Subtly Different Facial Expression Recognition And Expression Intensity Estimation

[1,2]James Jenn-Jier Lien
[1]Department of Electrical Engineering
University of Pittsburgh
Pittsburgh, PA 15260
jjlien@cs.cmu.edu
http://www.cs.cmu.edu/~jjlien

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

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

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

Abstract

We have developed a computer vision system, including both facial feature extraction and recognition, that automatically discriminates among subtly different facial expressions. Expression classification is based on Facial Action Coding System (FACS) action units (AUs), and discrimination is performed using Hidden Markov Models (HMMs). Three methods are developed to extract facial expression information for automatic recognition. The first method is facial feature point tracking using a coarse-to-fine pyramid method. This method is sensitive to subtle feature motion and is capable of handling large displacements with sub-pixel accuracy. The second method is dense flow tracking together with principal component analysis (PCA), where the entire facial motion information per frame is compressed to a low-dimensional weight vector. The third method is high gradient component (i.e., furrow) analysis in the spatio-temporal domain, which exploits the transient variation associated with the facial expression. Upon extraction of the facial information, non-rigid facial expression is separated from the rigid head motion component, and the face images are automatically aligned and normalized using an affine transformation. This system also provides expression intensity estimation, which has significant effect on the actual meaning of the expression.

1. Introduction

The face is a rich source of information about human behavior. Facial expression displays emotion [7], regulates social behavior [5], signals communicative intent [9], is computationally related to speech production [17], and reveals brain function and pathology [20]. To make use of the information afforded by facial expression, automated reliable and valid measurement is critical.

Most facial expression recognition systems use either complicated three-dimensional (3-D) wireframe face models to recognize and reproduce facial expressions [8,23] or analyze averaged optical flow within local regions (*e.g.*, forehead, brows, eyes, nose, mouth, cheek, and chin). A limitation of wireframe face models is that the initial alignment between the 3-D wireframe and the 2-D surface images is manual, which affects the accuracy of the recognition results. Additionally, it is impractical and difficult to use 3-D wireframe models when working with high-resolution images, large databases (*i.e.*, number of subjects or image sequences), or faces with complex geometric motion properties.

In contrast to the complex 3-D geometric models, optical flow-based approaches treat the facial expression recognition problem as 2-D. These approaches have been shown to track motion and classify prototypic emotion expressions [3,4,16,22,26]. A problem, however, is that the flow direction of each individual local face region is changed to conform to the flow plurality of the region [3, 22, 26] or averaged over an entire region [15, 16]. These systems are often insensitive to subtle motion because information about small deviations is lost. The recognition ability and accuracy of these systems may be reduced further when presented with less stylized expressions.

Most research in facial expression recognition is limited to six basic emotions (*i.e.*, joy, fear, anger, disgust, sadness, and surprise) posed by a small set of subjects [3,4,22,26]. These stylized expressions are classified into emotion categories rather than facial action. In everyday life, however, these six basic expressions

occur relatively infrequently. Human are capable of producing thousands of expressions that vary in complexity, intensity and meaning. Emotion or intention more often is communicated by subtle changes in one or two discrete features. For example, disagreement or anger may be communicated to an interactant by furrowed eyebrows (AU 4). The degree of anger experienced may be communicated by the expression intensity of the brow motion. Our goal is to develop a computer vision system, including both facial feature extraction and facial expression recognition based on FACS AUs, that is capable of automatically discriminating among subtly different facial expressions [11,12].

2. Extraction and Recognition System

Three methods are used to extract expression information (Figure 1). Feature point tracking and dense flow tracking are used to track facial motion since our goal is to recognize expressions varying in expression intensity in the spatio-temporal domain. The use of optical flow to track motion in the face is particularly appropriate because facial skin and features naturally have a great deal of texture. Facial feature point tracking is especially sensitive to subtle feature motion. Dense flow tracking together with principal component analysis (PCA) includes motion information from the entire face. Low-dimensional weight vectors represent the high-dimensional pixel-wise optical flows of each frame. These weight vectors are used to estimate expression intensity.

