

2.0 FACIAL EXPRESSION RECOGNITION SYSTEM OVERVIEW

Humans are capable of producing thousands of facial actions during communication that vary in complexity, intensity, and meaning. Emotion or intention is often communicated by subtle changes in one or several discrete features. The addition or absence of one or more facial actions may alter its interpretation. In addition, some facial expressions may have a similar gross morphology but indicate varied meaning for different expression intensities. In order to capture the subtlety of facial expression in nonverbal communication, we propose to develop a computer vision system with a user interface (Figure 2) that automatically extract features and their motion information, discriminate subtly different facial expressions, and estimate expression intensity. The system contains two components: extraction and recognition as shown in Figure 3. Three methods are developed for feature and motion extraction yielding symbol sequences to represent observed expressions. These symbol sequences are input to the recognition process, which is an HMM computation to give the maximum likelihood decision.

2.1 Three Methods of Feature Motion Extraction

Facial expression is produced by the activation of facial muscles, which are triggered by the nerve impulses. Facial muscle actions cause the movement and deformations of facial skin and facial features. In the interpretation of facial expression, it is these deformations which we observe, and from which we must deduce the underlying emotion. Three convergent approaches are used to extract expression information (Figure 3): (1) facial feature point tracking using the pyramid method, (2) dense flow tracking with principal component analysis (PCA), and (3) high gradient component analysis in the spatio-temporal domain. In order to allow recognition, this extracted expression



Figure 2 The user interface created by programming in C, Motif, X Toolkit and Xlib.

information must be converted into motion vectors so they may be passed to the recognition process (Figure 3).

