# REAL-TIME FACIAL FEATURE TRACKING FROM 2D+3D VIDEO STREAMS

*Filareti Tsalakanidou, Sotiris Malassiotis*

Informatics and Telematics Institute, Centre for Research and Technology Hellas
6th km Charilaou-Thermi Road, Thessaloniki 57001, Greece
filareti@iti.gr, malasiot@iti.gr

## ABSTRACT

This paper presents a completely automated 3D facial feature tracking system using 2D+3D image sequences recorded by a real-time 3D sensor. It is based on local feature detectors constrained by a 3D shape model, using techniques that make it robust under pose and partial occlusion. Several experiments conducted under relatively non-controlled conditions demonstrate the accuracy and robustness of the approach.

***Index Terms*** **—** Face recognition, tracking, feature extraction, geometric modeling

## 1. INTRODUCTION

The problem of detecting and tracking facial features from images and image sequences is important for a large range of applications including facial expression recognition and facial motion capture for expressive character production. Most techniques proposed in the literature use 2D image sequences [1, 2]. The problem with such techniques is that they are prone to illumination and pose variation changes that affect the perceived geometry of the face. Moreover, subtle skin deformations that characterize facial muscle movements such as wrinkles, furrows, bulges, etc are difficult to capture by a 2D camera due to problems caused by illumination, shadows, projection, etc. Some techniques try to alleviate these problems using deformable 3D face models [3] or multiple views [4] sacrificing however real-time performance.

In this paper, we propose using 2D+3D (brightness + depth) image sequences captured in real-time. Recent technological advances have made real time recording of good quality 3D data possible [5, 6]. Although the advantages of using 3D facial images are self evident, only few works have examined facial feature tracking from 3D sequences. In [7] dense deformable face models are used that are computationally expensive and require manual initialization. More similar to our approach is the work of Liebelt et al. [8] that also uses 2D+3D image sequences. However, our approach relies on local features only, while [8] relies largely on texture information since it employs Active Appearance Models.

In this paper, we employ a model-based feature tracker applied to sequences of 3D range images and corresponding grayscale images recorded by a novel real-time 3D sensor. To achieve real-time performance we use feature based 3D pose estimation followed by iterative tracking of 81 facial points using local appearance and surface geometry information. Special trackers are developed for important facial features such as the mouth and the eyebrows that account for the non-linearity in the movement of these features. The efficiency of the 3D face tracker is evaluated in a database with many subjects and sequences and promising results are obtained.

The paper is organized as follows. The face tracker is described in Section 2. Local detectors for the mouth and eyebrows are presented in Section 3. The performance of the face tracker is evaluated in Section 4. Finally, Section 5 concludes the paper.

## 2. FACE AND FACIAL FEATURE TRACKING

Our 3D face tracker extends the well-known Active Shape Model (ASM) technique [9] to handle 3D data and also cope with measurement uncertainty and missing data caused by occlusions and sensor errors. The ASM is a point distribution model (PDM) accompanied by a local appearance pattern for every point, which effectively models the shape of a class of objects, faces in our case. Point and local appearance distributions are obtained using a set of annotated training images. Any shape can then be expressed as the sum of a mean shape and a linear combination of basis shapes computed during training. Although ASMs have been demonstrated less accurate than Active Appearance Models (AAM), they have the advantage of robustness to illumination variations (using local gradient search) and are very efficient.

In each time instant we capture a grayscale image and a depth image. Pixel values of depth images represent the distance of the corresponding point from the camera plane. Using the one-to-one pixel correspondence of depth and grayscale images as well as camera projection parameters, we can directly associate every image point with its 3D coordinates and a texture value.

The shape $\mathbf{s}$ of the face is represented as a sequence of $n$=81 points corresponding to salient facial features (see Fig.1). The PDM is then expressed as

$$\mathbf{s} = \tilde{\mathbf{s}} + \sum_{i=1}^{m} a_i \mathbf{s}_i = \tilde{\mathbf{s}} + \mathbf{a} \cdot \mathbf{S} \tag{1}$$

where $\mathbf{s} = \{x_1, y_1, z_1, ..., x_n, y_n, z_n\}$ is the vector of $n$ landmark coordinates, $\mathbf{s}_i$ are the basis shapes computed by applying Principal Component Analysis to a set of manually annotated training examples, which are aligned to a common coordinate 3D frame (called model coordinate frame), $\tilde{\mathbf{s}}$ is the mean shape computed in the same space and $\mathbf{a}$ is a vector of shape parameters.

