In press, *Journal of Robotics and Autonomous Systems*

July 28/August 21, 1999

Detection, Tracking, and Classification of Action Units in Facial Expression

James Jenn-Jier Lien[1,2], http://www.cs.cmu.edu/~jjlien

Takeo Kanade[1], http://www.cs.cmu.edu/~tk

Jeffrey F. Cohn[1, 3], http://www.pitt.edu/~jeffcohn/jfc.html and http://www.cs.cmu.edu/~face

Ching-Chung Li[3], http://www.pitt.edu/~ccl

[1] Carnegie Mellon University, [2] Visionics Corporation, [3] University of Pittsburgh

Key words: face expression recognition, optical flow, high-gradient component detection, hidden Markov model, human-computer interaction.

Corresponding author: Jeffrey F. Cohn, PhD, University of Pittsburgh, 604 OEH, 4015 O'Hara Street, Pittsburgh, PA 15260.

1 Abstract

Most of the current work on automated facial expression analysis attempt to recognize a small set of prototypic expressions, such as joy and fear. Such prototypic expressions, however, occur infrequently, and human emotions and intentions are communicated more often by changes in one or two discrete features. To capture the full range of facial expression, detection, tracking, and classification of fine-grained changes in facial features are needed. We developed the first version of a computer vision system that is sensitive to subtle changes in the face. The system includes three modules to extract feature information: dense-flow extraction using a wavelet motion model, facial feature tracking, and edge and line extraction. The feature information thus extracted is fed to discriminant classifiers or hidden Markov models that classify it into FACS action units, the descriptive system to code fine-grained changes in facial expression. The system was tested on image sequences from 100 male and female subjects of varied ethnicity. Agreement with manual FACS coding was strong for the results based on dense-flow extraction and facial feature tracking, and strong to moderate for edge and line extraction.

1 Introduction

Most computer-vision-based approaches to facial expression analysis so far [e.g., 3,26,28] attempt to recognize only a small set of prototypic expressions of emotion, i.e. joy, surprise, anger, sadness, fear, and disgust. This focus follows from the work of Darwin [10] and more recently Ekman [13] and Izard et al.[20] who proposed that basic emotions have corresponding prototypic facial expressions. These expressions typically involve simultaneous changes in facial features in multiple regions of the face. In everyday life, however, such prototypic expressions occur relatively infrequently. Instead, emotion is communicated by changes in one or two discrete features, such as tightening the lips in anger or obliquely lowering the lip corners in sadness [7]. Change in isolated features, especially in the area of the brows or eyelids, is typical of paralinguistic displays; for instance, raising the brows signals greeting. To capture the subtlety of human emotion and paralinguistic communication, automated recognition of fine-grained changes in facial expression is needed.

The Facial Action Coding System (FACS) [14] is a human-observer-based system designed to detect subtle changes in facial features. Viewing videotaped facial behavior in slow motion, trained observers can manually FACS code all possible facial displays, which are referred to as action units (AU). More than 7,000 combinations of action units have been observed [12]. Although Ekman and Friesen [15] proposed that specific combinations of FACS action units represent prototypic expressions of emotion, it should be noted that emotion expressions are not part of FACS; they are coded in separate systems, such as EMFACS [19] or AFFEX [20]. FACS itself is purely descriptive, uses no emotion or other inferential labels, and provides the necessary ground truth with which to describe facial expression.

Several image analysis techniques that have proven useful in recognizing prototypic expressions show potential to recognize facial action units. One technique extracts motion

information by computing the difference in image intensity between successive frames in an image sequence. Using a small number of subjects, Bartlett et al. [1] showed that the motion information encoded in difference images was sufficient to recognize action units in the brow and eye regions. A potential problem with this technique is that it fails to capture pixel-wise correspondence between face images; different facial actions might produce identical patterns of intensity differences.

More precise motion information may be obtained by computing optical flow, which represents the direction and magnitude of motion. Mase [26] and Essa and Pentland [16] observed that increased flow was associated with certain action units in the brow and cheek regions but did not test the specificity of flow to action unit recognition. Recognition of expression from optical flow has remained at the level of prototypic expressions rather than the finer and more objective level of FACS action units. An exception was the study by Bartlett et al. [1] that recognized action units from optical flow in the brow and eye regions. This study used image data from subjects who were experts in recognizing and performing FACS action units, and the image sequences were manually time-warped and intensity graded prior to analysis.

Another technique extracts edges and lines from face images for detecting furrows and wrinkles. Using artificial enhancement by make-up, Terzopoulos and Waters [32] extracted contours that included the brows, eyelids, naso-labial furrows, and lips, but the specificity of these results to action unit recognition was not tested.

A question remains whether these techniques have sufficient and consistent sensitivity to subtle differences in facial displays. As noted above, intensity difference images fail to represent the direction and magnitude of change in pixel intensity, which may degrade action unit recognition. Optical flow methods are intended to overcome this problem but flow estimates

tend to be noisy. To reduce error of measurement, previous work typically aggregates measures of flow [3,26, 30] or disregards small changes that are below an assigned threshold [4], both of which tend to reduce sensitivity to small motion. Edge and line detectors are especially susceptible to noise. They encode permanent lines and furrows as well as transient changes in these features, which are more important to face expression analysis. To detect, track, and classify subtle changes in facial expression, convergent methods, which utilize multiple types of feature information, may provide more robust results.

We developed and implemented the first version of a face image analysis system that uses convergent feature extraction modules to automatically code input face image sequences into FACS action units. The focus of our present study was to compare the relative performance of each module. To ensure sufficient training and test data, we analyzed action units from 100 male and female subjects of varied ethnic background. We also compared the relative performance of two different approaches (discriminant analysis and hidden Markov modeling) to action unit recognition. Each feature extraction module and classification method demonstrated high and consistent sensitivity with different degree to FACS action units.

2 Face Image Analysis System

Our goal is to recognize FACS action units, which are the smallest visibly discriminable changes in facial expression. In the present study, we focus on 15 action units shown in Figure 1 that occur in the upper and lower face and that are common in emotion and paralinguistic communication. For instance, AU 4 is characteristic of negative emotion and mental effort, and AU 1+2 is a component of surprise. The action units chosen are relatively difficult to discriminate because they involve subtle differences in appearance: brow narrowing due to AU 1+4 versus AU 4, eye narrowing due to AU 6 versus AU 7, three separate action unit combinations involving AU 17, and mouth widening due to AU 12 versus AU 20.

Figure 2 depicts the overview of the analysis system. After the input image sequence is aligned (see 2.1), three modules extract feature information. Dense-flow extraction uses a wavelet motion model to track flow across the entire face image. Facial feature tracking tracks a small set of pre-selected features and is less computationally intensive than dense optical flow. By comparing the performance of these two motion-feature extraction modules, we can test whether motion information limited to the brows, eyes, and lips is sufficient for action unit recognition [8,9]. As the third module for feature extraction, high-gradient component detection uses a combination of horizontal, vertical, and diagonal line and edge feature detectors to detect and track changes in standard and transient facial lines and furrows.

Output from each of these three modules is fed to a discriminant classifier [11] or hidden Markov model (HMM) [29]. Discriminant analysis is a classic and well-validated approach in pattern recognition. The HMM performs well in a spatio-temporal domain and has demonstrated validity in gesture recognition [36] and speech recognition [29]. The present study compares the relative strengths of each of these approaches to feature extraction and action unit recognition. The comparative tests are critical to the development of efficient and robust face image analysis systems.

