Robust Multipose Face Detection in Images

Rong Xiao, Ming-Jing Li, and Hong-Jiang Zhang

***Abstract*—Automatic human face detection from images in surveillance and biometric applications is a challenging task due to the variances in image background, view, illumination, articulation, and facial expression. In this paper, we propose a novel three-step face detection approach to addressing this problem. The approach adopts a simple-to-complex strategy. First, a linear-filtering algorithm is applied to enhance detection performance by remove most nonface-like candidate rapidly. Second, a boosting chain algorithm is adopted to combine the boosting classifiers into a hierarchy "chain" structure. By utilizing the inter-layer discriminative information, this algorithm reveals higher efficiency than traditional approaches. Last, a postfiltering algorithm consists of image preprocessing; support vector machine-filter and color-filter are applied to refine the final prediction. As only small amount of candidate windows remain in the final stage, this algorithm greatly improves the detection accuracy with small computation cost. Compared with conventional approaches, this three-step approach is shown to be more effective and capable of handling more pose variations. Moreover, together with a two-level hierarchy in-plane pose estimator, a rapid multiview face detector is therefore built. The experimental results demonstrate the significant performance improvement using the proposed approach over others.**

***Index Terms*—AdaBoost, face detection, support vector machine (SVM).**

I. Introduction

FACE detection has been regarded as a challenging problem in the field of computer vision, due to the large intra-class variations caused by the changes in facial appearance, lighting, and expression. Such variations result in the face distribution to be highly nonlinear and complex in any space which is linear to the original image space [11]. Moreover, in the applications of real-life surveillance and biometric, the camera limitations and pose variations make the distribution of human faces in feature space more dispersed and complicated than that of frontal faces, which further complicates the problem of robust face detection.

Frontal face detection has been studied for decades. Sung and Poggio [16] built a classifier based on the difference feature vector which was computed between the local image pattern and the distribution-based model. Papageorgiou [2] developed a detection technique based on an over-complete wavelet representation of an object class. He first performed a dimensionality reduction to select the most important basis function and then trained a support vector machine (SVM) [18] to generate final prediction. Roth [3] used a network of linear units—SNoW learning architecture which is specifically tailored for learning in the presence of a very large number of features. Viola and Jones [12] developed a fast frontal face detection system. In their work, a cascade of boosting classifiers which were built on an over-complete set of Haar-like features integrated the feature selection and classifier design in the same framework.

Most nonfrontal face detector in the literature are based on the view-based method [1], in which several face models are built, each describes faces in a given range of view. Therefore, explicit three-dimensional (3-D) modeling is avoided. The work in [7] partitioned the views of face into five channels, and developed a multiview detector by training separate detector networks for each view. The authors of [9] studied the trajectories of faces in linear PCA feature spaces as they rotate and used SVMs for multiview face detection and pose estimation. The work in [6] used multiresolution information in different levels of wavelet transform. The system consists of an array of two face detectors in a view-based framework. Each detector is constructed using statistics of products of histograms computed from examples of the respective view. It has achieved the best detection accuracy in the literature, while it is very slow due to the computation complexity.

To address the problem of slow detection speed, Li *et al.* [15] proposed a coarse-to-fine, simple-to-complex pyramid structure, by combining the idea of boosting cascade and view-based methods. Although this approach improves the detection speed significantly, it is still stumped by the following problems. First of all, as the system computation cost is determined by the complexity and false positive (FP) rates of classifiers in the earlier stage, the inefficiency of AdaBoost significantly degrades the overall performance. Second, as each boosting classifier works separately, the useful information between adjacent layers are discarded, which hampers the convergence of the training procedure. Third, during the training process, more and more nonface samples collected by bootstrap procedures are introduced into the training set; thus it gradually increases the complexity of the classification. In the last stage pattern distribution between face and nonface become so complicated that can hardly be distinguished by Haar-like features. Finally, view-based method always suffers from the problems of high computation complexity and low detection precision.

In this paper, a novel approach to rapid face detection is presented. It uses a three-step algorithm based on a simple-to-complex strategy, and each step has different focus. In the first step, the classifier should be "simpler," rejecting negative samples with little computation over two to three features. As few features used in this step, training extensive algorithms with global optimization characteristics are affordable to obtain a high performance prefilter. In the second step, the classifier should be "efficient," reducing FP rate to the scale of 10^{-7} with as small computation cost as possible. As most prediction will be done

Manuscript received November 17, 2002; revised May 9, 2003.

The authors are with Microsoft Research Asia, Beijing 100080, China (e-mail: i-rxiao@microsoft.com; mjli@microsoft.com; hjzhang@microsoft.com).

Digital Object Identifier 10.1109/TCSVT.2003.818351

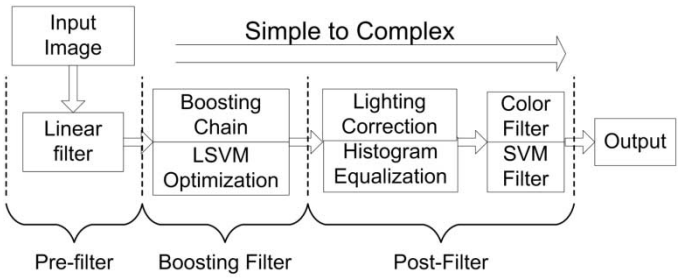

Fig. 1. Three-step face detector.

in this step, computation cost is critical to the overall detection speed. Therefore, a boosting chain filter with better convergence rate is proposed to substitute boosting cascade. In the last step, the classifier should be "accurate," removing FPs precisely. As most FPs in the candidate list are discarded in this step, a set of computation extensive algorithm could be applied without much computation load.

To enable the application of face detection in real-life surveillance and biometric applications, a multiview face detection system is designed based on the proposed approach. This system is able to handle pose variance in the range of $[-45°, 45°]$ both out-of-plane and in-plane rotation, respectively. In this system, first a two-level hierarchy in-plane pose estimator based on Haar-like features is built to alleviate the variance of in-plane rotation by dividing the input window into three channels. Second, an upright face detector based on the three-step algorithm is built, which enables the rapid multiview face detection in a single classifier.

The rest of the paper is organized as follows. Section II presented in detail the proposed three-step face detector framework. The multiview face detection system is presented in Section III. Section IV provides the experimental results, and conclusions are drawn in Section V.

II. Three-Step Face Detector

The differentiation of the proposed face detection approach from previous ones is its ability to detect faces rapidly with very low false alarm rates. The system architecture, shown in Fig. 1, consists of following components. First, a linear prefilter is used to increase the detection speed. Second, a boosting chain, developed from Viola's boosting cascade [12], is applied to remove most nonfaces from the candidates. After this procedure, the remaining candidate windows will typically be less than 0.001% in all scale. Finally, a color filter and an SVM filter are used to further reduce false alarms. Each of these components is described in detail in this section.

