IDIAP RESEARCH REPORT

REAL-TIME FACE DETECTION USING BOOSTING LEARNING IN HIERARCHICAL FEATURE SPACES

Dong Zhang [1] Stan Z. Li [2]
Daniel Gatica-Perez [1]

IDIAP–RR 03-70

DEC. 2003

SUBMITTED FOR PUBLICATION

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[1] IDIAP, Martigny, Switzerland
[2] Microsoft Research, Beijing, China

IDIAP Research Report 03-70

REAL-TIME FACE DETECTION USING BOOSTING LEARNING IN HIERARCHICAL FEATURE SPACES

Dong Zhang Stan Z. Li Daniel Gatica-Perez

DEC. 2003

SUBMITTED FOR PUBLICATION

Abstract. Boosting-based methods have recently led to the state-of-the-art face detection systems. In these systems, weak classifiers to be boosted are based on simple, local, Haar-like features. However, it can be empirically observed that in later stages of the boosting process, the non-face examples collected by bootstrapping become very similar to the face examples, and the classification error of Haar-like feature-based weak classifiers is thus very close to 50%. As a result, the performance of a face detector cannot be further improved. This paper proposed a solution to this problem, introducing a face detection method based on boosting in hierarchical feature spaces (both local and global). We argue that global features, like those derived from Principal Component Analysis, can be advantageously used in the later stages of boosting, when local features do not provide any further benefit, without affecting computational complexity. We show, based on statistics of face and non-face examples, that weak classifiers learned in hierarchical feature spaces are better boosted. Our methodology leads to a face detection system that achieves higher performance than the current state-of-the-art system, at a comparable speed.

1 Introduction

In pattern recognition terms, face detection is a two-class (face/non-face) classification problem. As the face manifold is highly complex, due to the variations in facial appearance, lighting, expressions, and other factors [2, 12], face classifiers that achieve good performance are very complex too.

The learning-based approach constitutes the most effective one for constructing face/non-face classifiers [9, 13, 6, 8]. Recently, Viola and Jones proposed a successful application of AdaBoost to face detection [18, 17, 16]. Li *et al.* extended Viola and Jones' work for multi-view faces using an improved boosting algorithm [5]. Both systems achieved a detection rate of about 91%, and a false alarm rate of 10^{-6} for frontal faces, with real-time performance on 320×240 images.

The appealing features of the methods (speed and good performance) can be explained by two factors. First, AdaBoost learning algorithms are used for learning of highly complex face/non-face classifiers. AdaBoost methods [4] learn a sequence of easily learnable weak classifiers, and boost them into a single strong classifier via a linear combination of them. Second, the real-time feature of the systems is achieved by an ingenuous use of techniques for rapid computation of a large number of simple Haar-like features [7, 17]. Moreover, the simple-to-complex cascade [18] or detector-pyramid [5] structures further speed up the computation.

In spite of their evident advantages, existing systems have limitations to achieve higher performance because weak classifiers become too weak in later stages of the boosting process. Current approaches use bootstrapping to collect non-face examples (false alarms) to re-train the detection system (e.g. as the input of the next layer in a cascade system). However, after the power of a strong classifier has reached a certain point, the non-face examples obtained by bootstrapping a learned strong classifier are very similar to the face patterns, in any space of the Haar-like features. It can be empirically shown that the classification error of Haar-like feature-based weak classifiers approaches 50%, and therefore boosting stops being effective in practice.

To address this problem, we propose a method for boosting-based face detection in which boosted weak classifiers are learned in a hierarchy of feature spaces. The power of weak classifiers can be increased by switching between these spaces, from local to global features, to an extent that boosting learning is still beneficial. In particular, we show that Principal Component Analysis (PCA) coefficients are quite effective at discriminating between face and non-face patterns, when embedded in a boosting algorithm at its later stages, unlike local features that do not provide any benefit. Although more expensive in computational terms, global features can be used only at very late stages of a cascade system, not affecting the real-time requirement. The result is a face detection system with higher detection rate and lower false alarm rate than a state-of-the-art system (tested on the CMU frontal face dataset), with negligible degradation in processing time. Our approach is illustrated in Fig1. The rest of paper is organized as follows: For sake of completeness, section 2 describes the Adaboost method. Section 3 describes Adaboost learning in the Haar-like feature space, and motivates our work based on the limitations of current methods. Section 4 introduces our approach. R esults and discussion are presented in Section 5. Section 6 provides some concluding remarks.

