Enhancing Training Set for Face Detection

Ruiping Wang[1, 2], Jie Chen[3], Shiguang Shan[1], Xilin Chen[1], Wen Gao[1, 2, 3]

[1] *ICT-ISVISION JDL for Face Recognition, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China*
[2] *Graduate School of the Chinese Academy of Sciences, Beijing, 100039, China*
[3] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China*
{rpwang, jchen, sgshan, xlchen, wgao}@jdl.ac.cn

Abstract

We present a novel method to enhance training set for face detection with nonlinearly generated examples from the original data. The motivation is from Support Vector Machines (SVM) that, for classification problems, examples lying close to class boundary usually have more influence and thus are more informative than those far from the boundary. We utilize a nonlinear technique — reduced set (RS) method and a new image distance metric to generate new examples, and then add them to the original collected database to enhance it. Extensive experiments show that the proposed approach has an encouraging performance.

1. Introduction

Face detection is a widely studied problem over the past decade [11]. Recently, the emphasis has been laid on data-driven learning-based techniques, such as [2, 4, 5, 7, 9]. The performance of these learning-based methods depends highly on the training set. Therefore, it is useful to explore methods to learn with the specific database in a robust way and train a face detector with good generalization property.

SVM tells us that only those boundary examples, called support vectors (SVs), are useful for the final decision [8]. Also, another popular algorithm AdaBoost pays more attention to the difficult training samples, which lie close to the decision boundary and thus are more likely to be misclassified, by assigning them larger weights. These give strong cues that boundary samples play a more important role than others in classification problems.

In this article, we adopt a nonlinear technique based on the RS method to generate new examples, which lie close to the face/non-face class boundary [6]. Then we add these examples to the original database. Furthermore, to improve the approximation performance of the RS method, we embed a new distance metric for images called IMED, which takes into account the spatial relationships of pixels [10], into the kernel used by the RS method. Finally, using the enhanced training set, we train an AdaBoost-based face detector with improved generalization performance.

The rest of this paper is organized as follows: After a review of the general theory of the RS method and IMED, the proposed dataset enhancing approach is described in section 2. Experiments are presented in section 3 and conclusion suggested in section 4.

2. RS variant based on IMED

In this section, after briefly describing the RS method and IMED metric, we propose our IMED based RS variant method for dataset enhancing.

The system overview is given in Fig.1. Original samples are used to generate reduced set vectors (RSVs) by the RS variant based on IMED for positive and negative classes respectively. Adding these RSVs to the original dataset, we get the enhanced training set for training final classifier.

2.1. Reduced set method

The RS method is one of the solutions [1, 3, 6] proposed to improve the run-time performance of SVM. SVM implicitly maps the data $(\boldsymbol{x}_1, y_1),...,(\boldsymbol{x}_l, y_l) \in \mathbf{X} \times \{\pm 1\}$ into a dot product feature space $\boldsymbol{F}$ via a (usually nonlinear) map $\boldsymbol{\Phi}: \mathbf{X} \to \boldsymbol{F}, \boldsymbol{x} \mapsto \boldsymbol{\Phi}(\boldsymbol{x})$, and computes a hyperplane separating the data in $\boldsymbol{F}$ by a large margin. A class of kernels $k(\boldsymbol{x}, \boldsymbol{x}')$ can be shown to compute the dot products in associated feature spaces, i.e., $k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{\Phi}(\boldsymbol{x}) \cdot \boldsymbol{\Phi}(\boldsymbol{x}'))$. The solution to this training problem is as follows:

$$f(\boldsymbol{x}) = \operatorname{sgn}(\sum_{i=1}^{l} \alpha_i y_i k(\boldsymbol{x}, \boldsymbol{x}_i) + b), \tag{1}$$

where $0 \le \alpha_i \le C, i = 1,...,l, \sum_{i=1}^{l} \alpha_i y_i = 0$ (the positive parameter C determines the trade-off between margin maximization and training error minimization). Those training examples $\boldsymbol{x}_i$ with $\alpha_i > 0$ are called support vectors. In this paper, we employ the Gaussian kernel:

