Recognizing Lower Face Action Units for Facial Expression Analysis

Ying-li Tian [1,3] Takeo Kanade[1] and Jeffrey F. Cohn[1,2]

[1] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213
[2] Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260
[3] National Laboratory of Pattern Recognition
Chinese Academy of Sciences, Beijing, China
Email: {yltian, tk}@cs.cmu.edu jeffcohn@pitt.edu

Abstract

Most automatic expression analysis systems attempt to recognize a small set of prototypic expressions (e.g. happiness and anger). Such prototypic expressions, however, occur infrequently. Human emotions and intentions are communicated more often by changes in one or two discrete facial features. In this paper, we develop an automatic system to analyze subtle changes in facial expressions based on both permanent (e.g. mouth, eye, and brow) and transient (e.g. furrows and wrinkles) facial features in a nearly frontal image sequence. Multi-state facial component models are proposed for tracking and modeling different facial features. Based on these multi-state models, and without artificial enhancement, we detect and track the facial features, including mouth, eyes, brow, cheeks, and their related wrinkles and facial furrows. Moreover we recover detailed parametric descriptions of the facial features. With these features as the inputs, 11 individual action units or action unit combinations are recognized by a neural network algorithm. A recognition rate of 96.7% is obtained. The recognition results indicate that our system can identify action units regardless of whether they occurred singly or in combinations.

1. Introduction

Recently facial expression analysis has attracted attention in the computer vision literature [4, 5, 8, 9, 11, 14]. Most automatic expression analysis systems attempt to recognize a small set of prototypic expressions (i.e. joy, surprise, anger, sadness, fear, and disgust) [9, 14]. In everyday life, however, such prototypic expressions occur relatively infrequently. Instead, emotion is communicated by changes in one or two discrete facial features, such as tightening the lips in anger or obliquely lowering the lip corners in sadness [2]. 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.

Ekman and Friesen [3] developed the Facial Action Coding System (FACS) for describing subtle changes in facial expressions. FACS consists of 49 action units, including those for head and eye positions. Thirty of these are anatomically related to contraction of a specific set of facial muscles. Although there are a small number of atomic action units, more than 7,000 combinations of action units have been observed [10]. FACS provides the necessary detail with which to describe facial expression.

Automatic recognition of action units is a difficult problem because there are no quantitative definitions and they appear in complex combinations. Previous work to directly recognize action units has used optical flow across the entire face or facial feature measurement [1, 8]. Bartlett *et al.* [1] built a hybrid system to recognize 6 action units in the upper face using 50 component projections, 5 feature measures, and 6 motion template matches. The system of Lien *et al.* [8] used dense-flow, feature point tracking and edge extraction to recognize action units.

In this paper, we develop an automatic action unit analysis system using facial features. Figure 1 depicts the overview of the analysis system. First, the head orientation and face position is detected. Then, subtle changes in the facial components are measured. Motivated by FACS action units, these changes are represented as a collection of mid-level feature parameters. Finally, action units are classified by a neural network using these parameters as the inputs.

To detect qualitative changes in facial expression, we develop a multi-state model based system for tracking facial features that uses convergent methods of feature analysis. We define the different head orientations and different component appearances as different states. For different head states, specific face components are used. For each face component, there are different states also. For each dif-

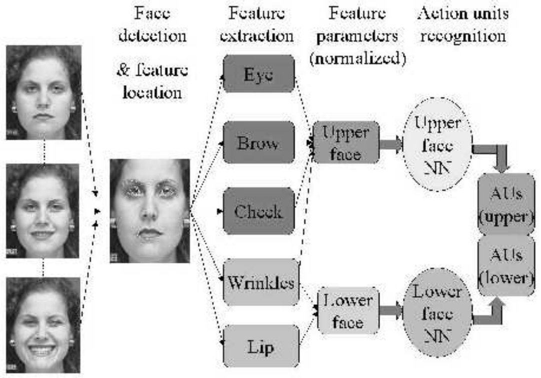


Figure 1. Feature based action unit recognition system.

ferent state, a different description and extraction method maybe used.