A. Basic Concepts of Detection With Boosting Cascade

In order to implement the rapid detector, the feature based algorithm is adopted in the prefilter and the boosting filter. Before continuing on the detail description, a few basic concepts are introduced here.

Haar-like features: Four types of Haar-like features, which are shown in Fig. 2 [12]. These features are computed by mean value difference between pixels in the black rectangles and pixels in the gray rectangles. Both are sensitive to horizontal and vertical variations, which are critical to capture upright frontal face appearance.

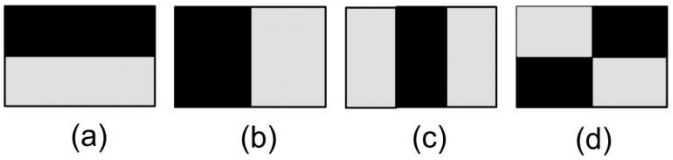

Fig. 2. Basic Haar-like features.

Weak Learner: A simple decision stump $h_t(x)$ is built on the histogram of the Haar-like feature f_t on the training set, where $h_t(x) = sign(p_t f_t(x) - \theta_t)$, θ_t is the threshold for the decision stump, and p_t is the parity to indicate the direction of decision stump.

Integral Image: To accelerate the computation of Haar-like features, an intermediate representation of the input image is defined in [12]. The value of each point (s, t) in an integral image is definesd as

$$ii(s,t) = \sum_{s' \le s, t' \le t} i(s', t') \tag{1}$$

where $i(s', t')$ is a grayscale value of the original image. Based on this definition, the sum of the pixels within rectangle in the original image could be computed within three sum operations.

Boosting Cascade: By combining boosting classifiers in a cascade structure, detector is able to rapidly discard most nonface like windows. Windows not rejected by the initial classifier are processed by a sequence of classifiers, each slightly more complex than the last. On a 640×480 images, containing more than one million face candidate windows in the image pyramid, with this structure, face are detected using an average of 270 microprocessor instructions per windows, which results in a rapid detection system

B. Linear Prefilter

Adaboost, developed by Freund and Schapire [19], has been proved to be a powerful learning method for face detection problem. Given $(x_1, y_1), \ldots, (x_n, y_n)$ as the training set, where $y_i \in \{-1, +1\}$ is the class label associated with example x_i, the decision function used by Viola [12] is

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x) + b\right). \tag{2}$$

In (2), α_t is a coefficient, b is a threshold, and $h_t(x)$ is a one-dimensional (1-D) weak learner defined in Section II-A.

In the case of $T = 2$, the decision boundary of (2) could be displayed in the two-dimensional (2-D) space, as shown in Fig. 3(a) and (b) . As only the sign information of $h_t(x)$ is used in (2), the discriminablity of the final decision function is greatly affected.

To address this problem, the decision function is rewritten in the following format:

$$H(x) = (a_1 f_1(x) > b_1) \wedge (a_2(f_1(x) + r f_2(x)) > b_2) \tag{3}$$

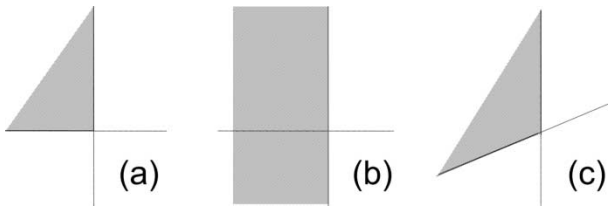

Fig. 3. Two-feature boosting classifier VS linear prefilter. (a), (b) Boundaries of boosting. (c) Decision boundary of the linear filter.

where α_i, b_i, and $r \in (-1, 1)$ are the coefficients which could be determined during the learning procedure. The final decision boundary is shown in Fig. 3(c).

The first term in (3) is a simple decision stump function, which can be learned by adjusting threshold according to the face/nonface histograms of this feature. The parameters in the second term could be acquired by linear SVM. The target recall could be achieved by adjusting bias terms b_i in both terms.

C. Boosting Filter

The boosting cascade proposed by Viola has proved to be an effective way to detect faces with high speed. During the training procedure, windows which are falsely detected as faces by the initial classifier are processed by successive classifiers. This structure dramatically increases the speed of the detector by focusing attention on promising regions of the image.

However, there are still two issues that require further investigation. One is how to utilize the historical knowledge in the previous layer; and the other is how to improve the efficiency of threshold adjusting. We propose a *boosting chain* with linear SVM optimization to address both issues.

1) Boosting Chain: In each layer of the boosting cascade, the classifier is adjusted to a very high recall ratio to preserve the overall recall ratio. For example, for a 20-layer cascade, to anticipate overall detection rates at 96% in the training set, the recall rate in each single will be 99.8% ($\sqrt[20]{0.96} = 0.998$) on the average. However, such a high recall rate at each layer is achieved with the penalty of decreasing sharp precision. As shown in Fig. 4, value b is computed for the best precision, and value a is the best threshold which satisfies the minimal recall rate requirement. During the threshold adjustment from value b to value a, the classifier's discriminability in the range $[a, +\infty]$ is lost. As the performance of most weak learners used in the boosting algorithm is near to random guess, such discriminative losing between the layers of boost cascade is critical to increase the converge speed of successive classifiers.

To address this issue, a chain structure of boosting cascade is proposed (as shown in Fig. 5). The algorithm is designed as follows:

The algorithm is designed as shown in Fig. 6, where the boosting chain is trained in a serial of boosting classifiers, and each classifier corresponds to a node of the chain structure. Different from the boosting cascade algorithm, the positive sample weights are directly introduced into the substantial learning procedure. For negative samples, collected by the bootstrap method, their weights are adjusted according to the classification errors of each previous weak classifier. Similar to

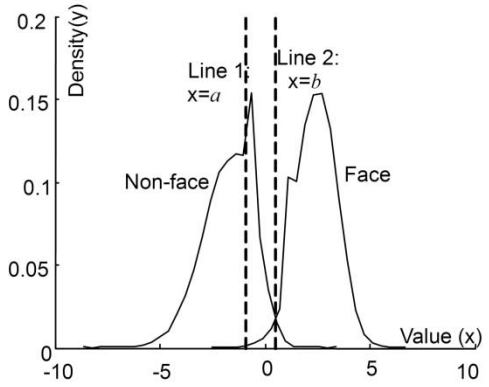

Fig. 4. Adjusting threshold for layer classifier.

the equation used in the boosting training procedure [13], the adjusting could be achieved by

$$w_j \leftarrow c \exp\left[-y_j \sum_{t=1}^{i} \Phi_t(\mathbf{x}_j)\right] \tag{4}$$

where y_j is the label of sample x_j, c is the initial weight for negative samples, and i is the current node index.

