos N. Kutulakos and Steven M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 2000.

[6] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1994.

[7] J. P. Lewis, Matt Cordner, and Nickson Fong. Pose space deformation : a unified approach to shape interpolation and skeleton-driven deformation. In *SIGGRAPH '00 : Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000.

[8] Bruno Mercier, Daniel Meneveaux, and Alain Fournier. A framework for automatically recovering object shape, reflectance and light sources from calibrated images. *International Journal of Computer Vision (IJCV)*, 2007.

[9] Bruce Merry, Patrick Marais, and James Gain. Animation space : A truly linear framework for character animation. *ACM Trans. Graph.*, 2006.

[10] Brice Michoud, Erwan Guillou, Héctor M. Briceño, and Saïda Bouakaz. Real-time marker-free motion capture from multiple cameras. In *ICCV*, 2007.

[11] Brice Michoud, Erwan Guillou, Hector Briceno Pulido, and Saida Bouakaz. 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*, 2008.

[12] Ivana Mikic, Mohan Trivedi, Edward Hunter, and Pamela Cosman. Human body model acquisition and tracking using voxel data. *International Journal on Computer Vision*, 2003.

[13] Laurent Moccozet, Fabien Dellas, and Nadia Magnenat-Thalmann. Animatable human body model reconstruction from 3d scan data using templates, 2004.

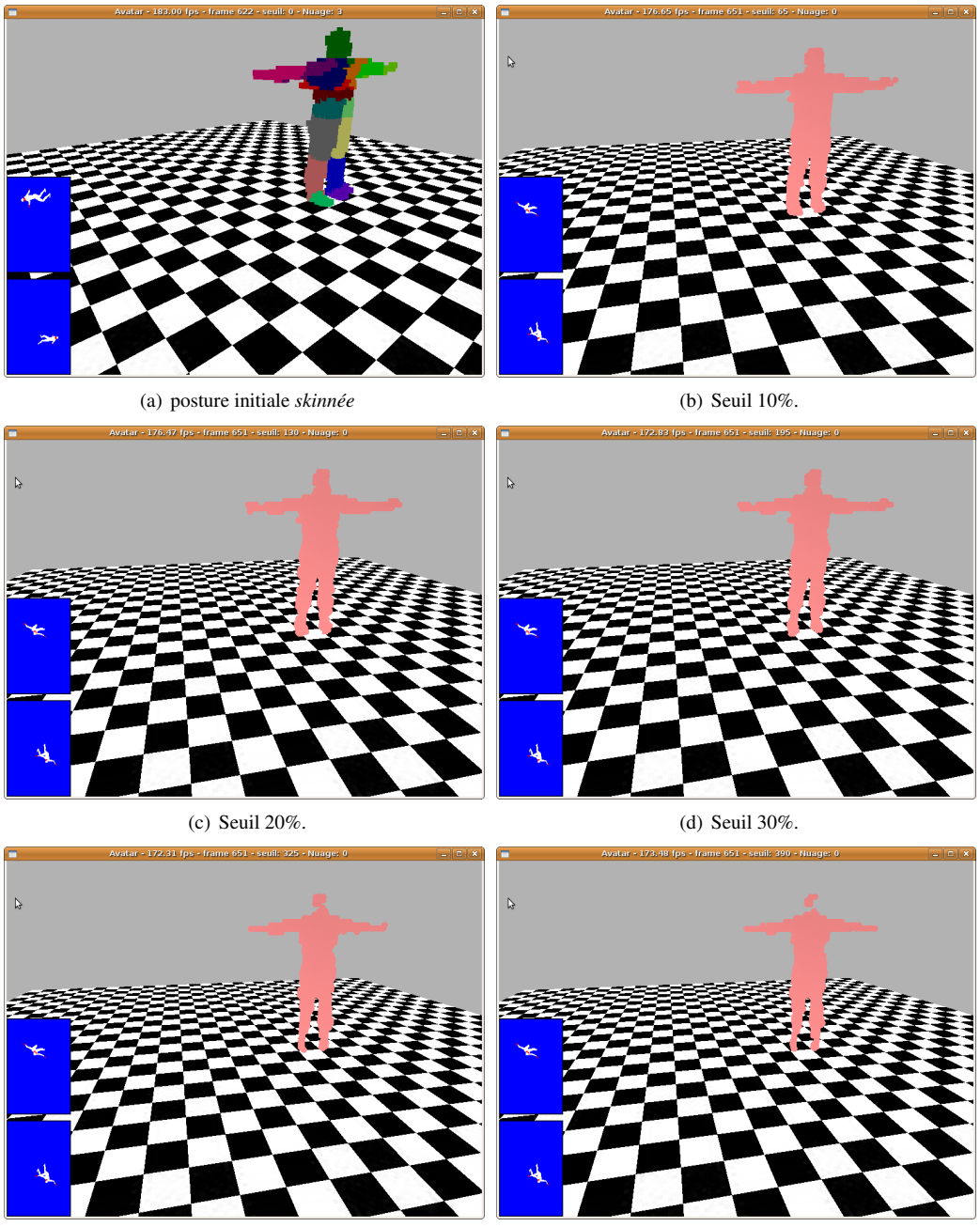
[14] Steven M. Seitz and Charles R. Dyer. Photorealistic scene reconstruction by voxel coloring. In *Proc. Computer Vision and Pattern Recognition Conf.*, 1997.

[15] Forstmann Sven and Ohya Jun. Skeletal animation by spline aligned deformation on the gpu. *IEICE technical report. Image engineering*, 2007.

[16] Xiaohuan Corina Wang and Cary Phillips. Multi-weight enveloping : least-squares approximation techniques for skin animation. In *SCA '02 : Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, 2002.

[17] Xiaosong Yang, Arun Somasekharan, and Jian J. Zhang. Curve skeleton skinning for human and creature characters. *Computer Animation and Virtual Worlds*, 2006.

[18] Chen Yisheng, Lee Jinho, Parent Rick, and Raghu Machiraju. Markerless monocular motion capture using image features and physical constraints. In *CGI '05 : Proceedings of the Computer Graphics International 2005*, 2005.



(a) posture initiale *skinnée*

(b) Seuil 10%.

(c) Seuil 20%.

(d) Seuil 30%.

(e) Seuil 50%.

(f) Seuil 60%.

FIGURE 7 – Reconstructions obtenues à partir de deux caméras.(a) la reconstruction initiale, (b-f) reconstruction offerte par notre système pour différentes valeurs de seuil. Il est important de noter que la valeur du seuil ne se fixe qu'on moment du rendu de la géométrie – *a posteriori* – et donc que toutes les reconstructions présentées sont disponible en une seule acquisition.