ROBUST AND REAL-TIME 3D-FACE MODEL EXTRACTION

Marc CHAUMONT, Brice BEAUMESNIL

LIUPPA / IUT Bayonne,
64100 Bayonne, France.
Marc.Chaumont@iutbayonne.univ-pau.fr
Brice.Beaumesnil@iutbayonne.univ-pau.fr

ABSTRACT

This article deals with 3D-face model and 3D-pose extraction from a small set of couples of 2D-3D corresponding-points. Major drawbacks of current 3D model extraction solutions are either the computationally complexity or the over-simplified modeling. As it happens, applications like face tracking or augmented reality need a rapid, robust and descriptive-enough solution. The solution we propose is based on a two step approach in which an approximation of a 3D-face model and a 3D pose is computed and then refined in order to extract more precise parameters. The contribution of this paper is to describe how to efficiently (rapidly and robustly) solve the problem of 3D-face model and 3D pose extraction. The results obtained show rapid and robust performances which could be exploited in a more global real-time face tracking application.

1. INTRODUCTION

In the specific case of face tracking, current solutions go from pixel-based to 3D model-based approaches. We believe that 3D information necessarily has a role to play during the tracking. 3D-pose and 3D-face model give a 3D information which may help in some ambiguous situations (occultation, face orientation, luminosity variation). This paper aims at improving the tracking techniques based on 3D-face model. More precisely, we propose a robust and rapid 3D-face model extraction and 3D-pose extraction.

The 2D features points stemming from automatic algorithms [1, 2] are often noisy (2D positions are un-precise) and their number is small. Our solution to extract a 3D-face model, with the knowledge of 2D features points, takes care of those difficult constraints and moreover is well-suited for real-time applications.

The solution is divided in two steps. The first step recovers an approximation of the 3D-face model (details are given in section 3), the second step deals with the improvement of this 3D-face model and the 3D-pose extraction (explanations are given in section 4).

2. GENERAL ENERGETIC FORMULATION

With a classical pinhole camera, the projection T of a 3D point $M'_i = (X'_i, Y'_i, Z'_i)^t$ (expressed in an object coordinate system) gives a 2D point $m'_i = (u'_i, v'_i)^t$ (expressed in an image coordinate system) which may be expressed in homogeneous coordinate by the equation 1 ([3] chap.5). f is the camera's focal length; k_u and k_v are the horizontal and vertical scale factors (measured in pixels/m); u_0 and v_0 are principal point coordinates; $(t_x, t_y, t_z)^t$ is the translation vector and $r_{ij:i,j\in[1,3]}$ are rotation matrix coefficients. Figure 1 illustrate the different coordinate systems and the projection of a 3D point M'_i to a 2D point m'_i.

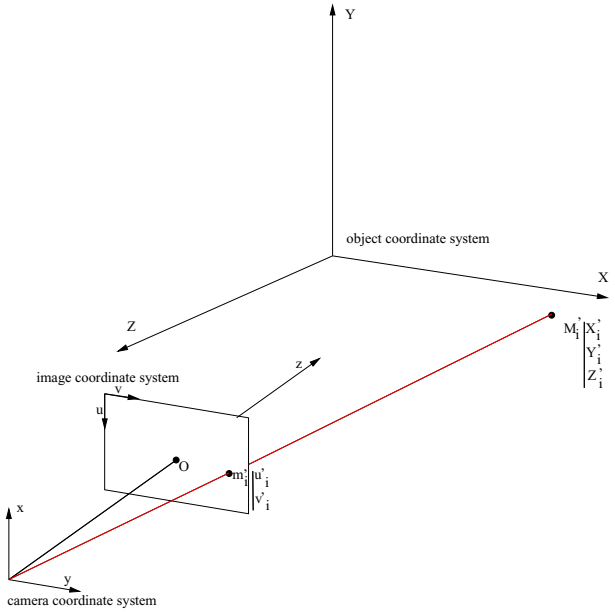


Fig. 1. The different coordinate systems

To extract the 3D-face model and the 3D-pose, we min-

0-7803-9134-9/05/$20.00 ©2005 IEEE

$$\begin{pmatrix} s.u'_i \\ s.v'_i \\ s \end{pmatrix} = \underbrace{\begin{pmatrix} \alpha_u & 0 & u_0 & 0 \\ 0 & \alpha_v & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}}_{intrinsic\ parameters} . \underbrace{\begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{extrinsic\ parameters} . \begin{pmatrix} X'_i \\ Y'_i \\ Z'_i \\ 1 \end{pmatrix} = T.M'_i,\ with \begin{cases} \alpha_u = -k_u.f \\ \alpha_v = k_v.f \end{cases} . \quad (1)$$

imize the distance error E (see equation 2) between the observed set of 2D image points $\{(u_i, v_i)^t\}$ and the projected set of points $\{(u'_i, v'_i)^t\}$. The projected set of points $\{(u'_i, v'_i)^t\}$ are obtained by projecting all the corresponding 3D-face model vertex using the T projection (see equation 1).