High gradient component (*i.e.* furrow) analysis in the spatio-temporal domain is used to recognize expressions by the presence of furrows. Facial motion produces transient wrinkles and furrows perpendicular to the motion direction of the activated muscle. The facial motion associated with a furrow produces gray-value change in the face image, which can be extracted by use of high gradient component detectors.

Because analysis of dynamic images produces more accurate and robust recognition than that of a single static image [2], expressions are recognized in the context of entire image sequences of arbitrary length. Hidden Markov Models (HMMs) [21] are used for facial expression recognition in image sequences of arbitrary length because they perform well in the spatio-temporal domain, robustly deal with the time warping problem (compared with [15]). Furthermore, the structure of HMMs provides a natural description for time dependent actions (*e.g.*, for facial expression [11,12], gesture [27] and speech recognition [21]).

2.1 Facial Action Coding System (FACS)

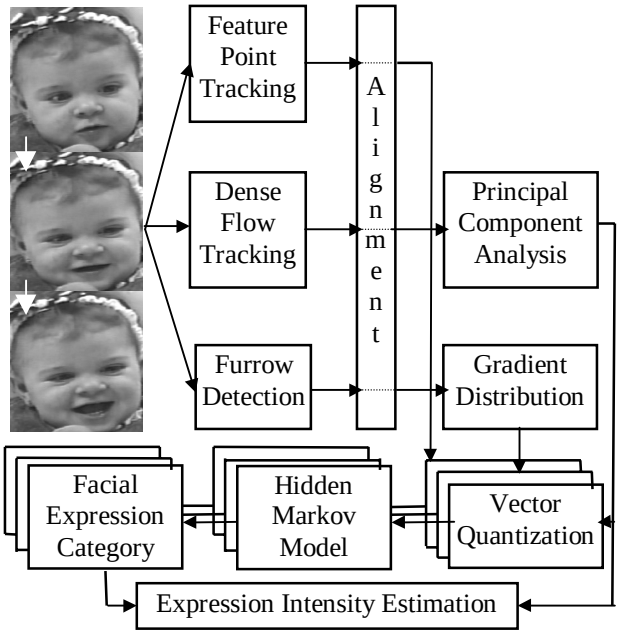


Figure 1. Block diagram of a facial expression recognition system.

Our approach to facial expression analysis is based on the Facial Action Coding System (FACS) [6], which is an anatomically based coding system that enables discrimination between closely related expressions. FACS divides the face into upper and lower regions and subdivides motion into action units (AUs). AUs are the smallest visibly discriminable muscle actions that combine to perform expressions. In the present study, three sets of subtly different facial expressions which occur frequently in everyday life are recognized, and their expression intensities are estimated (Table 1).

2.2 Rigid and Non-rigid Motion Separation and Geometric Normalization

Although all subjects are viewed frontally in our current research, some small out-of-plane head motion occurs with facial expressions. Additionally, face size varies among individuals. In order to separate non-rigid facial expression from rigid head motion, an affine transformation, which includes translation, scaling and rotation factors, is applied to each image. This normalizes the facial geometric position and enforces face magnification invariance. In an initial processing step, the images are automatically normalized to ensure that flows or gray values of each face image have close geometric correspondence with those of other images in the set. Face position and size are kept constant across subjects so that these variables do not interfere with

Table 1. Facial Action Coding System Action Units [6].

Upper Face Expressions		
AU4	AU1+4	AU1+2
		
Lower Face Expressions		
AU12	AU6+12+25	AU20+25
		
AU9+17	AU17+23+24	AU15+17
		

expression recognition.

The positions of all tracked points and image pixels in each frame are automatically normalized by warping them to a standard 2-D face model based on three facial feature points: the medial canthus of both eyes and the uppermost point on the philtrum (Figure 2). In addition, based on these three facial feature points, the original 490 x 640 (row x column) pixel display is automatically cropped to 417 x 385 pixels for each frame to remove the unnecessary background and keep the foreground face.

3. Three Extraction Methods

In our system, three methods are developed to automatically extract facial expression information: (1) facial feature point tracking using the coarse-to-fine pyramid method, (2) dense flow tracking together with PCA, and (3) high gradient component analysis in the spatio-temporal domain.

