Fully Automatic, Real-Time Detection of Facial Gestures from Generic Video

Marco La Cascia and Lorenzo Valenti
Dipartimento di Ingegneria Informatica
Università di Palermo
Palermo, Italy

Stan Sclaroff
Computer Science Department
Boston University
Boston, MA, USA

Abstract— **A technique for detection of facial gestures from low resolution video sequences is presented. The technique builds upon the automatic 3D head tracker formulation of [11]. The tracker is based on registration of a texture-mapped cylindrical model. Facial gesture analysis is performed in the texture map by assuming that the residual registration error can be modeled as a linear combination of facial motion templates. Two formulations are proposed and tested. In one formulation head and facial motion are estimated in a single, combined linear system. In the other formulation, head motion and then facial motion are estimated in a two-step process. The two-step approach yields significantly better accuracy in facial gesture analysis. The system is demonstrated in detecting two types of facial gestures: "mouth opening" and "eyebrows raising." On a dataset with lots of head motion the two-step algorithm achieved a recognition accuracy of 70% for the "mouth opening" and accuracy of 66% for "eyebrows raising" gestures. The algorithm can reliably track and classify facial gestures without any user intervention and runs in real-time.**

Keywords—facial gesture, head tracking

I. Introduction

Facial gesture and expression analysis is a very important task in different fields ranging form behavioral science to human-computer interaction. In recent years computer vision systems have been proposed to automate this task to some extent (for a survey, see [16, 9]).

Several techniques can classify images or sequences as belonging to one of a number of typical facial expressions representing some emotional state such as joy, sadness, anger and so on. To achieve this classification a variety of classifiers and facial motion features have been used. For example, optical flow is computed in [15, 21, 7, 4] from the image sequence and then used to detect the presence of expressions. Other techniques are based on facial motion features derived from flexible shapes [14], local parameterized models [2], active appearance models [6], tracking of pupils [10] or other feature points.

Other techniques are able to estimate facial motion features in terms of FACS [8] action units (AU). For example the work reported in [1, 12, 10, 13, 19] is an attempt to estimate one or more AUs activated separately or together. These techniques, to achieve useful results, usually require that sequences are acquired under particular conditions. An interesting comparison of facial motion feature extraction techniques used for detection of AUs is reported in [5].

To our knowledge only the techniques proposed in [10, 13] can work on sequences representing facial expressions under general conditions; however in [10] the use of an infrared camera with an infrared LED is required and in [13] a lot of user intervention for the initialization of the system is needed. The technique we propose is able to continuously extract facial motion features from generic low resolution (320x240) video sequences in a fully automatic fashion. Moreover, its low computational load allows for real-time operation (about 15 Hz) on a standard PC equipped with a USB camera. The only assumption is that the first frame of the sequence to analyze is a frontal view with a neutral facial expression.

The paper is organized as follows. Section 2 summarizes the technique for 3D head tracking. Section 3 describes the proposed approach for facial motion feature extraction. Section 4 describes the experimental evaluation. Finally Section 5 concludes the paper.

II. Head Tracking

To cope with head motion in the image sequence some sort of head tracking or image stabilization is needed. For example in [3] a 3D wireframe model of the face is employed. The user interactively selects facial features such as eye corners or mouth corners; once the model is fitted the tracking is performed as in [18]. Similarly in [7] a detailed 3D model of the face is used. In [13] head tracking is achieved using a cylindrical model [20] and user interaction is limited to marking several feature points in the first frame of the image sequence. In other cases a simple planar tracker [2] is used to cope with small in plane and out of plane rotation of the head.

In the proposed system, to stabilize the input sequence, we used a tracker based on registration of a texture-mapped cylindrical model [11]. We chose this tracker among others as it is able to stabilize the head in the presence of in plane and out plane rotation. Moreover, the tracker is completely automatic and runs in real time on a PC. The head is modeled as a texture mapped cylinder and tracking is formulated as an image registration problem in the cylinder's texture map image (Fig. 1). Incoming image frames I are mapped onto the cylinder texture map, according to the current 3D position and orientation of the cylinder $\mathbf{a}=[x\ y\ z\ \alpha\ \beta\ \gamma]^T$. We can write the

0-7803-8578-0/04/$20.00 ©2004 IEEE

texture image T corresponding to the image frame I as $T=\Gamma(I, a)$ where $\Gamma(I, a)$ is the function mapping an image I on the texture plane through a cylinder with position and orientation parameters a. T is then registered to the texture map T_0 corresponding to the initial frame I_0 and to the initial parameter a_0 of the model. To solve the registration problem, the residual error of registration $R=T-T_0$ (Fig. 2) is modeled as a linear combination of head motion templates b_i.