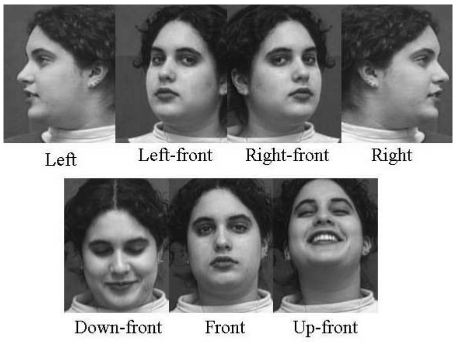
Based on the action unit description, we separately represent these facial features in two parameter groups for upper face and lower face. In this paper, we focus on recognizing lower face action units because the lips are much more deformable than eyes. Nine parameters are used to describe the lip shape, lip motion, lip state, and lower face furrows.

We employ a neural network to recognize the action units after the facial features are correctly extracted and suitably represented. Thirteen basic lower face action units and combinations (Neutral, AU9, AU 10, AU 12, AU 15, AU 17, AU 20, AU 25, AU 26, AU 27, AU 9+17, AU 10+17, and AU23+24) are identified by a single neural network. A previous attempt for a similar task [8] recognized 6 lower face AUs and combinations(AU 12, AU12+25, AU20+25, AU9+17, AU17+23+24, and AU15+17) with 88% average recognition rate by separate hidden Markov Models for each action unit or action unit combination. Compared to the previous results, the current system achieves a higher recognition accuracy with an average recognition rate of 96.71%. It is also able to identify action units regardless of whether they occurred singly or in combinations. As described in [10] over 7,000 AU combinations have been observed. Modeling all of these combinations is intractable at this time. Therefore, to demonstrate the robustness of our system to unmodeled AU combinations, we included in the test set various unmodeled AU combinations (including AU 12+25, AU 12+26, AU20+25, AU15+17+23, AU9+17+25, and AU10+17+23+24).

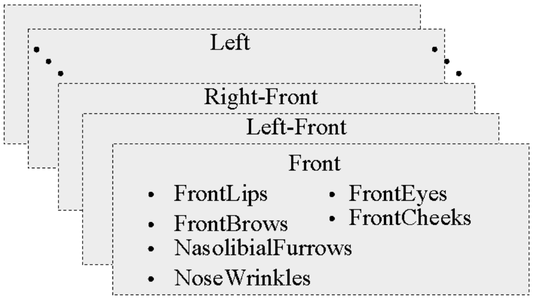
2. Multi-State Models for Face and Facial Components

2.1. Multi-state face model

Head orientation is a significant factor that affects the appearance of a face. Based on the head orientation, seven head states are defined in Figure 2. To develop more ro-



(a) Head state.



(b) Different facial components used for each head state.

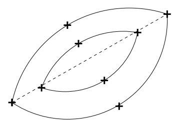
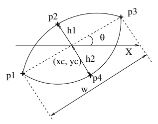
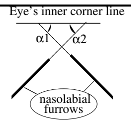
Figure 2. Multiple state face model. (a) The head state can be left, left-front, front, right-front, right, down, and up. (b) Different facial component models are used for different head states.

bust facial expression recognition system, head state will be considered. For the different head states, facial components, such as lips, appear very differently, requiring specific facial component models. For example, the facial component models for a front face include $FrontLips$, $FrontEyes$ (left and right), $FrontCheeks$(left and right), $NasolabialFurrows$, and $Nosewrinkles$. The right face includes only the component models $SideLips$, $Righteye$, $Rightbrow$, and $Rightcheek$. In our current system, we assume the face images are nearly front view with possible in-plane head rotations.

2.2. Multi-state lip and furrow models

Different face component models must be used for different head states. For example, a lip model of the front face doesn't work for a profile face. Here, we give the detailed facial component models in the nearly front-view lower face (Table 1). Both the permanent components such as lips and the transient components such as furrows are considered. Based on the different appearances of different components, specific geometric models are used to model

Table 1. Multi-state facial component models of a front face

Component	State	Description/Feature
Lip	Open	
	Closed	
	Tightly closed	
Furrow	Present	
	Absent	

the component's location, shape, and appearance. Each component employs a multi-state model corresponding to different component states. For example, a three-state lip model is defined to describe the lip states: open, closed, and tightly closed. Present and absent are use to model states of the transient facial features.

