Real-time Facial Expression Recognition in Image Sequences Using an AdaBoost-based Multi-classifier

Chin-Shyurng Fahn*, Ming-Hui Wu†, and Chang-Yi Kao‡
* National Taiwan University of Science and Technology, Taipei 10607, Taiwan
E-mail: csfahn@mail.ntust.edu.tw Tel: +886-02-2730-1215
† National Taiwan University of Science and Technology, Taipei 10607, Taiwan
E-mail: M9415054@mail.ntust.edu.tw Tel: +886-02-2733-3141 ext.7425
‡ National Taiwan University of Science and Technology, Taipei 10607, Taiwan
E-mail: D9515011@mail.ntust.edu.tw Tel: +886-02-2733-3141 ext.7425

***Abstract*—In this paper, a highly automatic facial expression recognition system without choosing characteristic blocks in advance is presented. The system is able to detect and locate human faces in image sequences acquired in real environments. To achieve efficient facial expression recognition, we evaluate the performance of three different classifiers using multi-layer perceptrons (MLPs), support vector machines (SVMs), and Adaboost algorithms (ABAs). From the experimental outcomes, we can observe that the average recognition rates obtained from both ABAs and MLPs are better than that from SVMs, but the training of MLPs takes quite a long time. Comparatively, ABAs have an advantage of facilitating the speed of convergence, which are chosen as the core technique to implement our strong facial expression classifier. Through conducting many experiments, the statistics of performance reveals that the accuracy rate of our facial expression recognition system reaches more than 90% for a single kind or multiple kinds of expressions appearing in an image sequence.**

I. INTRODUCTION

Robots can share the happiness with us [1]. To accomplish this, the interaction between humans and robots is critical techniques, especially for the face detection and facial expression recognition of an image sequence to which more and more researchers have been devoted [2-6]. Zhu et al. [7] adopted the hidden Markov model (HMM) as the classification scheme for facial expression recognition. The accuracy rate of their facial expression recognition system is satisfactory. However, the computation of moment invariants is very time-consuming, so it can not run in real time. Zhang and Ji [8] proposed a probabilistic framework to combine temporal and spatial information to recognize action units (AUs). Colmenarez et al. [9] presented a Bayesian probabilistic approach to recognizing faces and facial expressions. They possessed the mutual benefits in similarity measures between faces and facial expressions. Qin and He [10] took the technological advantages of both the support vector machine (SVM) and Gabor feature extraction for face recognition. The system proposed in [11] automatically detected frontal faces in an image sequence and classified them into seven classes in real time: neutral, anger, disgust, fear, joy, sadness, and surprise. A facial expression feature stream was used to train a parallel HMM structure in a similar fashion explained in [12], which provided a probabilistic model for temporal recurrent facial expression patterns.

To surmount the shortcomings as stated above, we attempt to develop an automatic facial expression recognition system that detects human faces and extracts facial features from an image sequence. This system is employed for recognizing six kinds of facial expressions: joy, anger, surprise, fear, sadness, and neutral of a computer user. In the expression classification procedure, we mainly compare the performance of different classifiers using multi-layer perceptions (MLPs), SVMs, and AdaBoost algorithms (ABAs). Through evaluating experimental results, the performance of ABAs is superior to that of the other two. According to this, we develop an AdaBoost-based multi-classifier used in our facial expression recognition system.

II. FACE AND FACIAL FEATURE DETECTION

In our system design philosophy, the skin color cue is an obvious characteristic to detect human faces. To begin with, we will execute skin color detection, then the morphological dilation operation, and facial feature detection. Subsequently, a filtering operation based on geometrical properties is applied to eliminate the skin color regions that do not pertain to human faces.

A. Color Space Transformation

Face detection is dependent on skin color detection techniques which work in one of frequently used color spaces. In the past, three color spaces YCbCr, HSI, and RGB have been extensively applied for skin color detection. Accordingly, we extract the common attribute from skin color regions to perform face detection.

The color model of an image captured from the experimental camera is composed of RGB values, but it's easy to be influenced by lighting. Herein, we adopt the HSI color space to replace the traditional RGB color space for skin color detection. We distinguish skin color regions from non-skin color ones by means of lower and upper bound thresholds. Via many experiments of detecting human faces, we choose the *H* value between 3 and 38 as the range of skin colors.

B. Connected Component Labeling

After the processing of skin color detection, we employ the linear-time connected-component labeling technique

proposed by Suzuki et al. [13] to complete the components connected. The following depicts their algorithm which consists of three parts: the first scan, forward scan, and backward scan. We resort to this algorithm for two main benefits: (i) it is based on only sequential local operations, so it does not require a search algorithm to solve label equivalences; (ii) the connectivity is achieved by simply reading and writing a one-dimensional table which stores label equivalences during the scans.

C. Face Region Verification

The detailed steps of face region verification are described in the following:

(i) Component size judgment
Our system deletes all this kind of connected components, if the pixel number of a connected component is smaller than 5,000 or greater than 50,000.

(ii) Aspect ratio judgment
Since the height of a human face is mostly greater than the width, we utilize the aspect ratio to verify face regions; that is, we discard the box with the height smaller than the width. In the light of experiments, the height of a human face is usually greater than or equal to the width and smaller than or equal to three times of the width. These criteria are expressed in (1) to locate probable face regions.

$$B_W \le B_H \text{ and } B_H \le 3B_W \quad (1)$$

where B_H and B_W are the height and width of a circumscribed box, respectively.

(iii) Face region segmentation
Our facial feature extraction method is mainly based on the normal positions of facial features. For example, the mouth lies in the lower half area of a face region, and the eyes lie in the upper half area. Therefore, we must clearly decide the lower boundary of a face region. The following prescribes the lower boundary of a face region.

$$F_L = \begin{cases} F_L, & \text{if } B_H / B_W \le 1.4 \\ F_U + 1.4 \cdot B_W, & \text{if } B_H / B_W > 1.4 \end{cases} \quad (2)$$

where F_L and F_U are the lower and upper boundaries of a face region, individually.

