Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 1173–1185
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Hybrid-boost learning for multi-pose face detection and facial expression recognition☆

Hsiuao-Ying Chen, Chung-Lin Huang*,[1], Chih-Ming Fu

Electrical Engineering Department, National Tsing-Hua University, Hsin-Chu, Taiwan

Received 3 November 2006; received in revised form 15 August 2007; accepted 17 August 2007

Abstract

This paper proposes a hybrid-boost learning algorithm for multi-pose face detection and facial expression recognition. To speed-up the detection process, the system searches the entire frame for the potential face regions by using skin color detection and segmentation. Then it scans the skin color segments of the image and applies the weak classifiers along with the strong classifier for face detection and expression classification. This system detects human face in different scales, various poses, different expressions, partial-occlusion, and defocus. Our major contribution is proposing the weak hybrid classifiers selection based on the Harr-like (local) features and Gabor (global) features. The multi-pose face detection algorithm can also be modified for facial expression recognition. The experimental results show that our face detection system and facial expression recognition system have better performance than the other classifiers.
© 2007 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Pattern classification; Adaboost; Hybrid-boost; Face detection; Face expression recognition

1. Introduction

Automatic face detection has many applications such as surveillance, human computer interface (HCI). Most of the published methods assume: *front-view pose, minimum out-of-plane head motion*, and *constant illumination* of which the illumination variation is the most difficult one. *Accuracy* and *efficiency* are two of the most important issues in evaluating a face detection system. Most of the previous face detection systems focus on the eyes as the most prominent feature of the face. Instead of treating the face detection as a binary classification problem, we propose a multi-class hybrid-boost learning algorithm which selects the most discriminative Gabor features and Harr-like features for multi-pose face detection and expression identification.

Most of the previous face detection researches [1–12] have many restrictions, such as no varying pose nor noisy defocus problem. Human face detection algorithms rely on the extracted facial features. The detected feature vector can also be applied for identifying the face in different poses and expressions. Viola et al. [3] introduce Adaboost with a cascade scheme and apply an integral image concept for face detection. They propose two-class AdaBoost learning algorithm for training efficient classifiers and a cascaded structure for rejecting non-face images.

Huang et al. [7] propose a novel tree-structured multi-view face detector (MVFD) called Vector Boosting, using the coarse-to-fine strategy to divide the entire face space into smaller and smaller subspaces. They developed a Width-First-Search (WFS) tree structure to achieve higher performance in both speed and accuracy. Li et al. [8] introduce the FloatBoost by using the floating search algorithm. There are basically three kinds of feature selection methods: Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), and Sequential Floating Search Method (SFSM). FloatBoost algorithm [11] uses SFSM to select features and the training time is five times longer than AdaBoost. Xu et al. [12] propose an MRC-boosting algorithm which may compute the most discriminative feature in close-form.

☆ This revised version is submitted to *Pattern Recognition* for peer review.

* Corresponding author. Tel.: +886 3 574 2585.
E-mail address: clhuang@ee.nthu.edu.tw (C. Huang).
[1] The corresponding author is also with Department of Informatics, Fo-Guang University, I-Lan, Taiwan.

0031-3203/$30.00 © 2007 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2007.08.010

Fig. 1. Examples the potential face regions.

Besides face detection in the image, there are other research interests regarding to the facial expression extraction [13–15]. Datcu et al. [16] propose a novel facial expression recognition method by using Relevance Vector Machine (RVM) [17,18]. In Ref. [19], Zhao et al. present a new texture modeling based on volume local binary patterns (VLBP) which can be applied for analyzing so-called *dynamic textures* and facial expression. The face expression extraction methods can be divided as *local analysis* [20,24] and *global analysis* [21,22]. The former focuses on some specific feature points and divides the face into three local graphical objects: right eye graph, left eye graph, and mouth graph. The latter determines the features by processing the entire face by boosting Harr feature based weak classifier.

Adaboost algorithm has also been widely applied for real time facial expression recognition [21,22]. Silapachote et al. [23] proposes a classification technique for face expression recognition using Adaboost to select the relevant global and local appearance feature with the most discriminant information. The other methods [20,24] analyze the internal representation of facial expressions based on collections of Action Units (AUs). The local analysis needs some additional verification step to avoid feature errors. Improper feature points deteriorate the recognition accuracy, and more feature points require more computation to fit the model to the face image. These restrictions make the systems more complicated and inadequate for real-time processing. Kanade et al. [25] present the CMU AU-coded face expression image database which is the most comprehensive testbed for comparative studies of facial expression.

Here, we propose a hybrid-boost learning which selects Gabor features (for global appearance) and Harr-like features (for local appearance) to provide the most discriminating information for the strong classifier in the final stage. Our face detection system locates the potential face regions by using skin color detection and segmentation, and then searches for the hybrid features for the multi-class strong classifier to detect the multi-pose face and different facial expressions. Our system is robust to various size, poses, expressions, and defocus problems. The experimental results show that our method has a better system performance than the other methods.

2. Segmentation of potential face regions

The 1st module, *potential face regions* segmentation, consists of skin color detection and segmentation. To identify the existence of human face, it scans the image to detect the skin color regions and remove unnecessary pixels. To reduce the search region, we need to locate the possible face region. In the captured human face images, we assume that the color distribution of human face is somehow different from that of the image background. Pixels belonging to face region exhibit similar chrominance values within and across different races [26]. However, the color of face region may be affected by different illumination. For skin-color detection, we analyze the color of the pixels in RGB color space to decrease the effect of illumination changes, and then classify the pixels into face-color or non-face color based on their hue component only.

Similar to Ref. [27], we analyze the statistics of skin color and non-skin color distributions from a set of training data to obtain the conditional probability density functions of skin color, and non-skin color. From 100 training images, we have the probability density function of color $\boldsymbol{c} = (r, g, b)$, which can be either face color and non-face color (i.e., $p(\boldsymbol{c}|face)$ and $p(\boldsymbol{c}|non\text{-}face)$). Based on color statistics, we use the Bayesian approach to determine the face-color region. Each pixel is assigned to the *face* or *non-face* class that gives the minimal cost when considering cost weightings on the classification decisions. The classification is performed by using Bayesian decision rule which can be expressed as: if $p(\boldsymbol{c}(i)|face)/p(\boldsymbol{c}(i)|non\text{-}face) > \tau$, then pixel i (with $\boldsymbol{c}(i) = (r(i), g(i), b(i))$ belongs to a face region, otherwise it is inside a non-face region, where $\tau = p(non\text{-}face)/p(face)$. After applying Bayesian classification on another 100 testing face images, we find that 90% of the correct classified facial pixels satisfying four constraints for the human skin color segmentation: (1) $r(i) > \alpha$, (2) $\beta_1 < (r(i) - g(i)) < \beta_2$, (3) $\gamma_1 < (r(i) - b(i)) < \gamma_2$, and (4) $\sigma_1 < (g(i) - b(i)) < \sigma_2$. Here, we choose $\alpha = 100$, $\beta_1 = 10$, $\beta_2 = 70$, $\gamma_1 = 24$, $\gamma_2 = 112$, $\sigma_1 = 0$ and $\sigma_2 = 70$.