$$E = \sum_i (u_i - u'_i)^2 + (v_i - v'_i)^2. \quad (2)$$

Note that the 3D-face model vertex used in the minimization (equation 2) belong to a "shaped 3D-face model". By the term "shaped 3D-face model", we mean that morphology and current emotions of the treated face are caught by the 3D-face model. To obtain this "shaped 3D-face model" we displace the vertex of a known average 3D-face model named CANDIDE-3 [4]. Figure 2 shows the CANDIDE-3 wireframe.

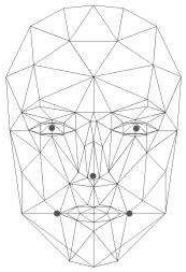


Fig. 2. Ahlberg CANDIDE-3 wireframe model

Thus, projected set of points $\{(u'_i, v'_i)^t\}$ (see equation 2) are the result of a shape displacement ($S_i.\sigma$), an animation displacement ($A_i.\alpha$) and a projection (T) of an CANDIDE-3 average 3D-face model as expressed in the following equation:

$$\begin{pmatrix} s.u'_i \\ s.v'_i \\ s \end{pmatrix} = T.[\underbrace{M_i + S_i\sigma + A_i.\alpha}_{M'_i}]. \quad (3)$$

S_i and and A_i are respectively the shape unit and the animation unit matrix, expressing the possible displacement of a vertex i. The displacement intensity is expressed by the weighting vectors σ and α. More details are given in Ahlberg's report [4]. Equation 2 minimization gives parameters T, σ, and α.

The minimization problem of equation 2, if processed directly, is difficult, not rapid enough for real-time applications and not always robust. We thus decompose the problem in two steps: first (section 3), the approximation of the 3D-face model shape (T, σ and α are coarsely computed) and second (section 4), the extraction of the 3D-pose and the improvement of the 3D-face model shape (T and 3D-model's shape are refined). Figure 3 summarise the computations order.

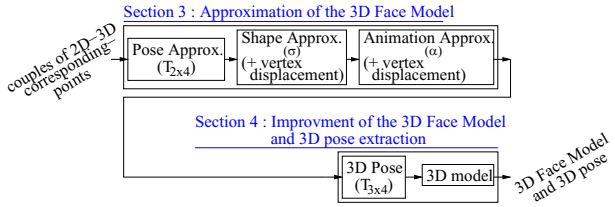


Fig. 3. General scheme of the 3D-face model and 3D pose extraction

3. 3D-FACE MODEL SHAPE APPROXIMATION

3.1. Pose approximation

The computation of projection T (given in equation 2) is not an easy task; a direct solving leads to a homogeneous linear system. Solutions in the literature such as [5] need a high number of couples of 2D-3D corresponding-points. In the case of Human face, there is a small number of salient points; projection T should then be simplified.

This simplification consists in supposing that all the 3D vertex are in a same 3D plan. This is a realist hypothesis when there are small depth differences between 3D points in comparison to the distance between the camera and the face. The projection T becomes a 2×4 matrix such that:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} (\alpha_u.r_{1i} + u_0.r_{3i})/t_z & \alpha_u.t_x/t_z + u_0 \\ (\alpha_v.r_{2i} + v_0.r_{3i})/t_z & \alpha_v.t_y/t_z + v_0 \end{pmatrix}_{i\in[1,3]} . \begin{pmatrix} M_i \\ 1 \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} a_0 & b_0 & c_0 & d_0 \\ a_1 & b_1 & c_1 & d_1 \end{pmatrix}}_{T_{2\times4}} . \begin{pmatrix} M_i \\ 1 \end{pmatrix}.$$