The local appearance model for each landmark $L_i$ is computed from image gradient information gathered in all 2D training images along the projection of a line that passes from $L_i$ and is

This work was supported by the EU FP6 project "PASION: Psychologically Augmented Social Interaction over Networks" (IST-027654).

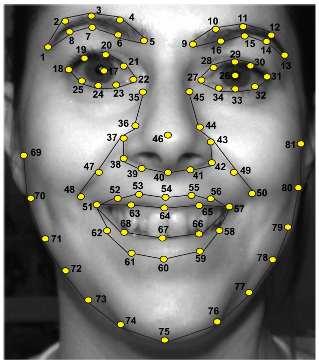

Figure 1. The 81 landmarks $L_i$ and corresponding segments of the ASM.

perpendicular to the facial contour that $L_i$ belongs to (e.g. eyebrow, mouth, etc). A set of shape contours is defined in terms of connectivity information between landmarks as illustrated in Fig. 1. After computing the gradient profiles of $L_i$ in all training images, we can build a local model of gradient changes associated with this landmark assuming a unimodal Gaussian distribution. The same procedure is applied for every landmark thus obtaining $n$ local appearance models.

Using Eq. 1, we can represent the shape of any face in the model coordinate frame. To express the same shape in the real-world coordinate frame we use

$$\mathbf{x} = \mathbf{R} \cdot \mathbf{s} + \mathbf{T} = \mathbf{R} \cdot (\tilde{\mathbf{s}} + \mathbf{a} \cdot \mathbf{S}) + \mathbf{T} \qquad (2)$$

where $\mathbf{R}$ is the 3D rotation matrix and $\mathbf{T}$ the 3D translation vector that rigidly align the model coordinate frame with the real-world coordinate frame and $\mathbf{x}$ represents the landmark coordinates in the real-world coordinate frame. By projecting $\mathbf{x}$ in the image plane, we obtain the corresponding 2D shape $\mathbf{v} = P(\mathbf{x})$, where $P$ represents a camera projection function that models the imaging process. $\mathbf{v}$ represents the landmark positions in the 2D image.

To estimate the landmark positions in a new pair of 2D and 3D images the following steps are taken:

1. Let $\mathbf{R}$ be the 3D rotation matrix and $\mathbf{T}$ the 3D translation vector that rigidly align the model with the face. A first estimate of these is obtained using the 3D face detection and pose estimation technique proposed in [10]. The shape parameters $\mathbf{a}$ are initialized to zero, i.e. we start from the mean face shape $\tilde{\mathbf{s}}$.

2. The current shape $\mathbf{s}$ is transformed to the real-world coordinate frame using the rigid transformation ($\mathbf{R}$, $\mathbf{T}$) and is subsequently projected on the 2D camera plane through $P$. A local search is then performed around each projected landmark position to find the point that best matches the local appearance model. To do this, first we compute the normal vector at the specific location. Then, we define a set of candidate pixels along this line and compute a local gradient vector for each of them exactly as in the case of training images. Similarity between extracted gradient profiles and the corresponding local appearance model is measured using the Mahalanobis distance. The point associated with the lowest distance is selected. The same procedure is applied for all landmarks and a set of new landmark positions is estimated in the 2D image. These are subsequently back-projected in the 3D space using the inverse projection function $P^{-1}$ and the $z$ values of the corresponding pixels of the depth image. Thus a new 3D shape $\mathbf{x}$ is defined in the real-world coordinate frame. Moreover, each landmark is associated with a weight set to be the reciprocal of the computed Mahalanobis distance. In case the $z$ value of a point is undefined, the median depth value in the neighborhood of this pixel is used. If no depth is defined in the greater area of this pixel, then a zero weight is assigned to this landmark, so that it is neglected in model estimation.

3. A new rigid transformation $(\mathbf{R}, \mathbf{T})$ aligning the new shape $\mathbf{x}$ with the current template $\mathbf{s}$ is estimated using Horn's quaternion method [11]. A new rectified shape $\mathbf{y} = \mathbf{R}^{-1} \cdot (\mathbf{x} - \mathbf{T})$ is computed in the model coordinate frame.

4. A new set of parameters $\mathbf{a}$ is estimated by minimizing $\|\tilde{\mathbf{y}} - \tilde{\mathbf{s}} - \mathbf{a} \cdot \mathbf{S}\|^2 + \lambda \cdot \|\mathbf{a}\|^2$, where the second term is a regularization constraint. A robust least squares approach using Huber's scheme is adopted. We also exclude points that may be occluded, for example points on the side of the face or nose, which may be easily determined using the estimated face orientation. Once a new set of parameters $\mathbf{a}$ is estimated, a new shape $\mathbf{s}$ is synthesized using Eq. 1.

