3D Model Based Gesture Acquisition Using a Single Camera

Patrick Horain and Mayank Bomb(*)
Groupe des Ecoles des Télécommunications
Institut National des Télécommunications, Département EPH
9 rue Charles Fourier, 91011 EVRY Cedex, France
http://www-eph.int-evry.fr
Patrick.Horain@int-evry.fr, mayank_bomb@yahoo.com

Abstract

We present a method for 3D human motion capture using a single camera, without markers and without a priori knowledge on gestures. It is based on registering a 3D articulated model on color images with respect to biomechanical constraints. Gestures regularization is discussed as a way to cope with projection ambiguities. Computation is reduced by registering only the moving parts of the body.

1. Introduction

1.1 State of the art

Gestures are an important and natural means of communication. Their acquisition allows animation of virtual actors, gesture coding possibly using MPEG-4, gesture based interface. It can be a basis for medical diagnosis. In the long run, gesture recognition could even allow sign language interpretation by machine vision.

There have been a lot of methods for gestures acquisition. Data gloves and electromagnetic or inertia sensors are frequently used as input devices but are expensive, fragile, and block natural interaction. Some computer vision algorithms rely on markers [1] but suffer from possible occlusions. Gesture acquisition by machine vision can rely on 2D aspect analysis [2], [3] or on 3D modeling [4]–[10]. 2D methods can recognize a limited number of gestures usually after some learning. 3D methods take advantage of the knowledge in the form of model geometry and try to compute explicitly the translations and rotations of joint. 3D methods may use two cameras [7], [8] or more [9], [10].

(*) M. Bomb has now returned to his studies in the Department of Electrical Engineering, Institute of Technology, Banaras Hindu University, Varanasi-221005, India (phone: +91-294-525234).

We are interested in gesture acquisition with a single camera. Cutler and Turk [3] use size and shape of motion blobs in the images for movement recognition. Note that blobs do not provide complete information about the attitude. Brand and Kettnaker [7] use Hidden Markov Models to infer 3D body orientation from low-resolution silhouettes after some training. Ouhaddi and Horain [10] use hand image sequences from a single camera to register a hand model.

We propose model based gesture acquisition and minimization of cost function for model registration. Kush and Huang [11] registered 3D hand model by minimizing a cost function by local variations of the degrees of freedom. Ohya and Kishino [12] have registered human body model by genetic algorithms. Gavrila and Davis [13] have searched hierarchically a tree representing all possible hand configurations in a discreet parameter space. Shimada *et al.* [14] use statistic dependence rules between successive configurations.

1.2 Overview of our method

We propose a method to capture 3D gestures using only one camera. It does away with markers and gloves, and does not require any *a priori* information on gestures. We use only static and dynamic biomechanical constraints to limit gestures.

It is based on matching image features with model features. We use color segmentation of images as a feature for model registration. An iterative algorithm is used to minimize the mismatch between the model projection and the image with respect to biomechanical constraints.

Projection of the model on the image plane can be similar for several attitudes leading to registration ambiguities. Gesture regularization is discussed as a way to handle these.

Finally, the computation cost is reduced by detecting the moving parts of the articulated model and by limiting the optimization process to the parameters of those parts.

2. Human body model

The articulated human upper body model we use has 23 degrees of freedom (Fig. 1). These parameters together form a vector $\mathbf{q}$ that completely defines the model attitude.

Each of these degrees of freedom has a higher and a lower limit resulting from static biomechanical constraints that prevent the model from attaining unnatural attitudes. Dynamic constraints form the basis of gesture regularization [15].

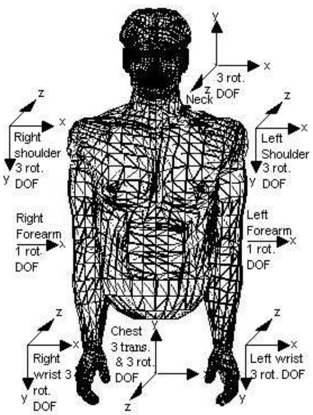

Figure 1. Upper body model with 23 degrees of freedom.

