Robust Multi-Pose Face Detection in Images

Rong Xiao, Ming-Jing Li, 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 non-face-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 the original cascade approaches [12]. Last, a post-filtering algorithm consists of image pre-processing; SVM-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 multi-view face detector is therefore built. The experimental results demonstrate the significant performance improvement using the proposed approach over others.

Keywords

Face Detection, AdaBoost, Support Vector Machine

† This is draft version of manuscript, which is for IEEE Transactions on Circuits and Systems for Video Technology.
* The corresponding authors are with Microsoft Research Asia, 3/F Beijing Sigma Center, No. 49 Zhichun Road, Hai Dian District, Beijing, 100080, China. (Tel: 86-10-62617711; Fax: 86-10-88097306; email: i-rxiao@microsoft.com, hjzhang@microsoft.com)

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. It 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. They 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. The SNoW learning architecture 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 is built on an over-complete set of Haar-like features that integrates the feature selection and classifier design in the same framework.

Most non-frontal 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 3D modeling is avoided. [7] partitioned the views of face into five channels, and developed a multi-view detector by training separate detector networks for each view. [9] studied the trajectories of faces in linear PCA feature spaces as they rotate, and used SVMs for multi-view face detection and pose estimation. The work in [6] used multi-resolution 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 alarm rates of classifiers in the earlier stage, the inefficiency of AdaBoost significantly degrades the overall performance. Secondly, as each boosting classifier works separately, the useful information between adjacent layers are discarded, which hampers the convergence of the training procedure. Thirdly, during the training process, more and more non-face 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 non-face become so complicated that can hardly be distinguished by Haar-like feature. 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 2-3 features. As few features used in this step, training extensive algorithms with global optimization characteristics are affordable to obtain a high performance pre-filter. In the second step, the classifier should be "efficient", reducing false positive rate to the scale of 10^{-7} with as small computation cost as possible. As most prediction will be done 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 false positives precisely. As most false positives 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 multi-view 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, firstly, a two-level hierarchy in-plane pose estimator based on Haar-like feature is built to alleviate the variance of in-plane rotation by dividing the input window into three channels. Secondly, an upright face detector based on the three-step algorithm is built, which enables the rapid multi-view 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 multi-view face detection system is presented in Section III. Section IV provides the experimental results and conclusion is 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 Figure 1, consists of following components: First, a linear pre-filter is used to increase the detection speed. Second, a boosting chain, developed from Viola's boosting cascade [12], is applied to remove most non-faces 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 a SVM filter are used to further reduce false alarms. Each of these components is described in detail in this section.

(Fig. 1 should be around here)

A. Basic concepts of Detection with Boosting Cascade

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

(Fig. 2 should be around here)

Haar-like feature: Four types of Haar-like features, which are shown in Figure 2 [12]. These features are computed by mean value difference between pixels in the black rectangles and pixels in the grey rectangles.

Both are sensitive to horizontal and vertical variations, which are critical to capture upright frontal face appearance.

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 $h_t(x) = sign(p_t f_t(x) - \theta_t)$, and θ_t is the threshold for the decision stump, p_t is the parity to indicate the direction of decision stump.

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

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

where $i(s',t')$ is grayscale value of the original image. Based on this definition, the mean 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 non-face 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 640x480 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. It results in a rapid detection system

B. Linear pre-filter

Adaboost, developed by Freund and Schapire [19], has been proved to be a powerful learning method for face detection problem. Given $(x_1, y_1),...,(x_n, y_n)$ 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:

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

In Equation (2), α_t is a coefficient, b is a threshold, $h_t(x)$ is a one-dimension 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 space, as shown in (a)-(b) of Figure 3. As only sign information of $h_t(x)$ is used in (2), the discriminablity of final decision function is greatly affected.

