A Robust Model-Based Approach for 3D Head Tracking in Video Sequences

Marius Malciu and Françoise Prêteux
Institut National des Télécommunications
ARTEMIS Project Unit
{Marius.Malciu , Francoise.Preteux}@int-evry.fr

Abstract

We present a generic and robust method for model-based global 3D head pose estimation in monocular and non-calibrated video sequences. The proposed method relies on a 3D/2D matching between 2D image features estimated throughout the sequence and 3D object features of a generic head model. Specifically, it combines motion and texture features in an iterative optimization procedure based on the downhill simplex algorithm. A proper initialization of the pose parameters, based on a block matching procedure, is performed at each frame in order to take into account large amplitude motions. For the same reason, we have developed a non-linear optical flow-based interpolation algorithm for increasing the frame rate. Experiments demonstrate that this method is stable over extended sequences including large head motions, occlusions, various head postures and lighting variations. The estimation accuracy is related to the head model, as established by using an ellipsoidal model and an ad hoc synthesized model. The proposed method is general enough to be applied to other tracking applications.

Keywords : 3D object model, monocular video sequences, head tracking, 3D pose estimation, 3D/2D registration, 3D/2D features, motion, optical flow, block matching, texture, temporal interpolation, downhill simplex method.

1. Introduction

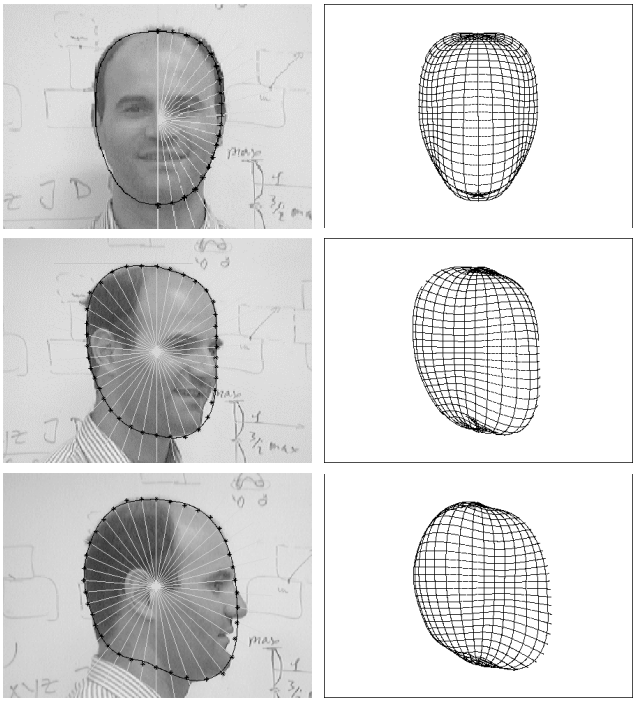
The main problems in model-based video coding techniques deal with object tracking and 3D pose estimation in complex scenes. Given a 3D global object model, pose estimation can be defined as recovering the model parameters (translations, rotations and scaling factor) so that the projected 3D model features match the 2D image features. In the particular case of human head tracking and related 3D pose estimation in video sequences, the encountered difficulties are due to head geometry and complex movements including large deformations. In addition, natural scenes often involve arbitrary and complex background and foreground texture information together with unknown and variable lighting conditions. The most common approach for model-based coding of facial image sequences [1-5] decomposes the head model tracking problem into two parts : 1) global model adaptation by taking into account the motion of the entire head, and 2) local model adaptation in order to simulate the local deformations of the face characteristic components. Here, we adopt such an approach and address the issue of 3D global model-based head pose estimation in monocular image sequences acquired by using an uncalibrated and mobile camera under non-stable lighting conditions. The principle of the method proposed here is a 3D/2D matching between 3D model features and 2D image features by using a downhill simplex-based optimization procedure [6]. Features based on motion (optical flow and block matching) and texture are taken into account. In order to guarantee the stability and the estimation accuracy in case of large amplitude motions and lighting condition variations, the initial frame rate is increased by applying a non-rigid and temporal interpolation procedure constrained by the displacement field and modeled as a multiple source wave motion [7].

