FACE DETECTION FROM FEW TRAINING EXAMPLES

Chunhua Shen[*]*, Sakrapee Paisitkriangkrai*[⋆]*, Jian Zhang*[⋆]

[*]NICTA, Canberra Research Laboratory, Canberra, ACT 2612, Australia
[⋆]NICTA, Neville Roach Laboratory, Sydney, NSW 2052, Australia

ABSTRACT

Face detection in images is very important for many multimedia applications. Haar-like wavelet features have become dominant in face detection because of their tremendous success since Viola and Jones [1] proposed their AdaBoost based detection system. While Haar features' simplicity makes rapid computation possible, its discriminative power is limited. As a consequence, a large training dataset is required to train a classifier. This may hamper its application in scenarios that a large labeled dataset is difficult to obtain. In this work, we address the problem of learning to detect faces from a small set of training examples. In particular, we propose to use covariance features. Also for better classification performance, linear hyperplane classifier based on Fisher discriminant analysis (FDA) is proffered. Compared with the decision stump, FDA is more discriminative and therefore fewer weak learners are needed. We show that the detection rate can be significantly improved with covariance features on a small dataset (a few hundred positive examples), compared to Haar features used in current most face detection systems.

Index Terms— Face detection, AdaBoost, object recognition

1. INTRODUCTION

Face detection (or more generic object detection) plays a critically important role in many computer vision applications such as intelligent video surveillance, vision based teleconference systems and content based image retrieval. It is challenging because of the variations of visual appearances, poses and illumination conditions *etc*. Since Viola and Jones proposed the real-time AdaBoost based face detector [1], a lot of incremental work has been conducted. Most of them have focused on improving the boosting method or accelerating the training process. For example, [2] proposed an improved FloatBoost method for better detection accuracy by introducing a backward feature selection step into the AdaBoost training procedure. [3] used forward feature selection for fast training. [4] significantly reduced decision stump's training time by approximation.

Comparatively less work has been done for finding better discriminative features. While Haar features' simplicity and the idea of *integral image* make rapid computation possible, Haar feature's discriminative power is limited. As a consequence, a large training dataset is usually required to train a classifier for reasonable detection accuracy. In many cases it is tedious to manually label thousands of images. With the size of the training data increasing, the computation complexity also goes up quickly. For a supervised learning algorithm, it consists two components: the feature and the classification algorithm. We try to address both issues in the context of face detection. In this work, we show that on small training datasets, using better features (here in particular, covariance features [5]) and more flexible weak classifiers, much better detection results can be achieved compared with Haar features with simple decision stump.

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Center of Excellence program.

Fig. 1: The first and second covariance region selected by AdaBoost. The figure displays the first two covariance features overlayed on human faces. The first covariance feature captures the information of the right eye while the second covariance feature describes the contour of the face, and the region around the nose.

Related work. Literature on face detection is abundant. We list a few that is relevant to our work here. Yang *et al*. [6] have provided a comprehensive survey of the field. Many face detection systems have been advocated. In [7], a neural network was trained for fast face detection. Romdhani *et al*. [8] proposed a cascaded support vector machine (SVM) structure to reduce the detection time. The work by Viola and Jones [1] is a break-through. They used very simple Haar features. Together with the idea of integral image, these features can be computed in constant time. AdaBoost was then used to select features and at the same time to build a strong classifier. Considering the problem's extreme imbalance nature, a cascade was proposed in order to avoid the flood of non-faces. All these systems have used several thousand face images and even more negative samples for training. Recently other features such as histogram of orientations (HOG) [9, 10] and covariance feature [11] have been proposed for pedestrian detection and better performance is achieved than Haar features [12]. In this work, we show that covariance feature is also powerful for face detection.

2. ALGORITHM

Our detection framework follows Viola and Jone's classical work [1]. The differences are: (1) we use covariance feature; (2) Since covariance feature is multidimensional, simple decision stump is no longer applicable. We adopt the weighted FDA as weak classifiers.