3. Image feature extraction

Registering a 3D model on images relies on matching projected model features with image features. For gesture acquisition, edges [6], [9], [10], motion [3], texture [16] or color [6] have been used.

Skin can be efficiently detected in images from its chrominance since it does not change with the orientation to light source [17]. Uniform colored clothes can also be extracted the same way. Thus, we used color segmentation on the C_bC_r components of the YC_bC_r color space.

3.1 Classification

The mean and covariance matrix of each class of chrominance are computed from examples. Pixels are classified into their closest class. Using a Mahalanobis metric [18], the distance from a pixel of chrominance $\mathbf{x} = (C_b, C_r)^T$ to the i^{th} class of mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ is $\mathbf{d}_i = \sqrt{\left(\mathbf{x} - \boldsymbol{\mu}_i\right)^T \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_i\right)}$.

If the smallest distance is greater than a given threshold, the pixel is considered as noise and not classified in any class. Fig. 2(a) shows an original image while Fig. 2(b) and (c) compare its segmentation in the RGB and YC_bC_r color spaces respectively.

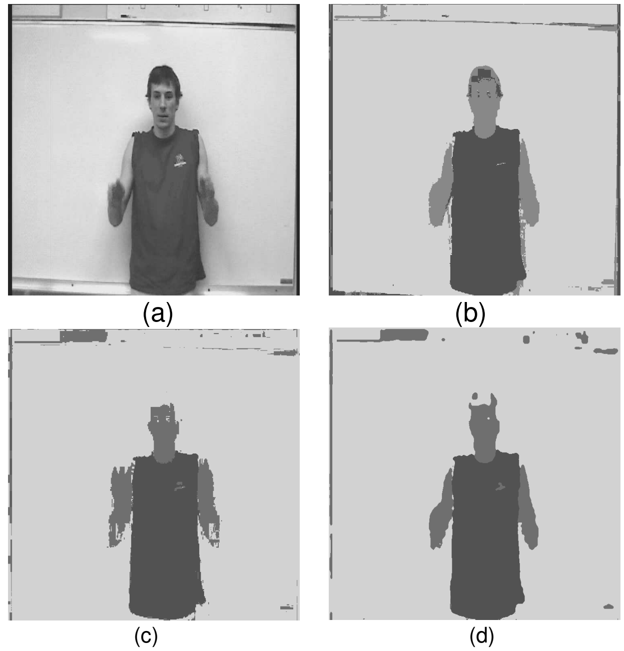

Figure 2. (a) A typical image from our video sequences. (b) Image segmented in RGB color space: shadow near the chest is classified incorrectly as skin. (c) Image segmented in YC_bC_r color space: misclassification has decreased significantly. (d) Final segmentation after mode filtering.

3.2 Mode filtering

Segmented image contains noise as seen in Fig. 2(c), which has to be removed for better feature extraction. We use mode-filtering [19] for this. For each pixel, we calculate the most frequent class in its neighborhood and assign the pixel to this class. The choice of the neighborhood size is a compromise between noise reduction and computation time. We used 13×13 neighborhood. A result is shown in Fig. 2(d).

4. Matching the model projection on the segmented image:

For each image in the sequence the model is registered by matching the color features of the projected model and the segmented image. The color of a particular model part is set to be the same as that of the corresponding part in the segmented image.

Let $\mathbf{q}_t^k$ denote the parameter vector at k^{th} iteration for the t^{th} image in the sequence. We project the model in the attitude defined by $\mathbf{q}_t^k$, onto the image plane using a z-buffer algorithm modified to retain the colors of model parts. Their matching is evaluated using a non-overlapping ratio:

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

where A_i is the set of pixels in the c^{th} color class, B_c^k is projection of the model parts associated with the c^{th} color class for the model attitude defined by $\mathbf{q}_t^k$, and m is the number of color classes.