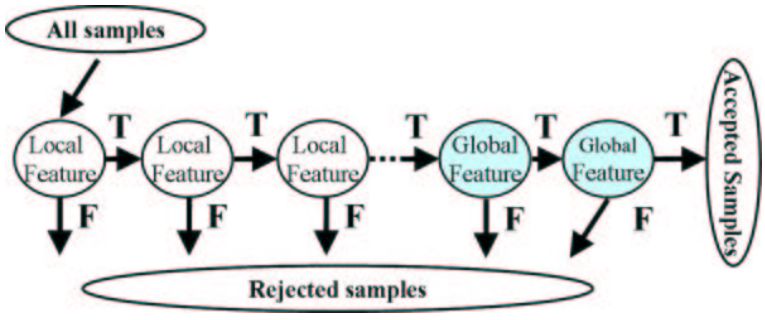

Figure 1: Face detection using both local and global features. A large majority of samples are rejected by the earlier layers in the cascade, which use simple local Haar-like features. A very small number of very difficult samples are verified by the later stages using PCA features.

2 AdaBoost Learning

The basic form of discrete AdaBoost [4] is for two class problems. A set of N labeled training examples is defined as $(x_1, y_1), \ldots, (x_N, y_N)$, where $y_i \in \{+1, -1\}$ is the class label for the example $x_i \in \mathbb{R}^n$. For face detection, x_i is an image sub-window of fixed size (*e.g.* 20×20 pixels) containing an instance of the face ($y_i = +1$) or non-face ($y_i = -1$) pattern. AdaBoost assumes that a procedure is available for learning a sequence of *weak classifiers* $h_m(x)$ ($m = 1, 2, \ldots, M$) from the training examples, with respect to the distributions $w_i^{(m)}$ of the examples. A *strong classifier* is a linear combination of the M weak classifiers,

$$H_M(x) = \sum_{m=1}^{M} \alpha_m h_m(x), \tag{1}$$

where $\alpha_m \geq 0$ are combining coefficients. The classification of x is obtained as $\hat{y}(x) = \mathrm{Sign}[\mathrm{H_M(x)} - \frac{1}{2}\sum_{\mathrm{m=1}}^{\mathrm{M}} \alpha_{\mathrm{m}}]$. The AdaBoost learning procedure is aimed to compute the sets of coefficients $\{\alpha_m\}$ and classifiers $\{h_m(x)\}$.

2.1 Learning Weak Classifiers

In the boosting framework, a strong high performance classifier is obtained by boosting simple weak classifiers. A weak classifier is constructed by thresholding one of the features according to the likelihoods (histograms) of the feature values for the positive samples and the negative samples,

$$h_k^{(M)}(x) = 1 \quad \text{if} z_k(x) > \tau_k^{(M)} \tag{2}$$

$$= 0 \quad \text{otherwise,} \tag{3}$$

where $z_k(x)$ is feature k extracted from x, and $\tau_k^{(M)}$ is the threshold for weak classifier k chosen to balance between detection rate and false alarm rate. The best weak classifier is the one for which the weighted error ϵ_M is minimized,

$$\epsilon_M = \sum_i w_i^{(M-1)} 1[\mathrm{sign}(H_M(x_i)) \neq y_i], \tag{4}$$

where $1[C]$ is one if C is true, or 0 otherwise. Defining

$$k^* = \arg\min_k \epsilon(h_{k^*}^{(M)}(x)), \tag{5}$$

the best weak classifier is given by

$$h_M(x) = h_{k^*}^{(M)}(x). \tag{6}$$

2.2 Boosting Weak Classifiers into a Strong One

AdaBoost learns to boost weak classifiers h_m into a strong one H_M by minimizing the upper bound on the classification error achieved by H_M. The bound can be derived as the following exponential loss function [10],

$$J(H_M) = \sum_i \mathrm{e}^{-y_i H_M(x_i)} = \sum_i \mathrm{e}^{-y_i \sum_{m=1}^{M} \alpha_m h_m(x)}. \tag{7}$$

