

Pedestrian Detection in Video Images via Error Correcting Output Code Classification of Manifold Subclasses

Qixiang Ye, *Member, IEEE*, Jixiang Liang, and Jianbin Jiao, *Member, IEEE*

Abstract—Pedestrian detection in images and video frames is challenged by the view and posture problem. In this paper, we propose a new pedestrian detection approach by error correcting output code (ECOC) classification of manifold subclasses. The motivation is that pedestrians across views and postures form a manifold and that the ECOC method constructs a nonlinear classification boundary that can discriminate the manifold from negative samples. The pedestrian manifold is first constructed with a local linear embedding algorithm and then divided into subclasses with a K -means clustering algorithm. The neighboring relationships of these subclasses are used to make the encoding rule for ECOCs, which we use to train multiple base classifiers with histogram of oriented gradient features and linear support vector machines. In the detection procedure, image windows are tested with all base classifiers, and their output codes are fed into an ECOC decoding procedure to decide whether it is a pedestrian or not. Experiments on three data sets show that the results of our approach improve the state of the art.

Index Terms—Error correcting output code (ECOC), manifold, pedestrian detection, support vector machine (SVM).

I. INTRODUCTION

DETECTION of pedestrians has attracted considerable attention in a wide variety of applications, such as intelligent video surveillance and pedestrian warning for driving assistance [1]–[5]. In recent years, the research of pedestrian detection has achieved some success in video surveillance systems, where the camera has a fixed viewpoint and captures a static background. However, pedestrian detection in driving warning systems is still an open problem because of the moving camera, complex backgrounds, varied illumination conditions in outdoor environments, and, in particular, a broad range of pedestrian views and postures.

Manuscript received January 24, 2011; revised May 12, 2011, July 3, 2011, and August 17, 2011; accepted August 27, 2011. Date of publication September 25, 2011; date of current version March 5, 2012. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2011CB706900 and Grant 2010CB731800 and by the National Science Foundation of China under Grant 60872143 and Grant 61039003. The Associate Editor for this paper was S. Sun.

Q. Ye and J. Jiao are with the College of Engineering and the Center of Engineering Technology, Graduate University of the Chinese Academy of Sciences, Beijing 100049, China.

J. Liang is with the College of Engineering, Graduate University of the Chinese Academy of Sciences, Beijing 100049, China.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TITS.2011.2167145

In existing pedestrian detection systems, feature representation and classifiers are two main problems being investigated.

The computational approach to vision by Maier claims that the primitives of visual information representation are simple components of forms and their local properties. Therefore, local features are most often investigated for pedestrian detection. These features include Haar-like features [5], histogram of oriented gradient (HOG), v-HOG features [13], [14], Gabor filter-based cortex features [15], covariance features [16], local binary pattern (LBP) features [17], HOG–LBP features [18], edgelet features [19], shapelet features [20], local receptive field features [21], multiscale orientation features [35], etc. A recent survey [3] has shown that, of the proposed features, various HOG features are most effective for pedestrian detection.

By using local statistics, the HOG features are robust to complex background and even robust to significant occlusion when using a part-based model [36]. There are also some improvements of HOG features for pedestrian detection, such as the combination with LBP features [18], the extension to nonrectangle blocks [37], etc. HOG features are also extended to other applications, such as the facial expression recognition [38]. In [39], Kamijo *et al.* use HOG features and a cascade of classifiers to parallel detect pedestrians in a multiple camera framework. The usage of multiple cameras can improve the view range of the system but cannot improve the detection performance of either the multiview or the multiposture pedestrians since when the parallel detection improves the detection rate, they also bring more false alarms to the system. In this paper, using HOG features as a representation, we discuss the problem of how to detect multiview and multiposture pedestrians more effectively; no matter, they are captured by one single camera or multiple cameras.