Then, we may apply the color thresholding followed by pixels grouping on the quantized face color regions. A merging stage is then iteratively performed on the set of homogeneous skin

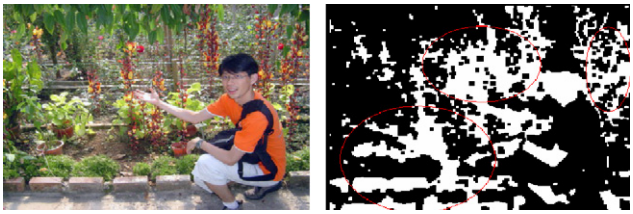

Fig. 2. Non-face regions are detected due to complex background and un-even illumination.

color pixels to provide a contiguous region as the candidate face area. Constraints of shape and size of face region are applied on each candidate face area to find the potential human faces. Since there may many small regions and holes in large regions, the next step is to perform the morphological filtering operation [28] to remove small areas and eliminate holes to get better potential face regions. As shown in Fig. 1, our face segmentation scheme is robust to various illumination conditions. It performs well for the darker outdoor image as well as the brightened indoor image.

However, under the conditions of complex background or uneven illumination, the results are not reliable. Then, we may skip the skin-color segmentation and search for the entire image frame to detect a human face. There is a trade-off between the missing rate (missing the real face regions) and the false alarm rate (non-face region mistreated as a potential face region). Here, we find the classification threshold to produce a higher false alarm rate and a lower missing rate (as shown in Fig. 2). Therefore, we need to use the following face detection algorithm to search more potential face regions for the real faces.

3. Hybrid-boost learning for face detection

Here, we propose the hybrid-boost learning which iteratively chooses the weak classifiers that minimize the exponential loss function. The weak classifiers consist of Gabor features and Harr-like features which characterize the salient visual properties such as spatial localization, orientation selectivity, and spatial frequency characteristics of the human faces. The Harr-like features are easier to be obtained than the Gabor features. Similar to Adaboost, the hybrid-boost learning algorithm will select a small number of weak classifiers to construct a strong classifier.

3.1. Hybrid feature pool

The potential face regions may have different sizes. To reduce the effects of variation in the distance and location, the input training images are normalized to 24 × 24 blocks. The object recognition system finds various features of the object and builds up a local neighborhood representation for each one of the selected features. Two related problems are involved in this process: (i) which features of the object should be used, and (ii) how to represent the information contained in their neighborhood [5]. There are many different type features, such as edges, corners, Gaussian derivatives, Gabor features, etc. Here, we propose a hybrid feature consisting of Gabor features (global feature) and Harr-like features (local features). The local features are acquired in the various-sized blocks, however, the global features are obtained in the normalized 24 × 24 blocks. The local features include the width and length of the block, whereas the global features include more detailed information of frequency and the orientation.

(a) *Gabor features*: The 2D isotropic Gabor function g is the product of a 2D Gaussian and a complex exponential function expressed as

$$g_{\theta,\lambda,\sigma}(x, y) = \exp\left\{-\frac{x^2+y^2}{2\sigma^2}\right\}\exp\left\{\frac{j\pi}{\lambda}(x\cos\theta + y\sin\theta)\right\}, \tag{1}$$

where θ represents the orientation, λ is the wavelength, σ denotes the scale. We use a parameter $\gamma = \lambda/\sigma$ instead of λ so that a change in σ corresponds to a true scale change in the Gabor function. Also, it is convenient to apply a 90° counterclockwise rotation, such that θ expresses the orthogonal direction to the Gabor function edges. Therefore, we define Gabor function as follows:

$$g_{\theta,\gamma,\sigma}(x, y) = \exp\left\{-\frac{x^2+y^2}{2\sigma^2}\right\}\exp\left\{\frac{j\pi}{\gamma\sigma}(x\sin\theta - y\cos\theta)\right\}. \tag{2}$$

By changing the parameters, we have different Gabor functions as shown in Fig. 3. By convolving a Gabor function with image pattern, we can evaluate their similarity based on the Gabor response. To emphasize different types of image characteristics, we vary the parameters σ, γ and θ of the Gabor function. The variation of θ changes the sensitivity to different edge and texture orientations. The variations of σ represent different "scales", and the variations of γ denote different sensitivity to high/low frequencies.

(b) *Harr-like features*: The Harr-like features is widely used by Adaboost learning algorithm [3]. More and more analysis use the rectangle features because these features consider a local region for the face and require less computation than the other features. The sum of the pixels which lie within the white rectangles are subtracted from the sum of pixels in the black rectangles. There are edges features, line features, and center-surround features. Harr-like features can be computed rapidly by using an intermediate representation called the integral image [3] as

$$ii(x, y) = \sum_{x'<x, y'<y} i(x', y'), \tag{3}$$

where $ii(x, y)$ is the integral image and $i(x, y)$, is the original image. Using the integral image, any rectangular sum can be computed in four array references. Clearly the difference between two rectangular sums can be computed in eight references. Since the two-rectangular features, like edge features, involve adjacent rectangular sums, they can be computed in six array references. In the same reason, the three-rectangular features defined above can be computed in eight array references.

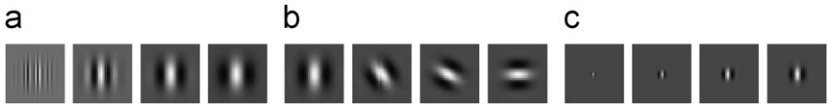

Fig. 3. Examples of Gabor functions. Each sub-figure shows the real part of Gabor function for different values of γ, θ, and σ: (a) $\gamma = \{1/2, 3/2, 5/2, 7/2\}$; (b) $\theta = \{0, \pi/6, \pi/3, \pi/2\}$; and (c) $\sigma = \{4, 8, 12, 16\}$.

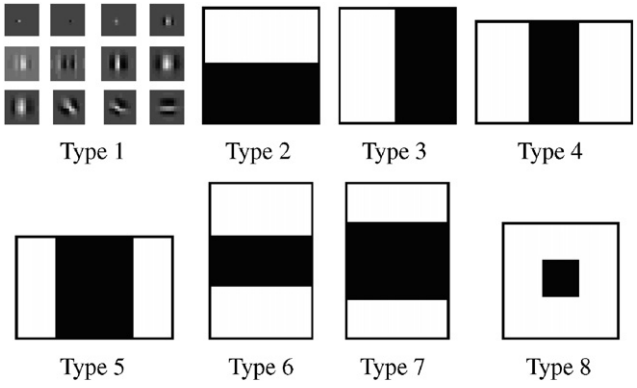

Fig. 4. Eight hybrid-feature types.

Harr-like features are sensitive to the presence of edges, bars, and other simple image structure. Different from Gabor filters, the available orientations are vertical and horizontal only.