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

$$H(x) = (a_1 f_1(x) > b_1) \wedge (a_2 (f_1(x) + r f_2(x)) > b_2), \quad (3)$$

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

(Fig. 3 should be around here)

The first term in Equation (3) is a simple decision stump function, which can be learned by adjusting threshold according to the face/non-face 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

Boosting cascade proposed by Viola has been 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 one is how to improve the efficiency of threshold adjusting. We propose a *boosting chain* with linear SVM optimization to address these two issues.

(Fig. 4 should be around here)

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 layers cascade, to anticipate a overall detection rates at 96% in 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 sharply precision decreasing. As shown in Figure 5, value *b* is computed for the best precision, and value *a is the best threshold which* satisfies the minimal recall 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 learner used in the boosting algorithm is near to random guess, such discriminative information discarded between the layers of boost cascade is critical to increase the converge speed of successive classifiers.

(Fig. 5 should be around here)

To address this issue, a chain structure of boosting cascade is proposed (as shown in Figure 4). The algorithm is designed as follows:

(Fig. 6 should be around here)

As shown in Fig 6, 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 is directly introduced into the substantial learning procedure. For negative samples, collected by bootstrap method, their weights are adjusted according to the classification errors of each previous weak classifier. Similar to the equation used in boosting training procedure [13], the adjusting could be done by:

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

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, result from previous node classifier is not discarded while training of 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 of boosting chain could be done in following manner:

(Fig 7. should be around here)

2). Linear optimization

In each step of 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 the balance between the targeted recall and false positive 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 post-optimization is proposed.

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

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

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

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:

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

subject to the constraints $\sum_i^n \beta_i y_i = 0$ and $C_i \geq \beta_i \geq 0$, $i = 1, ..., n.$ Coefficient C_i is set according to the classification risk w and trade-off 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 $\beta^0 = (\beta_1^0, \beta_2^0 ..., \beta_n^0)$. 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 Figure 8 illustrated the efficiency of this algorithm.

(Fig. 8 should be around here)

D. Post-filter

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

1). Image Pre-processing.

The processing procedure aims to alleviate background, lighting and contract variations. It consists of three steps [7]: First of all, a mask, which 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 subtracted the plane generated by this linear function, the lighting variations could be significantly reduced. Finally histogram equalization is performed. With this non-linearly 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 skin-tone color has been studied extensively in recent years [14]. In our system, YC_bC_r space is adopted due to its perceptually uniform. As Y component mainly represents image grayscale information which is quite irrelevant to skin-tone color, only C_b and C_r components are reserved for false alarm removal.

As shown in Figure 9(a), the color of face and non-face images is distributed as nearly Gaussian in C_bC_r space. 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:

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

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 constructed in this five dimension space to separate skin-tone color from the non-skin-tone 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 figure 9(b), the grayscale value of each pixels corresponding to its ratio to be skin-tone color in the training set. Therefore the darker the pixel is the less possible it will be a skin-tone 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 non-face images samples is performed. And it achieves a recall rate of 99.5% while removing more than one third of false alarms.

(Fig. 9 should be around here)

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 80600. In this space, SVM kernel can be removed by computing the linear decision function directly. With a simple weighting schema, Serra reduced 40% features without significant loss of classification performance.

Based on the wavelet analysis of the input image, a new approach to further feature reduction without loosing classification accuracy is proposed.

Wavelet transformation has been regarded as a complete image decomposition method with little correlation between each sub-band. This inspires a new way to reduce the redundancy of the feature space. The algorithm works as following. First, the wavelet transformation is performed on the input image. As shown in Figure 10, the original image of size 20x20 is divided into four sub-bands with size of 10*10 Then a hybrid second-degree polynomial SVM kernel, as shown in Equation (10), is proposed to reduce the redundancy of the feature space,

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

where each vector u_i and v_i corresponds to a sub-band of transformed image.

