Workshop on Perceptual User Interfaces, pp. 94-97, Banff, Alberta, Canada, October 19-21, 1997.

Automatically Recognizing Facial Expressions in the Spatio-Temporal Domain

[1,2]James J. Lien, [1]Takeo Kanade, [3]Adena J. Zlochower, [3]Jeffrey F. Cohn, and [2]Ching-Chung Li

[1]VASC, The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213
[2]Dept. of Electrical Engineering, University of Pittsburgh, Pittsburgh, PA 15260
[3]Dept. of Psychology, University of Pittsburgh, Pittsburgh, PA 15260

jjlien@cs.cmu.edu

Abstract

We developed a computer vision system that automatically recognizes facial action units (AUs) or AU combinations using Hidden Markov Models (HMMs). AUs are defined as visually discriminable muscle movements. The facial expressions are recognized in digitized image sequences of arbitrary length. In this paper, we use two approaches to extract the expression information: (1) facial feature point tracking, which is sensitive to subtle feature motion, in the mouth region, and (2) pixel-wise flow tracking, which includes more motion information, in the forehead and brow regions. In the latter approach, we use principal component analysis (PCA) to compress the data. We accurately recognize 93% of the lower face expressions and 91% of the upper face expressions.

1. Introduction

Facial expression provides cues about emotion and regulates interpersonal interaction. Because of its relevance to the study of psychological phenomena and the development of human-computer interaction (HCI), automated recognition of facial expression is an important addition to computer vision research. A number of automated facial expression recognition systems analyze six basic emotions (joy, fear, anger, disgust, sadness and surprise), and the associated expressions are classified into emotion categories rather than facial action [2, 10, 15]. In reality, humans are capable of producing thousands of expressions varying in complexity and meaning that are not fully captured with a limited number of expressions and emotion categories. Our goal is to recognize a variety of facial actions.

Automated recognition of individual motion sequences is a challenging task. Currently, most facial expression recognition systems use either complicated three-dimensional wireframe face models to recognize and synthesize facial expressions [5, 12] or use averaged optical flow within local regions (*e.g.*, forehead, eyes, nose, mouth, cheek, and chin) for recognition. In an individual region, the flow direction is changed to conform to the flow plurality of the region [2, 10, 15] or averaged over an entire region [7, 8]. Black and colleagues [2, 3] also assign parameter thresholds to their classification paradigm. These methods are relatively insensitive to subtle motion because information about small deviations is lost when their flow pattern is removed or thresholds are imposed. As a result, the recognition ability and accuracy of the systems may be reduced.

Our goal is to develop a system that recognizes both subtle feature motion and complex facial expressions. Our approach to facial expression analysis is based on the Facial Action Coding System (FACS) [4]. FACS separates expressions into upper and lower face action units (AUs), which are the smallest visibly discriminable muscle actions that combine to form expressions. We use optical flow to track facial feature points and pixel-wise facial motion. Use of optical flow to track motion is optimized in facial skin and features because they naturally have a great deal of texture.

2. Normalization

In our work, frontal views of all subjects are videotaped under constant illumination using fixed light sources, and none of the subjects wear eyeglasses. These constraints are imposed to prevent significant optical flow degradation.

Because subjects produce little out-of-plane motion, affine transformation is adequate to normalize face position and maintain face magnification invariance. We normalize the positions of all tracking points in each frame by mapping them to a standard two-dimensional face model based on three facial feature points: the medial canthus of both eyes and the uppermost point on the philtrum (see Figure 1).

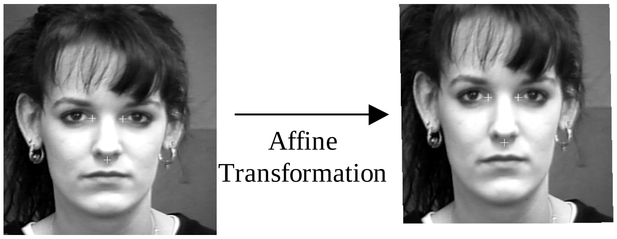
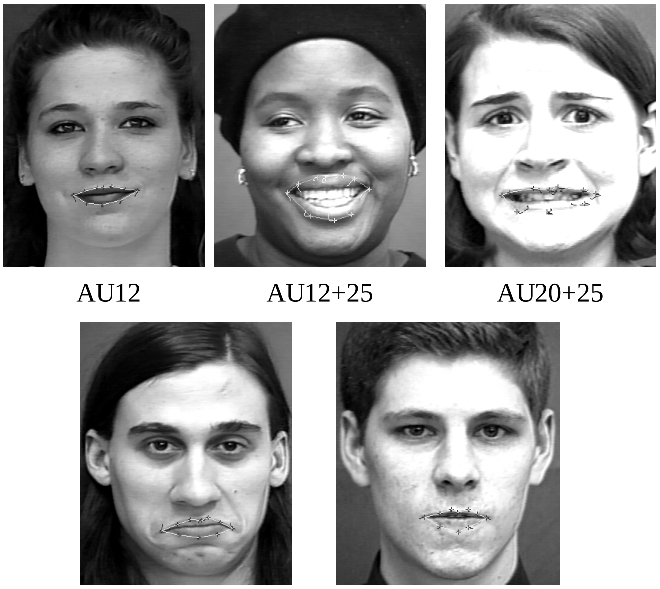


Figure 1. Normalization using affine transformation.

