Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Image and Vision Computing xx (2006) 1–15

www.elsevier.com/locate/imavis

Object detection using spatial histogram features

Hongming Zhang [a,*], Wen Gao [a,b], Xilin Chen [b], Debin Zhao [a]

[a] *Department of Computer Science and technology, Harbin Institute of Technology, No. 92, west Da-zhi street, Harbin 150001, China*
[b] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China*

Received 5 April 2005; received in revised form 22 November 2005; accepted 23 November 2005

Abstract

In this paper, we propose an object detection approach using spatial histogram features. As spatial histograms consist of marginal distributions of an image over local patches, they can preserve texture and shape information of an object simultaneously. We employ Fisher criterion and mutual information to measure discriminability and features correlation of spatial histogram features. We further train a hierarchical classifier by combining cascade histogram matching and support vector machine. The cascade histogram matching is trained via automatically selected discriminative features. A forward sequential selection method is presented to construct uncorrelated and discriminative feature sets for support vector machine classification. We evaluate the proposed approach on two different kinds of objects: car and video text. Experimental results show that the proposed approach is efficient and robust in object detection.

Keywords: Object detection; Spatial histogram features; Feature selection; Histogram matching; Support vector machine

1. Introduction

In computer vision community, object detection has been a very challenging research topic. Given an object class of interest T (target, such as pedestrian, human face, buildings, car or text) and an image P, object detection is the process to determine whether there are instances of T in P, and if so, return locations where instances of T are found in the image P. The main difficulty of object detection arises from high variability in appearance among objects of the same class. An automatic object detection system must be able to determine the presence or absence of objects with different sizes and viewpoints under various lighting conditions and complex background clutters.

Many approaches have been proposed for object detection in images under cluttered backgrounds. In most approaches, the object detection problem is solved within a statistical learning framework. First, image samples are represented by a set of features, and then learning methods are used to identify objects of interest class. In general, these approaches can be classified as two categories: global appearance-based approaches and component-based approaches.

Global appearance-based approaches consider an object as a single unit and perform classification on the features extracted from the entire object. Many statistical learning mechanisms are explored to characterize and identify object patterns. Rowley et al. [1] and Carcia and Delakis [2] use neural network approaches as classification methods in face detection. Based on wavelet features, Osuna et al. [3] and Papageprgiou and Poggio [4] adopt support vector machines to locate human faces and cars. Schneiderman and Kanade [5] use Naïve Bayes rule for face and non-face classification. Recently, boosting algorithms are applied to detect frontal faces by Viola and Jones [6], then are extended for multi-view face detection by Li et al. [7] and for text detection by Chen and Yuille [8]. Other learning methods used in object detection include probabilistic distribution [9,10], principal components analysis [11] and mixture linear subspaces [12].

Component-based methods treat an object as a collection of parts. These methods first extract some object components, and then detect objects by using geometric information. Mohan et al. [13] propose an object detection approach by components. In their approach, a person is represented by components such as head, arms, and legs, and then support vector machine classifiers are used to detect these components and decide whether a person is present. Naquest and Ullman [14] use fragments as features and perform object recognition with informative features and linear classification. Agarwal

[*] Corresponding author. Tel.: +86 10 58858300(313); fax: +86 10 58858301.

E-mail addresses: hmzhang@jdl.ac.cn (H. Zhang), wgao@jdl.ac.cn (W. Gao), xlchen@jdl.ac.cn (X. Chen), dbzhao@jdl.ac.cn (D. Zhao).

0262-8856/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2005.11.010

et al. [15] extract a part vocabulary of side-view cars using an interest operator and learn a Sparse Network of Winnows classifier to detect side-view cars. Fergus et al. [16] and Leibe et al. [17,18] also use interest operators to extract objects' parts and perform detection by probabilistic representation and recognition on many object classes, such as motorbikes, human faces, airplanes, and cars.

As opposed to a majority of the above approaches, the problem of detecting multi-class objects and multi-view objects has been recently gained great attention in computer vision community. Schneiderman and Kanade [5] train multiple view-based detectors for profile face detection and car detection. Lin and Liu [19] propose a multi-class boosting approach to directly detect faces of many scenarios, such as multi-view faces, faces under various lighting conditions, and faces with partial occlusions. Amit et al. [20] use a coarse to fine strategy for multi-class shape detection with an application of reading license plates. There are 37 object classes to be recognized, including 26 letters, 10 digits, and 1 special symbol. Li et al. [21,22] propose methods to learn a geometric model of a new object category using a few examples and detect multi-class objects by a Bayesian approach. To improve efficiency, Torralba et al. [23] introduce an algorithm for sharing features across object classes for multi-class object detection. Tu et al. [24] propose an image parsing framework to combine image segmentation, object detection, and recognition for scene understanding.

One visual task related to object detection is object recognition, whose goal is to identify specific object instances in images. Local descriptor-based methods are increasingly used for object recognition. Schiele [25] proposes to use Gaussian derivatives as local characteristics to create a multi-dimensional histogram as object representation, and then perform the task to recognize many 3D objects. Lowe [26] develops an object recognition system that uses SIFT descriptors based on local orientation histograms. However, these methods are designed to recognize a specific object rather than in generalization to categorize the object class.

Feature extraction for object representation plays an important role in automatic object detection systems. Previous methods have used many representations for object feature extraction, such as raw pixel intensities [1,2,27], edges [28], wavelets [3,4,29], rectangle features [6–8], and local binary pattern [30]. However, what kinds of features are stable and flexible for object detection still remains an open problem.

Motivated by the observation that objects have texture distribution and shape configuration, we propose spatial histogram based features (termed as spatial histogram features) to represent objects. As spatial histograms consist of marginal distributions of an image over local patches, the information about texture and shape of the object can be encoded simultaneously. In contrast to most features previously used, spatial histogram features are specific to the object class, since discriminative information of the object class is embedded into these features through measuring image similarity between the object class and the non-object class. In addition, computation cost of spatial histogram features is low. Our previous work [31] shows that spatial histogram features are effective and efficient to detect human faces in color images.

Based on object representation of spatial histogram features, we present an object detection approach using a coarse to fine strategy in this paper. Our approach uses a hierarchical object detector combining cascade histogram matching and a support vector machine to detect objects, and learns informative features for the classifier. First, we employ Fisher criterion to measure the discriminability of each spatial histogram feature, and calculate features correlation using mutual information. Then, a training method for cascade histogram matching via automatically selecting discriminative features is proposed. Finally, we present a forward sequential selection algorithm to obtain uncorrelated and discriminative features for support vector machine.

Unlike methods which use interest operators to detect parts prior to recognition of the object class, we apply the proposed object detector at anywhere in image scale space. Therefore, our method does not need figure-ground segmentation or object parts localization. In contrast to most systems which are designed to detect a single object class, our method can be applied to any type of object classes with widely varying texture patterns and varying spatial configurations. Extensive experiments on two different kinds of objects (car and video text) are conducted to evaluate the proposed object detection approach.

The rest of the paper is organized as follows. Section 2 outlines the proposed object detection approach. Section 3 describes spatial histogram features for object representation and provides quantitative measurement of spatial histogram features. Section 4 presents the methods of selecting informative features for object detection. Section 5 gives experiment results of car detection and video text detection. Section 6 concludes this paper.

2. Overview of the proposed object detection approach

The proposed approach is designed to detect multiple object instances of different sizes at different locations in an input image. Take car detection as an example, the overall architecture of the object detection approach is illustrated in Fig. 1. One essential component of the proposed approach is an object detector, which uses spatial histogram features as object representation. We call the object detector as spatial histogram features-based object detector (hereinafter referred as 'SHF-based object detector'). In our approach, the SHF-based object detector is formed as a hierarchical classifier which combines cascade histogram matching and support vector machine.

For object detection process, we adopt an exhaustive window search strategy to find multiple object instances in an input image. The process of object detection contains three phases: image pyramid construction (Step 1), object detection at different scales (Step 2), and detection results fusion (Step 3).

Initially, an image pyramid is constructed from the original image in the Step 1. The detector is applied at every location in the image in order to detect object instances anywhere in the input image. To detect objects larger than the fixed size, the input image is repeatedly reduced in scale, and the detector is applied at each scale. As illustrated in Fig. 1, the image pyramid is constructed by subsampling with a factor 1.2. Consequently, the proposed approach can detect objects of different scales. In particular, a small window (subwindow) with the fixed size scans the pyramid of images at different scales and each subimage is verified whether it contains an object instance.

