Acquisition du geste humain 3D par vision monoscopique

Patrick Horain Mayank Bomb

Groupe des Ecoles des Télécommunications
INT / EPH

9 rue Charles Fourier,
91011 Evry Cedex, France

Patrick.Horain@int-evry.fr, mayank_bomb@yahoo.com

Résumé

Nous présentons une méthode pour l'acquisition en 3D du geste humain au moyen d'une caméra unique, sans marqueur et sans connaissance a priori *sur les gestes observés. Celle-ci consiste à recaler un modèle 3D articulé du corps humain sur une séquence d'images en couleur tout en respectant des contraintes biomécaniques. La régularisation temporelle des gestes est étudiée pour permettre de lever les ambiguïtés de la projection dans les images. Le coût des calculs est réduit en recalant seulement les parties du corps en mouvement.*

Mots clefs

Acquisition du geste, mouvement articulé, suivi, segmentation, vision artificielle.

1 Introduction

Les gestes sont un complément important et naturel de la communication humaine. Les acquérir permet d'animer des acteurs virtuels, de les coder par exemple au moyen de la norme MPEG-4, de les utiliser dans une interface de commande ou, à plus long terme, de reconnaître la langue de signe.

Dans cet article, nous présentons une méthode pour acquérir les gestes en 3D par vision artificielle monoscopique, sans marqueur et sans connaissance *a priori* des gestes observés [1][2]. Celle-ci procède par recalage itératif d'un modèle 3D articulé du corps humain sur une séquence vidéo unique. Elle consiste à mettre en correspondance des caractéristiques extraites des images avec des caractéristiques du modèle par une méthode d'optimisation, tout en respectant les contraintes biomécaniques. Les projections du modèle dans les images pouvant être semblables pour des attitudes très différentes, nous discutons comment la régularisation des gestes permet de lever ces ambiguïtés.

2 Modélisation du corps humain

Notre modèle 3D articulé de la moitié supérieure du corps humain peut être animé suivant 23 degrés de liberté (Figure 1). Ces paramètres forment un vecteur **q** qui définit l'attitude du modèle.

Les variations de ces degrés de liberté sont limitées par des contraintes biomécaniques intégrées dans l'algorithme d'optimisation du recalage [3].