(Fig. 10 should be around here)

Therefore, for a 20x20 image, the dimensionality of vector u_i (v_i) is 100. As shown in Figure 10(c), this dimensionality is further reduced to 82 by cropping out the four corners of each sub-band window, which mainly consists of image background. Consequently, the dimensionality of the feature space of kernel $k'(u,v)$ is $p^* = 4*82*(82+3)/2=13940$. 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 MULTI-VIEW 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 arts automatic face recognition algorithm 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 Figure 2(a)-(d), are sensitive to the horizontal and vertical variations, directly handle 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 up-right position; then, an up-right 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 in detail the design of these two detectors.

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 following procedures: Firstly, Φ is divided into three sub-ranges, $\Phi_{-1} = [-45°,-15°]$, $\Phi_0 = [-15°,15°]$ and $\Phi_1 = [15°,45°]$. Secondly, 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 sub-ranges respectively. Thirdly, in-plane orientation of each window on the original image is estimated. Finally, based on the in-plane orientation estimation, the upright multi-view detector is applied to the estimated sub-range at the corresponding location.

As shown in Figure 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, each one covers the range of $[-45°,0°]$ and $[0°,45°]$. In this step, only one Haar-like feature, as shown in Figure 11, is used and results in the prediction accuracy of 99.1%. After that a finer prediction based on AdaBoost classifier with 6 Haar-like features is performed in each channel to obtain the final prediction of the sub-range.

(Fig. 11 should be around here)

B. Upright multi-view 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 variance, it's 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 to solve the problems of slow detection speed and high false alarm rates at the same time. Unfortunately, experiment 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 feature in characterizing multi-view faces.

To address this problem, three sets of new features based on integral image, which is shown in Figure 12, are proposed to enhance the discriminability of the basic Haar-like feature in Figure 2. Firstly, 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. Secondly, feature (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. 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 2^{nd} statistics, more informative features are available to distinguish the face pattern from the non-face pattern.

(Fig. 12 should be around here)

The introduction of the new features greatly increase the converge speed of training process. The experimental results show that nearly 69% features selected by boosting are new features, in which more than 40% 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 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 12000 non-face image and 8000 multi-view face images with out-of-plane rotation variations in the range of [-45º, 45º], were collected by cropping from various sources (mostly from WWW). A total number of about 80000 face training samples with size of 20x20 are generated from the 8000 face images by following random transformation: mirroring, four-direction shift with 1 pixels, 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 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 sub-sets. Sub-set A has 154 photos, and most of them are upright frontal faces with ideal lighting. Sub-set B contains 55 photos, which is selected from a typical home photo album. Sub-set C consists of 400 home photos with large pose variations and out-door lighting.

Moreover, CMU profiled face test set and PIE face database are used to demonstrate the effectiveness of the proposed algorithm while handling face with out-of-plane rotation. Sample images from these dataset are shown in Figure 17 and Figure 18.

B. Computational cost analysis

The computational costs of face detection are varied when the scale or content of input image changed. Obviously, such 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 sub-window 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 detector without post-filter and overall time cost $T_{b,i}$ are collected in Table 1.

(Table 1 should be around here)

Given each feature's average time cost ratio $R_{a,i} = T_{a,i}/n_i$, and post-filter'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:

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

As the variance of vector $R_a=\{R_{a,i}\}$ is very small, the computation costs $T_{a,i}$ could be roughly regarded as in direct proportion to the ADC value: $T_{a,i} \approx K * n_i$ *where* $K=E(R_{a,i})$. Moreover, as the post-filter's time cost ratios are very small, $R_{b,i} << 1\%$, in most cases the computation cost of post-filter 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 the performance of computer hardware.

C. Performance evaluation of 3-Step structure.

1). Pre-filter

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

(Figure 13 should be around here)

2). Boosting filter

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

(Table 2 should be around here)

(Figure 14 should be around here)

In order to sidestep any differences resulting from the underlying infrastructure systems of detector [10], a training set of 18000 images (8000 faces and 10000 non-faces) and a test set of 15000 images (5000 faces, and 10000 non-faces) have been used to evaluate these algorithms. The images are 20*20 grayscale and aligned by eye center.