D. Pupils Detection

We exploit the rule that the probable positions of pupils are approximately situated in a face region by 0.5~0.8 time of the height and 0.15~0.85 time of the width referring to the lower left corner of the face region. The tone of their eyes regions is comparatively dark to the skin color. Hence, we can adopt this characteristic to judge the positions of pupils. We can observe that the eyebrows are located above the pupil, and the hair is also above the pupil or on its left side usually. Therefore, we start from the lower right corner of the left half image to leftwards search the first white pixel in row-major order. Then from this pixel towards both the left and up, set up a square region of 10 × 10 pixels. From calculating the center of gravity of the white pixels in this region, the position of the left pupil is received; likewise, we can search out the position of the right pupil.

E. Center of a Mouth Detection

We utilize the rule that the probable position of a mouth lies approximately in the face image by 0.05~0.55 time of the height and 0.2~0.8 time of the width referring to the lower left corner of the face region. According to our observation and experiments, the color of a lip is usually darker than the skin color, and it has a greater red component but a smaller blue one. For locating the region of a mouth, we apply the *H* value to detect lip-colored areas by use of lower and upper bound thresholds. Through doing many experiments, we choose the *H* value between 0 and 6 as the range of lip colors. Finally, from calculating the center of all black pixels, the position of the center of the mouth is acquired.

Ⅲ. FACIAL LANDMARKS EXTRACTION

After detecting a face region and finding its pupils and center of a mouth, we will perform facial landmarks extraction which is the crucial step of an expression recognition system. In general, each facial expression contains plenty of distinctive features. If we distinguish facial expressions by extracting all the distinctive features, it will spend a lot of execution time. Hence, we alternatively choose some landmarks which could represent the changes of facial expressions. It then reduces much computational load to reach our expression recognition system in real time. First, we draw the proper ranges of eyes, mouth, and eyebrows related to the positions of pupils and center of a mouth. Next, we perform both the binarization and edge detection operations on the above ranged images and find 16 landmarks on a human face as shown in Fig. 1 to obtain the characteristic information of facial expressions.

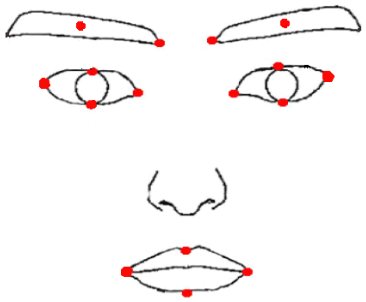


Fig. 1 Facial landmarks on a human face.

A. Landmarks Extraction of Eyes

According to our experiments, we can observe that most of the upper horizontal bounds of eye regions lie beyond the base line through the pupils by 0.2 time of the unit of length, the lower horizontal bounds of eye regions lie below by 0.4 time of the unit of length, the inner vertical bounds of eye regions lie inwards from the central points of the two pupils by 0.33 time of the unit of length, and the outer vertical bounds of eye regions lie outwards by 0.5 time of the unit of length. Fig. 2 shows the proper rectangular ranges of the left and right eye regions.

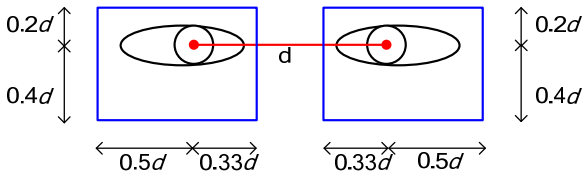
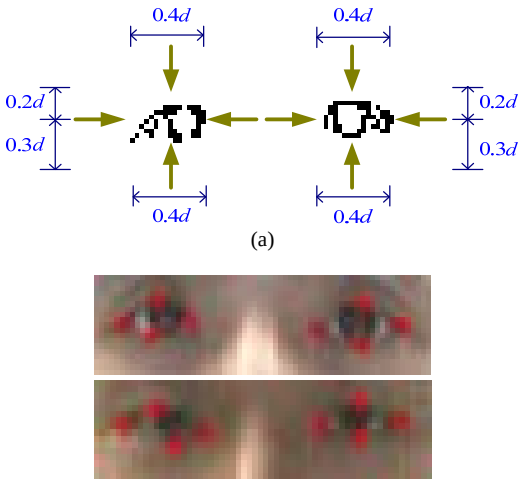


Fig. 2 The rectangular ranges of eye regions.

In the sequel, we have to conduct the binarization and Sobel edge detection processing in these ranges of eye regions. Because our experimental environment is often influenced by varied illumination; for example, day, night or power of the fluorescent lamp, we must alter the thresholds due to environmental changes, and then make our system attain better performance. After this, we carry out the logical "AND" operation on the two binary eye region images that are respectively derived from the binarization and edge detection processing.

Our system would obtain eight landmarks on both eyes from the candidate landmarks, which represent part of facial expression features about eyes. Fig. 3(a) illustrates the range of searching the candidate landmarks of eyes, and some results of extracting the landmarks of eyes are indicated in Fig. 3(b).



(a)

(b)

Fig. 3 Finding the landmarks of eyes: (a) the searching range; (b) some locating results.

B. Landmarks Extraction of Eyebrows

It is a little difficult to extract the landmarks of eyebrows, because eyebrows often appear in different positions for distinct expressions such as anger, sadness, and joy. Hence, the regular position and range of an eyebrow are hard to define. To overcome this, we utilize the color difference of eyebrows and the skin. It can be observed that most of the upper horizontal bounds of eyebrow regions lie beyond the base line through the pupils by 0.75 time of the unit of length, the lower horizontal bounds of eyebrow regions lie beyond by 0.13 time of the unit of length, the inner vertical bounds of eyebrow regions lie inwards from the central points of the two pupils by 0.45 time of the unit of length, and the outer vertical bounds of eyebrow regions lie outwards by 0.6 time of the unit of length. Fig. 4 shows the proper rectangular ranges of the left and right eyebrow regions.