The extracted features on labeled samples are usually fed into a classifier to learn detection models. In the classifiers, the linear support vector machine (SVM) is the most popular classifier [13], [26]. Its combination with boost algorithm, such as MPLBoost [30], demonstrates the state-of-the-art performance. However, when we need to detect multiview and multiposture pedestrians in a system, linear SVMs are challenged. It is observed in experiments that pedestrians of continuous view and posture variation form a manifold, which is difficult to classify linearly from the negatives. The method requires multiview and multiposture pedestrians to be correctly classified with a linear SVM in the training process, often leading to overfitting. Some nonlinear classification methods such as kernel SVMs [13],

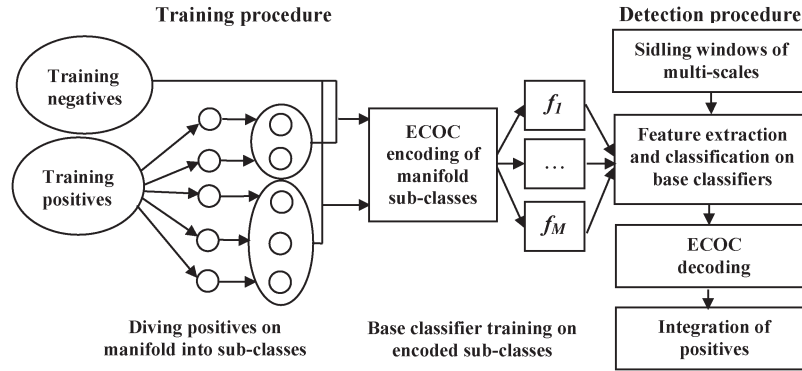


Fig. 1. Training and detection procedures of the proposed pedestrian detection approach.

[31] are options to this problem, but these methods are generally much more computationally expensive than linear methods. In addition, selection of a proper kernel for very high dimensional samples is not a trivial task.

On the other hand, some approaches use a “divide-and-conquer” strategy to deal with the view and posture problem by first dividing training positives into subclasses and then training or designing multiclass models for detection [6]–[9], [23]–[25]. In [9] and [37], tree structure boosting classifiers are developed to detect multiview pedestrians in images. Although they are partially successful, as far as classifier learning is concerned, those methods consider little about the relation among subclasses and cannot always reduce both bias and variance of the learned classification models.

Error correcting output codes (ECOCs) were created as a general framework to handle multiclass problems [10], [11]. The classification is performed according to a set of binary ECOCs. By introducing suitable coding rules, ECOCs can reach a nonlinear classification while reducing both bias and variance of the learned classification models. In many applications, the ECOC framework is justified to be simple but more effective than other multiclass extensions [10]–[12].

In this paper, we formulate the multiview and multiposture pedestrian detection as a manifold classification problem. The manifold learning is first employed to cluster the pedestrian samples into several subclasses, each of which is more compact than the original class and then can be well modeled with a linear classifier. Then, the ECOC is used to encode the relationships among the neighboring subclasses obtained by manifold learning. One or multiple subclasses together with the negative class are modeled with a base classifier, and all base classifiers are integrated by ECOC coding to perform the final nonlinear classification. To the best of our knowledge, this is the first time that ECOC classification has been applied to pedestrian detection. This is also the main difference between our approach with that of [16]. In [16], extracted pedestrian features are transformed into a manifold tangent space of manifold for classification. The authors reported that classification in the tangent space can reduce the effect of views and postures.

The contributions of this paper are summarized as follows: The multiview and multiposture pedestrian detection problem is converted to a manifold subclass classification problem. A simple but effective solution is proposed by introducing manifold-advised ECOC classification.

The remainder of this paper is organized as follows: The methodology for pedestrian detection is presented in Section II. Experimental results are provided in Section III, and conclusions are made in Section IV.

II. METHODOLOGY

In this section, we first present an overview of the proposed pedestrian detection approach and then describe the feature representation, the construction of manifold subclasses, and the ECOC classification in detail.

A. Overview of the Proposed Pedestrian Detection Approach

The proposed pedestrian detection approach contains training and detection procedures, as shown in Fig. 1. Before either training or detection, HOG features need to be extracted to represent pedestrians (see Section II-B). In the training procedure, we first construct a manifold on which we can divide the pedestrian samples of different views and postures into subclasses using clustering (see Section II-C). These subclasses are encoded together with negative samples to train base classifiers with linear SVMs of soft margins ($C = 0.01$) [32]. According to the ECOC encoding rules (see Section II-D), M base classifiers will be trained to form a base classifier set $\{f_m(x)\}$, $m = 1, \dots, M$, where $f_m(x) = \text{Sign}(w_m^T \cdot x + b_m)$ are SVM classifiers of normal vector w_m^T and threshold b_m . Adjusting the value of b_m will balance the classification error rates on positives and negatives. Hard examples from test images will be added into the training set to improve the performance of the detector. The method is then retrained using this augmented set to produce the final detector. The training procedure is shown in Fig. 1.