From the Detection-FP rate curve shown in Figure 14, the boosting chain approach outperforms Adaboost cascade and FloatBoost cascade with similar ADC values. It works especially well at higher recall rate. This property will greatly enhance the efficiency of the post-filtering procedure. In addition, from Table 2, the boosting chain algorithm again achieves the best performance. Compared with the result reported in [12], where only 7-8 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). Post-filters

SVM classifier with two-degree polynomial kernel for face detection has been well studied over years. In this section, experimental comparison between the proposed new hybrid kernel and standard approach are made in Table 3. The differences between two classifier are subtle, and in most cases, the standard two-

degree polynomial kernel are slightly better in recall rates and worse in false positive rates. However, as discussed in Section II (D), hybrid kernel is superior with only 17.3% computational and storage costs.

(Table 3 should be around here)

Different from SVM filter, Color filter is more conservative. In most time, it improves the detection precision without the significant loss of recall rates. Such a property makes it a good supplement to SVM filter. Figure 15 depict the experimental results of such a combination.

(Figure 15 should be around here)

D. Face detection on non-frontal data set

Three test sets have been collected from CMU PIE database to evaluate the performance of our system on handling non-frontal 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 non-frontal face images with out-of-plane rotation poses 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 4.

(Table 4 should be around here)

According to Table 4, the results on test set 1 and test set 2 are much better than the result from test set 3. It reveals that the proposed system is sensitive to the out-of-plane rotation.

(Fig. 16 should be around here)

D. Performance comparisons

1). MIT+CMU frontal face test set

In Figure 16, the experimental results from upright multi-view 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 24x24), Rowley [8] (Neural network with training samples of size 20x20), Roth-Yang [3] (SNoW-based face detector), and Schneiderman [6] (AdaBoost on wavelet coefficients). From the experimental results in

Figure 16, our system outperforms the results given in 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 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 5, where the symbol "NP" stands for the meaning of "without post-filtering", and "SP" stands for the meaning of "with only SVM-filtering".

In Table 5, the expected higher recall rates have been achieved in all experiments at a very light penalty of the precision lose. 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 pre-filtering and boosting filtering processes are designed to preserve recall ratio effectively.

E. Discussions

From experiment results shown in Figure 13 - 18, and Table 1-5, it is seen the performance of proposed approach and the multi-pose face detection system in following aspects:

a). According to the results from Figure 13, Figure 14, Table 2 and Table 5, the pre-filtering and boosting filtering in the three-step approach are effective to achieve high recall ratio while maintaining comparable false alarm ratio. Although, as shown in Table 5, such recall ratio improvement is penalized by the decreasing of precision rate, this shortcoming is overcome by applying the post-filters, as indicated in the third and fourth experiments in Table 5.

b). SVM-filter and color-filter are designed to reduce false alarms without significant lose of recall ratio. In these experiments, SVM-filter proves that it is effective as a post-filter. It removed most remaining false alarms at a cost of loosing 4% recall ratio on the average.

c). The color-filter is robust enough to improve the precision without recall ratio decreasing in all three testing set.

d). From the recall-precision curve shown in Figure 16, the three-step approach outperforms Viola [12] and Rowley [8] algorithm. It works especially well at low false alarm rate. This reveals the efficiency of the post-filtering procedure.

e). Also, in Figure 16, it is noticed that the accuracy from the results of the Roth-Yang [3] and Schneiderman [6] algorithms are superior to that of others. However, such performance improvement is penalized by the drastically deceasing of detection speed.

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

To conclude, in the three-step face detection framework, pre-filter accelerates the detection speed, boosting chain increases recall rates and post-filters improve precision rates. By integrating these characteristics, the proposed system demonstrates its superior to the boosting cascade approach.

V. CONCLUSIONS

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: pre-filtering, focused on improving detection speed with a linear filter; a linear SVM optimized boosting chain filter, aimed to remove most non-face candidate 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 multi-view 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". *Proc. of IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, pp. 84-91, June 1994. Seattle, Washington.