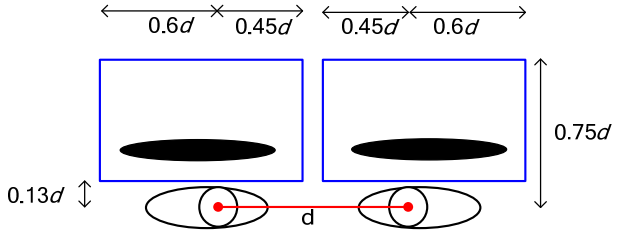
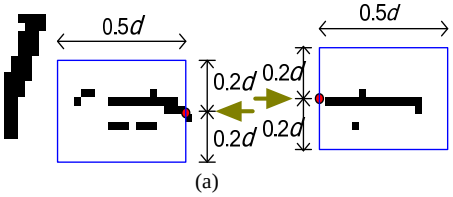


Fig. 4 The rectangular ranges of eyebrow regions.

First of all, we must respectively accomplish the binarization and Sobel edge detection processing in the ranges of eyebrow regions, and perform the logical "AND" operation on the two resulting binary images. The hair usually lies in the periphery or the top of a face, which is prone to disturb eyebrow regions. Therefore, we only define one landmark on the inner rim and another on the center of an eyebrow.

Our system would find four landmarks on a pair of eyebrows from the candidate landmarks, which symbolize part of facial expression features about eyebrows. Fig. 5(a) illustrates the range of searching the candidate landmarks of eyebrows, and some outcomes of extracting the landmarks of eyebrows are indicated in Fig. 5(b).



(a)

(b)

Fig. 5 Finding the landmarks of eyebrows: (a) the searching range; (b) some locating results.

C. Landmarks Extraction of a Mouth

In accordance with our observation and experiments, the lip color is usually darker than the skin color, so we apply this characteristic to extract the landmarks of a mouth. It is also observed that most of the upper horizontal bounds of mouths lie beyond their central points by 0.35 time of the unit of length, the lower horizontal bounds lie below by 0.55 time of the unit of length, the left vertical bounds of mouths lie leftwards from the central points by 0.6 time of the unit of length, and the right vertical bounds lie rightwards by 0.6 time of the unit of length. Fig. 6 shows the proper rectangular range of a mouth region.

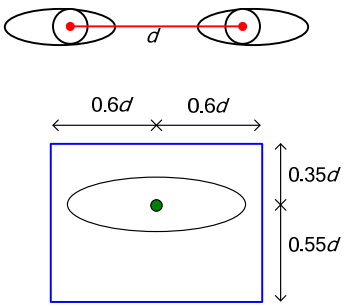
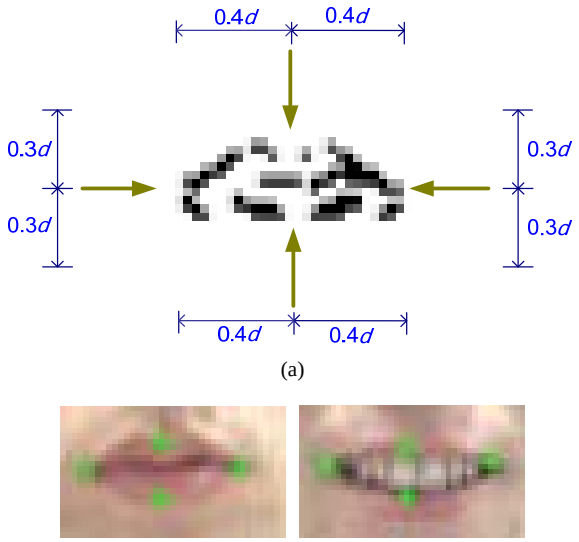


Fig. 6 The rectangular range of a mouth region.

Following that, we first respectively carry out the binarization and Sobel edge detection processing in the range of a mouth. In this phase, we alter the thresholds along with environmental changes to make our system attain better results. Then the logical "AND" operation is performed on the two binary images to result in a refined binary edged image of a mouth.

Our system would acquire four landmarks on a mouth from the candidate landmarks, which stand for part of facial expression features about the mouth. Fig. 7(a) illustrates the range of searching the candidate landmarks of a mouth, and some results of extracting the landmarks of mouths are indicated in Fig. 7(b).



(a)

(b)

Fig. 7 Finding the landmarks of a mouth: (a) the searching range; (b) some locating results.

IV. Facial Expression Recognition

The design philosophy of classification is based on the difference between one kind of facial expressions and a neutral facial expression. From the 16 landmarks of a human face, we compute 16 characteristic distances which represent a kind of expressions. Then we subtract the 16 characteristic distances of a neutral facial expression from those of a certain kind of expressions to acquire its corresponding 16 displacement values. After performance evaluation, we implement a classifier to recognize six kinds of expressions using AdaBoost algorithms (ABAs) rather than multi-layer perceptrons (MLPs) and support vector machines (SVMs) [14].

A. Feature Manipulations

In the light of our observation, the landmarks of eyebrows and mouths will emerge more obvious displacement for the six kinds of facial expressions except the neutral. For example, when the people are joyful, the facial landmarks on the left and right corners of a mouth are raised up and drawn apart to both sides, while the people are angry, the facial landmarks on the inner rims of eyebrows are pressed inwards and downwards. Because the locations of facial landmarks affected by each kind of facial expressions are not the same, in order to recognize facial expressions effectively, we must understand the relationship between a kind of facial expressions and its displacement of the corresponding facial landmarks. Table I shows the distinguishing features of different facial expressions.

Table I
The Distinguishing Features Of Different FaciaL Expressions

Facial Expression	Distinguishing Feature
Joy	1. The corners of a mouth are raised up. 2. The width of a mouth becomes large. 3. Eyes are a little diminished.
Anger	1. Two eyebrows are close to each other. 2. The interval between eyebrows appears vertical lines. 3. Eyes open widely.
Surprise	1. The eyebrows are raised up. 2. The chin is fallen down. 3. The height of a mouth becomes large.
Fear	1. Two eyebrows are close to each other or raised up. 2. The width of a mouth becomes large. 3. Eyes open widely.
Sadness	1. The corners of a mouth are fallen down. 2. Two eyebrows are a little close to each other. 3. Eyes are a little diminished. 4. The upper lip is carried up.

