Computers and Mathematics with Applications 64 (2012) 1311–1323

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Fast pedestrian detection system with a two layer cascade of classifiers

Ying-Che Kuo*, Zih-Yi Yang, Chih-Hung Yen

Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung 411, Taiwan

ARTICLE INFO

Keywords:
Pedestrian detection
Fisher linear discriminant analysis
Haar-like feature
Covariance matrix descriptor

ABSTRACT

This work presents a novel pedestrian detection system that uses Haar-like feature extraction and a covariance matrix descriptor to identify the distinctions of pedestrians. An approach that adopts an integral image is also applied to reduce the computational loads required in both the Haar-like feature extraction and evaluation of the covariance matrix descriptor. Based on the Fisher linear discriminant analysis (FLDA) classification algorithm, the proposed system can classify pedestrians efficiently. Additionally, the detection procedure of the proposed system is accelerated using a two-layer cascade of classifiers. The front end, constructed based on Haar-like features, can select candidate regions quickly wherever pedestrians may be present. Moreover, the back end, constructed based on the covariance matrix descriptor, can determine accurately whether pedestrians are positioned in candidate regions. If a region tests positive through the two-layer cascade classifiers, pedestrian images are likely captured.

Test video sequences during the experiments are taken from a test set of the INRIA person database, using 30 input frames per second, each with a resolution of 320 × 240 pixels. Experimental results demonstrate that the proposed system can detect pedestrians efficiently and accurately, significantly contributing to efforts to develop a real time system.

1. Introduction

Most vision-based pedestrian detection systems utilize a pattern recognition scheme. The detection schemes of such detection systems can be divided into feature extraction and classifier construction. In feature extraction, the selected feature is extracted from numerous samples as a prerequisite for the subsequent classifier learning and training. A trained classifier is then oriented on how to identify features as specified from a scanned image. Such an identification approach is found to be effective for specific object recognitions, because pattern recognition is extensively adopted in diverse applications such as face recognition [1–3], pedestrian detection [4–8], vehicle detection [9] as well as object detection and classification [10–13].

In pattern recognition, pedestrian detection is a more challenging task than regular object detections, e.g. vehicle or face recognition, owing mainly to the following reasons: (1) Visions may vary with gestures, largely because a human body is not a rigid object; (2) Due to the wide variety of colors and textures; and (3) Pedestrians are in highly irregular surroundings. Therefore, of priority concern is how to implement a pedestrian detection system based on computer vision technology, capable of describing accurately the distinguishing features of pedestrians as retrieved from captured image. Numerous works on pedestrian recognition perform classifications by adopting a Haar-like feature and support vector machine (SVM) [10]. Moreover, the histogram of oriented gradient (HOG) [5], based on a histogram, has been developed as a feature extraction approach with an excellent recognition rate. Also, a covariance descriptor derived from a cross product operation of features has been implemented in recent years to perform object recognitions [14,15] and pedestrian detections [4,6]. Related performances have been thoroughly reviewed with respect to combinations of various pedestrian

* Corresponding author.
E-mail address: kuoyc@ncut.edu.tw (Y.-C. Kuo).

0898-1221/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2012.03.075

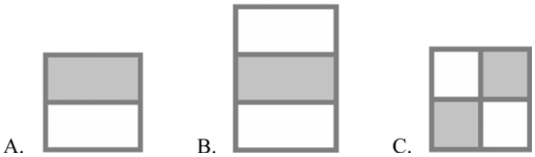


Fig. 1. Three commonly seen types of feature patterns.

feature extraction approaches (e.g. LRF [6], HOG [5,6], Haar-like [1,2], and region covariance [4,14,15]) and classification algorithms (e.g. AdaBoost [16] and SVM [17]).

Despite achieving a satisfactory recognition rate, pedestrian detection systems [4–8] are limited in that completing a task for an image size of 320 × 240 pixels takes several seconds. Of particular concern is how to embed such a detection system into a real time driver assistance system without posing risk to pedestrians passing by. Tuzel et al. [7] suggested that combining a covariance descriptor and an integral image is the most feasible means of describing pedestrian features. This is because, according to Paisitkriangkrai [4], the robustness under such a combination is maintained by using a covariance descriptor in terms of the alteration in pedestrian appearances, gestures and illuminations. Although the accuracy and the detection efficiency, as provided by a covariance descriptor combined with an AdaBoost classifier in Paisitkriangkrai [6], are somewhat lower than those provided by a combination of HOG and SVM, a superior robustness is demonstrated and a higher accuracy is maintained for a low resolution image relative to HOG. Importantly, a covariance descriptor can accelerate the computation process via an integral image evaluation. In sum, the covariance descriptor is a highly effective means of extracting distinguishing features in a pedestrian detection system.

As for efforts to reduce the computational load in a robust descriptor (e.g., a covariance and HOG), a pedestrian detector is often introduced to the front end of a pedestrian classifier, simply for the swift identification of a candidate area in order to rule out regions not containing pedestrian image. Simultaneously, the computational load of the back end is reduced as well. Similar to most other investigations, this work extracts Haar-like features in the front end of a detection system as a two layer cascade of classifiers as done so in [4], since the computational load of extracting features from a scanned image is relatively low.

Despite the acceleration of feature extractions in the front end, the operational complexity must be simplified and the overall efficiency improved as well. Also, despite their excellent recognition rate, machine learning classification algorithms (e.g., AdaBoost, SVM and neural networks) are viewed as inapplicable to a real time system, owing to an excessively long duration classification procedure. Therefore, this work adopts Fisher linear discriminant analysis [18] for detection classification, thus preserving the recognition rate but with a smaller computation load.

The proposed pedestrian detection system implements a two-layer cascade of classifiers. Largely responsible for rapid identification of candidate regions where pedestrians may be present, the front end trains a Fisher classifier through means of a Haar-like feature extraction approach combined with a Fisher linear discriminant analysis. This Fisher classifier has a detection efficiency nearly six times higher than that of an AdaBoost classifier. The back end is responsible mainly for a further confirmation that a pedestrian(s) is indeed detected in the above mentioned candidate region. In this work, experimental results demonstrate that the pedestrian recognition rate of the proposed system can reach as high as 96%.

2. Feature extraction

This work aims to speed up the detection process by use of the approach of an integral image [1] combined with a covariance descriptor [14] together with a Haar-like extraction.

2.1. Haar-like feature with integral image

As a digital image feature well applied to object recognition, a Haar-like feature is extracted out of a rectangular region, also known as a rectangular feature. Firstly proposed by Viola and Jones [2] and applied to human face recognition [1], the rectangular feature is named after Haar, for it resembles much of the Haar wavelet in the wavelet transform.