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|^2 / 2\sigma^2). \tag{2}$$

Given a vector $\boldsymbol{\Psi} \in \boldsymbol{F}$, it is expanded in images of input patterns $\boldsymbol{x}_i \in \mathbf{X}$,

$$\boldsymbol{\Psi} = \sum_{i=1}^{N_x} \alpha_i \boldsymbol{\Phi}(\boldsymbol{x}_i). \tag{3}$$

with $\alpha_i \in \mathbb{R}$. To reduce the evaluating complexity, the RS method seeks to approximate it by an expansion $\boldsymbol{\Psi}' = \sum_{i=1}^{N_z} \beta_i \boldsymbol{\Phi}(\boldsymbol{z}_i)$, with $N_z \ll N_x, \beta_i \in \mathbb{R}$, and *reduced*

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

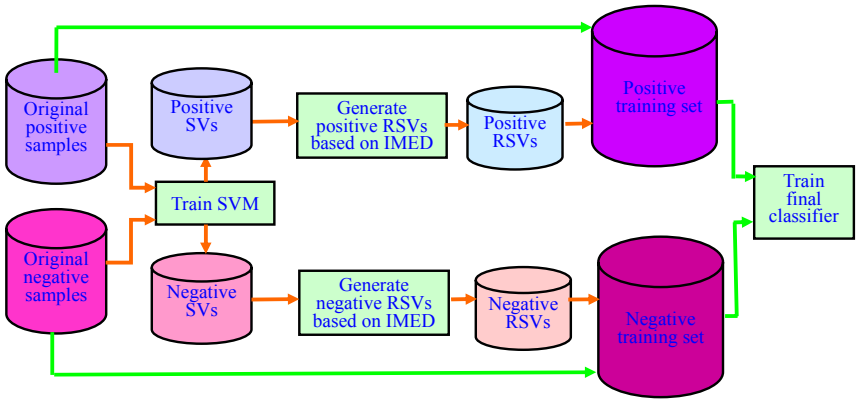


Figure 1. The flow chart of the proposed method.

set vectors (RSVs) $z_i \in \mathbf{X}$. To this end, one can minimize $\|\Psi - \Psi'\|^2$, which can be simply computed in terms of the kernel. An iterative approach to compute vectors z_i and coefficients β_i has been developed in [6]. Finally, for any N_z, the obtained expansion can be plugged into the SVM decision function to yield $f(\boldsymbol{x}) = \mathrm{sgn}(\sum_{j=1}^{N_z} \beta_j k(\boldsymbol{x}, \boldsymbol{z}_j) + b)$ to approximate (1).

2.2. The image Euclidean distance

The image Euclidean distance (IMED) has the following properties: relative insensitivity to small perturbation, simplicity of computation, and efficiency to be embedded in most powerful image recognition techniques [10].

Given two images $\boldsymbol{x}, \boldsymbol{y}$ with the size $M \times N$, the IMED between them $d_{IMED}(\boldsymbol{x}, \boldsymbol{y})$ is written by:

$$d^2_{IMED}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^T G (\boldsymbol{x} - \boldsymbol{y}), \tag{4}$$

where $G = (g_{ij})_{MN \times MN}$ is an MN^{th} order symmetric and positive definite matrix, and each element g_{ij} is:

$$g_{ij} = f(|P_i - P_j|) = \frac{1}{2\pi\sigma^2} \exp\left(\frac{-|P_i - P_j|^2}{2\sigma^2}\right), \tag{5}$$

where σ is the width parameter (in our case set to 1), P_i, P_j, $i, j = 1, 2, \ldots, MN$ are pixels and $|P_i - P_j|$ is their distance on the image lattice. An algorithm to compute IMED rapidly is also discussed in [10].

2.3. The improved RS variant

The RS method was originally proposed to speed up SVM in test phase. It computes an expansion $\boldsymbol{w}' = \sum_{j=1}^{N_z} \beta_j \Phi(\boldsymbol{z}_j)$ with a set of vectors $\boldsymbol{z}_j$ and coefficients β_j to approximate the weight vector $\boldsymbol{w} = \sum_{i=1}^{N_x} \alpha_i y_i \Phi(\boldsymbol{x}_i)$ $(\alpha_i > 0, y_i = \pm 1)$ in the decision function of SVM (N_x is the number of SVs). Since the vectors $\boldsymbol{x}_i$ and $\boldsymbol{z}_j$ belong to the same space, it implies that, the class label of $\boldsymbol{z}_j$ is determined by the sign of its corresponding coefficient β_j [6]. Namely, those vectors with positive coefficients are interpreted as the positive examples, otherwise negative ones.