[2] C. P. Papageorgiou, M. Oren, and T. Poggio. "A general framework for object detection". *Proc. of International Conf. on Computer Vision*, 1998.

[3] D. Roth, M. Yang, and N. Ahuja. "A snowbased face detection". *Neural Information Processing*, 12,2000

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

[5] F. Fleuret and D. Geman. "Coarse-to-fine face detection". *International Journal of Computer Vision* 20 (2001) 1157-1163

[6] H. Schneiderman and T. Kanade. "A Statistical Method for 3D Object Detection Applied to Faces and Cars". *Proc. IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, 2000

[7] H. A. Rowley, S. Baluja, and T. Kanade. "Neural network-based face detection". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (1998), pages 22-38.

[8] H. A. Rowley. *Neural Network-Based Face Detection*, Ph.D. thesis. CMU-CS-99-117, http://www-2.cs.cmu.edu/ ~har/ thesis.ps.gz

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

[10] M. Alvira, and R. Rifkin, "An Empirical Comparison of SNoW and SVMs for Face Detection", CBCL Paper #193/AI Memo #2001-004, Massachusetts Institute of Technology, Cambridge, MA, January 2001.

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

[12] P. Viola and M. Jones. "Robust real time object detection". *IEEE ICCV Workshop on Statistical and Computational Theories of Vision*, Vancouver, Canada, July 13, 2001.

[13] R. E. Schapire. "The boosting approach to machine learning: An overview". *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

[14] R. L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face Detection in Color Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence* Vol.24, No.5, pp 696-706, 2002.

[15] S. Z. Li, *et al*. "Statistical Learning of Multi-View Face Detection". *Proc. of the 7th European Conf. on Computer Vision*. Copenhagen, Denmark. May, 2002.

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

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

[18] V. N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, Inc., New york, 1998.

[19] Y. Freund and R. E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of Computer and System Sciences*, 55(1):119--139, August 1997.

Figure Lists:

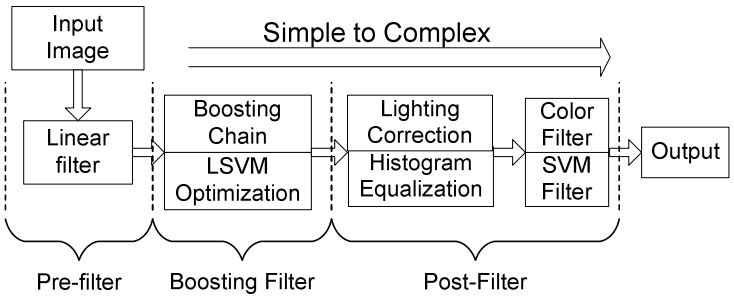


Figure 1**:** Three-step face detector

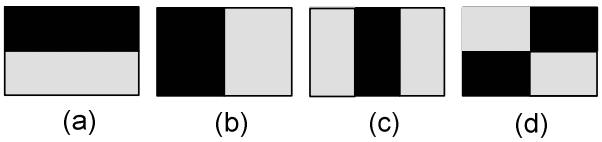


Figure 2**:** The basic Haar-like features

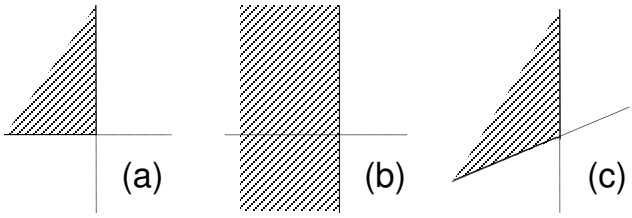


Figure 3**:** Two-feature boosting classifier VS linear pre-filter. (a) and (b) are boundaries of boosting, while (c) is the decision boundary of linear filter.