In the detection procedure, a test image is repeatedly reduced in size by a factor of 1.2, resulting in pyramid images. Then, sliding windows of multiscales are extracted from each pyramid image and fed to all the base classifiers, obtaining the binary ECOCs. These codes will be classified as negative or positive by an ECOC decoding procedure (see Section II-D).

All image windows classified as positive are integrated into the original image as detection results. When the overlapping area of two windows classified as positive exceeds 80%, they will be merged. The detection procedure is shown in the right part of Fig. 1.

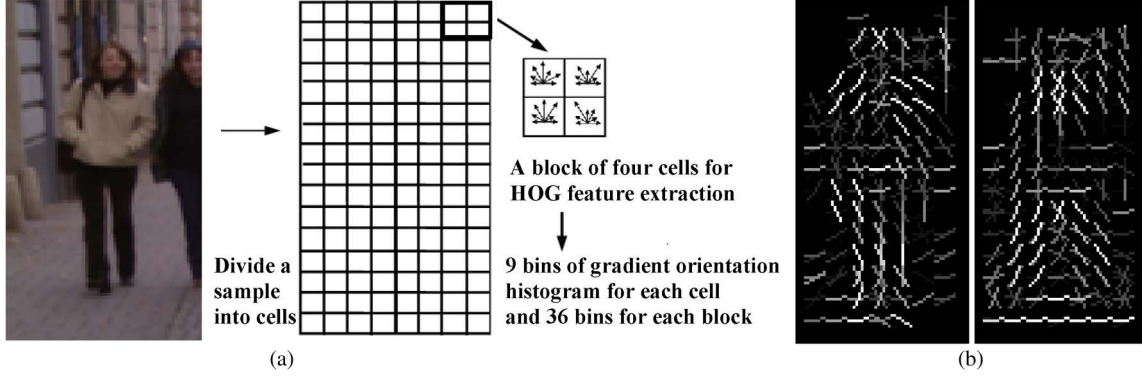


Fig. 2. HOG feature illustration. (a) Feature extraction and (b) visualization of HOG features for two pedestrian images from different subclasses. The brighter a stroke, the larger the HOG feature value.

B. Pedestrian Representation

HOG features, as proposed by Dalal and Triggs [13], are kind of state-of-the-art features for pedestrian representation and are employed as representation in this paper. As shown in Fig. 2(a), when extracting HOG features, a 64×128 training sample is divided into cells of size 8×8 pixels, and each group of 2×2 cells is integrated into a block in a sliding fashion and blocks overlap with each other. We first calculate the gradient orientation of each pixel. In each cell, we calculate nine-dimensional HOG features by calculating the nine-bin histogram of gradient orientations of all pixels in this cell. Each block contains four cells, on which 36-dimensional features are extracted. Each sample is represented by 105 blocks, on which 3780-dimensional HOG features are extracted. Fig. 2(b) shows the visualization of HOG features.

C. Pedestrian Manifold Construction and Division

A manifold embedding method is used to convert pedestrian samples from a very high dimensional space to a low dimensional embedded space. Samples in the embedded space are then divided into subclasses with a K -means clustering algorithm.

Local linear embedding (LLE) is employed to construct the pedestrian manifold [27]. It computes low dimensional and neighborhood-preserving embeddings of high dimensional inputs by mapping them into a global coordinate with lower dimensionality, as illustrated in Fig. 3. Given n pedestrian samples $\{x_i, i = 1, \dots, n$ in the 3780-dimensional input space X , LLE starts with finding the k nearest neighbors, based on the Euclidean distance, for each vector $x_i, 1 \leq i \leq n_i$, where n_i 's denote the indices of the k nearest neighbors of sample i . LLE identifies the optimal local convex combinations of the nearest neighbors to represent each original sample. This is equivalent to minimizing the objective as

$$\arg \min_{W_{ij}} \sum_i \left(x_i - \sum_{j \in n_i} w_{i,j} x_j \right)^2 \quad (1)$$

where $\sum_{j \in n_i} w_{i,j} = 1, w_{i,j} > 0$. The foregoing objective function can be solved as a least-square problem. Next, LLE considers an embedded space. Let z_i be the embedding of x_i in the embedded space. The embedded space has a dimensional-