3. Lower Face Feature Extraction

Contraction of the facial muscles produces changes in both the direction and magnitude of the motion on the skin surface and in the appearance of permanent and transient facial features. Examples of permanent features are the lips, eyes, and any furrows that have become permanent with age. Transient features include any facial lines and furrows that are not present at rest. We assume that the first frame is in a neutral expression. After initializing the templates of the permanent features in the first frame, both permanent and transient features can be tracked and detected in the whole image sequence regardless of the states of facial components. The tracking results show that our method is robust for tracking facial features even when there is large out of plane head rotation.

3.1. Lip features

A three-state lip model is used for tracking and modeling lip features. As shown in Table 1, we classify the mouth states into open, closed, and tightly closed. Different lip templates are used to obtain the lip contours. Currently, we use the same template for open and closed mouth. Two parabolic arcs are used to model the position, orientation,

(a) (b) (c) (d)

Figure 3. Permanent feature tracking results for different expressions. (a) Narrowing eyes and opened smiled mouth. (b) Large open eye, blinking and large opened mouth. (c) Tight closed eye and eye blinking. (4) Tightly closed mouth and blinking.

and shape of the lips. The template of open and closed lips has six parameters: lip center (xc, yc), lip shape ($h1$, $h2$ and w), and lip orientation (θ). For a tightly closed mouth, the dark mouth line connecting lip corners is detected from the image to model the position, orientation, and shape of the tightly closed lips.

After the lip template is manually located for the neutral expression in the first frame, the lip color is obtained by modeling as a Gaussian mixture. The shape and location of the lip template for the image sequence is automatically tracked by feature point tracking. Then, the lip shape and color information are used to determine the lip state and state transitions. The detailed lip tracking method can be found in paper [12].

Some permanent facial feature tracking results for different expressions are shown in Figure 3. More facial tracking results can be found at http://www.cs.cmu.edu/~face.

3.2. Transient features

Facial motion produces transient features. Wrinkles and furrows appear perpendicular to the motion direction of the activated muscle. These transient features provide crucial information for 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. When lines and furrows become permanent facial features, contraction of the corresponding muscles produces changes in their appearance, such as deepening or lengthening. The presence or absence of the furrows in a face image can be determined by geometric feature analysis [8, 7], or by eigen-analysis [6, 13]. Kwon and Lobo [7] detect furrows by snake to classify pictures of people into different age groups. Lien [8] detected whole face horizontal, vertical and diagonal edges for face expression recognition.

In our system, we currently detect nasolabial furrows, nose wrinkles, and crows feet wrinkles. We define them in two states: present and absent. Compared to the neutral frame, the wrinkle state is present if the wrinkles appear, deepen, or lengthen. Otherwise, it is absent. After obtaining the permanent facial features, the areas with furrows related to different AUs can be decided by the permanent facial feature locations. We define the nasolabial furrow area as the area between eye's inner corners line and lip corners line. The nose wrinkle area is a square between two eye inner corners. The crows feet wrinkle areas are beside the eye outer corners.

We use canny edge detector to detect the edge information in these areas. For nose wrinkles and crows feet wrinkles, we compare the edge pixel numbers E of current frame with the edge pixel numbers E_0 of the first frame in the wrinkle areas. If E/E_0 large than the threshold T, the furrows are present. Otherwise, the furrows are absent. For the nasolabial furrows, we detect the continued diagonal edges. The nasolabial furrow detection results are shown in Fig. 4.

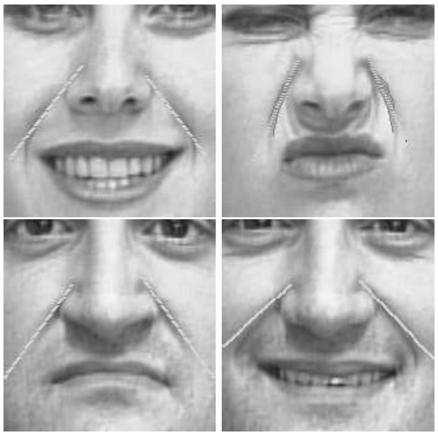


Figure 4. Nasolabial furrow detection results. For the same subject, the nasolabial furrow angle(between the nasolabial furrow and the line connected eye inner corners) is different for different expressions.

4. Lower Face Feature Representation

Each action unit of FACS 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 and deform shape or location of the facial components. In order to recognize the subtle changes of face expression, it is necessary to represent the facial features in a group of suitable parameters.

We define nine parameters to represent the lower face features from the tracked facial features. Of these, 6 parameters describe the permanent features of lip shape, lip state and lip motion, and 3 parameters describe the transient features of the nasolabial furrows and nose wrinkles.

