# THREE DIMENSIONAL MODEL BASED RIGID TRACKING OF A HUMAN HEAD

*Nikolaos Sarris, Dimitrios Makris and Michael G. Strintzis*

Information Processing Laboratory,
Department of Electrical and Computer Engineering,
Aristotle University of Thessaloniki
540 06 Thessaloniki, Greece
Phone: +30 31 996349; Fax: +30 31 996398
Email: {nikos, dmakris}@dion.ee.auth.gr

## ABSTRACT

Keywords: *3D model based coding, rigid head tracking, 3D motion estimation, feature tracking*

The presented work proposes substantial improvements to established methods for the two-dimensional tracking of an image and the rigid adaptation of a three-dimensional face model to a 2D projection. These are combined to produce a robust system for the rigid 3D tracking of a human head which is promising even for extreme rotation conditions where most methods fail. Our technique, while using a 2D feature-based image tracking algorithm introduced by Kanade, Lucas and Tomasi, is based on the assessment of the tracked features considering their suitability for tracking, the success in tracking them in the current frame and their reliability judged from their 3D position in a model adapted to the previous frame. Having determined the optimum set of tracked features on the current image frame these are used to readapt the 3D human face model which has been initialised to fit to the face manually detected in the first frame of the sequence.

## 1. INTRODUCTION

Our work aims to determine the rigid motion of a generic 3D human face model so that it adapts to a 2D face image assuming a perspective projection camera model of known focal length. The 3D rigid motion is defined as a 3×3 rotation matrix **R** and a 3D translation vector **T**, such as:

$$\begin{bmatrix} X_i' \\ Y_i' \\ Z_i' \end{bmatrix} = \mathbf{R} \cdot \begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix} + \mathbf{T} \qquad (1)$$

where $(X_i,Y_i,Z_i)$ are the initial positions of the model nodes and $(X_i',Y_i',Z_i')$ are the required positions of the model nodes so that their projections coincide with the 2D positions $(x_i,y_i)$ of the corresponding points in the image frame. These projections assuming a pinhole camera of focal length $f$ are given by:

$$\frac{x_i}{f} = \frac{X_i}{Z_i}, \qquad \frac{y_i}{f} = \frac{Y_i}{Z_i} \qquad (2)$$

We assume that an initial correspondence of a set of rigid 3D model nodes (such as ear, nose, or forehead nodes) to a set of 2D feature points in the first frame of the image sequence exists and we aim to track the set of feature points in subsequent frames and estimate the 3D motion of the model using the correspondences of the most reliably tracked feature points.

Various methods have been proposed for the tracking of motion in images using dense optic flow, block matching, Fourier transforms or pel recursive methods, as described in [1]. In this work a feature based approach has been preferred both for its merits in speed as well as for its natural suitability to this knowledge based node-point correspondence problem. In general, the features to be tracked may be corners, edges or points and the selection of the particular set of features must be such that they may be tracked easily and reliably. This work has mainly used the algorithm proposed by Kanade, Lucas and Tomasi, referred to as the KLT algorithm [4], enhanced in the following ways: The algorithm is not allowed to select and track any features from the image but only a set of those which correspond to rigid 3D nodes of the face model. These features, while tracked in subsequent frames of the image sequence, are sorted according to their suitability for tracking (KLT 'trackability' metric), their likelihood to have been correctly tracked (KLT 'dissimilarity' metric), and their reliability considering the orientation of the 3D normal vector of the corresponding node in the previous frame.

Having a set of 2D feature points at our disposal many are the methods which have been proposed for

the computation of the 3D motion parameters (**R, T**) of the face model [2], and their accuracy in the solution depends highly on the reliability of the given feature correspondences. Having ensured from the previous step that the selected feature correspondences are the best possible for the given tracking method we employ the method proposed in [5] and [6] enhanced to include the focal length of our camera and estimate **R** and **T** up to a scaling factor. Using the 3D model coordinates in the previous frame, however, we compute this scaling factor and determine an absolute solution for **R** and **T**.

Finally, we compute the 2D positions of those features which were considered not to have been reliably tracked, by projecting the known corresponding 3D nodes.

## 2. FEATURE TRACKING

In the proposed system a fast and reliable method to track 2D features was crucial, thus, an enhanced version of the already successful Kanade-Lucas-Tomasi (KLT) algorithm was implemented.