Each Haar-like feature is composed of a number of adjacent rectangles, referred to as a feature pattern. Each feature pattern consists of a white rectangular block and a grey one, and the feature of such a pattern is defined as the sum of pixel values in the white blocks minus the sum in the grey blocks. Illustrated in Fig. 1 are the commonly seen types of feature patterns. According to the number of rectangles covered, these patterns are further classified into three types. Type A refers to the two-rectangle feature (horizontally or vertically), which is defined as the difference between the sums of the pixels covered by two rectangular regions, while type B refers to the three-rectangle feature (horizontally or vertically), computing the difference between the sums covered by rectangles on both ends and the central one, and type C refers to the four-rectangle feature computing the difference between two diagonal pairs of rectangles.

In the process of Haar-like feature extractions, it is required to evaluate the sum of pixel values in individual rectangular regions. Yet, during the extraction of a number of distinct rectangular patterns, overlapped blocks are recounted over and over again. For this reason, the idea of *integral image* is adopted to promote the extraction efficiency. The corresponding

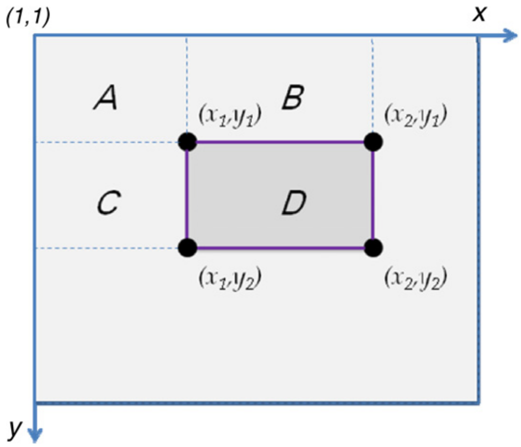


Fig. 2. Application of an integral image to a Haar-like feature extraction.

pixel value at a point (x, y) in an original image is represented as $i(x, y)$, and the integral image $ii(x, y)$ bounded between points $(1, 1)$ and (x, y) is defined as

$$ii(x, y) = \sum_{x' \le x, y' \le y} i(x', y'). \tag{1}$$

Setting $s(x, -1) = 0$ and $ii(-1, y) = 0$, with $s(x, y)$ representing the sum of the pixel values along the y axis in the original image $i(x, y)$, then the integral is derived as a solution to a set of recursive Eqs. (2) and (3).

$$s(x, y) = s(x, y - 1) + i(x, y) \tag{2}$$

$$ii(x, y) = ii(x - 1, y) + s(x, y). \tag{3}$$

As illustrated in Fig. 2, following the calculation of an integral image, the sum of the pixel values within the rectangular region D can be evaluated by the values at points (x_1, y_1), (x_1, y_2), (x_2, y_1), and (x_2, y_2). The value of the integral image at (x_2, y_2) (denoted as $ii(x_2, y_2)$) is the sum of pixel values covered by rectangles A, B, C and D, while the value of the integral image at (x_1, y_1) is that by the rectangle A, and the value of integral image at (x_1, y_2) is that by the rectangles A and C. Hence, the sum of the pixel values contained by D is expressed as Eq. (4).

$$D = ii(x_2, y_2) + ii(x_1, y_1) - [ii(x_2, y_1) + ii(x_1, y_2)]. \tag{4}$$

Through the approach of the integral image, it merely takes $M \times N \times 2$ iterations to evaluate all the integral images $ii(x, y)$ from an image of size $M \times N$ pixels, following which the sum of the pixel values in an arbitrarily sized rectangle can be computed involving as little effort as simple arithmetic operations. In so doing, the recount problem on overlapped blocks is thus removed, i.e. it is an approach speeding up the Haar-like extraction process by far.

2.2. Region covariance with integral image

Proposed by Tuzel et al. [14], a covariance matrix descriptor, also referred to as a region covariance, was applied to object detection as well as texture recognition. Such a descriptor is suggested as an effective way to describe mutual relations by means of a covariance between various image features in a raw image. Aimed at pedestrian detection, this work adopts a total of eight features, constituting a feature matrix as $\left[x \quad y \quad |I_x| \quad |I_y| \quad \sqrt{I_x^2 + I_y^2} \quad |I_{xx}| \quad |I_{yy}| \quad \tan^{-1} \frac{|I_x|}{|I_y|} t \right]$, where I represents an original image, (x, y) the coordinates of a pixel, $|I_x|$ and $|I_y|$ the first order derivatives with respect to x and y respectively, $|I_{xx}|$ and $|I_{yy}|$ the second order derivatives with respect to x and y respectively, $\sqrt{I_x^2 + I_y^2}$ the image gradient strength, and $\tan^{-1} \frac{|I_x|}{|I_y|}$ the gradient direction.

In most cases, a covariance descriptor is in essence a region descriptor since a small rectangular region of interest, named as $R(1, 1; W, H)$ which is bounded by up-left coordinate $(1, 1)$ and down-right coordinate (W, H) in an image, is involved in the computation process. The featured image corresponding to a specific feature i extracted out of $R(1, 1; W, H)$ is denoted as f_i, $i = 1, 2, \ldots, D$. Moreover, the featured image f_i is composed of $f_{x,y}(i)$, $1 \le x \le W$, $1 \le y \le H$. $f_{x,y}(i)$ is the pixel value in the featured image. As expressed in Eq. (5), μ symbolizes the vector composed of $\mu(i)$, the mean value of respective featured images. As written in Eq. (6), $F_{x,y}$ signifies the vector made up of D number of features pertaining to the pixel in (x, y)-coordinate of $R(1, 1; W, H)$.

Hence, a covariance matrix $C_{R(1,1;W,H)}$ associated with a rectangle $R(1,1;W,H)$ in an image is defined as Eq. (7). Accordingly, the covariance matrix $C_{R(1,1;W,H)}$ is a matrix of dimension $D \times D$, where $D = 8$ denotes the number of features extracted in this work.

$$\mu = [\mu(1), \mu(2), \mu(3), \ldots, \mu(D)] \tag{5}$$

where $\mu(i) = E[f_i] = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H} f_{x,y}(i)$

$$F_{x,y} = [f_{x,y}(1), f_{x,y}(2), f_{x,y}(3), \ldots, f_{x,y}(D)] \tag{6}$$

$$C_{R(1,1;W,H)} = \frac{1}{WH-1}\sum_{x=1}^{W}\sum_{y=1}^{H}(F_{x,y}-\mu)^T(F_{x,y}-\mu). \tag{7}$$

Very analogous to the case of the Haar-like feature evaluation, it is very likely that: (1) the problem of re-evaluating the mean, or the sum, of pixel values cannot be obviated, and (2) the possibility of overlapped rectangular regions (feature patterns) cannot be excluded. Though the covariance descriptor necessitates complicated operations, the operation efficiency can be considerably promoted in the event that an integral image approach, as referred to previously, can be applied to this case.