Next we decompose $\boldsymbol{w}$ into two parts to group positive SVs (faces) and negative SVs (non-faces) separately as:

$$\boldsymbol{w} = \boldsymbol{w}^p + \boldsymbol{w}^n, \tag{6}$$

where

$$\boldsymbol{w}^p = \sum_{k=1}^{N_x^p} \alpha_k^p \Phi(\boldsymbol{x}_k^p), \tag{7}$$

$$\boldsymbol{w}^n = \sum_{l=1}^{N_x^n} \alpha_l^n \Phi(\boldsymbol{x}_l^n), \tag{8}$$

$$N_x = N_x^p + N_x^n, \tag{9}$$

and $\alpha_k^p > 0$ corresponding to SVs $\boldsymbol{x}_k^p$ of positive class while $\alpha_l^n < 0$ to $\boldsymbol{x}_l^n$ of negative class. N_x^p and N_x^n are the numbers of positive and negative SVs respectively.

In our RS variant version, we will approximate $\boldsymbol{w}^p$ and $\boldsymbol{w}^n$ separately rather than treating $\boldsymbol{w}$ as a whole as in the original RS method, which tries to perform vector clustering in both classes simultaneously in some sense. Since our new goal is to generate more discriminative examples to enhance the original dataset, especially the positive face set, we conduct the similar clustering as [6] for both classes individually, namely, we compute

$$\boldsymbol{w}^{p\prime} = \sum_{m=1}^{N_z^p} \beta_m \Phi(\boldsymbol{z}_m) \tag{10}$$

$$\boldsymbol{w}^{n\prime} = \sum_{n=1}^{N_z^n} \beta_n \Phi(\boldsymbol{z}_n) \tag{11}$$

to approximate $\boldsymbol{w}^p$ and $\boldsymbol{w}^n$ respectively. As the sign of coefficient β_m (β_n) determines the class of $\boldsymbol{z}_m$ ($\boldsymbol{z}_n$), in our case, we found empirically that, during the process of sequential approximation, all β_m in $\boldsymbol{w}^{p\prime}$ are positive whilst all β_n in $\boldsymbol{w}^{n\prime}$ are negative. This means, when conducting vector clustering in these two classes separately, we obtain a sequence of reduced set vectors all with the corresponding class label for each class. This result also

seems reasonable in some sense. Fig.2 provides a sketch of the framework of our dataset enhancing algorithm.

Given: Original training set TS_O, with TS_O^p / TS_O^n is the $+/-$ samples set respectively, $TS_O = TS_O^p \cup TS_O^n$

1. Compute IMED metric matrix G.
2. Train SVM using TS_O to get the weight vector $\boldsymbol{w}$ and its $+/-$ part $\boldsymbol{w}^p / \boldsymbol{w}^n$.
3. Perform the IMED embedded RS method to $\boldsymbol{w}^p / \boldsymbol{w}^n$, and compute the $+/-$ RSV set RS^p / RS^n.
4. Add RS^p / RS^n to TS_O^p / TS_O^n respectively to obtain TS_E^p / TS_E^n, and the enhanced set $TS_E = TS_E^p \cup TS_E^n$.
5. Use TS_E to train the final face detector.

Figure 2. Our approach for dataset enhancing (we use $+/-$ to represent positive/negative respectively).

Another issue in the RS method is the approximation performance. It suggests, during iteration procedure, numerical instabilities can be approached by restarting the iteration with different starting values [6]. However, this benefits little to the approximation property. Within this paper, we will embed the IMED into the Gaussian kernel (2) used in the RS method for the generation of RSVs. By replacing $\|\boldsymbol{x}-\boldsymbol{y}\|^2$ by $(\boldsymbol{x}-\boldsymbol{y})^T G(\boldsymbol{x}-\boldsymbol{y})$, we obtain the final kernel function: $k_{IMED}(x,y) = \exp(\frac{-(\boldsymbol{x}-\boldsymbol{y})^T G(\boldsymbol{x}-\boldsymbol{y})}{2\sigma^2})$, and it improves the approximation performance greatly.