Insert Figures 1 & 2 About Here

2.1 Image alignment

Expressive changes in the face often co-occur with head movement. People raise their head in surprise [6] or turn toward a friend while beginning to smile [21]. Expression may also vary as a result of individual differences in facial proportions [17]. Row A of Figure 3 shows an example of real input image sequence that includes both rigid motion (movement of head position and orientation) and non-rigid motion (facial expression). The infant subject turns his head to the right and pitches his head up, while he smiles (AU 12), opens his mouth (AU 25 and

then AU 26), contracts the orbicularis oculi muscle (AU 6) which raises the cheeks and narrows the eye opening, and raises his brows. In Row B, each image is the difference between the image directly above it and the first image in Row A. White areas indicate apparent motion, which is evidently a mixture of rigid (head motion) and non-rigid motion (facial expression).

Completely removing the effects of head movement from the input image sequence would be very difficult. It may even require a complicated transformation that is dependent on the knowledge of the exact shape of the individual face. When, however, out-of-plane rotation of the head is small, either an affine or a perspective transformation of images can align images so that face position, size, and orientation are kept relatively constant across subjects, and these factors do not interfere significantly with feature extraction. The affine transformation is computationally faster, but the perspective transformation gives more accurate warping for a higher degree of out-of-plane rotation.

Insert Figure 3 About Here

The perspective transformation is the transformation that relates two views of a rigid planar object [31]. Mathematically, if image $I_0(x)$ and image $I(x')$ are two views of a planar object then the two coordinate systems are related by

$$x' = \frac{m_0 x + m_1 y + m_2}{m_6 x + m_7 y + 1}$$
$$y' = \frac{m_3 x + m_4 y + m_5}{m_6 x + m_7 y + 1}$$

where the parameters m_0 through m_7 can be estimated by one of several standard methods [e.g., 34, 31].

Row C of Figure 3 shows the resultant images obtained by perspectively transforming the original images in Row A. Note that the subject's head is warped close to the original orientation. Row D shows the intensity difference between the first frame and each of the

subsequent transformed images in Row C. Comparison between Row B and D demonstrates that the transformed images retains motion due to facial expression (eyebrow raising, eye narrowing, cheek raising, and lip motion) while rigid motion was mostly eliminated.

2.2 Dense Flow Extraction

In FACS, each action unit is anatomically related to contraction of a specific facial muscle. For instance, AU 12 (oblique raising of the lip corners) results from contraction of the zygomaticus major muscle, AU 20 (lip stretch) from contraction of the risorius muscle, and AU 15 (oblique lowering of the lip corners) from contraction of the depressor anguli muscle. Such muscle contractions produce motion in the overlying skin. Optical flow can detect the magnitude and direction of this motion. Wu, Kanade, Cohn, & Li [34] developed a method to compute dense flow using a coarse-to-fine Cai-Wang [5] wavelet representation. The wavelet motion model represents motion vectors by a linear combination of hierarchical basis functions. The Cai-Wang basis functions directly transform any function into wavelet coefficients from coarse to fine scales. This differs from the conventional usage of the wavelet transform, which proceeds from fine to coarse for decomposition and then from coarse to fine for reconstruction. Referring to [34] for details, the Cai-Wang wavelet-based dense flows are sensitive to small motion and stable in a smoothly textured region. Computation is relatively slow. On an SGI-Irix workstation, it takes approximately 20 minutes of processing per frame pair where a frame consists of 640 x 490 pixels. Figure 4 shows an example of dense flow extraction using this method.

Insert Figure 4 About Here

Dense flow extraction produces a (u, v)-vector field for each frame; if the image size is nxm, we can consider the flow field as two nxm-dimensional vectors by concatenating all horizontal and vertical motion vectors individually. To reduce the number of dimensions, we perform principal components analysis (PCA) and then project each vector field onto the

component axes. Because action units in the upper and lower face are relatively independent, we perform this process separately on dense flow for each region.

The upper face, defined as the area above the infra-orbital furrows inclusive, is 110 x 240 pixels. Within this region, the first 10 principal components for horizontal flow and the first 10 principal components for vertical flow accounted for over 90% of the variation in dense flow. Projecting the upper-face vector fields onto these principal component axes resulted in two 10-dimensional vectors, which we concatenated to form a 20-dimensional feature vector to represent the flow in the upper-face region in each frame.

In the lower face, the first 15 horizontal and 15 vertical principal components accounted for over 90% of the variation in dense flow. Similarly to the upper-face region, by projecting the vector fields onto these principal component axes, we obtain a 30-dimensional feature vector for the lower-face region in each frame.

2.3 Facial-feature tracking

Not only is obtaining dense flow for the whole face image computationally intensive, but also the features like PCA coefficients represent aggregate properties of facial motions. It may be more advantageous or appropriate to compute features of motion for a small set of localized facial features. Previous work suggests that the motion in the brow, eyes, and lips is sufficient for the recognition of many action units [8]. We selected 38 points in these facial feature areas: 6 points around the contours of the brows, 8 around the eyes, 14 around the nose, and 10 around the mouth. In the first frame, these points are manually marked. Comparison of feature-point markings by two operators showed that the mean inter-observer error in manual marking is 2.29 and 2.01 pixels in the horizontal and vertical dimensions, respectively. The Pearson correlation for inter-observer reliability is 0.97 and 0.93, respectively. Automated marking of the features in the initial frame has been partially implemented in recent work [33].

The movement of a feature is tracked automatically. The Lucas-Kanade algorithm [25] is a standard technique to estimate the feature-point movement efficiently with sub-pixel accuracy when displacement is small. Given an n x n feature region R in frame $I_t(x,y)$ in the sequence, the displacement vector $\boldsymbol{d} = (d_x, d_y)$ in the next frame $I_{t+1}(x,y)$ is obtained by minimizing the residual $E(\boldsymbol{d})$;

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

The Lucas-Kanade algorithm gives $\boldsymbol{d} = (d_x, d_y)$ as the solution of

$$\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} d_x \\ d_y \end{bmatrix} = \begin{bmatrix} e_x \\ e_y \end{bmatrix}$$

where

$$H_{11} = \sum \left(\left(\frac{\partial I_t}{\partial x} \right)^2 + \left(\frac{\partial I_{t+1}}{\partial x} \right)^2 \right)$$

$$H_{12} = H_{21} = \sum \left(\left(\frac{\partial I_t}{\partial x} \right) \left(\frac{\partial I_t}{\partial y} \right) + \left(\frac{\partial I_{t+1}}{\partial x} \right) \left(\frac{\partial I_{t+1}}{\partial y} \right) \right)$$

$$H_{22} = \sum \left(\left(\frac{\partial I_t}{\partial y} \right)^2 + \left(\frac{\partial I_{t+1}}{\partial y} \right)^2 \right)$$

$$e_x = \sum \left(I_t - I_{t+1} \right) \left(\frac{\partial I_t}{\partial x} + \frac{\partial I_{t+1}}{\partial x} \right)$$

$$e_y = \sum \left(I_t - I_{t+1} \right) \left(\frac{\partial I_t}{\partial y} + \frac{\partial I_{t+1}}{\partial y} \right)$$

The region size used in the algorithm is 13 by 13. Although the original idea of the algorithm assumes a small displacement, when combined with an iterative image pyramid [28], rapid and large displacements of up to 100 pixels (such as those found in sudden mouth opening) can be

tracked robustly while maintaining sensitivity to subtle (sub-pixel) facial motion [22,23]. On a 300 MHz Pentium II computer, tracking 38 features requires approximately 1 second per frame pair, which is a significant improvement in processing time over dense-flow extraction.

Insert Figure 5 About Here

Figure 5 shows an example of facial feature tracking. The subject's face changes from neutral (AU 0) to brow raise (AU 1+2), eye widening (AU 5) and jaw drop (AU 26), both of which are characteristic of surprise. Line segments trailing from the features represent trajectories of features during the image sequence.