For defining these parameters, we first define the basic coordinate system. Because the eye's inner corners are the most stable features in a face and relatively insensitive to facial expressions, we define the x-axis is the line connected two inner corners of eyes and the y-axis is perpendicular to x-axis. All the parameters of lip motion and the nasolabial furrows are calculated in this coordinate system and obtain the ratio by comparing to the neutral frame. In order to remove the effects of the different size of face images in different image sequences, we use the parameter ratios instead of directly using the parameters.

We notice that if the nasolabial furrow is present, there are different angles between the nasolabial furrow and x-axis for different action units. For example, the nasolanial furrow angle of AU9 or AU10 is larger than that of AU12. So we use the angle to represent its orientation if it is present. Although the nose wrinkles are located in the upper face, but we classify the parameter of them in the lower face feature because it is related to the lower face AUs.

The definitions of lower face parameters are listed in Table 2. These feature data are affine aligned by calculating them based on the line connected two inner corners of eyes and normalized for individual differences in facial confor-

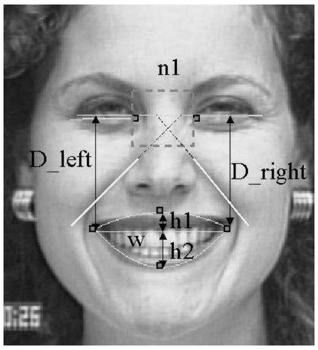


Figure 5. Lower face features. $h1$ and $h2$ are the top and bottom lip heights; w is the lip width; D_{left} is the distance between the left lip corner and eye inner corners line; D_{right} is the distance between the right lip corner and eye inner corners line; $n1$ is the nose wrinkle area.

mation by converting to ratio scores. The physic meanings of the parameters are shown in Figure 5.

5. Lower Face Action Unit Recognition

5.1. Lower face action units

The descriptions of the basic lower face action units to be recognized are shown in Table 3.

5.2. NN structure

We used a three layer neural network with one hidden layer to recognize the lower face action units. Unlike previous methods [8, 14] which build a separate model for each expression or action unit, we build a single model for all the basic lower face action units and AU combinations.

Action units can occur either singly or in combinations. The action unit combinations may be additive, in which case combination does not change the appearance of the constituents, or nonadditive, in which case the appearance of the constituents does change. Lien *et al.* [8] separately modeled each action unit combination no matter whether the combination is additive or nonaddtive. Combinations which are not explicity modeled therefore cannot be recognized, even for those additive combinations.

We trained two neural networks. The first one ignores the nonadditive combinations and only models the basic single action units. The second one separately models some nonadditive combinations such as AU9+17 and AU10+17 besides the basic single action units. The additive combinations of the basic single action units can be correctly recognized. We found both networks achieved high recognition accuracy but the second one is better.

5.3. NN inputs

The inputs of the neural network are the lower face feature parameters shown in Table 2. Seven parameters are used except two parameters of the nasolabial furrows. We don't use the angles of the nasolabial furrows because they are varied much for the different subjects. Generally, we use them to analyze the different expressions of same subject.

Table 2. Representation of lower face features for AUs recognition

Permanent features		
Lip height (r_{height})	Lip width (r_{width})	Left lip corner motion (r_{left})
$r_{height} = \frac{(h1+h2)-(h1_0+h2_0)}{(h1_0+h2_0)}$. If r_{height}>0, lip height increases.	$r_{width} = \frac{w-w_0}{w_0}$. If r_{width}>0, lip width increases.	$r_{left} = -\frac{D_{left}-D_{left0}}{D_{left0}}$. If r_{left}>0, left lip corner move up.
Right lip corner (r_{right})	Top lip motion (r_{top})	Bottom lip motion(r_{btm})
$r_{right} = -\frac{D_{right}-D_{right0}}{D_{right0}}$. If r_{right}>0, right lip corner move up.	$r_{top} = -\frac{D_{top}-D_{top0}}{D_{top0}}$. If r_{top}>0, top lip move up.	$r_{btm} = -\frac{D_{btm}-D_{btm0}}{D_{btm0}}$. If r_{btm}>0, bottom lip move up.
Transient features		
Left nasolibial furrow angle (Ang_{left})	Right nasolibial furrow angle (Ang_{right})	State of nose wrinkles (S_{nosew})
Left nasolibial furrow present with angle Ang_{left}.	Left nasolibial furrow present with angle Ang_{right}.	If $S_{nosew} = 1$, nose wrinkles present.