AdaBoost constructs $h_m(x)$ by stagewise minimization of Eq. 7. Given the current $H_{M-1}(x) = \sum_{m=1}^{M-1} \alpha_m h_m(x)$, and the newly learned weak classifier h_M, the best combining coefficient α_M for the new strong classifier $H_M(x) = H_{M-1}(x) + \alpha_M h_M(x)$ minimizes the cost

$$\alpha_M = \arg\min_\alpha J(H_{M-1}(x) + \alpha h_M(x)). \tag{8}$$

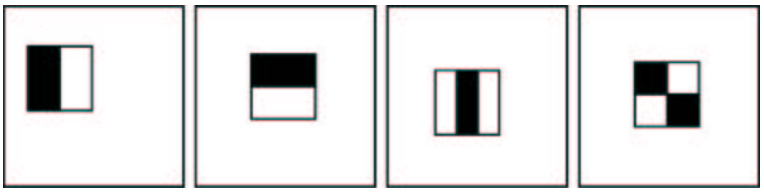

Figure 2: The four types of simple Haar-like features defined on a sub-window x. The rectangles are centered at (x, y), with size $dx \times dy$. A feature is defined as the weighted sum of the pixels in the rectangles; the weight is $+1$ for pixels in the white area and -1 for pixels in the black area.

The minimizer is

$$\alpha_M = \log \frac{1 - \epsilon_M}{\epsilon_M}, \tag{9}$$

where ϵ_M is the weighted error, defined in Eq. 4.

Each example is re-weighted after each round of boosting learning, *i.e.* $w_i^{(M-1)}$ is updated according to the classification performance of H_M,

$$\begin{aligned} w^{(M)}(x, y) &= w^{(M-1)}(x, y) \exp\left(-\alpha_M y h_M(x)\right) \\ &= \exp\left(-y H_M(x)\right), \end{aligned} \tag{10}$$

which will be used for calculating the weighted error or another cost for training the weak classifier in the next round. In this way, a more difficult example will be associated with a larger weight so that it will be more emphasized in the next round of learning.

3 AdaBoost Learning in Haar-like Feature Space

In the early stage of face detection, the weak classifiers, which perform simple classification, are derived based on histograms of four basic types of Haar-like features, shown in Fig. 2. A total of 45891 such features can be derived for a sub-window of size 20×20, for all admissible locations and sizes. Such features can be computed very efficiently [11] from the integral image defined in [18]. The task of face detection is to classify every possible sub-window. A vast number of sub-windows result from the scan of the input image. For efficiency reasons, it is crucial to discard as many non-face sub-windows as possible at the earliest stages, so that as few sub-windows as possible are further processed by later stages. Following a simple-to-complex strategy [1, 3], the detector cascade [17] quickly eliminates non-faces and greatly improves the efficiency, as opposed to applying a full detector to every sub-window.

However, the power of classification of the described system is limited when the weak classifiers derived based on simple local features become too weak to be boosted, especially in the later stages of the cascade training. Empirically, we have observed that when the discriminating power of a strong classifier reaches a certain point, *e.g.* a detection rate of 90%, and a false alarm rate of 10^{-6}, non-face examples collected by bootstrapping become very similar to those of face examples in terms of the simple local Haar-like features. The histograms of the face and non-face examples for any feature can hardly be differentiated, and the empirical probability of misclassification of the weak classifiers approaches 50%. At this stage, boosting becomes ineffective because the weak learners are too weak to be boosted. This issue has been discussed in the past by Probably Approximately Correct (PAC) learning theory [15]. A specific example of this fact is illustrated in Fig. 3, for the training set described in Section 5.

One way to address this problem is via the derivation of a stronger weak classifier in another feature space, which is more powerful and complementary with the local Haar-like feature space. We propose to boost in PCA coefficient space. As we show in the next section, weak classifiers in this global feature space have sufficient classification power for boosting learning to be effective in the later stages of a cascade system.

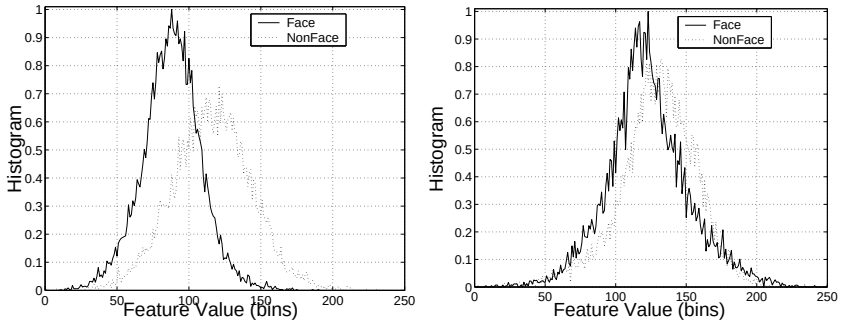

Figure 3: Left: Histogram of the face and non-face examples for the 5th haar-like feature selected by boosting learning. The error rate is significantly lower than 50%; Right: Histogram of the face and non-face examples for the 1648th haar-like feature selected by boosting learning. The error rate is close to 50%.

