Automated Facial Expression Recognition Based on FACS Action Units

[1,2]James J. Lien
[1]Department of Electrical Engineering
University of Pittsburgh
Pittsburgh, PA 15260
jjlien@cs.cmu.edu

[2]Takeo Kanade
[2]Vision and Autonomous Systems Center
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
tk@cs.cmu.edu

[3]Jeffrey F. Cohn
[3]Department of Psychology
University of Pittsburgh
jeffcohn@vms.cis.pitt.edu

[1]Ching-Chung Li
[1]Department of Electrical Engineering
University of Pittsburgh
ccl@vms.cis.pitt.edu

Abstract

Automated recognition of facial expression is an important addition to computer vision research because of its relevance to the study of psychological phenomena and the development of human-computer interaction (HCI). We developed a computer vision system that automatically recognizes individual action units or action unit combinations in the upper face using Hidden Markov Models (HMMs). Our approach to facial expression recognition is based on the Facial Action Coding System (FACS), which separates expressions into upper and lower face action. In this paper, we use three approaches to extract facial expression information: (1) facial feature point tracking, (2) dense flow tracking with principal component analysis (PCA), and (3) high gradient component detection (i.e., furrow detection). The recognition results of the upper face expressions using feature point tracking, dense flow tracking, and high gradient component detection are 85%, 93%, and 85%, respectively.

1. Introduction

Facial expression provides sensitive cues about emotion and plays a major role in human interaction and non-verbal communication. The ability to recognize and understand facial expression automatically may facilitate communication.

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 [6, 17] or consider only averaged local motion. Using vision techniques, however, it is difficult to design a motion-based three-dimensional face model that accurately represents facial geometric properties. Also, the initial adjustment between the three-dimensional wireframe and the surface images is manual, which affects the accuracy of the recognition results. This type of recognition system becomes even more impractical and complicated when working with high-resolution images, large databases, or faces with complex geometric motion properties.

Other systems 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 [3, 15, 20] or averaged over an entire region [11, 12]. Black and colleagues [3, 4] 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.

Current recognition systems [3, 15, 20] analyze six prototypic expressions (joy, fear, anger, disgust, sadness and surprise) and classify them into emotion categories, rather than facial action. 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 develop a system that robustly recognizes both subtle feature motion and complex facial expressions [8].

2. System Structure

Our system uses three approaches to recognize facial action (Figure 1). Two of the approaches use optical flow to track facial motion. The use of optical flow is optimized for our purposes because facial skin and features naturally have great deal of texture. Two optical flow approaches are used to extract expression information: (1) facial feature point tracking, which is sensitive to subtle feature motion, and (2) dense flow tracking, which includes more facial motion information. In the latter approach, we use principal component analysis (PCA) to process the dense flows. Facial motion produces transient wrinkles and furrows perpendicular to the motion direction of the activated muscle. The facial motion associated with the furrows produces gray-value changes in the face image. The information obtained from the gray-value changes using (3) high gradient component detection is also used to recognize expression.