Next, with reference to the 16 landmarks on a human face as shown in Fig. 8, we produce 16 characteristic distances which are the main features used for recognizing facial expressions and calculated in the following way:

$$D^1 = |M_1 - M_2| \tag{3.1}$$

$$D^2 = |M_3 - M_4| \tag{3.2}$$

$$D^3 = |(M_1 + M_2)/2 - M_3| \tag{3.3}$$

$$D^4 = |(M_1 + M_2)/2 - M_4| \tag{3.4}$$

$$D^5 = |EB_1 - EB_3| \tag{3.5}$$

$$D^6 = |EB_2 - EB_4| \tag{3.6}$$

$$D^7 = |EB_1 - P_l| \tag{3.7}$$

$$D^8 = |EB_2 - P_l| \quad (3.8)$$

$$D^9 = |EB_3 - P_r| \quad (3.9)$$

$$D^{10} = |EB_4 - P_r| \quad (3.10)$$

$$D^{11} = |E_3 - E_4| \quad (3.11)$$

$$D^{12} = |E_7 - E_8| \quad (3.12)$$

$$D^{13} = |E_2 - M_1| \quad (3.13)$$

$$D^{14} = |E_6 - M_2| \quad (3.14)$$

$$D^{15} = |E_1 - E_5| \quad (3.15)$$

$$D^{16} = |(P_l + P_r)/2 - M_4| \quad (3.16)$$

where E_1, E_2, …, E_8 are the landmarks of eyes; EB_1, EB_2, …, EB_8 are the landmarks of eyebrows; M_1, M_2, …, M_8 are the landmarks of a mouth; P_l and P_r are the positions of the left and right pupils, respectively.

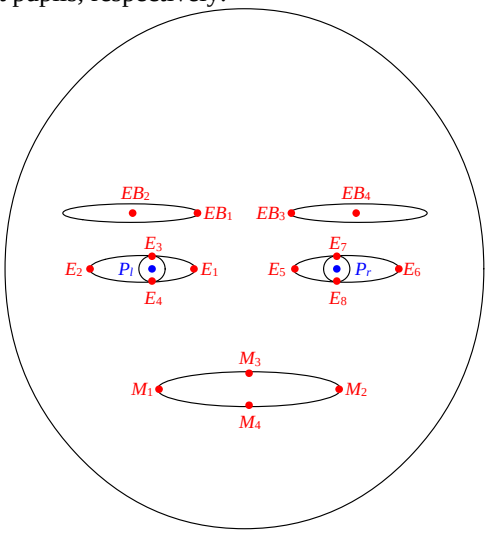


Fig. 8 The 16 landmarks used to generate 16 characteristic distances on a human face.

Because the size of faces extracted by our recognition system is varied, the derived characteristic distances are irregular. Hence, we need to normalize these characteristic distances that can make our recognition system more accurate. First of all, we take the distance d between two pupils as the unit of length, because it will not be changed with different facial expressions. All the original 16 characteristic distances are divided by the unit of length to obtain 16 normalized characteristic distances expressed as follows.

$$D^{i'} = D^i / d \quad (4)$$

where D^i, i=1, 2, …,16 are the original characteristic distances.

And then we apply these characteristic distances to get the displacement values that are fed to the classifier. In a sequence of human face images with neutral facial expressions, say 10 frames, we compute a mean value which is saved as a reference value for each normalized characteristic distance, and subtract the 16 reference values from the 16 normalized characteristic distances in the subsequent frames as depicted in Eq. (5). Such 16 displacement values act as the facial expression features inputted to our recognition system.

$$S_j^{i'} = D_j^{i'} - D_r^{i'},\ j = 11, 12, \ldots \quad (5)$$

where $D_j^{i'}$ is the i-th normalized characteristic distance in the j-th frame and $D_j^{i'}$ is the i-th reference characteristic distance.

B. The AdaBoost Algorithm

The AdaBoost algorithm (ABA) was proposed in the literature of computational learning theory in 1996 [15]. It has two different versions: one is used for binary classification problems and the other is to deal with the problems with more than two classes. The ABA generates a hypothesis whose error on the training set is small by combining many hypotheses whose errors may be large (but still better than random guessing). Fig. 9 is the generalized version of the ABA for binary classification problems.

Given a training sample set:

$\tilde{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ with $x_i \in X$ and $y_i \in \{-1, +1\}$.

Initialize the distribution: $\tilde{D}_1^{(i)} = 1/m, i = 1,2,\ldots,m$.

For t = 1,2,…,T:

(1)Train the weak classifier using the distribution $\tilde{D}_t^{(i)}, i = 1,2,\ldots,m$.

(2)Get the weak hypothesis $g_t : X \to \{-1,+1\}$.

(3)Update the distribution

$$\tilde{D}_1^{(i)} = \tilde{D}_t^{(i)} \exp(-\eta_t y_i g_t(x_i)) / Z_t, i = 1,2,\ldots,m$$

where Z_t is a normalization factor (guaranteed that $\tilde{D}_{t+1}^{(i)}$ is still a distribution)

and $\eta_t = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t}$ with $\varepsilon_t = \sum_{i=1}^{m} \tilde{D}_t^{(i)} [y_i \neq g_t(x_i)]$.

End For

Output the final hypothesis:

$$G(x) = sign\left(\sum_{t=1}^{T} \eta_t g_t(x)\right).$$

Fig. 9 A generalized version of the ABA.

C. The Weak Classifier

The weak classifier is the essential part of an ABA. Each weak classifier produces the answer “yes” or “no” for particular features. The ABA is very flexible, and can be improved by combining a sequence of weak classifiers, each of whose associated conditional probabilities is determined by the output of the previously tuned weak classifiers. Such boosting and random subspace methods have been designed as decision trees, where they often produce an ensemble of classifiers, which is superior to a single classification rule. Herein, we adopt Classification and Regression Trees (CARTs) as the structure of a weak classifier tuned by the ABA.

