Proceedings of the 7th IEEE Robot and Human Communication Workshop. September 30-October 2, 1998, Takamatsu, Japan

Beyond Prototypic Expressions: Discriminating Subtle Changes in the Face

Jeffrey F. Cohn
Department of Psychology
University of Pittsburgh
4015 O'Hara Street,
Pittsburgh PA 15260
jeffcohn@vms.cis.pitt.edu

James Jenn-Jier Lien
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
jjlien@cs.cmu.edu

Takeo Kanade
Departments of Computer Science
and Electrical Engineering,
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
tk@cs.cmu.edu

Wei Hua
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
hua@cs.cmu.edu

Adena Zlochower
Department of Psychology
University of Pittsburgh
Pittsburgh, PA 15260
adena@vms.cis.pitt.edu

Abstract

Current approaches to automated analysis have focused on a small set of prototypic expressions (e.g., joy or anger). Prototypic expressions occur infrequently in everyday life, however, and emotion expression is far more varied. To capture the full range of emotion expression, automated discrimination of fine-grained changes in facial expression is needed. We developed and implemented a computer vision system, Automated Face Analysis, that is sensitive to subtle changes in the face. Three convergent modules extract feature information and discriminate FACS action units using Hidden Markov Models. The modules include feature-point and dense-flow tracking and high-gradient component detection. In image sequences from 100 young adults, action units and action unit combinations in the brow, eye, and mouth regions were selected for analysis if they occurred a minimum of 25 times in the image database. Selected facial features were automatically tracked using hierarchical algorithms to estimate optical flow and high-gradient components. Image sequences were randomly divided into training and test sets. Automated Face Analysis demonstrated high concurrent validity with manual FACS coding.

1. Introduction

The face is a rich source of information about human behavior. Facial displays indicate emotion and pain [7], regulate interpersonal behavior [5], reveal brain function and pathology [17], and signal changes with development in children (e.g., [9]). To make use of the information afforded by facial expression, reliable, valid, and efficient methods of measurement are critical.

Computer-vision-based approaches to facial expression analysis (e.g., Black and Yacoob [2]; Mase [15]) discriminate between a small set of prototypic expressions of emotion. This focus follows from the work of Darwin [6] and more recently Ekman [7], who proposed that "basic emotions" (i.e., joy, surprise, anger, sadness, fear, and disgust) each have a prototypic facial expression, involving changes in facial features in multiple regions of the face, which facilitates analysis. In everyday life, prototypic expressions may occur relatively infrequently, and emotion more often is communicated by changes in one or two discrete features, such as tightening the lips, which may communicate anger [4]. To capture the subtlety of human emotion and non-verbal communication, automated discrimination of fine-grained changes in facial expression is needed.

The Facial Action Coding System (FACS) [8] is a human-observer-based system designed to detect subtle changes in facial features. FACS consists of 44 anatomically based "action units" which individually or in combination can represent all visibly discriminable expressions. Seminal work by Mase [15] and Essa and Pentland [10] suggested that FACS action units could be detected from differential patterns of optical flow. Essa and Pentland [10] found increased flow in muscle regions associated with action units in the brow and in the mouth region. The specificity of optical flow to action

unit discrimination, however, was not tested. Discrimination of facial expression remained at the level of emotion prototypes rather than the finer and more objective level of FACS action units. An exception, Bartlett et al. [1], discriminated action units in the brow and eye regions in a small number of subjects.

Current methods of estimating optical flow may lack sensitivity to subtle motion, which is needed to discriminate expressions at this more fine-grained and objective level. Work to date has used aggregate measures of optical flow within relatively large feature regions (e.g., mouth or cheeks), including modal flow and mean flow within the region [2, 18, 14,15, 22]. Black and colleagues [2, 3] also disregard subtle changes in flow that are below an assigned threshold. Information about small deviations is lost when the flow pattern is aggregated or thresholds are imposed. As a result, the accuracy of discriminating FACS action units may be reduced.