Meanwhile, results from previous node classifier are not discarded while training the subsequential new classifier. Instead, the previous classifier is regarded as the first weak learner of the current boosting classifier. Therefore, these boosting classifiers are linked into a "chain" structure with multiple exits for negative patterns. The evaluation algorithm of boosting chain is shown in Fig. 7.

2) Linear Optimization: In each step of the boosting chain, performance at the current stage involves a tradeoff between accuracy and speed. The more features used, the higher detection accuracy achieved. At the same time, classifiers with more features require more time to evaluate. The naïve optimization method used by Viola is to simply adjust threshold for each classifier to achieve a balance between the targeted recall and FP rates. However, as mentioned before, this method frequently results in a sharp increase in false rates. To address this issue, a new algorithm based on linear SVM for postoptimization is proposed.

Alternatively, the final decision function of AdaBoost in (2) could be regarded as the linear combination of weak learners $\{h_1(x), h_2(x), \ldots, h_T(x)\}$.

Each weak learner $h_t(x)$ will be determined after the boosting training. When it is fixed, the weak learner maps the sample x_i from the original feature space F to a point

$$x_i^* = h(x_i) = \{h_1(x_i), h_2(x_i), \ldots, h_T(x_i)\} \tag{5}$$

in a new space F^* with new dimensionality T. Consequently, the optimization of α_t parameter can be regarded as finding an optimal separating hyper-plane in the new space F^*. The optimization is obtained by the linear SVM algorithm to resolve the following quadratic programming problem:

$$\text{Maximize: } L(\beta) = \sum_{i=1}^{n} \beta_i - \frac{1}{2} \sum_{i,j=1}^{n} \beta_i \beta_j y_i y_j (h(x_i) \cdot h(x_j)) \tag{6}$$

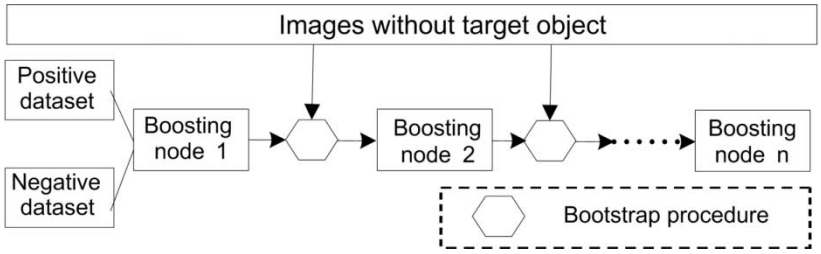

Fig. 5. Boosting chain structure.

Assume: P positive training set, $p=|P|$
N_i ith negative training set, $n_i=|N_i|$
f_i maximum false positive rate of ith layer
d_i minimum detection rate of ith layer
w_j weighting of sample x_j
F overall false positive rate.
Φ_i ith boosting classifier in the cascade

1. Initialize: $i=0$, $F_0=1$, $\Phi=\{\}$
   $w_j=1/p$ for all positive sample x_j, $w_j=1/n_i$ for all negative sample x_j;
2. While $F_i>F$
   a) $i=i+1$
   b) Taining Φ_i to meet the f_i and d_i requirements on validation set.
      i. Using initial weights w_j, training set P and N_i
      ii. Train a node classifier Φ_i
   c) Node classifier optimization (in Section II C.2)
   d) $F_i=F_{i-1}*f_i$, $\Phi=\Phi\cup\{\Phi_i\}$
   e) Evaluate *boosting* chain Φ on non-face image set, and put false detections into the set N_{i+1}
   f) For each *sample* x_j in set N_{i+1}, update weight w_j for Φ_{i+1} according to Equation (4).

Fig. 6. The training algorithm for building a boosting chain filter.

1. Given an example x, evaluate the boosting chain with M node
2. Initialize $s = 0$
3. Repeat for $i = 1$ to M:
   a) $s = s + \sum_{t=1}^{m_i} \alpha_{i,t} h_{i,t}(x)$
   b) if ($s < b_i$) then exit with negative response.
4. Exit with positive response.

Fig. 7. Evaluate the boosting chain.

subject to the constraints $\sum_i^n \beta_i y_i = 0$ and $C_i \geq \beta_i \geq 0$, $i = 1, \ldots, n$. Coefficient C_i is set according to the classification risk w and tradeoff constant C over the training set

$$C_i = \begin{cases} wC, & \text{if } x_i \text{ is a face pattern} \\ C, & \text{otherwise.} \end{cases} \tag{7}$$

The solution of this maximization problem is denoted by $\beta^0 = (\beta_1^0, \beta_2^0 \ldots, \beta_n^0)$. Then the optimized α_t will be given by $\alpha_t = \sum_{i=1}^n \beta_i y_i h_t(x_i)$.

By adjusting the bias term b and classification risk w, the optimized result is found. Experimental results in Fig. 8 illustrated the efficiency of this algorithm.

D. Postfilter

Due to the variations of image patterns and the limitations of Haar-like features, there still remain many false alarms after processing. In this step, a set of image preprocessing methods are first applied to the candidate windows to reduce pattern variations, and then two filters based on color information and wavelet features are applied to further reduce false alarms.

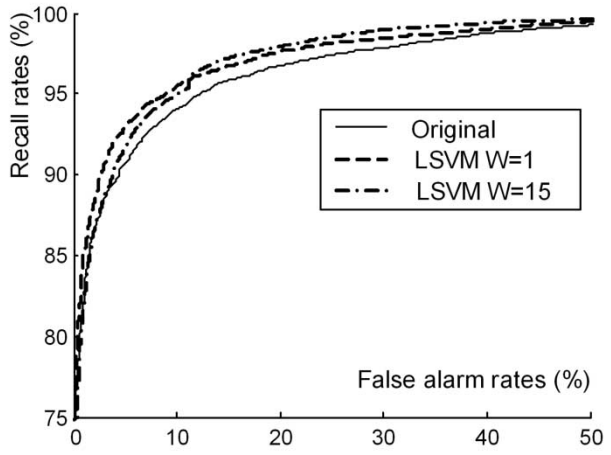

Fig. 8. ROC curves comparing the original Boosting chain algorithm with the LSVM optimization algorithm with different weights.