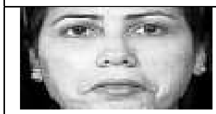
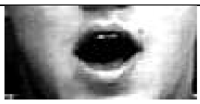
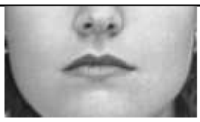
5.4. NN outputs

The outputs of the neural network are 13 lower face action units or AU combinations including 10 basic lower face action units (neutral, AU9, AU10, AU12, AU15, AU17, AU20, AU25, AU26, and AU27) and 3 combinations (AU23+24, AU9+17, and AU10+17). The basic lower face action units are shown in Table 3. When an action unit occurred in combination with other action units that may modify the single AU's appearance, we call these kind of combinations as nonadditive combinations. For the nonadditive combinations, we analysis them as independent action units such as AU9+17 and AU10+17. Additive combinations are not modeled separately. We use AU 23+24 instead of AU23 and AU24 because they almost occur together.

5.5. Training and test data set

We use the data of *Pitt-CMU AU-Coded Face Expression Image Database*, which currently includes 1917 image sequences from 182 adult subjects of varying ethnicity, performing multiple tokens of 29 of 30 primary FACS action units. 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 6+12+25). Each expression sequence began from a neutral face. For each sequence, action units were coded by a certified FACS coder.

Table 3. Description of the basic lower face action units or combination

AU 9	AU 10	AU20
		
The infraorbital triangle and center of the upper lip are pulled upwards. Nose wrinkling is present.	The infraorbital triangle is pushed upwards. Upper lip is raised. Nose wrinkle is absent.	The lips and the lower portion of the nasolabial furrow are pulled pulled back laterally. The mouth is elongated.
AU 15	AU 17	AU12
		
The corner of the lips are pulled down.	The chin boss is pushed upwards.	Lip corners are pulled obliquely.
AU 25	AU 26	AU27
		
Lips are relaxed and parted.	Lips are relaxed and parted; mandible is lowered.	Mouth stretched, open and the mandible pulled downwards.
AU 23+24	neutral	
		
Lips tightened, narrowed, and pressed together.	Lips relaxed and closed.	

Total 463 image sequences from 122 adults (65% female, 35% male, 85% European-American, 15% African-American or Asian, ages 18 to 35 years) are processed for lower face action unit recognition. Some of the image sequences are with more action unit combinations such as AU9+17, AU10+17, AU12+25, AU15+17+23, AU9+17+23+24, and AU17+20+26. For each image sequence, we use the neutral frame and two peak frames. 400 image sequences are used as training data and 63 different image sequences are used as test data.

5.6. Recognition results

The recognition results of 63 image sequences are shown in Table 4. The average recognition rate is 96.71%. 100% recognition rate is obtained for each basic lower face action unit except AU10, AU17, and AU26. All the mistakes of AU26 are confused by AU25. It is reasonable because both AU25 and AU26 are with parted lips. But for AU26, the mandible is lowered. We did not use the jaw motion information in current system. All the mistakes of AU10 and AU17 are caused by the image sequences with AU combination AU10+17. Two combinations AU10+17 are classified to AU10+12. One combination of AU10+17 is classified as AU10 (missing AU17). The combination AU 10+17 modified the single AU's appearance. The neural network needs to learn the modification by more training data of AU 10+17. There are only ten examples of AU10+17 in 1220 training data in our current system. More data about AU10+17 is collecting for future training. Our system is able to identify action units regardless of whether they occurred singly or in combinations. Our system is trained 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.

Table 4. Lower face action unit recognition results. 0 means neural expression.

AU	No.	Correct	false	Missed	Confused	Recognition rate
0	63	63	-	-	-	100%
9	16	16	-	-	-	100%
10	12	11	-	1	-	91.67%
12	14	14	2	-	-	100%
15	12	12	-	-	-	100%
17	36	34	-	-	2 (AU12)	94.44%
20	12	12	-	-	-	100%
25	50	50	5	-	-	100%
26	14	9	-	-	5 (AU25)	64.29%
27	8	8	-	-	-	100%
23+24	6	6	-	-	-	100%
Total	243	235	7	1	7	96.71%

For evaluating the necessity of including the nonadditive combinations, we also train a neural network using 11 basic lower face action units as the outputs. For the same test data set, the average recognition rate is 96.3%. We found that separately model the nonadditive combinations is helpful to increase action unit recognition accuracy.