By canceling, from equation 2, each E's partial derivative in function of T's parameters, we obtain 2 linear systems (σ and α are set to zero). The first linear system is given below (the second linear system is obtained by re-

placing u_i by v_i and a_0 by a_1, b_0 by b_1 and so on):

$$\begin{pmatrix} \sum_i X_i^2 & \sum_i X_i.Y_i & \sum_i X_i.Z_i & \sum_i X_i \\ \sum_i X_i.Y_i & \sum_i Y_i^2 & \sum_i Y_i.Z_i & \sum_i Y_i \\ \sum_i X_i.Z_i & \sum_i Y_i.Z_i & \sum_i Z_i^2 & \sum_i Z_i \\ \sum_i X_i & \sum_i Y_i & \sum_i Z_i & \sum_i 1 \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ b_0 \\ c_0 \\ d_0 \end{pmatrix} = \begin{pmatrix} \sum_i X_i.u_i \\ \sum_i Y_i.u_i \\ \sum_i Z_i.u_i \\ \sum_i u_i \end{pmatrix}. \tag{4}$$

Those two systems are solved by using classical linear algebra tools. Note that the matrix involved in that system is very small (matrix size=4×4); its computation and its inversion is very rapid. A more robust $T_{2\times4}$ projection may be computed by adding a robust function into equation 2. The robust function may for example take the image gradient into account.

3.2. Shape approximation

Once $T_{2\times4}$ projection is computed, shape adaptation is processed. The minimization problem of equation 2, is solved by fixing $T_{2\times4}$ projection. Equation 2 is re-written such that:

$$\begin{aligned} E &= \sum_i [U_i - N.S_i.\sigma]^t.[U_i - N.S_i.\sigma], \\ &\quad with\ U_i = \begin{pmatrix} u'_i \\ v'_i \end{pmatrix} - T.\begin{pmatrix} M_i \\ 1 \end{pmatrix}, \\ &\quad and \quad N = T_{2\times3}. \end{aligned} \tag{5}$$

We obtain a linear system (equation 6) by canceling the partial derivative $\frac{\partial E}{\partial \sigma}$:

$$\underbrace{\left(\sum_i S_i^t.N^t.N.S_i\right).\sigma}_{A} = \underbrace{\sum_i S_i^t.N^t.U_i}_{B}. \tag{6}$$

Solution is such that $\sigma = (A^tA)^{-1}A^t.B$. First, note that couples of 2D-3D corresponding-points used for the filling of matrix A and vector B should be chosen as non animated points. Second, remark that (A^tA) may be non-invertible because of a too small number of couples of 2D-3D corresponding-points. The matrix A when non-invertible owns zero filled lines and columns. The diagonal coefficient, where a cross zero-line and a zero-column appears, could be set a non-zero value. Indeed, the corresponding σ's coefficient is not influenced by a single of the set of couples of 2D-3D corresponding-points. This σ's coefficient will thus be equal to zero. Third, one should take care that the solution belongs to the valid domain; each σ's coefficient belongs to the range $[-1, 1]$. The matrix involved in the system is small and sparse; its computation and its inversion are very rapid. The same reasoning may be done for α computation.

4. 3D-FACE MODEL POSE AND 3D-FACE MODEL SHAPING

In the previous section, we explained how to rapidly obtain a first approximation of a 3D-pose (matrix $T_{2\times4}$) and a first shaping of a 3D-face model. Our objective is now to recover the depth information (t_z) and to extract a more descriptive pose.

4.1. 3D-pose extraction

To extract extrinsic parameters (rotation and translation), we still have to solve the equation 2. A well known result is that intrinsic parameters may roughly be approximated without important reconstruction error [6, 7]. Intrinsic parameters are thus coarsely fixed[1] and extrinsic parameters are extracted with the well known POSIT DeMenthon algorithm [8].

4.2. 3D final shaping

Once extrinsic parameters are computed, we observe that the mapping between 2D points and 3D corresponding-points is not totally correct. This un-correct mapping is due to the approximation made on T (explained in subsection 3.1) and the shape and animation units displacements which do not fully capture the specific shape of the treated face. To obtain the exact mapping we displace each 3D point separately with taking caution to erroneously localize 2D image points (outlier points). At this stage, z-coordinates Z_i of each M_i points should not move anymore; indeed z-coordinates correspond to the object depth. The unknown X_i and Y_i coordinates are then easily obtained by solving the linear equation 7 for each couple of 2D-3D corresponding-points. A final check is processed to prevent 3D-mesh turnaround and strong vertex displacements.