The objective of the present study was to implement the first version of our automated method of facial analysis and to assess its concurrent validity with manual FACS coding. Unlike previous systems that use aggregate measures of flow within large feature windows, our system tracks the movement of closely spaced feature points and imposes no arbitrary thresholds. In addition, to optimize system performance, three separate modules are used to extract convergent information about facial features. The descriptive power of feature tracking is evaluated by comparing the results of Hidden Markov Modeling or discriminant analysis with those of manual FACS coding.

2. Automated Face Analysis

Automated Face Analysis includes three convergent modules for feature extraction (Figure 1). One module, feature-point tracking, tracks the movement of pre-selected feature points within very small feature windows (currently 13 by 13 pixels). The feature points to be tracked are in regions of high texture and represent underlying muscle activation of closely related action units. In an earlier study [23], we found that motion in a small number of pre-selected feature points was sufficient to discriminate between closely related action units. The second module, dense-flow tracking, quantifies flow across the entire face. Dense flow tracking has reduced sensitivity for the localized, small motion in which feature-point tracking excels, but extracts information within larger regions. The third module tracks changes in facial lines and furrows (i.e., high gradient component detection). Facial motion produces transient lines and wrinkles which are indicative of the underlying action units. Hidden Markov Modeling (HMM) or discriminant analysis is performed on the feature point, dense flow, and high-gradient component data for action unit discrimination and intensity estimation.

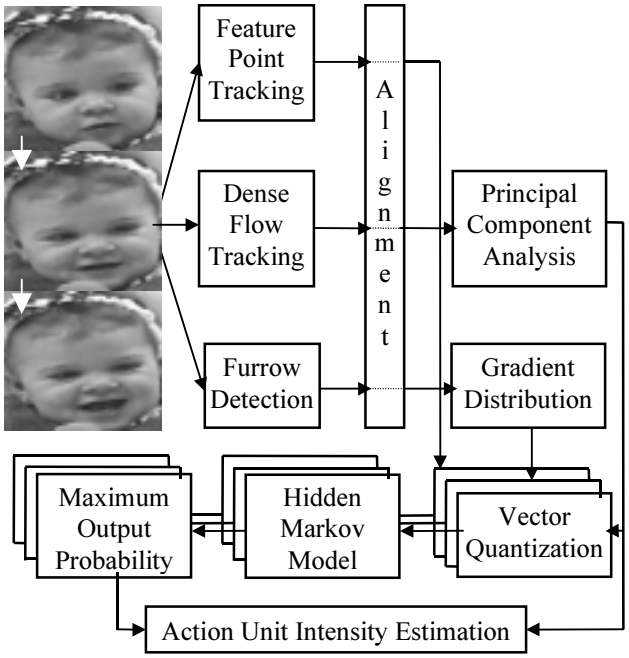


Figure 1. Automated Face Analysis

2.1 Facial Action Coding System (FACS)

Our approach to facial expression analysis is based on the Facial Action Coding System (FACS) [8], which is a comprehensive, anatomically based coding system. 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, fifteen FACS action units and action unit combinations are discriminated and their intensities estimated (Figure 2).

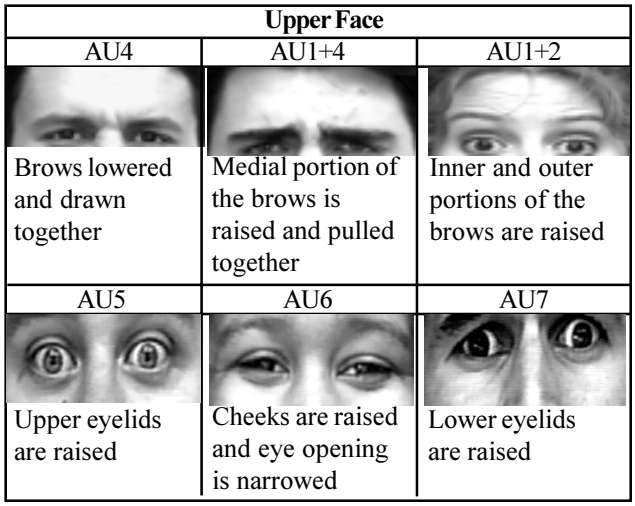


Figure 2. FACS Action Units (AUs)