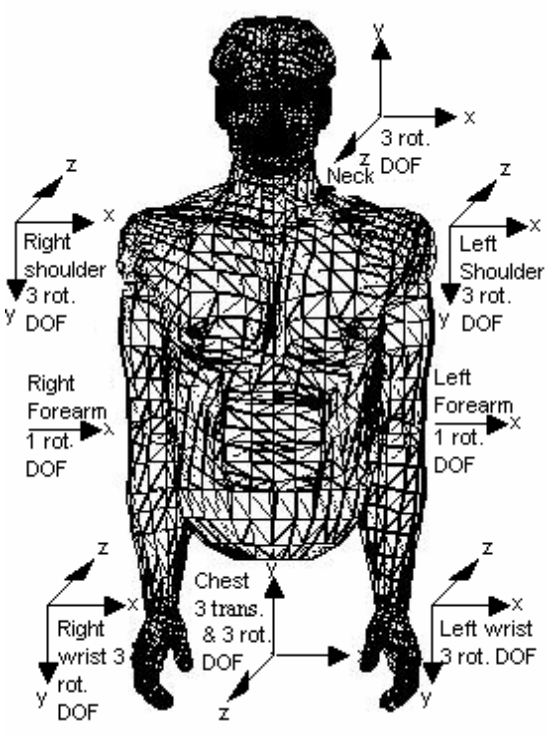


Figure 1 – *Modèle 3D du buste avec 23 degrés de liberté*

3 Extraction des caractéristiques dans les images

Ce modèle 3D est recalé sur les images en superposant sa projection avec des caractéristiques extraites des images. Celles-ci peuvent être les contours [4], le mouvement apparent [5], les textures [6], la couleur [3]…
Nous utilisons une segmentation des images par la couleur. La peau et les vêtements de couleur uniforme peuvent efficacement être détectées par leur chrominance dans l'espace de couleurs YC_bC_r [7] puisque celle-ci ne change pas avec l'orientation relative de la source lumineuse.
Pour chaque classe de couleur c, la moyenne $\boldsymbol{\mu}_c$ et la matrice de covariance $\boldsymbol{\Sigma}_c$ des chrominances C_bC_r sont apprises sur des exemples. Chaque pixel $\mathbf{x} = (C_b, C_r)^T$ est attribué à la classe dont il est le plus proche selon une métrique de Mahalanobis [8] :

$$\mathbf{d}_c = \sqrt{\left(\mathbf{x} - \boldsymbol{\mu}_c\right)^T \boldsymbol{\Sigma}_c^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_c\right)}.$$

Finalement, le bruit de la segmentation est réduit par un filtrage modal [9] : chaque pixel reçoit le numéro de la classe de couleur la plus fréquente dans son voisinage (Figure 2).

4 Recalage du modèle sur les images

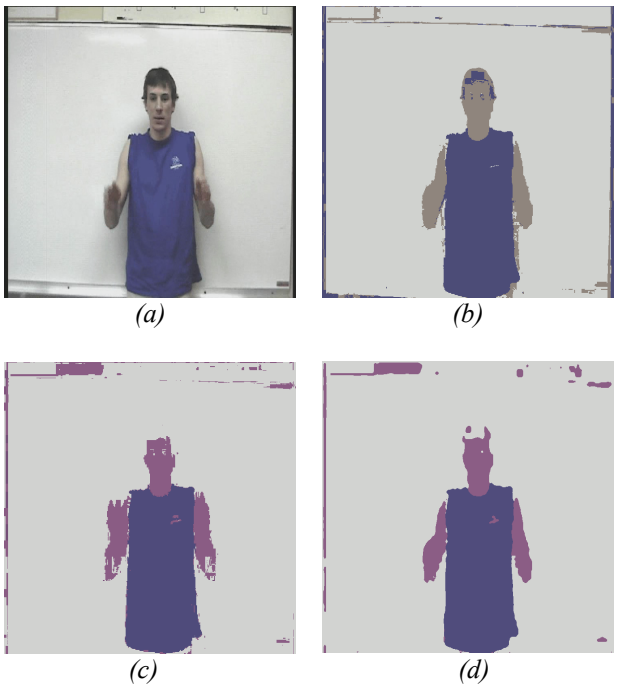
Chaque segment articulé du modèle est associé à une classe de couleur. Le recalage est effectué en rapprochant la projection de ces segments des régions de l'image segmentée. La meilleure correspondance est recherchée pour chaque image de la séquence suivant un algorithme itératif d'optimisation.
Pour la $k^{\text{ème}}$ itération sur l'image t de la séquence, le modèle est projeté sur l'image dans l'attitude définie par un vecteur de paramètres $\mathbf{q}_t^k$ suivant un algorithme de *z-buffer* modifié pour mémoriser les couleurs des segments du modèle. Ce recalage est alors évalué par son taux de non-recouvrement :

$$F(\mathbf{q}_t^k) = \prod_{c=1}^{m} \left(\frac{\left|A_c \cup B_c^k\right| - \left|B_c^k \cap A_c\right|}{\left|A_c \cup B_c^k\right|} \right)^{\frac{1}{m}}$$

où A_c est l'ensemble des pixels dans la $c^{\text{ème}}$ classe de couleur, B_c^k est la projection des segments du modèle associés à la $c^{\text{ème}}$ classe pour l'attitude du modèle définie par $\mathbf{q}_t^k$ et m est le nombre de classes de couleur.