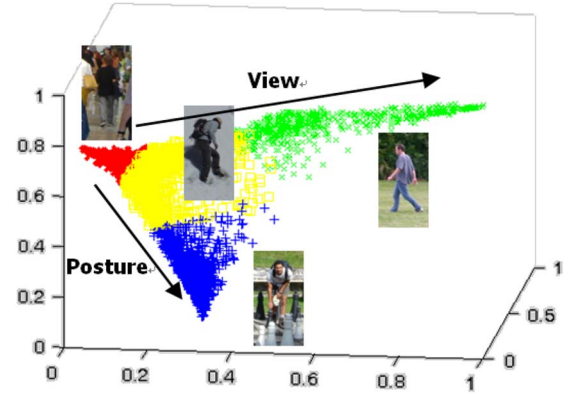


Fig. 3. Pedestrian manifold. Points of different colors denote samples from different pedestrian subclasses.

ity $d \ll D$. z_i is calculated such that the following objective function is minimized:

$$\arg \min_z \sum_i \left(z_i - \sum_{j \in n_i} w_{i,j} z_j \right)^2 \quad (2)$$

Note that the foregoing is equivalent to finding an embedded space such that local convex representations are preserved. It can be shown that with some additional conditions, which make the problem well defined, the task of minimization can be accomplished by solving a sparse eigenvector problem. More specifically, the d eigenvectors associated with the d smallest nonzero eigenvalues provide an ordered set of orthogonal coordinates centered on the origin [27]. d is the dimensionality of the embedding space. We set $d = 3$, in which case, the manifold could be visualized.

It can be seen in Fig. 3 that the manifold is not compact and nonlinear, which makes it difficult to train a linear classifier such as a linear SVM to classify pedestrians. We propose to divide the manifold into subclasses, each of which is more compact and approximately linear and, therefore, can be better modeled with a linear SVM classifier.

A standard K -means clustering algorithm is employed to perform the manifold division. Suppose that there are L subclasses on the manifold labeled $\{1, 2, \dots, L\}$. The pair-wise Geodesic distances [27], which are calculated from pair-wise Euclidean distances, are used as the clustering measure. The

TABLE I
CODING MATRIX OF PEDESTRIAN SUBCLASSES AND NEGATIVES

	f_1	f_2	f_3	f_4	f_5	f_6	...	f_M
Negatives (0)	0	0	0	0	0	0
Sub-class 1 (1)	1	0	0	0	1	0
Sub-class 2 (2)	0	1	0	0	1	1
Sub-class 3 (3)	0	0	1	0	0	1
...
Sub-class l (l)	0	0	0	1	0	0
...
Sub-class L (L)	0	0	0	1	0	0

sample that has the minimum summarization of Geodesic distance to all samples uses the subclass center. L is initially set to 2 and then is increased by 1 until the performance is optimized.

The reason for performing clustering on the manifold embedding instead of the original high dimensional feature space is the curse of dimensionality. Because of the high dimensionality of the original feature space, samples in such a space are very sparse, which makes it difficult to perform clustering analysis. According to the manifold's property [27], the spatial topology of samples in the embedded space is an approximation to that of the original feature space. Therefore, we can divide the samples by clustering them in the embedded space to avoid the curse of dimensionality.

D. ECOC Classification of Manifold Subclasses

Suppose that the pedestrian manifold is divided into L subclasses; together with a negative class, we have a total of $L + 1$ subclasses to be recognized. Consequently, we formulate the detection problem as a multiclass classification problem.

There are many different approaches to reduce a multiclass problem to a binary classification problem. The simplest approach considers the comparison of each class to all the others. Other research suggests that the comparison of all possible pairs of all $L + 1$ classes [10] can be accomplished by solving $L(L + 1)/2$ binary problems. Dieteerich and Bakiri [28] presented a framework in which the classification is performed according to a set of binary ECOCs. The outputs of all the classifiers are combined for decoding. In [11], the authors justified that the ECOC framework can reduce both the variance and the bias of classification models, showing its superiority on the multiclass problem over the other methods.

ECOC Encoding: For ECOC classification, we need to construct a coding matrix, as shown in Table I. Each row of the coding matrix defines codes for a positive subclass or the negative class. Each column defines a partition of subclasses (coded by 0 or 1 according to its subclass membership). From the view of learning, the coding matrix is interpreted as a set of $L + 1$ binary learning problems: one for each column.

