PERGAMON

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition 38 (2005) 651 – 659

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Improved feature reduction in input and feature spaces

Frank Y. Shih*, Shouxian Cheng

Computer Vision Laboratory, Department of Computer Science, College of Computing Sciences, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA

Received 12 May 2004; received in revised form 25 October 2004; accepted 25 October 2004

Abstract

In this paper, we present an improved feature reduction method in input and feature spaces for classification using support vector machines (SVMs). In the input space, we select a subset of input features by ranking their contributions to the decision function. In the feature space, features are ranked according to the weighted support vector in each dimension. By applying feature reduction in both input and feature spaces, we develop a fast non-linear SVM without a significant loss in performance. We have tested the proposed method on the detection of face, person, and car. Subsets of features are chosen from pixel values for face detection and from Haar wavelet features for person and car detection. The experimental results show that the proposed feature reduction method works successfully. In fact, our method performs better than the methods of using all the features and the Fisher's features in the detection of person and car. We also gain the advantage of speed.

Keywords: Feature reduction; Feature ranking; Support vector machine; Object detection

1. Introduction

Feature extraction and reduction are two primary issues in feature selection that is essential in pattern classification. Feature extraction is used to achieve high classification rates by extracting features to represent objects from raw data. Feature reduction is used to select a subset of features with preservation or improvement of classification rates. In general, it intends to speed up the classification process by keeping the most important class-relevant features.

Support vector machines (SVMs) [1,2] are founded from a mathematical point of view. While most classifiers (e.g., Bayesian, neural networks, and radial basis function (RBF)) are trained to minimize the empirical risk, SVMs are implemented to minimize the structural risk. Osuna et al. [3] applied SVMs to face detection. Heisele et al. [4] trained a 2nd-degree polynomial SVM using 10,038 faces and 36,220 non-faces, and achieved a higher detection rate than [3]. However, they need to spend several minutes to search an image for faces at different scales. Heisele et al. [5] presented two following methods to speed up face detection using SVMs: hierarchical classification and feature reduction.

Principal component analysis (PCA) features are used to reduce the dimensionality in input space. Weston et al. [6] developed a feature reduction method by minimizing the bounds on the leave-one-out error. Evgenious et al. [7] introduced a method for feature selection based on the observation that the most important features are the ones that separate the hyperplane the most.

In this paper, we present the method of feature reduction in the input and feature spaces to achieve a fast non-linear SVM without a significant loss in performance. The rest of the paper is organized as follows. Sections 2 and 3 describe the feature reduction methods in input and feature spaces, respectively, for face detection. In Section 4, we develop a fast non-linear SVM by the combination of input and feature spaces for face detection. In Section 5, we apply the

* Corresponding author. Tel.: +1 973 596 5654; fax: +1 973 596 5777.
E-mail address: shih@njit.edu (F.Y. Shih).

0031-3203/$30.00
doi:10.1016/j.patcog.2004.10.004

proposed method for person and car detection. Finally, we provide discussions and conclusions in Section 6.

2. Feature reduction in input space

2.1. Feature ranking

Principal component analysis (PCA) is widely used in image representation for dimensionality reduction. To obtain m principal components, we multiply a transformation matrix of $m \times N$ by an input pattern of $N \times 1$. The computation is costly. In this section, we propose a method of feature reduction in the input space in order to save computational time.

One way of feature reduction is using Fisher's criterion to choose a subset of features that possess a large between-class variance and a small within-class variance. For face detection, we use the within-class variance as

$$\sigma_i^2 = \frac{\sum_{j=1}^{l}(g_{j,i} - m_i)^2}{l-1}, \tag{1}$$

where l is the total number of samples, $g_{j,i}$ is the ith dimensional gray value of sample j, and m_i is the mean value of the ith dimension. We use Fisher's score for between-class measurement as

$$S_i = \left| \frac{m_{i,face} - m_{i,nonface}}{\sigma_{i,face}^2 + \sigma_{i,nonface}^2} \right|. \tag{2}$$

By selecting the features with the highest Fisher's scores, we can retain the most discriminative features between face and non-face classes.