5 Optimisation itérative

Le modèle est recalé sur les images par une procédure itérative de minimisation d'une fonction de coût $E(\mathbf{q}_t^k)$, qui peut-être choisie égale au taux de non-recouvrement :



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

Figure 2 - *(a) Exemple d'image extraite d'une séquence vidéo. (b) Image segmentée dans l'espace de couleur RVB : les ombres (adjacentes au torse) sont mal classées. (c) Image segmentée sur la chrominance. (d) Segmentation finale après filtrage modal.*

$$E(\mathbf{q}_t^k) = F(\mathbf{q}_t^k).$$

Le gradient de cette fonction n'est pas calculable facilement et ne peut être utilisé dans l'algorithme d'optimisation. Les méthodes d'optimisation ont été comparées pour l'enregistrement de gestes par Ouhaddi *et al.* qui montrent que la méthode de descente de simplexe est plus performante. Les contraintes biomécaniques peuvent être intégrées efficacement dans l'algorithme en contraignant les évolutions du simplexe d'états à l'intérieur d'un domaine convexe [3]. Nous avons utilisé cette méthode.

6 Régularisation des gestes

Il peut arriver que la projection du modèle sur l'image soit la même pour plusieurs attitudes du modèle, ce qui constitue une ambiguïté (Figure 3). De plus, le bruit des images peut causer un recalage saccadé du modèle sur la séquence d'images. L'acquisition des gestes doit donc être régularisée.
Le filtrage de Kalman est souvent employé pour le suivi [10][11], mais les ambiguïtés possibles dans la projection modèle se traduisent par de grandes covariances sur la mesure par recalage, d'où un gain faible et un suivi médiocre. Le problème vient de ce que le filtrage Kalman

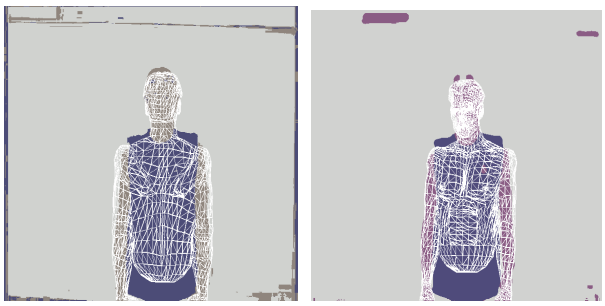


Figure 3 – *Exemple de recalage ambigu : que le modèle soit de face ou de dos, sa projection est presque la même.*

effectue la mesure et la régularisation dans des étapes successives et indépendantes.
Lowe [12] a proposé une méthode de régularisation qui fusionne ces étapes. Elle repose sur la connaissance des écarts types σ_i des variations des paramètres q^i qui jouent le rôle de contraintes biomécaniques dynamiques statistiques. Le terme de régularisation est alors défini par :

$$R(\mathbf{q}_t^k) = \sqrt{\left(\sum_{i=1}^{n} \left(\frac{\hat{q}_t^{i-} - q_t^{k,i}}{\sigma_i} \right)^2 \right)}$$

où $\mathbf{q}_t^k$ est le vecteur de paramètres d'attitude pour la $t^{\text{ème}}$ image évalué à l'itération k, $\hat{q}_t^{i-}$ est la prédiction de l'$i^{\text{ème}}$ paramètre de $\mathbf{q}_t$, $q_t^{k,i}$ est la valeur obtenue à la $k^{\text{ème}}$ itération pour l'ième paramètre.
Ce terme est introduit dans la fonction de coût :

$$E(\mathbf{q}_t^k) = F(\mathbf{q}_t^k) + R(\mathbf{q}_t^k).$$

En cas de l'ambiguïté, il pénalise le recalage sur une attitude du modèle qui serait éloignée de celle prédite.
Les écarts types des variations des paramètres sont fixés par observation d'un corpus de langue des signes [13].

7 Détection du mouvement

L'optimisation, qui doit être répétée pour chaque image de la séquence, est très coûteuse. Pour réduire ces calculs, il est utile de détecter les parties du modèle en mouvement et de limiter le processus d'optimisation à leurs paramètres.
Les images successives de classes de couleur sont comparées (Figure 4) pour déterminer quelles régions colorées se déplacent en avant-plan. L'optimisation est limitée aux paramètres des segments du modèle projeté qui leur sont associés.