The displacements of tracked points, when concatenated, forms feature vectors. In the analyses of the brow region, the measurements consist of the horizontal and vertical displacements of the 6 feature points around the brows (3 on the upper contour of each brow; see Figure 4). In the analyses of the eye region, the measurements consist of the horizontal and vertical displacements of the 8 feature points around the eyes. In analyses of the mouth region, the measurements consist of the horizontal and vertical displacements of the 10 feature points around the mouth and 4 on either side of the nostrils because of the latter's relevance to the action of AU 9. Therefore, each measurement is represented by a *2p* dimensional vector by concatenating *p* feature displacements (where $p = 32$); that is $\boldsymbol{D} = (\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots\ldots, \boldsymbol{d}_p) = (d_{1x}, d_{1y}, d_{2x}, d_{2y}, \ldots d_{px}, d_{py})$.

2.4 High-Gradient Component Analysis

Facial motion produces transient wrinkles and furrows perpendicular to the motion direction of the activated muscle. These transient features provide information relevant to the recognition of action units. Contraction of the *corrugator* muscle, for instance, produces vertical furrows between the brows, which is coded in FACS as AU 4, while contraction of the medial portion of the *frontalis* muscle (AU 1) causes horizontal wrinkling in the center of the forehead.

Some of these lines and furrows may become permanent with age. Permanent crows-feet wrinkles around the outside corners of the eyes, which is characteristic of AU 6 when transient, are common in adults but not in infants [37]. When lines and furrows become permanent facial features, contraction of the corresponding muscles produces changes in their appearance, such as deepening or lengthening.

To detect these features, we apply gradient filters (first-order derivative) of different orientations in several facial regions. Figure 6 shows 3 x 5 horizontal and 5 x 3 vertical filters and 5 x 5 diagonal filters used for forehead and nasolabial furrow regions. Similarly, 5 x5 diagonal filters are used for the region of the nasolabial furrow and 3 x 3 filters for small furrows around the lips and chin. Prior to applying these gradient filters, a 5 x 5 Gaussian filter is used to smooth the image data. On a 300 MHz Pentium PC, the processing time is approximately four frames per second.

Insert Figure 6 About Here

The high-gradient components produced by transient skin or feature deformations need to be separated from permanent characteristics of the individual's face. For this purpose, the output of the gradient filters for the current frame is subtracted from that of the first frame (Figure 7).

Insert Figure 7 About Here

The results of high-gradient component detection are turned into feature vectors by the following method. The forehead and lower face regions of the normalized face image are divided into sixteen blocks each. The mean and covariance values of the gradient filter output are calculated for each block. For upper- and lower-face expression recognition, these mean and variance values are concatenated to form a 32-dimensional vector for each frame.

3 Action Unit Recognition

The results of feature extraction from an image sequence of facial expression by the three modules are now represented by feature vector sequences. We want to classify each sequence into one of the predetermined action units. We use discriminant analysis and Hidden Markov Models (HMM) as decision making mechanisms. To classify the facial feature tracking data, we use both discriminant analysis and HMM. For the data from dense flow and the data from high-gradient component detection, we use only HMM. For both discriminant analysis and HMM, the data were divided into training and test sets.

3.1 Classification by discriminant analysis

Discriminant analysis of feature vectors computes dimensions along which phenomena differ and obtain classification functions that predict class membership. Discriminant analysis was used in the current study only with facial feature tracking.

The discrimination among action units is done by computing and comparing the *a posteriori* probabilities of action units AU_k; given measurement $\boldsymbol{D}$. That is,

$$\boldsymbol{D} \in AU_k \quad \text{if } p(AU_k \mid \boldsymbol{D}) > p(AU_j \mid \boldsymbol{D}) \qquad \text{for all } j \neq k$$

where

$$p(AU_i \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid AU_i)\,p(AU_i)}{p(\mathbf{D})} = \frac{p(\mathbf{D} \mid AU_i)\,p(AU_i)}{\sum_{j=1}^{K} p(\mathbf{D} \mid AU_j)\,p(AU_j)}$$

and K denotes the total number of action units under consideration. The Bayes discriminant function between AU_i and AU_j can be expressed by the log-likelihood ratio

$$f_{ij}(\mathbf{D}) = \log \frac{p(\mathbf{D} \mid AU_i)}{p(\mathbf{D} \mid AUj)}$$

and

$$\boldsymbol{D} \in \text{AU}i \qquad \text{if} \qquad f_{ij}(\text{D}) > \log \frac{p(AUj)}{p(AUi)}$$

Each $p(\boldsymbol{D}|\text{AU}_i)$ is assumed to be a multivariate Gaussian distribution $N(\boldsymbol{m}_i, \boldsymbol{å}_i)$, where the mean $\boldsymbol{m}_i$ and the covariance matrix $\boldsymbol{å}_i$ are estimated by the sample mean and sample covariance matrix of the training data. Under the Gaussian assumption, this discriminant function is a quadratic discriminant function in general; but if the covariance matrices $\boldsymbol{å}_i$ and $\boldsymbol{å}_j$ are equal, it is reduced to a linear discriminant function. *A priori* probabilities $p(\text{AU}_i)$'s are assumed to be equal, since we wish to generalize our results to other samples in which the relative frequency of action units is not known.

3.2 Classification by hidden Markov models

A hidden Markov model describes the statistical behavior of a process that generates time series data having certain statistical characteristics. The motion vector sequences extracted in the face analysis system can be encoded via vector quantization [24] into sequences of a finite set of symbols. Such symbol sequences representing various action units are modeled by discrete hidden Markov models. A discrete HMM has N states $\{s_1, s_2, \ldots s_N\}$ and M observation symbols $\{o_1, o_2, \ldots o_M\}$. At time t, the HMM occupies a state s_t and may undergo a state transition from the state $s_t=i$ to a state $s_{t+1}=j$ at time $t+1$ with a state transition probability $a_{ij}=P(s_{t+1}=j|s_t=i)$. Associated with each state is a set of observation symbols $o_t=k$, $(k=1,2,\ldots,M)$ with their respective observation probabilities $b_i(k)=P(o_t=k|s_t=i)$. Starting from an initial state $s_1=i$ with probability $\boldsymbol{p}_i=P(s_1=i)$, the process undergoes a sequence of state transitions over a time duration T and generates an observation symbol sequence $O=(o_1,o_2,\ldots,o_T)$ with a certain probability. Let $A=\{a_{ij}\}$, $(i,j=1,2,\ldots,N)$, denote the set of state transition probabilities of the process, $B=\{b_j(k)\}$,

($j=1,2,\ldots,N$; $k=1,2,\ldots,M$), denote the set of observation symbol probabilities, and $\boldsymbol{P}=\{\boldsymbol{p}_i\}$, ($i=1,2,\ldots,N$), denote the set of initial state probabilities, where

$$\sum_{j=1}^{N} a_{ij} = 1, \qquad \sum_{j=1}^{N} b_j(k) = 1, \qquad \sum_{i=1}^{N} \boldsymbol{p}_i = 1,$$

The HMM is thus specified by a triplet $\boldsymbol{l}=(\boldsymbol{P},A,B)$. The model parameters can be trained from a set of training symbol sequences by applying the Baum-Welch algorithm [29]. Given a model λ, an observation symbol sequence O may be generated from one or more state sequences $S=(s_1,s_2,\ldots,s_T)$, each with its own probability $P(S|\boldsymbol{l})=\delta_{s_1}\prod_{t=2}^{T} a_{s_{t-1}s_t}$. The probability of the symbol sequence O observed under a specific state sequence S is $P(O|S,\boldsymbol{l})=\prod_{t=1}^{T} b_{s_t}(o_t)$. Thus, the probability of the observation symbol sequence O generated by the given model λ is given by

$$P(O|\boldsymbol{l}) = \sum_{all\ S} P(O \mid S, \boldsymbol{l}), P(S \mid \boldsymbol{l})$$

$$= \sum_{all\ S} \delta_{s_1} b_{s_1}(o_1) \prod_{t=2}^{T} a_{s_{t-1}s_t} \quad b_{s_t}(o_t)$$