In Step 2, all subimages are passed to the SHF-based object detector. Firstly, spatial histogram features are generated from the image patches. Secondly, cascade histogram matching is performed to all subwindow images for coarse classification. It eliminates a large number of subimages in background as non-objects and provides almost all subwindows of object instances to the fine detection stage. Finally, the support vector machine classification is applied to each remained window to identify whether or not it contains an object instance. If a subimage is indicated as an object instance by the object detector, it is mapped to the image of corresponding scale at the pyramid.

Step 3 is a stage for detection results fusion, in which overlapped object instances of different scales are merged into final detection results. Since the object detector is insensitive to small changes in translation and scale, multiple detections usually occur around each object instance in the pyramid images. We use a grouping method to combine overlapping detections into final detection results. As shown in Fig. 1, all detections at different scales are firstly mapped to the original image scale, resulting in a detection map. The detection map is a binary image, which contains object candidate regions and background. A grouping algorithm is applied to label the detection map into disjoint regions. In addition, some very small regions are removed because very small regions usually correspond to false detections. Each region yields a single final detection. The position and size of each final detection is the average of the positions and sizes of all detections in the region. As a result, the number of object instances and their locations and scales are reported in the output image.

The SHF-based object detector is constructed through a coarse to fine strategy. For any input image patch of the fixed size, the SHF-based object detector initially produces spatial

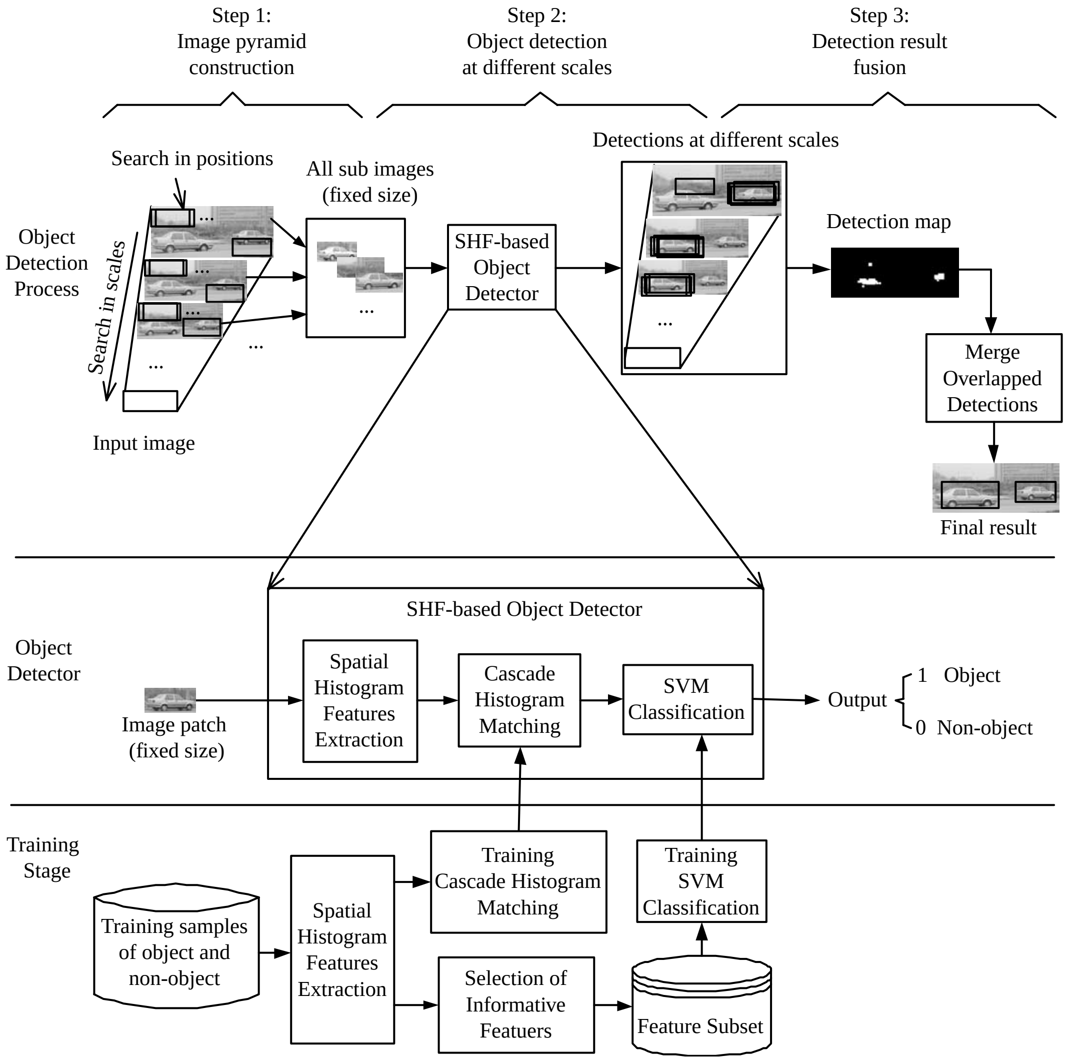

Fig. 1. Architecture of the proposed object detection approach.

g_1	g_2	g_3
g_4	g_0	g_5
g_6	g_7	g_8

Fig. 2. Neighborhood for basic LBP computation.

histogram features from the image path, and then perform hierarchical classification with cascade histogram matching method and support vector machine. As a result, the SHF-based object detector generates an output that indicates whether or not the input image patch is an object instance. In the object detection process, cascade histogram matching method and support vector machine play different roles. Cascade histogram matching quickly locates object candidate instances, and support vector machine precisely verifies the object candidate instances. During training stage, a lot of samples of object and non-object are used to select informative spatial histogram features and to train the SHF-based object detector.

3. Spatial histogram features

Object representation and feature extraction are essential to object detection. In this section, we describe a novel object pattern representation combining texture and spatial structures. Specially, we model objects by their spatial histograms over local patches and extract class specific features for object detection. Moreover, we quantitatively analyze spatial histogram features by discriminating feature analysis and features correlation measurement.

3.1. Spatial histograms

In our approach, a subwindow contains a grey sample image with a certain size. Local binary pattern (LBP) is used to preprocess sample images. Local binary pattern is a relatively new and simple texture model and it has been proved to be a very powerful feature in texture classification [32,33]. LBP is invariant against any monotonic transformation of the gray scale. As illustrated in Fig. 2, basic LBP operator uses neighborhood intensities to calculate the region central pixel value.

The 3×3 neighborhood pixels are signed by the value of center pixel:

$$s(g_0, g_i) = \begin{cases} 1, & g_i \geq g_0, \\ 0, & g_i < g_0, \end{cases} \quad 1 \leq i \leq 8. \tag{1}$$

The signs of the eight differences are encoded into an 8-bit number to obtain LBP value of the center pixel:

$$\mathrm{LBP}(g_0) = \sum_{i=1}^{8} s(g_0, g_i) 2^{i-1}. \tag{2}$$

For any sample image, we compute histogram-based pattern representation as follows. First we apply variance normalization on the gray image to compensate the effect of different lighting conditions, then we use basic local binary pattern operator to transform the image into an LBP image, and finally

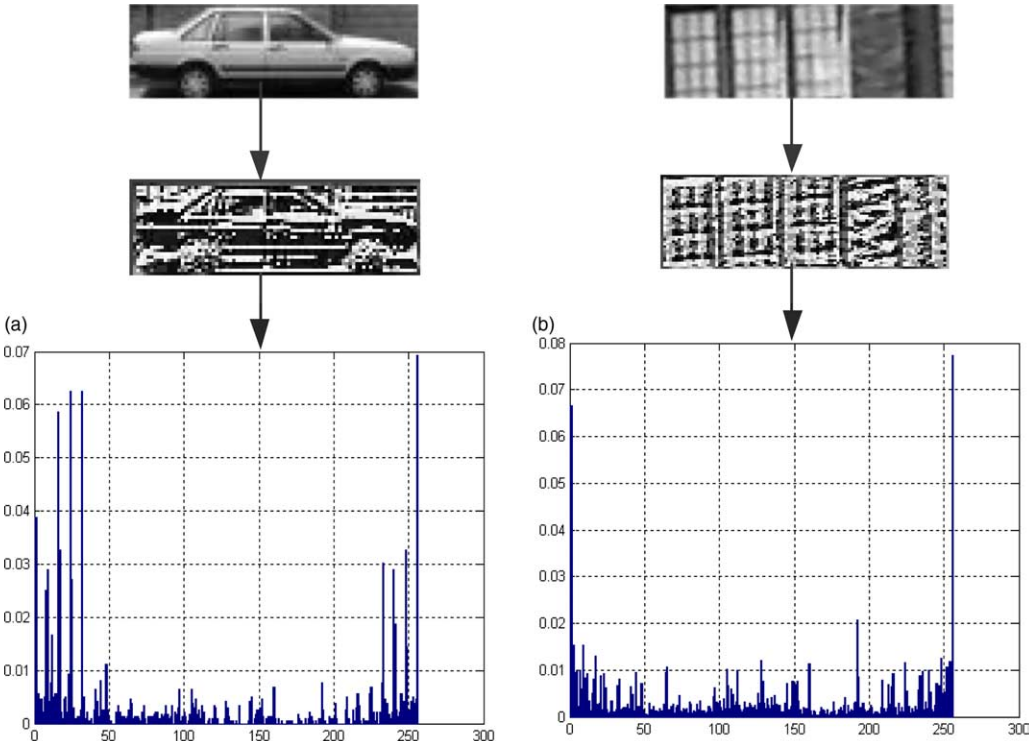

Fig. 3. Examples of two image samples. (a) An image sample of a side-view car, its LBP image and histogram. (b) A non-car image sample, its LBP image and histogram.