In Section 2, we propose an analytically-based approach for generating 3D head-like surfaces with an arbitrary degree of approximation. Section 3 describes the temporal interpolation procedure based on an undulatory motion modeling. In Section 4, the principles underlying the 3D estimation procedure are introduced. The cost function to be minimized is defined as an error function involving optical flow and texture features. A visibility principle is then taken into account within the

cost function, in order to increase the robustness and accuracy of the 3D pose estimation. The initialization step is performed by applying a block matching procedure on each frame. In Section 5, the obtained results are presented and discussed for both simulated and real image sequences.

2. 3D head modeling

The most popular approaches for synthesizing 3D head-like surfaces are based on analytical representations or polygonal meshes. We adopt the former representation which offers the advantage of compactness and easy manipulation. In our experiments, the non-deformable 3D head is modeled as an ellipsoidal surface or, alternatively, as an *ad hoc* Fourier-synthesized surface. In order to obtain a head-like closed surface, a set of points corresponding to the head profile, viewed according to three projections, is fitted by a limited Fourier expansion of the surface.



(a) (b)

Figure 1. (a) Fitted head profiles and (b) Head-like synthesized surface, with respect to three viewpoints.

The approximation accuracy depends on the order of the expansion and on the sampling density along the profile. Figure 1 shows the surface obtained by using a Fourier expansion of order 4 using spherical coordinates. Results are similar for higher orders.

3. Undulatory-based modeling for sequence interpolation

By using the displacement field computed between two sequence frames, a non-linear and non-rigid interpolation method for increasing the frame rate is developed as follows [7]. Let $I_0(\mathbf{x})$ and $I_1(\mathbf{x})$ be two similar images and $\alpha \in [0\,,1]$ a real number. Here, $\mathbf{x}$ represents the 2D spatial coordinates in the image plane. Let us assume that we aim at generating an intermediate image $I_\alpha(\mathbf{x})$ in a continuous fashion, so that it corresponds to the first image for $\alpha = 0$ and to the second one if $\alpha = 1$. Using an optical flow algorithm, we compute the displacement fields $\mathbf{v}_{01}(\mathbf{x})$ (respectively $\mathbf{v}_{10}(\mathbf{x})$) between frames I_0 and I_1 (respectively I_1 and I_0). The image I_α is then generated pixel by pixel in the following way : for each location $\mathbf{x}$, the k nearest neighbors among the points $\mathbf{x}+\alpha\cdot\mathbf{v}_{01}(\mathbf{x})$ and $\mathbf{x}+(1-\alpha)\cdot\mathbf{v}_{10}(\mathbf{x})$ are selected. Denoting by $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_k, \mathbf{y}_k)$ these pixel/neighbor pairs, let $l(i)$ be the image index defined as follows :

$$l(i) = \begin{cases} 0, & \text{if } \mathbf{y}_i = \mathbf{x}_i + \alpha\cdot\mathbf{v}_{01}(\mathbf{x}_i) \\ 1, & \text{if } \mathbf{y}_i = \mathbf{x}_i + (1-\alpha)\cdot\mathbf{v}_{10}(\mathbf{x}_i) \end{cases},$$

for $i = 1, 2, \ldots, k$. The α-intermediate image at location $\mathbf{x}$ is computed as the following linear combination :

$$I_\alpha(\mathbf{x}) = \frac{\displaystyle\sum_{i=1}^{k} \frac{1}{d(\mathbf{x},\mathbf{y}_i)} I_{l(i)}(\mathbf{x}_i)}{\displaystyle\sum_{i=1}^{k} \frac{1}{d(\mathbf{x},\mathbf{y}_i)}} .$$

Here, d denotes some distance in the 2D plane, chosen in practice as the L^1-distance.