5. Optimization and cost function

Registration is achieved by iteratively minimizing a cost function. The cost function $E(\mathbf{q}_t^k)$ can be the non-overlapping ratio: $E(\mathbf{q}_t^k) = F(\mathbf{q}_t^k)$.

Because the gradient of $F(\mathbf{q}_t^k)$ cannot be easily computed, we do not use gradient-based methods. Ouhaddi has compared optimization methods for gesture registration. We follow his choice of the downhill simplex method. It allows easy integration of the static biomechanical constraints by bounding the simplex transformations inside the convex domain defined by the constraints [6].

The condition for stopping iterations may be specified by maximum number of iteration or a minimum required change in the value of cost function from one iteration to the next under which the convergence is assumed.

6. Gesture regularization

It may happen that the model projection on the image plane is the same for different model attitudes (Fig. 3). This ambiguity may lead to incorrect registration. Moreover, noisy image segmentation may cause shaky model attitudes in successive images. Gesture regularization helps with these problems.

Kalman filtering has been used for gesture regularization [9], [10]. If $\mathbf{q}_t$ denote our set of $n = 23$ parameters and $\dot{\mathbf{q}}_t$ their velocities in the t^{th} image frame, then we define the state vector as: $\mathbf{x}_t = [\mathbf{q}_t \quad \dot{\mathbf{q}}_t]^T$.

Kalman filtering models the state vector $\mathbf{x}_t$ as a linear prediction $\hat{\mathbf{x}}_t^-$ added with noise $\mathbf{w}_{t-1}$: $\mathbf{x}_t = \hat{\mathbf{x}}_t^- + \mathbf{w}_{t-1}$ where $\hat{\mathbf{x}}_t^- = \mathbf{A}\mathbf{x}_{t-1}$ and $\mathbf{A}$ is a $2n \times 2n$ matrix. Using first order linear difference equations, we write:

$$\hat{\mathbf{q}}_t^- = \mathbf{q}_{t-1} + \dot{\mathbf{q}}_{t-1} \quad (1) ; \qquad \hat{\dot{\mathbf{q}}}_t^- = \dot{\mathbf{q}}_{t-1} \quad (2).$$

Thus we get $\mathbf{A} = \begin{bmatrix} \mathbf{I}_n & \mathbf{I}_n \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix}$.

Innovation from this prediction is computed from measurement. In our case, this measurement is achieved from image registration. Because of the previously mentioned possible ambiguities in model projection, the measurement we achieve by registration may suffer a large deviation, resulting in a small Kalman gain and poor tracking. Unfortunately, Kalman filtering keeps measurement and regularization as successive independent steps.

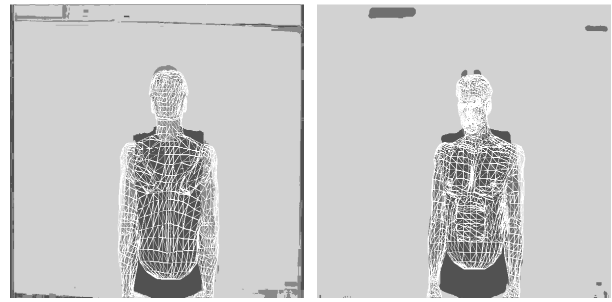

Figure 3. Example of a registration ambiguity: when the model is facing in front or is turned 180° its projection remains mostly the same.

Starting from the same prediction $\hat{\mathbf{x}}_t^-$, we choose a method similar to the one proposed by Lowe [20]. As a difference to Kalman filtering, Lowe's method integrates measurement and regularization. The algorithm uses a prediction of new values of parameters similar to the Kalman filter. These are predicted from the position and attitude parameters of the previous image using (2). However, velocities are now determined from the difference between the attitude parameters in two previous images.