In a bid to evaluate a covariance descriptor by means of an integral image, referring to Eq. (7), the (i,j) entry, $C_{R(1,1;W,H)}(i,j)$, of the covariance matrix is defined as Eq. (8).

$$C_{R(1,1;W,H)}(i,j) = \frac{1}{WH-1}\sum_{x=1}^{W}\sum_{y=1}^{H}(f_{x,y}(i)-\mu(i))(f_{x,y}(j)-\mu(j)) \tag{8}$$

where $1 \le i, j \le D$.

Eq. (8) is expanded and rearranged into Eq. (9).

$$C_{R(1,1;W,H)}(i,j) = \frac{1}{WH-1}\left\{\sum_{x=1}^{W}\sum_{y=1}^{H}f_{x,y}(i)f_{x,y}(j) - \frac{1}{WH}\left(\sum_{x=1}^{W}\sum_{y=1}^{H}f_{x,y}(i)\right)\left(\sum_{x=1}^{W}\sum_{y=1}^{H}f_{x,y}(j)\right)\right\}. \tag{9}$$

It is found from Eq. (9) that ahead of the evaluation of the covariance matrix in $R(1,1;W,H)$, both the integral image $P_{X,Y}(i)$ of $\{f_{x,y}(i)\}_{i=1,2,3,\ldots,D}$, namely all the featured image components, and the integral image $Q_{X,Y}(i,j)$ of $\{f_{x,y}(i)f_{x,y}(j)\}_{i,j=1,2,3,\ldots,D}$, namely the product of two arbitrary featured image components, must be evaluated. Integral images are respectively expressed as Eq. (10) and Eq. (11).

$$P_{X,Y}(i) = \sum_{x=1}^{X}\sum_{y=1}^{Y}f_{x,y}(i) \quad 1 \le i,\ j \le D \tag{10}$$

$$Q_{X,Y}(i,j) = \sum_{x=1}^{X}\sum_{y=1}^{Y}f_{x,y}(i)f_{x,y}(j) \quad 1 \le i,\ j \le D. \tag{11}$$

Derived from Eqs. (10) and (11), $P_{x,y}(i)$ is converted into a D dimensional vector $P_{x,y}$ as given in Eq. (11), and $Q_{x,y}(i,j)$ into a matrix $Q_{x,y}$ of dimension $D \times D$ as in Eq. (13).

$$P_{x,y} = [P_{x,y}(1), P_{x,y}(2), l \ldots, P_{x,y}(D)] \tag{12}$$

$$Q_{x,y} = \begin{bmatrix} Q_{x,y}(1,1), & Q_{x,y}(1,2), & \ldots, & Q_{x,y}(1,D) \\ Q_{x,y}(2,1), & Q_{x,y}(2,2), & \ldots, & Q_{x,y}(2,D) \\ \vdots & & \ddots & \vdots \\ Q_{x,y}(D,1), & Q_{x,y}(D,2), & \ldots, & Q_{x,y}(D,D) \end{bmatrix}. \tag{13}$$

Substitution of Eqs. (12) and (13) back into Eq. (9) gives $C_{R(1,1;X,Y)}$, as expressed in Eq. (14), the covariance descriptor associated with the rectangle $R(1,1;X,Y)$ defined by the up-left coordinate $(1,1)$ and the down-right coordinate (X,Y) in an image.

$$C_{R(1,1;X,Y)} = \frac{1}{XY-1}\left[Q_{X,Y} - \frac{1}{XY}\left(P_{X,Y}\right)^T P_{X,Y}\right]. \tag{14}$$

Just as in the Haar-like feature extraction, all it takes to get $C_{R(x_1,y_1;x_2,y_2)}$, as given in Eq. (15), is merely to locate four reference points in a featured image and perform simple arithmetic operations for an arbitrary sized and position $R(x_1,y_1;x_2,y_2)$.

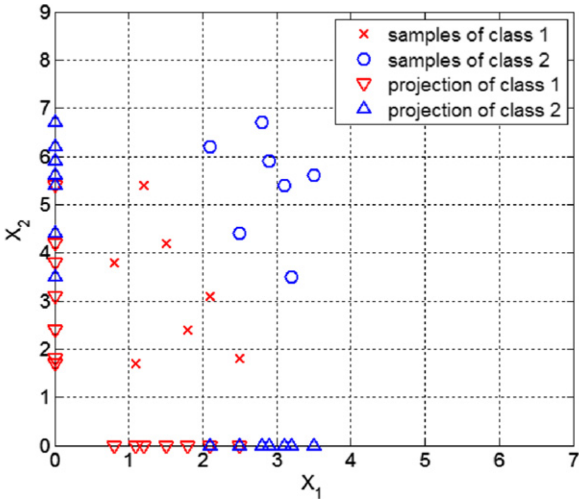


Fig. 3. Indistinguishable projected samples onto the $\mathbf{X_1}$ and $\mathbf{X_2}$ axes.

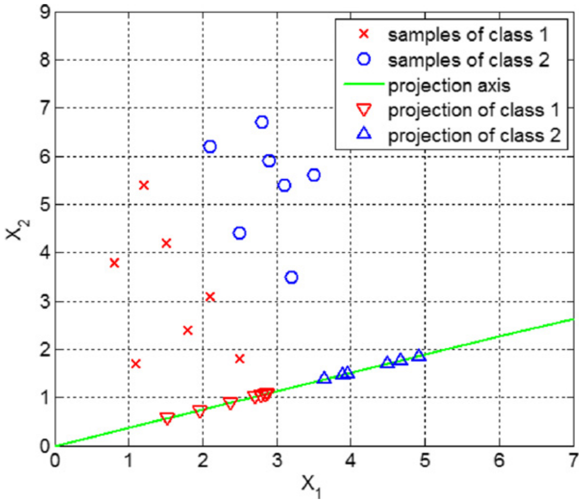


Fig. 4. Projected samples with the optimal separability through the optimal projection vector.

$$C_{R(X1,Y1;X2,Y2)} = \frac{1}{(X2-X1)\,(Y2-Y1)-1}\Bigg[Q_{X2,Y2}+Q_{X1,Y1}-Q_{X2,Y1}-Q_{X1,Y2}-\frac{1}{(X2-X1)\,(Y2-Y1)} \times \left(P_{X2,Y2}+P_{X1,Y1}-P_{X2,Y1}-P_{X1,Y2}\right)^T\left(P_{X2,Y2}+P_{X1,Y1}-P_{X2,Y1}-P_{X1,Y2}\right)\Bigg]. \tag{15}$$

3. Fisher linear discriminant analysis

Proposed by Fisher [18], the Fisher linear discriminant analysis (FLDA) refers to a technology of group separation, mapping a high dimensional pattern sample onto a one dimensional space with the highest recognizability. It is ensured that the new subspace is of the maximum "between classes distance" and the minimum "within classes distance". Consequently, a pattern sample acquires the optimum separability.