1) Image Preprocessing: The processing procedure aims to alleviate background, lighting, and contract variations. It consists of three steps [7]. First of all, a mask, which is generated by cropped out the four edge corner from the window, is applied to the candidate region. Then a linear function is selected to estimate the intensity distribution on the current window. By subtracting the plane generated by this linear function, the lighting variations could be significantly reduced. Finally, histogram equalization is performed. With this nonlinearly mapping, the range of pixel intensities is enlarged and thus somewhat improves the contrast variance which caused by camera input difference.

2) Color-Filter: Modeling skintone color has been studied extensively in recent years [14]. In our system, YC_bC_r space is adopted due to its perceptually uniform. As the Y component mainly represents image grayscale information which is quite irrelevant to skintone color, only C_b and C_r components are reserved for false alarm removal.

As shown in Fig. 9(a), the color of face and nonface images is distributed as nearly Gaussian in C_bC_r space. A two-degree polynomial function will be an effective decision function for this problem. For any point (c_b, c_r) in the C_bC_r space, the decision function can be written as

$$F(c_r, c_b) = \text{sign}\left(a_1 c_r^2 + a_2 c_r c_b + a_3 c_b^2 + a_4 c_r + a_5 c_b + a_6\right) \tag{8}$$

which is a linear function in the feature space with dimension $(c_r^2, c_r c_b, c_b^2, c_r, c_b)$. Consequently, a linear SVM classifier is

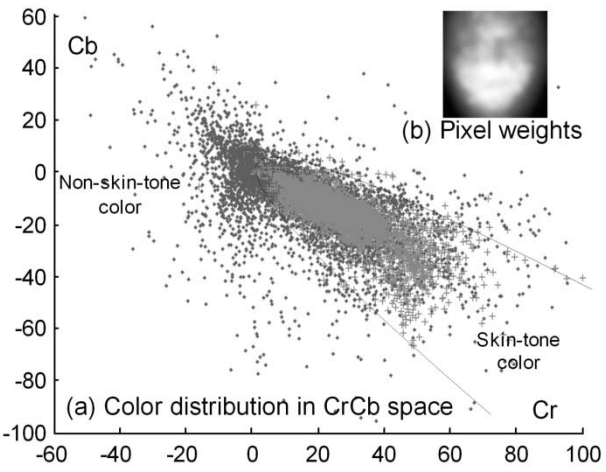

Fig. 9. Two-degree polynomial color filter in C_rC_b space. The pixel weights are shown at the top right. The darker the pixel is, the less important it will be.

constructed in this five-dimensional space to separate skintone color from the nonskintone color.

For each face training sample, classifier $F(c_r, c_b)$ is applied to each pixel of face image. Statistics results can be therefore collected in Fig. 9(b), the grayscale value of each pixels corresponding to its ratio to be skintone color in the training set. Therefore the darker the pixel is the less possible it will be a skintone color. Therefore, only 50% pixels with large grayscale value are included to generate the mean value for color-filtering. An experiment over 6423 face and 5601 nonface images samples is performed and achieves a recall rate of 99.5% while removing more than one third of false alarms.

3) SVM-Filter: SVM is a technique for learning from examples that is well-founded in statistical learning theory. Due to its high generalization ability, it has been widely used in area of object detection since 1997 [4]. However, kernel evaluation in SVM classifier is very time consuming and frequently yields to slow detection speed. Serra [17] proposed a new feature reduction algorithm to solve this problem. This work inspires a new way to reduce kernel size. For any input image u, v, the two-degree polynomial kernel is defined as

$$k(u, v) = (s(u \cdot v) + b)^2. \tag{9}$$

Serra extended it into a feature space with dimension $p = m^*(m + 3)/2$, where m is the dimensionality of sample u. For example, a sample with dimensionality 400 will be mapped into the feature space with dimensionality 80 600. In this space, the SVM kernel can be removed by computing the linear decision function directly. Moreover, with a simple weighting schema, Serra reduced the dimensionality to 40% without significant loss of classification performance.

In this section, based on the wavelet analysis of the input image, a new approach to further feature reduction without losing classification accuracy is proposed.

Wavelet transformation has been regarded as a complete image decomposition method with little correlation between each subband. This inspires a new way to reduce the redundancy of the feature space. The algorithm works as follows. First, the wavelet transformation is performed on the input image. As shown in Fig. 10, the original

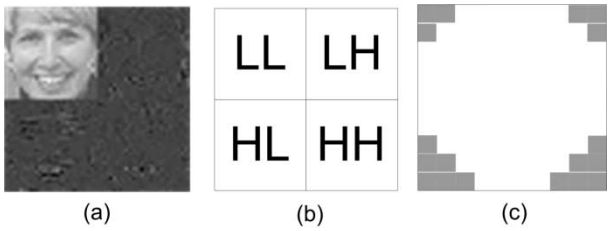

Fig. 10. Wavelet feature extraction. (a), (b) One-level wavelet transform. (c) Mask for cropping.

image of size 20×20 is divided into four subbands with a size of $10 * 10$. Then a hybrid second-degree polynomial SVM kernel, as shown in (10), is proposed to reduce the redundancy of the feature space

$$k'(u, v) = \sum_{0 \le i < 4} \left(s_i u_i^T v_i + r_i\right)^2 \tag{10}$$

where each vector u_i and v_i corresponds to a subband of transformed image.

Therefore, for a 20×20 image, the dimensionality of vector $u_i(v_i)$ is 100. As shown in Fig. 10(c), this dimensionality is further reduced to 82 by cropping out the four corners of each subband window, which mainly consists of image background. Consequently, the dimensionality of the feature space of kernel $k'(u, v)$ is $p^* = 4 * 82 * (82 + 3)/2 = 13\,940$. This results in a more compact feature space with much smaller (29%) features than Serra's approach, while similar classification accuracy is achieved in this space.

III. A Robust Multiview Face Detection System

In real life surveillance and biometric applications, human faces appeared in images have a large range of pose variances. We consider the pose variance in the range of out-of-plane rotation $\Theta = [-45°, 45°]$ and in-plane rotation $\Phi = [-45°, 45°]$, since state-of-the-art automatic face recognition algorithms are still not sufficiently robust to recognize detected face with poses out of these ranges.

Conventionally, it is very difficult to handle both of these variations in one classifier. Moreover, as Haar-like features, shown in Fig. 2(a)–(d), are sensitive to the horizontal and vertical variations, directly handling in-plane rotation is extremely difficult for boosting approaches. We address this problem by first applying an in-plane orientation detector to determine the in-plane orientation of a face in an image with respect to the upright position; then, an upright face detector this is capable of handling out-plane rotation variations in the range of $\Theta = [-45°, 45°]$ is applied to the candidate window with the orientation detected before. This section presents the design of these two detectors in detail.