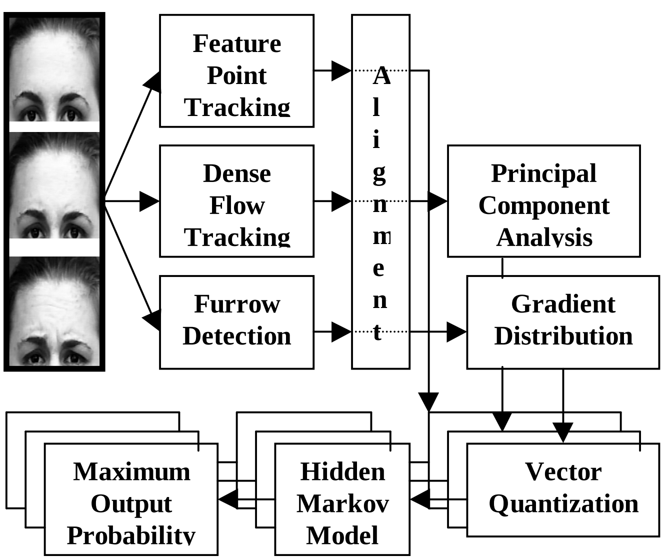


Figure 1. System structure

The expressions are recognized in the context of the entire image sequence since analysis of a dynamic image produces more accurate and robust recognition of facial expression than that of a single static image [2]. Hidden Markov Models (HMMs) [16] are used for facial expression recognition because they perform well in the spatio-temporal domain and are analogous to human performance (*e.g.*, for speech [16] and gesture recognition [21]).

We use the Facial Action Coding System (FACS) [5] to identify facial action. FACS is an anatomically based coding system that enables discrimination between closely related expressions. FACS divides the face into upper and lower face action and further subdivides motion into action units (AUs). AUs are defined as visibly discriminable muscle movements that combine to produce expressions.

Our current approach recognizes upper face expressions in the forehead and brow regions. Table 1 describes the action units associated with three brow movements.

Table 1. Description of action units in the brow region.

Action Unit		Description
AU 4		Brows are lowered and drawn together
AU 1+4		Inner parts of the brows are raised and drawn medially
AU 1+2		Entire brow is raised

2.1 Normalization

Though all subjects are viewed frontally in our current research, some out-of-plane head motion occurs with the non-rigid face motion (facial expressions). Additionally, face size varies among individuals. To eliminate the rigid head motion from facial expressions, an affine transformation, which includes translation, scaling and rotation factors, is adequate to normalize the face position and maintain face magnification invariance. The images are normalized prior to processing to ensure that flows of each frame have exact geometric correspondence. Face position and size are kept constant across subjects so that these variables do not interfere with expression recognition.

The positions of all tracking points or image pixels in each frame are normalized 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 (Figure 2).

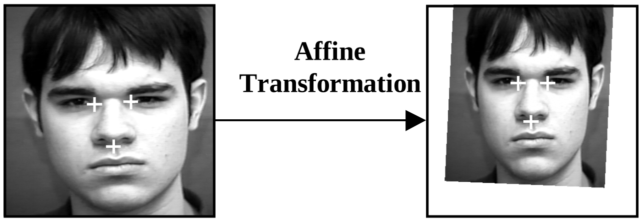


Figure 2. Normalization

2.2 Facial Feature Point Tracking

Facial expressions are recognized based on selected facial feature points that represent underlying muscle activation. The movement of facial feature points is tracked across an image sequence using Lucas-Kanade's optical flow algorithm, which has previously been shown to have high tracking accuracy [10, 14].

A computer mouse is used to manually mark 8 facial feature points around the contours of both brows in the first frame of each image sequence (see Figure 3). Each point is the center of a 13x13-flow window (image size: 417 x 385; row x column pixels) that includes the horizontal and vertical flows. Because of large facial feature motion displacement (e.g., brows raised suddenly), we use the pyramidal (5-level) optical flow approach. This approach deals well with large feature point movement (100-pixel displacement between two frames) and is sensitive to subtle facial motion (sub-pixel displacement), such as eyelid movement. The facial feature points are tracked automatically in the remaining frames of the image sequence (Figure 4).

In our current research, we recognize upper face expressions based on the displacement of 6 feature points at the upper boundaries of both brows. 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 6 feature points located at both upper brows, the resulting 6-dimensional horizontal displacement vector by 6-dimensional vertical displacement vector is concatenated to produce a 12-dimensional displacement vector.

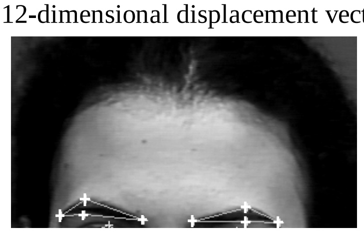


Figure 3. Feature points at upper face.