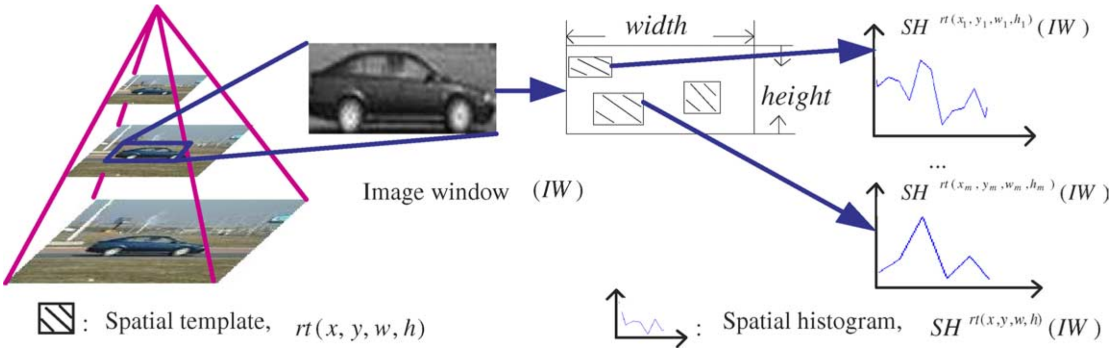
width
height
SH rt(x1, y1, w1, h1) (IW)
...
SH rt(xm, ym, wm, hm) (IW)
Image window (IW)
: Spatial template, rt(x, y, w, h)
: Spatial histogram, SH rt(x,y,w,h) (IW)

Fig. 4. Spatial distribution is encoded by spatial histograms in image scale space.

we compute histogram of the LBP image as representation. Fig. 3 shows a sample image of a side-view car and a non-car sample image, their LBP images and histograms.

It is easy to prove that histogram, a global representation of image pattern, is invariant to translation and rotation. However, histogram is not adequate for object detection since it does not encode spatial distribution of objects. For some non-object images and object images, their histograms can be very similar or even identical, making histogram not sufficient for object detection.

In order to enhance discrimination ability, we introduce spatial histograms, in which spatial templates are used to encode spatial distribution of object patterns. As illustrated in Fig. 4, we use an image window with a fixed size (*width*, *height*) to sample object patterns in image scale space, and then encode pattern spatial distribution by spatial templates. Each template is a binary rectangle mask, shown in Fig. 5. We denote each template as rt(x,y,w,h), where (x,y) is the location of the top left position of the mask, while w and h are the width and height of the mask, respectively.

For a single spatial template rt(x,y,w,h), we model subimage within the masked window by histogram. We call this kind of histograms as *spatial histograms*. For a sample image P, its spatial histogram associated with the template rt(x,y,w,h) is denoted as $\mathrm{SH}^{\mathrm{rt}(x,y,w,h)}(P)$.

3.2. Object features extracted from spatial histograms

A lot of methods can be used to measure similarity between two histograms, such as quadratic distance, Chi-square distance and histogram intersection [25]. In this paper, we adopt histogram intersection for its stability and computational inexpensiveness. Similarity measurement by intersection of two histograms [34] is calculated as

$$D(H_1, H_2) = \frac{\sum_{i=1}^{K} \min(H_1^i, H_2^i)}{\sum_{i=1}^{K} H_1^i}, \tag{3}$$

where H_1 and H_2 are two histograms, and K is the number of bins in the histograms.

Suppose a database with n object samples and a spatial template, we represent object histogram model over the spatial template by the average spatial histogram of the object training samples, defined as

$$\mathrm{SH}^{\mathrm{rt}(x,y,w,h)} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{SH}^{\mathrm{rt}(x,y,w,h)}(P_j), \tag{4}$$

where P_j is an object training sample, and rt(x,y,w,h) is the spatial template. For any sample P, we define its *spatial histogram feature* $f^{\mathrm{rt}(x,y,w,h)}(P)$ as its distance to the average object histogram, given by

$$f^{\mathrm{rt}(x,y,w,h)}(P) = D(\mathrm{SH}^{\mathrm{rt}(x,y,w,h)}(P), \mathrm{SH}^{\mathrm{rt}(x,y,w,h)}). \tag{5}$$

An object pattern is encoded by a spatial template set {rt(1),…,rt(m)}, where m is the number of spatial templates. Therefore, an object sample is represented by a spatial histogram feature vector in the spatial histogram feature space:

$$F = [f^{\mathrm{rt}(1)}, \ldots, f^{\mathrm{rt}(m)}]. \tag{6}$$

As the masks can vary in positions and sizes in the image window, the exhaustive set of spatial histogram features is very large. Therefore, the spatial histogram feature space completely encodes texture and spatial distributions of objects. In addition, spatial histogram feature is a kind of object class specific features, since it encodes sample's similarity to object histogram models.

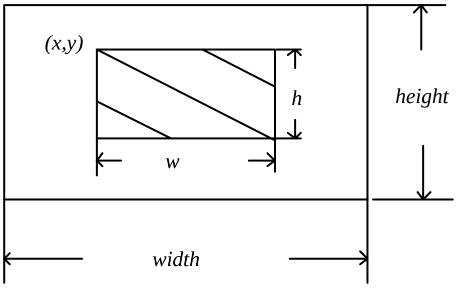

Fig. 5. Image window and spatial template.

3.3. Feature discriminating ability

Each type of spatial histogram feature has discriminating ability between object pattern and non-object pattern. To demonstrate this property, we take a spatial histogram feature of side-view car pattern as an example. The size of sample image is 100×40 pixels. The spatial template is rt(40,20,20,20), which is within the 100×40 image window and locates a 20×20 mask in position (40,20). The car model over this spatial template $\mathrm{SH}_{\mathrm{car}}^{\mathrm{rt}(40,20,20,20)}$ is generated by 200 car samples. The spatial histogram feature $f^{\mathrm{rt}(40,20,20,20)}$ is the testing feature.

Fig. 6 shows the testing feature's distribution over an image sample set containing 2000 car samples and 15,000 non-car samples. In horizontal axis, feature value stands for the value range of the testing feature. In vertical axis, frequency reflects the feature value's distribution of car class samples and non-car class samples. For each class, frequency is the numbers of the samples with feature values over the total number of the corresponding class samples.

As shown in Fig. 6, we use a threshold to classify car and non-car on the testing feature. By setting the threshold to 0.7, we retain 99.1% car detection rate with false alarm rate 45.1% and threshold 0.8 produces 93.8% car detection rate with false alarm rate 12.1%.