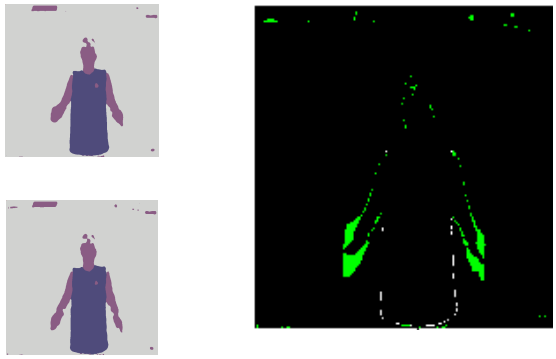


Figure 4 – *Détection du mouvement : les images successives segmentées (à gauche) sont comparées (à droite).*

8 Résultats et conclusion

Des résultats sur quelques images extraites d'une séquence sont présentés en Figure 5. D'autres résultats sont visibles à l'adresse [14]. Les gestes sont correctement acquis même en cas d'auto-occultation.
Cette méthode permet d'acquérir les gestes humains sans marqueur et avec une seule caméra. Elle ne requiert pas la connaissance *a priori* d'un vocabulaire de gestes attendus.
Les données obtenues peuvent être utilisées pour animer des acteurs virtuels [15], voire pour reconnaître un sous-ensemble de la langue des signes [14].

Remerciements

Ce travail a été possible en partie grâce au soutien de l'INRIA à travers l'Action de Recherche Coopérative *Vers un système d'interprétation de la langue des signes française*.

Références

[1] J. M. Rehg and T. Kanade. Visual tracking of high DOF articulated structures: an application to human tracking. In *Proceedings of 3rd ECCV*, vol. 2, pages 37-46, Stockholm, 1994.

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

[3] H. Ouhaddi et P. Horain. Modélisation 3D du geste par la vision. Dans *Actes des Journées CORESA'99*, pages 277-283, Sophia-Antipolis, juin 1999. http://www-eph.int-evry.fr/~horain.

[4] R. Plaenkers and P. Fua. Model-based silhouette extraction for accurate people tracking. In *Proc. ECCV 2002*, pages 325-, Copenhagen, Denmark, May 2002. http://vrlab.epfl.ch/Publications/pdf/Plaenkers_Fua_ECCV_02.pdf.

[5] R. Cutler and M. Turk. View-based interpretation of real-time optical flow for gesture recognition. In *Proc. IEEE FG'98*, Nara, Japan, 1998. http://www.cs.ucsb.edu/~mturk/research.htm.
[6] F. Lerasle, G. Rives et M. Dhome. Suivi des membres corporels par vision multi-oculaire. Dans *Actes RFIA'98*, vol. I, pages 193-199, Clermont-Ferrand, 1998.
[7] S. Kewei, F. Xitian, C. Anni, and S. Jingao. Automatic face segmentation in YCrCb images. In *Proc. 5th Asia-Pacific Conf. on Communications* and *4th Optoelectronics and Communications Conf.*, vol. 2, pages 916–919, 1999.
[8] M. Partridge and M. Jabri. Face recognition using a new distance metric. In *Proc. IEEE Signal Processing Society Workshop Neural Networks for Signal Processing X*, vol. 2, pages 584–593, 2000.
[9] E. R. Davis. *Machine Vision, Theory Algorithms Practicalities*. 2nd Edition, Academic Press, 1995.
[10] Q. Delamarre and O. Faugeras. 3D Articulated Models and Multi-View Tracking with Physical Forces. *Computer Vision and Image Understanding*, 81(3) : 328–357, 2001. http://www-sop.inria.fr/robotvis/personnel/qdelam/PUBLI.html.
[11] I. Kakadiaris and D. Metaxas, "Model-based estimation of 3D human motion", *IEEE tPAMI,* 22(12) 1453–1459, Dec. 2000.
[12] D. G. Lowe. Robust model-based motion tracking through the integration of search and estimation. *International Journal of Computer Vision*, 8(2) : 113-122, 1992.
[13] IVT - ABAQUE-micro - je.tu.il. *Les signes de Mano*. Cd-rom, http://www.ivtcscs.org/produits.
[14] P. Horain. *Vers l'acquisition du geste par vision artificielle pour l'interprétation de la langue des signes*. http://www-eph.int-evry.fr/~horain/ARC-LSF, octobre 2002.
[15] P. Horain. *Télémondes : Télévirtualité et mondes virtuels habités*. http://www-eph.int-evry.fr/~horain/Telemondes, mars 2002.