A. In-Plane Rotation Estimator

Conventionally, the problem of in-plane rotation variations can be solved by training a pose estimator to rotate the window to an upright position [7]. This method results in the slow processing speed due to its high computation cost over pose correction on each candidate window. In this paper, another approach is therefore adopted, which consists of the following procedures

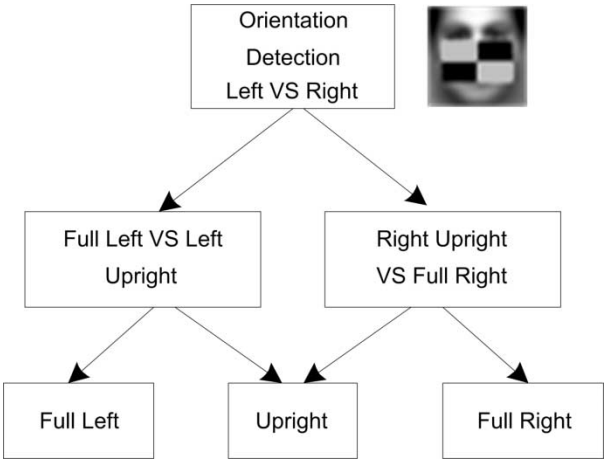

Fig. 11. In-plane pose estimation based on Haar-like features.

First, Φ is divided into three subranges, $\Phi_{-1} = [-45°, -15°]$, $\Phi_0 = [-15°, -15°]$, and $\Phi_1 = [15°, 45°]$. Second, the input image is in-plane rotated by $\pm 30°$. In this way, there are totally three images including the original image, and each corresponds to one of the three subranges, respectively. Third, in-plane orientation of each window on the original image is estimated. Finally, based on the in-plane orientation estimation, the upright multiview detector is applied to the estimated subrange at the corresponding location.

As shown in Fig. 11, the design of the pose estimator adopts the coarse-to-fine strategy [5]. The full range of in-plane rotation is first divided into two channels, and each one covers the range of $[-45°, 0°]$ and $[0°, 45°]$. In this step, only one Haar-like feature, as shown in Fig. 11, is used and results in the prediction accuracy of 99.1%. After that, a finer prediction based on AdaBoost classifier with six Haar-like features is performed in each channel to obtain the final prediction of the subrange.

B. Upright Multiview Face Detector

The use of in-plane pose prediction narrows down the face pose variation in the range of out-of-plane rotation Θ and in-plane rotation Φ_0. With such a variance, it is possible to detect upright faces in a single detector based on the proposed three-step algorithm. Other than the view-based methods, this architecture is promising for solving the problems of slow detection speed and high false alarm rates at the same time. Unfortunately, experimental results show that the boosting training procedure in Section II-C tends to converge slowly and is easy to over-fit. It reveals the limitation of Haar-like features in characterizing multiview faces.

To address this problem, three sets of new features based on an integral image, which is shown in Fig. 12, are proposed to enhance the discriminability of the basic Haar-like feature in Fig. 2. First, three features in the first row are proposed in which (a) enhances the ability to characterize vertical variations. Similarly, (b) and (c) are cable of capture the diagonal variations. Second, features (d)-(e) are more general, which do not require the rectangles in features are adjacent. As such features overwhelm the feature set with an extra degree of freedom dx, an extra constrain of mirror invariant is added to reduce the size of feature set while the most informative features are preserved.

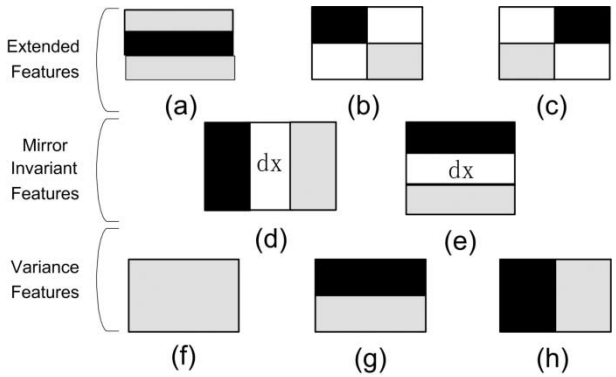

Fig. 12. Three sets of new features used in this system.

Finally, a set of three variance features are proposed to capture texture information of facial pattern. Different from the previous features, variance value instead of mean value of pixels in the feature rectangles is computed. With the utilizing of such second statistics, more informative features are available to distinguish the face pattern from the nonface pattern.

The introduction of the new features greatly increases the convergence speed of the training process. The experimental results show that nearly 69% of features selected by boosting are new features, in which more than 40% of the features are variance features. Therefore, the efficiency of those new features is demonstrated.

IV. Experimental Results

In this section, we evaluate the performance of our proposed pose-invariant face detection approach. We first analyze the performance of the proposed system, followed by the performance comparisons between the proposed three-step approach and four typical kinds of well-known face detectors in the literature.

A. Data Set

More than 12 000 nonface image and 8000 multiview face images with out-of-plane rotation variations in the range of $[-45°, 45°]$ were collected by cropping from various sources (mostly from the World Wide Web). A total number of about 80 000 face training samples with size of 20×20 are generated from the 8000 face images by following random transformation: mirroring, four-direction shift with one pixel, in-plane rotation within 15 degrees, and scaling within 20% variations.

Two image databases were used to evaluate the proposed algorithm and to compare it with other algorithms. One is the MIT+CMU frontal face test set [8], which is composed of 125 grayscale images containing 483 labeled frontal faces. The other is photo test sets collected by ourselves on various sources, and it could be divided into three subsets. Subset A has 154 photos, and most of them are upright frontal faces with ideal lighting. Subset B contains 55 photos, which are selected from a typical home photo album. Subset C consists of 400 home photos with large pose variations and outdoor lighting.