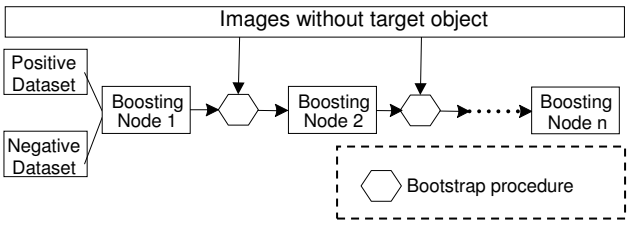


Figure 4**:** Boosting chain structure

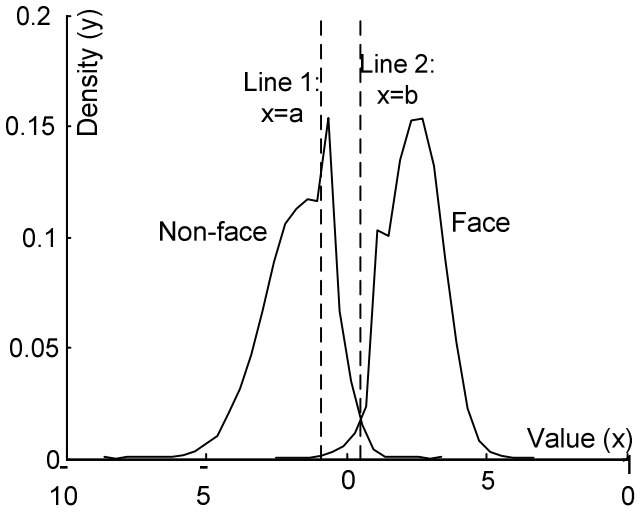


Figure 5**.** Adjusting threshold for layer classifier.

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 positive 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.
      – Using initial weights w_j, training set P and N_i
      – 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).

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

---

a). Given an example x, evaluate the boosting chain with M node
b). Initialize $s = 0$
c). 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.
d). Exit with positive response.

Figure 7: Evaluate the boosting chain

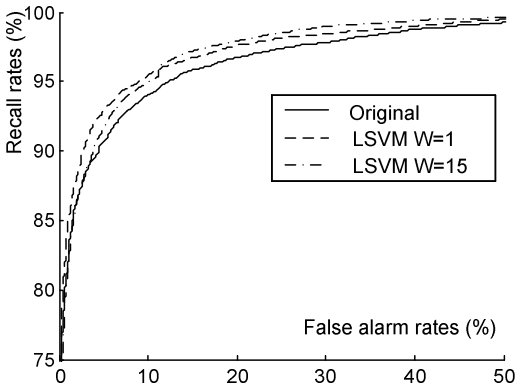


Figure 8**:** The ROC curves comparing the original Boosting chain algorithm with the LSVM optimization algorithm with different weights.

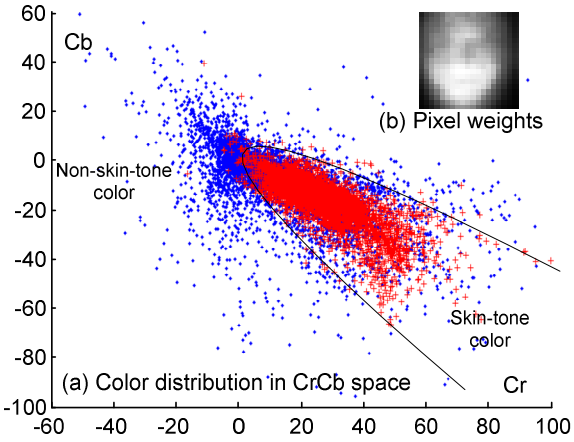


Figure 9**:** Two-degree polynomial color filter in C_rC_b space. The pixel weights are shown at top-right. The darker the pixel is the less important it will be.

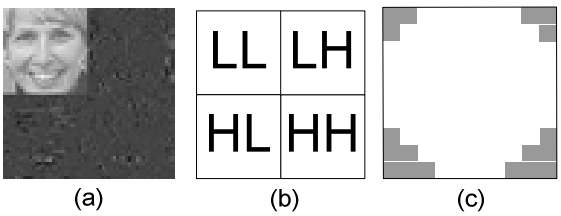


Figure 10: Wavelet feature extraction. (a)-(b) represents the one-level wavelet transform, (c) defines the mask for cropping.