To improve Fisher's method, we propose a 2nd-degree polynomial SVM with kernel $K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})^2$. The decision function for a pattern $\mathbf{x}$ is defined as

$$\begin{aligned} f(\mathbf{x}) &= \sum_{i=1}^{s} \alpha_i y_i (1 + \mathbf{x}_i \cdot \mathbf{x})^2 + b \\ &= \sum_{i=1}^{s} \alpha_i y_i (1 + x_{i,1}x_1 + x_{i,2}x_2 + \cdots + x_{i,k}x_k \\ &\quad + \cdots + x_{i,N}x_N)^2 + b, \end{aligned} \tag{3}$$

where s is the total number of support vectors, $\mathbf{x}_i$ is the ith support vector, and $x_{i,k}$ and x_k are respectively the kth dimension for the support vector $\mathbf{x}_i$ and the pattern $\mathbf{x}$. The component in the kth dimension (where $k = 1, 2, \ldots, N$) is

$$\begin{aligned} f(\mathbf{x}, k) &= \sum_{i=1}^{s} \alpha_i y_i [2x_k x_{i,k}(1 + x_{i,1}x_1 + \cdots + x_{i,k-1}x_{k-1} \\ &\quad + x_{i,k+1}x_{k+1} + \cdots + x_{i,N}x_N) + x_k^2 x_{i,k}^2] \\ &= \sum_{i=1}^{s} \alpha_i y_i [2x_k x_{i,k}(1 + x_{i,1}x_1 \\ &\quad + \cdots + x_{i,N}x_N) - x_k^2 x_{i,k}^2]. \end{aligned} \tag{4}$$

We use the largest m contributions to the decision function out of the original N features. The contribution can be obtained by

$$F(k) = \int_V f(\mathbf{x}, k)\, \mathrm{d}P(\mathbf{x}), \tag{5}$$

where V denotes the input space and $P(\mathbf{x})$ denotes the probability distribution function. Since $P(\mathbf{x})$ is unknown, we approximate $F(k)$ using a summation over the support vectors as

$$\begin{aligned} F(k) = \sum_{i=1}^{s} \Bigg| \sum_{j=1}^{s} \alpha_j y_j [2x_{i,k}x_{j,k}(1 + x_{j,1}x_{i,1} \\ + \cdots + x_{j,N}x_{i,N}) - x_{i,k}^2 x_{j,k}^2] \Bigg|. \end{aligned} \tag{6}$$

2.2. Experimental results

We adopt a face image database from the Center for Biological and Computational Learning at Massachusetts Institute of Technology (MIT), which contains 2429 face training samples, 472 face testing samples, and 23,573 non-face testing samples. We randomly collected 15,228 non-face training samples from the images that do not contain faces. The size of all these samples is 19×19. A 2nd-degree polynomial SVM with kernel $K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})^2$ is used in our experiments.

In order to remove background pixels, a mask is applied to extract only the face. Prior to classification, we perform image normalization and histogram equalization. The image normalization is used to normalize the gray-level distribution by the Gaussian function with zero mean and one variance. The histogram equalization uses a transformation function equal to the cumulative distribution to produce an image whose gray levels have a uniform density. Fig. 1 shows (a) a face image, (b) the mask, and (c) and (d) the images after normalization and histogram equalization, respectively.

Fig. 2 shows the receiver operating characteristic (ROC) curves of using two following methods to obtain PCA features: one uses both face and non-face training samples,

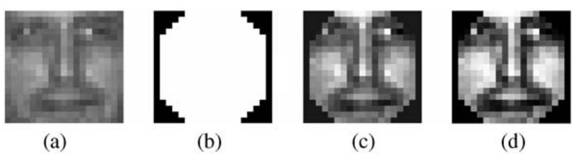

Fig. 1. (a) Original face image, (b) the mask, (c) normalized image, and (d) histogram equalized image.