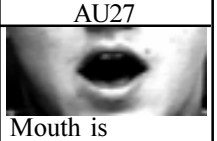
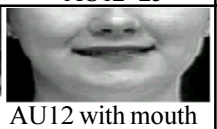
Lower Face		
AU25	AU26	AU27
 Lips are relaxed and parted	 Lips are relaxed and parted; mandible is lowered	 Mouth is stretched open and the mandible pulled down
AU12	AU12+25	AU20+25
 Lip corners are pulled obliquely	 AU12 with mouth opening	 Lips are parted and pulled back laterally
AU9+17	AU17+23+24	AU15+17
 The infraorbital triangle and center of the upper lip are pulled upwards and the chin boss is raised (AU17)	 AU17 and lips are tightened, narrowed, and pressed together	 Lip corners are pulled down and chin is raised

Figure 2 continued. FACS Action Units (AUs)

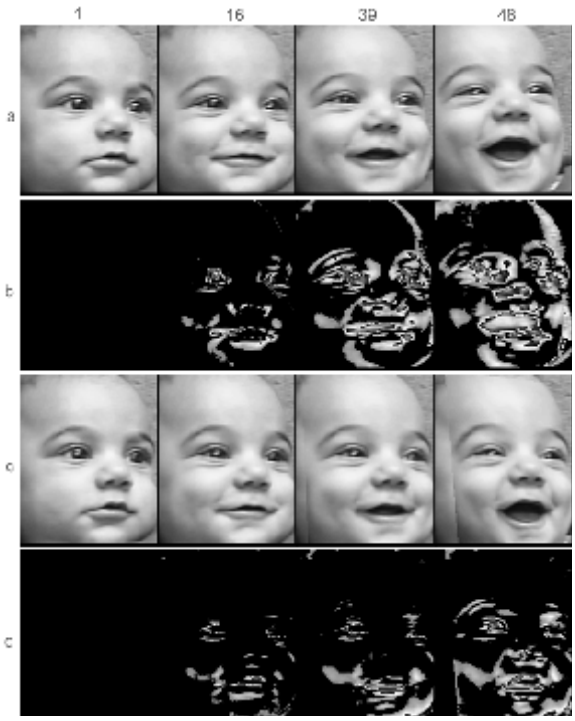


Figure 3. Perspective Alignment

2.2 Image Alignment

To remove the effects of spatial variation in face position, slight head rotation, and facial proportions, images are warped automatically to a standard face model using either an affine or perspective transformation [21]. By automatically controlling for face position, orientation, and magnification in this initial processing step, feature-point displacement, dense flow, and high-gradient component vectors in each frame have close geometric correspondence. Face position and size are kept constant across subjects so that these variables do not interfere with action unit discrimination.

An example of image alignment by perspective transformation can be seen in Figure 3. In the original image sequence (Row A), the subject turns his head to the right while beginning to smile (AU 12), pitches the head up, increases mouth opening (AU 25/26), and narrows the eyes and raises the cheeks (AU 6) and brows (AU 1+2). The intensity difference images in Row B illustrate the confounding of rigid and non-rigid motion in the original image sequence (Row A). Row C shows the perspective transformation of the original image sequence in Row A, and row D the intensity differences between Row C and the initial frame in Row A. In comparison with the intensity images of the original series (Row B), facial features no longer appear doubled, and white areas have specificity for non-rigid motion.

2.3 Feature Extraction

Three modules extract information about transient and permanent facial features and their change over time: feature-point tracking, dense-flow tracking with PCA, and high-gradient component detection.