4 AdaBoost Learning in PCA Feature Space

When the local Haar-like features reach their limit, we would like to use another representation that is more discriminative between face and non-face examples. A fruitful alternative is to recourse to a global representation in the late stages of cascade boosting learning, such that these two feature spaces, one local and one global, complement each other.

Principal Component Analysis (PCA) is a classic technique for signal representation, used in the past for face recognition [14]. PCA can be summarized as follows. Given a set of face examples in $\mathbb{R}^N$ represented by column vectors, the mean face vector is substracted to obtain the vectors $\mathbf{x}_i \in \mathbb{R}^N$ $(i = 1, \ldots, m)$. The covariance matrix is then computed as $\mathbf{C} = \frac{1}{m}\sum_{j=1}^{m} \mathbf{x}_j \mathbf{x}_j^T$. Linear PCA diagonalizes the covariance matrix by solving the eigenvalue problem $\lambda \mathbf{v} = \mathbf{C}\mathbf{v}$, *i.e.*

$$\lambda(\mathbf{x}_i \cdot \mathbf{v}) = (\mathbf{x}_i \cdot \mathbf{C}\mathbf{v}) \quad \forall i = 1, \ldots, m, \tag{11}$$

The eigenvalues are then sorted in descending order, and the first $M \leq N$ principal components $\mathbf{v}_k$ $(1 \leq k \leq M)$ are used as the basis vector of a lower dimensional subspace, forming the transformation matrix $\mathbf{T}$ (Fig. 4). The projection of a point $\mathbf{x} \in \mathbb{R}^N$ into the M-dimensional subspace can be calculated as $\theta = (\theta_1, \ldots, \theta_M) = \mathbf{x}^\top \mathbf{T} \in \mathbb{R}^M$. Its reconstruction from η is $\hat{\mathbf{x}} = \sum_{k=1}^{M} \theta_k \mathbf{v}_k$, and constitutes the best approximation of the $\mathbf{x}_1, \ldots, \mathbf{x}_m$ in any M-dimensional subspace in the minimum squared error sense.

In Adaboost learning, each weak classifier is constructed based on the histogram of a single feature derived from PCA coefficients $(\theta_1, \ldots, \theta_M)$. At each round of boosting, one PCA coefficient -the most effective to discriminate the face and non-face classes- is selected by Adaboost. Note that the boosting algorithm select features derived from PCA based on their ability to discriminate face and non-face samples, rather than on the rank of their eigenvalues. Therefore, some PCA features corresponding to small eigenvalues may be selected in the earlier stages than those with larger eigenvalues. As we mentioned earlier, the distributions of the two classes in the Haar-like feature space almost completely overlap in the later stages of the cascade training. In that case, we propose to switch features spaces and construct weak classifiers in the PCA space. Empirically, we have found that in such space, the distributions of the face and non-face classes have smaller overlap, given the same set of non-faces obtained by bootstrapping, and used for training in later classifiers in the cascade. This situation is illustrated in Fig. 5. We can observe that the two classes are better separated, and therefore we can expect that weak classifiers based on the PCA coefficients are “boostable”. One question regarding cascade boosting in hierarchical feature spaces is at which stage of the cascade we should decide to switch from Haar-like to PCA features (we refer to such stage as the *switching stage*). It is well-known that PCA features are much more expensive in computationl terms than the Haar-like features. On one extreme. if we used PCA features in the very early stages of boosting, we would have to extract PCA features from a very large number of sub-windows, and the speed of the face detection system would be unacceptably slow. On

Figure 4: 15 out of total 400 eigenfaces ranked according to the eigenvalues (from left to right, up to down, eigenvalues in descending order).