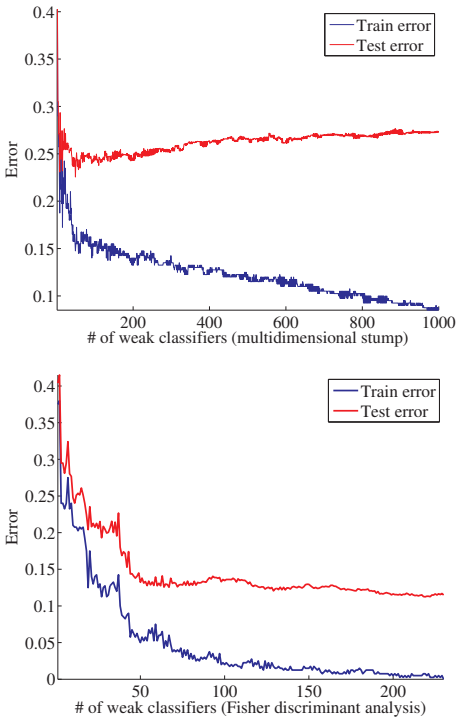

Fig. 2: AdaBoost learning with decision stump (top) and FDA (bottom) as weak classifiers on the *banana dataset* used in [13]. It is clearly shown that boosting FDA is much better than boosting multidimensional stump in both training and testing performance.

Covariance features. Tuzel *et al*. [5] have proposed region covariance in the context of object detection. Instead of using joint histograms of the image statistics (b^d dimensions where d is the number of image statistics and b is the number of histogram bins used for each image statistics), covariance is computed from several image statistics inside a region of interest. This results in a much smaller dimensionality. The correlation coefficient of two random variables X and Y is given by

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)\mathrm{var}(Y)} = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y} \quad (1)$$

$$\begin{aligned}\mathrm{cov}(X,Y) &= \mathbf{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \\ &= \frac{1}{n-1}\sum_k (X_k-\mu_X)(Y_k-\mu_Y), \end{aligned} \quad (2)$$

where $\mathrm{cov}(\cdot,\cdot)$ is the covariance of two random variables; μ is the sample mean and σ is the sample variance. Correlation coefficient is commonly used to describe the information we gain about one random variable by observing another random variable.

In this work, the 7D image statistics used in this experiment are pixel location x, pixel location y, image intensity $\mathbf{I}(x,y)$, first order partial derivative of the intensity in horizontal and vertical direction, $|\mathbf{I}_x|$ and $|\mathbf{I}_y|$, second order partial derivative of the intensity in horizontal and vertical direction $|\mathbf{I}_{xx}|$ and $|\mathbf{I}_{yy}|$. The covariance descriptor of a region is a 7×7 matrix. Due to the symmetry, only upper triangular part is stacked as a vector and used as covariance descriptors. A vector of covariance descriptors is projected onto a 1D space using weighted FDA algorithm. AdaBoost [14] is then applied to select the best rectangular region w.r.t. the weak learner that best classifies training samples with minimal classification error. The best weak learner is added to a cascade. Weak learners are added until the predefined classification accuracy is met. The descriptors encode information of the correlations of the defined features inside the region. The experimental results show that the covariance region selected by AdaBoost are physically meaningful and can be easily interpreted as shown in Figure 1. The first selected feature focuses on the human right eye while the second selected feature focuses on the contour of the face and nose. This finding is consistent with other researchers' results. For example, in [15] it was found that the SIFT [16] word being most probable for faces is the region around eye.

Note that the technique is different from [5], where the covariance matrix is directly used as the feature and the distance between features is calculated on the Riemannian manifold[1]. However, eigendecomposition is involved for calculating the distance on the Riemannian manifold. We instead vectorize the symmetric matrix and measure the distance in the Euclidean space, which is much faster.