(c) *Hybrid features*: For each feature $\mathbf{x}$, one weak classifier is configured. The proposed hybrid feature $\mathbf{x}$ includes both the Gabor feature and Harr-like feature. For Gabor features, besides position (x, y), there are three other parameters, σ, γ and θ, but for Harr-like feature, there are two other parameters, height (H) and width (W). The height and width parameter are {3, 6, 9, 12, 15, 18, 21, 24} and will round to the filter boundary, so eight Hs and Ws are used. First, with fixed image size (24×24), the scale parameter σ is a constant as $\sigma = 1.2$. Second, we select eight orientations and eight frequencies, so that the sets of values for γ and θ are $\Upsilon = \{0.8, 0.9, \ldots 1.5\}$, $\Theta = \{0, \pi/8, 2\pi/4, \ldots 7\pi/8\}$. So, there are four different parameters for the hybrid features, including position parameters x, y and filter parameters p_1 and p_2 (i.e., $p_1 = \gamma$, $p_2 = \theta$ for Gabor feature and $p_1 = H$ and $p_2 = W$ for Harr-like feature). Finally, we define the feature as $\mathbf{x} = (t, x, y, p_1, p_2)$, where $t = 1$ indicates Gabor feature and $t = 2$–8 specifies Harr-like feature as shown in Fig. 4.

3.2. Soft-decision function for weak classifiers

The output of conventional weak classifier is Boolean value, i.e., $h(\mathbf{x}) = \text{sign}[f(\mathbf{x}) - c]$, where $f(\mathbf{x})$ the response of the feature $\mathbf{x}$, and c is the threshold. However, the response distribution of the hybrid feature $f(\mathbf{x})$ is a complex distribution such as Multiple Gaussian model. Instead of selecting a simple threshold function or a hard decision function for $h(\mathbf{x})$, we define a soft decision function for each weak classifier for class ω_l or the hybrid feature $\mathbf{x}$, i.e., $h(\mathbf{x}, \omega_l)$. We create a pool of *2D soft-decision function for weak classifiers* before the hybrid-boost learning. To apply a piece-wise approximation to the continuous decision function, we define a function b which converts the response $f(\mathbf{x})$ to an index j, indicating the jth histogram bin of all the possible responses of the feature $\mathbf{x}$, i.e., $\{f(\mathbf{x})\}$, which are divided into n bins, $j = 1, \ldots, n$. The response value of $f(x)$ in each type has been normalized to [0, 1], so we define that $b(u) = j$ if $u \in [(j-1)/n, j/n)$. Here, we define the soft decision function of weak classifier for different classes as $h(\mathbf{x}, \omega_l)$ which is determined by the *posteriori density* $P(\omega_l|b(f(\mathbf{x})))$ where ω_l represents the lth class, and $l = 1, \ldots, k$. From Bayes' formula, we have

$$P(\omega_l|b(f(\mathbf{x}))) = \frac{P(b(f(\mathbf{x}))|\omega_l)P(\omega_l)}{P(b(f(\mathbf{x})))}, \tag{4}$$

where $P(b(f(\mathbf{x})))$ is the histogram of the response of feature $\mathbf{x}$ for all training data, $P(\omega_l)$ is the priori of class l, and $P(b(f(\mathbf{x}))|\omega_l)$ is the conditional probability. Before hybrid-boost learning, we create a *posteriori density* $P(\omega_l|b(f(\mathbf{x})))$ for each feature $\mathbf{x}$ as shown in Fig. 5. The response $f(\mathbf{x})$ is derived by adding a monotonic kernel function profile $K(r)$ which assigns a small weight to the response at a location further away from the center.

$$f(\mathbf{x}) = f(\mathbf{x})K(d/R) \quad \text{with } K(r) = 1 - (r)^2 \text{ for } r < 1, \quad \text{and } K(r) = 0 \text{ otherwise}, \tag{5}$$

where $r = (x^2 + y^2)^{1/2}$ and $R = (h^2 + w^2)^{1/2}$. We use h and w to represent the height and width of the test image block, and we select $h = w = 24$. The soft decision function for weak classifier is denoted as

$$\text{If} \quad \underset{i}{\text{Argmax}}\{P(\omega_i|b(f(\mathbf{x})))\} = l \quad \text{then} \quad h(\mathbf{x}, \omega_l) = 1, \quad \text{and } h(\mathbf{x}, \omega_j) = -1 \quad \text{for } j \neq l. \tag{6}$$

Here, we define six classes ($k = 6$) for human face detection (i.e., 90°, −45°, 0°, 45°, 90°, and non-face class), and set the number of bins $n = 20$. For each feature vector, there is a *posteriori* function $\text{P}(\omega_i|b(f(\mathbf{x})))$ generated in the training phase. The vector $\mathbf{x}$ is defined as $\mathbf{x} = (t, x, y, p_1, p_2)$, where t represents the type of features, i.e. $1 \leqslant t \leqslant 8$, (x, y) is the position of the feature with dimension 24×24, p_1 and p_2 are filter parameters quantized to eight different values. Since there are many possible features and *posteriori* functions (i.e., $8 \times 24 \times 24 \times 8 \times 8 = 294912$), the hybrid-boost learning algorithm selects only the most discriminant features from these hybrid features. It is important to choose suitable granularity (the number of bins) for a piece-wise function. The finer the granularity is, the more accurate the decision function can be obtained with lower estimation error, but more noise sensitivity.

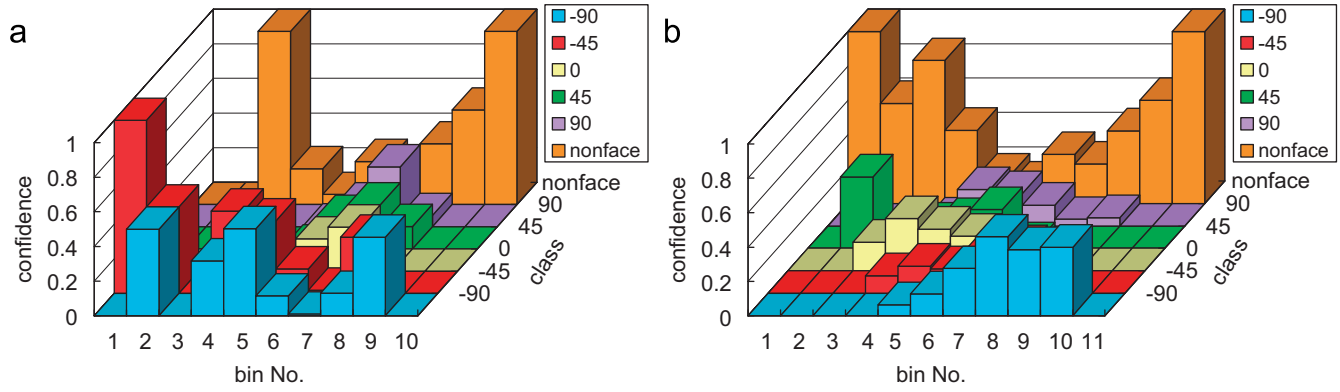

Fig. 5. The decision function $h(\mathbf{x}, l)$ for weak classifiers: (a) Gabor feature $\mathbf{x} = (t, x, y, p_1, p_2) = (1, 11, 6, 4, 3)$; and (b) Harr-like feature $\mathbf{x} = (t, x, y, p_1, p_2) = (2, 0, 12, 2, 1)$.