At each iteration of the optimization algorithm we get an updated list of parameters $\mathbf{q}_t^k$. Lowe proposes a regularization term derived from the standard deviation σ_i expected for parameter i, i=1 to n. These can be regarded as statistical dynamic biomechanical constraints. The regularization term is defined as $R(\mathbf{q}_t^k) = \sqrt{\left(\sum_{i=1}^{n} \left(\frac{\hat{q}_t^{i-} - q_t^{k,i}}{\sigma_i} \right)^2 \right)}$ where, $\hat{q}_t^{i-}$ is the value predicted for the i^{th} parameter from (1), $q_t^{k,i}$ is the value obtained for i^{th} parameter at the k^{th} optimization iteration.

We use $R(\mathbf{q}_t^k)$ as another term in the cost function to be minimized. The cost function then becomes:

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

In case of ambiguity, this term introduces a penalty for that model attitude which is far from the prediction.

We determined the standard deviations of parameters from movements observed in about 100 signs of the French sign language [21].

7. Determining the moving parts

Optimization has to be repeated for each image in the sequence and hence requires much computation. To reduce the execution time, the moving parts are detected and optimization is restricted to their parameters.

Motion detection can be achieved by many algorithms. Cutler and Turk [3] define motion blobs. Little and Boyd [22] use optical flow characteristics to detect motion. Bobick and Davis [23] use motion energy images (MEI) and motion history images (MHI) as templates for expression recognition. Black *et al.* [24] propose a robust estimation of optical flow by representing the brightness variations as a probabilistic mixture of camera motion, illumination phenomena, specular reflections and iconic changes specific to objects being viewed.

We detect the moving parts by comparing the color indexes of successive images. This allows us to determine which color region moves in front of which other region (Fig. 4). The model parts that are projected on the moving region are to be selected for registration.

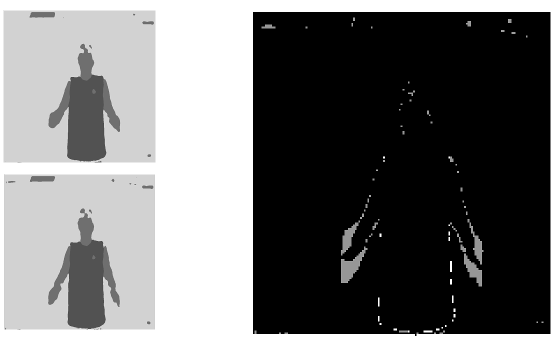

Figure 4. Motion detection: left images show segmentations with arms moving down; the right image is the result of motion detection.

8. Results and conclusion

Results on an image sequence are shown in Fig. 5. Self-occlusions are correctly processed.

This method allows us to get rid of the cumbersome markers, detectors and gloves, and just use a single camera. It does not require any *a priori* knowledge on gestures.

However it is computationally expensive because of the registration using iterative minimization of a complex functional. Another limitation comes from our color segmentation step that requires that clothes, skin and background be of different colors. Finally Lowe's regularization method might be integrated with the Kalman filtering. We are currently working on this issue.

Our results can be used to animate virtual actors and are being considered for recognition of a limited subset of the sign language [25].