3.1 Facial Feature Point Tracking Using the Coarse-to-Fine Pyramid Method

Because facial features have high texture and represent underlying muscle activation, optical flow can be used to track movement of feature points, and facial expressions can be recognized based on the motion of these feature points. Feature points located around the contours of the brows, eyes, nose, mouth, and below the lower eyelids are manually marked in the first frame of each image sequence using a computer mouse (Figure 3). Each feature point is the center of a 13 x 13 flow window which is used to compute the horizontal and vertical flow of the feature.

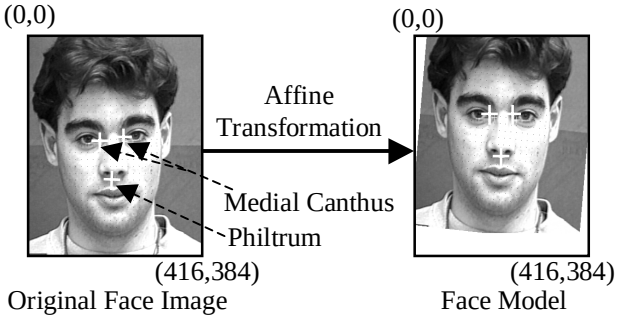


Figure 2. Facial image normalization.

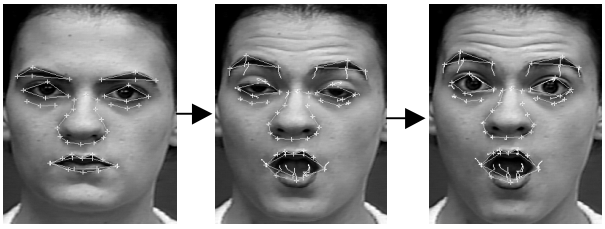


Figure 3. Facial feature point tracking sequence.

The movement of facial feature points is automatically tracked across an image sequence using Lucas-Kanade's optical flow algorithm, which has high tracking accuracy [14] (Figure 3). The pyramidal (5 level) optical flow method [19] is used for tracking because it robustly manages large facial feature motion displacement, such as mouth opening or brows raised suddenly. This method deals well with large feature point movement (100 pixel displacement between two frames) while maintaining its sensitivity to subtle (sub-pixel) facial motion.

In this study, upper face expressions are recognized based on the displacements of 6 feature points at the upper boundaries of both brows, and lower face expressions are recognized based on the displacements of 10 feature points around the mouth. The displacement of each feature point is calculated by subtracting its normalized position in the first frame from its current normalized position. The 6- and 10-dimensional horizontal displacement vectors and 6- and 10-dimensional vertical displacement vectors are concatenated to form 12- and 20-dimensional displacement vectors for the upper and lower facial expressions, respectively. The 12- and 20-dimensional displacement vectors of the upper and lower face represent the facial motion of each frame.

3.2 Dense Flow Tracking together with Principal Component Analysis

The facial feature point tracking of previous section is sensitive to subtle feature motion and tracks large

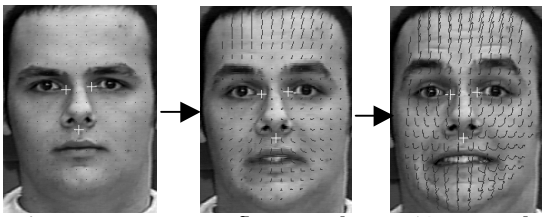


Figure 4. Dense flow tracking. (Compared with Figure 3: same upper face expression but different lower face expressions.)

displacements well. In addition, it is useful to measure the motion of the entire face, including the forehead, cheek and chin regions. To include this detailed motion information, each pixel of the entire face image is tracked using dense flow [25] (Figure 4).

Because we have a large image database in which the motion of consecutive frames in a sequence is strongly correlated, the high-dimensional pixel-wise flows of each frame need to be compressed to their low-dimensional representations without losing significant characteristics and inter-frame correlation. 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 identify individuals or objects [10, 18, 24], facial motion is analyzed using dense flow - not gray value - to ignore differences across individual subjects (compared with [1]). To ensure that the pixel-wise flows of each frame have relative geometric correspondence, an affine transformation is used to automatically warp the pixel-wise flows of each frame to the 2-D face model.