Weighted Fisher discriminant analysis as weak learners. It is well known that the weak classifier plays an important roll for an ensemble learning algorithm such as Boosting and Bagging [17]. Decision stump is the one of the simplest classifiers. It selects the most discriminative dimension and discard all the other dimensions' information. In other words, it projects the original multidimensional data onto one of its axis. This treatment may drop a lot useful information for classification. Levi and Weiss [10] have adopted multidimensional stumps to train a boosted detector. In [18], linear SVMs are used as weak classifiers. The drawback is its heavy training complexity. We instead adopt FDA as weak learners. FDA projects the training data onto the direction which most separates the two classes by maximizing the Fisher score. In this way more information is exploited than multidimensional decision stump. Moreover, it has close-form solution, therefore it is much more faster in training than SVMs. After projection, the offset of the linear classifier is obtained by exhausted search as in [1].

To show the better classification capability, we have trained a boosted classifier on an artificial 2D dataset with the multidimensional decision stump and FDA as weak classifiers respectively. Figure 2 shows the results. AdaBoost with the multidimensional stump's training error is still around 0.1 after 1000 rounds training. In contrast, The training error of AdaBoost with FDA drops to zero after 230 rounds. More importantly, the testing error of AdaBoost with FDA is much lower than using multidimensional stumps. After 300 rounds' training, the testing error of multidimensional stumps increases slightly, which is an indicator of overfitting.

3. EXPERIMENTS

The experimental section is organized as follows. First, the dataset used in this experiment, including how the performance is analyzed, are described. Experiments and the parameters used to achieve optimal results are then discussed. Finally, experimental results and analysis of different techniques are compared.

MIT + CMU frontal face test set. We tested our face detectors on the low resolution faces dataset, MIT + CMU frontal face test set. The complete set contains 130 images with 507 frontal faces. Detections are considered true or false positives based on the area of overlap with ground truth bounding boxes. To be considered a correct detection, the area of overlap between the predicted bounding box and ground truth bounding box must exceed 50%. Multiple detections of the same face in an image are considered false detections.

[1]Covariance matrices are symmetric and positive semi-definite, hence they reside in the Riemannian manifold.

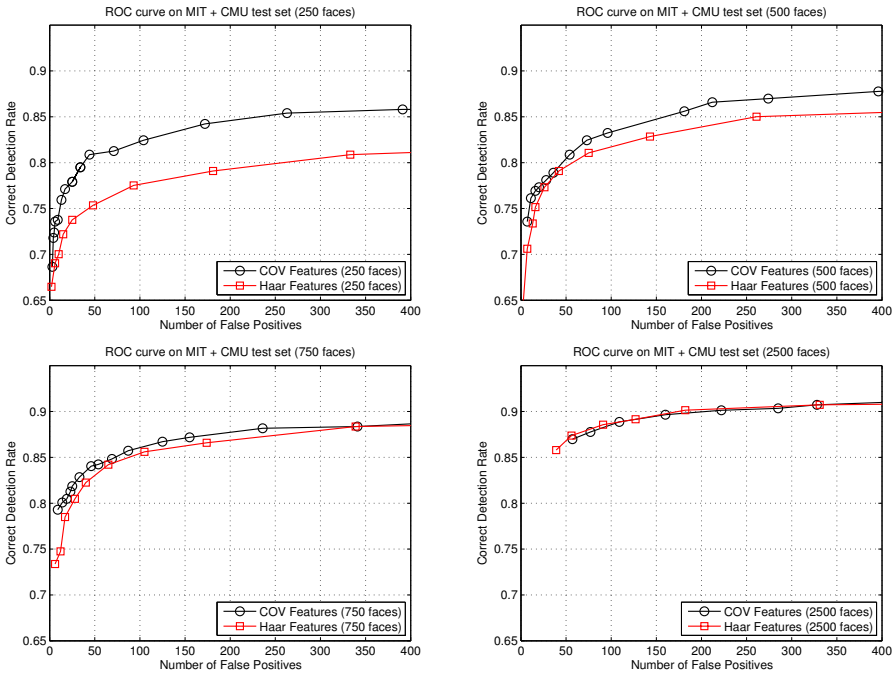

Fig. 3: ROC curves for our algorithm on the MIT + CMU test set [7]. The detector was run using a step size of 1 pixel and a scale factor of 1.2.