By defining a forward variable $\boldsymbol{a}_t(i)$ that is the probability of the observation sequence $O_1^t=(o_1,\ldots o_t)$ up to time t and state i at time t,

$$\boldsymbol{a}_t(i)=P(O_1^t, s_t=i|\boldsymbol{l}),$$

$P(O|\boldsymbol{l})$ can be efficiently computed by a recursive algorithm:

$$\boldsymbol{a}_1(i) = \boldsymbol{p}_i b_i(o_1), \qquad (i=1,\ldots,N)$$

$$\boldsymbol{a}_{t+1}(j) = \sum_{i=1}^{N} \boldsymbol{a}_t(i) a_{ij} b_j(o_{t+1}), \qquad (t=1,\ldots,T\text{-}1;\ j=1,\ldots,N)$$

and $$P(O|\lambda) = \sum_{j=1}^{N} \boldsymbol{a}_T(j)$$

For each class of AU, one HMM model is constructed. Let these models be designated as $\boldsymbol{l}_i$, *(i=1,…,C).* A given observation sequence *O* may be evaluated for the probabilities of individual models $P(\boldsymbol{l}_i|O)$, and be assigned to the model with the highest probability. Applying the Bayes rule and assuming equal *a priori* probability of each model $P(\boldsymbol{l}_i)$, the pattern classification can be performed by maximizing the likelihood function $P(O|\boldsymbol{l}_i)$ as discussed above.

As an example, consider the sequences of displacement vectors in the upper face region (from 6 features in the brow region) extracted by the facial feature tracking method. The displacement vectors are vector-quantized into 16 coded symbols (*M*=16). A separate second-order 3-state left-to-right HMM, like the one shown in Figure 8(a), was constructed for each of the three action units, AU4, AU1+4, and AU1+2 [22,23]. For example, a sample symbol sequence (2,2,2,2,2,2,9,9,9,1,1,1,1,1,1,1,1,1) of AU4 must be generated by its HMM with a certain probability. Using the selected training sequence and applying the Baum-Welch reestimation algorithm, we obtained HMM parameters *A* and *B* for each HMM, λ_{AU4}, λ_{AU1+4} and λ_{AU1+2}. Because the left-to-right HMM is used, the initial state probabilities are $\boldsymbol{p}_1$=1, $\boldsymbol{p}_2$=$\boldsymbol{p}_3$=*0*. Each element in *A* and *B* is initialized with a uniformly distributed small value and normalized to meet the statistical constraints. After each iteration, the probability values are smoothed and renormalized to meet the statistical constraints. As an example, the following probabilities for λ_{AU4} were obtained:

a_{11}=0.627,	a_{22}=0.705,	a_{33}=1.0
a_{12}=0.253,	a_{23}=0.295,	a_{13}=0.120
$b_1(2)$=1.0	$b_2(2)$=0.999	$b_3(1)$=0.309
$b_2(9)$=0.001	$b_3(9)$=0.691	and all other $b_j(k)$=0.0

Similarly, a third order 4-state left-to-right HMM as shown in Figure 8(b) has been trained for each AU in the lower face region. These HMM models each represent the most likely AUs and AU combinations, and are used to evaluate the encoded feature data for automatic recognition of expression units.

Insert Figures 8a and 8b About Here

4 Experimental Results

4.1 Data set used

Facial behavior in 100 adults (65% female, 35% male, 85% European-American, 15% African-American or Asian, ages 18 to 35 years) was recorded. 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. Each frame in the sequence was digitized into a 640 by 490 pixel array with 8-bit precision for gray scale values. For each sequence, action units were coded by a certified FACS coder. Seventeen percent of the data were coded by a second certified FACS coder for comparison. Agreement between the two coders was quantified with coefficient kappa ***(k)***, which is the proportion of agreement above what would be expected to occur by chance [18]. The mean kappa ***(k)*** for inter-coder agreement was 0.86.

Action units that are important to the communication of emotion and that occurred a minimum of 25 times were selected for analysis. This frequency criterion ensured sufficient data for training and testing. When an action unit occurred in combination with other action units that may modify the single AU's appearance, the combination of AUs, rather than the single action unit, was the unit of analysis.

For facial-feature tracking with discriminant analysis, we used 872 samples of 15 action units or action unit combinations that occurred in 504 image sequences. The samples were randomly divided into training and cross-validation sets. However, if an action unit occurred in more than one image sequence from the same subject, all of the samples of that action unit by that subject were assigned to the training set. Thus, for each action unit, samples from the same subject belonged exclusively either to the training or the cross-validation set but not both. This strict criterion ensured that the training and the cross-validation set were uncorrelated with respect to subjects for each action unit, and thus that what was recognized by our method was the action unit rather than the subject.

For feature-point tracking with HMM and dense-flow extraction with HMM, we used samples of 9 action units or action unit combinations that occurred in 240 image sequences. For high-gradient component detection with HMM, we used samples of 5 action units or action unit combinations that occurred in 220 image sequences of 85 subjects. The samples were randomly divided into training and cross-validation sets in a similar manner as above. For all methods, the results presented below are for test sets only.

Insert Tables 1 Through 5 About Here

4.2 Classification results

The agreement of action unit recognition between manual FACS coding and each method was quantified. In addition to the percentage of correct recognition, we present coefficient kappa (κ), which is the proportion of agreement above what would be expected to occur by chance [18]. In preliminary analyses, subjects' race and gender did not affect the classification accuracy and therefore were not included as factors in the discriminant analyses, HMMs, and classification results reported below.

In the brow region, three action units or action unit combinations (AU 1+2, AU 1+4, and AU 4) were analyzed (Table 1). Ninety-two percent (92%) were correctly classified by dense-flow extraction with HMM, 91% by facial-feature tracking with discriminant analysis, 85% by facial feature-tracking with HMM, and 88% by high-gradient component detection with HMM. The corresponding kappa coefficients were $\kappa = 0.87, 0.87, 0.78$, and 0.82, respectively. Accuracy was higher for dense-flow extraction with HMM and facial-feature tracking with discriminant analysis than for facial-feature tracking with HMM and high-gradient component detection with HMM.

In the eye region, analysis was limited to facial-feature tacking with discriminant analysis (Table 2). Three action units (AU 5, AU 6, and AU 7) were classified with 88% accuracy (corresponding $\kappa = 0.82$). The disagreements that occurred were primarily between AU 6 and AU 7, which are difficult to discriminate for manual FACS coders as well.

In the mouth region, 6 action units were analyzed by dense-flow extraction with HMM, 9 action units by facial-feature tracking with discriminant analysis, 6 action units by facial-feature tracking with HMM, and two action units by high-gradient component detection (Tables 4 & 5). Accuracy was above 80% for each module. The percentage correctly classified by dense-flow extraction with HMM was 92%, $\kappa = 0.91$. The percentage correctly classified by facial-feature tracking with discriminant analysis was 81%, $\kappa = 0.79$, and by facial-feature tracking with HMM was 0.88, $\kappa = 0.86$. The percentage correctly classified by high-gradient component detection with HMM was 81%, $\kappa = .60$.

Discussion

Previous studies have primarily used optical flow to recognize facial expression [1,16,26,30,35]. Sample sizes in these studies have been small, and with the exception of

Bartlett et al. [1], they have 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 three convergent modules for feature extraction, dense-flow extraction, facial feature tracking, and high-gradient component detection. All three were found sensitive to subtle motion in facial displays.