The KLT algorithm is based on the minimisation of the sum of squared intensity differences between a past and a current feature window, which is performed using a Newton-Raphson minimisation method. Although the KLT algorithm has proved to yield satisfactory results on its own in our system it is very important to assess the results of tracking so that the optimum set of feature correspondences is used in the stage of the model adaptation. For this reason, we sort the tracked feature points according to two criterions introduced by and closely related to the operation of the KLT algorithm, and we introduce a third criterion related to the nature of the 3D model to be adapted:

*Trackability*: The ability of a 2D feature point to be tracked reliably, which is related with the textureness of its window. Mathematically, trackability is defined as examining the minimum eigenvalue of each 2 by 2 gradient matrix [3], [4].

*Dissimilarity*: The sum of squared intensity differences, which indicates how well the feature has been tracked [3], [4].

*Reliability*: The projection of the normal to the node N, vector, on the projection ray, where N belongs to the 3D model as adapted to the previous frame, and corresponds to the 2D feature point under question. The importance of this last criterion is justified by two observations based on three-dimensional geometry:

According to our right-hand 3D coordinate system, in order for a 3D feature node to be visible in the image frame, this projection of its normal has to be positive. If this projection is close to 1, the tangent to the node surface is almost parallel to the image plane and therefore, it is very unlikely for this feature to be occluded in the next (which is the current) frame. However, if it is close to 0, there is a considerable possibility for the feature to be occluded in the next frame, because of a slight rotation of the head, and if it is not occluded it is very likely that it will lie close to the borderline of the face and may therefore be associated with the background.

Moreover, it can be shown that the greater the projection of the normal is, the greater area in the neighbourhood of the node is projected onto the image plane and the less change there is in that area between frames: Let a feature lie in the centre of a plane surface of area $E$ and let the position of this surface in the 3D space be such that the angle between the normal to the surface vector and the projection ray be $\grave{e}$. The projection of the normal will then be equal to $cos\grave{e}$. The part of the surface which will be projected onto the image plane will generally be proportional to $Ecos\grave{e}$. If the surface now rotates in such a way that $\grave{e}$ changes (rotation around x and/or y axis), the projected part of the surface will also change at a rate which will be proportional to the angle $\grave{e}$ as:

$$\frac{d(E\cos q)}{dq} = -E\sin q \qquad (3)$$

Thus, the smaller the angle $\grave{e}$ (i.e. the greater z-component), the smaller the rate of change of the 2D projected part of the 3D neighbourhood of the feature node. Considering that the KLT and most 2D tracking algorithms, are based on the 2D image plane texture around the features it is clear that the greater the z-component, the less change there will be in the projected surfaces around the node and thus, the more successful the tracking will eventually be.

## 3. 3D RIGID MOTION ESTIMATION

Having the set of the most reliably tracked feature points at our disposal we need to compute the 3D motion parameters (**R, T**) of the face model which will adapt its projection to the given one in the current image frame. Our approach is based on calculating an approximate initial solution by an enhanced version of the 8-point algorithm introduced in [5] and [6], and then optimising this solution by an iterative minimisation algorithm, as recommended in [7].

The 8-point algorithm is modified to include the focal length, $f$, of the camera which we do not

assume to be equal to 1. This involves the modification of equation (6) and the matrix **A** of equation (36) both found in [5], as described in Eq.(4), which is a result of the epipolar constraint, and Eq.(5):

$$\begin{bmatrix} x_i' & y_i' & f \end{bmatrix} \cdot \mathbf{E} \cdot \begin{bmatrix} x_i \\ y_i \\ f \end{bmatrix} = \mathbf{0} \tag{4}$$

$$\mathbf{A} = \begin{bmatrix} x_1'x_1 & x_1'y_1 & x_1'f & y_1'x_1 & y_1'y_1 & y_1'f & x_1f & y_1f & 1 \\ x_2'x_2 & x_2'y_2 & x_2'f & y_2'x_{21} & y_2'y_2 & y_2'f & x_2f & y_2f & 1 \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ x_n'x_n & x_n'y_n & x_n'f & y_n'x_n & y_n'y_n & y_n'f & x_nf & y_nf & 1 \end{bmatrix} \tag{5}$$

where $(x_i,y_i)$ and $(x_i',y_i')$ are the 2D feature correspondences of the (at least) eight most reliably tracked features.

The rotation matrix **R** and the translation vector **T** up to a scaling factor, are determined from the essential matrix **E** as described in [6].