By mapping, a high dimensional vector can be compressed into a one dimensional scalar. Yet, such mapping will inevitably make distinctive samples indistinguishable, as presented in Fig. 3, unless they are properly mapped.

As demonstrated in Fig. 4, samples are as intended mapped onto the optimal projection axis through FLDA, that is, the optimal separability is demonstrated.

In other words, the issue that FLDA addresses is to search for an optimal projection vector W_{opt} out of samples in pursuit of the optimal separability in the mapped 1-D domain.

FLDA is represented as Eq. (16),

$$J(\phi) = \frac{\phi^T S_B \phi}{\phi^T S_W \phi} \tag{16}$$

where ϕ represents a n-dimensional vector, S_B the "between classes scatter matrix" and S_W the "within classes scatter matrix". Scatter matrices are respectively defined as Eq. (17) and Eq. (18).

$$S_B = \sum_{i=1}^{c} n_i (u_i - \bar{x})(u_i - \bar{x})^T \tag{17}$$

$$S_W = \sum_{i=1}^{c} \sum_{x_k \in \text{class}_i} (u_i - x_k)(u_i - x_k)^T \tag{18}$$

where $\bar{x}$ denotes the overall mean of the sample cases, u_i the mean of the i class, c the number of the class, n_i the quantity of samples of the i class and x_k the k-th sample.

It is known from Eq. (16) that ϕ is identically the optimal projection vector W_{opt} when $J(\phi)$ is maximized. As the first step to find W_{opt}, both the eigenvalue and the eigenvector are given as Eq. (19).

$$[V, D] = \mathbf{eig}(S_W^{-1} S_B) \tag{19}$$

where V and D represent the eigenvalue and the eigenvector of $S_W^{-1} S_B$ respectively. Then W_{opt} is evaluated as the eigenvector corresponding to the maximum eigenvalue.

The Fisher classifier $F(X)$ employed in this work is defined as Eq. (20).

$$F(X) = \begin{cases} \text{class}_1, & (W_{\text{opt}})^T X > \theta \\ \text{class}_2, & (W_{\text{opt}})^T X < \theta \end{cases} \tag{20}$$

where W_{opt} denotes the optimal projection vector, X the input sample, and θ the classifier threshold defined as Eq. (21).

$$\theta = \frac{n_1 \tilde{u}_{1+} n_2 \tilde{u}_2}{n_1 + n_2} \tag{21}$$

n_i the number of samples with class i, $\tilde{u}_i$ is the mean of samples with classes i.

4. Pedestrian detection system

For the implementation of an efficient pedestrian detection system, Haar-like features and a covariance descriptor are adopted as the extraction approach. A two layer cascade of classifiers is designed to detect pedestrians for two distinct feature extraction approaches using the Fisher linear discriminant analysis. The front end, constructed based on Haar-like feature extraction, is made able to swiftly select candidate regions where pedestrians may be present, and the back end, constructed based on the covariance matrix descriptor, is able to accurately see whether or not there are pedestrians positioned in candidate regions. If such a candidate region is tested positive, a pedestrian is then located.

4.1. Pedestrian image data set

In this work, the pedestrian image data base is the one taken out of INRIA [19], where there are 2416 positive and 1218 negative samples in the training set, and 1126 positive and 453 negative samples in the test set.

4.2. Select the area for feature extraction

(1) *Haar-like feature*: A pedestrian featured image model is built in this work as an effective auxiliary in the determination of a Haar-like feature extraction frame. As illustrated in Fig. 5(a), a pedestrian featured image of 64 × 128 pixels is evaluated as the mean of all the 2416 trained positive samples. Such an image is then segmented into 8 × 16 square cells in 8 × 8 pixels, following which the pedestrian featured image model, presented in Fig. 5(b), is built as the average pixel values in each square cell.

For the purpose of extracting key features out of the model in Fig. 5(b), three types of Haar-like rectangular filters employed, shown in Fig. 6, are oriented in this same way as the rectangular feature patterns adopted. The goal is simply to recognize distinctive features horizontally, vertically and diagonally, according to which the feature patterns can be properly sized and positioned.

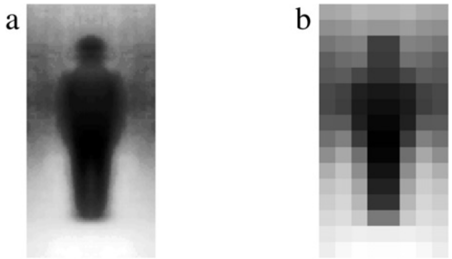


Fig. 5. Feature extraction in a pedestrian image model.

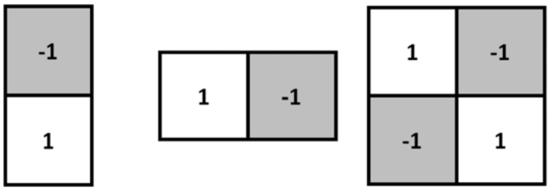


Fig. 6. Filters along the vertical, the horizontal and diagonal directions.

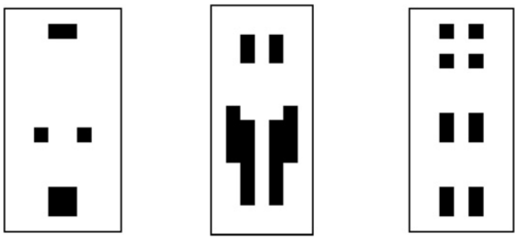


Fig. 7. Distinctive feature spots.

As illustrated in Fig. 7, the locations of steep variation in gradient in Fig. 5(b) are marked black. That is, such pixels represent the spots of distinctive features.

As a result, there are as many as 12 four-rectangular Haar-like features, together with five and six rectangular features along the vertical and the horizontal directions respectively in this work and shown in Fig. 8.

(2) *Covariance descriptor*: A covariance descriptor is formed as extractions of covariance matrices, containing mutual relations between a great number of featured images, edge information, gradient direction, and filter responses of an input image. A covariance descriptor is said to be an appearance based feature detector, simply due to the reason that the featured images, as stated above, are referred to as appearance-based information.