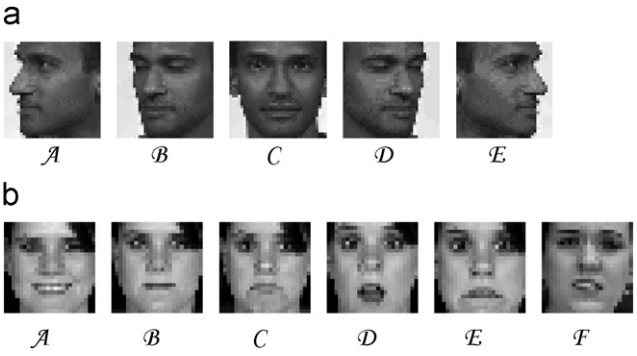

Fig. 6. (a) Five different posted faces; and (b) six different types of expressions.

4. Multi-pose face detection and expression recognition

Many face detection techniques can detect the frontal upright faces in wide variety of images. However, most of the methods can only deal with the front-pose faces. Here, we introduce a multi-pose face detection method to detect the faces in various poses. Instead of treating the face detection or expression recognition as a cascade binary classification problem, we propose a multi-class classification algorithm to solve the multi-pose face detection problem. This algorithm selects a small number of weak classifiers from a large weak classifier pool. However, when the feature dimensions are large, the training process will be very time-consuming.

As shown in Fig. 6, we illustrate the formulation for profile face detection and expression recognition, where the face data are categorized into five classes of different posed faces (i.e., pose angles: −90°, −45°, 0°, 45°, and 90°) and six classes of different expressions (i.e., happy, anger, sad, surprise, fear, and disgust).

4.1. The multi-class hybrid-boost learning algorithm

Different from the two-class Adaboost, we propose a hybrid-boost algorithm for multi-pose face detection which can detect the human face and determine its pose simultaneously. Here, we use the multi-class hybrid-boost learning algorithm to select the features from the hybrid feature set. Similar to Adaboost, hybrid-boost learning selects a small number of weak classifiers from a large weak classifier pool to form a stronger classifier. In each round of boosting, one feature is selected as a weak classifier. The multi-class hybrid-boost learning algorithm is shown as:

Define: χ is the sample space and y is the label set. A sample of a multi-class multi-label problem is a pair $(\mathbf{x}, Y)$, where $\mathbf{x} \in \chi$, $Y \subseteq y$. Weak hypothesis h_t: $\chi \times y \to \Re$ and $\alpha_t \in \Re$.
Label: For $Y \subseteq y$, define $Y[l]$ for $l \in y$ as $Y[l] = \begin{cases} 1 & \text{if } l \in Y, \\ -1 & \text{if } l \notin Y. \end{cases}$
Input: (1) n training samples: $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$, and (2) the number of iterations T.
Initialize: $D_1(i, l) = 1/(nk)$ where $i = 1, 2, \ldots, n$, and k indicates the number of classes

- For $t = 1, \ldots, T$
  - $r_t = \max_j \sum_{i,l} D_j(i, l) Y_i(l) h_j(\mathbf{x}_i, l)$
  - Let $\alpha_t = \frac{1}{2} \ln\left(\frac{1+r_t}{1-r_t}\right)$
  - Update the distribution:

$$D_{t+1}(i, l) = \frac{D_t(i, l) \exp(-\alpha_t Y_i[l] h_t(\mathbf{x}_i, l))}{Z_t}, \quad (7)$$

where Z_t is a normalization factor so that D_{t+1} is a $p.d.f$.

Under the distribution D_t, we will select a weak classifier h_t: $\chi \times y \to [-1, 1]$ from the pool of weak classifiers with maximum value of $\sum_{i,l} D_t(i, l) Y_i(l) h_t(\mathbf{x}_i, l)$. After T rounds of boosting, we will have T weak classifiers. At learning stage, for each weak classifier, we build up the distribution of the samples which represents the probability of misclassification. By updating the distribution, the weights α_t of misclassified samples will become larger. If a sample is misclassified for many times, its weight will become larger and larger. The larger the weight is, the harder the sample can be classified correctly. Here, T weak classifiers are constructed and the final strong classifier is a weighted linear combination of the T weak classifiers.

Fig. 7. Training samples: (a) human face with glasses; and (b) human face with moustache.

4.2. Different-posed face detection

In hybrid-boost learning procedure, we need to collect a large image database in different poses and manually categorize the images. For the multi-class hybrid-boost learning, we rescale all the training images to 24×24. In testing procedure, for a 320×240 input color image, we segment the potential face region, and then use a fast scanning method to detect human face in different sized regions which will be rescaled to 24×24 for classification. The training set consists of faces in different poses ($-90°$, $-45°$, $0°$, $45°$, and $90°$) selected from FERET image database and the non-face images. The face image database contains face with glasses and moustache as shown in Fig. 7.

FERET face database: The FERET image database [29] are selected which consists of 14051 eight-bit grayscale images of human faces with different poses ranging from frontal to left and right profiles. This database includes at least of 1000 persons with five or more pictures per person which correspond to different poses under several kinds of illuminations. In the experiments, 500 images per-pose have been selected for the learning process. However, the categories of pose in the database are not clearly differentiable, including quarter left and right ($\pm 22.5°$), half left and right ($\pm 67.5°$), profile left and right ($\pm 90°$), and some other non-regular poses with the angel $\pm 10°$, $\pm 15°$, $\pm 25°$, $\pm 45°$, $\pm 60°$. In our experiment, we categorize the images to each pose subjectively. The categorized images for each pose may not be accurate, the error tolerance is $\pm 15°$. Besides, we also choose 2500 non-face images (500 non-face images and 2000 misplaced face images) for hybrid-boost learning.

(b) *The strong classifier*: Here, we choose T weak classifiers after T rounds of boosting. A weak classifier is composed of a feature $\mathbf{x}$ and a weight α. We use the pre-selected weak classifiers $\{h_t(\mathbf{x}, i) \mid t = 1, \ldots, T\}$ and the weights α_t to construct i strong classifiers, $i = 1, \ldots, 6$. During multi-class face detection, we define the hypothesis H_l that the response of lth strong classifier is the strongest

$$H_l = 1 \quad \text{if } l = \underset{i}{\text{argmax}} \left(\sum_t \alpha_t h_t(\mathbf{x}, i) \right), \quad H_j = 0 \text{ for } j \neq l, \tag{8}$$

where $j = 1, \ldots, 6$. Next, we double check the strong classifier by calculating the confidence as

$$Conf_{H_l} = \left| \frac{\sum_t \alpha_t h_t(\mathbf{x}, l)}{\sum_t \alpha_t} \right|. \tag{9}$$

The $Conf_{Hl}$ determines whether hypothesis $H_l = 1$ is acceptable or not. If the confidence of the strong classifier is not higher than certain threshold Th then hypothesis H_l is no longer valid ($H_l = 0$). In our experiment, we have six strong classifiers for multi-pose face detection. However, the strong classifier with the highest response does not necessarily indicate the correct class. So, we define the following two decision rules for selecting the correct strong classifier. First, we define the outputs of the six rated strong classifiers as

$$V_{ai} = \sum_t \alpha_t h_t(\mathbf{x}, i), \tag{10}$$

where $i = 1, 2, \ldots, 6$, and a_i indicates the *ith* place strong classifier. The 1st place strong classifier must meet the following two rules to indicate an accurate detection.

$$\text{Rule 1}: \quad Conf_{H_{ai}} > Th_1 \quad \text{for } i = 1 \text{ and } 2,$$
$$\text{Rule 2}: \quad V_{a1} - V_{a2} > Th_2,$$

where V_{a1} and V_{a2} represent the output responses of the first place and the second place strong classifiers. Thus, these two rules must be all satisfied otherwise we will consider the test image belonging to the non-face class. Rule 1 requires that the confidence of output class is larger than a threshold Th_1 (similar to the two-class problem). If the difference between the two output responses is not large enough, then it is not distinguishable. These rules are used to increase the detection rate and lower the false alarm rate of the face detection.