For each module, accuracy in the test sets was 80% or higher. The one previous study to demonstrate accuracy for discrete facial actions [1] used extensive manual preprocessing of the image sequences and was limited to upper face action units (i.e., ones in the brow and eye regions) with only twenty subjects. Our present study dealt with action units in both the upper and lower face, and tested with the large number of subjects, which included African-Americans and Asians in addition to European-Americans, thus providing a sufficient test of how well the initial training analyses generalized to new image sequences. Also, pre-processing was limited to manual marking in the initial digitized image. The level of agreement between the face image analysis system and manual FACS coding was comparable to that achieved between manual FACS coders. The inter-method disagreements 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 the face image analysis system is close to manual FACS coding for the type of image sequences and action units analyzed here.

Several factors may account for the lack of 100% agreement for action units recognition. The restricted number of features in facial feature tracking and the lack of integration across modules may be attributed to reduced accuracy. From a psychometric perspective, integrating the results of two or three modules can be expected to optimize recognition accuracy. Also it should be noted that there is the inherent subjectivity of human FACS coding, which attenuates the reliability of human FACS codes.

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 unfelt 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. The computerized face image analysis system can track quantitative changes on a frame-by-frame basis. Sub-pixel changes may be measured, and the temporal dynamics of facial expression could be determined.

Two challenges in analyzing facial images are the problems of compensating head motion and the need to segment facial displays within the stream of behavior. Perspective alignment performs well for mild out-of-plane rotation [34], but better models are needed for larger out-of-plane motion. Segmenting facial displays within the stream of behavior is a focus of future research.

In summary, the face image analysis system demonstrated concurrent validity with manual FACS coding. In the test set, which included subjects of mixed ethnicity, average recognition accuracy for 15 action units in the brow, eye, and mouth regions was 81% to 91%. This is comparable to the level of inter-observer agreement achieved in manual FACS coding and represents advancement over the existing computer-vision systems that can recognize only a small set of prototypic expressions that vary in many facial regions. With continued development, the face image analysis system will reduce or eliminate the need for manual coding in behavioral research and contribute to the development of multi-modal computer interfaces that can understand human emotion and paralinguistic behavior.

6 Acknowledgement

This research was supported by grant number R01 MH51435 from the National Institute of Mental Health.

7 References

[1] M.S. Bartlett, J.C. Hager, P. Ekman, T.J. Sejnowski, Measuring facial expressions by computer image analysis. Psychophysiology, 36 (1999), 253-264.

[2] J.N. Bassili, The role of facial movement and the relative importance of upper and lower areas of the face, Journal of Personality and Social Psychology 37 (1979) 2049-2059.

[3] M.J. Black, Y. Yacoob, Recognizing facial expressions under rigid and non-rigid facial motions, International Workshop on Automatic Face and Gesture Recognition, Zurich, 1995, 12-17.

[4] M.J. Black, Y. Yacoob, A.D. Jepson, D.J. Fleet, Learning Parameterized Models of Image Motion, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1997, 561-567.

[5] W. Cai, J. Wang, Adaptive multiresolution collocation methods for initial boundary value problems of nonlinear PDEs. SIAM Journal of Numerical Analysis, 33, (1996) 937-970.

[6] L.A. Camras, L. Lambrecht, L., G.F. Michel. Infant "surprise" expressions as coordinative motor structures. Journal of Nonverbal Behavior, 20, (1996), 183-195.

[7] J.M. Carroll, J.A. Russell, Facial Expressions in Hollywood's Portrayal of Emotion, Journal of Personality and Social Psychology 72, (1997) 164-176.

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

[9] J.F. Cohn, J.J. Lien, T. Kanade, W. Hua, & A. Zlochower, Beyond prototypic expressions: Discriminating subtle changes in the face, Proceedings of the IEEE Workshop on Robot and Human Communication (ROMAN'98), 1998, 33-39, Takamatsu, Japan.

[10] C. Darwin, The Expression of Emotion in Man and Animals, University of Chicago, 1872/1965.

[11] R.O. Duda, P.E. Hart, Pattern Classification and Analysis, NY: Wiley, 1973.

[12] P. Ekman, Methods for measuring facial action, In K.R. Scherer & P. Ekman (Eds.), Handbook of Methods in Nonverbal Behavior Research, Cambridge: Cambridge University 1982, 45-90.

[13] P. Ekman, Facial expression and emotion, American Psychologist 48 (1993) 384-392.

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

[15] P. Ekman, W.V. Friesen, Facial Action Coding System Investigator's Guide, Consulting Psychologist Press, Palo Alto, CA, 1978.

[16] I.A. Essa, A. Pentland, A Vision system for observing and extracting facial action parameters, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1994.

[17] L.G. Farkas, I.R. Munro, Anthropometric Facial Proportions in Medicine, Springfield, Illinois: Charles C. Thomas, 1987.

[18] J.L. Fleiss, Statistical Methods for Rates and Proportions, NY: Wiley, 1981.

[19] W.V. Friesen, P. Ekman, EMFACS-7: Emotional Facial Action Coding System, Unpublished manuscript, University of California at San Francisco, 1983.

[20] C. Izard, L. Dougherty, E. Hembree, A System for Identifying Emotion by Holistic Judgments (AFFEX), Newark, DE: University of Delaware: Instructional Resources Center, 1983.

[21] R.E. Kraut, R. Johnson, Social and emotional messages of smiling: An ethological approach. Journal of Personality and Social Psychology, 37 (1979), 1539-1553.

[22] J.J. Lien, T.K. Kanade, A.Z. Zlochower, J.F. Cohn, C.C. Li, Subtly different facial expression recognition and expression intensity estimation, Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 853-859, Santa Barbara: CA, June 23-25, 1998. Available at http://www.cs.cmu.edu/~jjlien

[23] J.J. Lien, Automatic recognition of facial expressions using hidden Markov models and estimation of expression intensity. Technical Report CMU-R1-TR-31, Carnegie Mellon University, 1998. Available at http://www.cs.cmu.edu/~jjlien

[24] Y. Linde, A. Buzo, R. Gray, An algorithm for vector quantizer design, IEEE Tranactions on Communications, COM-28 (1) (1980).

[25] B.D. Lucas, T. Kanade, An iterative image registration technique with an application in stereo vision, Seventh International Joint Conference on Artificial Intelligence, 1981, 674-679.

[26] K. Mase, Recognition of facial expression from optical flow, IEICE Transactions, E74 (1991) 3474-3483.

[27] C. Padgett, G.W. Cottrell, B. Adolphs, Categorical perception in facial emotion classification, Proceedings of The Cognitive Science Conference, 18 (1996) 249-253.

[28] 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, 1995.

[29] L. Rabiner, B.H. Juang, Fundamentals of speech recognition, Englewood Cliffs, NJ: Prentice Hall, 1993.

[30] M. Rosenblum, Y. Yacoob, 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.

[31] R. Szeliski and S.B. Kang, Recovering 3D shape and motion from image streams using nonlinear least squares. Journal of Visual Communication and Image Representation. Vol 5, No.1 (1994), 10-28.

[32] D. Terzopoulos and K. Waters, Analysis of facial images using physical and anatomical models, Proceedings of the IEEE International Conference on Computer Vision, December 1990, 727-732.

[33] Y. Tian, Summary of the facial expression analysis system, Technical reports, Robotics Institute, Carnegie Mellon University, December 1998.

[34] Y.T. Wu, T. Kanade, J.F. Cohn, C.C. Li., Optical flow estimation using wavelet motion model, Proceedings of the IEEE International Conference on Computer Vision, 1998, 992-998.