Accordingly, in the process of feature extractions for a covariance descriptor, distinctive parts in pedestrian appearance are employed, such as the entire body, a combination of the head and shoulders, hands and the main part of human body, and legs. Presented in Fig. 9 are the five feature extraction ranges selected in this work for a covariance descriptor.

4.3. Training process

On the basis of the images captured in a detection window, training and recognition of the pedestrian detection system are conducted in two distinct ways to extract features, i.e. Haar-like feature extraction and a covariance descriptor approach. That is, it necessitates two distinct approaches in the training process of a Fisher classifier.

(1) *Haar-like based classifier training processes*: As illustrated in Fig. 10, a Haar-like based classifier is trained as follows.

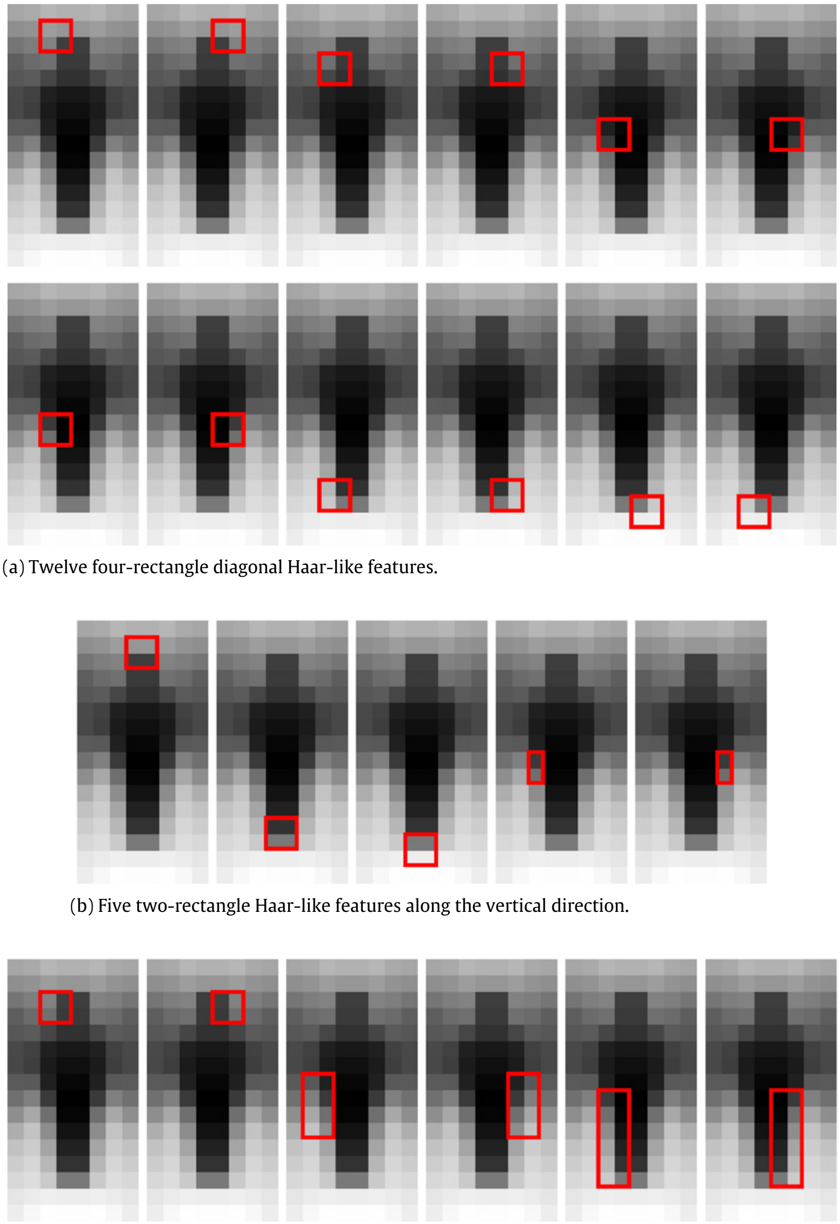
(a) Evaluate the integral images of the positive and negative images taken out of a sample set,

(b) within the integral image of a positive sample, an integral image of size 64 × 128 pixels is selected out of the pedestrian location, while within that of a negative sample, an identical sized image is selected in an arbitrary position,

(c) evaluate all the Haar-like features, a 23-dimensional vector, since there are 23 distinct feature patterns,

(d) repeat steps (a)–(c) until the required number of training samples are collected, and

(e) perform the Fisher linear discriminant analysis on all the 23-dimensional Haar-like features, following which an optimal projection vector W_{opt} as well as a classifier threshold θ is gained.



(a) Twelve four-rectangle diagonal Haar-like features.

(b) Five two-rectangle Haar-like features along the vertical direction.

(c) Six two-rectangle Haar-like features along the horizontal direction.

Fig. 8. Twenty-three Haar-like feature regions selected for filtering pedestrian.

(2) *Covariance descriptor based classifier training processes*: The information is represented as a 23-dimensional vector in a detection window of a Haar-like feature classifier. Due to the fact that a 36-dimensional vector is generated for each feature extraction frame by the classifier of a covariance descriptor, there is a total of five 36-dimensional vectors denoting the data contained in a detection window on account of five covariance descriptor employed within the extraction range. As illustrated in Fig. 11, the training process is stated as follows.

(a) An image of size 64 × 128 pixels is extracted at the pedestrian position out of a positive row image, while the same task is repeated at an arbitrary location out of a negative raw image,

(b) by means of Eq. (5), both the raw images extracted in step a are converted into featured images, including (x, y) coordinates, the absolute value, $|I_x|$, of the first order partial derivative with respect to x, the absolute value, $|I_y|$, of the first order partial derivative with respect to y, the absolute value, $|I_{xx}|$, of the second order partial derivative with respect to x, the absolute value, $|I_{yy}|$, of the second order partial derivative with respect to y, the gradient strength $\sqrt{I_x^2 + I_y^2}$ and the gradient direction $\tan^{-1}\frac{|I_x|}{|I_y|}$,

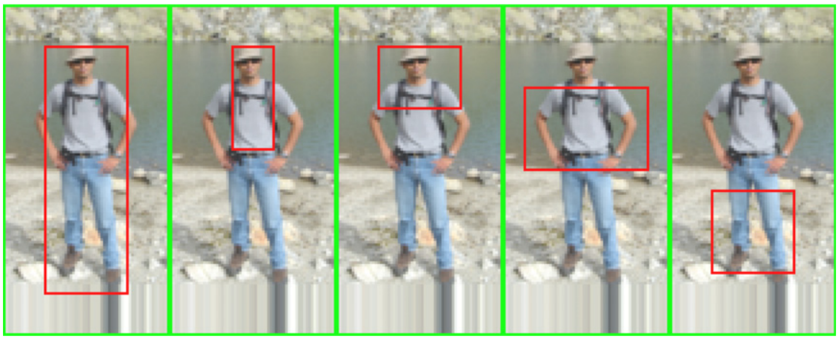


Fig. 9. Five feature extraction ranges for covariance descriptor.