As shown in Fig. 8(a), we find that the system without the above two constrains creates many candidate faces which are false alarms. Even with the two constraints, there are still some candidate faces as shown in Fig. 8(b) of which we want to choose the most accurate one. If these candidate faces are all close each other (i.e., the distance between every two candidate faces is less than certain threshold), we may select the one with the largest response as the correct one. We assume that ith candidate face is found at $\boldsymbol{s}_i = (x_i, y_i)$, $i = 1, \ldots, n$. The third constraint is defined as

$$\begin{aligned} \text{Rule 3:} \quad & \text{If } |x_i - x_j| + |y_i - y_j| < Th_3 \\ & \text{and} \quad l = \underset{\text{i}}{\text{Argmax}}\, V_{a_i} \text{ then } \mathrm{H}_{al} = 1 \\ & \text{and } \mathrm{H}_{ai} = 0 \text{ for } i \neq l, \end{aligned} \tag{11}$$

where $i = 1, \ldots, n$, $j = 1, \ldots, n$ and $i \neq j$. The thresholds Th_1, Th_2, and Th_3 are experimentally determined based on the number of iterations (T). With the above three rules, our face detection system may locate the face and identify its pose accurately as shown in Fig. 8(c). The blue box indicates $\pm 45°$ face detection, and the red box indicates the front face detections.

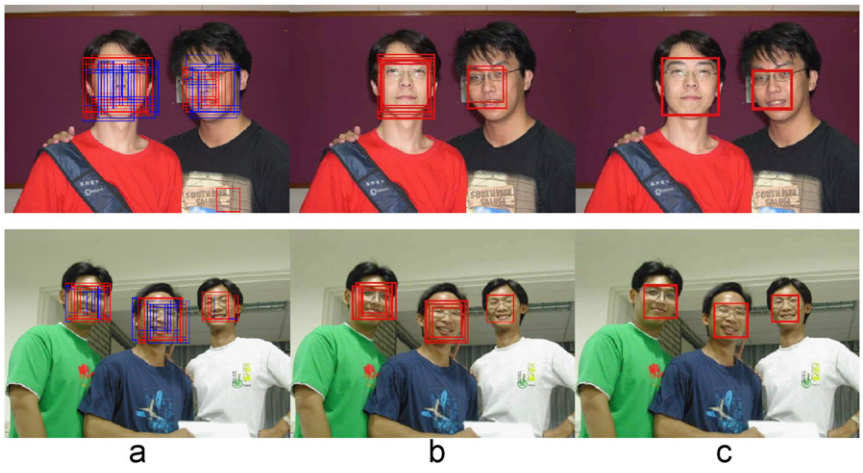

Fig. 8. (a) The detection results without applying the constraints; (b) the detection results with the two decision rules; and (c) The final results using three rules.

4.3. Facial expression recognition

Similar to multi-pose face detection, we may also apply the hybrid-boost learning for facial expression recognition and then compare the results with other facial expression system. The hybrid features are considered appropriate for facial expressions recognition which consists of the local features (Harr features) and the global features (Gabor features). Facial expressions are difficult due to many uncertain factors, such as the expression of disgust often looks like the expression of anger, or culture influence. Women are expected to be more emotionally extrovert while men are intended to be more guarded "poker face". We focus our attention on some portions of the faces because expressions are mostly localized to region near the eyes and the mouth. A smile is mostly shown by a person mouth, while anger is partly shown by a person eyes. Our approach can single out discriminative features both at global level and multiple local levels.

The training data of the facial expression classifiers come from Cohn and Kanade Facial Expression Database [25]. The database consists of 100 university students whose ages range from 18 to 30 years. Sixty-five percent are female, 15% are African–American, and 3% are Asian or Latino. Subjects began each display with a neutral face. Image sequences from neutral to target display are digitized into 640 × 480 pixel arrays with 8-bit precision for grayscale values. To reduce the effects of the distance variations, we need to normalize the training images. The input training images are normalized to a standard size (24 × 24 pixels) with seven different expressions (happy, anger, sad, surprise, fear, disgust and neutral) of the normalized images on Cohn and Kanade facial expression database.

5. Experimental results and discussions

Our system can also be implemented by using AMD 3000+ CPU and the image size is 320 × 240 pixels. In the first frame of a video sequence, we apply the face detector to search the entire image for the presence of different scale faces and the corresponding poses simultaneously. Once the face is detected, the face tracking is used to search and identify the face in the following image frames. The face tracking is also a face detection process with much smaller search area based on the detected face in the previous frame. It requires 300–350 ms for detecting a face and 50–60 ms for face tracking. In the experiments, we compare our method with the other methods to prove that our system is more robust.

5.1. Experimental results with training database

The human face detection is defined as a six-class categorization problem of which the five classes are the faces with five pose angles, i.e., −90°, −45°, 0°, 45°, 90°, and a non-face class. Here, we use FERET image database [29] for hybrid-boost training. All the face samples in FERET image database are normalized to 24 × 24 pixel without color information. Furthermore, to enhance the overall performance, we use 500 training images per class with several conditions including: (1) a slight rotation, (2) wearing glasses, (3) with different illuminations, (4) different races. Totally, we have 3000 face samples for the hybrid-boost learning.

Here, we develop a multi-class learning algorithm with six categories of poses ($k = 6$) and 100 weak classifiers ($T = 100$) of which 20 weak classifiers are Gabor features and the others are Harr-like features. These classifiers are applied to the FERET image database for testing. Because the testing data is part of the training set, it achieves a high correct rate up to 99.44%. Since, there are only 500 training images which are non-face images, the false alarm rates is much higher than what we expected. The experimental results are shown in Table 1. To decrease the false alarm rate, we add 2000 miss-classified images into the non-face class for another phase of learning. Then, we test a total of 5000 images, and find that the false alarm rate is reduced to 1.72%, and the average detection rate is 99.24%.

Then, we test our method by using "leave-one-group-out" cross validation. In cross-validation, the training set is randomly

Table 1
The experimental results

Input Class	−90°	−45°	0°	45°	90°	Non-face	P(%)	M(%)	FA (%)
−90°	498	1	1	0	0	0	99.6	0	N/A
−45°	1	497	1	0	0	1	99.4	0.2	N/A
0°	0	0	499	0	0	2	99.6	0.4	N/A
45°	1	0	0	496	1	2	99.2	0.4	N/A
90°	0	0	0	1	497	2	99.4	0.4	N/A
Non-face	22	16	12	9	18	423	84.6	N/A	15.4

P: detection rate; M: missing rate; FA: false alarm rate.