5. Steps 2-5 are repeated until convergence of the fitting error $e = \|\mathbf{y} - \mathbf{s}\|$ or until a number of iterations is reached. Then a new real-world shape $\mathbf{x}$ is computed using Eq. 2.

For each subsequent frame, initialization is performed based on the previous frame, i.e. we start from step 2 using $\mathbf{R}$, $\mathbf{T}$ and $\mathbf{s}$ estimated in the previous frame. If the model has not converged, we re-initialize the tracker, i.e. we start from step 1 and repeat face detection, pose estimation and model fitting. For faster convergence we use a multi-resolution scheme with three layers.

## 3. LOCAL FEATURE DETECTORS

The proposed tracker achieves small localization errors per landmark, however there are cases where localization of individual features such as the eyebrows and the mouth is not accurate enough. This is due to the inadequacy of the linearity assumption in the PDM, but also due to the uni-modal distribution chosen for local appearance variations (e.g. appearance of teeth when opening the mouth). Instead of resorting to non-linear modeling techniques, we propose a set of dedicated local facial feature detectors.

### 3.1. Local eyebrows detector

A local 3D ASM with 16 landmarks corresponding to the eyebrows boundaries is used, which is initialized using the eyebrows estimation provided by the global tracker. Furthermore, we adopt a different local appearance model as follows.

For each candidate landmark position, we compute the average intensities $S_1$ and $S_2$ above and below this feature point inside small rectangular boxes aligned with the eyebrow contour. Our goal is to find the point maximizing $S_1 - S_2$, i.e. the contrast between bright and dark areas (skin and eyebrow). In addition, we ask that $S_1 - S_2 > T_1$ and $S_2 < T_2$. The first condition implies that the landmark point should lie in an area of adequate gradient change. The second is used to overcome the problem of shadows, which results in selecting points lying in the border of shadowed and non-shadowed skin areas instead of lying in the border of eyebrow and skin areas. $T_1$ and $T_2$ are experimentally chosen from training images.

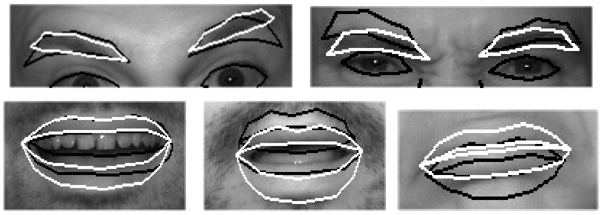

Figure 2. Examples of eyebrow and lip boundary localization using the global 3D ASM model (black line) and local detectors (white line).

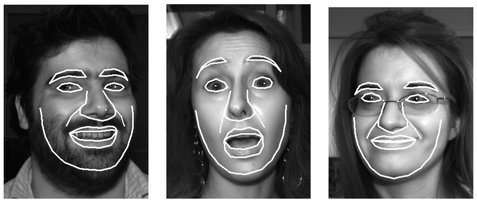

Figure 3. Examples of facial feature tracking results using the proposed global tracker and local feature detectors.

Since a bad initial estimation may prevent the local model from converging, we perform several local fittings with slightly perturbed initial positions and choose the one minimizing the fit error. The proposed local eyebrow detector enhances significantly the estimation provided by the global ASM especially in cases where the eyebrows are raised or lowered (see Fig. 2).

### 3.2. Local mouth detector

Lip boundary localization is also problematic due to the multi-modal nature of local mouth appearance in the inner lip boundaries, since their local gradient patterns are significantly affected by whether the mouth is open or closed. The problem is more severe when the mouth is open and the teeth are visible, since in this case the boundary between the teeth and the dark area of the mouth cavity is erroneously recognized as a lip boundary.

To overcome this problem, we propose a two-step approach for localizing lip boundaries. First, a two-class Support Vector Machine classifier with an RBF kernel is used to decide whether the mouth is open or closed. This classifier is trained using 16-dimensional feature vectors of local gradient measurements in the area of the mouth, obtained from 240 labelled images. After the mouth is classified as open or closed, an open or closed mouth local 3D ASM is fitted on the face to localize the position of outer and inner lip boundaries. Model fitting is based on image gradient profiles. However, we do not only consider points along the normal but also points in a narrow zone aligned with the normal. Examples of improved mouth localization are shown in Fig. 2.