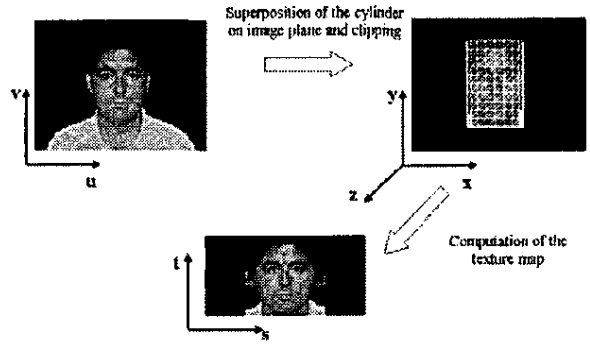

Figure 1. Cylindrical model and texture map

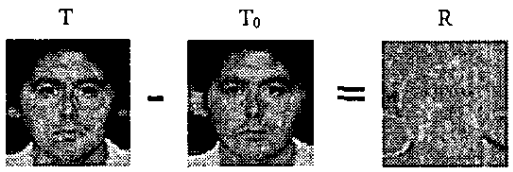

Figure 2. Residual error of registration

Head motion templates b_i are computed at the first frame of the sequence as the residual error of registration corresponding to small displacements n_i of the model along the six degrees of freedom: $b_i=T_0-\Gamma(I_0, a_0-n_i)$. Fig. 3 shows an example of head motion templates corresponding to a small displacement ±d and ±2d of the head position.

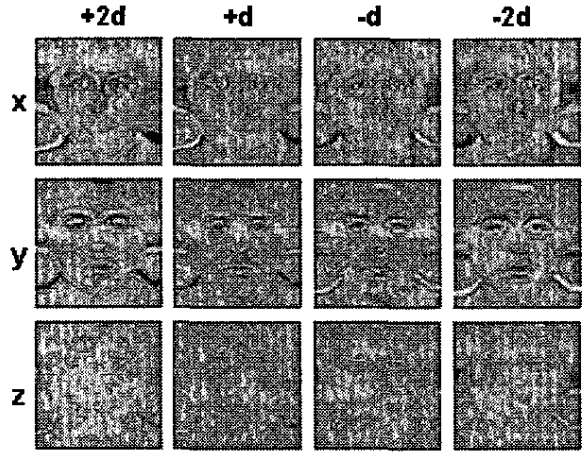

Figure 3. Example head motion templates

Fast and stable on-line tracking is then achieved via weighted least squares minimization of the registration error. The least square solution of the term $W(T-T_0)\approx Bx$ is computed as $x=[B^TW^TWB]^{-1}\ B^TW^TWR$ where the matrix $B=[b_1\ b_2...b_k]$ and W is a diagonal matrix having on the diagonal the elements of the binary image that masks the face region in the input image (Fig. 9(a)). Finally the increment in the cylinder parameters can be written as $\Delta a=N_a\ x$ where $N_a=[n_1\ n_2...n_k]$. Note that, in the above formulation, the texture images T, T_0, R, and the templates b_i are represented as column vectors containing the pixels in raster scan order. .

The complete formulation of the tracker also includes lighting correction via illumination templates and a regularization term. For further details see [11].

III. Facial Gesture Extraction

To achieve automatic, user-independent and real-time detection of facial gestures, we extended the formulation of the head tracker in [11]. In what follows two different extensions of the head tracker are presented. A comparison of results obtained is then reported. The system is tested in detecting two types of facial gestures: "mouth opening" and "eyebrows raising."

A. Integrated tracking and analysis

As described above, the head tracker uses the registration error resulting from head motion to obtain an estimate of the head's rigid motion (3D position and orientation). We could also use the registration error resulting from the mouth opening, or eyebrows raising to detect such facial gestures.