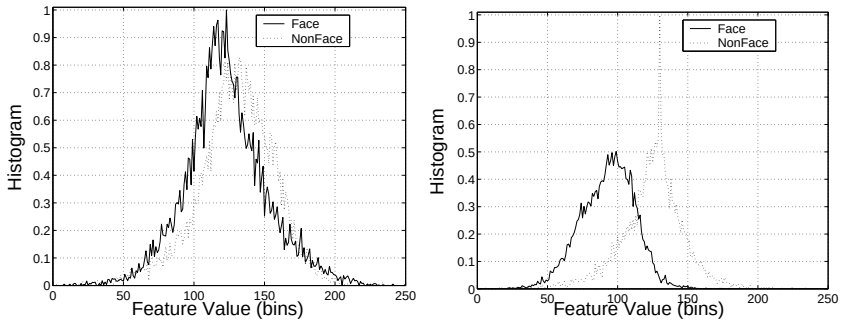

Figure 5: Left: Histogram of the face and non-face examples for the 1648th haar-like feature selected by boosting learning, whose error rate is almost 50% (same as Fig. 3(right)). Right: Histogram of the face and non-face examples for the PCA features selected at the same stage of boosting, whose error rate is significantly lower than 50%.

the other extreme, if we used PCA features in the very late stages of boosting, the performance improvement gained from their usage would be limited. Therefore, we determine the switching stage based on the tradeoff between speed and performance improvement. In experiments, we compare the performance of boosting in the single Haar-like feature space, and boosting in the hierarchical feature space, based on the comparable speed of the two systems.

5 Results

For training purposes, a total of 11,341 face examples were collected from various sources, covering out-of-plane rotation in the range $[-20°, +20°]$. All faces were manually aligned by the eyes position. For each aligned face example, five synthesized face examples were generated by a random in-plane-rotation in the range $[-20°, +20°]$, random scaling in the range $[-10\%, +10\%]$, random mirroring and random shifting to +1/-1 pixel. This created a training set of 56,705 face examples. The face examples were then cropped and re-scaled to 20×20 pixels. For non-face examples, over 10^{20} instances were collected for training from over 100,000 large images containing no faces.

In our experiments, two face detection systems were used. The first one was trained using only the Haar-like features. We refer to this system as *S-Boost* as it was only applied in a *Single* feature space. The second system was trained using both Haar-like features and PCA features. We refer to it as *H-Boost* due to the *Hierarchical* feature spaces we use. We compared the two classifiers on the complete CMU frontal face test set, which is publicly available [1]. The test set is composed of 130 images containing 510 faces, and has been also used to report results by the state-of-the-art systems [9, 17].

A face detection system can make two types of errors: a *false alarm* (FA), when the system accepts an *non-face* sample, and a *false rejection* (FR), when the system rejects a *face* sample. The performance of the face detection system is often measured in terms of these two errors, as follows:

$$\text{FAR} = \frac{\text{number of FAs}}{\text{number of non-face samples}} \times 100\%, \tag{12}$$

$$\text{FRR} = \frac{\text{number of FRs}}{\text{number of face samples}} \times 100\%. \tag{13}$$

Furthermore, the Receiver Operating Curve (ROC) is the set of operating points yielding the maximal detection rate ($100\% - \text{FRR}$) for a given false alarm rate (FAR).

Fig. 6 shows the ROC curves for both classifiers. Since changing the switching stage of H-Boost will affect both the system performance and speed, the mean and standard deviation were used to measure the performance of H-Boost, which were obtained by running the system 10 times with slightly different switching stages. We can see that H-Boost performs consistently better than S-Boost. On one hand, the detection rate of H-Boost is always higher than that of S-Boost, given the same number of false alarms. On the other hand, for a given detection rate, the false alarms of H-Boost are always fewer than those of S-Boost. The results suggest that the higher performance of H-Boost reflects the benefit of the usage of the PCA features in the late stages of boosting, which are more effective to discriminate face and non-face examples. Note that the processing speed of the mean curve for H-Boost is comparable with that of S-Boost (see details later in this section).