The classical CART algorithm was proposed by Breiman et al. in 1984 [16]. It builds a binary decision tree which splits

a single variable at each node for predicting continuous dependent variables (regression) and categorical predictor variables (classification). The leaves and nodes of the decision tree represent the results of classification and the prediction rules, respectively. The CART algorithm conducts a thorough search recursively for all variables whose values are categorized into two groups using a threshold to find out an optimal splitting rule for each node. We can regard the classification via a decision tree as a tree traversal process. A node of the CART is constructed by use of the following rules. Given a training sample set $\tilde{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, where each x_i belongs to an instance space $X \in R^n$ (each vector with dimensionality n; $x_i=(x_{i1}, x_{i2}, \ldots, x_{in})$) and each label y_i belongs to a finite label space $Y \in \{-1, 1\}$:

Rule 1. For each feature (all dimensions), determine a threshold which separates the sample set $\tilde{S}$ with a minimal classification error.

Rule 2. Select the *j*-th feature with the minimal error and build a CART node.

(i) Set up the branch condition: $\xi_j > threshold_j$.

(ii) Arrange the branches that are connected with leaves to perform respective classification.

And suppose that the classification error associated with a leaf is the probability of a sample being misclassified. We stop the tree traversal when encountering a misclassification. The whole CART is constructed by means of the following steps:

Step 1. Construct the root of a CART with the minimal error node.

Step 2. Select the leaf with the largest error.

Step 3. Construct a node using only those samples which are associated with the chosen leaf.

Step 4. Replace the chosen leaf by the new constructed node.

Step 5. Repeat Steps 2-4 until all leaves have no error or reach the predefined conditions.

Fig. 10 illustrates an example of the CART where four nodes are constructed.

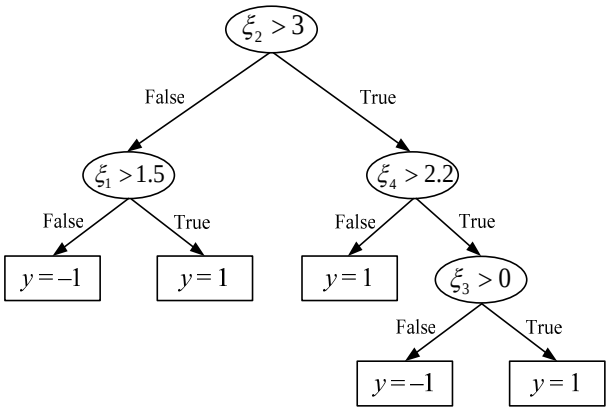


Fig. 10 Illustration of a 4-split CART.

D. Our Proposed AdaBoost-Based Multi-Classifier

Because the ABA is primarily applied to a binary classifier, we further develop a bottom-up hierarchical classification structure for the recognition of multi-class facial expressions. This structure is possessed of a good property that we can update the models by training only new added data, without modifying the whole models trained earlier. Fig. 11(a) shows an AdaBoost-based classifier M_1 that recognizes two kinds of expressions.

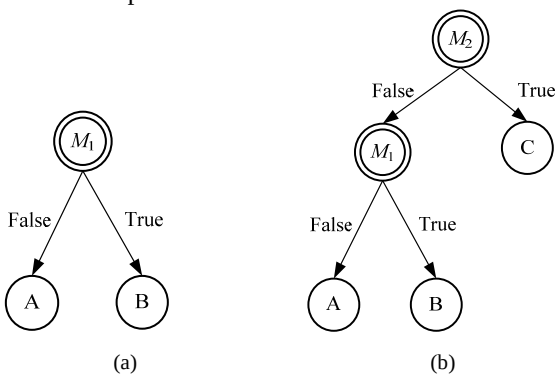


Fig. 11 Illustration of classifier expansion: (a) an AdaBoost-based binary classifier; (b) an AdaBoost-based ternary classifier from expanding (a).

When we add another kind of expressions, we just construct the node M_2 by taking both the expressions "*A*" and "*B*" as the negative samples and the expression "*C*" as the positive sample as shown in Fig. 11(b). Hence, we can utilize this structure to recognize three kinds of facial expressions. From this example, we can see that when a new expression put to the structure, we just construct at most two nodes. By feeding the features of a facial expression to the strong classifier trained with the ABA, we can acquire a weight of the final prediction. The facial expressions will be recognized if this weight is positive with respect to one kind of expressions by the strong classifier. In our facial expression recognition system, there are six kinds of expressions, including joy, anger, surprise, fear, sadness, and neutral, to be classified. According to the extracted features, the classification of facial expressions is sometimes ambiguous. Therefore, the seventh leaf standing for the "Other" kind of expressions is required in our AdaBoost-based multi-classifier as shown in Fig. 12.

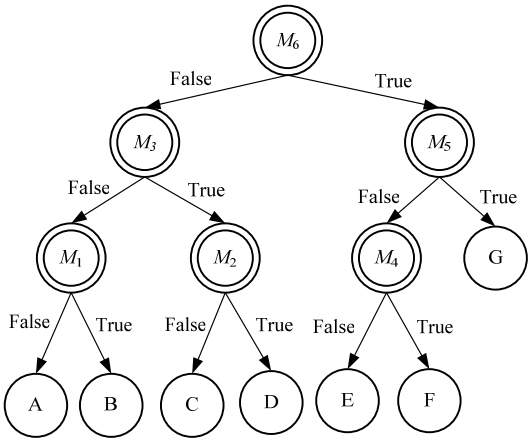


Fig. 12 An example of a multi-classifier for discriminating seven classes.

V. Experimental results

The experimental results consist of three main parts: face detection using skin color segmentation, facial expression recognition using an ABA, and sequential composite expressions recognition using the ABA together with the "Majority voting" scheme. The hardware and software used in these experiments are listed in Table II.