Naturally, one might consider some other metrics, such as the tangent distance and the generalized Hausdorff distance. However, both of them require complicated computation and are difficult to be combined with other recognition techniques. Furthermore, for the noiseless images, IMED can still promote the resulting system's performance [10]. All these advantages make IMED a better choice than other metrics for our problem.

3. Experiments

Several experiments are conducted to evaluate our method on two datasets, the MIT database of a small size and a comparatively larger one collected by ourselves.

3.1. Experiments on the MIT database

This dataset consists of a training set of 6,977 images (2,429 faces and 4,548 non-faces) and a test set of 24,045 images (472 faces and 23,573 non-faces) [12]. All samples are resized to 20×20 grayscale.

We then apply the first 2 steps in Fig.2 to the original training set. Each sample is represented by a 400-dimensional intensity feature vector. The SVM selects 1,148 SVs including 475 faces and 673 non-faces.

In order to compare the approximation property of the original RS (*called ED-RSV*) with the IMED embedded RS method (*IMED-RSV*), which adopts $k_{IMED}(\boldsymbol{x},\boldsymbol{y})$ as the kernel, we take the weight vector $\boldsymbol{w}$ as a whole to be approximated as in [3, 6]. Here, note that for both methods approximation errors are computed in terms of the same kernel of Eq.(2). The RSVs generating process is stopped when the approximation error difference between the neighboring two generated vectors, say $|\varepsilon_{i+1}-\varepsilon_i|$, falls below a specified threshold. In our case, we set the condition as $|\varepsilon_{i+1}-\varepsilon_i|/\varepsilon_i < 0.001$.

By this means, we obtain two sets of RSVs, 300 samples for each set. Herein, the *ED-RSV* set consists of 100 faces and 200 non-faces, and the *IMED-RSV* set includes 102 faces and 198 non-faces. Comparison result is given in Fig.3(a), which indicates that, by embedding IMED in the RS method, one can get consistent improvement of approximation performance. Thus, the *IMED-RSV* set includes more discriminating information for the classification problem, and they are able to achieve better generalization property.

To evaluate these two sets, we add them to the original training set respectively to get two enhanced sets. Note that the two enhanced sets have very close numbers of examples for each class. We then use the original training set and the two enhanced sets to train three AdaBoost-based [9] face detectors, which are then evaluated on the test set of MIT database. In Fig.3(b), we illustrate the ROC curves for comparison. From these curves one can conclude that: first, RSVs, as new generated samples, have much enriched the original dataset and hence improved the detector performance; second, by embedding IMED during the calculating of RSVs, we can further promote the ability to increase discrimination of the generated RSVs distinctly.

Table 1. Some reduced set vectors by *IMED-RSV*.

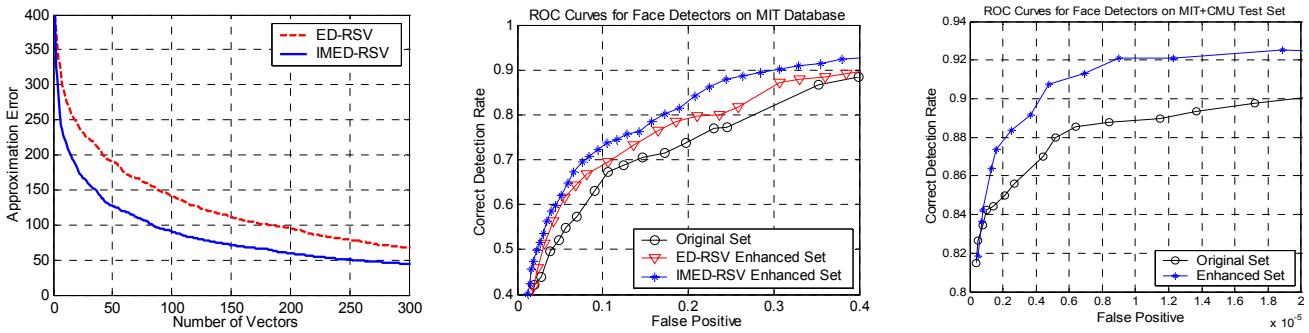
Faces					
Non-faces					

Tab.1 shows some reduced set vectors generated by *IMED-RSV*. We can notice that these new samples do really resemble the true pattern of their class label.