Figure 4. Feature point tracking

2.3. Dense Flow Tracking and Principal Component Analysis

Though the feature point tracking approach is sensitive to subtle feature motion and tracks large displacement well, information from areas not selected (e.g., the forehead, cheek and chin regions) is lost. To include more detailed and robust motion information from larger regions of the face, we use Wu's dense flow algorithm [19] to track each pixel of the entire face image (Figure 5).

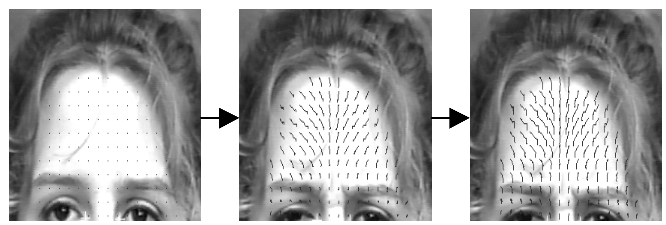


Figure 5. Dense flow sequence.

Because we have a large image database in which consecutive frames of the sequences are strongly correlated, the high dimensional pixel-wise flows of each frame 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, 7, 13, 18], we analyze facial motion using optical flow -not the gray value- to ignore differences across individual subjects.

Before using PCA, we need to ensure that the pixel-wise flows of each frame have relative geometric correspondence. We use affine transformation and automatically map the images to the 2-dimensional face model. 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 6).

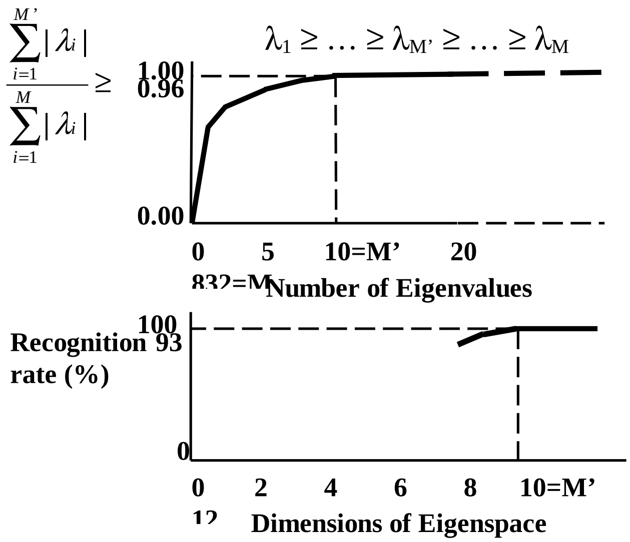


Figure 6. Computation of eigenflow number

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 (Figure 7). The 10-dimensional horizontal-flow weighted vector and the 10-dimensional vertical-flow weighted vector are concatenated to form a 20-dimensional weighted vector that corresponds to each flow-based frame. In this case, the compression rate is about 83:1.

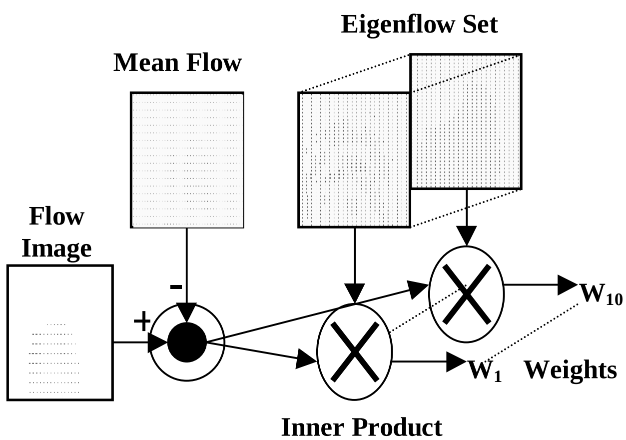


Figure 7. Horizontal weight vector computation.

2.4. High Gradient Component Detection

Facial motion produces transient wrinkles and furrows perpendicular to the motion direction of the activated muscle. The facial motion associated with the furrows produces gray-value changes in the face image. High gradient components (i.e., furrows) of the face image are extracted by using line or edge detectors. Figure 8 shows an example of the high gradient component detection. A gray value of 0 corresponds to black and 255 to white.