Table II
The Hardware And Software used In The Facial Expression Recognition System

Hardware	Software
CPU: Pentium4 3.2GHz	Tool: Borland C++ Builder 6.0
RAM: 512MB	MATLAB 7.2
Camera: Logitech Quick-Cam Pro 4000	Operating System: Microsoft Windows XP

The place where we completed the experiments is at National Taiwan University of Science and Technology. The subjects are all the graduate students in the Image Processing and Pattern Recognition Laboratory. In the first of this section, we will illustrate our facial expression database. Then we will show the face detection results from captured image sequences. Subsequently, we will compare three different classifiers of recognizing facial expressions using MLPs, SVMs, and ABAs. Finally, we will show the sequential composite facial expressions recognition result from various image sequences.

A. The Facial Expression Database

Up to now, no standard database has been generally acknowledged by international researchers in the field of facial expression recognition, but there are some databases commonly used in experiments; for example, Database Japanese Female Facial Expressions (JAFFE) [17], Cohn-Kanade Facial Expression Database [18], Ekman-Hager Facial Action Exemplars [19], and The CMU Pose, Illumination, and Expression (PIE) Database of Human Faces [20]. Of these databases, some have a single image frame of facial expressions but not continuous image sequences, and some have gray images but not colored images. They are not all suitable for testing our facial expression recognition system. Therefore, we set up one small-scale database of facial expressions by ourselves using the web camera "Logitech Quick-Cam Pro 4000" to take image sequences with the resolution of 320 × 240 pixels.

In addition, this database employed in our system is different from the ordinary databases of facial expressions. Such databases usually store static face images one by one. In consequence, the facial features are extracted from only a single face image of their databases each time. On the contrary, in our system we directly extract facial features from an image sequence without storing image frames. This method could not only accelerate the speed of setting up the facial expression database, but save the waste of the hard disk space.

At present, the facial features of ten persons (eight males and two females) are stored in our database, and each person has 1,200 materials comprising joy, anger, surprise, fear, sadness, and neutral, each of which contains 200 materials. That is, there are 12,000 materials totally in our facial expression database. Fig. 13 shows some image samples of six kinds of expressions, where the faces may be panned from -30° to 30° and tilted from -10° to 10°.

Fig. 13 The training image samples of six kinds of expressions.

B. The Results of Face Detection

The face detection procedure is accomplished by the following processes in order: skin color detection, morphological operation, connected component labeling, component size judgment, aspect ratio judgment, and proper face region segmentation. The course of this procedure probably costs 0.06 seconds. Some face detection results are shown in Fig. 14.

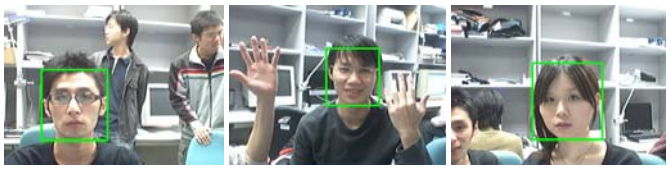


Fig. 14 The results of face detection in complex backgrounds.

We perform the face detection experiments in three image sequences of 300, 400, and 500 continuous frames, respectively. There are three different subjects including two males and one female in these sequences. The correct rates of face detection of these three image sequences are almost identical. In addition, we define the "Error rate" as the percentage of regarding an inhuman face as a human face, and the "Miss rate" is the percentage which someone's face appears in an image sequence but the system has not detected it out. Table III shows the results of face detection rates of the above experiments. The reason why the errors of face detection occur is that the light is insufficient or the colors of some regions are close to the skin color. Consequently, all the erroneous face candidates may not meet the conditions such as the aspect ratio and the size of the box bounding a face.

TABLE III
THE STATISTICS OF FACE DETECTION RATES

Exp.	The Total Number of Faces	The Number of Detected Faces	Correct Rate	Error Rate	Miss Rate
I	300	295	98.33%	1.30%	0.37%
II	400	393	98.25%	1.43%	0.32%
III	500	492	98.40%	1.25%	0.35%

C. The Results of Facial Expression Recognition

Our system is designed to recognize facial expressions on an image sequence, but actually, the recognition is achieved by judging a single image each time. Herein, we mainly compare the performance of different classifiers using MLPs, SVMs, and ABAs, and then conclude their pluses and minuses. Since all these machine learning methods are supervised ones, we have to acquire some samples to train the classifiers. During the training, we employ 10-fold cross-validation to estimate the accuracy of different system models.

Owing to the limited length of a piece of writing, we only describe our AdaBoost-based multi-classifier below. Because the ABA we adopt is a binary classifier, we propose a bottom-up hierarchical classification structure consisting of properly arranged ABAs for facial expression recognition. Such a decision tree of recognizing the six kinds of expressions is similar to that based on SVMs [21]. The total training time for the condition of 12,000 samples by means of 16 CART splits and 300 iterations of the ABA is about 4 minutes. The detailed experimental data are recorded in Table IV. And we can see that the accuracy rates of recognizing expressions are better and evener than those resulting from SVMs.

TABLE IV
THE FACIAL EXPRESSION RECOGNITION RESULTS FROM ABAS

Expression type	Recognition result						
	Neutral	Joy	Anger	Surprise	Fear	Sadness	Other
Neutral	1923	1	29	2	10	28	7
Joy	21	1929	0	1	35	3	11
Anger	7	1	1971	0	15	3	3
Surprise	0	0	0	1995	5	0	0
Fear	4	119	1	1	1835	15	24
Sadness	26	8	44	0	9	1907	6
Other	2	3	4	1	8	6	1976

To further show the performance of the above experiments, we will introduce the definition of precision and recall rates as depicted in Table V. It means that the most expressions can be classified correctly.