Using PCA and focusing on the (110 x 240 pixels) upper face region, 10 "eigenflows" are created (Figure 5) (10 eigenflows from the horizontal- and 10 eigenflows from the vertical direction flows [11,12]). These eigenflows are defined as the 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. The compression rate is 83:1.

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, producing a 10-dimensional weight vector (Figure 6). The 10-dimensional horizontal-flow weight vector and the 10-dimensional vertical-flow weight vector are concatenated to form a 20-dimensional weight vector for each flow-based frame.

3.3 High Gradient Component Analysis in the Spatio-Temporal Domain

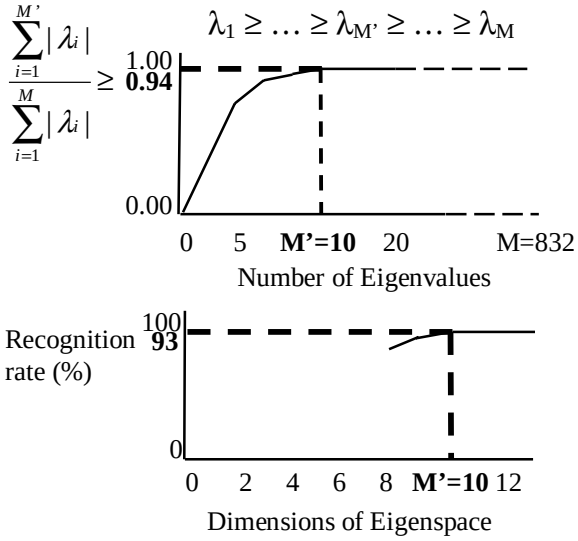


Figure 5. Computation of eigenflow number for vertical direction dense flows.

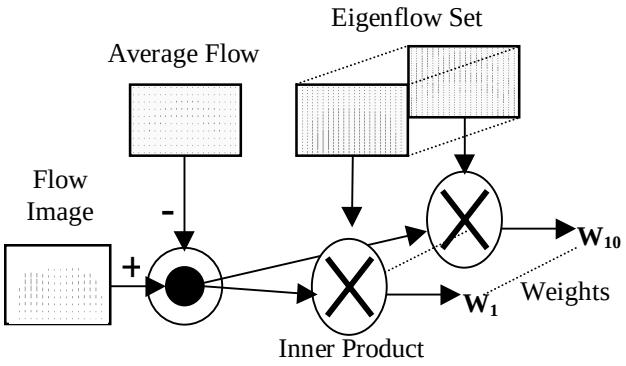


Figure 6. Vertical-flow weight vector computation for the upper face expressions.

Facial motion produces transient wrinkles and furrows perpendicular to the motion direction of the activated muscle. The facial motion associated with these furrows produces gray-value changes in the face image. High gradient components of the face image are extracted with a variety of line or edge detectors. After normalization of each 417 x 385 pixel image, a 5 x 5 Gaussian filter is used to smooth the image. 3 x 5 horizontal line and 5 x 3 vertical line detectors are used to detect horizontal lines (*i.e.*, high gradient components in the vertical directions) and vertical lines in the forehead region, respectively; 5 x 5 diagonal line detectors are used to detect diagonal lines along the nasolabial furrow; and 3 x 3 edge detectors are used to detect high gradient components around the lips and on the chin region.

To verify that 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 with corresponding points within a 3 x 3 region of the first frame. If the absolute value of the difference in gradient intensity between these points is higher than a threshold value, 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. All other pixels are assigned a value of 0. An example of the procedure for extracting high gradient components on the forehead region is shown in Figure 7. A gray value of 0 corresponds to black and 255 to white.

The forehead (upper face) and lower face regions of the normalized face image are each divided into 16 blocks (Figure 8). 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 and lower face expression recognition, mean and variance values are concatenated to form two 32-dimensional mean-variance vectors for each frame.