[35] Y. Yacoob, L. Davis, Computing spatio-temporal representations of human faces, In Proc. Computer Vision and Pattern Recognition, Seattle, WA, June 1994, 70-75.

[36] J. Yamato, J. Ohya, K. Ishii, Recognizing human action in time-sequential images using hidden Markov models, Proceedings of the IEEE International Conference on Computer Vision, 1992, 379-385.

[37] A.J. Zlochower, J.F. Cohn, J.J. Lien, T. Kanade, Automated face coding: A computer vision based method of facial expression analysis in parent-infant interaction, International Conference on Infant Studies, Atlanta, Georgia, April 1998.

Table 1. Proportion of Agreement Between Each Module and Manual Coding in the Brow Region

Manual Coding		AU 1+2 	AU 1+4 	AU 4 
		Dense-Flow Extraction with HMM		
AU 1+2	(21)	1.00	.00	.00
AU 1+4	(15)	.07	.80	.13
AU 4	(22)	.00	.09	.93
		Facial-Feature Tracking with Discriminant Analysis		
AU 1+2	(43)	.95	.05	.00
AU 1+4	(19)	.00	.74	.26
AU 4	(32)	.00	.03	.97
		Facial-Feature Tracking with HMM		
AU 1+2	(25)	.92	.08	.00
AU 1+4	(25)	.08	.76	.16
AU 4	(25)	.00	.12	.88
		High-Gradient Component Detection		
AU 1+2	(50)	.86	.14	.00
AU 1+4	(30)	.17	.80	.03
AU 4	(45)	.00	.04	.96

Note. In this and the following tables, the number of samples of each AU appears in parentheses. Only test data are presented.

Table 2. Proportion of Agreement Between Facial-Feature Tracking with Discriminant Analysis and Manual Coding in the Eye Region

Manual Coding		AU 5 	AU 6 	AU 7 
		Facial-Feature Tracking with Discriminant Analysis		
AU 5	(28)	.93	.00	.07
AU 6	(33)	.00	.82	.18
AU 7	(14)	.00	.07	.93

Table 3. Proportion of Agreement Between Dense-Flow Extraction with HMM and Facial-Feature Tracking with HMM and Manual Coding in the Mouth Region

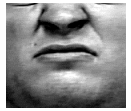
Manual Coding		AU12 	AU6+12+25 	AU20+25 	AU15+17 	AU17+23+24 	AU9+17 
				Dense-Flow Extraction with HMM			
AU12	(15)	1.00	.00	.00	.00	.00	.00
AU6+12+25	(15)	.00	.87	.13	.00	.00	.00
AU20+25±16	(15)	.00	.13	.87	.00	.00	.00
AU 15+17	(15)	.00	.00	.00	.93	.07	.00
AU17+23+24	(15)	.00	.00	.00	.13	.87	.00
AU 9+17±25	(15)	.00	.00	.00	.00	.00	1.00
				Facial-Feature Tracking with HMM			
AU12	(25)	1.00	.00	.00	.00	.00	.00
AU6+12+25	(25)	.00	.84	.16	.00	.00	.00
AU20+25±16	(25)	.00	.20	.80	.00	.00	.00
AU 15+17	(25)	.00	.00	.00	.88	.12	.00
AU17+23+24	(25)	.00	.00	.00	.00	.92	.08
AU 9+17±25	(25)	.00	.00	.00	.04	.12	.84

Table 4. Proportion of Agreement Between Facial-Feature Tracking with Discriminant Analysis and Manual Coding in the Mouth Region

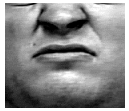
		AU 27	AU 26	AU25	AU12	AU12+25	AU20+25	AU15+17	AU17+23+24	AU9+17
Manual Coding										
		Facial-Feature Tracking with Discriminant Analysis								
AU 27	(29)	.79	.10	.03	.00	.00	.07	.00	.00	.00
AU 26	(20)	.30	.52	.18	.00	.00	.00	.00	.00	.00
AU 25	(22)	.00	.14	.73	.00	.00	.00	.14	.00	.00
AU 12	(18)	.00	.00	.00	.83	.17	.00	.00	.00	.00
AU 12+25	(35)	.00	.00	.03	.00	.81	.17	.00	.00	.00
AU20+25±16	(26)	.00	.03	.00	.00	.08	.89	.00	.00	.00
AU 15+17	(36)	.00	.00	.00	.00	.00	.03	.92	.06	.00
AU17+23+24	(12)	.00	.00	.00	.08	.00	.00	.00	.92	.00
AU 9+17±25	(17)	.00	.00	.00	.00	.00	.00	.00	.00	1.00

Table 5. Proportion of Agreement Between High-Gradient Component Detection with HMM and Manual Coding in the Mouth Region

Manual Coding		AU 12+25	AU 9+17
		High-Gradient Detection	
			
AU 12+25	(50)	.84	.16
AU 9+17	(30)	.33	.77

.

Figure Captions

Figure 1. Examples of FACS Action Units.

Figure 2 Overview of Face Image Analysis System.

Figure 3. Example of perspective alignment. Row A shows the original images. Row B shows the difference between the first and each subsequent image in Row A. Row C shows the perspective alignment of the original images. Row D shows the corresponding difference images.

Figure 4. Example of dense flow extraction.

Figure 5. Example of manually located features (leftmost image) and automated tracking (two images on the right). The subject's expression changes from neutral (AU 0) to brow raise (AU 1+2) and mouth open (AU 26).

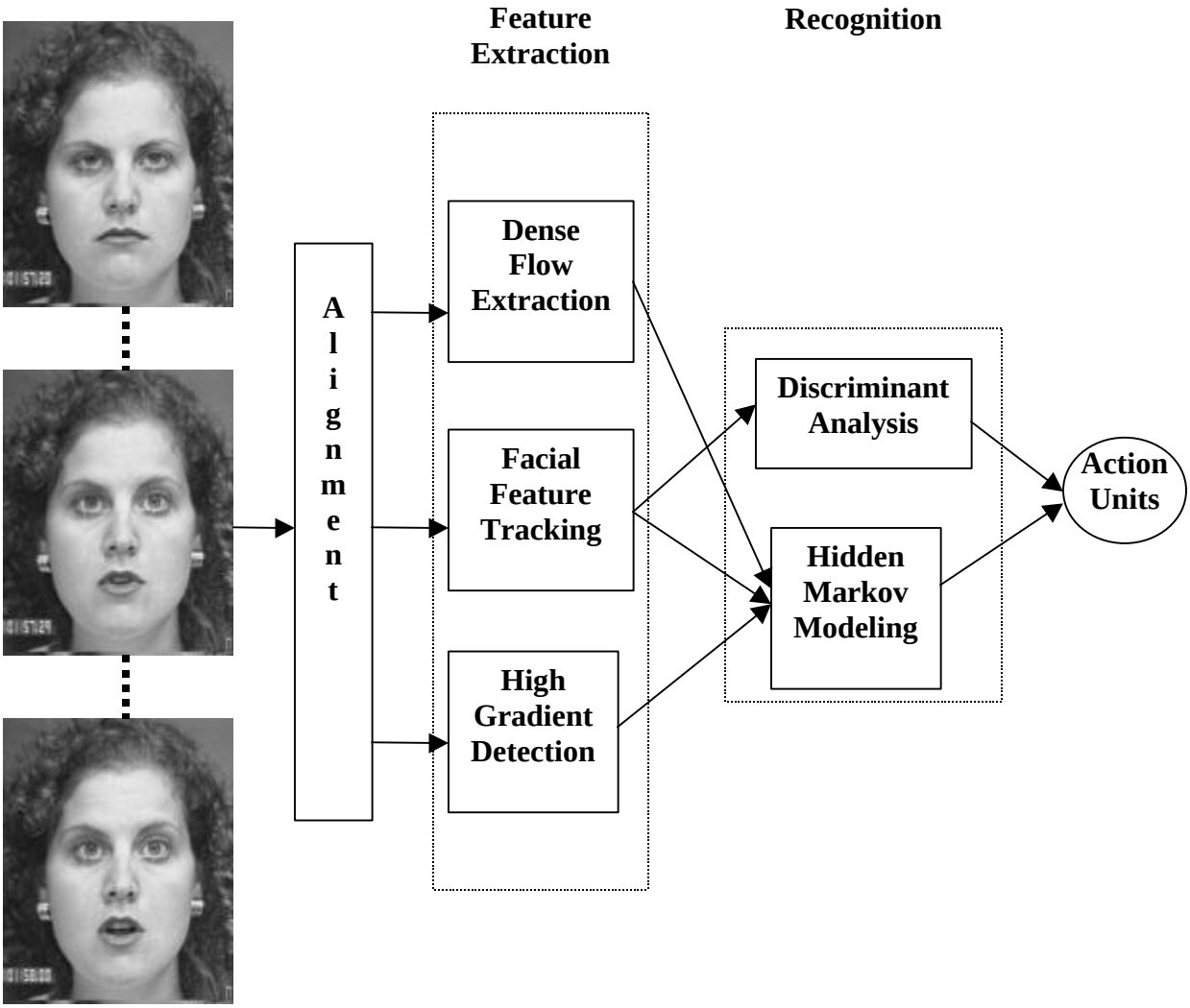
Figure 6. Method of furrow detection.