### 3.3. Combining global and local feature position estimates

To incorporate the information provided by the local feature detectors into the global model, the fitting algorithm presented in Section 2 is modified as follows: after step 2, the parts of shape $\mathbf{x}$ corresponding to eyebrows and mouth are replaced with the improved estimates. Then we continue with step 3. Using the proposed 2D+3D ASM and dedicated local detectors very good localization accuracy may be achieved even under moderate face poses as can be seen in Fig. 3.

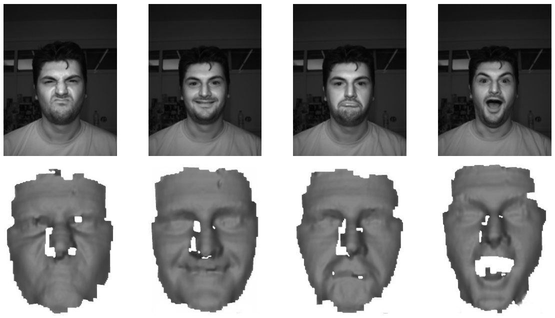

Figure 4. Examples of grayscale images and corresponding 3D models of the facial expression database. The latter are generated from the recorded 3D images.

## 4. EXPERIMENTAL RESULTS

Our experiments were conducted on a 2D+3D image database (800 sequences of 52 participants) that was recorded using a prototype 3D sensor [6] and for the aim of automatic facial action coding [12]. Therefore each sequence depicts the subject mimicking a facial action or facial expression. The resolution of captured images is 582×782 pixels, while the accuracy of depth data is better than 0.3*mm* for objects standing at a mean distance of 60*cm* (see Fig. 4). The duration of each sequence is about 30*s* with a framerate of about 5 fps.

To train the global 3D shape model as well as the local detectors we used a set of 400 image pairs depicting an action unit or facial expression at its peak. To test the face tracker, we use another set of 600 images, where we manually mark the positions of facial landmarks. The estimated feature positions are compared against their ground-truth positions. Using the proposed face tracker, we achieve a mean localization error of 5.35 pixels and standard deviation 2.2, when the mean face dimensions are 280×370 pixels. On the contrary using the global detector only, the corresponding error is 7.8 pixels. We also compare the 2D+3D tracker against a 2D only ASM with the same 81 landmarks. In this case, we obtain a localization error of 10.2 pixels, which is mainly attributed to erroneous estimation of open mouth landmarks.

In Fig. 5, we plot the localization error of the algorithm for selected landmarks $L_i$ (see Fig. 1) in sequences depicting raising of eyebrows and surprise. Estimated feature positions are compared against their actual positions, which were manually defined in sequence frames. We also plot the movement of facial landmarks with respect to their position in the first frame.

Unlike similar 2D techniques, the proposed face tracker achieves good performance even if there is a lot of head movement as demonstrated by the following experiment. We have recorded an image sequence showing a human subject smiling and laughing while rotating her head up to 30° to the left and right (see Fig. 6). In Fig. 7 the tracking errors for selected facial landmarks in this sequence are shown. In case of larger poses, where half of the face is occluded, the tracking error increases significantly.

The algorithm runs on 5 frames per second on an Intel Core Duo 2.0 GHz PC with 4GB RAM.

## 5. CONCLUSION

In this paper, a novel real-time face tracker based on active shape models and a set of special local detectors for the eyebrows and

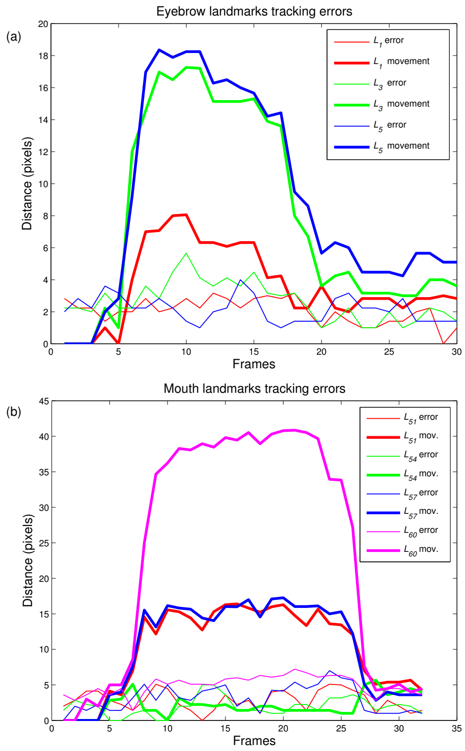

Figure 5. Tracking errors of eyebrow and mouth landmarks $L_i$ (see Fig. 1) in sequences depicting a) raising of eyebrows and b) surprise (jaw dropped + eyebrows raised). Movements of facial landmarks with respect to their positions in the first frame are also shown.