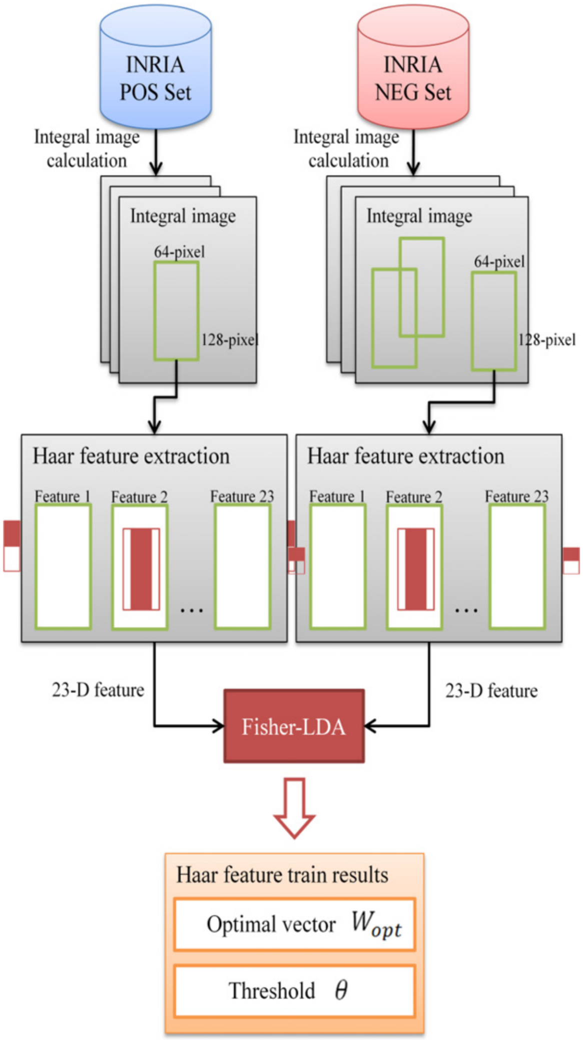


Fig. 10. A Haar-like based classifier training process.

(c) evaluate individual covariance matrices of dimension 8 × 8 built according to five feature extraction ranges within a covariance descriptor. Since such covariance matrices are symmetric, merely the entries above the main diagonal are taken during matrix evaluations. Thus, each sample's feature is represented by five 36-dimensional vectors,

(d) repeat steps (a)–(c) until the required number of training samples are collected, and

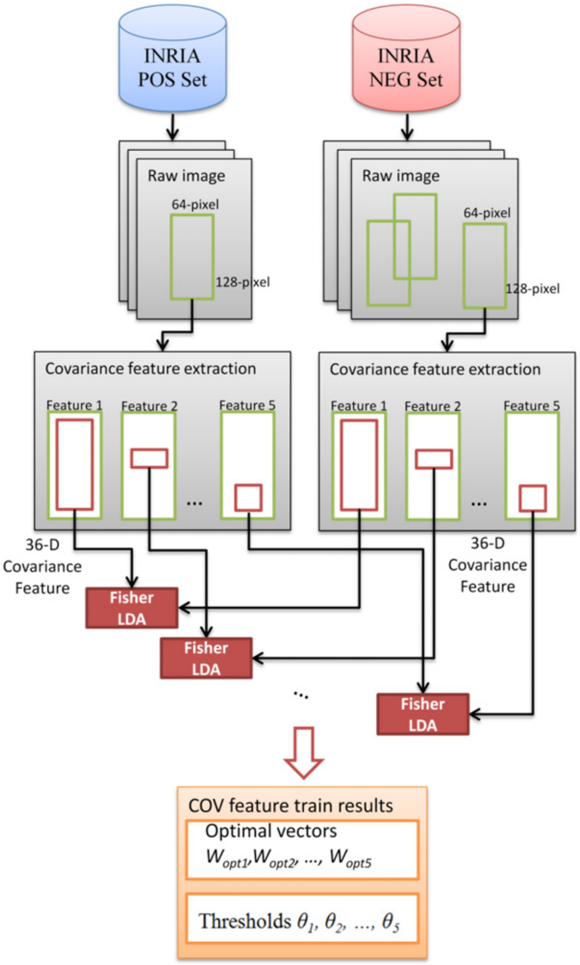


Fig. 11. A covariance descriptor based classifier training process.

(e) perform in the end the Fisher linear discriminant analysis on all the covariance descriptors associated with respective feature extraction ranges, following which five sets of optimal projection vector W_{opt} as well as a classifier threshold θ are found.

4.4. Cascade FLDA classifier

A Haar-like feature classifier is concatenated with the other, namely a region covariance feature classifier, such that an accurate and efficient pedestrian detection is reached. The former is designed to be able to identify candidate areas that may contain pedestrians, while the latter is able to accurately recognize the pedestrian images therein. The detection result is affirmative if a block is tested as containing pedestrians by such two layer classifiers.

The aspect ratio of 1:2 in a detection window is made in practical applications, for the reason that an image of 14 × 128 pixels is employed to train a classifier. The detection window increases the width from 40 to 80 pixels with an eight pixel increment, shifts each time by one-quarter of the window width, and scans the entire image from left to right and then up to down. Illustrated in Fig. 12 is the pedestrian detection process where all the images contained in each detection window are applied to classifiers for test, and the coordinates as well as the dimensions of such a detection window are recorded in the event that a pedestrian is considered detected. Subsequently, overlapped detection windows merged are displayed on the raw image.

5. Experiments and results

The detection efficiency in this work is compared with those stated in [4]. There is a review on the detailed comparison between various extraction approaches, i.e. the Haar-like feature approach, the region covariance descriptor, HOG and LRF, and on the architecture comparison between AdaBoost, SVM and neural network. Besides, the classifier structure with the highest efficiency is suggested as a reference for successive detection system works.