In order to combine FAR and FRR into one number, the *error rate* is defined as $\text{ER} = \frac{\text{FAR}+\text{FRR}}{2}$. Fig. 7 and Fig. 8 show the curves of the error rate as a function of the number of the selected weak classifiers in the switching stage from Haar-like features to PCA features. The switching stage for Fig. 7 and Fig. 8 is the 12th stage of the cascade. We can see that as the number of selected weak classifiers increased, the error rate always decreased. However, from the 265th weak classifier on, the error rate decreased only marginally for S-Boost, which indicates that any further selected weak classifiers could not discriminate face and non-face samples well. As a result, the selected weak classifiers contribute very little to the final strong classifier. On the contrary, switching from Haar-like space to PCA space decreased the error rate significantly. For H-Boost, boosting learning continued selecting weak learners in PCA space that discriminate face and non-face samples

[1] http://www-2.cs.cmu.edu/ har/faces.html

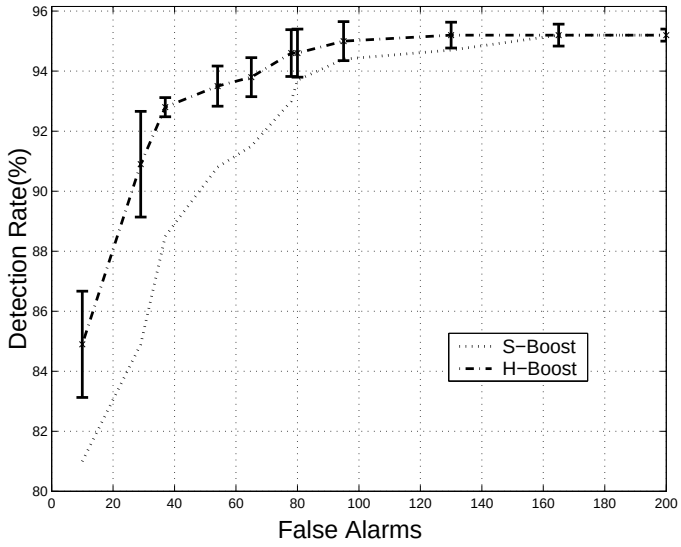

Figure 6: Comparison of ROC curves for S-Boost and H-Boost on the CMU test set. The curve of H-Boost shows the mean and standard deviation of the detection rate for a given false alarm rate. The speeds for S-Boost and the mean curve in H-Boost are comparable.

well, thus the error rate continues to decrease. We have obtained similar results on an separate testing set, composed of 2,000 face samples and 10,000 non-face samples which were not used in the training stages, as shown in Fig. 8.

Next, we compare the speed of both face detection systems. The speed of H-Boost corresponds to that of the mean curve in Fig. 6. We computed the speed of S-Boost and H-Boost using a Pentium-P4 2.6GHz, 512MB RAM computer. Since the face detector scans across the image at multiple scales and locations, the choice of the starting scale and the step size affects the detector speed significantly (refer to [17] for details). Using a starting scale of 1.2 and a step size of 1.25, both systems can process 15 frames per second for 320×240-pixel images.

There are two facts that make the computational complexity of H-Boost comparable to that of S-Boost. First, a large majority of sub-windows are rejected by the first several layers in the cascade, so only a very small number of sub-window candidates will be verified by the later stages using PCA features. Second, the

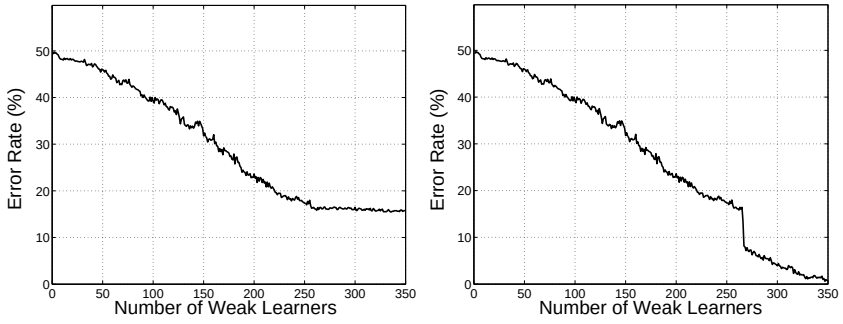

Figure 7: Left: The error rate (on training set) as a function the number of the selected weak classifiers using only Haar-like features. Right: The error rate (on training set) as a function of the number of the selected weak classifiers using both Haar-like features and PCA features.