This interpolation method can be related to wave propagation theory. Here, according to the superposition principle, the intermediate image is generated by a group of waves emitted from various sources (pixels) located in the neighborhood of the current pixel. We have shown that such a generating process is able to overcome the limitations of standard linear interpolation methods and to control the continuity of the resulting displacement field.

Using this method, we have properly interpolated both rigid head motion and facial deformations. The corresponding results are presented in Figure 2, where the displacement fields have been computed using the Quénot's algorithm for estimating optical flow [8].

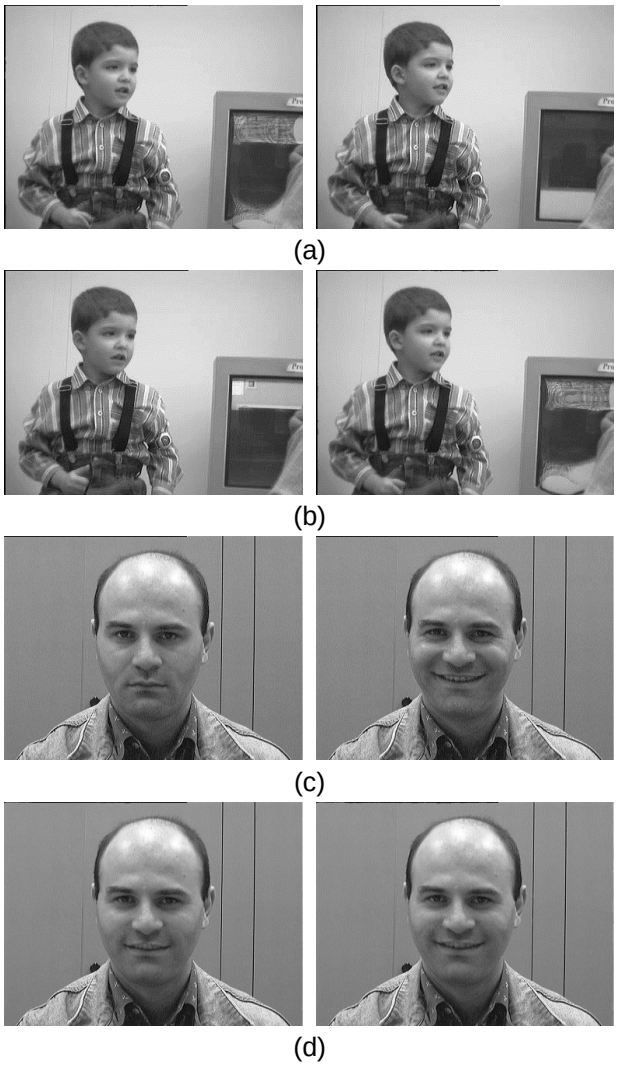


Figure 2. (a,c) Original and (b,d) interpolated images.

4. 3D/2D matching principle and 3D pose estimation

The developed 3D head pose estimation method is based on a 3D/2D matching between 3D model features and 2D image features. The model features taken into account correspond to the geometry of the 3D model, while the image features refer to the computed optical flow in the head region and to the head texture. The texture feature is solely used as a measure of the matching accuracy between model and image data. The optical flow contributes to the matching process at various levels. Basically, optical flow provides useful synthetic and global information for guiding and controlling the matching procedure. Moreover, optical flow is also used to increase the frame rate when large displacements occur. Finally, displacement field discontinuities can be exploited in order to discriminate occluded and non-occluded head regions.

Within the 3D/2D matching procedure, the parallel projection model is considered. In the 3D camera coordinate system, the pose of a rigid object is defined by means of three angles, two translation parameters and a scaling factor, measured with respect to a reference object position.