Experiment parameters. Our training sets contain 250, 500, 750 and 2,500 frontal human faces obtained from the internet. The faces are scaled and aligned to a base resolution of 24×24 pixels. We used approximately 8,500 non-face training images to bootstrap the negative samples. In this experiment, we use 7,000 covariance filters sampled uniformly from the entire set of rectangle filters. Each filter consists of four parameters, namely, x-coordinate, y-coordinate, width and height. A strong weak classifier consisting of several weak classifiers is built in each stage of the cascade. In each stage, weak classifiers are added until the learning goal is met. In this experiment, we set the minimum detection rate in each stage to be 99.5% and the maximum false positive rate to be 50%. The non-face samples used in each stage of the cascade are collected from false positives of the previous stages of the cascade (bootstrapping). The cascade training algorithm terminates when there are not enough negative samples to bootstrap. In this experiment, we set the scaling factor to 1.2 and window shifting step to 1. The technique used for merging overlapping windows is similar to [1]. For training and testing Haar-like wavelet features, we use the fast AdaBoost implementation proposed by Wu *et al*. [3].

Results. Figure 3 shows a comparison between the ROC curves produced by our covariance features and the ROC curves from Haar-like wavelet features [1], which serves as a baseline. We construct the ROC curve by repeatedly adding one node to the cascades at a time. The curve between two consecutive points is approximated by a line segment. The ROC curves show that covariance features significantly outperform Haar-like wavelet features when the training database size is small (*e.g*., less than 500 faces). However, when the database size is large, the performance of both features are very similar. These results indicate that the type of features plays a crucial role in face detection performance, especially when the number of training samples is small. As the number of samples grows, the performance difference between the two techniques decreases.

We believe that Haar features are not discriminant enough to separate the two classes (face and non-face). As a result, it does not generalize well when the database size is small. Covariance features, on the other hand, are much more discriminant and extremely powerful in separating a patch of faces from non-faces. Figure 4 shows the AdaBoost with FDA's accumulated rejection rate over different cascade levels. It can be seen that the first 3 levels of cascade can reject more than 90% of the non-face samples already. Also it can be observed that for different sizes of training data, the accumulated rejection rate over the first 6 cascade levels is very similar. Based on this observation, we may design more efficient and accurate detectors using multi-layer boosting with heterogeneous features. We leave this topic for future research.

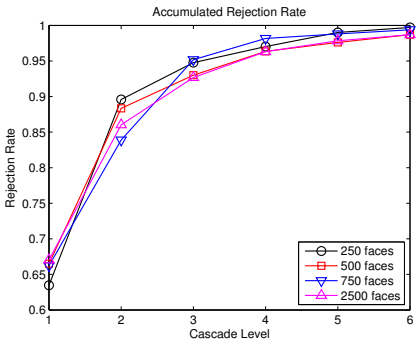

Fig. 4: The accumulated rejection rate over the first 6 cascade levels using the covariance feature.

Based upon our observations, using small training size does not only ease a process of face labeling, which is a rather tedious and time-consuming process, but also results in a smaller and simpler classifier (Figure 6). The classifier trained using 250 faces contains only 201 covariance features in total while the classifier trained using 2500 faces contains three times as many weak classifiers as the one with 250 faces. Training with small datasets reduces the training time. Hence the second advantage is that the resulted simpler final

Fig. 5: Some detection results of our face detectors trained using only 250 faces on MIT + CMU test images. Note that there are very few false positives and negatives.

classifier reduces the detection time. Also simpler classifiers are less likely to be overfit.