6. Conclusion and Discussion

In this paper, we developed a neural network based facial expression recognition system. By training the network, was able to learn the correlations between facial feature parameter pattern and specific action units. Although often correlated, these effects of muscle contraction potentially provide unique information about facial expression. Action units 9 and 10 in FACS, for instance, are closely related expressions of disgust that are produced by variant regions of the same muscle. The shape of the nasolabial furrow and the state of nose wrinkles distinguishe between them. Change in the appearance of facial features also can affect the reliability of measurements of pixel motion in the face image. Closing of the lips or blinking of the eyes produces occlusion, which can confound optical flow estimation. Unless information about both motion and feature appearance are considered, accuracy of facial expression analysis and, in particular, sensitivity to subtle differences in expression may be impaired. To measure both types of information, we developed a multi-state method of tracking facial features that uses convergent methods of feature analysis and has high sensitivity and specificity for subtle differences in facial expression. All the facial features are represented in a group of features parameters. Eleven basic lower face action units are recognized and 96.71% of action units (neutral, AU 9, AU 10, AU12, AU 15, AU 17, AU20, AU25, AU 26, AU 27 and AU 23+24) were correctly classified. Those disagreements that occurred were primarily on AU 10+17 and AU 26.

In summary, the face image analysis system demonstrated concurrent validity with manual FACS coding. The multi-state model based convergent-measures approach was proved to capture the subtle changes of facial features. In the test set, which included subjects of mixed ethnicity, average recognition accuracy for 11 basic action units in the lower face was 96.71% regardless of these action units occur singly or in combinations. 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.

Acknowledgements

The authors would like to thank Zara Ambadar, Bethany Peters, and Michelle Lemenager for processing the images. We also thank Simon Baker for his comments and suggestions on earlier versions of this paper. This work is supported by NIMH grant R01 MH51435.

References

[1] M. Bartlett, J. Hager, P.Ekman, and T. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophysiology*, 36:253–264, 1999.

[2] J. M. Carroll and J. Russell. Facial expression in hollywood's portrayal of emotion. *Journal of Personality and Social Psychology.*, 72:164–176, 1997.

[3] P. Ekman and W. V. Friesen. *The Facial Action Coding System: A Technique For The Measurement of Facial Movement*. Consulting Psychologists Press Inc., San Francisco, CA, 1978.

[4] I. A. Essa and A. P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transc. On Pattern Analysis and Machine Intelligence*, 19(7):757–763, JULY 1997.

[5] K. Fukui and O. Yamaguchi. Facial feature point extraction method based on combination of shape extraction and pattern matching. *Systems and Computers in Japan*, 29(6):49–58, 1998.

[6] M. Kirby and L. Sirovich. Application of the k-l procedure for the characterization of human faces. *IEEE Transc. On Pattern Analysis and Machine Intelligence*, 12(1):103–108, Jan. 1990.

[7] Y. Kwon and N. Lobo. Age classification from facial images. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 762–767, 1994.

[8] J.-J. J. Lien, T. Kanade, J. F. Chon, and C. C. Li. Detection, tracking, and classification of action units in facial expression. *Journal of Robotics and Autonomous System*, in press.

[9] K. Mase. Recognition of facial expression from optical flow. *IEICE Transc.*, E. 74(10):3474–3483, 0ctober 1991.

[10] K. Scherer and P. Ekman. *Handbook of methods in nonverbal behavior research*. Cambridge University Press, Cambridge, UK, 1982.

[11] D. Terzopoulos and K. Waters. Analysis of facial images using physical and anatomical models. In *IEEE International Conference on Computer Vision*, pages 727–732, 1990.

[12] Y. Tian, T. Kanade, and J. Cohn. Robust lip tracking by combining shape, color and motion. In *Proc. Of ACCV'2000*, 2000.

[13] M. Turk and A. Pentland. face recognition using eigenfaces. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[14] Y. Yacoob and L. S. Davis. Recognizing human facial expression from long image sequences using optical flow. *IEEE Trans. On Pattern Analysis and machine Intelligence*, 18(6):636–642, June 1996.