We adopt Fisher criterion to measure discriminating ability of each spatial histogram feature. For a spatial histogram feature f_j $(1\leq j\leq m)$, suppose that we have a set of N samples $x_1,x_2,\ldots,x_N$, where each x_i is a scalar value of the spatial histogram feature. In the data set, N_1 samples are in object subset labelled ω_1 and N_2 samples in non-object subset labelled ω_2. Between-class scatter S_b is the distance between the two classes given by

$$S_\mathrm{b}=(m_1-m_2)^2, \tag{7}$$

where $m_i=\frac{1}{N_i}\sum_{x\in\omega_i}x, i\in\{1,2\}$. Within-class scatter S_i for each class is computed as

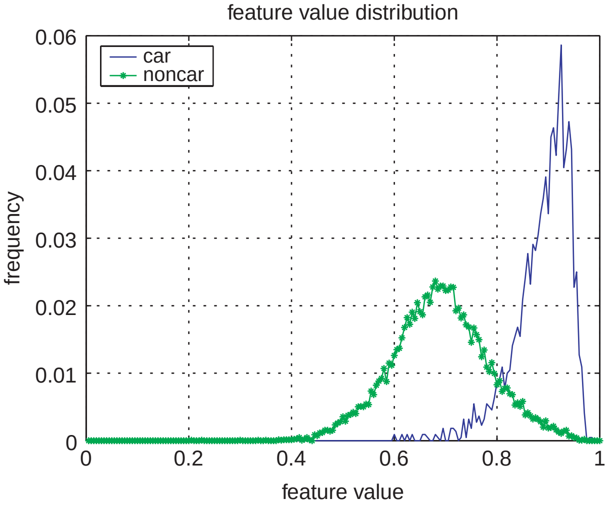

Fig. 6. Feature distribution.

$$S_i=\frac{1}{N_i}\sum_{x\in\omega_i}(x-m_i)^2,\quad i\in\{1,2\}, \tag{8}$$

then total within-class scatter S_w is defined by

$$S_\mathrm{w}=S_1+S_2. \tag{9}$$

Thus, Fisher criterion of the spatial histogram feature f_j is the ratio of the between-class to the total within-class scatter, given by

$$J(f_j)=\frac{S_\mathrm{b}}{S_\mathrm{w}}. \tag{10}$$

The greater Fisher criterion is, the more discriminative the spatial histogram feature is.

3.4. Features correlation measurement

An efficient feature set requires not only each feature has strong discriminating ability, but also they are mutually independent. There are many methods to calculate features correlation, such as mutual information [35,36] and correlation coefficient [37]. We employ mutual information to measure features correlation, since mutual information is a natural indicator of statistical dependence between random variables and takes into account the amount of information shared between variables.

A spatial histogram feature of samples is a random variable, which expresses distance between each sample's spatial histogram and associated object histogram model. For any spatial histogram feature f_j $(1\leq j\leq m)$, we denote it as a variable X, its entropy is defined as

$$H(X)=-\int p(x)\log_2 p(x)\mathrm{d}x, \tag{11}$$

where $p(x)$ is probability density of the variable. In our approach, we adopt a Parzen window density estimate technology [38] to approximate the probability density using a set of training samples of the feature f_i.

Given two spatial histogram features f_1 and f_2, the mutual information of these two features is defined by

$$I(f_1|f_2)=H(f_1)+H(f_2)-H(f_1,f_2), \tag{12}$$

where $H(f_1,f_2)$ is the joint entropy of f_1 and f_2.

It is obvious that $I(f_1|f_2)=I(f_2|f_1)$ and $0\leq I(f_1|f_2)\leq H(f_1)$. Therefore, we calculate feature correlation between the two features f_1 and f_2 as

$$\mathrm{Corr}(f_1,f_2)=\frac{I(f_1|f_2)}{H(f_1)}. \tag{13}$$

Let F_s be a feature subset, we calculate the correlation between a feature $f_\mathrm{m}\notin F_\mathrm{s}$ and F_s as follows:

$$\mathrm{Corr}(f_\mathrm{m},F_\mathrm{s})=\max\{\mathrm{Corr}(f_\mathrm{m},f_k)|\,\forall f_k\in F_\mathrm{s}\}. \tag{14}$$

4. Learning informative features for object detection

We apply a hierarchical classification using cascade histogram matching and support vector machine to object detection. Since the spatial histogram feature space is high dimensional as mentioned in Section 3, it is crucial to get a compact and informative feature subset for efficient classification. In this section, the selection methods of informative features based on discriminability and features correlation are presented.

4.1. Cascade histogram matching

Histogram matching is a direct method for object recognition. In this method, a histogram model of an object pattern is first generated for one spatial template. If the histogram of a sample is close to the model histogram under a certain threshold, the sample is classified as an object pattern. Let P is a sample and its spatial histogram feature with one template $\mathrm{rt}(x,y,w,h)$ is $f^{\mathrm{rt}(x,y,w,h)}(P)$, P is classified as object pattern if $f^{\mathrm{rt}(x,y,w,h)}(P)\geq\theta$, otherwise P is classified as non-object pattern. θ is the threshold for classification.

Histogram matching with one spatial template is far from acceptable by an object detection system. We select most informative spatial histogram features and combine them in a cascade form to perform histogram matching. We call this classification method as cascade histogram matching. If n spatial histogram features $f_1,\ldots,f_n$ with associated classification thresholds $\theta_1,\ldots,\theta_n$ are selected, the decision rule of cascade histogram matching is as follows:

$$H(P)=\begin{cases}1 \text{ object}, & \text{if}(f_1(P)\geq\theta_1)\wedge\ldots\wedge(f_n(P)\geq\theta_n),\\ 0 \text{ non-object}, & \text{otherwise.}\end{cases} \tag{15}$$

We measure each feature's contribution by its Fisher criterion and detection rate, and propose a training method for cascade histogram matching. Detection rate is the classification accuracy on a positive object samples set. This method selects a discriminative feature set F_{select} with a classification threshold set ThreSet to construct a cascade histogram matching classifier.

Suppose that we have (1) spatial histogram features space $F=\{f_1,\ldots,f_m\}$, (2) positive and negative training sets: SP and SN, (3) positive and negative validation sets: $\mathrm{VP}=\{(x_1,y_1),\ldots,(x_n,y_n)\}$ and $\mathrm{VN}=\{(x'_1,y'_1),\ldots,(x'_k,y'_k)\}$, where x_i and x'_i are samples with m-dimensional spatial histogram feature vectors, $y_i=1$ and $y'_i=0$, (4) acceptable detection rate: D. The method for training cascade histogram matching is given in the following procedure:

(1) Initialization: $F_{\text{select}}=\phi$, $\text{ThreSet}=\emptyset$, $t=0$, and set two classification accuracy parameters to zero, i.e. $\text{Acc(pre)}=0$, $\text{Acc(cur)}=0$;
(2) Compute Fisher criterion $J(f)$ using training sample sets SP and SN, for each feature $f\in F$;
(3) Find spatial histogram feature f_t which has maximal Fisher criterion value, i.e.
$$f_t=\arg\max_{f_j}\{J(f_j)|f_j\in F\};$$
(4) Perform histogram matching with f_t on the validation set $V=\mathrm{VP}\cup\mathrm{VN}$, find a threshold θ_t such that the detection rate d on the positive validation set VP is greater than D, i.e. $d\geq D$;
(5) Compute the classification accuracy on the negative validation set VN
$$\text{Acc(cur)}=1-\frac{1}{k}\sum_{i=1}^{k}|H(x'_i)-y'_i|.$$

Here, $H(x)$ is the classification output by histogram matching with f_t and θ_t, and $H(x)\in\{0,1\}$;

(6) If the classification accuracy satisfy condition: $\text{Acc(cur)}-\text{Acc(pre)}\leq\epsilon$ (ϵ is a small positive constant), the procedure exits and returns F_{select} and ThreSet, otherwise process following steps:
   (a) $\text{Acc(pre)}=\text{Acc(cur)}$, $\text{SN}=\phi$, $F_{\text{select}}=F_{\text{select}}\cup\{f_t\}$, $F=F\backslash\{f_t\}$, $\text{ThreSet}=\text{ThreSet}\cup\{\theta_t\}$, $t=t+1$;
   (b) Bootstrap: perform cascade histogram matching with F_{select} and ThreSet on an image set containing no target objects, put false detections into SN;
   (c) Go to (2) and continue next iteration step.

4.2. Support vector machine for object detection

Cascade histogram matching is the coarse object detection stage and it can obtain high detection rate, however, the false positive rate is still high. For the sake of improvement of detection performance, we employ support vector machine (SVM) classification as a fine object detector.

An SVM [39] performs pattern recognition for a two-class problem by determining the separating hyper plane that maximizes the distance to the closest points of a training set. In our approach, we first adopt an SVM method as the evaluation classifier in selecting informative spatial histogram features, and then use the selected feature set to train an SVM for object detection using the Libsvm software [40].

By integrating discriminability and features correlation, we use a forward sequential selection method to iteratively select a feature subset F_{select} for classification. Initially, F_{select} is set to be empty. In each iteration, this method firstly chooses an uncorrelated spatial histogram feature with large Fisher criterion, then uses a classifier to evaluate the performance of the selected feature subset, and finally adds a feature which has maximum classification accuracy to F_{select}.