3. Facial Feature Point Tracking

We select facial feature points that represent underlying muscle activation and track their movement across an image sequence using optical flow. In our current work, we recognize the following lower face expressions in the mouth: AU 12 (lip corners pulled obliquely), 12+25 (lip corners pulled obliquely and mouth opened), 20+25 (lips stretched and mouth opened), 15+17 (lip corners depressed and chin raised), and 17+23+24 (lips tightened and pressed and chin raised). See Figure 2 for an illustration of these AUs.


AU12 AU12+25 AU20+25

AU15+17 AU17+23+24

Figure 2. Facial feature point tracking of lower face expressions.

We use a computer mouse to manually select 10 facial feature points around the lip contour in the first frame of each image sequence. Each point is the center of a 13x13-flow window that includes the horizontal and vertical flows. By using the hierarchical optical flow tracking method [6], the facial feature points are tracked automatically in the remaining frames of the image sequence. The displacement of each feature point is calculated by subtracting its normalized position in the first frame from its current normalized position. Since each frame has 10 feature points surrounding the lip region, the resulting 10-dimensional horizontal displacement vector by 10-dimensional vertical displacement vector is concatenated to produce a 20-dimensional displacement vector.

4. Pixel-wise Tracking and Principal Component Analysis

To recognize pixel-wise motion in the upper face, we use Wu's pixel-wise optical flow algorithm [14] to track the entire face image (417 x 385 = row x column pixels). Currently, the following upper face expressions are recognized in the forehead and brow regions: AUs 4 (brows lowered), 1+4 (inner part of the brow raised and drawn together), and 1+2 (entire brow raised). See Figure 3.

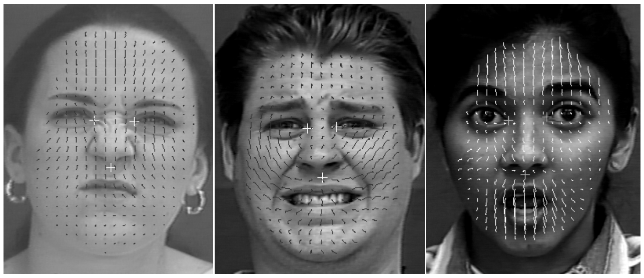
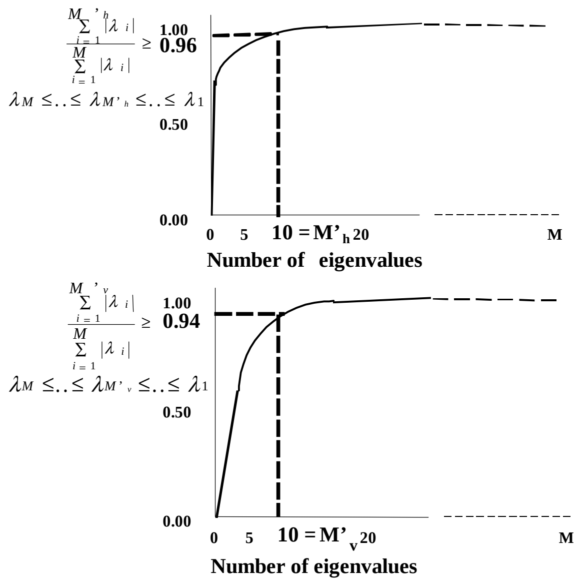

AU4 AU1+4 AU1+2

Figure 3. Pixel-wise tracking of upper face expressions.

Because we have a large image database in which consecutive frames of the sequences are strongly correlated, the expressions need to be compressed to their low-dimensional representations without losing the significant characteristics and inter-frame correlations. Principal component analysis (PCA) has excellent properties for our purposes, including image data compression and maintenance of a strong correlation between two consecutive motion frames. Since our goal is to recognize expression rather than identifying individuals or objects [1, 9, 13], we analyze facial motion using optical flow -not the gray value- to ignore differences across individual subjects.