2.3.1 Feature-point tracking

Because facial features have high texture and represent underlying muscle activation, optical flow may be used to track movement of feature points. Facial expressions are discriminated on the motion of these 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 4).

The movement of feature points is automatically tracked in the image sequence using the Lucas-Kanade algorithm [13]. Given an n by n feature region R and a gray-scale image I, the algorithm solves for the displacement vector $\boldsymbol{d} = (d_x, d_y)$ of the original n by n feature region by minimizing the residual $E(\boldsymbol{d})$, which is defined as

$$E(\boldsymbol{d}) = \sum_{\boldsymbol{x} \in R} [\, I_{t+1}(\boldsymbol{x}+\boldsymbol{d}) - I_t(\boldsymbol{x}) \,]^2$$

where $\boldsymbol{x} = (x, y)$ is a vector of image coordinates. The Lucas-Kanade algorithm performs the minimization efficiently by using spatio-temporal gradients, and the displacements d_x and d_y are solved with sub-pixel accuracy. The region size used in the algorithm currently is 13-by-13. The algorithm is implemented by using an iterative hierarchical 5-level image-pyramid [25], with which rapid and large displacements of up to 100 pixels (e.g., as found in sudden mouth opening) can be robustly tracked while maintaining sensitivity to subtle (sub-pixel) facial motion.

Action units are discriminated from the displacements of

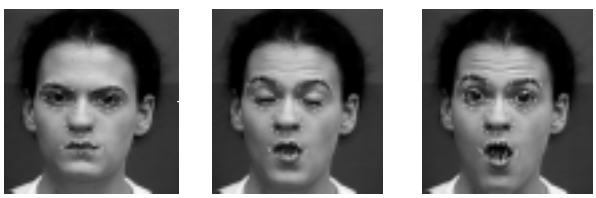


Figure 4. Feature-Point Tracking

six feature points in the brows, eight around the eyes, and ten 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-, 8- and 10-dimensional horizontal and vertical displacement vectors are concatenated to form 12-, 16-, and 20-dimensional displacement vectors that represent the facial motion of each frame.

2.3.2 Dense Flow Tracking with Principal Component Analysis (PCA)

Feature-point tracking is sensitive to subtle feature motion and tracks large displacement well but omits information from the forehead, cheek and chin regions, which also contain important expression information. To include information from these regions, the entire face image is tracked using dense flow [21] (Figure 5).

Because our goal is to discriminate expression rather than identify individuals or objects [11, 16, 20], facial motion is analyzed using dense flow rather than gray value. Compared with Bartlett et al. [1], estimates of facial motion by this approach are unaffected by differences between subjects. To ensure that the pixel-wise flows of each frame have close geometric correspondence, flows are automatically warped to the 2-D face model. The high-dimensional pixel-wise flows of each frame are compressed by PCA to low-dimensional representations without losing significant characteristics or inter-frame correlation.

Using PCA and focusing on the upper face region, ten

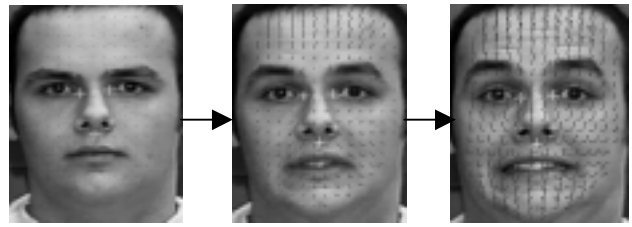


Figure 5. Dense-Flow Tracking

horizontal and ten vertical "eigenflows" are computed [12]. These eigenflows are defined as the eigenvectors corresponding to the ten largest eigenvalues of the 832 x 832 covariance matrix constructed by 832 flow-based training frames from 44 training image sequences.