4. Expression Recognition and Expression Intensity Estimation

The 12- and 20-dimensional training displacement vectors from feature point tracking, the 20-dimensional training weight vectors from the dense flow tracking together with PCA, and the 32- and 32-dimensional training mean-variance vectors from the high gradient component detection are each vector quantized [13]. HMMs are then trained. Because the HMM set represents the most likely individual action unit (AU) or AU combinations, it can be employed to evaluate the test-input sequence. The test-input sequence is evaluated by selecting the maximum likelihood decision value from the HMM set.

After recognizing an input facial expression sequence, the expression intensity of an individual frame in this sequence is estimated using the correlation property of PCA. That is, the minimum distance between two projected points (weight vectors) in eigenspace has the maximum correlation or motion similarity. The sum-of-squared-difference (SSD) is used to find the frame with the best match in expression (motion) intensity from any training sequence having the same expression as the test frame (Figure 9). Since the expression intensity of the frame from the training set has been previously ascertained, the relative expression intensity of the test

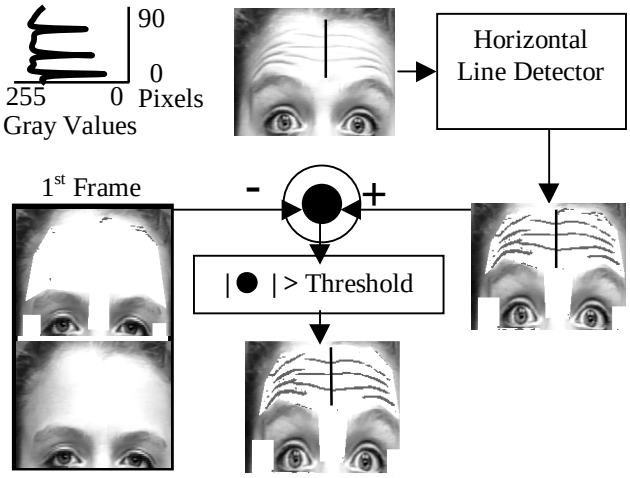


Figure 7. The procedure for horizontal line detection in the spatio-temporal domain at the forehead (upper face) region.

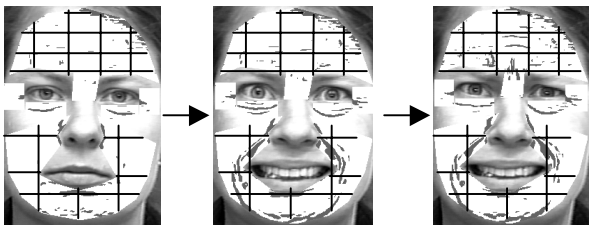


Figure 8. Quantization of the high gradient components.

expression can be determined.

5. Experimental Results

For this study, frontal views of all subjects were videotaped under constant illumination using fixed light sources in order to minimize optical flow degradation, and none of the subjects wore eyeglasses. Previously untrained subjects were video recorded performing a series of expressions, and the image sequences were coded by certified FACS coders. Facial expressions were analyzed in digitized image sequences of arbitrary length (expression sequences from neutral to peak varied from 9 to 47 frames).

Subjects were 85 males and females (Asian, Euro- and African-American) between the ages of 1 and 35 years. 300 image sequences were analyzed. Recognition accuracy did not vary between males and females, and Euro- and African-Americans.

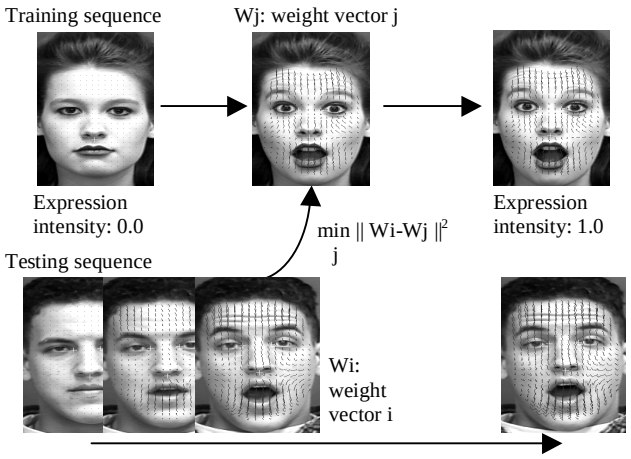


Figure 9. Expression intensity estimation.