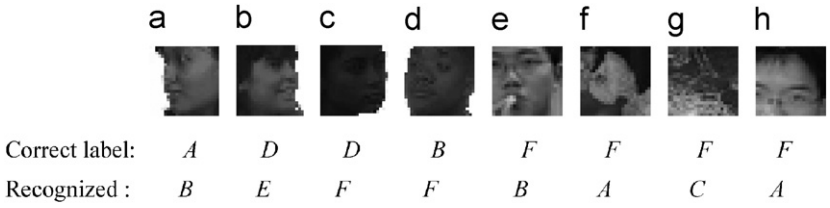

Fig. 9. The miss-classified examples.

Table 2
The experimental results based on "leave-one-group-out" cross-validation

Class	P (%)	M (%)	FA (%)
−90°	93.84	0.12	N/A
−45°	94.13	0.78	N/A
0°	93.64	0.63	N/A
45°	94.37	0.76	N/A
90°	93.95	0.62	N/A
Non-face	94.83	N/A	4.54

divided into m disjoint pattern sets (or groups) of equal size n/m, where n is the total number of training patterns. The classifier is trained and then tested m times, each time with a different set (n/m patterns) held out as a validation set (or testing set), and the rest $n - n/m$ patterns as the training set. The estimated performance is the average of these m tests. Here, we divide the whole training data, which consists of 1000 face images for each pose, and 5000 non-face images, into four groups ($m = 4$). This experiment consists of 4 different hybrid-learning processes and 4 testing processes. The performance is shown in Table 2 which is the average of the results of four different experiments. As shown in Table 2, the average detection rate is reduced to about 94%.

The false alarms are perhaps due to lower illumination or occlusion, and some ambiguous or undistinguishable posed face images may also confuse the learning algorithm. The boundary between two classes of different-posed face images cannot be very accurately defined. The labeled poses in Fig. 9(a) and (b) are −90° and 45°; but they are recognized as −45° and 90°. Fig. 9(c) and (d) are samples that too dark to be recognized. Fig. 9(e)–(h) show the false alarms.

5.2. Comparisons

Tables 3 and 4 show that the learning algorithm using Gabor features performs better detection rate than the others but it has the highest false alarm rate. To illustrate the advantages of using hybrid features, we show some results in Fig. 10. Here, the red square box shows the detected human face is class C (i.e., frontal face), and the blue box illustrates the detected human face is class B or class D (i.e., profiles in ±45°), and the white box illustrates the detected human face is class A or class E (i.e., profiles in ±90°). Some detected faces with large angle of rotation are categorized as class B or class D which are still treated as the correct detections. Fig. 10 shows that the hybrid-boost learning generates better classifiers with lower false alarm rate.

5.3. Testing on real-life photos

Here, we test our system on real-life photos simple/complex backgrounds as shown in Fig. 11. The input image size is 400 × 300 pixels. The executing time for each image depends on the skin color region detection and segmentation. Since the extracted skin color region is accurate, it takes less than 1 s to locate the correct face position. There are 87 test color images (includes 162 human faces of 100 different persons) of which only seven faces cannot be found and four faces are misplaced. Experimental results demonstrate an average detection rate 93.4% (151/162). Most of the errors are due to the color of faces is not right or the orientation of faces is too large. Since the face detection is applied only on the possible face regions rather than the entire image frame, the number of false alarms is reduced. The false alarm rate is lower than 0.1%.

Fig. 12 shows some cases of miss-identification or false alarm. For case A in Photo I and case E in Photo IV, there are

Table 3
Comparison with the single feature classifier on the same training database

Feature type	AP (%)	FA (%)	*M* (%)	MM (%)	Time/image (ms)
Gabor	99.48	4.96	0.28	0.24	6.8
Harr-like	99.16	1.72	0.36	0.48	2.33
Hybrid	99.24	1.72	0.36	0.4	3.0

AP: average correct detection rate; FA: false alarm rate; *M*: missing rate; MM: mismatch rate.

Fig. 10. The 1st row shows the face detection results using Gabor feature; the 2nd row shows results using Harr-like feature; and the 3rd row shows results using Hybrid feature.

Table 4
Testing results on 24 × 24 gray level image database independent of the training set

Feature type	AP (%)	FA (%)	*M* (%)	MM (%)
Gabor	94.1	8	4	1.9
Harr-like	93.15	3.4	3.85	3
Hybrid	93.65	3.2	3.5	2.85

human faces missed. It is because the face region is either too dark or too bright which cannot be defined as skin color. For case *B* in Photo II, the entire image is very much blurred, and the face region is not clear either. Therefore, the left person of the image cannot be properly detected. For case *C* in Photo III, the face of the person wearing sunglasses with long hair is not detected. For case *D* and case *E* in Photo IV, the potential face region after skin color segmentation of the undetected person is fragmented so that the face regions are not identified. The inaccurate skin color segmentation make the following face detector miss-identify the human faces. For case *F* in photo V, there is one miss and one false alarm occurring at the same time. For case *F* in Photo V and case *G* in Photo VI, there are also two false alarms (classified as −45° and 0°).

5.4. Testing on web-camera video

Here, we also do the real-time face detection experiments using the video captured from a webcam as shown in Fig. 13(a)–(e). It requires 300–350 ms to detect a face for a 320–240 color image and 50–60 ms for tracking. Fig. 13(a) shows the face detection in different poses. Although the boundary between the frontal face and the profile face is not

Fig. 11. Pose recognition results on real-life photos with simple/complex backgrounds.

Fig. 12. Some error examples in pose recognition results on real-life photos (I)–(VI).