Each flow-based frame of an expression sequence is pro-

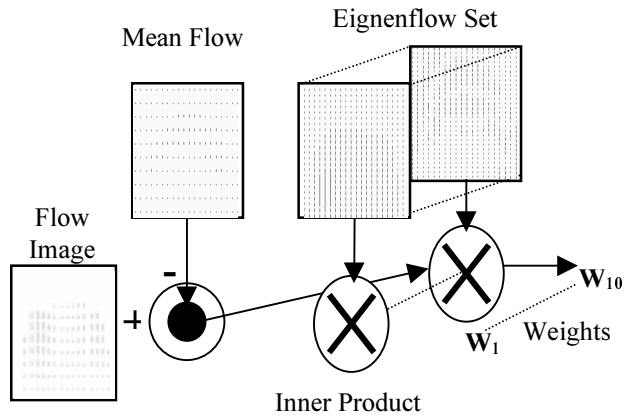


Figure 6. Computation of Vertical Flow Vector

jected onto the flow-based eigenspace by taking the inner product of each element of the eigenflow set, producing a 10-dimensional weighted vector (Figure 6). The ten-dimensional horizontal-flow weighted vector and the ten-dimensional vertical-flow weighted vector are concatenated to form a twenty-dimensional weighted vector for each flow-based frame.

2.3.3. High Gradient Component Analysis

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 (*i.e.*, furrows) of the face image are extracted with a variety of line or edge detectors. After normalization of each image, a 5 x 5 Gaussian filter is used to smooth the image. 3 x 5 (row x column) 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 by

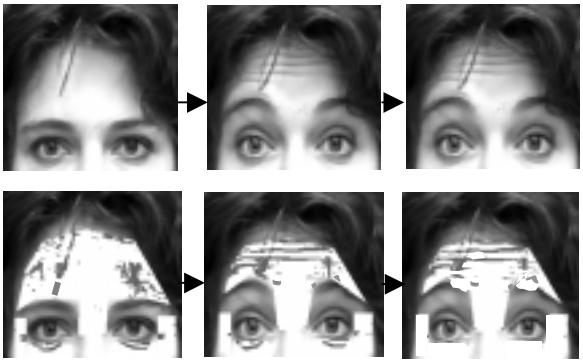


Figure 7. High-Gradient Component Detection

permanent characteristics 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 the threshold value, it is considered a valid high-gradient component produced by facial expression. All other high-gradient components are ignored (see Figure 7). In the former case, the high gradient component (pixel) is assigned a value of one. In the latter case, the pixels are assigned a value of zero.

The forehead (upper face) and lower face regions of the normalized face image are divided into thirteen and sixteen blocks, respectively. 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 also calculated. For upper- and lower-face expression recognition, mean and variance values are concatenated to form 26- and 32-dimensional mean and variance vectors, respectively, for each frame.

4. Experimental Results

Facial behavior was recorded in 100 adults (65% male and 15% African American or Asian, ages 18 to 35 years). Subjects sat directly in front of the camera and performed a series of facial expressions that included single action units (e.g., AU 12, or smile) and combinations of action units (e.g., AU 1+2, or brow raise). Each expression sequence began from a neutral face. Six of the expressions were based on descriptions of prototypic emotions. Action units were coded by certified FACS coders. Inter-observer reliability was quantified with coefficient kappa [19], which corrects for chance agreement. Mean κ was 0.86.

Action units in the brow and mouth regions were selected for analysis if they occurred a minimum of twenty-five times. When an action unit occurred in combination with other action units that may modify its appearance, the combination rather than the single action unit was the unit of analysis. The action units we analyzed represent key components of emotion and related paralinguistic displays. In each region, the actions chosen included similar appearance changes (e.g., brow narrowing due to AU 1+4 versus AU 4 and mouth widening due to AU 12 versus AU 20). For feature-point tracking, results are for 15 action units from 504 image sequences; for dense-flow tracking, three action units from 115 image sequences; and for high-gradient component detection, five action units from 260 image sequences. All data were divided into a training and a test set.

Tables 1 through 3 present test results of the feature-point tracking module of Automated Face Analysis. For all three modules, the average discrimination rate in the test image sequences was 80% or higher, which represents excellent agreement with manual FACS coding.