3.2. Experiments on the large database

To verify the effectiveness of our approach further, we carry out experiments on a larger dataset with 15,000 frontal face and 25,000 non-face samples, each normalized to 20×20 grayscale. It will be computationally expensive to train a single SVM on such a large set. For simplicity, we split it into 5 subsets of equal size.

As demonstrated in Fig.2, we perform the first 3 steps for each subset (the step 1 needs to be performed only once). Grouping RSVs from all subsets, we get a total of 910 faces and 1,535 non-faces, which are then added to the original set to obtain the final enhanced set. Both the original and the enhanced set are then used to train the AdaBoost-based face detectors. To arrive at the better performance, we adopt bootstrap [7] to increase non-face



(a) (b) (c)

Figure 3. (a) Approximation error as a function of the number of generated reduced set vectors. (b) The ROC curves on the MIT test set. (c) The ROC curves on the MIT+CMU test set.

Figure 4. Some results of our trained detector.

training samples for both detectors. The bootstrap is carried out several times on a set of 10,964 images containing no faces, and gives the two detectors equal opportunities.

The resulting two detectors are evaluated on the MIT+CMU frontal face test set, which consists of 130 images showing 507 upright faces [4]. The detection performances are compared in Fig.3(c). It shows that the detector based on the enhanced set outperforms the one based on the original set significantly. We also attribute these results to the reduced set vectors enriching the original dataset and thus improving the generalization performance of the final face detector. Some results of our trained detector based on the enhanced set are demonstrated in Fig.4.

4. Conclusion and future work

We propose a novel technique based on the RS method to enhance the training set for face detection. We introduce a new distance metric IMED for images, and embed it in the kernel function for the generation of RSVs to improve the approximation property of the RS method. Test results demonstrate that the detector trained by the proposed enhanced set achieves better performance than the one by the original dataset. There are several avenues for the future work: (a) It may be possible to find other more efficient metrics instead of IMED for the RS method; (b) The effect of changing the number of RSVs needs to be investigated further; (c) We will extend the proposed algorithm to the multi-class problems.

5. Acknowledgements

This research is partially sponsored by Natural Science Foundation of China under contract No.60332010 and "100 Talents Program" of CAS. The first author also would like to gratefully acknowledge support from Natural Science Foundation of China under contract No.60473043.

6. References

[1] C. J. C. Burges. Simplified support vector decision rules. In 13th Intl. Conf. on Machine Learning, pages 71–77, 1996.
[2] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. CVPR, 1997.
[3] S. Romdhani, P. Torr, B. Scholkopf, and A. Blake. Computationally efficient face detection. ICCV, 2001.
[4] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. PAMI, 20:23–38, 1998.
[5] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to face and cars. CVPR, 2000.
[6] B. Scholkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Muller, G. Ratsch, and A. Smola. Input space vs. feature space in kernel-based methods. IEEE Transactions on Neural Networks, 10(5):1000 - 1017, 1999.
[7] K. K. Sung, and T. Poggio. Example-Based Learning for View-Based Human Face Detection. PAMI. pp. 39-51. 1998.
[8] V. Vapnik. The Nature of Statistical Learning Theory. Springer, N.Y., 1995.
[9] P. Viola and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. CVPR, 2001.
[10] Liwei Wang, Yan Zhang, and Jufu Feng. On the Euclidean Distance of Images. PAMI, 2005
[11] M. H. Yang, D. Kriegman, and N. Ahuja. Detecting Faces in Images: A Survey. PAMI, vol. 24, pp. 34-58. 2002.
[12] http://www.ai.mit.edu/projects/cbcl/software-dataset/index.html