B. Computational Cost Analysis

The computational costs of face detection are varied when the scale or content of input image changed. Obviously, such

TABLE I
COMPUTATION COSTS ANALYSIS. THREE MODELS WITH DIFFERENT COMPLEXITY ARE EVALUATED OVER THE SAME TEST SET. IN TEST A, TIME COSTS FROM BOOSTING CHAIN WITH A PREFILTER ARE COLLECTED, AND IN TEST B TIME COSTS FROM THE OVERALL SYSTEMS ARE COLLECTED

Model No	ADC n	TestA T_a	TestB T_b	RatioA R_a	RatioB R_b
1	33.3	387.67s	388.42s	11.64	0.19%
2	8	96.78s	97.47s	12.10	0.71%
3	19.6	222.1s	222.82s	11.37	0.32%

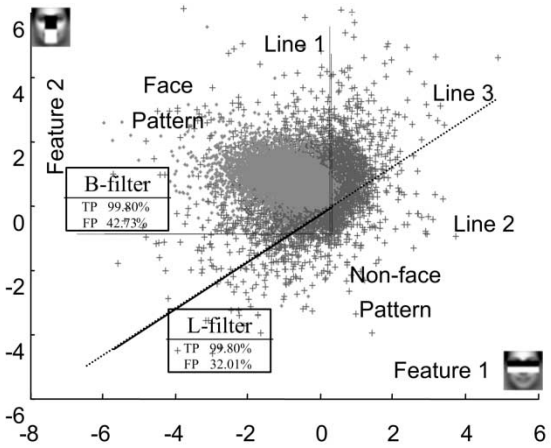

Fig. 13. The comparison between a prefilter and two feature boosting gives the experimental results of two kinds of classifiers. The B-filter is the boosting filter, the L-filter is the linear prefilter, TP is true positive rates, and FP is false positive rates.

variations are determined by the complexity of the detection model and input image. In order to represent such complexity, a value, called average detection complexity (ADC), is defined as how many features are expected to be used on average to predict whether an input subwindow contains a face. In this experiment, three detection models with different complexity are evaluated in the photo set, which contains 300 images. The ADC values n_i, time costs $T_{a,i}$ of a detector without a postfilter, and overall time cost $T_{b,i}$ are collected in Table I.

Given each feature's average time cost ratio $R_{a,i} = T_{a,i}/n_i$ and postfilter's time cost ratio $R_{b,i} = (T_{b,i} - T_{a,i})/T_{b,i}$, each model's overall time cost could be defined as

$$T_i = T_{b,i} = \frac{n_i^* R_{a,i}}{(1 - R_{b,i})}. \tag{11}$$

As the variance of vector $R_a = \{R_{a,i}\}$ is very small, the computation costs $T_{a,i}$ could be roughly regarded to be in direct proportion to the ADC value: $T_{a,i} \approx K^* n_i$ where $K = E(R_{a,i})$. Moreover, as the postfilter's time cost ratios are very small, $R_{b,i} \ll 1\%$, in most cases the computation cost of postfilter can be omitted. Consequently, and the overall computational cost is represented as

$$T_i = T_{b,i} \approx k^* n_i \tag{12}$$

where K is a constant related to the performance of computer hardware.

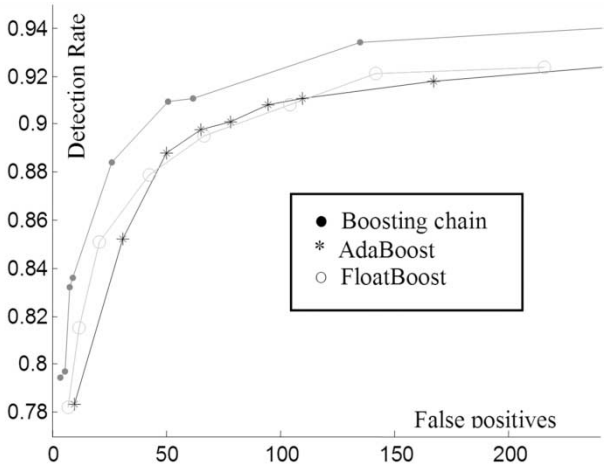

Fig. 14. Detection rates for various numbers of FPs on the MIT + CMU test set. All detectors are constructed in an 11-layer cascade.

TABLE II
AVERAGE NUMBER OF FEATURES USED IN FACE DETECTION IN THE MIT-CMU TEST SET

Boosting Chain	FloatBoost Cascade	Boosting Cascade
18.1	18.9	22.5

C. Performance Evaluation of the Three-Step Structure

1) Prefilter: To compare with the boosting approach, a set of experiments have been performed. As shown in Fig. 13, the linear filter reduces false alarm ratea by more than 25%, while the same recall rate and comparable computation cost are maintained.

2) Boosting Filter: Three detectors based on boosting chain, FloatBoost cascade [15], and Adaboost cascade have been implemented on the same training set for the comparison. The FP-Detection rate curve over the MIT+CMU test is shown in Fig. 14, and the ADC values of each detector are listed in Table II.

In order to sidestep any differences resulting from the underlying infrastructure systems of the detector [10], a training set of 18 000 images (8000 faces and 10 000 nonfaces) and a test set of 15 000 images (5000 faces, and 10 000 nonfaces) have been used to evaluate these algorithms. The images are $20 * 20$ grayscale and are aligned by the eye center.

From the Detection-FP rate curve shown in Fig. 14, the boosting chain approach outperforms Adaboost cascade and FloatBoost cascade with similar ADC values. It works especially well at higher recall rates. This property will greatly enhance the efficiency of the postfiltering procedure. In addition, from Table II, the boosting chain algorithm again achieves the best performance. Compared with the result reported in [12], where only seven or eight features required on the average to predict a window, the AdaBoost detector implemented here used much more features due to the complexity of the training set.

3) Postfilters: An SVM classifier with a two-degree polynomial kernel for face detection has been well studied over

TABLE III
COMPARISON OF THE TWO-DEGREE POLYNOMIAL SVM POSTFILTER ON PHOTO TEST SETS. R = recall, F = FP RATES

	Set A		Set B		Set C	
	R	F	R	F	R	F
Hybrid 2d-polynomial	98.68	25	95.95	26.65	91.79	11.64
2d- polynomial	99.34	28.61	94.59	27.74	92.86	13.28

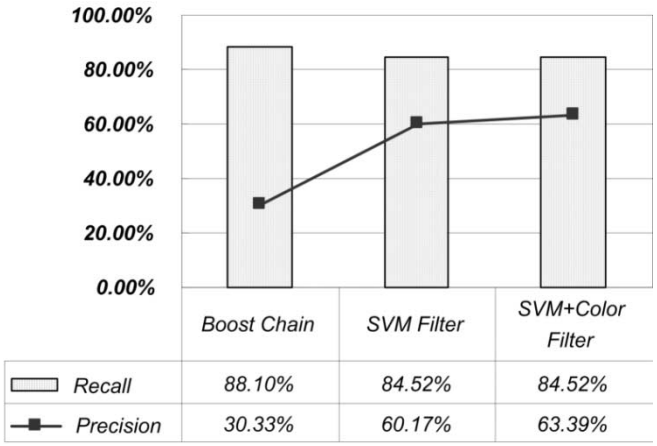

Fig. 15. Experimental results of postfiltering.