Suppose that we have (1) a spatial histogram features space $F=\{f_1,\ldots,f_m\}$, (2) a training sample set $s=\{(x_1,y_1),\ldots,(x_n,y_n)\}$ and a testing sample set $v=\{(x'_1,y'_1),\ldots,(x'_k,y'_k)\}$,where x_i and x'_i are samples with m-dimensional spatial histogram feature vectors, $y_i\in\{0.1\}$ and $y'_i\in\{0.1\}$ for negative and positive samples, respectively. The selection of feature subset F_{select} is performed as the following procedure:

Fig. 7. Some samples of training car images.

(1) Find f^* with maximum Fisher criterion, $F_{\text{select}}=\{f^*\}$ and $F_{\text{ori}}=F\backslash\{f^*\}$;

(2) Set classification accuracy: Acc(pre) = 0;

(3) Compute Fisher criterion $J(f)$ and feature correlation $\text{Corr}(f,F_{\text{select}})$ on the training sample set s, for each feature $f\in F_{\text{ori}}$;

(4) Compute Thre as follows:

$$\begin{cases} \text{MinCorr} = \min\{\text{Corr}(f,F_{\text{select}})|f\in F_{\text{ori}}\}, \\ \text{MaxCorr} = \max\{\text{Corr}(f,F_{\text{select}})|f\in F_{\text{ori}}\}\ , \\ \text{Thre} = \text{MinCorr}*(1-\alpha)+\text{MaxCorr}*\alpha, \end{cases}$$

here α is a balance weight $(0<\alpha<1)$, we choose $\alpha=0.2$ in experiments;

(5) Find $f'\in F_{\text{ori}}$ with large Fisher criterion as below:

$$f' = \arg\max_{f_j}\{J(f_j)|\text{Corr}(f_j,F_{\text{select}})\leq\text{Thre}\};$$

(6) Train an evaluation classifier C on the training set s, using f' and F_{select}. It should be noted that C can be any type of classifier, such as artificial neural network, Naïve Bayes classifier, and nearest neighbor classifier. In our experiment, we use SVM as the evaluation classifier.

(7) Evaluate the classifier C on the testing samples set v, and compute the classification accuracy:

$$\text{Acc(cur)} = 1-\frac{1}{k}\sum_{i=1}^{k}|C(x_i')-y_i'|.$$

Here, $C(x)$ is classification output by the classifier C using f' and F_{select}, and $C(x)\in\{0,1\}$;

(8) If the classification accuracy satisfy condition: $\text{Acc(cur)}-\text{Acc(pre)}\leq\epsilon$ (ϵ is a small positive constant), the procedure exits and returns F_{select} that contains the selected features, otherwise process following steps:
   (a) $\text{Acc(pre)}=\text{Acc(cur)}$, $F_{\text{select}}=F_{\text{select}}\cup\{f'\}$, $F_{\text{ori}}=F_{\text{ori}}\backslash\{f'\}$,
   (b) Go to (3) and continue next iteration step.

After running the above feature selection algorithm, we train an SVM classifier for object detection in images using the selected feature set F_{select}. The SVM classifier and the cascade histogram matching constitute the final object detector based on the coarse to fine strategy as shown in Fig. 1.

5. Experimental results

In order to evaluate the effectiveness of the proposed approaches, we conduct experiments of two different object detection tasks. One is to detect side-view car, which has semi-rigid structure with special componential configuration. The other is text detection in video frames. Text region is mainly a texture pattern without any obvious componential structure.

Some performance measures are used to evaluate object detection systems: (1) detection rate is defined as the number of correct detections over the total number of positives in data set; (2) false positive rate is the number of false positives over the total number of negatives in data set; (3) precision is the number of correct detections over the sum of correct detections and false positives.

5.1. Car detection

Side-view cars consist of distinguishable parts such as wheels, car doors, and car windows. These parts are arranged in a relatively fixed spatial configuration. Side-view cars have enormous changes in configurations because of various design styles.

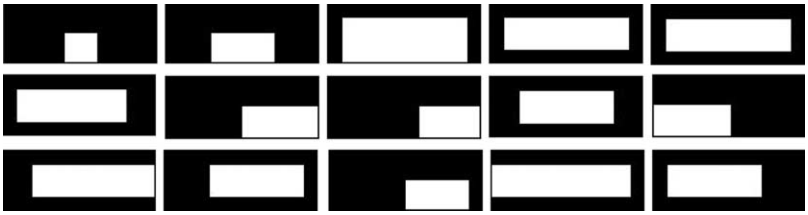

Fig. 8. Selected 15 spatial templates for cascade histogram matching in car detection.

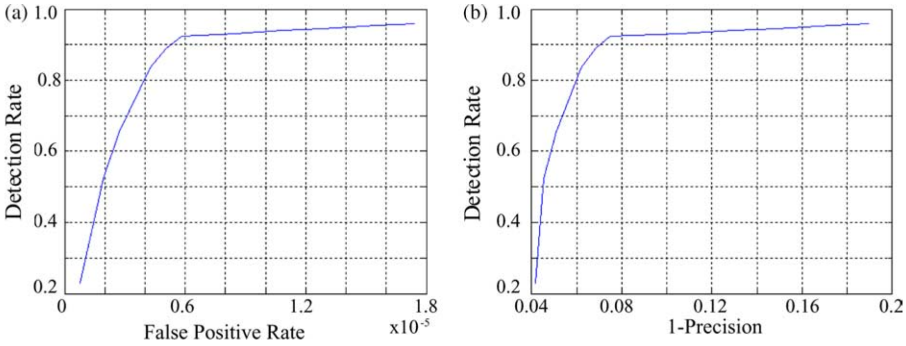

Fig. 9. (a) ROC and (b) RPC curves obtained on UIUC car detection test set A.

We build a training image database with 2725 car samples and 14,968 non-car samples, each 100×40 pixel in size. 500 car sample images are from the training image set from the UIUC image database for car detection [41]. Other car images are collected from video frames and web sites. We also construct a validation set containing 1225 car images and 7495 non-car images for training cascade histogram matching and selection of informative classification features for SVM. Some car samples are shown in Fig. 7.

The exhaustive spatial template set within 100×40 image window is very large: 3,594,591. However, this spatial template set is over complete, and most spatial templates are with small and meaningless size or mutually overlapped. To reduce redundant and meaningless spatial templates, the mask is moved in steps of size 5 pixels in the horizontal and vertical directions and only those spatial templates, whose masks are multiple times the size of 10×10, are used in car detection. In total, 270 spatial templates in a 100×40 image sample are evaluated to extract spatial histogram features. In our experiment, we use the sample database to train cascade histogram matching. As a result, 15 spatial templates (see Fig. 8) are learned for cascade histogram matching to reject most non-car instances in the coarse detection stage. In order to improve the detection accuracy, 25 spatial templates are learned for SVM classification with RBF (Radial Basis Function) kernel in the fine detection stage to perform object verification.

We test our system on two test image sets from the UIUC image database for car detection [41]. The first set (Test Set A) consists of 170 images containing 200 cars with roughly the same size as in the training images. The second set (Test Set B) consists of 108 images containing 139 cars with different sizes. The test sets are difficult for detection since they contain partially occluded cars, and cars that have low contrast with backgrounds.

To get understanding of the overall performance of the car detection system, we report the receiver operating characteristics (ROC) curves and recall-precision curves (RPC) of test sets A and B as shown in Fig. 9 and Fig. 10. These curves are obtained by changing classification thresholds of cascade histogram matching and SVM, and then running our car detection system on the test image sets.

For analysis of the performance of different steps in the coarse to fine object detection system, we conduct experiments using two schemes on test set A. The first scheme is to use cascade histogram matching without SVM classification, the second is to use the combination of cascade histogram matching with SVM classification. Table 1 is the results of experiments using two sets of different classification thresholds. The experimental results show that cascade histogram matching method gets high detection rates, however false positive rate is still high and detection precision is low. As the fine object detection method, the SVM classification improves the detection precisions without significant loss of detection rates.

In Table 2 and Table 3, the experimental results on test sets A and B are compared with the results reported on the same data sets from Agrawal et al. [15]. From the experimental results, our car detection approach outperforms the results reported by Agrawal et al. [15] with higher detection rate and lower false detections number.