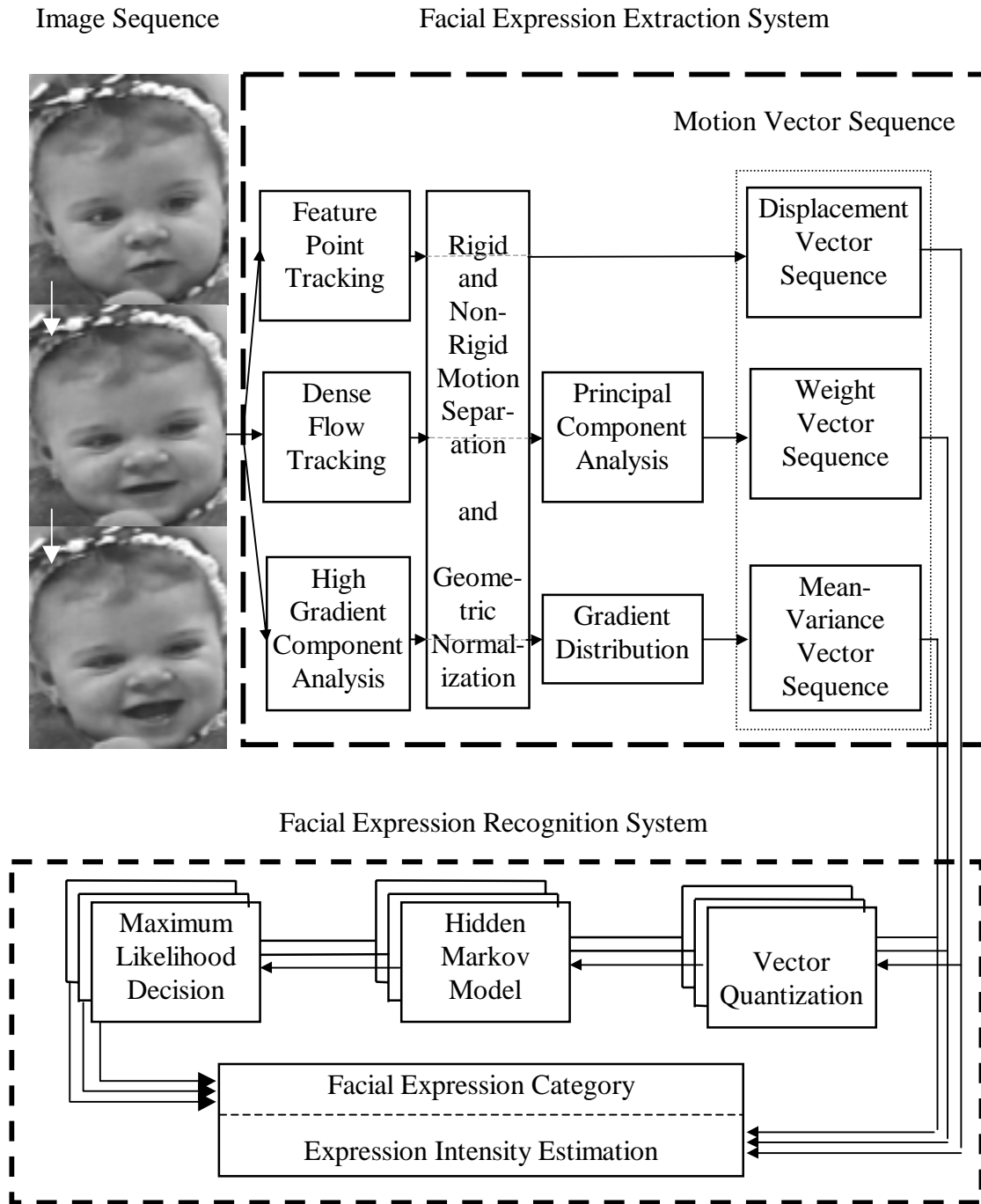


Figure 3 Block diagram of a facial expression recognition system.

Feature point tracking and dense flow tracking are used to track facial motion for recognition of expressions varying in intensity in the spatio-temporal domain. Frontal views of subjects (none wears eyeglasses) are videotaped under constant illumination, although lighting may vary across subjects particularly when we videotape on different days. These constraints are imposed to prevent significant degradation in optical flow calculation.

Facial feature point tracking using the pyramid method is especially sensitive to subtle feature motion and is also able to track a large displacement of feature motion in subpixel accuracy. Facial feature point is based on facial features in regions of brows, eyes, nose, and mouth. However, the forehead, cheek and chin regions also have important expression information. Dense flow tracking is used to include motion information from the entire face. The use of optical flow to track motion is advantageous because facial features and skin naturally have a great deal of texture. Using the principal component analysis, a low-dimensional weight vector in eigenspace can be obtained to represent the high-dimensional dense flows of each frame. Based on the displacement and weight vectors, the motion information is converted to symbol sequences from which we can recognize facial expressions, and is applied to estimate the expression intensity.

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

2.2 Recognition Using Hidden Markov Models

Modeling facial expression needs to take into account the stochastic nature of human facial expression involving both the human mental state, which is hidden or immeasurable, and the human action, which is observable or measurable. For example, different people

with the same emotion may exhibit very different facial actions, expression intensities and durations. Individual variations notwithstanding, a human observer can still recognize what emotion is being expressed, indicating that some common element underlies each motion. Therefore, the purpose of facial expression modeling is to uncover the hidden patterns associated with specific expressions from the measured (observable) data. Facial expression modeling requires a criterion for measuring a specific expression. It is desirable to analyze a sequence of images to capture the dynamics ⁽⁵⁾. Expressions are recognized in the context of an entire image sequence of arbitrary length. We will develop a recognition system based on the stochastic modeling of the encoded time series describing facial expressions, which should perform well in the spatio-temporal domain, analogous to the human performance.

In order to model subtly different facial expressions having different durations (arbitrary length of image sequence), the Hidden Markov Model (HMM) is developed to recognize expressions based on the maximum likelihood decision criterion. A key problem is to determine the HMM topology for the facial expressions under consideration. Some other advantages of using HMMs are: HMM computations converge quickly making it practical for real time processing, it may evaluate an input sequence of uncertain category to present a low output probability, and a multi-dimensional HMM may be developed to integrate individual HMMs to give a robust and reliable recognition. The correspondence between facial expressions and elements of the HMM is shown in Table 1.

Facial expression and speech represent human visual and audio actions, respectively ⁽⁸⁸⁾. The HMM technique has been successfully applied to model all known phonemes (the basic units of speech). Elementary HMMs of phonemes have then combined to represent words, and then sentences ^(59,76,79). Speech may be considered as two- or three-dimensional signals: frequency and amplitude change with time. Facial expressions may be considered as three (or four)-dimensional signals: a time sequence of images. So a set of elementary HMMs will be developed to model various “expression units” of individual

Table 1 Correspondence between facial expressions and elements of the Hidden Markov Model.