If $|\mathbf{T}|\neq 0$ we can determine this scaling factor by using the depths of the nodes corresponding to the tracked features. Thus, as a consequence of Eq.(1), we will have:

$$\begin{bmatrix} \frac{1}{f}\mathbf{x_i'} & -\mathbf{T} \end{bmatrix} \cdot \begin{bmatrix} z_i' \\ |\mathbf{T}| \end{bmatrix} = \frac{z_i}{f} \cdot \mathbf{R} \cdot \mathbf{x_i} \tag{6}$$

From Eq.(6), we may form the following overdetermined system:

$$\begin{bmatrix} \frac{1}{f}\cdot\mathbf{x_1'} & \mathbf{0} & ... & \mathbf{0} & -\mathbf{T} \\ \mathbf{0} & \frac{1}{f}\cdot\mathbf{x_2'} & ... & \mathbf{0} & -\mathbf{T} \\ ... & ... & ... & ... & ... \\ \mathbf{0} & \mathbf{0} & ... & \frac{1}{f}\cdot\mathbf{x_n'} & -\mathbf{T} \end{bmatrix} \cdot \begin{bmatrix} z_1' \\ z_2' \\ ... \\ ... \\ z_n' \\ |\mathbf{T}| \end{bmatrix} = \begin{bmatrix} \frac{z_1}{f}\cdot\mathbf{R}\cdot\mathbf{x_1} \\ \frac{z_2}{f}\cdot\mathbf{R}\cdot\mathbf{x_2} \\ ... \\ \frac{z_n}{f}\cdot\mathbf{R}\cdot\mathbf{x_n} \end{bmatrix} \tag{7}$$

This calculation of **R** and **T** is taken as an initial solution which is further optimised by the minimisation (shown in Eq.(8)) of the distances of the projections of the feature points given in Eq.(9) from their actual 2D positions found by the tracking procedure. This minimisation is performed by a modification of the Levenberg Marquadt algorithm.

$$\min \sum_{\substack{\text{all} \\ \text{feature} \\ \text{points}}} \left( \begin{bmatrix} X_i \\ Y_i \end{bmatrix} - \begin{bmatrix} \hat{X}_i \\ \hat{Y}_i \end{bmatrix} \right)^2 \tag{8}$$

$$\hat{X} = f\frac{x'}{z'} + X_0 \quad \hat{Y} = f\frac{y'}{z'} + Y_0 \tag{9}$$

Having determined **R** and **T**, the new 3D coordinates for all the model nodes are known.

As the number of useful rigid features on a human face is usually quite small and our approach needs at least 8 well-tracked features, there is an immediate need of replacing any lost features. We assume the 2D coordinates of any feature to be those calculated by projecting the 3D nodes, thus, a reliable replacement always requires a reliable motion tracking. Moreover, as there is always the possibility that the real features are not visible due to occlusions by head rotation, we have to check again the sign and magnitude of the previously mentioned component of the feature node's normals. If it is negative the feature point is definitely invisible on the image frame. Finally, even if the normal component is positive but small, there is a great possibility for the re-projected feature to be close to the borderline of the face, in which case, a small calculation error may cause the feature to be reprojected onto the background, which of course is undesirable.

## 4. RESULTS

Results have been shown to be promising even for extreme rotation conditions where most methods have proved to fail. In figures 1 and 2 the results of tracking feature points in subsequent frames and their use for the adaptation of a generic 3D face model, are illustrated.

## 5. ACKNOWLEDGEMENTS

This work was supported by the European CEC Project VIDAS (ACTS project 057) and the GSRT project PABE. The help of COST 254 is also gratefully acknowledged.

## 6. REFERENCES

[1] F. Dufaux, F. Moscheni, *"Motion Estimation Techniques for Digital TV: A Review and a New Contribution"*, IEEE Proc. Vol 83, No 6, June 1995, pp 858-876.

[2] J.K. Aggarwal, N. Nandhakumar, *"On the Computation of motion from Sequences of Images – A Review"*, IEEE Proc, Vol 76, No8, August 1988, pp. 917-935.

[3] C. Tomasi, T. Kanade, *"Detection and Tracking of Point Features, Shape and Motion from Image Streams: a Factorization Method – Part 3"*, School of Computer Science, Carnegie Mellon University, Pittsburgh, April 1991.

[4] J. Shi, C. Tomasi, *"Good Features to Track"*, in Proceedings, IEEE Conference on Computer Vision and Pattern Recognition, 1994, pp. 593-600.