After normalization of each 417x385-pixel image, a 5x5 Gaussian filter is used to smooth the image. 3x5 line detectors are used to detect the horizontal lines (high gradient components in the vertical direction) in the forehead region (Figure 8).

To be sure the high gradient components are produced by transient skin or feature deformations – and not a permanent characteristic of the individual's face – the gradient intensity of each detected high gradient component in the current frame is compared to corresponding points within a 3x3 region of the first frame. If the absolute value of the difference in gradient intensity between these points is higher than the threshold value (10 in our case), it is considered a valid high gradient component produced by facial expression. All other high gradient components are ignored. In the former case, the high gradient component (pixel) is assigned a value of 1. In the latter case, the pixels are assigned a value of 0.

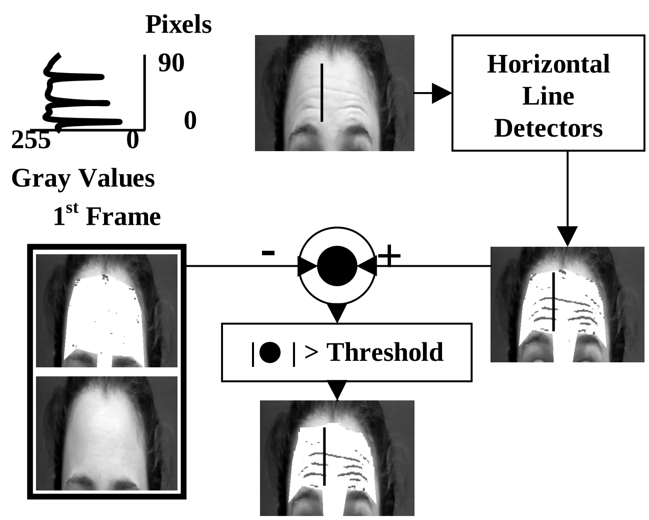


Figure 8. High gradient component detection.

The forehead region of the normalized face image is divided into 13 blocks (Figure 9). The mean value of each block is calculated by dividing the number of pixels having a value of 1 by the total number of pixels in the block. The variance of each block is calculated as well.

For upper face expression recognition, 13 mean values and 13 variance values are concatenated to form a 26-dimensional mean and variance vector for each frame.

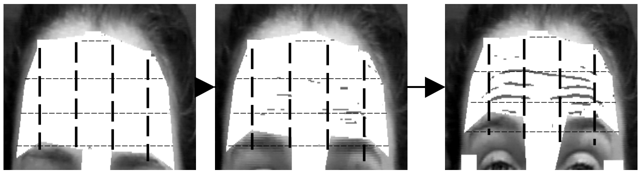


Figure 9. Quantization of high gradient components.

2.5. Recognition Using Hidden Markov Models

After separately vector quantizing [9] the 12-dimensional training displacement vectors from feature point tracking; the 20-dimensional training weighted vectors from the PCA, and the 26-dimensional training mean and variance vectors from the high gradient component detection, the corresponding facial expression HMM sets representing the upper face expressions are trained. Because the HMM set represents the most likely individual AU or AU combination, it can be employed to evaluate the test-input sequence. The test-input sequence is evaluated by selecting the maximum output probability value from the HMM set.

3. Experimental Results

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 minimize optical flow degradation. Previously untrained subjects are video recorded performing a series of expressions, and the image sequences are coded by certified FACS coders. Facial expressions are analyzed in digitized image sequences of arbitrary length (expression sequences from neutral to peak vary from 9 to 44 frames).

60 subjects, both male and female, from the larger database were used in this study. The study includes more than 260 image sequences and 5000 images. Subjects ranged in age (18-35) and ethnicity (Caucasian, African-American, and Asian/Indian).

The average recognition rate of upper face expression using feature point tracking was 85% (Table 2). The average recognition rate using dense flow tracking was 93% (Table 3), and the average recognition rate using high gradient component detection was 85% (Table 4).