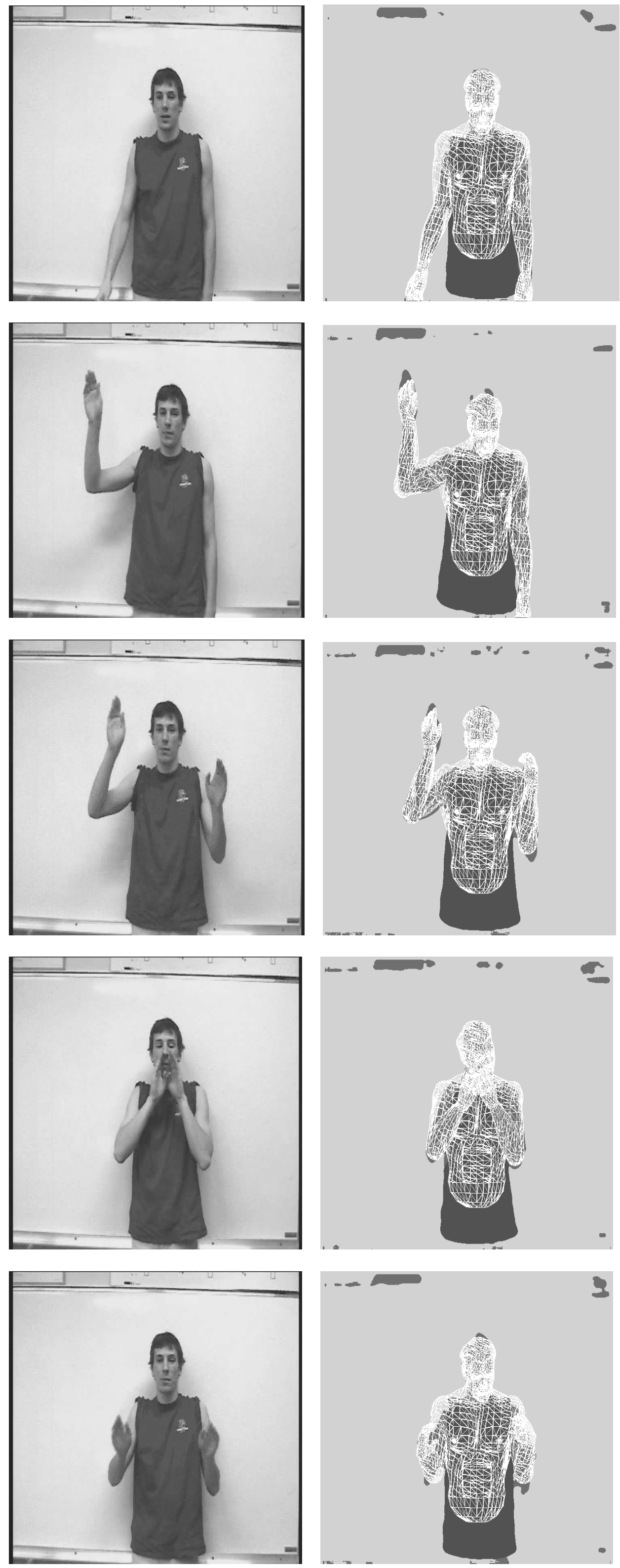

Figure 5. Results of gesture registration on an image sequence. Original images are shown on the left and corresponding registration images are shown on the right.

10. References

[1] Y. Iwai, T. Hata, and M. Yachida, "Gesture recognition based on subspace method and hidden Markov model", *Proc. the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 1997, vol. 2, pp. 960–966.

[2] J. Martin and J. L. Crowley, "An appearance based approach to gesture recognition", *Proc. 9th ICIAP, Lecture notes in computer science 1311*, 1997, Springer Verlag, pp. 340–347. Available from http://www-prima.imag.fr/~jmartin/english-publis.html.

[3] R. Cutler and M. Turk, "View-based interpretation of real-time optical flow for gesture recognition", *Proc. IEEE FG'98*, 1998. Available from http://www.cs.ucsb.edu/~mturk/research.htm.

[4] M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video", *IEEE tPAMI*, vol. 22, no. 8, 2000, pp. 844–851.

[5] J. M. Rehg and T. Kanade, "Visual tracking of high DOF articulated structures: an application to human tracking", *Proc. 3rd ECCV*, 1994, vol. 2, pp. 37–46.

[6] H. Ouhaddi and P. Horain, "3D hand gesture tracking by model registration", *Proc. IWSNHC3DI'99*, 1999. Available from http://www-eph.int-evry.fr/~horain.

[7] R. Plaenkers and P. Fua, "Model-based silhouette extraction for accurate people tracking", *Proc. ECCV 2002*, p. 325 ff. Available at http://vrlab.epfl.ch/Publications/pdf/Plaenkers_Fua_ECCV_02.pdf.

[8] A. V. Nefian, R. Grzeszczuk, and V. Eruhimov, "A statistical upper body model for 3D static and dynamic gesture recognition from stereo sequences", *Proc. ICPR'01,* 2001, vol. 2, pp. 286–289.