The average recognition rate of upper face expressions was 85% by feature point tracking, 93% by dense flow tracking with PCA, and 85% by high gradient component analysis in the spatio-temporal domain. These results are based on 60, 44, and 100 training image sequences and 75, 75, and 160 testing image sequences, respectively (Table 2). The average recognition rate of lower face expressions was 88% by feature point tracking and 81% by high gradient component analysis (based on 120 and 50 training image sequences, and 150 and 80 testing image sequences, respectively) (Table 2). Results for dense flow tracking together with PCA are not yet available for the lower face.

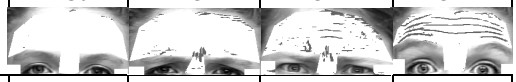
6. Conclusion

We have developed a computer vision system that automatically recognizes facial expressions based on FACS action units. To optimize system performance, three methods extract facial motion: feature point tracking, dense flow tracking together with PCA, and high gradient component analysis in the spatio-temporal domain.

The coarse-to-fine pyramidal optical flow method for feature point tracking is an easy, fast, and accurate way to track facial feature motion. It tracks large displacement well and is sensitive to subtle feature motion in sub-pixel accuracy. To track motion across the entire face, dense flow together with PCA is used. PCA compresses the high-dimensional pixel-wise flows to a low-dimensional weight vector for each frame. Unlike feature point tracking, dense flow tracking introduces insensitivity to small local motions and is subject to error due to occlusion (*e.g.* hair covering the forehead) or

Table 2. Recognition results.

Upper Face Expression Recognition			
Human	**Feature Point Tracking**		
	AU4	AU1+4	AU1+2
AU4	**22**	3	0
AU1+4	4	**19**	2
AU1+2	0	2	**23**
Human	**Dense Flow Tracking with PCA**		
AU4	**23**	2	0
AU1+4	3	**22**	0
AU1+2	0	0	**25**

Human	**High Gradient Component Analysis**			
	AU0	AU4	AU1+4	AU1+2
				
AU0	**26**	4	0	0
AU4	5	**43**	2	0
AU1+4	0	1	**24**	5
AU1+2	0	0	7	**43**

Lower Face Expression Recognition						
Human	**Feature Point Tracking**					
AUs	12	6+12 +25	20+25	9+17	15+17	17+23+24
12	**25**	0	0	0	0	0
6+12+25	0	**21**	4	0	0	0
20+25	0	5	**20**	0	0	0
9+17	0	0	0	**22**	0	3
15+17	0	0	0	0	**23**	2
17+23+24	0	0	0	3	1	**21**

Human	**High Gradient Component Analysis**	
	1: AU12 or AU6+12+25 	**2**:AU9+17 or AU17+23+24 
1	**42**	8
2	7	**23**

discontinuities between the face contour and background or appearance of tongue or teeth when the mouth opens. Additionally, processing time is prolonged in dense flow tracking (98% computing time of this system) because of recursive computation in the wavelet-based approach (multiple basis functions) we employ.

High-gradient component analysis in the spatio-temporal domain is sensitive to change in transient facial features (*e.g.*, furrows), but is subject to error from individual differences in subjects. Younger subjects, especially infants, show less furrowing than older ones, which reduces the information value of high gradient

component detection.

Although all three methods resulted in some recognition error, the pattern of errors was encouraging. That is, the error results were classified into the expression most similar to the target (*e.g.*, AU4 is confused with AU1+4 but not AU1+2). Because each method has strengths and weaknesses, feature point tracking, dense flow tracking together with PCA, and high gradient component analysis can be used in combination to produce a more robust and accurate recognition system. A focus of current work is the implementation of a multi-dimensional HMM to integrate these three methods.