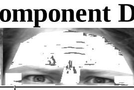
Table 2. Recognition results of feature point tracking.

Human	Feature Point Tracking		
	AU 4	AU 1+4	AU 1+2
AU 4	**22**	3	0
AU 1+4	4	**19**	2
AU 1+2	0	2	**23**

Table 3. Recognition results of dense flow tracking.

Human	Dense Flow Tracking		
	AU 4	AU 1+4	AU 1+2
AU 4	**23**	2	0
AU 1+4	3	**22**	0
AU 1+2	0	0	**25**

Table 4. Recognition results of high gradient component detection.

Human	High Gradient Component Detection			
				
	**26**	4	0	0
	5	**43**	2	0
	0	1	**24**	5
	0	0	7	**43**

4. Conclusions and Future Work

We have developed a computer vision system that automatically recognizes a number of upper face expressions. To increase system robustness, we use three approaches to extract facial motion: feature point tracking, dense flow tracking with PCA, and high gradient component detection. The pyramidal optical flow method for feature point tracking is an easy, fast and accurate way to track facial motion. It tracks large displacement well and is sensitive to subtle feature motion.

Because motion information in unselected regions (e.g., forehead, cheek, and chin) is lost, we use dense flow to track motion across the entire face. PCA is used to compress the high-dimensional pixel-wise flows to low-dimensional weighted vectors. Unlike feature point tracking, dense flow tracking with PCA introduces motion insensitivity and increases processing time. Additionally, because every pixel is analyzed in dense flow tracking, occlusion (e.g., appearance of tongue or teeth when the mouth opens) or discontinuities between the face contour and background may affect the tracking and recognition results. Because of the individual problems of each of the approaches, use of feature point tracking, dense flow tracking, and high gradient component detection with HMMs in combination may produce a more robust and accurate recognition system.

Though all three approaches using HMMs resulted in some recognition error, the pattern of the errors is encouraging. That is, the error results were classified into the expression most similar to the target (i.e., AU 4 was confused with AU 1+4 but not AU 1+2).

In future work, more detailed and complex action units will be recognized. Our goal is to increase the processing speed of the dense flow approach, estimate expression intensity based on dense flow and PCA approach, and separate rigid and non-rigid motion more robustly. Our recognition system can be applied to lip-reading, the combination of facial expression recognition and speech recognition, development of tele- or video-conferencing, human-computer interaction (HCI), and psychological research (i.e., to code facial behavior).

Acknowledgement

This research was supported by NIMH grant R01 MH51435. We would also like to thank Adena Zlochower for reviewing the manuscript.

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] J.N. Bassili, “Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Areas of the Face,” Journal of Personality and Social Psychology, Vol. 37, pp. 2049-2059, 1979.

[3] 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.

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

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

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

[7] M. Kirby and L. Sirovich, "Application of the Karhuneh-Loeve Procedure for the Characterization of Human Faces," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, No. 1, January 1990.

[8] J.J. Lien, T. Kanade, A.J. Zlochower, J.F. Cohn, and C.C. Li, "Automatically Recognizing Facial Expressions in the Spatia-Temporal Domain," Perceptual User Interface Workshop, pp. 94-97, Banff, Alberta, Canada, 1997.

[9] Y. Linde, A. Buzo, and R. Gray, "An Algorithm for Vector Quantizer Design," IEEE Trans. on Communications, Vol. COM-28, NO. 1, January 1980.

[10] 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.

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

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

[13] 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.

[14] C.J. Poelman, "The Paraperspective and Projective Factorization Methods for Recovering Shape and Motion," Technical Report CMU-CS-95-173, Carnegie Mellon University, Pittsburgh, PA, July 1995.

[15] 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.

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

[17] 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.

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

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

[20] 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.

[21] J. Yang, "Hidden Markov Model for Human Performance Modeling," Ph.D. Dissertation, University of Akron, August 1994.