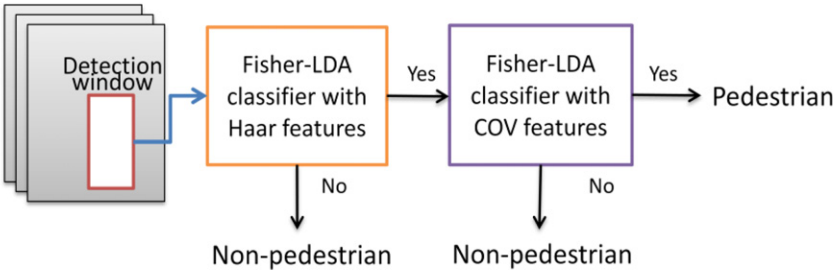

Detection window
Fisher-LDA classifier with Haar features
Yes
Fisher-LDA classifier with COV features
Yes
Pedestrian
No
No
Non-pedestrian
Non-pedestrian

Fig. 12. A pedestrian detection process.

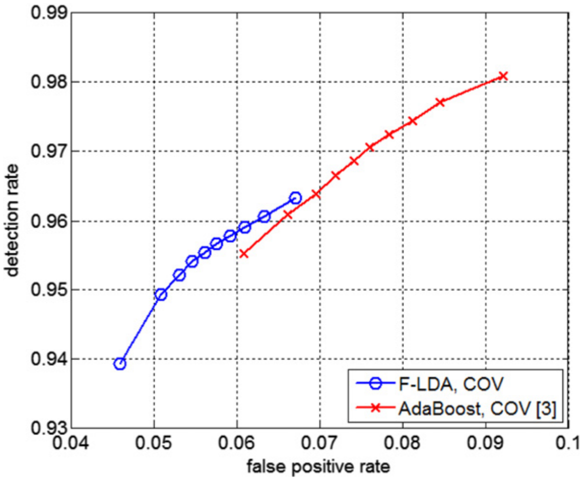


Fig. 13. Performance comparison between the proposed Fisher classifier and the AdaBoost classifier.

Table 1
Detection efficiency comparison between combinations of various feature extractions and classifiers.

	Number of detection windows per second
Haar-like, F-LDA	2640
Haar-like, AdaBoost	461
COV feature, F-LDA	546
COV feature, AdaBoost	260

As the first step of conducting experiments, by use of the INRIA person data base, classifiers are trained by positive samples, taken from a training set, and random samples, selected out of a negative sample training set. Subsequently, positive samples, taken from a test set, and negative random samples, selected out of a negative sample test set, are tested. Each time up to 2000 images are randomly selected out of a positive sample training set in the INRIA person data base, and an identical number of negative samples are randomly selected out of a negative sample training set as well. To be fair, the Fisher and AdaBoost classifiers are trained with the same set of samples.

Plotted in Fig. 13 is the comparison on 2000 experimental results conducted by a Fisher classifier and an AdaBoost classifier, both built based on a covariance descriptor.

As can be seen from Fig. 13, the Fisher classifier is of a detection rate between 0.94 and 0.964 and a false positive rate between 0.046 and 0.068, while the AdaBoost classifier [4] is of a detection rate between 0.955 and 0.981, and a false positive rate between 0.061 and 0.092. A superior performance is represented by the curve toward the upper left corner. Albeit the AdaBoost classifier provides a detection rate up to 0.981, a figure much higher than that provided by the Fisher classifier, two such classifiers render comparable performances.

Tabulated in Table 1 are performance comparisons between various combinations of feature extraction approaches and classifiers. Implemented in MATLAB, all the experiments, as tabulated in Table 1, are conducted on the same PC, equipped with an Intel Core Duo processor at 1.83 GHz and with 2 GB memory. It is noteworthy that the FLDA based classifier, combined either with the Haar-like feature extraction or with a covariance descriptor, demonstrates a detection efficiency twice as high as that provided by the AdaBoost classifier. In simple terms, a superior system detection efficiency is reached at the cost of a marginal detection accuracy.

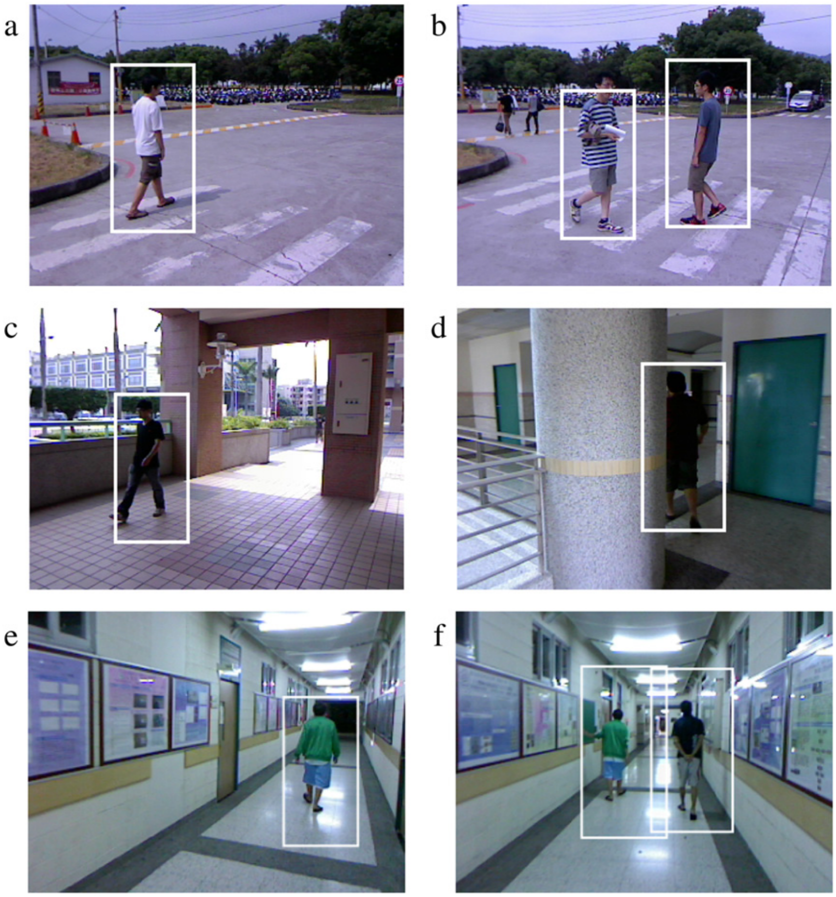


Fig. 14. Detection results with various surroundings and in various illuminations.

Presented in Fig. 14 are photos taken by a camera on the move. Marked in Fig. 14(a) and (b) are pedestrians in regular circumstances, while in (c) and (d) are those in shadowed conditions, and in (e) and (f) are those in non-uniform illumination. Some detection results of photos of the USC Pedestrian Detection Test Set [20] are also presented in Fig. 15. Based on the detection results, the proposed system is demonstrated to be efficient as well as accurate in pedestrian detection.