	Facial Expression	Hidden Markov Model
Hidden Process	Mental State	Model State
Observable	Expression (Facial Action)	Symbol Sequence
Temporal Domain	Dynamic Behavior	A Network of State Transition
Characteristics	Expression	State Transition Probability and Symbol Probability
Recognition	Expression Similarity	The Confidence of Output Probability

AUs or AU combinations, such as illustrated in Figure 4. Based on combinations of elementary HMMs, we will be able to recognize continuously varying facial expressions. A comparison of modeling facial expressions with modeling speech using HMMs is listed in Table 2.

2.3 Facial Action Coding System and “Expression Units”

The proposed automatic of facial expression analysis follows the anatomically based Facial Action Coding System (FACS) ⁽³⁴⁾, which is the most comprehensive method for coding facial expressions by psychologists. With FACS, observers can manually code discrete deformations of the face (movements of the facial muscle and skin) which are referred to as action units (AUs). Basically, FACS divides the face into upper and lower facial expressions and subdivides motion AUs. FACS consists of 44 basic AUs, with 14 additional AUs for head and eye positions as shown in Table 3. AUs are the smallest visibly discriminable muscle actions that individuate or combine to produce characteristic facial expressions which can be recognized from the image. More than 7000










Upper Facial Expressions		
<u>AU4:</u> Brows are lowered and drawn together.	<u>AU1+4:</u> Medial portion of the eyebrows is raised (AU1) and pulled together (AU4).	<u>AU1+2:</u> Inner (AU1) and outer (AU2) portions of the brows are raised.
		
Lower Facial Expressions		
<u>AU12:</u> Lip corners are pulled up and backward.	<u>AU6+12+25:</u> Cheek raised (the lower-eye and infra-orbital furrows are raised and deepened, and the eye opening is narrowed) (AU6), and AU12 with mouth opening (AU25).	<u>AU20+25:</u> Lips are parted (AU25), pulled back laterally, and may be slightly raised or pulled down (AU20).
		
<u>AU9+17:</u> The infra-orbital triangle and center of the upper lip are pulled upwards (AU9), and the chin boss and lower lip are pulled upwards (AU17).	<u>AU17+23+24:</u> The chin boss is raised, which pushes up the lower lip (AU17); the lips are tightened, narrowed (AU23), and pressed together (AU24).	<u>AU15+17:</u> Lip corners are pulled down and stretched laterally (AU15), and chin boss is raised which pushes up the lower lip (AU17).
		

Figure 4 “Expression units” of subtly different facial expressions in our study (taken from ⁽³⁴⁾).

Table 2 Comparison of modeling facial expressions with modeling speech using HMMs.

	Speech	Facial Expressions
Human Action	Audio Action	Visual Action
Dimension (Including Time Series)	2-dimensional Signals	3 or 4-dimensional Signals
Action Unit	Phoneme	Expression Unit: Individual AUs or AU Combinations
HMM Unit	1st-order 3-state HMM	2nd-order 3-state HMM for Upper Facial Expression and 3rd-order 4-state HMM for Lower Facial Expression *.
HMM Unit Combinations	One Word	One Basic Facial Expression (<i>e.g.</i> , joy)
Concatenated HMM Unit Combinations	Sentences	Continuously Varying Basic Facial Expressions

* Obtained in this research.

have been observed. According to FACS, each AU corresponds to an activity in a distinct muscle, with the exception of AU4^(34,107). Even though the one-to-one mapping of individual AUs to distinct muscle activities is a basic assumption of the FACS, AUs enable discrimination between closely related expressions. By discriminating “expression units (individual AUs or AU combinations)”, we can simulate and understand individual mechanics of the facial muscles. In the present study we consider, three upper facial “expression units” and six lower facial “expression units” which are shown in Figure 4. They are frequently occurring facial expressions containing subtle differences. They will be studied for automatic recognition and estimation of their intensities.

Table 3 Action Units (AUs) in the Facial Action Coding System (FACS) ⁽³⁴⁾.