To extend the formulation it is sufficient to add a series of facial expression templates corresponding to different magnitudes of these gestures. Templates can be computed off-line from a set of registered texture maps using for example PCA [5] or, similarly, Gaussian blurring the texture maps corresponding to different magnitudes of a typical instance of the expression (Fig. 4) after subtracting the average image. Examples of the templates for the "eyebrows raising" gesture, computed via PCA and via Gaussian blurring are shown in Figs. 5 and 6, respectively.

This approach is extremely fast as it only adds a few multiplications to the basic head tracker formulation. The extended formulation is straightforward. We can consider the eyebrows raising templates of Fig. 6 as vectors e_1, e_2, e_3, and e_4, and the similarly computed mouth opening template as vectors m_1, m_2, m_3, and m_4. We can assume valid the approximation $W(T-T_0) \approx (T_m\ x_m + T_e\ x_e + B\ x)$ where $T_m = [m_1\ m_2\ m_3\ m_4]$ and $T_e = [e_1\ e_2\ e_3\ e_4]$.

If we define $M = [T_m\ T_e\ B]$ and $c = [x_m\ x_e\ x]^T$ the weighted least square minimization of the residual leads to the solution $c=[M^TW^TWM]^{-1}\ M^TW^TW\ R$. The subvector x gives the new 3D position and orientation of the cylinder, while the subvectors x_m and x_e are a measure respectively of the mouth opening and eyebrows raising.

How these vectors are related to the corresponding FACS AUs remains a topic for future investigation. However, analysis on tens of sequences showed that quantitative

information about "mouth opening" and "eyebrows raising" gestures is clearly conveyed by the vectors $\mathbf{x}_m$ and $\mathbf{x}_e$. Figs. 7 and 8 show graphs of the magnitude of $\mathbf{x}_m$ and $\mathbf{x}_e$ for two example sequences in our test set. Both facial gestures are clearly identified. This behavior was consistent across tens of test sequences as is reported in Section 4.

Figure 4. Frames from "eyebrows raising" sequence.

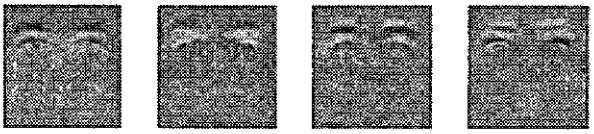

Figure 5. First four components of eyebrows raising.

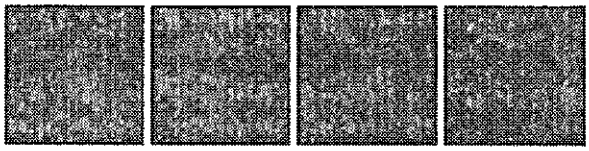

Figure 6. Eyebrows raising templates computed by Gaussian blurring.

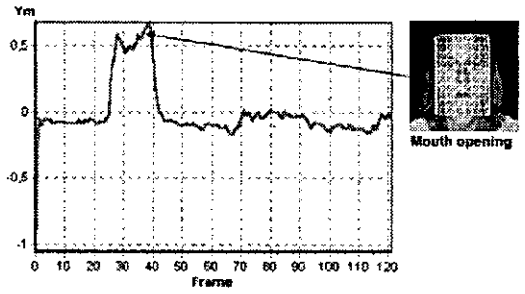

Figure 7. Results for a "mouth opening" gesture.

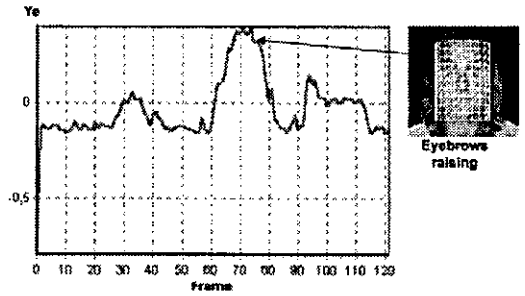

Figure 8. Results for an "eyebrows raising" gesture.