5. Discussion

Previous studies have used optical flow to discriminate facial expression [11, 17]. Sample sizes in these studies have been small, and with the exception of Bartlett et al. [1], this work has focused on the recognition of molar expressions, such as positive or negative emotion or emotion prototypes (e.g., joy, surprise, fear). We developed and implemented two optical-flow based approaches and one for high-gradient component detection. All three were sensitive to subtle motion in facial displays. Feature-point tracking was tested in 504 images sequences from 100 subjects and achieved a level of precision that was as high as or higher than that of previous studies and comparable to that of human coders. The other two modules were tested in a smaller number of image sequences and showed comparable precision.

Accuracy in the test sets was 80% or higher in each region. The one previous study to demonstrate accuracy for discrete facial actions [1] used extensive pre-processing of image sequences and was limited to upper face action units in ten subjects. In the present study, pre-processing was limited to manual marking with a pointing device in the initial digitized image, facial behavior included action units in both the upper and lower face, and the large number of subjects, which included African-Americans and Asians in addition to Caucasians, provided a sufficient test of how well the initial training analyses generalized to new image sequences. Feature-point tracking demonstrated moderate to high concurrent validity with human FACS coding.

The level of inter-method agreement for action units was

comparable to that achieved in tests of inter-observer agreement in FACS. Moreover, the inter-method disagreements that did occur were generally ones that are common in human coders, such as the distinctions between AU 1+4 and AU 4 or AU 6 and AU 7. These findings suggest that Automated Face Analysis and manual FACS coding are at least equivalent for the type of image sequences and action units analyzed here.

One reason for the lack of 100% agreement at the level of action units is the inherent subjectivity of human FACS coding, which attenuates the reliability of human FACS codes. Two other possible reasons were the restricted number of optical flow feature windows in feature-point tracking and the lack of aggregation across modules. From a psychometric perspective, aggregating the results of each module can be expected to optimize discrimination accuracy.

In human communication, the timing of a display is an important aspect of its meaning. For example, the duration of a smile is an important factor in distinguishing between felt and false positive emotion. Until now, hypotheses about the temporal organization of emotion displays have been difficult to test. Human observers have difficulty in locating precise changes in behavior as well as in estimating changes in expression intensity. Automated Face Analysis by contrast can track quantitative changes on a frame-by-frame basis [24]. Small pixel-wise changes may be measured, and the temporal dynamics of facial expression can be determined.

Two challenges in analyzing facial displays are the problems of controlling for rigid motion and segmenting displays within the stream of behavior. Perspective alignment performs well for mild out-of-plane rotation, but more complex models may be needed for larger out-of-plane motion. Segmenting displays is a focus of current research. Central to the development of Automated Face Analysis is what we learn from substantive applications. These include a cross-cultural study of emotion regulation in European-American, Japanese, and Chinese infants, two separate studies of adult emotion regulation, and a quantitative analysis of facial motion in patients with facial nerve dysfunction. Pending applications include video-conferencing, human-computer interaction, and forensic use.

6. Acknowledgement

This research was supported by grant number R01 MH51435 from the National Institute of Mental Health to Jeffrey Cohn and Takeo Kanade.

7. 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." In D. Touretski, M. Mozer, and M. Hasselmo (Eds.), *Advances in Neural Information Processing Systems,* Vol. 8 , pp. 823-829, MIT, 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*, pp. 12-17, Zurich, 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] J.M. Carroll and J.A. Russell, "Facial Expressions in Hollywood's Portrayal of Emotion*," Journal of Personality and Social Psychology*, Vol. 72, pp. 164-176, 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] C. Darwin, *The Expression of Emotion in Man and Animals*, University of Chicago, 1872/1965.

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

[8] P. Ekman and W.V. Friesen, *Facial Action Coding System,* Consulting Psychologist Press, Palo Alto, CA, 1978.