Before using PCA, the images are automatically normalized using affine transformation to ensure that the pixel-wise flows of each frame have exact geometric correspondence. Using PCA and focusing on the (110 x 240 pixels) upper face region, 10 "eigenflows" are created (10 eigenflows from the horizontal- and 10 eigenflows from the vertical direction flows). These eigenflows are defined as the most prominent eigenvectors corresponding to the 10 largest eigenvalues of the 832 x 832-covariance matrix constructed by 832 flow-based training frames from the 44 training image sequences (see Figure 4).



$$\frac{\sum_{i=1}^{M'_h} |\lambda_i|}{\sum_{i=1}^{M} |\lambda_i|} \geq 0.96 \qquad \lambda_M \leq .. \leq \lambda_{M'_h} \leq .. \leq \lambda_1$$

$$\frac{\sum_{i=1}^{M'_v} |\lambda_i|}{\sum_{i=1}^{M} |\lambda_i|} \geq 0.94 \qquad \lambda_M \leq .. \leq \lambda_{M'_v} \leq .. \leq \lambda_1$$

Figure 4. Computation of eigenflow number. Top figure refers to horizontal eigenflows and bottom figure refers to vertical eigenflows.

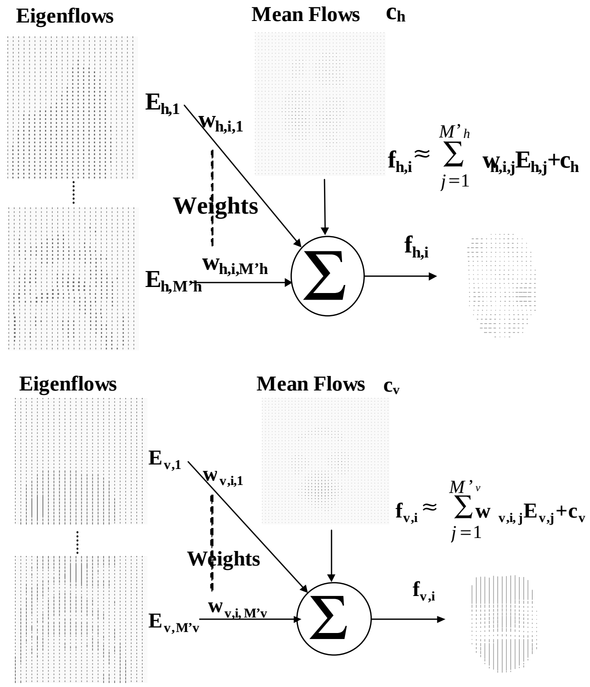
Each flow-based frame of the expression sequences is projected onto the flow-based eigenspace by taking its inner product with each element of the eigenflow set, which produces a 10-dimensional weighted vector (see Figure 5). The 10-dimensional horizontal-flow weighted vector and the 10-dimensional vertical-flow weighted vector are concatenated to form a 20-dimensional weighted vector.

5. Recognition Using Hidden Markov Models

Mase and Pentland [7] use a similar flow-based PCA approach for lip-reading recognition. They use a template matching method that analyzes the minimum value of the sum-squared-difference (SSD) between the projected flow curve of the test word and that of the word templates in the two-dimensional eigenspace. In this case, time warping is an essential preprocessing step. This approach is impractical for our purposes because the length of our image sequences is arbitrary (it varies between 9 and 44 frames) and the projected flow curve is in a higher dimensional eigenspace.

We employ Hidden Markov Models (HMMs) [11] for facial expression recognition because they perform well in the spatio-temporal domain and are analogous to human performance (*e.g.,* for speech and gesture recognition).

After separately vector quantizing the 20-dimensional training displacement vectors from the feature point tracking and the 20-dimensional training weighted vectors from the PCA, we train two sets of facial expression HMMs representing the lower (mouth) and upper face expressions (forehead and brows), respectively. Because the HMM set represents the most likely individual AU or AU combination, it can be employed to evaluate the test-input sequence. We evaluate the test-input sequence by selecting the maximum output probability value from the HMM set.



$$f_{h,i} \approx \sum_{j=1}^{M'_h} w_{h,i,j} E_{h,j} + c_h$$

$$f_{v,i} \approx \sum_{j=1}^{M'_v} w_{v,i,j} E_{v,j} + c_v$$

Figure 5. Eigenflows and their corresponding weighted vectors. Top figure is the horizontal direction and bottom figure is the vertical direction.

6. Experimental Results

The database consists of 80 subjects, both male and female, with more than 140 image sequences and 2800 images. Subjects range in age (18-35) and ethnicity (Caucasian, African-American, and Asian/Indian). Certified FACS coders coded the videotaped image sequences.

Using feature point tracking in the mouth region, 93% of the 43 lower face test image sequences (based on 40 training image sequences) are correctly recognized (see Table 1). Using pixel-wise flows with PCA in the forehead and brow regions, 91% of the 47 upper face test image sequences (based on 44 training image sequences) are accurately recognized (see Table 2).