Unfortunately, while the results were consistently good for sequences showing a very small head motion, results were not satisfactory for sequences where head motion and facial gestures occur at the same time. This is easily explained by the strong coupling that exists between the head motion templates and facial motion templates. This coupling makes the matrix $\mathbf{M}^T\mathbf{W}^T\mathbf{W}\mathbf{M}$ very close to singular and can lead to instability in tracking. In what follows a variant of the algorithm is presented that can overcome this problem.

B. Expression analysis in texture maps

To decouple the system we analyze the residual error in two separate steps. During the first step the residual $\mathbf{R}$ is modeled in terms of the head motion templates $\mathbf{B}$ and the new value of the cylinder's parameter vector $\mathbf{a}^+$ is estimated as in Section 2. In the second step the incoming image $\mathbf{I}$ is registered according to the estimated parameter $\mathbf{a}^+$ and the new residual $\mathbf{R}^+ = \Gamma(\mathbf{I}, \mathbf{a}^+)\text{-}\mathbf{T}_0$ is modeled in terms of the facial motion templates $\mathbf{T}_m$ and $\mathbf{T}_e$.

Again, the approximation $\mathbf{W}(\mathbf{R}^+) \approx (\mathbf{T}_m\ \mathbf{x}_m + \mathbf{T}_e\ \mathbf{x}_e)$ is assumed to be valid. Solving by least squares we obtain $[\mathbf{x}_m\ \mathbf{x}_e]^T = [\mathbf{F}^T\mathbf{W}^T\mathbf{W}\mathbf{F}]^{-1}\ \mathbf{F}^T\mathbf{W}^T\mathbf{W}\ \mathbf{R}^+$ where $\mathbf{F} = [\mathbf{T}_m\ \mathbf{T}_e]$.

Another advantage of this approach is that it is possible to avoid the disturbance deriving from the residual error in regions of the texture map other than the eyes and mouth. Our experiments confirm that using a mask $\mathbf{W}$ as shown in Fig. 9(b) instead of the one in Fig. 9(a) yields improved classification of the facial gestures considered.

(a) (b)

Figure 9. Masks used for head tracking (a) and facial gesture analysis (b).

Finally, we note that this approach to decoupling the facial motion estimation can be exploited further. For example, it would be straightforward to subdivide the residual error registration into two steps, one using a mask that includes only the region around the eyes and the other only the region around the mouth. A similar approach has been used in [17] to improve face recognition.

IV. Experimental Results

To evaluate the proposed techniques we collected 50 sequences representing six subjects performing different facial gestures and head movements. Sequences were captured with a low-cost USB webcam at 15 fps and 320x240 resolution. The sequence durations range from 50 to 300 frames (about 3 to 20 secs.).

The sequences include "mouth opening" and "eyebrows raising" gestures with different amounts of head motion (translations, as well as in plane and out of plane rotations). We considered three kinds of sequences: A) sequences with very small head motion, B) sequences with significant head motion and facial gestures occurring while head is almost

steady, C) sequences with significant head motion and facial gestures occurring simultaneously.

Tables 1 and 2 report the gesture detection results for these test sets. A facial gesture was considered detected whenever the magnitude of $\mathbf{x}_m$ or $\mathbf{x}_e$ exceeded a fixed threshold that was the same for all sequences. As can be seen, the two step texture map analysis algorithm achieved consistently better detection of the gestures tested. The improved performance is particularly pronounced for the most challenging sequences C) that show facial gestures occurring simultaneously with significant head motion.

TABLE I. RESULTS FOR "MOUTH OPENING" DETECTION.

Algorithm	Sequence Group	Correct detected	Misses	False Positive
Integrated approach	A	72%	28%	0%
	B	46%	34%	20%
	C	27%	42%	21%
Texture map analysis	A	88%	12%	0%
	B	72%	16%	12%
	C	50%	35%	15%

TABLE II. RESULTS FOR "EYEBROWS RAISING" DETECTION.

Algorithm	Sequence Group	Correct detected	Misses	False Positive
Integrated approach	A	70%	30%	0%
	B	43%	36%	21%
	C	25%	47%	28%
Texture map analysis	A	86%	14%	0%
	B	70%	18%	12%
	C	42%	33%	22%

V. CONCLUSIONS

The head tracker of [11] was extended to enable extraction of facial gestures from generic, low resolution video sequences in real time and without any user intervention. Testing on a challenging dataset shows that the approach provides a level of accuracy that is acceptable for many applications.