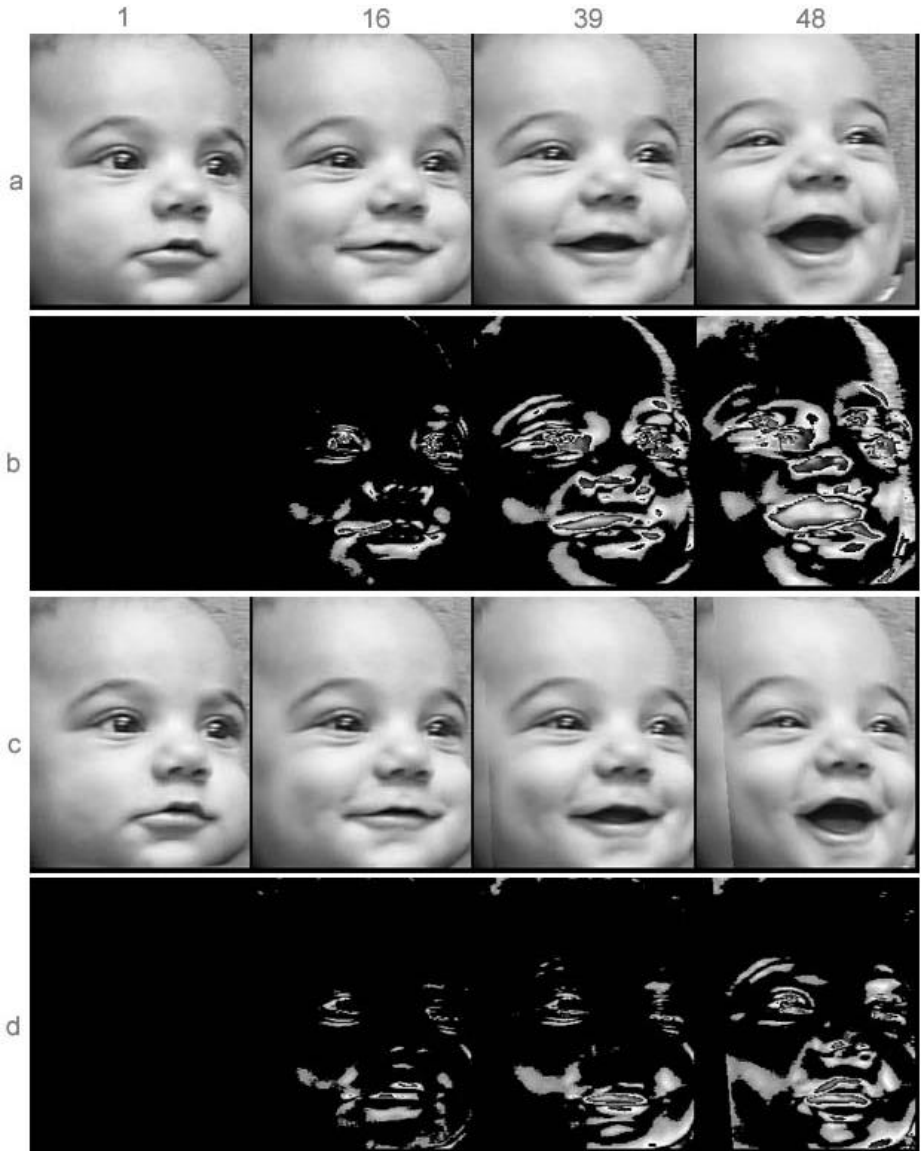
Figure 7. Method of thresholding.

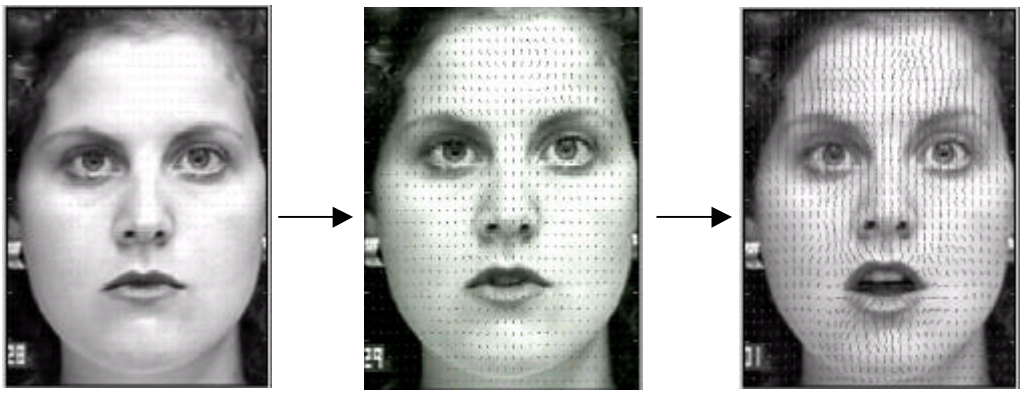
Figure 8a. The 2^{nd}-order 3-state left-right Hidden Markov Model for each of the upper facial expression units AU4, AU1+4 and AU1+2.

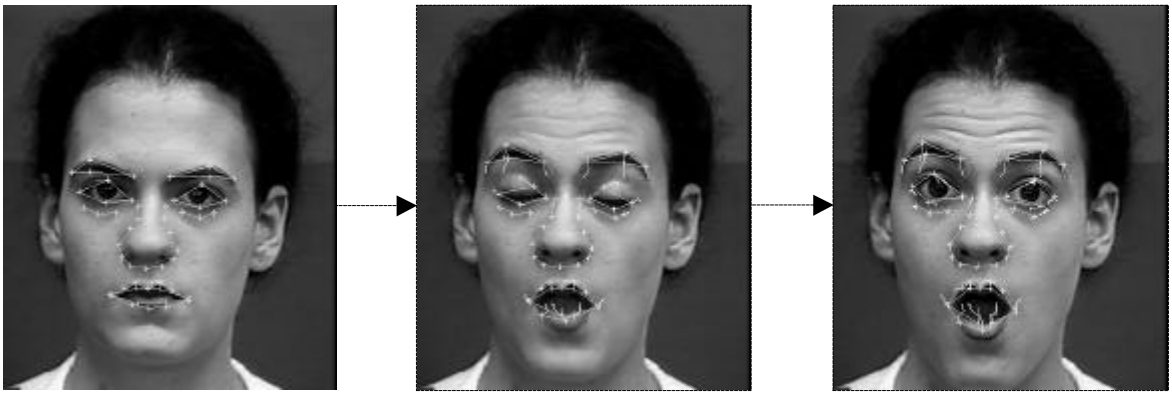
Figure 8b. The 3rd-order 4-state left-right Hidden Markov Model for each of the lower facial expression units AU12, AU6+12+25, AU20+25, AU9+17, AU15+17 and AU17+23+24.