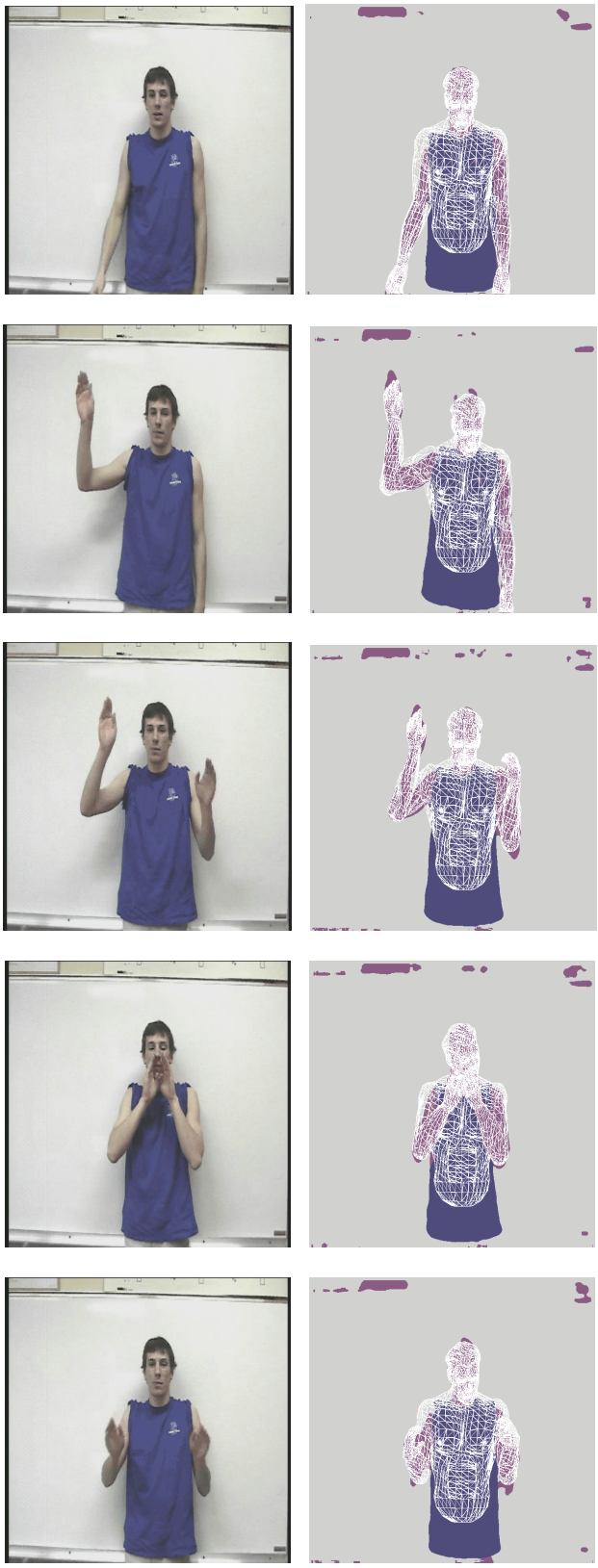


Figure 5 – *Exemples de recalages du modèle (à droite) sur une séquence d'images (à gauche).*