<u>Upper Face</u>		<u>Lower Face</u>		<u>Miscellaneous</u>	
<u>AU</u>	<u>Label</u>	<u>AU</u>	<u>Label</u>	<u>AU</u>	<u>Label</u>
1	Inner Brow Raise	9	Nose Wrinkle	8	Lips Toward
2	Outer Brow Raise	10	Upper Lip Raise	19	Tongue Show
4	Brow Lower	11	Nasolabial Deepen	21	Neck Tighten
5	Upper Lid Raise	12	Lip Corner Pull	29	Jaw Thrust
6	Cheek Raise	13	Sharp Lip Pull	30	Jaw Sideways
7	Lids Tight	14	Dimple	31	Jaw Clench
41	Lids Droop	15	Lip Corner Depress	32	Bite (Lip)
42	Lids Slit	16	Lower Lip Depress	33	Blow
43	Lids Closed	17	Chin Raise	34	Puff
44	Squint	18	Lip Pucker	35	Cheek Suck
45	Blink	20	Lip Stretch	36	Tongue Bulge
46	Wink	22	Lip Funnel	37	Lip Wipe
		23	Lip Tight	38	Nostril Dilate
		24	Lip Press	39	Nostril Compress
		25	Lips Part		
		26	Jaw Drop		
		27	Mouth Stretch		
		28	Lip Suck		
 <u>Head Position</u>			 <u>Eye Position</u>		
<u>AU</u>	<u>Label</u>	<u>AU</u>	<u>Label</u>		
51	Turn Left	61	Left		
52	Turn Right	62	Right		
53	Head Up	63	Up		
54	Head Down	64	Down		
55	Tilt Left	65	Walleye		
56	Tilt Right	66	Cross-eye		
57	Forward				
58	Back				

2.4 Rigid and Non-Rigid Motion Separation and Geometric Normalization

For facial expression recognition, two main issues in image processing will affect the recognition results: separation of non-rigid facial expression from rigid head motion, and facial geometric correspondence to keep face size constant across subjects. Both processes are necessary in order to ensure that these variables do not interfere with expression recognition. Though all subjects are viewed frontally in our current research, some out-of-plane head motion (*e.g.*, yaw rotations or less than ± 10 degree pitch rotations) may occur with facial expressions. Furthermore, face size varies among individuals. For elimination of the above-mentioned rigid head motion from non-rigid facial expression, an affine transformation (which includes translation, scaling and rotation factors) is adequate to normalize the face geometric position and maintain face magnification invariance. Face images are automatically normalized with the affine transformation to ensure that optical flows or gray values of individual frames have close geometric correspondence in order to achieve consistent recognition performance.

In the first frame of each image sequence, we manually select three facial feature points for image normalization: medial canthus of both eyes and the uppermost point on the philtrum as shown in Figure 5. These three points will carry only rigid motion components accompanied with the head motion. Each of these points forms the center of a 13 x 13 pixel flow window, and they are automatically tracked in the remaining frames of each image sequence. Based on these three facial feature points, the original 490 x 640 (row x column) pixel display is cropped to 417 x 385 pixels for each frame to keep the foreground face and remove the unnecessary background. The positions of all tracking facial feature points, dense flows, or image gray values for each frame are then normalized by warping them onto a standard two-dimensional face model based on the affine transformation \mathcal{J} (Figure 5) given as follows:

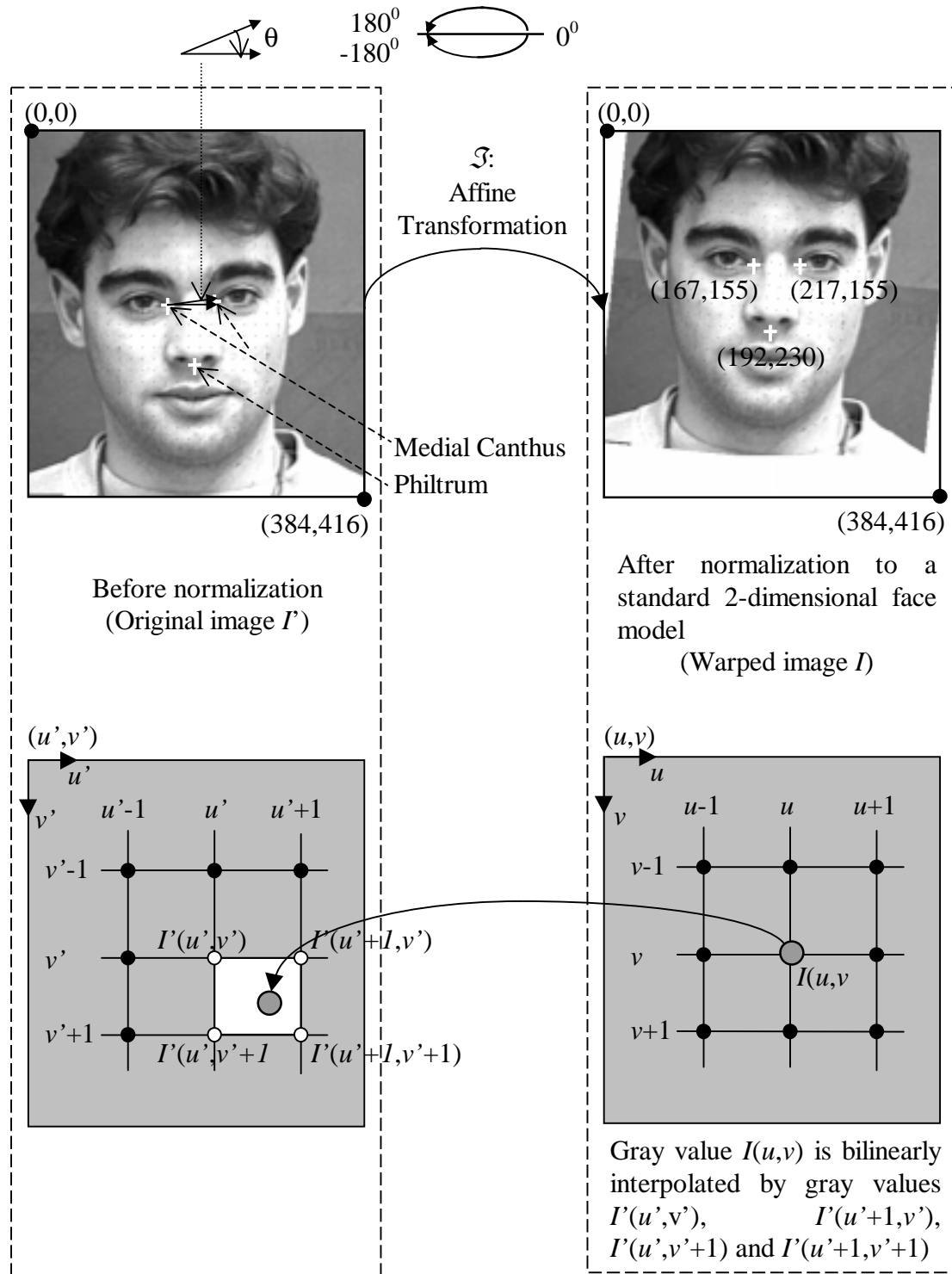


Figure 5 Normalization of each face image to a standard 2-dimensional face model.

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} S_u & 0 \\ 0 & S_v \end{bmatrix} \begin{bmatrix} u' \\ v' \end{bmatrix} + \begin{bmatrix} D_u \\ D_v \end{bmatrix} \quad (2-1)$$

where

$$S = \begin{bmatrix} S_u & S_v \end{bmatrix} = \begin{bmatrix} \frac{w}{w'} & \frac{h}{h'} \end{bmatrix} \quad (2-2)$$

$$D = \begin{bmatrix} D_u & D_v \end{bmatrix} = \begin{bmatrix} d_u - d'_u & d_v - d'_v \end{bmatrix} \quad (2-3)$$

Here, u and v are the horizontal and vertical positions of the two-dimensional face model coordinates, and u' and v' are the horizontal and vertical positions of the original image coordinates. The upper-left corner of each frame, including the face model image is denoted as (0,0). In the standard face model, the top point of the philtrum is the rotation center whose position is $(d_u, d_v) = (192, 230)$, and the position of the medial canthus of the right eye is (167,155) and that of the left eye is (217,155); the width w between the medial canthi of both eyes is 50 pixels and the height h from the level of the medial canthi to the top point of the philtrum is 75 pixels. The horizontal scaling is given by the parameter S_u which is computed as the ratio of the distance w at the face model to that distance w' at the original face image. The vertical scaling given by S_v is computed as the ratio of the distance h at the face model to that distance h' at the original face image. The horizontal and vertical displacements (translations) are represented by D_u and D_v , respectively, and are measured from the top point of the philtrum in the original face image (d'_u, d'_v) to that in the face model (d_u, d_v) . The angle of rotation of the line connecting the medial canthi of both eyes in the original face image from the corresponding horizontal line in the face model is represented by θ , where the clockwise rotation is negative and the counterclockwise rotation is positive. The pixel positions of each image are integer-valued, but the warped positions after the affine transformation are, in general, not integer-valued. So the gray value at each integer-valued pixel of the warped image needs to be estimated by bilinear interpolation based on the gray values of its four nearest neighbor pixels in the original image as shown in Figure 5.