The estimation step is achieved by applying an iterative updating procedure. Assuming that the estimated pose vector of the model in the n^{th} sequence frame, denoted by $\hat{\mathbf{p}}_n$, is known, an error-function which measures the discrepancy between the 3D model features corresponding to an arbitrary pose $\mathbf{p}$ and the $n+1^{\text{th}}$ frame features, denoted by $E(\mathbf{p},\hat{\mathbf{p}}_n)$, is defined. An updated estimate of the pose vector, $\hat{\mathbf{p}}_{n+1}$, minimizing $E(\mathbf{p},\hat{\mathbf{p}}_n)$ with respect to $\mathbf{p}$, is then computed by using the downhill simplex method [6].

The error-function is defined according to the following steps :

- when the sole texture is taken into account, a 2D-3D head texture mapping from this frame onto the model surface is performed, yielding a textured model. The difference between the projected texture of the model and the texture of the current frame, denoted by $\varepsilon_{\text{texture}}$, provides a matching accuracy measure :

$$\varepsilon_{\text{texture}}(\mathbf{p},\hat{\mathbf{p}}_n) = \int_{\mathbf{x}\in I_n} \left\| F_{n+1}(\pi(\tau_{\mathbf{p}}(\mathbf{x}))) - F_n(\pi(\tau_{\hat{\mathbf{p}}_n}(\mathbf{x}))) \right\|.$$

Here, F_n and F_{n+1} denote the previous and the current frame respectively, π is the projection transform, $\tau_{\mathbf{p}}$ is the 3D geometrical transform which maps the model from the reference position into the $\mathbf{p}$-position, I_n is the set of all the visible points of the model in the $\hat{\mathbf{p}}_n$-position and $\|\cdot\|$ denotes the L^1-norm.

Concerning optical flow-based 3D head tracking, a matching accuracy measure on the current frame, denoted by $\varepsilon_{\text{optical flow}}$, is defined as the difference between the displacement field estimated in the head region (using some optical flow algorithm) and the projected 3D displacement field induced by the model rigid motion :

$$\varepsilon_{\text{optical flow}}(\mathbf{p},\hat{\mathbf{p}}_n) = \int_{\mathbf{x}\in I_n} \left\| \hat{\mathbf{v}}(\pi(\tau_{\hat{\mathbf{p}}_n}(\mathbf{x}))) - \pi(\tau_{\mathbf{p}}(\mathbf{x}) - \tau_{\hat{\mathbf{p}}_n}(\mathbf{x})) \right\|,$$

where $\hat{\mathbf{v}}$ denotes the displacement field computed via

the optical flow algorithm.

A third possibility consists in combining both texture and optical flow for defining a measure of the discrepancy between the 3D model features and the current frame features. In this case, the error ε is expressed as a linear combination of $\varepsilon_{\text{texture}}$ and $\varepsilon_{\text{optical flow}}$: $\varepsilon = a \cdot \varepsilon_{\text{texture}} + b \cdot \varepsilon_{\text{optical flow}}$.

- the model points close to occluding contours, that may vanish from a frame to the next one, are a possible source of estimation errors. We define a visibility index as a positive function on the model surface in a given position, which penalizes the points close to occluding contours; whenever the model is convex, a simple way to define such a function is to consider the projection of the model surface normal to the perpendicular direction on the image plane.
- the estimation procedure may be seriously influenced by occluding objects that move fast with respect to the head. The presence of such an object will break the regularity of the displacement field in the head region. Consequently, a simple rule can be derived in order to detect the occluded regions : if in a region the displacement field computed via the optical flow algorithm is similar to the displacement field induced by the model rigid motion, then the region is not occluded and vice versa. By using a simple distance-based classification of the displacement field related to the model one, we associate a binary occlusion index with each point of the head model in the pose estimated for the previous frame.

Combining these three principles, the error-function *E* is defined as one of the three previously-defined error measures $\varepsilon_{\text{texture}}$, $\varepsilon_{\text{optical flow}}$ or ε , where the integrands are locally weighted by the product of the visibility index and the occlusion index. An accurate 3D head pose estimate corresponds to the global minimum of *E*. Detecting this global minimum strongly depends on the initialization of the minimization procedure. Here, the initialization step is achieved by performing a block matching on the head pixels between the previous and the current frames.