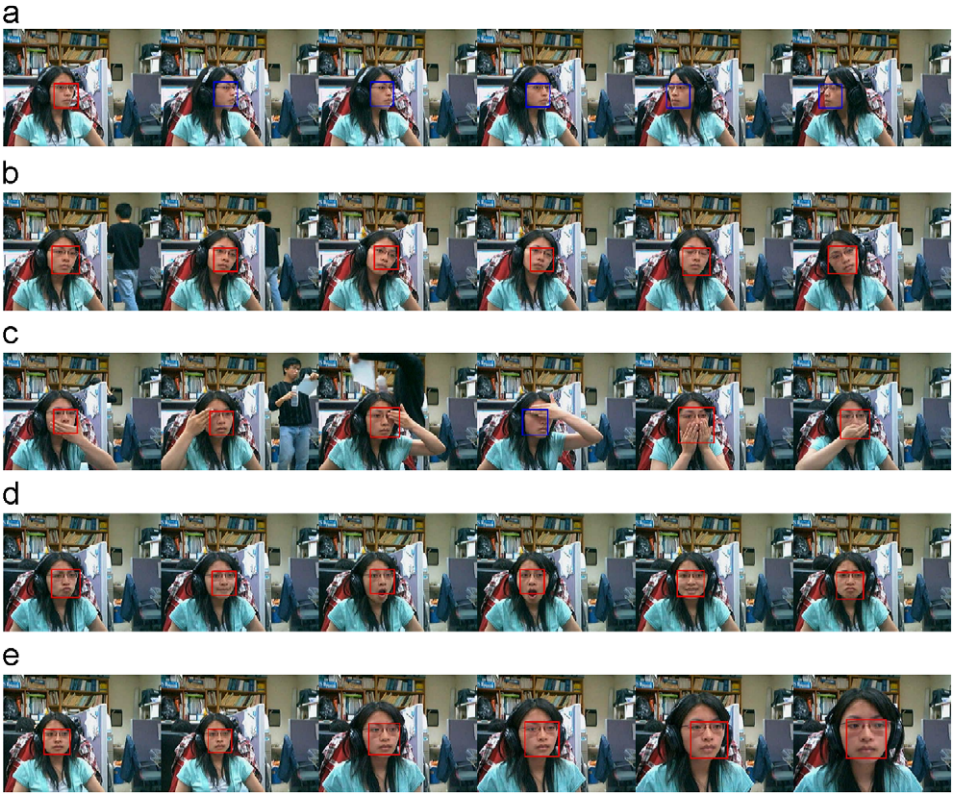

Fig. 13. Face detection of: (a) different-posed faces; (b) slightly rotated faces; (c) partially occluded faces; (d) various expressions in the frontal and 10°-posed faces; and (e) different-scaled faces.

obvious, we may say that the detection is almost correct. Fig. 13(b) shows that the detected faces are slightly rotated. In Fig. 13(c), we can detect the faces partial occluded by hands. The system is robust for the mouth-occluded cases, but not unstable for the eye-occluded cases. Fig. 13(d) shows the real-time face detection with various expressions in the 10°-posed

Table 5
Comparison:

(a) system configuration	Detection rate %	CPU	Frame rate (Hz)	Feature		
				Type	Number	Resolution
Our detector	94.7	AMD 3000+(÷2.0 GHz)	10–20	Hybrid feature	100/294 912	24*24 pixels
OpenCV detector	90 ↑	Pentium IV 2.8 GHz	15–25	Harr-like feature		24*24 pixels
MPT detector	90	Pentium IV 3.0 GHz	24	Gabor feature	900/165 888	48*48 pixels

(b) performance	Acceptable rotate angle (Deg)			Occlusion		Expression
	ROP (Deg)	RIP (Deg)	Up–Down	Mouth	Eyes	
Our detector	±90	±20	±20	○	△	○
OpenCV detector	±45	±20	±20	○	△	○
MPT detector	±25	±15	±15	×	×	△

○: accurate detection, △: not accurate detection; ×: detection fails.

Table 6
The results with $T = 200$

Input Classified	Happy	Anger	Sad	Surprise	Fear	Disgust	neutral	$P(\%)$	$M(\%)$
Happy	423	1	0	0	6	1	4	97.2	0.9
Anger	0	205	11	0	0	5	7	89.9	3.1
Sad	1	18	303	0	5	1	10	89.6	2.9
Surprise	0	0	0	347	0	0	2	99.4	0.6
Fear	6	2	1	0	202	0	0	95.7	0
Disgust	4	1	0	0	0	118	3	89.6	2.4

P: correct classification rate; and M: missing rate.

face and the frontal face. Fig. 13(e) shows the detection results of the different-sized of faces.

5.5. Comparisons with other real-time detector

Here, we compare our method with the other real-time face detectors, such as OpenCV [31] and MPT [32]. The face detection in OpenCV based on [3,30] uses simple Harr-like features and a cascade of boosted tree classifiers. MPT face detector extracts a square-bounding box around each face in the processed image. Face detection is applied to each image of the sequence without any face tracking. The main drawback is the processing time even though MPT works nearly in real time for pictures of size (320 × 200 pixels). The comparisons are shown in Table 5, where "ROP" is "rotation out of plane" and "RIP" is "rotation in plane".

In Table 5(a), we can find that every system has its own advantages. The computation time of our detector and OpenCV's detector varies. In our system, the processing time for each frame depends on the potential face region and the sizes of the sub-window. Similarly, the computation of the OpenCV face detection system depends on the stage number and the sizes of the sub-window. In Table 5(b), we test the three detection algorithms in different conditions, including the face rotation, occlusion and the face with different expressions. Our test video includes mouth or eyes occlusion cases. We find that the eye occlusion makes the most influence on the detection results. Table 5(b) shows that our system has the best performance. Once the possible face regions are found, the processing time of the face detection process is much reduced. The reduced computation time is nearly proportional to the ratio of the detected skin color area to entire image frame. The computation time of skin color segmentation is much less than the processing time of the face detector, therefore it can be neglected.

5.6. Facial expression recognition

Similarly, we apply the hybrid-boost learning on the Facial Expression Database [25] and select the classifiers for the facial expression recognition. We manually selected 1687 images from the data set and labeled as one of the six basic emotions (happy, anger, sad, surprise, disgust, and fear) and 460 images for neutral. The system requires 7 ms to process a 24 × 24 images. It has 93.1% average correction rate when the testing data is same as the training data as shown in Table 6. We have also compared our experimental results with the other methods [15,16] using the same training database as shown in Table 7. The average correct recognition rate (AR) of our is higher than the [16] and the processing rate is lower than the other method [15].

Finally, we show some facial expression identification results of image with simple/complex backgrounds by using the

Table 7
Comparison with the same training database

Methods	Feature type	Method		Time/image	AR(%)
		Feature selection	Classification		
Ours	Hybrid feature	Adaboost	Adaboost	6 ms	93.1
[15]	Gabor feature	Adaboost	SVM	< 10 ms	93.3
[16]	Facial model	AUs	RVM		90.84

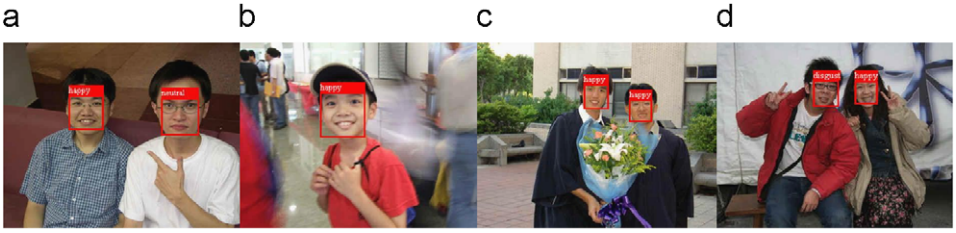

Fig. 14. Facial expression recognition for real-life photos with simple/complex backgrounds: (a) happy and mutual; (b) happy; (c) happy; and (d) disgust and happy.

hybrid-boost learning classifiers. As shown in Fig. 14(a)–(c), we have correctly identified the facial expressions, whereas in 14(d), we have misclassified a facial expression of disgust.

6. Conclusions and feature works

We have introduced a multi-posed face detection and expression identification system which is more robust than the other proposed face detection system and facial expression system. Our system is based on hybrid-boost multi-class learning algorithm as well as three decision rules which generates higher detection rate and lower false alarm rate. The experimental results show that the system has better performance than the others using Harr-like feature or Gabor feature.