[9] R.N. Emde, T.J. Gaensbauer, & R.J. Harmon, "Emotional Expression in Infancy: A Biobehavioral Study," *Psychological Issues*, Vol. 10, p. 37, 1996.

[10] I.A. Essa and A. Pentland, "A Vision System for Observing and Extracting Facial Action Parameters*," IEEE Computer Vision and Pattern Recognition*, 1994.

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

[12] J.J. Lien, A.J. Zlochower, J.F. Cohn, C.C. Li, and T. Kanade, "Automatically Recognizing Facial Expressions in the Spatio-Temporal Domain*," Proceedings of the Workshop on Perceptual User Interfaces*, Canada, 1997.

[13] B.D. Lucas and T. Kanade, "An Iterative Image Registration Technique With an Application in Stereo Vision," Seventh International Joint Conference on Artificial Intelligence, pp. 674-679, 1981.

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

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

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

[17] W.E. Rinn, "The Neuropsychology of Facial Expression: A Review of the Neurological and Psychological Mechanisms for Producing Facial Expressions," *Psychological Bulletin*, Vol. 95, pp. 52-77, 1984.

[18] 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.

[19] D. Terzopoulos and K. Waters, "Analysis of Facial Images

Using Physical and Anatomical Models," *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 727-732, Dec. 1990.

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

[21] Y.T. Wu, T. Kanade, J.F. Cohn, and C.C. Li., "Optical Flow Estimation Using Wavelet Motion Model," *Proceedings of the International Conference on Computer Vision (ICCV)*, 1998.

[22] Y. Yacoob and L. Davis, "Computing Spatio-Temporal Representations of Human Faces," *In Proc. Computer Vision and Pattern Recognition*, pp. 70-75, Seattle, WA, June 1994.

[24] J.F. Cohn and A. Zlochower, A. "Computerized analysis of facial expression: Feasibility of automated discrimination," *American Psychological Society*, NY, June 1995.

[25] C.J. Poelman. "The paraperspective and projective factorization method for recovering shape and motion," *Technical Report CMU-CS-95-173*, Carnegie Mellon University, Pittsburgh, PA.

		Automated Face Analysis		
Manual Coding	N	 AU 1+2	 AU 1+4	 AU 4
AU 1+2	43	**.95**	.02	.02
AU 1+4	22	.05	**.86**	.09
AU 4	65	.02	.08	**.91**

Table 1. Agreement Between Feature Point Tracking and Manual Coding in the Brow Region

		Automated Face Analysis		
Manual Coding	N	 AU 5	 AU 6	 AU 7
AU 5	33	**.97**	.00	.03
AU 6	36	.06	**.81**	.14
AU 7	20	.00	.15	**.85**

Table 2. Agreement Between Feature Point Tracking and Manual Coding in the Eye Region

		Automated Face Analysis								
Manual Coding	N	 AU 27	 AU 26	 AU 25	 AU 12	 AU 12+25	 AU 20+25	 AU 15+17	 AU 17+23+24	 AU 9+17
AU 27	29	**.83**	.10	.03	.00	.00	.03	.00	.00	.00
AU 26	24	.25	**.54**	.21	.00	.00	.00	.00	.00	.00
AU 25	22	.00	.05	**.86**	.00	.00	.00	.09	.00	.00
AU 12	18	.00	.00	.00	**.78**	.22	.00	.00	.00	.00
AU 12+25	35	.00	.00	.03	.00	**.86**	.11	.00	.00	.00
AU 20+25	31	.00	.00	.00	.00	.16	**.81**	.03	.00	.00
AU 15+17	36	.00	.00	.00	.00	.00	.00	**.94**	.06	.00
AU 17+23+24	12	.00	.00	.00	.00	.00	.00	.08	**.92**	.00
AU 9+17	17	.00	.00	.00	.00	.00	.00	.00	.00	**1.00**

Table 3. Agreement Between Feature Point Tracking and Manual Coding in the Mouth Region