[5] R.Y. Tsai, T.S. Huang, *"Uniqueness and Estimation of Three-Dimensional Motion Parameters of Rigid Objects with Curved Surfaces"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-6, No. 1, January 1984, pp. 13-26.

[6] J. Weng, T.S. Huang, N.Ahuja, *"Motion and Structure from Two Perspective Views: Algorithms, Error Analysis, and Error Estimation"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 11, No. 5, May 1989, pp.451–474.

[7] J. Weng, N. Ahuja and T. S. Huang, "Optimal Motion and Structure Estimation", IEEE Trans. PAMI, vol. 15, no.9, pp.864-884, Sept. 1993.

*Figure 1: Tracking of Rigid feature Points*

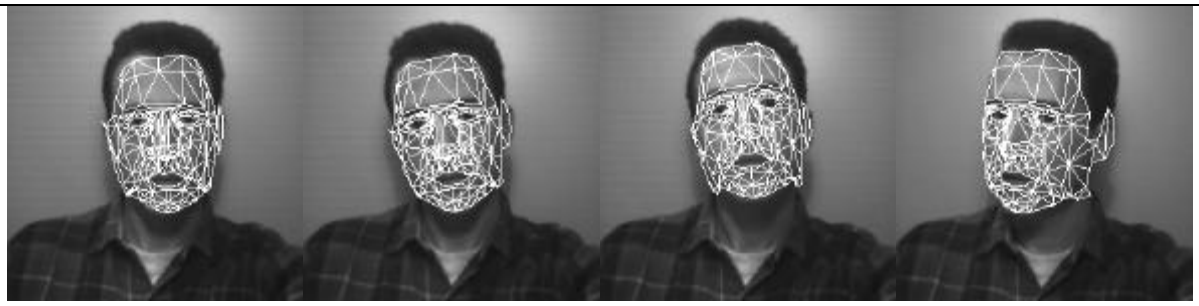

*Figure 2: Adaptation of 3D face model*

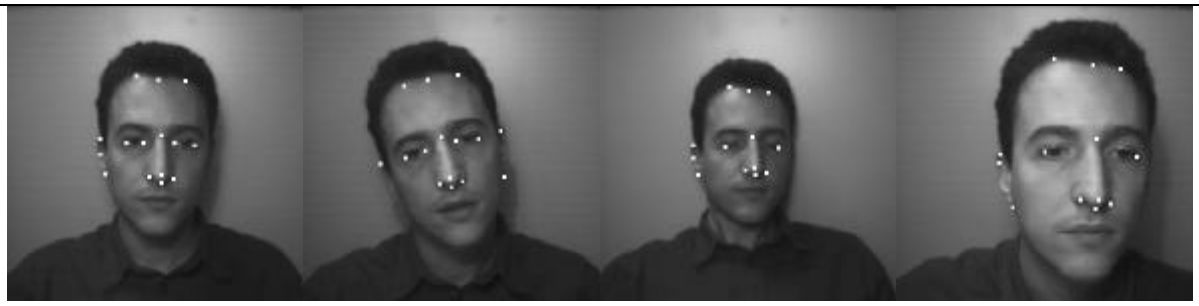

*Figure 3: Tracking of Rigid feature Points*

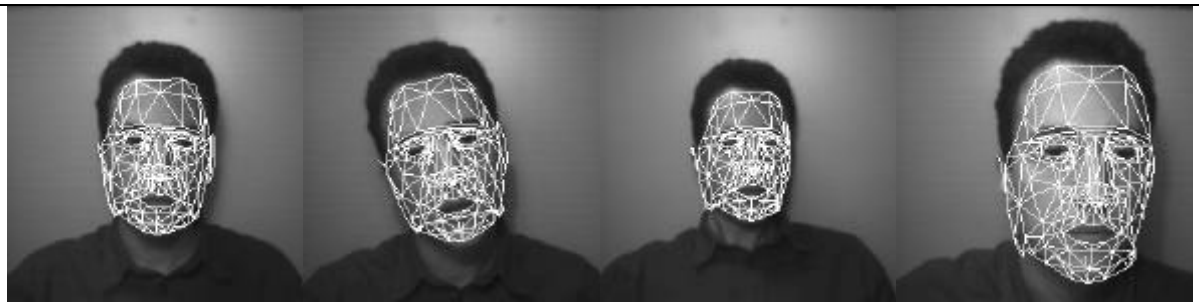

*Figure 4: Adaptation of 3D face model*