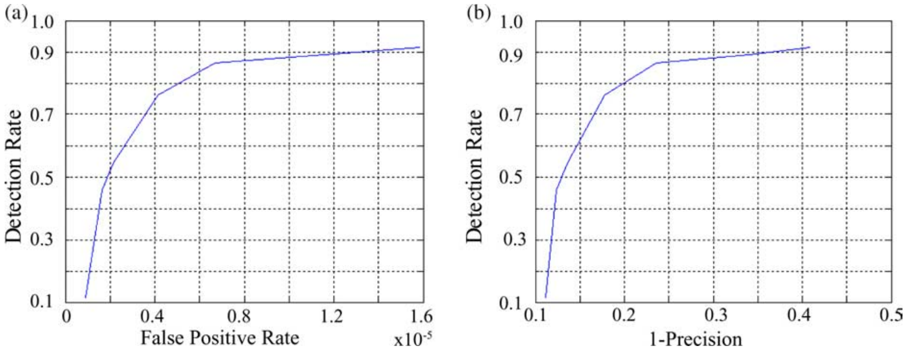

Fig. 10. (a) ROC and (b) RPC curves obtained on UIUC car detection test set B.

Table 1
Car detection results on test set A

	No. of correct detections	No. of false detections	Detection rate (%)	Precision (%)
Without SVM	164	187	82.0	46.7
With SVM	158	11	79.0	93.4
Without SVM	196	441	98.0	30.7
With SVM	193	45	96.5	81.1

Table 2
Comparison of car detection results on test set A

Method	No. of correct detections	No. of false detections	Detection rate (%)	Precision (%)
Agarwal et al. [15]	183	557	91.50	24.73
Our approach	193	45	96.50	81.1

Table 3
Comparison of car detection results on test set B

Method	No. of correct detections	No. of false detections	Detection rate (%)	Precision (%)
Agarwal et al. [15]	112	1216	80.58	8.43
Our approach	120	37	86.33	76.43

Table 4
Comparison of car detection results on test set A reported in literature

	Agarwal et al. [15]	Fergus et al. [16]	Leibel et al. [17]	Leibel et al. [18]	Our approach
ERR	77.0%	88.5%	97.5%	91.0%	92.5%
Scale inv.	Yes	Yes	No	Yes	Yes

Table 4 shows the experimental results of RPC equal error rate on the test set A and scale invariance from different car detection systems [15–18]. Except Agrawal et al. [15] and our work, many systems conduct experiments only on test set A which contains cars roughly in single-scale. Our system achieves a performance of recall-precision curve (RPC) equal error rate 92.5%, which outperforms Agrawal et al. [15], Fergus et al. [16] and Leibel et al. [18]. The system of Leibel et al. [17] obtains a performance of equal error rate 97.5% by using a complete and flexible representation of the object class with an implicit shape model. Based on detection confidences over pixels, the system can combine object categorization and segmentation. However, it is only capable of detecting side-view cars in single-scale with a small tolerance for scale changes. Although the accuracy is lower than that of Leibel et al. [17], our system still compares favorably to Leibel et al. [17] and can detect objects in multi-scales.

In Fig. 11, some car detection results are given. These images contain highly variable side-view cars with different sizes under complex backgrounds. As shown in Fig. 12(a) and (b), some cars are far small than 100×40, so they are often missing detected. Some false car detections are presented in Fig. 12(c) and (d).

5.2. Video text detection

Text detection is the process of detecting and locating regions that contain texts from a given image. We apply the proposed approach to detect text in video frames. A text block pattern is an image window 50×20 in size. A text region classifier is constructed using spatial histogram features. We detect text lines by two steps. First, we scan the image at multiple scales by the text region classifier to produce a text region map. Second, we segment text regions into distinct text lines using vertical segmentation algorithm similar to [42].

We build a training database with 2936 text images extracted form video frames and 12,313 non-text images, each 50×20 pixels in size. We also construct a validation set containing 2012 text images and 6865 non-text images for training cascade histogram matching and learning of informative classification features for SVM. Some text samples are shown in Fig. 13.

After reducing redundant spatial templates, 130 spatial templates are evaluated to extract spatial histogram features. We use the training data set to select informative spatial histogram features for text detection. As a result, 17 spatial templates are learned for cascade histogram matching. 32 spatial histogram features are learned for an SVM with RBF kernel to improve text detection performance.

The final text detection system is tested on a video text detection test set used in Hua et al.'s work [43,44]. The video text set contains 45 video frames extracted from the MEPG-7 Video Content Set. There are totally 158 text blocks in the 45 frames and 128 of them are human-recognizable. Hua et al.'s work [44] and our work use the 128 human-recognizable text blocks as ground truth data. The video text detection test set is

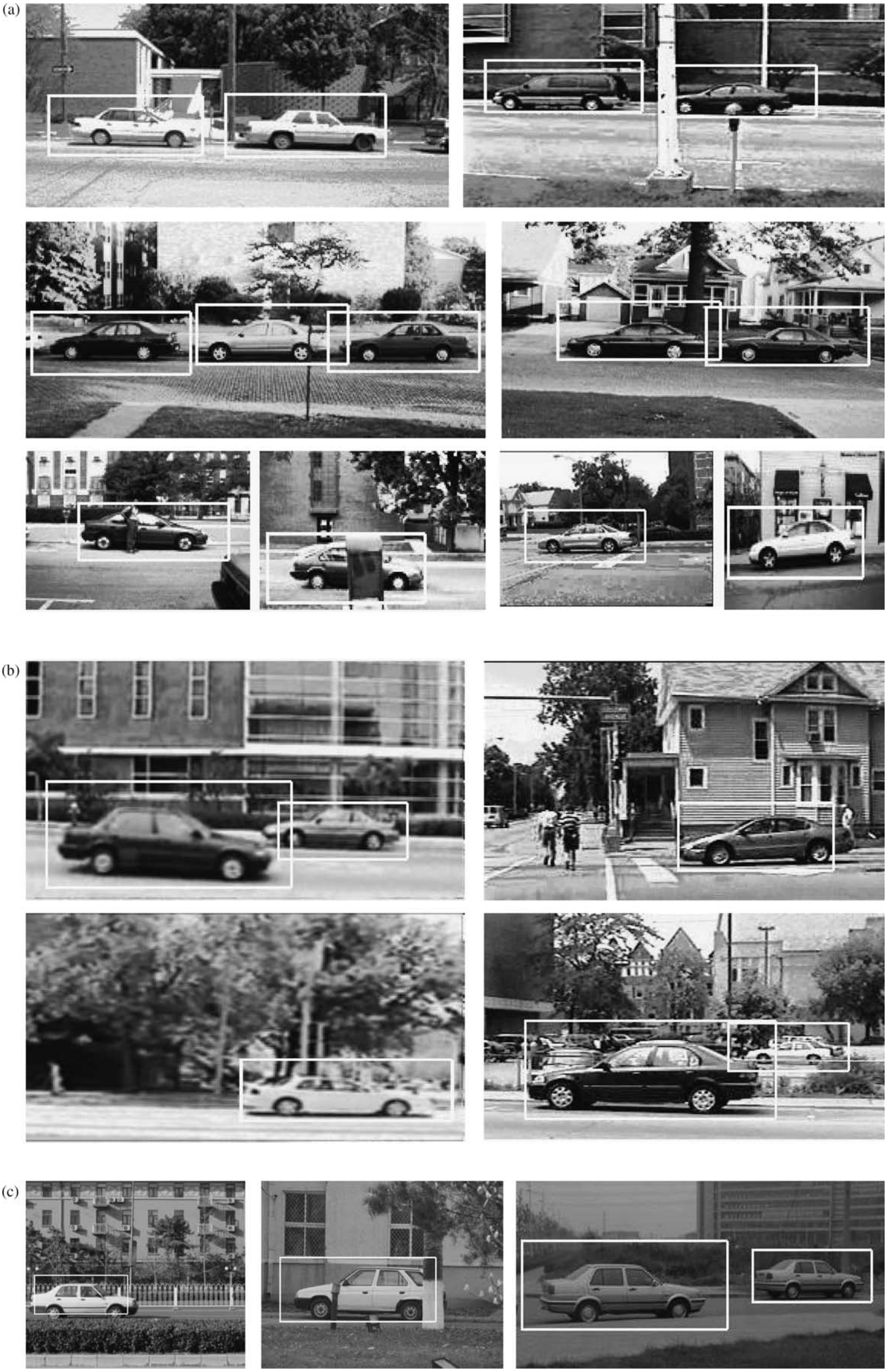

Fig. 11. Some experiment results of car detection on (a) UIUC test set A, (b) UIUC test set B, and (c) some other digital photos.