Nevertheless, our technique can still be extended or improved on several fronts. For example we believe that a larger set of facial motion features could be extracted by simply using a larger set of facial gesture templates. Moreover these facial motion features should be related to FACS AUs. In the future we also plan to extend the formulation to a two camera system to improve the performance of the tracker.

ACKNOWLEDGMENTS

Stan Sclaroff's participation in this project was supported in part through U.S. National Science Foundation Grants IIS 0208876 and IIS 9912573.

REFERENCES

[1] M.S. Bartlett et al., "Measuring facial expressions by computer image analysis", Psychophysiology, No. 36, 1999.

[2] M.J. Black and Y. Yacoob, "Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion", Int. Journal Computer Vision, Vol. 25, No. 1, pp. 23-48, 1997.

[3] I. Cohen et al., "Facial expression recognition from video sequences: temporal and static modeling", Computer Vision and Image Understanding, Vol. 91, pp. 160-187, 2003.

[4] J.F. Cohn et al., "Feature-point tracking by optical flow discriminates subtle differences in facial expression", Proc. of IEEE Conf. on Automatic Face and Gesture Recognition, 1998.

[5] G. Donato et al., "Classifying Facial Actions", IEEE Trans. on PAMI, Vol. 21, No. 10, pp. 974-989, 1999.

[6] F. Dornaika and J. Ahlberg, "Effective Active Appearance Model for Real-Time Head and Facial Feature Tracking", Proc. of IEEE Workshop on Analysis and Modeling of Faces and Gestures, 2003.

[7] I. Essa and A. Pentland, "Coding, Analysis, Interpretation, and Recognition of Facial Expressions", IEEE Trans. on PAMI, Vol. 19, No. 7, pp. 757-763, 1997.

[8] P. Ekman and W.V. Friesen, "Facial Action Coding System (FACS)", Consulting Psychologists Press, Palo Alto, 1978.

[9] B. Fasel and J. Luettin, "Automatic Facial Expression Analysis: A Survey", Pattern Recognition, Vol. 36, No. 1, 2003.

[10] A. Kapoor et al., "Fully Automatic Upper Facial Action Recognition", Proc. of IEEE Workshop on Analysis and Modeling of Faces and Gestures, 2003.

[11] M. La Cascia et al., "Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Robust Registration of Texture-Mapped 3D Models", IEEE Trans. on PAMI, Vol. 22, No. 4, pp. 322-336, 2000.

[12] J. Lien, "Automatic recognition of facial expressions using hidden markov models and estimation of expression intensity", Ph.D. Thesis, Carnegie Mellon University, 1998.

[13] T. Moriyama et al., "Automatic Recognition of Eye Blinking in Spontaneously Occurring Behavior", Proc. of Int. Conf. on Pattern Recognition, 2002.

[14] A. Lanitis et al., "Automatic Interpretation and Coding of Face Images Using Flexible Models", IEEE Trans. on PAMI, Vol. 19, No. 7, pp. 743-756, 1997.

[15] K. Mase, "Recognition of Facial Expresions from Optical Flow", IEICE Trans. E74, No. 10, pp. 3474-3483, 1991.

[16] M. Pantic and L.J.M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art", IEEE Trans. on PAMI, Vol. 22, No. 12, pp. 1424-1445, 2000.

[17] A. Pentland et al., "View-based and modular eigenspaces for face recognition", Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, 1994.

[18] H. Tao and T.S. Huang, "Connected vibrations: a modal analysis approach to non-rigid motion tracking", Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, 1998.

[19] Y.-li Tian et al., "Recognizing Action Units for Facial Expression Analysis", IEEE Trans. on PAMI, Vol. 23, No. 2, pp. 1-19, 2001.

[20] J. Xiao et al., "Robust full motion recovery of head by dynamic templates and re-registration techniques", Proc. of IEEE Conf. on Automatic Face and Gesture Recognition, 2002.

[21] Y. Yacoob and L.S. Davis, "Recognizing human facial expressions from long image sequences using optical flow", IEEE Trans. on PAMI, Vol. 18, No. 6, pp. 636-642, 1996.