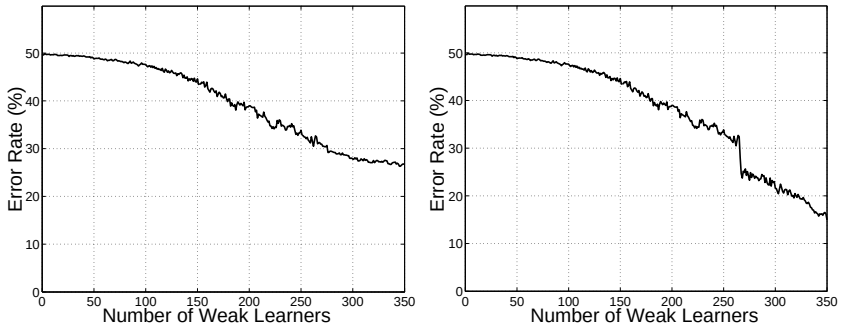

Figure 8: Left: The error rate (on testing set) as a function of the number of the selected weak classifiers using only Haar-like features. Right: The error rate (on testing set) as a function of the number of the selected weak classifiers using both Haar-like features and PCA features.

number of selected PCA features is far less than that of the Haar-like features selected at the same stage.

6 Conclusions

The paper introduced a novel boosting-based face detection algorithm in hierarchical feature spaces. Motivated by the fact that the weak learners based on the simple Haar-like features are too weak in the later stages of the cascade, we propose to boost PCA features in the later stages. The global PCA feature space complements the local Haar-like feature space. The algorithm selects the most effective features from PCA features using boosting, instead of ranking them according to their eigenvalues. The experiments on the CMU face test set showed that the proposed metholdogy can achieve better performance than a current state-of-the-art, single feature, Adaboost-based detection system, at a comparable speed.

7 Acknowlegments

This project was initiated when D. Zhang was with Microsoft Research Asia, and has partially funded by the Swiss NCCR on Interactive Multimodal Information Management (IM2).

References

[1] Y. Amit, , D. Geman, and K. Wilder. “Joint induction of shape features and tree classifiers”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:1300–1305, 1997.

[2] M. Bichsel and A. P. Pentland. “Human face recognition and the face image set’s topology”. *CVGIP: Image Understanding*, 59:254–261, 1994.

[3] F. Fleuret and D. Geman. “Coarse-to-fine face detection”. *International Journal of Computer Vision*, 20:1157–1163, 2001.

[4] Y. Freund and R. 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.

[5] S. Z. Li, L. Zhu, Z. Q. Zhang, A. Blake, H. Zhang, and H. Shum. “Statistical learning of multi-view face detection”. In *Proceedings of the European Conference on Computer Vision*, Copenhagen, Denmark, May 28 - June 2 2002.

[6] E. Osuna, R. Freund, and F. Girosi. “Training support vector machines: An application to face detection”. In *CVPR*, pages 130–136, 1997.

[7] C. P. Papageorgiou, M. Oren, and T. Poggio. “A general framework for object detection”. In *Proceedings of IEEE International Conference on Computer Vision*, pages 555–562, Bombay, India, 1998.

[8] D. Roth, M. Yang, and N. Ahuja. “A snow-based face detector”. In *Proceedings of Neural Information Processing Systems*, 2000.

[9] H. A. Rowley, S. Baluja, and T. Kanade. “Neural network-based face detection”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–28, 1998.

[10] R. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. “Boosting the margin: A new explanation for the effectiveness of voting methods”. *The Annals of Statistics*, 26(5):1651–1686, October 1998.

[11] P. Y. Simard, L. Bottou, P. Haffner, and Y. L. Cun. “Boxlets: a fast convolution algorithm for signal processing and neural networks”. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11, pages 571–577. MIT Press, 1998.

[12] P. Y. Simard, Y. A. L. Cun, J. S. Denker, and B. Victorri. “Transformation invariance in pattern recognition - tangent distance and tangent propagation”. In G. B. Orr and K.-R. Muller, editors, *Neural Networks: Tricks of the Trade*. Springer, 1998.

[13] K.-K. Sung and T. Poggio. “Example-based learning for view-based human face detection”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

[14] M. A. Turk and A. P. Pentland. “Eigenfaces for recognition”. *Journal of Cognitive Neuroscience*, 3(1):71–86, March 1991.

[15] L. Valiant. “A theory of the learnable”. *Communications of ACM*, 27(11):1134–1142, 1984.

[16] P. Viola and M. Jones. “Asymmetric AdaBoost and a detector cascade”. In *Proceedings of Neural Information Processing Systems*, Vancouver, Canada, December 2001.

[17] P. Viola and M. Jones. “Rapid object detection using a boosted cascade of simple features”. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii, December 12-14 2001.

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