Constructing the coding matrix (ECOC encoding) is not a trivial task. Empirical and heuristic methods are proposed in the previous work [10], [28]. In this paper, the manifold subclasses are encoded in terms of their neighboring relationships on the manifold, which is called a manifold advised ECOC. If two or more subclasses are neighboring on the manifold, then they will be put together as a combined class. Then, there is a column in

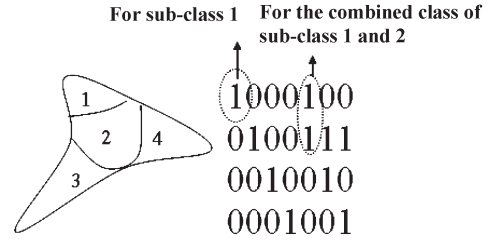


Fig. 4. Encoding four subclasses.

which the codes corresponding to these neighboring subclasses are “1,” and the codes to the other subclasses are “0,” as shown in Table I. Here, the neighboring relationships of subclasses are discovered with a nearest neighbor analysis. Two subclasses are determined to be neighbors if any sample in a subclass is one of the k nearest neighbors when constructing the manifold.

When training the base classifiers with respect to the coding matrix, we need to combine the pedestrian samples of neighboring subclasses together to form a positive set

$$\{(x_1, f_j(x_1)), \dots, (x_{ip}, f_j(x_{ip})), \dots, (x_{iq}, f_j(x_{iq}))\} \quad (3)$$

where ip and iq denote the sample indices of two neighboring subclasses on the manifold. The combined subclasses are put together with the negative training set, and then, a linear SVM is employed to learn the base classifier $f_j(x)$, $j = 1 \dots M$. Let $\{(x_i, l_i)\}$, $i = 1, \dots, N$ be a set of training samples where instance x_i belongs to the feature space X , and label l_i takes values from a set of subclass labels; we define $\{C_l\}_{l=0, \dots, L}$ as $L + 1$ distinct codewords, each of which has a length of M and corresponds to a row in Table I. M base classifiers $\{f_1(x), f_2(x), \dots, f_M(x)\}$ need to be trained, each of which corresponds to a column in Table I. In the training procedure, if the j th bit of C_l is 1 ($C_{lj} = 1$), then the base classifier $f_j(x)$ outputs 1; otherwise, $f_j(x)$ outputs 0.

Fig. 4 illustrates the encoding of four neighboring subclasses on a manifold. According to the neighbor relations of subclasses, we can construct the ECOC coding matrix as shown in the figure. In the coding matrix, the first column contains one nonzero element corresponding to the row of subclass 1. Therefore, the column is for subclass 1. The fifth column contains two nonzero elements corresponding to the rows of subclasses 1 and 2. Therefore, the column is for the combined class made up of subclasses 1 and 2.

ECOC Decoding: Given a test sample x , the learned base classifiers can be applied to the sample to compute a binary test output vector $C = \{f_1(x), \dots, f_M(x)\}$. Then, we can determine which codeword C_l is the closest to the test output vector C using the Hamming distance. This is called ECOC decoding. In the decoding process, the test sample x is assigned to the subclass of the smallest Hamming distance, as follows:

$$H_l(x) = \|C_l - C\|_1 = \sum_{j=1}^M |C_{lj} - f_j(x)| \quad (4)$$

$$l(x) = \arg \min_l \{H_l(x) | l = 0, 1, \dots, L\} \quad (5)$$

$$F(x) = \begin{cases} 0, & \text{if } l(x) = 0 \\ 1, & \text{otherwise} \end{cases} \quad (6)$$

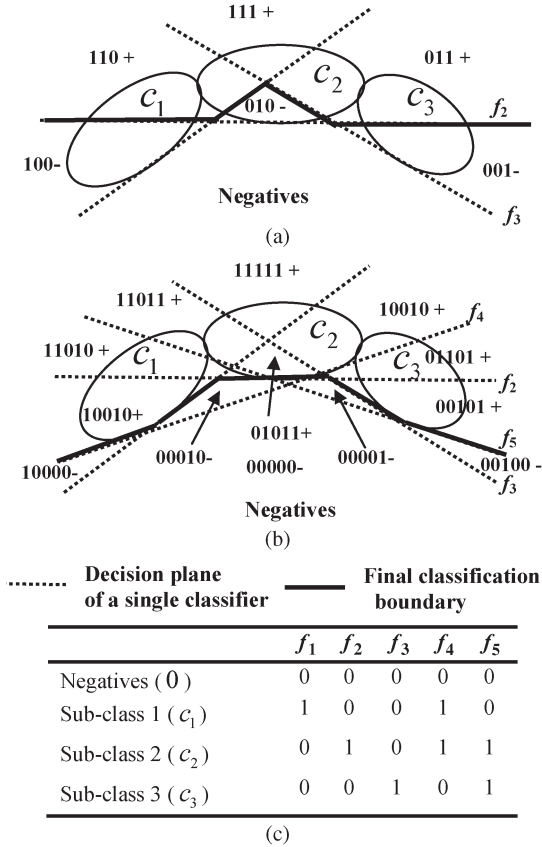


Fig. 5. Comparison of classification boundaries of (a) the voting scheme and (b) the ECOC approach. (c) Coding matrix for three neighboring subclasses shown in (a) and (b).