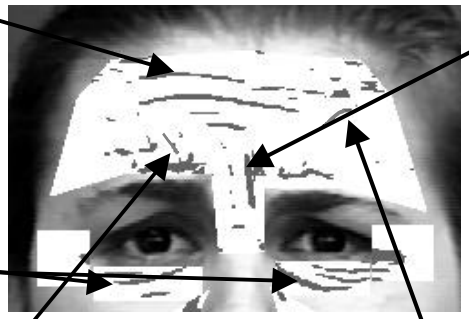
Upper Face Action Units		
AU4	AU1+4	AU1+2
		
Brows lowered and drawn together	Medial portion of the brows is raised and pulled together	Inner and outer portions of the brows are raised
AU5	AU6	AU7
		
Upper eyelids are raised	Cheeks are raised and eye opening is narrowed	Lower eyelids are raised
Lower Face Action Units		
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


Feature Extraction
Recognition
Alignment
Dense Flow Extraction
Facial Feature Tracking
High Gradient Detection
Discriminant Analysis
Hidden Markov Modeling
Action Units


1
16
39
48
a
b
c
d







Horizontal Line Detection

1	1	1	1	1
0	0	0	0	0
-1	-1	-1	-1	-1

Vertical Line Detection

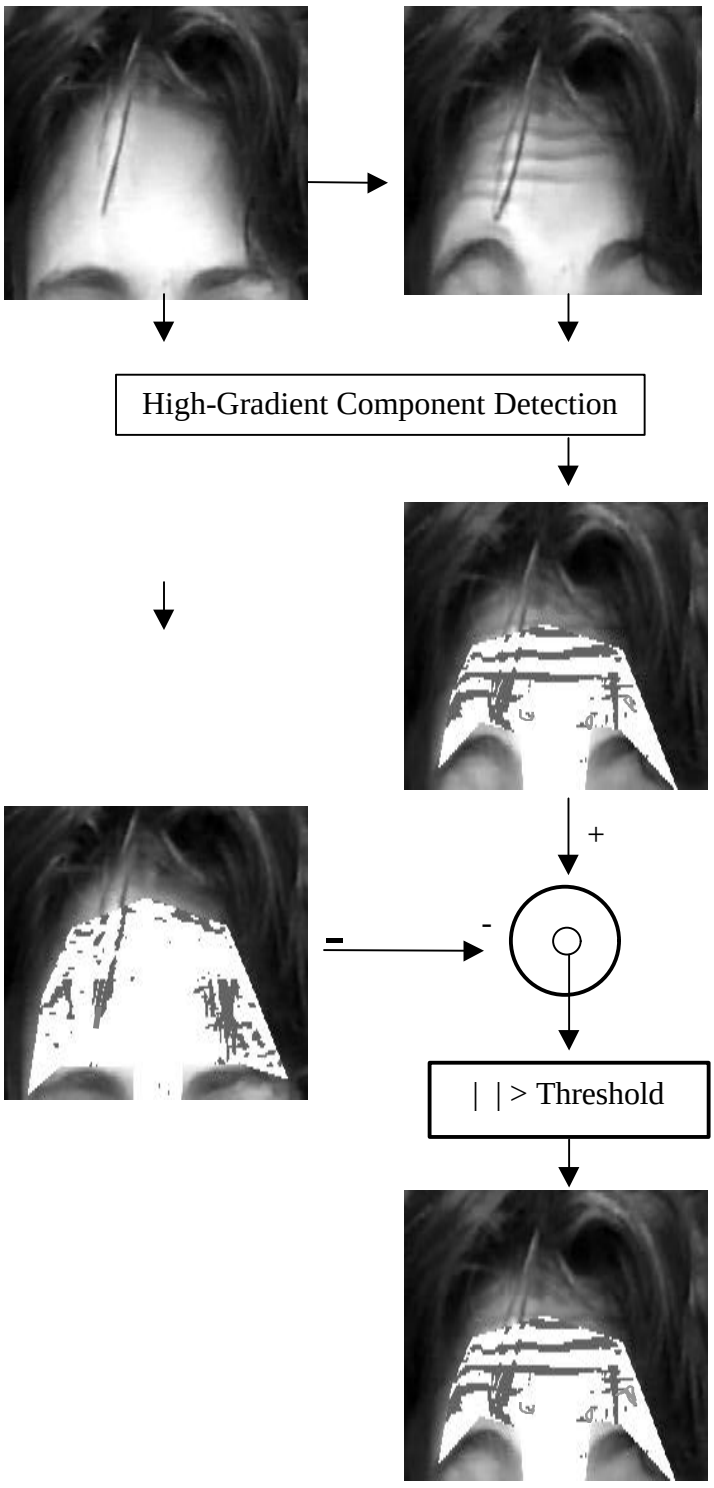
1	0	-1
1	0	-1
1	0	-1
1	0	-1
1	0	-1

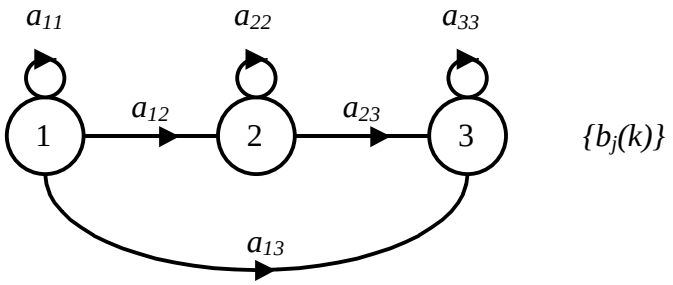
135^0 Diagonal Line Detection

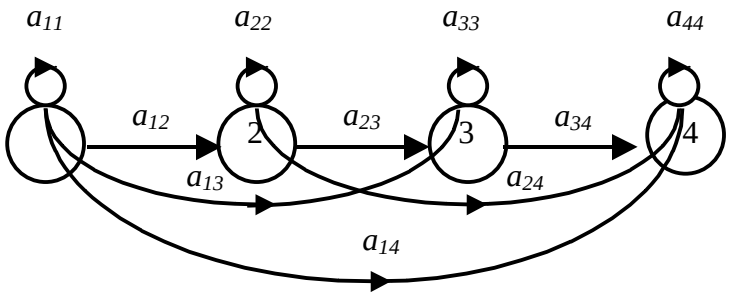
0	1	0	0	0
-1	0	1	0	0
0	-1	0	1	0
0	0	-1	0	1
0	0	0	-1	0

45^0 Diagonal Line Detection

0	0	0	-1	0
0	0	-1	0	1
0	-1	0	1	0
-1	0	1	0	0
0	1	0	0	0


High-Gradient Component Detection
+
-
| | > Threshold


a_{11}
a_{22}
a_{33}
1
a_{12}
2
a_{23}
3
a_{13}
$\{b_j(k)\}$


a_{11}
a_{22}
a_{33}
a_{44}
a_{12}
2
a_{23}
3
a_{34}
4
a_{13}
a_{24}
a_{14}