$$\begin{pmatrix} ((u_i - u_0).r_{31} - \alpha_u.r_{11}) & ((u_i - u_0).r_{32} - \alpha_u.r_{12}) \\ ((v_i - v_0).r_{31} - \alpha_v.r_{21}) & ((v_i - v_0).r_{32} - \alpha_v.r_{22}) \end{pmatrix} \cdot \begin{pmatrix} X_i \\ Y_i \end{pmatrix} = \begin{pmatrix} (\alpha_u.r_{13} - (u_i - u_0).r_{33}).Z_i + \alpha_u.t_x - (u_i - u_0).t_z \\ (\alpha_v.r_{23} - (v_i - v_0).r_{33}).Z_i + \alpha_v.t_y - (v_i - v_0).t_z \end{pmatrix}. \tag{7}$$

5. RESULTS

The principal steps of our technique for the extraction of the 3D-face model and its 3D pose are illustrated in Figure 4. Figure 4(a) shows the first image of the *Foreman* sequence. Figure 4(b) is the result of the shape approximation step (sub-section 3.2); grey points represent some face-features (their location has been set manually but could have been obtained automatically using feature detectors), and black

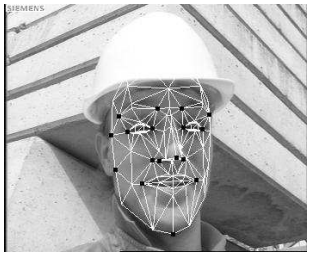
[1] f is set to 0.05, k_u and k_v are set to 5000, point $(u_0, v_0)^t$ is set to the image center

points represent the corresponding-points obtained by projecting the vertex of the 3D-face model. Figure 4(c) shows the mesh of the 3D-face model projected onto the *Foreman* image. We can notice that the mapping between 2D points and 3D corresponding-points is not totally correct. This un-correct mapping is due to an approximation on T (subsection 3.1) and to the shape animation units which are too much general. This un-correct mapping is corrected by the 3D final shaping explained in sub-section 4.2. Figure 4(d) is the WRML representation of the final 3D-face model with its 3D pose.

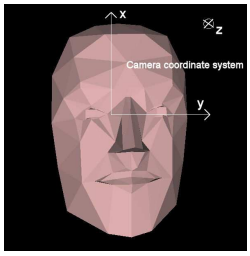
Those results are interesting for face tracking. First, the 3D-face model is obtained very rapidly without any triangulation. Second, extrinsic parameters (rotation and translation) are well adapted to model the face trajectory and predict the face position at time t knowing position at time $t-1$. Moreover, in the case of occultation, we could reasonably guess that the 3D model position do not change much and we could recover the face more easily when it re-appeared.

(a) Image from *Foreman* sequence

(b) 2D points (grey) and projected 3D points (black) (step : shape approximation)



(c) Mesh of the 3D-face model



(d) Illustration of the 3D-face model seen by the camera

Fig. 4. Illustration of some steps of the extraction of the 3D-face model and its 3D pose

6. CONCLUSION

In this paper we deal with the problem of rapid and robust 3D-face model and 3D-pose extraction. To that purpose, we use an average 3D-face model and few couples of 2D-3D corresponding-points. A succession of robust and rapid computations lead to a 3D-face model exactly fitted and a complete 3D projective camera pinhole model (camera's parameters, rotation and translation).

7. REFERENCES

[1] J. Ahlberg, "Facial feature extraction using deformable graphs and statistical pattern matching," in *Swedish Symposium on Image Analysis, SSAB*, Mar. 1999.

[2] J. Rurainsky and P. Eisert, "Template-based eye and mouth detection for 3D video conferencing," in *International Workshop on Very Low Bitrate Video, VLBV*, Sept. 2003, pp. 23–31.

[3] R. Horaud et O. Monga, *Vision par ordinateur : Outils fondamentaux*, Editions Herms, Eyrolles, 1995.

[4] J. Ahlberg, "CANDIDE-3 - un updated parameterised face," Tech. Rep., Department of Electrical Engineering, Linköping University, Jan. 2001.

[5] O. D. Faugeras and G. Toscani, "The calibration problem for stereo," in *Computer Vision and Pattern Recognition, CVPR*, 1986 June, pp. 15–20.

[6] S. Bougnoux, "From projective to euclidean space under any practical situation, a criticism of self-calibration," in *International Conference on Computer Vision, ICCV*, Jan. 1998, pp. 790–798.

[7] L.F. Cheong and C-H. Peh, "Characterizing depth distortion due to calibration uncertainty," in *European Conference on Computer Vision, ECCV*. June 2000, vol. 1842, pp. 664–677, Springer-Verlag.

[8] D. DeMenthon and L.S. Davis, "Model-based object pose in 25 lines of code," *International Journal of Computer Vision, IJCV*, vol. 15, no. 1, pp. 123–141, June 1995.