where $F(x)$ denotes the final classifier, $F(x) = 1$ denotes a pedestrian, and $F(x) = 0$ denotes a negative.

E. Discussion of ECOC Classification

ECOC classification is a supervised learning method that applies binary learning algorithms to solve multiclass problems. Any learning algorithm that can handle two-class problems could be employed as a base classifier. ECOC classification can be viewed as a compact form of “voting” among multiple classifiers. The main advantage of this voting is that the errors committed by each of the base classifiers are substantially uncorrelated [28]. When making the ECOC code matrix, each codeword will be well separated in Hamming space from the other codewords, and each column should be uncorrelated with all the other columns. Kong [29] justified that with ECOC, we can obtain a smaller variance and bias of classification models than with the other schemes, such as voting of one-versus-rest classifiers or classifier pairs, etc. Given an effective encoding rule, ECOC is justified to be a simple but effective classifier combination scheme compared with the boosting or mixture-of-expert method.

As previously mentioned, our coding rule encodes neighboring subclasses on the manifold. Fig. 5 provides an explanation for the reason for this rule. Fig. 5(a) illustrates the classification boundaries of voting, and Fig. 5(b) is the illustration of classification boundaries of ECOC, which use the coding matrix in the

TABLE II
DATA SETS

Datasets	Training samples	Test images /samples
SDL [29]	7550 positives and 5769 negatives	258 /1688
TUD- Brussels [30]	1303 positives and 5000 negatives from 386 images	509 /1397
INRIA [13]	2478 positives and 12180 negatives	288 /589



Fig. 6. Positive and negative training samples.

table of Fig. 5(c). Subclasses are relaxed to be convex shapes. The feature space is divided into subspaces, and samples in these subspaces will be classified into positive (+) or negative (−) samples, with the ECOC decoding by (4)–(6). For example, if the base classifiers output is {11010} for the samples in a subspace, then the samples will be classified to subclass 2 since the output of base classifiers is closest to {10010} in all of the codewords {00000, 10010, 10011, 00101}. This analysis can be extended to all the other subspaces. Consequently, ECOC with the proposed coding strategy can optimize the classification boundary between positives and negatives, as shown in Fig. 5(b). Finally, we could obtain a nonlinear classification boundary that can well discriminate the positive manifold from the negatives.

III. EXPERIMENTAL RESULTS

In this section, we describe the data sets used in our experiments, evaluate the proposed approach on these data sets, and provide in-depth analysis of the detection results.

A. Data Sets

There are three data sets used in our experiments, as described in Table II. One is the System Development Laboratory of Graduate University of Chinese Academy of Sciences (SDL) data set with 7550 positives and 5769 negatives for training set [29] with front/side views and running, sporting, and bicycling postures. Some of the training examples are shown in Fig. 6. In the SDL data set, there are 258 images with 1688 samples for testing, containing multiple views and postures, such as walking, sporting, running, etc. It is publicly available online (<http://coe.gucas.ac.cn/SDL-HomePage/resource.asp>) [29]. The