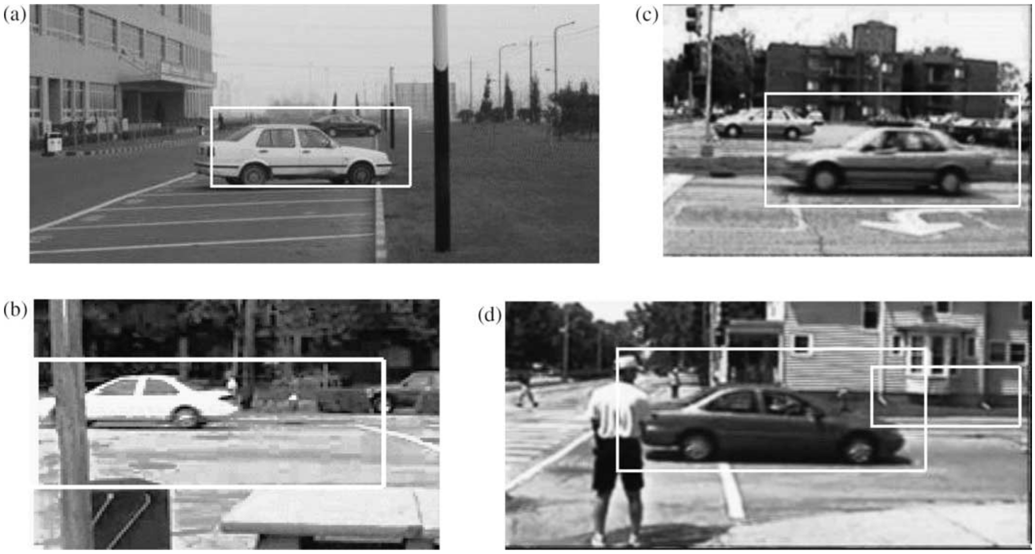

Fig. 12. Car detection results: missing detections (a and b) and false detections (c and d).

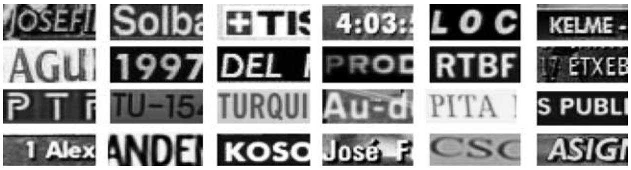

Fig. 13. Some training samples of text images.

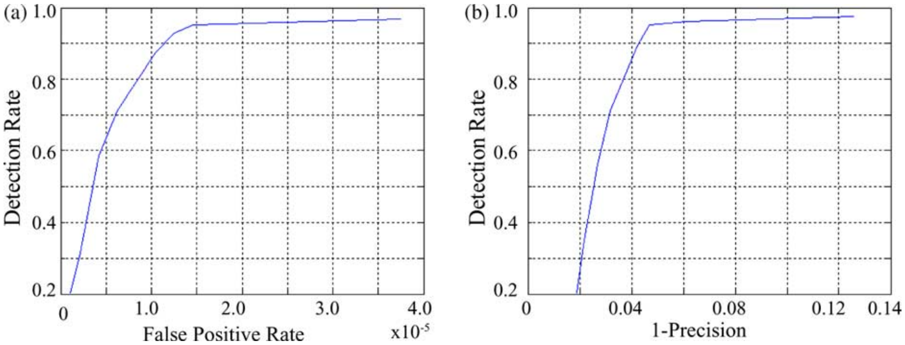

Fig. 14. (a) ROC curve and (b) RPC curve obtained on video text test set.

Table 5
Text detection results on the video text set

	No. of text blocks	No. of correct detections	No. of false detections	Detection rate (%)	Precision (%)
Without SVM	128	78	112	60.9	41.0
With SVM	128	72	2	56.3	97.2
Without SVM	128	119	237	93.0	33.4
With SVM	128	114	5	89.0	95.7

now available at http://www.cs.cityu.edu.hk/(liuwy/PE_VTDetect/

Fig. 14 reports the ROC curve and RPC curve that are obtained on the video text detection test set. These curves characterize the overall performance of the video text detection system. The maximum detection rate is 96.8% with 10 false detections and the RPC equal error rate is 95.3%.

Similar to car detection, we also conduct experiments to analyze performance of cascade histogram matching and support vector machine. Table 5 shows results of experiments

Table 6
Comparison of text detection results on the video text set

Method	No. of text blocks	No. of correct detections	No. of false detections	Detection rate (%)	Precision (%)
Hua et al. [44]	128	117	6	91.4	95.1
Our approach	128	122	6	95.3	95.3

using two sets of different classification thresholds. These results show that high detection rates and low detection precision are obtained during the cascade histogram matching stage. In the fine object detection stage, the SVM improves detection precisions without significant loss of detection rates.

In Table 6, the experimental results on the video text detection test set are compared with the results reported on the same date set from Hua et al. [44]. From the experimental results, our text detection approach outperforms the approach reported by Hua et al. [44] with higher detection rate.

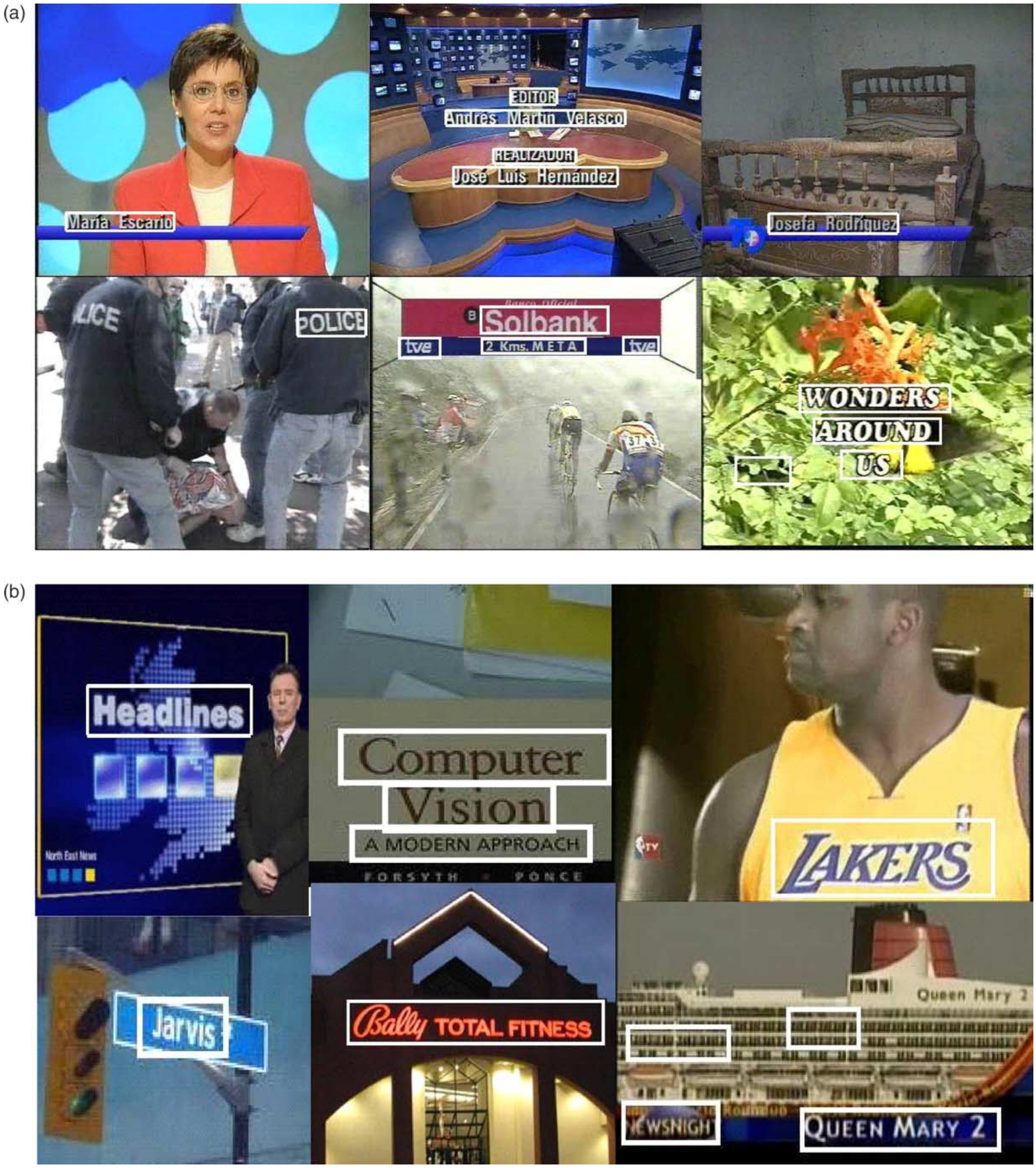

Fig. 15. Text detection results on (a) the video text set and (b) still images.