5. Results

In order to perform a quantitative evaluation of the proposed 3D pose estimation algorithm [9], three synthetic image sequences were generated as follows :

- 3D head pose parameters were automatically extracted in three real image sequences corresponding to a slowly moving person ("Sorin" sequence), a rapidly moving child ("Corneliu" sequence) and a very rapidly moving person ("Armel" sequence);
- a 3D mesh head model, textured in a realistic manner, was animated with respect to these 3D pose parameters and
- the textured and animated head model was projected into a video sequence representing an office background captured with a mobile camera.

Several synthetic images generated in this way are presented in Figure 3.

Figure 3. Images synthesized from the test sequences.

A statistical analysis of the pose parameter estimation errors has been performed in order to evaluate the performances of the proposed algorithm. Figure 4 shows the error distribution functions of the head pose parameters computed for all the synthetic images (more than 600). Here, α , β and γ denote the absolute errors of the rotation angle estimates (expressed in degrees), with respect to *z*, *x* and *y* axis, respectively, in a coordinate system having the *x* and the *y* axis in the image plane; t_x and t_y represent the absolute errors of the translations estimates (in pixels) and s is the relative error of the scale factor estimate (expressed in percents). For all the test sequences, registration failure occurs after a few tens of frames when only optical flow-based estimation is performed. As a general remark, combining texture and optical flow information leads to a more accurate pose estimation, especially when fast head rotations occur. In our experiments, the values of the weighting coefficients *a* and *b* are 1 and 2 respectively. For 90% of test images estimation errors are less than 3º for rotation angles, 2 pixels for translation components and 8% for the scaling factor.

The 3D pose estimation algorithm was tested on 6 natural image sequences of different types: inside or outside scenes, moving or static camera/background, and