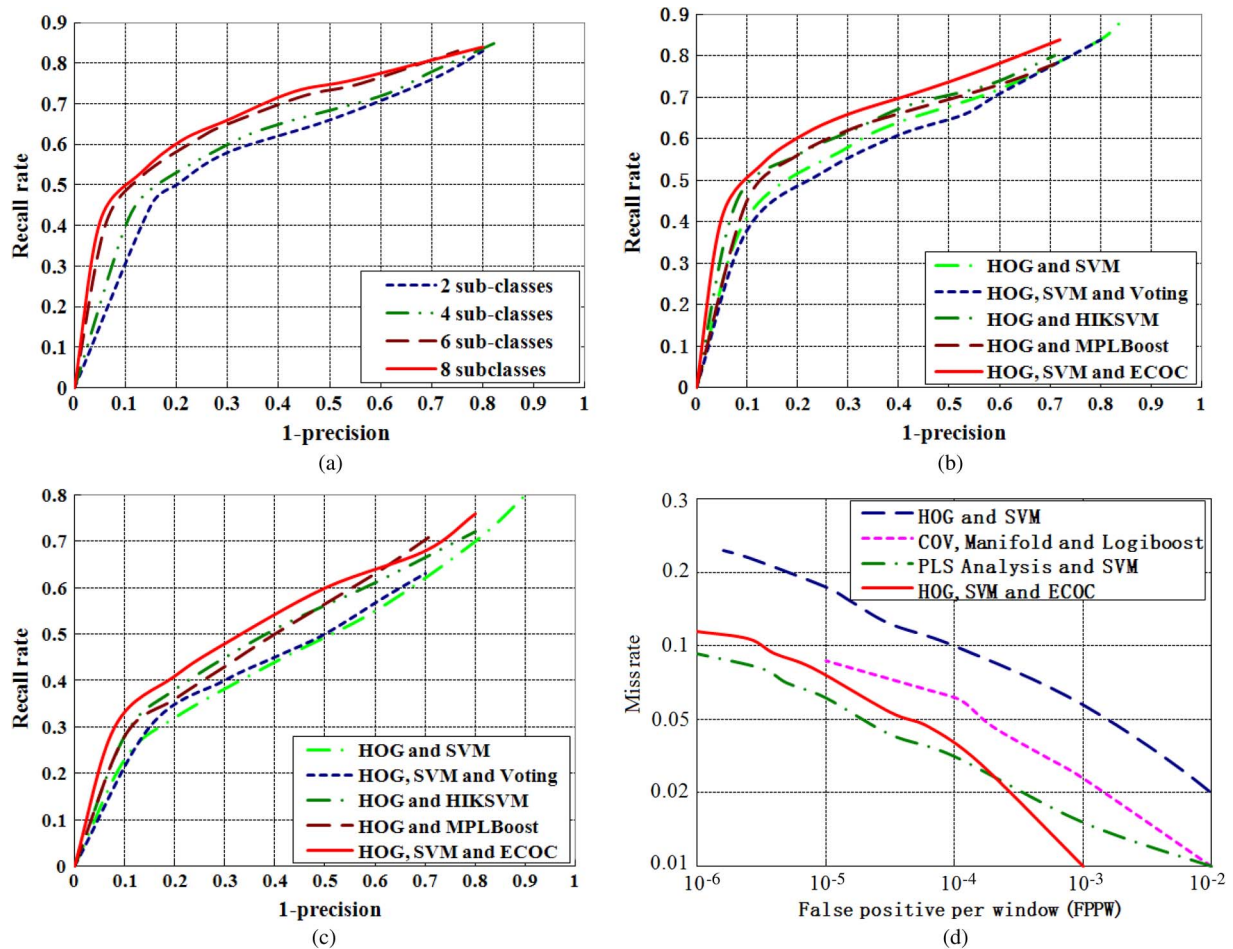


Fig. 7. Performance and comparisons. (a) Performance under different manifold subclasses numbers. Performance comparison on (b) SDL data set, (c) TUD-Brussels data set, and (d) INRIA data set.

second data set is the Technique University of Darmstadt (TUD)-Brussels data set, captured from a moving platform for driving warning systems. The third data set is the INRIA data set [9], which has been widely used for pedestrian/human detection evaluation in recent years. Samples (2478 positives and 12 180 negatives) selected from 1218 person-free training photos provide the initial negative set. There are 288 images with 589 multiview and multiposture pedestrian samples for testing.

B. Detection Performance and Comparisons

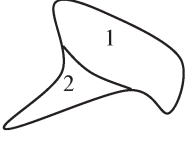
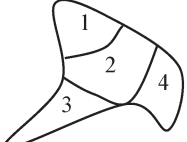
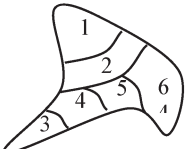
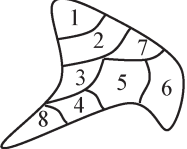
SDL Data Set: On this data set, the detection performance of ECOCs under different subclass numbers has been evaluated. Recall rate and 1.0 precision are used to evaluate the performance, and the results are shown in Fig. 7(a). In Table III, we illustrate the spatial relationships of 2–4–6–8 subclasses on the manifold and then present the ECOC according to the neighboring relationships. It can be seen that with an increasing number of subclass, the neighboring relations become complex, and then more base classifiers are needed. In Fig. 7(a), it can be seen that the performance of eight subclasses is much higher than that of two subclasses, showing that with the increasing number of subclass, the compactness and linearity of the subclasses increase, and then, the detection performance also increases. When the number of subclass increases from