In Fig. 15, some video text detection results are given. These examples include text lines with considerable variations in contrast, intensity and texture. Some text lines are missing detected because of their small size or low contrast. Some false detected text lines are similar to text in contrast and texture.

5.3. Performance time

We implement the object detection methods on a conventional PC. Operating on 320×240 pixel images, the side-view car detection method proceeds at 9 frames per second on a Pentium IV 3.2 GHz CPU. For 320×240 pixel images, the text detection method has an average detection speed of 5 frames per second using a Pentium IV 3.2 GHz CPU. These results show that the proposed object detection methods are suitable for practical applications.

6. Conclusions

In this paper, we have presented a spatial histogram feature-based object detection approach. This method automatically selects informative spatial histogram features, and learns a hierarchical classifier by combining cascade histogram matching and a support vector machine to detect objects in images. Extensive object detection experiments show high detection rates with relatively low numbers of false detections. These results illustrate the high discriminant power of spatial histogram features and the effectiveness and robustness of the hierarchical object detection approach.

The proposed approach is able to detect not only the objects that consist of distinguishable parts in spatial configurations, such as side-view cars, but also the objects without fixed part-based configurations, such as text lines in video frames. In summary, the results show that the object representation using spatial histogram features is general to different kinds of object classes, and our feature selection methods are efficient to extract informative class-specific features for object detection.

As a direct extension of this work, we are currently investigating spatial histogram features for multi-class objects detection. The ongoing experiments dealing with multi-view human face detection are very encouraging.

Acknowledgements

The authors would like to thank the anonymous reviewers for their invaluable comments and suggestions. They also would like to thank Dr Jie Yang for his helpful discussion. This research is partially sponsored by Natural Science Foundation of China under contract No. 60332010, the 100 Talents Program of CAS, ShangHai Municipal Sciences and Technology Committee (No. 03DZ15013), and ISVISION Technologies Co., Ltd.

References

[1] H.A. Rowley, S. Baluja, T. Kanade, Neural network-based face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1) (1998) 29–38.

[2] C. Garcia, M. Delakis, Convolutional face finder: a neural architecture for fast and robust face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (11) (2004) 1408–1423.

[3] E. Osuna, R. Freund, F. Girosi, Training support vector machines: an application to face detection, Proceedings of the 1997 IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 130–136.

[4] C.P. papageprgiou,T. Poggio, A training object system: car detection in static images, MIT AI Memo No. 180, 1999.

[5] H. Schneiderman, T. Kanade, A statistical method for 3D object detection applied to faces and cars, Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition 1 (2000) 746–751.

[6] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition 1 (2001) 511–518.

[7] S.Z. Li, L. Zhu, Z.Q. Zhang, A. Blake, H.J. Zhang, H. Shum, Statistical learning of multi-view face detection, Proceedings of the Seventh European Conference on Computer Vision 4 (2002) 67–81.

[8] X.R. Chen, A. Yuille, Detecting and reading text in natural scenes, Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR2004) (2004) 366–373.

[9] K.K. Sung, T. Poggio, Example-based learning for view-based human face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1) (1998) 39–50.

[10] C.J. Liu, A Bayesian discriminating features method for face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (2003) 725–740.

[11] B. Menser, F. Muller, Face detection in color images using principal component analysis, Proceedings of the Seventh International Congress on Image Processing and its Applications, 1999, pp. 13–15.

[12] M.H. Yang, N. Ahuja, D. Kriegman, Face detection using mixtures of linear subspaces, Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 70–76.

[13] A. Mohan, C. Papageorgiou, T. Poggio, Example-based object detection in images by components, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (4) (2001) 349–361.

[14] M.V. Naquest, S. Ullman, Object recognition with informative features and linear classification, Proceedings of the Ninth International Conference on Computer Vision, 2003, pp. 281–288.

[15] S. Agarwal, A. Awan, D. Roth, Learning to detect objects in images via a sparse, part-based representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (11) (2004) 1475–1490.

[16] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsurpervise scale-invariant learning, Proceedings of the Ninth International Conference on Computer Vision and Pattern Recognition 2 (2003) 264–271.

[17] B. Leibe, A. Leonardis, B. Schiele, Combined object categorization and segmentation with an implicit shape model, ECCV2004 Workshop on Statistical Learning in Computer Vision, 2004.

[18] B. Leibe, B. Schiele, Scale-invariant object categorization using a scale-adaptive mean-shift search, Proceedings of the DAGM'04 Annual Pattern Recognition Symposium, 3175, Springer LNCS, Berlin, 2004, pp. 145–153.

[19] Y.Y. Lin, T.L. Liu, Robust face detection with multi-class boosting, Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2005) 1 (2005) 680–687.

[20] Y. Amit, D. Geman, X.D. Fan, A coarse-to-fine strategy for multiclass shape detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (12) (2004) 1606–1621.

[21] F.F. Li, R. Fergus, P. Perona, A bayesian approach to unsupervised one-shot learning of object categories, Proceedings of the Ninth International Conference on Computer Vision, 2003, pp. 1134–1141.

[22] F.F. Li, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, 2004 Conference on Computer Vision and Pattern Recognition Workshop on Generative-Model Based Vision, 2004.

[23] A.Torralba, K.P. Murphy, W.T. Freeman, Sharing features: efficient boosting procedures for multiclass object detection, Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2004), 2004, pp. 762–769.

[24] Z.W. Tu, X.R. Chen, A.L. Yuille, S.C. Zhu, Image parsing: unifying segmentation, detection and recognition, International Journal of Computer Vision 63 (2) (2005) 113–140.

[25] B.Schiele, Object recognition using multidimensional receptive field histograms, PhD Thesis, I.N.P. Grenoble. English translation, 1997.

[26] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[27] K.I. Kim, K. Jung, J.H. Kim, Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (12) (2003) 1631–1639.

[28] F. Bernhard, K. Christian, Real-time face detection using edge-orientation matching, Proceedings of the Third International Conference Audio- and Video-Based Biometric Person Authentication, 2001, pp. 78–83.

[29] C. Garcia, G. Tziritas, Face detection using quantized skin color regions merging and wavelet packet analysis, IEEE Transactions on Multimedia 1 (3) (1999) 264–277.

[30] A. Hadid, M. Pietikäinen, T. Ahonen, A discriminative feature space for detecting and recognizing faces, Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, pp. 797–804.

[31] H.M. Zhang, D.B. Zhao, Spatial histogram features for face detection in color images, 5th Pacific Rim Conference on Multimedia, Lecture Notes in Computer Science 3331 (2004) 377–384.

[32] M. Pietikäinen, T. Ojala, Z. Xu, Rotation-invariant texture classification using feature distributions, Pattern Recognition 33 (2000) 43–52.

[33] T. Ojala, M. Pietikäinen, T. Mäenpä, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Macine Intelligence 24 (7) (2002) 971–987.

[34] M. Swain, D. Ballard, Color indexing, International Journal of Computer Vision 7 (1991) 11–32.

[35] Y.M. Wu, A.D. Zhang, Feature selection for classifying high-dimensional numerical data, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2004), 2004, pp. 251–258.

[36] T.W.S. Chow, D. Huang, Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information, IEEE Transactions on Neural Networks 16 (1) (2005) 213–224.

[37] C.R. Rao, Linear Statistical Inference and Its Applications, Wiley, New York, 1973.

[38] N. Kwak, C.H. Choi, Input feature selection by mutual information based on parzen window, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (12) (2002) 1667–1671.

[39] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[40] C.C. Chang, C.J. Lin, Libsvm–a library for support vector machines, www.csie.ntu.edu.tw/cjlin/libsvm, 2004.

[41] UIUC Image Database for Car Detection, http://l2r.cs.uiuc.edu/cogcomp/Data/Car/, 2004.

[42] R. Lienhart, A. Wernicked, Localizing and segmenting text in images and videos, IEEE Transactions on Circuits and Systems for Video Technology 12 (4) (2002) 236–268.

[43] X.S. Hua, W.Y. Liu, H.J. Zhang, An automatic performance evaluation protocol for video text detection algorithms, IEEE Transactions on Circuits and Systems for Video Technology 14 (4) (2004) 498–507.

[44] X.S. Hua, W.Y. Liu, H.J. Zhang, Automatic performance evaluation protocol for video text detection algorithms, Proceedings of the International Conference on Document Analysis and Recognition, 2001, pp. 545–550.