TABLE V
DEFINITION OF PRECISION AND RECALL RATES

Notation	Definition
Precision	$\frac{\text{True positive}}{\text{True positive} + \text{False positive}}$
Recall	$\frac{\text{True positive}}{\text{True positive} + \text{False negative}}$
True positive	$\text{Result} \cap \text{Ground truth}$
False positive	$\text{Result} \cap \overline{\text{Ground truth}}$
False negative	$\overline{\text{Result}} \cap \text{Ground truth}$

TABLE VI
THE RECALL AND PRECISION RATES OF THE THREE CLASSIFIERS

Expression Type	MLP	
	Recall Rate	Precision Rate
Neutral	96.6%	96.6%
Joy	94.7%	94.7%
Anger	95.9%	95.9%
Surprise	99.7%	99.7%
Fear	93.8%	93.8%
Sadness	93.5%	93.5%
Other	95.4%	95.4%
Expression Type	**SVM**	
	Recall Rate	Precision Rate
Neutral	94.2%	94.2%
Joy	91.5%	91.5%
Anger	94.3%	94.3%
Surprise	99.9%	99.9%
Fear	86.1%	86.1%
Sadness	88.2%	88.2%
Other	91.3%	91.3%
Expression Type	**ABA**	
	Recall Rate	Precision Rate
Neutral	96.2%	96.2%
Joy	96.5%	96.5%
Anger	98.6%	98.6%
Surprise	99.8%	99.8%
Fear	91.8%	91.8%
Sadness	95.4%	95.4%
Other	98.8%	98.8%

In the calculation, 2,000 materials are viewed as the correct samples, and the other 10,000 materials are regarded as the wrong samples. Table VI records the precision and recall rates of the above experiments, and Table VII summarizes the system performance using three different classification techniques. By observing Tables VI and VII, we can find that the average recognition rates obtained from both the MLPs and ABAs are better than that from SVMs. Especially, the accuracy rates are raised for recognizing the expressions “Joy,” “Fear,” and “Sadness.” And we inspect

that all the recognition rates received from MLPs are very even, but the training of MLPs takes quite a long time. Except the accuracy rates of recognizing expressions "Neutral" and "Fear," the other accuracy rates obtained from the ABAs are superior to those from the MLPs.

On the other hand, we compare SVMs with ABAs, and the performance of the former is worse than that of the latter. It is due to the ABAs constituting a strong classifier composed of some weak classifiers which have greater adaptability. The goal of an SVM is to find the best hyperplane to group the input data into two classes. It can act as a weak classifier used in the ABAs. On the whole, the classification result obtained from the ABAs is better than that from the SVMs. In consequence, we choose ABAs as the classification method to realize our facial expression recognition system.

TABLE VII
SYSTEM PERFORMANCE OF THE THREE CLASSIFIERS

Classification Technique	MLP	SVM	ABA
Average Recognition Rate	95.7 %	92.2 %	96.7 %
Average Training Time	25 Min	11 Min	8 Min

The following tests our system on image sequences. Each of which just has two kinds of expressions. Table VIII lists the test image sequences of composite facial expressions. To classify multiple kinds of expressions in a single image sequence, we report the classification result for each process unit. Herein, we simply treat a single frame as a process unit which is classified into a kind of expressions. The classification result of a process unit which is parenthesized by parentheses stands for the change of expressions in an image sequence. Therefore, the process unit at such a moment easily makes the recognition system ambiguous.

TABLE VIII
TEST IMAGE SEQUENCES OF COMPOSITE FACIAL EXPRESSIONS

Composite expression type	Video label
"Neutral" and "Joy"	N-J
"Surprise" and "Anger"	Sur-A
"Anger" and "Fear"	A-F
"Neutral" and "Sadness"	N-Sad
"Surprise" and "Joy"	Sur-J
"Fear" and "Sadness"	F-Sad

In Table IX, the expression replied from the system, which is printed in a font of boldface, means a misclassification result. As mentioned above, except the failure in classifying the process unit at the moment of expression changes, the other failures are caused by the high similarity between two kinds of expressions, especially for the expressions "Joy" and "Fear."

TABLE IX
CLASSIFICATION RESULTS OF SEQUENTIAL COMPOSITE FACIAL EXPRESSIONS

Video label	Sequential classification result
N-J	Neutral Neutral Neutral Neutral (**Other**) Joy Joy Joy Joy Joy
Sur-A	Surprise Surprise Surprise Surprise Surprise (Anger) Anger Anger **Other** Anger
A-F	Anger Anger Anger Anger (**Other**) Fear **Joy** Fear Fear Fear
N-Sad	Neutral Neutral Neutral Neutral Neutral (Sadness) Sadness Sadness Sadness Sadness
Sur-J	Surprise Surprise Surprise Surprise Surprise (**Fear**) Joy Joy Joy Joy Joy Joy Joy
F-Sad	Fear Fear Fear Fear Fear Fear (Sadness) **Neutral** Sadness Sadness Sadness Sadness

In this experiment, the total number of test process units is 65 and the number of process units correctly classified is 59. The correct classification rate is about 90.7%. Fig. 15 shows an example frame of the test image sequences, each of which only has two kinds of facial expressions for simplifying demonstration.

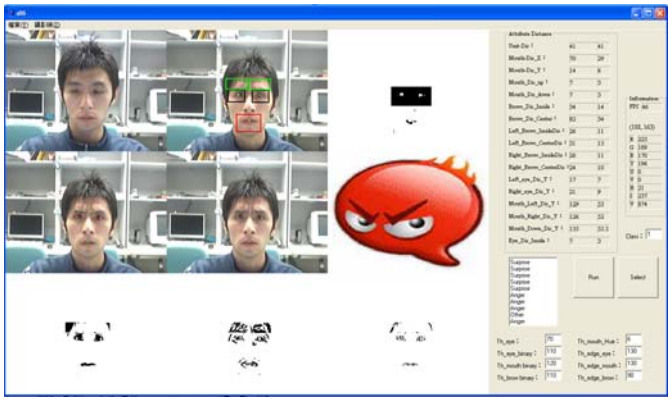


Fig. 15 Demonstration of the facial expression changing from "Surprise" to "Anger."

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, we have presented a highly automatic facial expression recognition system in which a face detection procedure is first able to detect and locate human faces in image sequences acquired in real environments. We need not label or choose characteristic blocks in advance. In the face detection procedure, some geometrical properties are applied to eliminate the skin color regions that do not belong to human faces. It requires no too much miscellaneous calculation and could accelerate the processing speeds of the facial expression recognition system. In the facial feature extraction procedure, we only perform both the binarization and edge detection operations on the proper ranges of eyes, mouth, and eyebrows to obtain the 16 landmarks of a human face to further produce 16 characteristic distances which represent a kind of expressions. It can effectively reduce the influence of noises originated from the other ranges and lower the wrong situation of extracting the landmarks to increase the recognition rate of the whole system.