TABLE IV
DETECTION RESULTS ON FACES WITH OUT-OF-PLANE ROTATION

	Pie Frontal	Pie half- profiled	Pie profiled
Recall	91.28	90.14	6.175
Precision	96.12	94.32	63.99

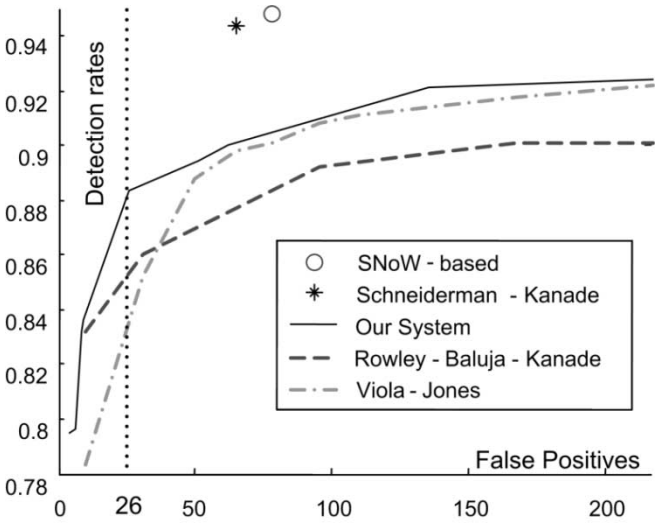

Fig. 16. Detection rates for various numbers of FPs on the MIT + CMU test set.

TABLE V
COMPARISON OF OUR SYSTEM, VIOLA–JONES BOOSTING CASCADE ON PHOTO TEST SETS. R = recall, P = precision

Algorithms	Set A		Set B		Set C	
	R	P	R	P	R	P
Viola	82.58	96.24	73.81	52.99	61.97	22.92
3-Step(NP)	98.06	88.37	88.1	30.33	78.87	23.35
3-Step(SP)	97.4	96.77	84.52	60.17	72.39	70.6
3-Step	97.4	98.68	84.52	63.39	72.39	75.81

years. In this section, experimental comparison between the proposed new hybrid kernel and standard approach are made in Table III. The differences between two classifiers are subtle, and, in most cases, the standard two-degree polynomial kernel is slightly better in recall rates and worse in FP rates. However, as discussed in Section II-D, hybrid kernel is superior with only 17.3% computational and storage costs.

Different from the SVM-filter, color-filter is more conservative. Most of the time, it improves the detection precision without the significant loss of recall rates. Such a property makes it a good supplement to the SVM-filter. Fig. 15 depicts the experimental results of such a combination.

D. Face Detection on the Nonfrontal Data Set

Three test sets have been collected from the CMU PIE database to evaluate the performance of our system on handling nonfrontal faces. The first set is the frontal set which contains face images with out-of-plane rotation poses within the range of $[-20°, 20°]$. The second set is the half-profiled set which contains nonfrontal face images with out-of-plane rotation poses of less than 45°. The third set is the profiled set which containing of face images with out-of-plane rotation poses greater than 45°. The experimental results are depicted in Table IV.

According to Table IV, the results on test set 1 and test set 2 are much better than the results from test set 3. It reveals that the proposed system is sensitive to out-of-plane rotation.

E. Performance Comparisons

1) MIT + CMU Frontal Face Test Set: In Fig. 16, the experimental results from an upright multiview detector (in Section III-B) is compared with the results reported on the same data set from Viola-Jones [12] (boosting cascade with training samples of size 24×24), Rowley [8] (Neural network with training samples of size 20×20), Roth-Yang [3] (SNoW-based face detector), and Schneiderman [6] (AdaBoost on wavelet coefficients). From the experimental results in Fig. 16, our system outperforms the results achieved by Viola [12] and Rowley [8]. Although the accuracy is lower than that of Roth-Yang [3] and Schneiderman [6], our system is approximately 15 times faster than most approaches (except Viola's, which is about the same speed as ours).

2) Photo Test Sets: To evaluate the performance of the proposed approach with comparison to Viola–Jones algorithm on the three sets of real life images, we have implemented the Viola–Jones algorithm as a baseline with the same training set as our current system. Experimental results are shown in Table V, where the symbol "NP" stands for the meaning of "without postfiltering" and "SP" stands for the meaning of "with only SVM-filtering."

In Table V, the expected higher recall rates have been achieved in all experiments with only a very slight loss of

Fig. 17. Sample experimental results using our method on images from CMU-MIT frontal, rotated, and profiled face database.

precision. Compared to that of Viola's approach, the decrease in precision on test set B, due to complex backgrounds in these photos, indicates that the approach is not always optimal to maintain the high precision ratio. This is because the prefiltering and boosting filtering processes are designed to preserve recall ratio effectively.

F. Detection Result on Sample Images

To demonstrate the effectiveness of the proposed algorithm while handling face with out-of-plane rotation, two face image sets, "CMU profiled" and "PIE," are used as the test set. Detection results on sample images from these sets are shown in Figs. 17 and 18.

G. Discussions

From experiment results shown in Figs. 13–18 and Tables I–V, it is seen the performance of proposed approach and the multipose face detection system in following aspects.

1) According to the results from Figs. 13 and 14 and Tables II and V, the prefiltering and boosting filtering in the three-step approach are effective to achieve a high recall ratio while maintaining a comparable false alarm ratio. Although, as shown in Table V, such a recall ratio improvement is penalized by the decrease in precision; this shortcoming is overcome by applying the postfilters, as indicated in the third and fourth experiments in Table V.
2) SVM-filter and color-filter are designed to reduce false alarms without significant loss of recall ratio. In these experiments, the SVM-filter proves that it is effective as a postfilter. It removed most remaining false alarms at a cost of losing 4% recall ratio on the average.
3) The color-filter is robust enough to improve the precision without the recall ratio decreasing in all three testing sets.
4) From the recall-precision curve shown in Fig. 16, the three-step approach outperforms those of Viola [12] and Rowley [8], and it works especially well at low false alarm rates. This reveals the efficiency of the postfiltering procedure.
5) Also, in Fig. 16, it is noticed that the accuracy of the results of the Roth-Yang [3] and Schneiderman [6] algorithms are superior to that of others. However, such