6. Conclusions

This work presents a novel pedestrian detection system. The proposed system consists of two modules, the front end (i.e. a selection module for a candidate region) and a back end (i.e. a pedestrian classification module). Largely responsible for rapid identification of candidate regions, the front end trains the Fisher classifier through means of a Haar-like feature extraction approach combined with Fisher linear discriminant analysis. This Fisher classifier is nearly six times more efficient than an AdaBoost classifier in terms of detection efficiency. The back end is responsible mainly for a further confirmation that a pedestrian(s) is indeed detected in the above mentioned candidate region. The pedestrian classification module consists of five Fisher classifiers trained by a covariance descriptor together with the Fisher linear discriminant analysis. Just as in the case of the front end, such a back end provides a classification efficiency twice as high as that of an AdaBoost classifier. Experimental results demonstrate that the Fisher classifier has an efficiency comparable to that of the AdaBoost one, even providing a higher efficiency in a specific context.

Although the robustness of pedestrian detection is maintained to a certain degree against illumination variation, recognition accuracy degrades somewhat in dim light. This discrepancy is simply because a nearly zero difference between the sums of the pixel values covered by neighboring rectangles is found during a Haar-like feature extraction. Moreover, evaluation in covariance descriptor via the featured image, based on which quantities (e.g., first as well as second partial derivatives, gradient strength, and gradient direction) cannot be determined accurately, simply for the same reason as in the case of Haar-like feature extraction. Efforts are underway in our laboratory to examine the feasibility of an IR camera not only as a solution to pedestrian detection in shadowed conditions, but also as a means of improving the recognition rate in high illumination.

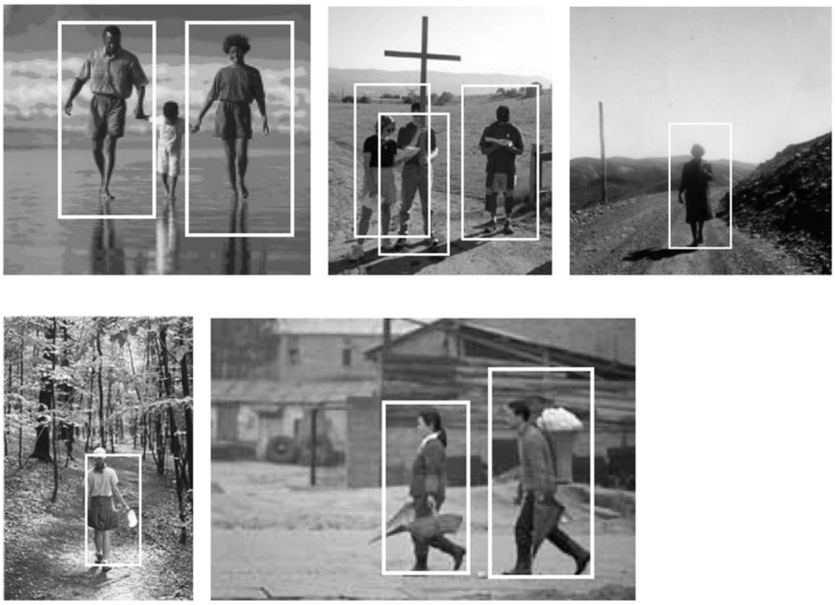


Fig. 15. Detection results of photos of the USC Pedestrian Detection Test Set.

Acknowledgments

The authors would like to thank the National Science Council of the Republic of China, Taiwan for financially supporting this research under Contract No. NSC 99-2221-E-167-028.

References

[1] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.

[2] P. Viola, M.J. Jones, Rapid object detection using a boosted cascade of simple features, in: IEEE Int'l Conf. on Computer Vision and Pattern Recognition, Kauai, Hawaii, 2001.

[3] N.S. Pai, S.P. Chang, An embedded system for real-time facial expression recognition based on the extension theory, Int. J. Comput. Math. Appl. 61 (8) (2011) 2101–2106.

[4] S. Paisitkriangkrai, C. Shen, J. Zhang, Fast pedestrian detection using a cascade of boosted covariance features, IEEE Trans. Circuits Syst. Video Technol. 18 (8) (2008) 1140–1151.

[5] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Int'l Conf. on Computer Vision and Pattern Recognition, 2005, pp. 886–893.

[6] S. Paisitkriangkrai, C. Shen, J. Zhang, Performance evaluation of local features in human classification and detection, IET Comput. Vis. 2 (4) (2008) 236–246.

[7] O. Tuzel, F. Porikli, P. Meer, Human detection via classification on Riemannian manifolds, in: IEEE Int'l Conf. on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[8] S. Munder, D.M. Gavrila, An experimental study on pedestrian classification, IEEE Trans. Pattern Anal. Mach. Intell. 28 (11) (2006) 1863–1868.

[9] Y.C. Kuo, N.S. Pai, Y.F. Li, Vision-based vehicle detection for a driver assistance system, Int. J. Comput. Math. Appl. 61 (8) (2011) 2096–2100.

[10] C.P. Papageorgiou, M. Oren, T. Poggio, A general framework for object detection, in: Int'l Conf. on Computer Vision, 1998.

[11] R.A. Fisher, The use of multiple measures in taxonomic problems, Ann. Eugenics 7 (1936) 179–188.

[12] Z. Xue, Y. Shang, A. Feng, Semi-supervised outlier detection based on fuzzy rough *C*-means clustering, Int. J. Math. Comput. Simul. 80 (9) (2010) 1911–1921.

[13] W. Chen, T. Liu, B. Wang, Ultrasonic image classification based on support vector machine with two independent component features, Int. J. Comput. Math. Appl. 62 (7) (2011) 2696–2703.

[14] O. Tuzel, F. Porikli, P. Meer, Region covariance: a fast descriptor for detection and classification, in: Proc. Eur. Conf. Comput. Vis., 2006, pp. 589–600.

[15] Y.W. Pang, Y. Yuan, X.L. Li, Gabor-based region covariance matrices for face recognition, IEEE Trans. Circuits Syst. Video Technol. 18 (7) (2008) 989–993.

[16] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. System Sci. 55 (1) (1997) 119–139.

[17] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995).

[18] R.A. Fisher, The use of multiple measures in taxonomic problems, Ann. Eugenics 7 (1936) 179–188.

[19] INRIA person dataset. http://pascal.inrialpes.fr/data/human/.

[20] USC Pedestrian Detection Test Set. http://iris.usc.edu/Vision-Users/OldUsers/bowu/DatasetWebpage/dataset.html.