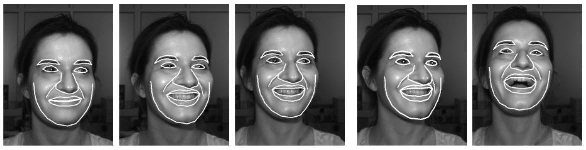

Figure 6. Example of image sequence showing a happy expression under pose variations. White lines correspond to tracking results.

the mouth were presented. The proposed techniques effectively combine 3D face geometry and 2D appearance data to achieve increased accuracy and robustness under facial expressions and pose variations as demonstrated by experiments conducted in a face database with many subjects and different expressions.

This is our first step towards real-time 3D dense motion capture of the face. Thus future work will focus on capturing non-rigid 3D facial deformation in texture-less areas such as the cheeks as well as making the algorithms robust to large head poses where half of the face may be occluded.

## 6. REFERENCES

[1] F. Dornaika and F. Davoine, "Simultaneous facial action tracking and expression recognition in the presence of head motion," *International Journal of Computer Vision*, vol. 76, no. 3, pp. 257–281, March 2008.

[2] L. Gu and T. Kanade, "A generative shape regularization model for robust face alignment," in *Proc. 10th European Conf. on Computer Vision*, Oct. 2008, pp. 413–426.

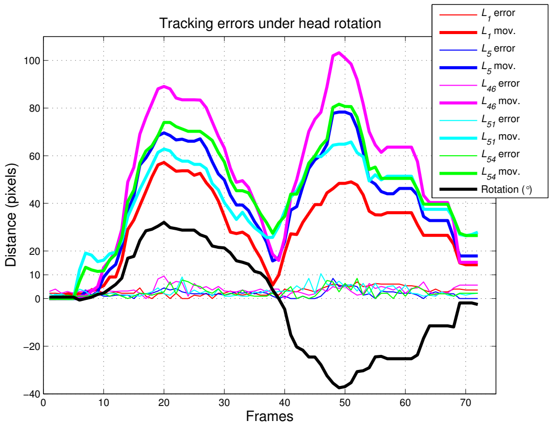

Figure 7. Tracking errors of selected facial landmarks $L_i$ in the sequence shown in Fig. 6. The bold black line represents head rotation around the vertical axis $y$ (positive/negative values correspond to rotations to the right/left (30° to -35° )).

[3] S. B. Gokturk, J.-Y. Bouguet, C. Tomasi, and B. Girod, "Model-based face tracking for view-independent facial expression recognition," in *Proc. 5th IEEE Conf. on Automatic Face and Gesture Recognition*, May 2002, pp. 287–293.

[4] S. Von Duhn, L. Yin, M. J. Ko, and T. Hung, "Multiple-view face tracking for modeling and analysis based on non-cooperative video imagery," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[5] S. Zhang and P. Huang, "High-resolution, real-time 3D shape acquisition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop*, June 2004, vol. 3, p. 28.

[6] D. Modrow, C. Laloni, G. Doemens, and G. Rigoll, "A novel sensor system for 3D face scanning based on infrared coded light," in *Proc. SPIE Conf. on Three-Dimensional Image Capture and Applications*, Jan. 2008, vol. 6805.

[7] X. Huang, S. Zhang, Y. Wang, D. Metaxas, and D. Samaras, "A hierarchical framework for high resolution facial expression tracking," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop*, June 2004, vol. 1, p. 22.

[8] J. Liebelt, J. Xiao, and J. Yang, "Robust AAM fitting by fusion of images and disparity data," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, June 2006, vol. 2, pp. 2483 – 2490.

[9] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 743–756, July 1997.

[10] S. Malassiotis and M. G. Strintzis, "Robust real-time 3D head pose estimation from range data," *Pattern Recognition*, vol. 38, no. 8, pp. 1153–1165, Aug. 2005.

[11] B. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society of America A*, vol. 4, no. 4, pp. 629–642, April 1987.

[12] F. Tsalakanidou and S. Malassiotis, "Robust facial action recognition from real-time 3D streams," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, June 2009, pp. 4–11.