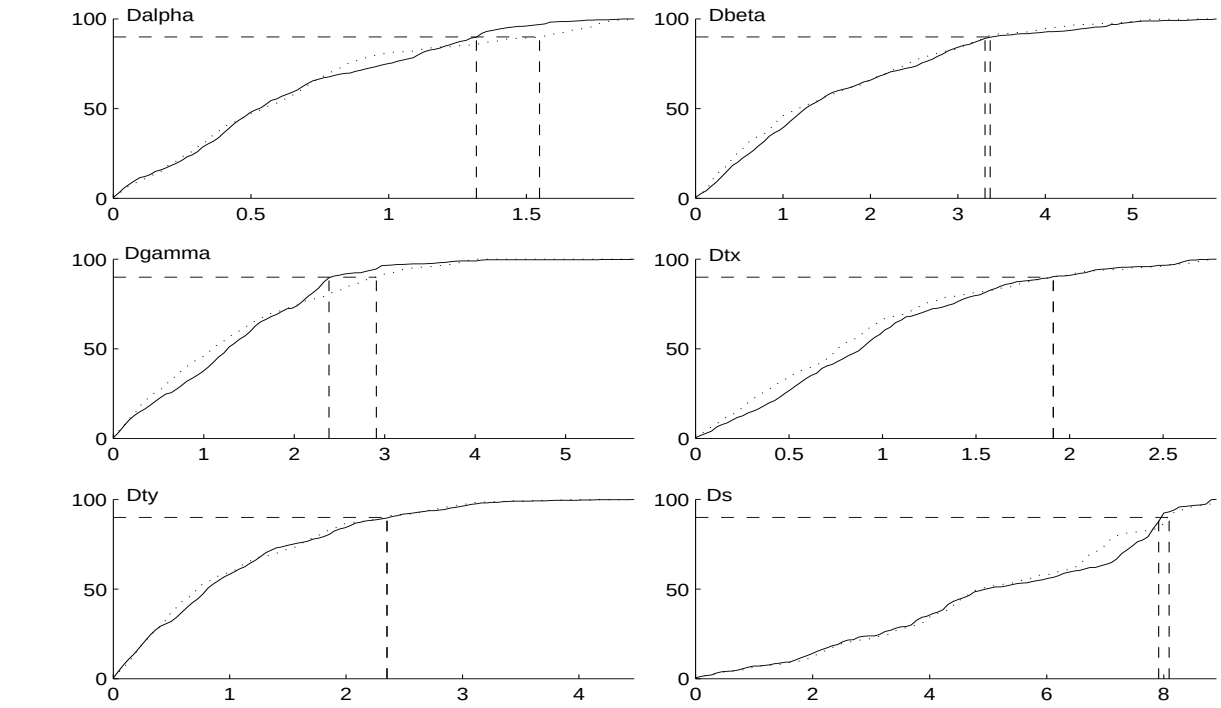


Figure 4. Error distribution functions of the head pose parameters corresponding to the synthetic image sequences. Dalpha (Dbeta, Dgamma, Dtx, Dty and Ds, respectively) denote the distribution function of alpha (beta, gamma, tx, ty and s, respectively).

presence or absence of head occlusion. Several testing results are presented in Figure 5, where the head model in its estimated pose has been superimposed in white. The experiments demonstrate the stability of the algorithm and the accuracy of the 3D pose estimation for the following configurations :

- mobile camera;
- mobile background;
- large amplitude camera zoom;
- partial head occlusions;
- non-stable lighting conditions.

6. Conclusion and future work

The proposed 3D head pose estimation method relies on a 3D/2D matching between 2D image features estimated throughout the sequence and 3D object features of a generic head model. In order to take into account large amplitude head motions, we have developed a non-linear optical flow-based interpolation algorithm for increasing the frame rate. We have demonstrated that this method is stable over extended sequences including large head motions, occlusions and various head postures. Even though this technique has been designed for model-based object tracking in the context of head tracking, the method is general enough to be applied to other tracking problems. Our future work will deal with the analysis and synthesis of facial image sequence in model-based image coding.

References

[1] H. Li, P. Roivainen, R. Forchheimer, "3-D motion estimation in model-based facial image coding", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6), June 1993, pp. 545-555.

[2] G. Bozdagi, A. M. Tekalp, L. Onural, "3D Motion estimation and wireframe adaption including photometric effects for model-based coding of facial image sequences", *IEEE Transactions on Circuits and Systems for Video Technology*, 4(3), June 1994, pp. 246-256.

[3] K. Aizawa, T. S. Huang, "Model-based image coding: advanced video coding techniques for very low bit-rate applications", *Proceedings of the IEEE*, 83(2), February 1995, pp. 259-271.

[4] L. Zhang, "Automatic adaptation of a face model using action units for semantic coding of videophone sequences", *IEEE Transactions on Circuits and Systems for Video Technology*, 8(6), October 1998, pp. 781-795.

[5] M. J. T. Reinders, P. J. L van Beek., B. Sankur, J.C.A. van der Lubbe, "Facial feature localization and adaptation of a generic face model for model-based coding", *Signal Processing: Image Communication*, 7, 1995, pp. 57-74.

[6] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical recipes in C: the art of scientific computing*, Cambridge University Press, 1998.

[7] M. Malciu, L.-T. Nessi, F. Prêteux; "Pose 3D du visage dans des séquences vidéos : estimation robuste par modèle d'objet", accepted for *12ème Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle RFIA'2000*, Paris, February 2000.

[8] G. M. Quénot, "The orthogonal algorithm for optical flow detection using dynamic programming", *Proceedings IEEE Int. Conference on Acoustics, Speech and Signal Processing*, San Francisco, CA, March 1992.

[9] F. Prêteux, M. Malciu, "3D head tracking in video sequences: A robust approach", accepted for publication in *Journal of Electronic Imaging*.

Figure 5: 3D head pose estimation results for the (a) "Sorin", (b) "Corneliu" and (c) "Armel" sequences, using a Fourier-synthesized head model.