References

[1] M.-H. Yang, D.J. Kriegman, N. Ahuja, Detecting faces in images: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 24 (1) (2002) 34–58.
[2] E. Hjelmas, B.K. Low, Face detection: a survey, Comput. Vision Image Understanding 83 (2) (2001) 236–274.
[3] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, IEEE CVPR, 2001, pp. 511–518.
[4] M. Jones, P. Viola, Fast multi-view face detection, Technical Report. TR2003-96, Mitsubishi Electric Research Laboratories, July 2003.
[5] P. Moreno, A. Bernardino, J. Santos-Victor, Gabor parameter selection for local feature detection, IbPRIA 2005, Lecture Notes in Computer Science 3522 (2005) 11–19.
[6] G. Guo, H.J. Zhang, Boosting for fast face recognition, IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Face and Gesture (RATFG-RTS'01), 2001.
[7] C. Huang, H. Ai, Y. Li, S. Lao, Vector boosting for rotation invariant multi-view face detection, Proceedings of the 10th IEEE ICCV, 2005.
[8] S.Z. Li, Z.Q. Zhang, Floatboost learning and statistical face detection, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004) 1112–1123.
[9] R. Xiao, M.-J. Li, H.-J. Zhang, Robust multi-pose face detection in images, IEEE Trans. Circuits Systems Video Technol. 31–41, 2004.
[10] R. Verschae, J. Ruiz del Solar, A hybrid face detector based on an asymmetrical adaboost cascade detector and a Wavelet-Bayesian-detector, Lecture Notes in Computer Science, vol. 2686, Springer, Berlin, 2003, pp. 742–749.
[11] Y.Y. Lin, T.L. Liu, Robust face detection with multi-class boosting, Proceedings of IEEE CVPR, 2005.
[12] X.Xu, T. Huang, Face recognition with MRC-boosting, Proceedings of IEEE, ICCV, 2005.
[13] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognition 36 (2003) 259–275.
[14] M. Pantic, L.L.M. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1424–1455.
[15] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Recognizing facial expression—machine learning and application to spontaneous behavior, CVPR, 2005, pp. 568–573.
[16] D. Datcu, L.J.M. Rothkrantz, Facial expression recognition with relevance vector machines, ICME, 2005.
[17] M.E. Tipping, Sparse Bayesian learning and the relevance vector machine, Mach. Learn. (2001), pp. 211–244.
[18] W.M. Yu, T. Du, K.B. Lim, Comparison of the support vector machine and relevant vector machine in regression and classification problems, Control, Automation, Robotics, and Vision Conference (ICARCV), December 2004.
[19] G. Zhao, M. Pietikäinen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29(6) (2007) 915–928.
[20] H. Gu, Q. Ji, Information extraction from image sequences of real-world facial expressions, Mach. Vision Appl. 2005, pp. 105–115.
[21] Y. Wang, H. Ai, B. Wu, C. Huang, Real time facial expression recognition with adaboost, Proceedings of IEEE of ICPR, Cambridge, 2004.
[22] S. U. Jung, D.H. Kim, K. H. An, M. J. Chung, Efficient rectangle feature extraction for real-time facial expression recognition based on adaboost, IEEE/RSJ International Conference on Intelligent Robotics and Systems (IROS), 2005.
[23] P. Silapachote, D. R. Karuppiah, A. R. Hanson, Feature selection using adaboost for face expression recognition, Proceedings of the 4th IASTED International Conference on Visualization, Image, and Image Processing, Spain, September 2004.
[24] P.S. Aleksic, A. Katsaggelos, Automatic facial expression recognition using facial animation parameters and multistream HMMs, IEEE Trans Inf. Forensics Secu. 1 (1) (2006).

[25] T. Kanade, J.F. Cohn, Y. Tian, Comprehensive database for facial expression analysis, Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (AFGR'00), Grenoble, France, 2000, pp. 46–53.
[26] D. Chai, K.N. Ngan, Face segmentation using skin-color may in video application, IEEE Trans. CAS VT 9 (4) (1999) 551–564.
[27] D. Chai, S. L. Phung, A. Bouzerdoum, Skin color detection for face localization in human-machine communications, Sixth International Symposium on Signal Processing and its Applications, vol. 1, August 2001, pp. 343–346.
[28] R.C. Gonzalez, R.E. Woods, Digital Image Processing, Addison-Wesley Publishing Company, Reading, MA, 1992.
[29] P.J. Phillips, H.J. Moon, S.A. Rizvi, P.J. Rauss, The FERET evaluation methodology for face recognition algorithms, Pattern Anal. Mach. Intell. 22 (7) (2000) 1090–1104.
[30] R. Lienhart, J. Maydt, An extended set of Harr-like features for rapid object detection, ICIP2002, 2002.
[31] Intel. Open Source Computer Vision Library, 2000.
[32] Machine Perception Toolbox (MPT) ⟨http://mplab.ucsd.edu/grants/project1/free-software/MPTWebSite/API/⟩.

About the author—HSIAO-YING CHEN received her B.S., degree in electrical and control engineering from the National Chaio-TungUniversity, Hsinchu, Taiwan, and MS degree in electrical engineering from the National Tsing-Hua University, Hsinchu, Taiwan, in 2004, and 2006, respectively. Currently, she is with Media Technology, Inc., Hsinchu. Her research interests are pattern recognition and computer vision.

About the author—CHUNG-LIN, Dr. Huang received his B.S. degree in Nuclear Engineering from the National Tsing-Hua University, Hsin-Chu, Taiwan, ROC, in 1977, and M.S. degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan, ROC, in 1979, respectively. He obtained his Ph.D. degree in Electrical Engineering from the University of Florida, Gainesville, FL, USA, in 1987. From 1987 to 1988, he worked for the Unisys Co., Orange County, CA, USA, as a Project Engineer. Since August 1988 he has been with the Electrical Engineering Department, National Tsing-Hua University, Hsin-Chu, Taiwan, ROC. Currently, he is a Professor in the same department. In 1993 and 1994, he had received the Distinguish Research Awards from the National Science Council, Taiwan, ROC. In November 1993, he received the best paper award from the ACCV, Osaka, Japan, and in August 1996, he received the best paper award from the CVGIP Society, Taiwan, ROC. In December 1997, he received the best paper award from IEEE ISMIP Conference held Academia Sinica, Taipei. In 2002, he received the best paper annual award from the Journal of Information Science and Engineering, Academia Sinica, Taiwan. His research interests are in the area of image processing, computer vision, and visual communication. Dr. Huang is a senior member of IEEE.

About the author—CHIH-MING FU received his B.S., M.S., and Ph.D. degrees in electrical engineering from the National Tsing-Hua University, Hsinchu, Taiwan, ROC., in 1999, 2001, and 2006, respectively. Currently, he is with Cheertek Technology, Inc., Hsinchu. His research interests are wavelet analysis, signal processing, image/video processing, pattern recognition, and multimedia communication. In August 2006, he received the best paper award from the CVGIP Society, Taiwan.