During the development of the facial expression classification procedure, we evaluate three machine learning methods: MLPs, SVMs, and ABAs. We combine ABAs with CARTs, which selects weak classifiers and integrates them into a strong classifier automatically. It not only takes less training time than the other machine learning methods do, but also enhances the classification capability. Thus, we can update training samples to handle different situations, but need not spend much computational cost. According to these, we select the ABA as the classifier of the facial expression recognition system. The throughput obtained is from 5 to 8 frames per second, and the performance of the system is very satisfactory, whose recognition rate achieves more than 90%. Hence, the facial expression recognition system we proposed is quite closed to a real-time facial expression recognition one.

Some future works are worth investigating to attain better performance. In our current feature extraction procedure, only color and edge cues are adopted, and we will focus on adding some other cues such as the texture features of a human face to make it more robust. In the facial expression classification procedure, if the number of expressions that should be recognized increases, the execution time will also increase with it. We will replace the AdaBoost-based binary classifier by the one with the ability of classifying more than two classes to overcome this problem. Moreover, in the facial expression classification procedure, the crux of the matter is that people's expression changes usually have continuity with the elapsed time. If we can consider the time information in this procedure, it will raise the whole reliability of the facial expression recognition system. In the near future, we will employ the techniques of hidden Markov models (HMMs) [22] or conditional random fields (CRFs) [23] for improving the accuracy of facial expression recognition.

ACKNOWLEDGMENT

The authors would like to thank the National Science Council of Taiwan for her support in part under Grant NSC95-2213-E-011-105.

REFERENCES

[1] Y. Sakagami et al., "The intelligent ASIMO: System overview and integration," in *Proc. of the IEEE/RSJ Int. Conf. on Intell. Robots Syst.*, Saitama, Japan, vol. 3, pp. 2478-2483, 2002.

[2] G. L. Foresti, C. Micheloni, L. Snidaro, and C. Marchiol, "Face detection for visual surveillance," in *Proc. of the 12th IEEE Int. Conf. on Image Anal. Process*., Udine, Italy, pp. 115-120, 2003.

[3] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proc. of the IEEE Int. Conf. on Automat. Face Gesture Recogni.*, Sophia Antipolis, France, pp. 454-459, 1998.

[4] M. H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. on Pattern Anal. Machine Intell.*, vol. 24, no. 1, pp. 34-58, 2002.

[5] G. Yang and T. S. Huang, "Human face detection in a complex background," *Pattern Recogni.*, vol. 27, no. 1, pp. 53-64, 1994.

[6] K. C. Yow and R. Cipolla, "Feature-based human face detection," *Image Vision Comput.*, vol. 15, no. 9, pp. 712-735, 1997.

[7] Y. Zhu, L. C. De Silva, and C. C. Ko, "Using moment invariants and HMM in facial expression recognition," in *Proc. of the 4th IEEE Southwest Symp. on Image Anal. and. Interpret.*, Singapore, pp. 305-309, 2000.

[8] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Trans. on Pattern Anal. Machine Intell.*, vol. 27, no. 5, pp. 699-714, 2005.

[9] A. Colmenarez, B. Frey, and T. S. Huang, "A probabilistic framework for embedded face and facial expression recognition," in *Proc. of the Int. Conf. on Comput. Vision Pattern Recogni.*, New York, NY, pp. 592-597, 1999.

[10] J. Qin and Z. S. He, "A SVM face recognition method based on Gabor-featured key points," in *Proc. of the 4th Int. Conf. on Mach. Learning Cybern.*, Chongqing, China, pp. 18-21, 2005.

[11] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction," in *Proc. of the Int. Conf. on Comput. Vision Pattern Recogni. Worksho*p, San Diego, CA, vol. 5, pp. 53-58, 2003.

[12] M. E. Sargin et al., "Prosody-driven head-gesture animation," in *Proc. of the Int. Conf. on Acoust., Speech, and Signal Process.*, vol. 2, Istanbul, Turkey, pp. 677-680, 2007.

[13] K. Suzuki, I. Horiba, and N. Sugie, "Linear-time connected-component labeling based on sequential local operations," *Comput. Vision Image Understand.*, vol. 89, no. 1, pp. 1-23, 2003.

[14] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 3rd Ed., San Diego: Academic Press, 2006.

[15] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. of the 13th Int. Conf. on Mach. Learning*, Bari, Italy, pp. 148-156, 1996.

[16] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Boca Raton: Chapman and Hall, 1984.

[17] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. of the 3rd IEEE Int. Conf. on Automat. Face Gesture Recogni.*, Nara, Japan, pp. 200-205, 1998.

[18] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. of the 4th IEEE Int. Conf. on Automat. Face Gesture Recogni.*, Pittsburgh, PA, pp. 46-53, 2000.

[19] P. Ekman, J. Hager, C. H. Methvin, and W. Irwin, "Ekman-Hager facial action exemplars," unpublished.

[20] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database of human faces," *Tech. Report CMU-RI-TR-01-02*, Robotics Inst., Carnegie Mellon Univ., Pittsburgh, PA, 2001.

[21] M. H. Wu, "A facial expression recognition system based on the facial landmarks extracted from an image sequence," *Master Thesis*, Dept. of Comput. Sci. and Inform. Eng., Nat. Taiwan Univ. of Sci. and Tech., Taipei, Taiwan, 2008.

[22] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE Acoust. Speech Signal Process. Mag.*, vol. 3, no. 1, pp. 4-16, 1986.

[23] Y. Wang and Q. Ji, "A dynamic conditional random field model for object segmentation in image sequences," in *Proc. of the IEEE Comput. Soci. Conf. on Comput. Vision Pattern Recogni.*, vol. 1, New York, NY, pp. 264-270, 2005.