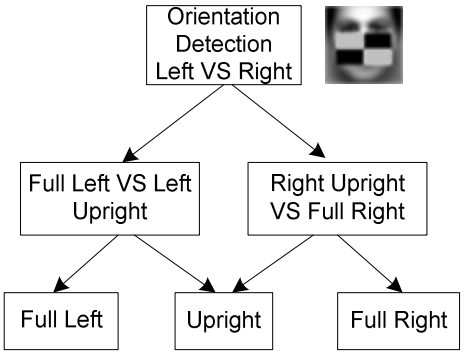


Figure 11: In-plane pose estimation based on Haar-like feature

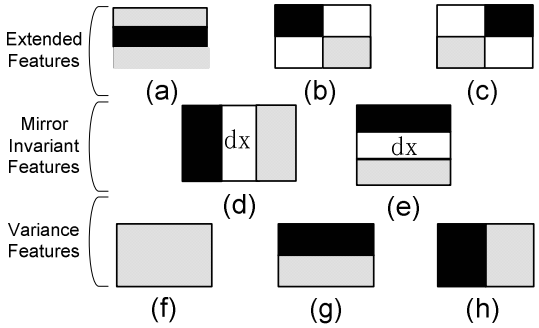


Figure 12**:** Three sets of new features used in this system.

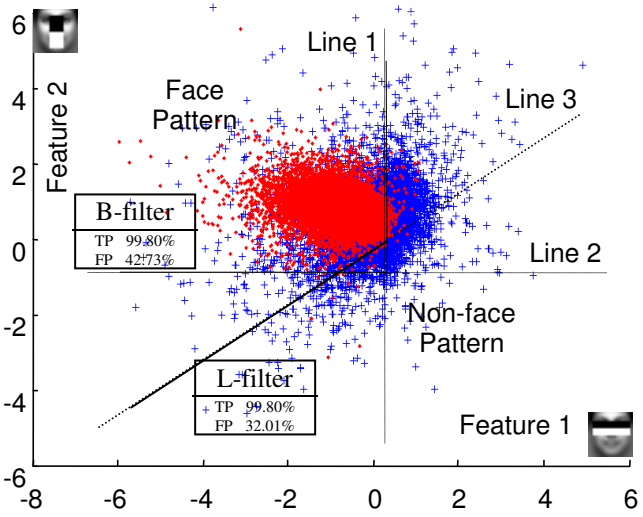


Figure 13: The comparison between pre-filter and 2 feature boosting gives the experimental results of two kinds of classifiers. B-filter is the boosting filter, L-filter is linear pre-filter, TP is true positive rates, and FP is false positive rates.

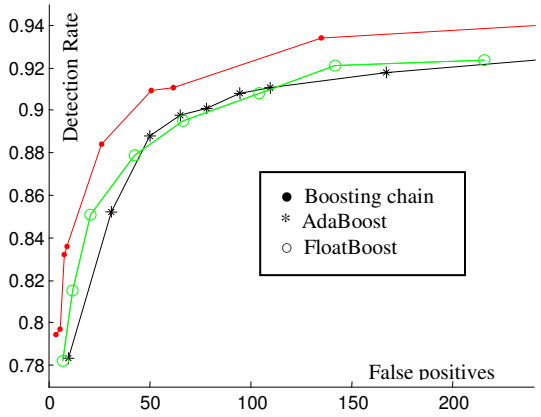


Figure 14: Detection rates for various numbers of false positives on the MIT+CMU test set. All detectors are constructed in 11 layer cascade.