In future work, we will recognize more detailed and complex action units, increase the processing speed of dense flow analysis, interpolate expression intensity, and separate rigid and non-rigid motion more robustly. Potential applications include assessment of nonverbal behavior in clinical and research settings, speech recognition in combination with lip-reading, teleconferencing, and human-computer interface/ interaction. In addition, automated quantitative assessment of facial expression (*i.e.* expression intensity estimation) can inform work in facial animation (analysis and synthesis).

Acknowledgements

This research is supported by NIMH grant R01 MH51435. Thanks to David LaRose for his help, comments, and encouragement. Thanks to Adena J. Zlochower for her help with FACS.

References

[1] M.S. Bartlett, *et al*., "Classifying Facial Action," *Adv. in Neural Info. Proc. Sys. 8*, MIT Press, 1996.

[2] J.N. Bassili, "Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Areas of the Face," *J. of Personality and Social Psy.*, Vol. 37, pp. 2049-2059, 1979.

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

[4] M.J. Black, *et al*., "Learning Parameterized Models of Image Motion," *CVPR*, 1997.

[5] J.F. Cohn and M. Elmore, "Effect of Contingent Changes in Mothers' Affective Expression on the Organization of Behavior in 3-Month-Old Infants," *Infant Behavior and Development*, Vol. 11, pp. 493-505, 1988.

[6] P. Ekman and W.V. Friesen, "The Facial Action Coding System," *Consulting Psy. Press*, CA, 1978.

[7] P. Ekman, "Facial Expression and Emotion," *American Psychologist*, Vol. 48, pp. 384-392, 1993.

[8] I.A. Essa, "Analysis, Interpretation and Synthesis of Facial Expressions," *Perceptual Computing TR 303*, MIT Media Laboratory, Feb. 1995.

[9] A.J. Fridlund *Human Facial Expression: An Evolutionary View*, Academic Press, CA, 1994.

[10] M. Kirby and L. Sirovich, "Application of the Karhuneh-Loeve Procedure for the Characterization of Human Faces," *IEEE Trans. on PAMI* 12, No. 1, 1990.

[11] J.J. Lien, T. Kanade, A.J. Zlochower, J.F. Cohn, and C.C. Li, "Automatically Recognizing Facial Expressions in the Spatio-Temporal Domain," *Workshop on Perceptual User Interfaces*, pp. 94-97, Banff, Alberta, Canada, October 19-21, 1997.

[12] J.J. Lien, T. Kanade, J.F. Cohn, and C.C. Li, "Automated Facial Expression Recognition Based on FACS Action Units," *Third IEEE International Conference on Automatic Face And Gesture Recognition*, Nara, Japan, April 14-16, 1998.

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

[14] B.D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," *Proc. of the 7th Intl. Joint Conf. on AI*, 1981.

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

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

[17] D. McNeil, "So you think gestures are nonverbal?" *Psychological Review*, *92*, 350-371, 1985.

[18] H. Murase and S.K. Nayar, "Visual Learning and Recognition of 3-D Objects from Appearance," *IJCV*, 14, pp. 5-24, 1995.

[19] C.J. Poelman, "The Paraperspective and Projective Factorization Methods for Recovering Shape and Motion," *Ph.D. dissertation*, Carnegie Mellon University, CMU-CS-95-173, July 1995.

[20] W.E. Rinn,. "The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions." *Psychological Bulletin*, *95*, pp. 52-77, 1984.

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

[22] M. Rosenblum, Y. Yacoob and L.S. Davis, "Human Emotion Recognition from Motion Using a Radial Basis Function Network Architecture," *Proc. of the Workshop on Motion of Non-rigid and Articulated Objects*, Austin, TX, Nov. 1994.

[23] D. Terzopoulos and K. Waters, "Analysis of Facial Images Using Physical and Anatomical Models," *ICCV*, pp. 727-732, Dec. 1990.

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

[25] Y.T. Wu, T. Kanade, J. F. Cohn, and C.C. Li, "Optical Flow Estimation Using Wavelet Motion Model," *ICCV*, 1998.

[26] J. Yacoob and L. Davis, "Computing Spatio-Temporal Representations of Human Faces," *CVPR*, pp. 70-75, 1994.

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