7. Conclusion

We have developed a computer vision system that automatically recognizes a series of complex facial expressions. Our recognition system may be applied to psychological research (i.e., to code facial behavior), lip-reading and speech analysis, development of tele- or video-conferencing, and human-computer interaction (HCI). We use two approaches to extract facial motion: feature point tracking and pixel-wise flow tracking with PCA. Feature point tracking is an easy way to track facial motion, and it is sensitive to subtle feature motion. However, using this method, motion information in unselected regions (e.g., forehead, cheek, and chin) is lost. To compensate for this shortcoming, we track pixel-wise flows across the entire face and use PCA to compress the high-dimensional pixel-wise flows to low-dimensional weighted vectors. Unlike feature point tracking, pixel-wise flow tracking with PCA may introduce motion insensitivity. In future work, we will combine both methods to design a more robust system.

Table 1: Lower face expression recognition based on 43 test sequences. The average recognition rate is 93%.

HMM **Human**	**AU 12**	**AU 12+ 25**	**AU 20+ 25**	**AU 15+ 17**	**AU17 +23+ 24**	**Recognition Rate**
AU12	**9**	0	0	0	0	**100%**
AU12+ 25	0	**8**	0	0	0	**100%**
AU20+ 25	0	2	**7**	0	0	**78%**
AU15+ 17	0	0	0	**8**	1	**89%**
AU17+ 23+24	0	0	0	0	**8**	**100%**

Table 2: Upper face expression recognition based on 47 test sequences. The average recognition rate is 91%.

HMM **Human**	**AU4**	**AU1+4**	**AU1+2**	**Recognition Rate**
AU4	**11**	1	0	**92%**
AU1+4	0	**7**	1	**88%**
AU1+2	0	2	**25**	**93%**

Acknowledgement

This research is supported by NIMH grant R01 MH51435 to Jeffery Cohn and Takeo Kanade.

References

1. M.S. Bartlett, P.A. Viola, T.J. Sejnowski, B.A. Golomb, J. Larsen, J.C. Hager and P. Ekman, "Classifying Facial Action," Advances in Neural Information Processing Systems 8, MIT Press, Cambridge, MA, 1996.

2. M.J. Black and Y. Yacoob, "Recognizing Facial Expressions under Rigid and Non-Rigid Facial Motions," International Workshop on Automatic Face and Gesture Recognition, Zurich, pp. 12-17, 1995.

3. M.J. Black, Y. Yacoob, A.D. Jepson, and D.J. Fleet, "Learning Parameterized Models of Image Motion," Computer Vision and Pattern Recognition, 1997.

4. P. Ekman and W.V. Friesen, "The Facial Action Coding System," Consulting Psychologists Press, Inc., San Francisco, CA, 1978.

5. I.A. Essa, "Analysis, Interpretation and Synthesis of Facial Expressions," Perceptual Computing Technical Report 303, MIT Media Laboratory, February 1995.

6. B.D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," Proceedings of the 7th International Joint Conference on Artificial Intelligence, 1981.

7. K. Mase and A. Pentland, "Automatic Lipreading by Optical-Flow Analysis," Systems and Computers in Japan, Vol. 22, No. 6, 1991.

8. K. Mase, "Recognition of Facial Expression from Optical Flow," IEICE Transactions, Vol. E74, pp. 3474-3483, 1991.

9. H. Murase and S.K. Nayar, "Visual Learning and Recognition of 3-D Objects from Appearance," International Journal of Computer Vision, 14, pp. 5-24, 1995.

10. M. Rosenblum, Y. Yacoob and L.S. Davis, "Human Emotion Recognition from Motion Using a Radial Basis Function Network Architecture," Proceedings of the Workshop on Motion of Non-rigid and Articulated Objects, Austin, TX, November 1994.

11. L.R. Rabiner, "An Introduction to Hidden Markov Models," IEEE ASSP Magazine, pp. 4-16, January 1986.

12. D. Terzopoulos and K. Waters, "Analysis of Facial Images Using Physical and Anatomical Models," IEEE International Conference on Computer Vision, pp. 727-732, December 1990.

13. M. Turk and A. Pentland, "Eigenfaces for Recognition," Journal of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71-86, 1991.

14. Y.T. Wu, T. Kanade, J. F. Cohn, and C.C. Li, “Optical Flow Estimation Using Wavelet Motion Model,” ICCV, 1998.

15. J. Yacoob and L. Davis, "Computing Spatio-Temporal Representations of Human Faces," In Proc. Computer Vision and Pattern Recognition, CVPR-94, pp. 70-75, Seattle, WA, June 1994.