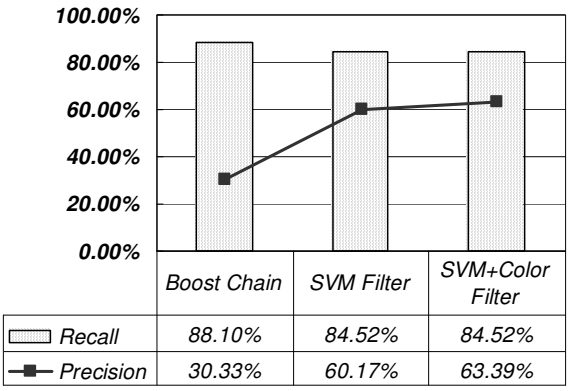


Figure 15: The experimental results of post filtering.

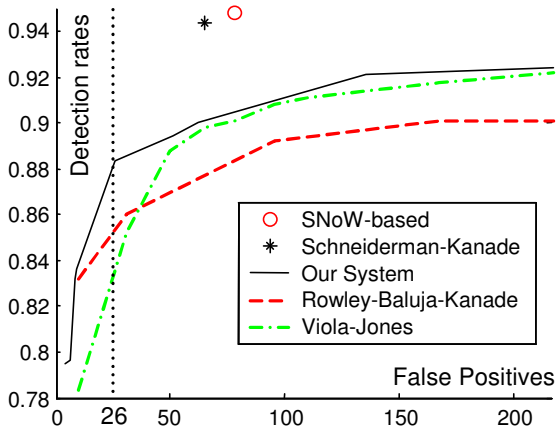
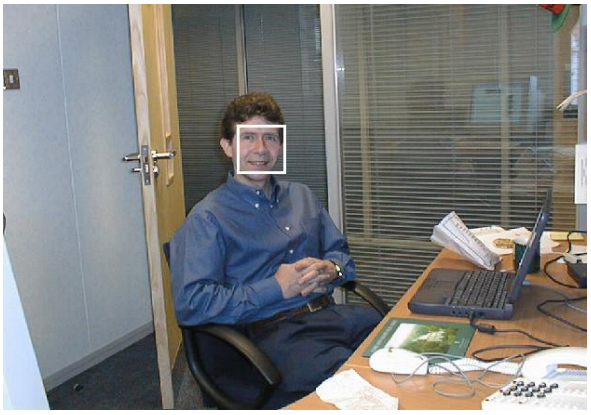
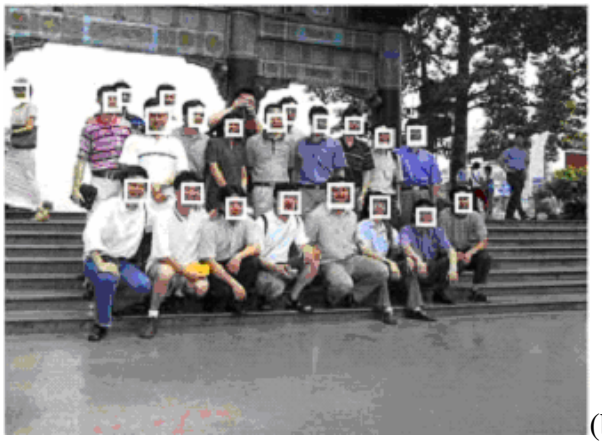


Figure 16: Detection rates for various numbers of false positives on the MIT+CMU test set

Figure 17: Sample experiment results using our method on images from CMU-MIT frontal, rotated and profiled face database



(a)



(b)

(c)

Figure 18. Sample experiment results from three digital photo set. Images (a), (b) and (c) are collected from photo set A, B and C respectively.

Table Lists:

Tabel 1: Computation Costs Analysis. Three models with different complexity are evaluated over the same test set. In TestA, time costs from boosting chain with pre-filter are collected, in TestB, 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	388.42	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%

Tabel 2 Average number of feature used in face detection on MIT-CMU Test set

Boosting Chain	FloatBoost Cascade	Boosting Cascade
18.1	18.9	22.5

Table 3: Comparison of two-degree polynomial SVM post-filter on photo test sets. R=recall, F=false positive 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

Table 4: 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

Table 5: 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