[9] Q. Delamarre and O. Faugeras, "3D articulate models and multi-view tracking with physical forces", *CVIU journal*, 2001, vol. 81, pp. 328–357. Available from http://www-sop.inria.fr/robotvis/personnel/qdelam/PUBLI.html.

[10] I. Kakadiaris and D. Metaxas, "Model-based estimation of 3D human motion", *IEEE tPAMI,* Dec. 2000, vol. 22, no. 12, pp. 1453–1459.

[11] J. J. Kush and T. S. Huang, "Vision based modeling and tracking for virtual teleconferencing and telecollaboration", *Proc. ICCV 95*, 1995, pp. 666–671.

[12] J. Ohya and F. Kishino, "Human posture estimation from multiple images using genetic algorithm", *Proc. 12th ICPR*, 1994, vol. 1, pp. 750–753.

[13] D. M. Gavrila and L. S. Davis, "3D model based tracking of humans in action: a multi-view approach", *Proc. IEEE CVPR*, San Francisco, CA, 1996, pp. 73–80.

[14] N. Shimada, Y. Shirai, Y. Kuno, and J. Miura, "Hand gesture estimation and model refinement using monocular camera – Ambiguity limitations by inequality constraints", *Proc. IEEE FG'98*, 1998, pp. 268–273. Available from http://www-cv.mech.eng.osaka-u.ac.jp/~shimada/research/research2.html.

[15] M. Iiyama, Y. Kameda, and M. Minoh, "Estimation of the location of joint points of human body from successive volume data", *Proc. ICPR'00*, 2000, vol. 3, pp. 699-702. Available from http://www.computer.org/proceedings/icpr/0750/Volume 3/0750toc.htm.

[16] F. Lerasle, G. Rives, and M. Dhome, "Human body limbs tracking by multi-ocular vision", *Actes RFIA'98*, vol. I, 1998, Clermont-Ferrand, France, pp. 193–199.

[17] S. Kewei, F. Xitian, C. Anni, and S. Jingao, "Automatic face segmentation in YCrCb images", *Proc. 5th Asia-Pacific Conf. on Communications* and *4th Optoelectronics and Communications Conf.*, vol. 2, 1999, pp. 916–919.

[18] M. Partridge and M. Jabri, "Face recognition using a new distance metric", *Proc. IEEE Signal Processing Society Workshop Neural Networks for Signal Processing X*, vol. 2, 2000, pp. 584–593.

[19] E. R. Davis, *Machine Vision, Theory Algorithms Practicalities*, 2nd Edition, Academic Press, 1995.

[20] D. G. Lowe, "Robust model-based motion tracking through the integration of search and estimation", *Int. Journal of Computer Vision*, vol. 8, no. 2, 1992, pp. 113-122.

[21] *Les signes de Mano*, Cd-rom, IVT - ABAQUE-micro - je.tu.il. Available from http://www.ivtcscs.org/produits.

[22] J. J. Little and J. E. Boyd, "Recognizing people by their gait, the shape of motion", *Videre: Journal of Computer Vision Research,* Winter 1998, vol. 1, no. 2, The MIT Press. Available from http://www.cs.ubc.ca/nest/lci/papers/1998/publications.html.

[23] A. F. Bobick and W. D. Davis, "Action recognition using temporal templates", *Computational Imaging and Vision*, *Motion Based Recognition*, vol. 9, M. Shah and R. Jain, Ed. Kluwer Academic Publishers, 1997, pp. 125–146.

[24] M. Black, D. Fleet and Y. Yacoob, "Robustly Estimating Changes in Image Appearance", Computer Vision and Image Understanding, 2000, vol. 78, pp. 8–31, Available from http://citeseer.nj.nec.com/black00robustly.html.

[25] P. Horain, "Vers l'acquisition du geste par vision artificielle pour l'interprétation de la langue des signes", Feb. 2002. Available at http://www-eph.int-evry.fr/~horain/ARC-LSF.