Our findings in this work are consistent with the experimental results in [10], which used the HOG feature and the multidimensional decision stump as weak classifiers.

In Figure 5, we show some detection results of our face detectors trained using 250 faces on MIT + CMU frontal face test sets. Note that there are very few false positives and negatives on a detector trained with such few training examples.

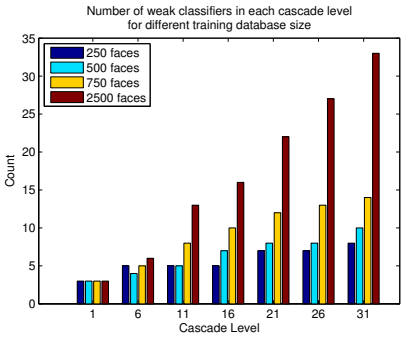

Fig. 6: The number of weak classifiers in different cascade levels.

4. CONCLUSION

In this work, we have proposed a new approach for face detection. Current object detectors heavily depend on large scale training data to model the variations of the target object. Detectors that rely on small labeled training data are thus needed. We show in this work that discriminative features are critically important for the success of such detecting systems. In particular we have shown that covariance features plus boosted FDA significantly improve the capability of the detector to learn from a small number of labeled examples.

In the future we will research on designing more efficient and accurate detection systems using heterogeneous features, based upon the findings from this work.

References

[1] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comp. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[2] S. Z. Li and Z. Zhang, "FloatBoost learning and statistical face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1112–1123, 2004.

[3] J. Wu, S. C. Brubaker, M. D. Mullin, and J. M. Rehg, "Fast asymmetric learning for cascade face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 369–382, 2007.

[4] M.-T. Pham and T.-J. Cham, "Fast training and selection of Haar features using statistics in boosting-based face detection," in *Proc. IEEE Int. Conf. Comp. Vis.*, Rio de Janeiro, Brazil, 2007.

[5] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Proc. Eur. Conf. Comp. Vis.*, Graz, Austria, May 2006, vol. 2, pp. 589–600.

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

[7] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 23–38, 1998.

[8] S. Romdhani, P. Torr, B. Schölkopf, and A. Blake, "Computationally efficient face detection," in *Proc. IEEE Int. Conf. Comp. Vis.*, Vancouver, 2001, vol. 2, pp. 695–700.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, San Diego, CA, 2005, vol. 1, pp. 886–893.

[10] K. Levi and Y. Weiss, "Learning object detection from a small number of examples: The importance of good features," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, Washington, DC, 2004, vol. 2, pp. 53–60.

[11] O. Tuzel, F. Porikli, and P. Meer, "Human detection via classification on Riemannian manifolds," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, Minneapolis, MN, 2007.

[12] S. Paisitkriangkrai, C. Shen, and J. Zhang, "An experimental evaluation of local features for pedestrian classification," in *Proc. Int. Conf. Digital Image Computing - Techniques and Applications*, Adelaide, Australia, 2007.

[13] G. Rätsch, T. Onoda, and K.-R. Müller, "Soft margins for AdaBoost," *Mach. Learn.*, vol. 42, no. 3, pp. 287–320, 2001.

[14] R. E. Schapire, "Theoretical views of boosting and applications," in *Proc. Int. Conf. Algorithmic Learn. Theory*, London, UK, 1999, pp. 13–25, Springer-Verlag.

[15] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images," in *Proc. IEEE Int. Conf. Comp. Vis.*, Beijing, China, 2005, vol. 1, pp. 370–377.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comp. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[17] R. Meir and G. Rätsch, *An introduction to boosting and leveraging*, pp. 118–183, Advanced lectures on machine learning. Springer-Verlag, New York, NY, USA, 2003.

[18] Q. Zhu, S. Avidan, M. Yeh, and K.-T. Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, New York, 2006, vol. 2, pp. 1491–1498.