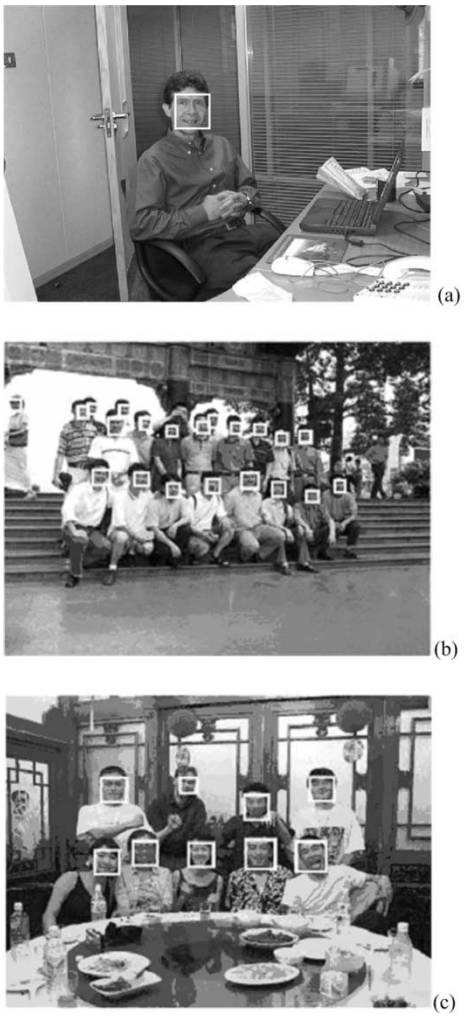

Fig. 18. Sample experimental results from three digital photo sets. Images (a), (b), and (c) are collected from photo sets A, B, and C, respectively.

performance improvement is penalized by the drastically decreasing detection speed.

6) Experimental results from three test sets of real-life images reveal the robustness and the high accuracy of the proposed face detection system.

To conclude, in the three-step face detection framework, a prefilter accelerates the detection speed, the boosting chain increases recall rates. and postfilters improve precision rates. By integrating these characteristics, the proposed system demonstrates its superior performance to that of the boosting cascade approach.

V. Conclusion

In this paper, a novel framework for rapid and pose-invariant face detection has been presented. In this framework, face detection is divided into three steps: prefiltering, focused on improving detection speed with a linear filter, a linear SVM optimized boosting chain filter, aimed at removing most nonface candidates while maintaining a high recall rate, and post filtering, targeted at further reducing false alarms. Based on this framework, and together with a two-level hierarchy in-plane pose estimator, a real-time system for multiview face detection in photos has been built.

The experiment results from most testing sets have shown the robustness and superiority of the proposed system. Also, we believe the generic framework presented in this paper can be applied to other classification problems in computer vision.

References

[1] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces of face recognition," in *Proc. IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, June 1994, pp. 84–91.

[2] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proc. 6th Int. Conf. Computer Vision*, Jan. 1998, pp. 555–562.

[3] D. Roth, M. Yang, and N. Ahuja, "A SNoW-based face detection," in *Advances in Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K.-R. Muller, Eds. Cambridge, MA: MIT Press, 2000, vol. 12, pp. 855–861.

[4] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection," in *Proc. IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, June 1997, pp. 130–136.

[5] F. Fleuret and D. Geman, "Coarse-to-fine face detection," *Int. J. Computer Vision*, vol. 20, pp. 1157–1163, 2001.

[6] H. Schneiderman and T. Kanade, "A statistical method for 3D object detection applied to faces and cars," in *Proc. IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, June 2000, pp. 746–751.

[7] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 22–38, 1998.

[8] Neural Network-Based Face Detection, H. A. Rowley.*http://www-2.cs.cmu.edu/~har/thesis.ps.gz* [Online]

[9] J. Ng and S. Gong, "Performing multi-view face detection and pose estimation using a composite support vector machine across the view sphere," in *Proc. IEEE Int. Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, Corfu, Greece, Sept. 1999, pp. 14–21.

[10] M. Alvira and R. Rifkin, "An Empirical Comparison of SNoW and SVM's for Face Detection," Massachusetts Inst. of Technology, Cambridge, MA, CBCL Paper #193/AI Memo #2001–004, 2001.

[11] M. Bichsel and A. P. Pentland, "Human face recognition and the face image set's topology," *CVGIP: Image Understanding*, vol. 59, pp. 254–261, 1994.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. 2001 IEEE Computer Soc. Computer Vision and Pattern Recognition*, vol. 1, HI, Dec 2001, pp. 511–518.

[13] R. E. Schapire. The boosting approach to machine learning: An overview. presented at Proc. MSRI Workshop on Nonlinear Estimation and Classification [Online]

[14] R. L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 696–706, May 2002.

[15] S. Z. Li *et al.*, "Statistical learning of multi-view face detection," in *Proc. 7th Eur. Conf. Computer Vision.*, Copenhagen, Denmark, May 2002, pp. 67–81.

[16] T. Poggio and K. K. Sung, "Example-based learning for view-based human face detection," in *Proc. ARPA Image Understanding Workshop*, vol. II, 1994, pp. 843–850.

[17] T. Serre *et al.*, "Feature selection for face detection," in *AI Memo 1697*: Massachusetts Institute of Technology, 2000.

[18] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[19] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Computer Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

Rong Xiao received the Ph.D. degree from Nanjing University, Nanjing, China, in 2001.

He joined Microsoft Research China as an Associate Researcher in July 2001 to pursue his research interests in statistical learning, face detection and recognition. His research interests include machine learning and object detection and tracking.

Ming-Jing Li received the B.S. degree in electrical engineering from the University of Science and Technology of China, Beijing, in 1989 and the Ph.D. degree in pattern recognition from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 1995.

He joined Microsoft Research China in July 1999. His research interests include handwriting recognition, statistical language modeling, search engines, and image retrieval.

Hong-Jiang Zhang received the B.S. degree from Zhengzhou University, China, and the Ph.D. degree from the Technical University of Denmark, both in electrical engineering, in 1982 and 1991, respectively.

From 1992 to1995, he was with the Institute of Systems Science, National University of Singapore, where he led several projects in video and image content analysis and retrieval and computer vision. He also worked at MIT Media Lab in 1994 as a Visiting Researcher. From 1995 to 1999, he was a Research Manager at Hewlett-Packard Labs, where he was responsible for research and technology transfers in the areas of multimedia management; intelligent image processing and Internet media. In 1999, he joined Microsoft Research Asia, Beijing, China, where he is currently a Senior Researcher and Assistant Managing Director in charge of media computing and information processing research. He has authored three books, over 260 referred papers and book chapters, seven special issues of international journals on image and video processing, content-based media retrieval, and computer vision, as well as over 45 patents or pending applications.

Dr. Zhang is a Member of ACM. He currently serves on the editorial boards of five IEEE/ACM journals and a dozen committees of international conferences.