6 to 8, the performance marginally increases since the number of training samples in each subclass drops. This also affects the performance of the trained classifiers. This shows that both subclass linearity and number of subclasses will affect the performance of base classifiers, as two opposite factors. Therefore, a proper number of subclass should experimentally be selected so that the preceding two factors are balanced.

When evaluating the detection performance, four representative methods, including HOG+SVM [13], HOG+SVM+voting, HOG+HISVM [31], and HOG+MPLBoost (a boosting method on SVMs) [30], are selected for comparison. In all of the methods, HOG features are employed, and a sliding window classification scheme is used for detection. The voting scheme is a combination of multiple SVM classifiers that are trained individually. HIKSVM is a kernel SVM method for nonlinear classification with high efficiency, and MPLBoost is a weighted voting scheme for combing multiple strong classifiers. Fig. 7(b) shows that our proposed approach reports a higher performance on the SDL data set.

TUD-Brussels Data Set: This data set contains video images from a driving assistant system. The test images are captured video frames from a driving scene with cluttered background and are of walking and bicycling postures. In Fig. 7(c), we compare our approach (with six subclasses) with the aforementioned methods. It can be seen that the proposed approach

TABLE III
ECOC CODING MATRIX EXAMPLE

Sub-classes on manifold	Base classifier number	ECOCs
	3	011 101
	7	1000100 0100111 0010010 0001001
	15	100000100000000 010000111100011 001000010010010 000100001011011 000010000101101 000001000000100
	21	10000000100000000000 01000000111100000000 001000000101111001000 000100000000101110000 000010000000011011110 000001000000000000011 0000001000011000001101 000000010000000100000

outperforms the other detection methods. In Fig. 7(c), when the precision is 0.5, the highest recall rate is about 60% reported by our approach.

INRIA Data Set: This data set is popular for pedestrian detection evaluation and uses a miss rate and false positives per window (FPPW) criteria. In Fig. 7(d), we compare our approach (with four subclasses) with several state-of-the-art approaches, including HOG+SVM [13], COV+Manifold+Logiboost [16], and partially least square (PLS) analysis [22]. While we are able to run the implementation for the method [13], curves for methods [16], [22] are obtained from their reported results using the same training samples and test images in comparison. When the training set is small, it is observed that the case of four subclasses reports the best performance. As shown in Fig. 7(d), our approach reports the best results on the data set. Compared with the FPPW rate of 10^{-4} , it has a 4% miss rate, which is about 6% lower than HOG+SVM and 3% lower than COV+Manifold+Logiboost. When the FPPW is 10^{-6} , it can be seen that the miss rate of our approach is also lower than the other three methods, whereas it is a little higher than PLS analysis, which uses more powerful feature representation and performs a feature dimension reduction operation before classification.

It is observed that when test samples, such as samples of SDL and TUD-Brussels data sets, contain more view and posture variation, the advantages of our approach become more obvious. When the test samples are near frontal view with little posture variation, our approach has small advantage over the classic SVM+HOG method.

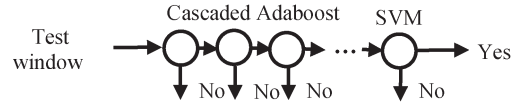


Fig. 8. Speed up detection with the coarse-to-fine base classifier.

TABLE IV
COMPRISE OF DROP OF DETECTION RATE AND DETECTION SPEED

Cascade number	Speed images/seconds	Drop of detection rate
0	0.13	0
5	1.4	2.4%
10	4.7	3.1%
20	12.9	5.6%
27	26.3	7.2%

C. Detection Speed

Given D dimensional feature vectors, the time complexity of a linear SVM classification is $O(D)$. It needs an inner product operation between a feature vector and the norm vector. The time complexity of a kernel SVM is $O(S \cdot D)$, where S is the number of support vectors. This means that we need S inner product operations between the feature vector and the support vectors. The time complexity of our ECOC+SVM is $O(M \cdot D)$ by multiplying the sample feature vector with M linear SVMs. It can be seen that the time complexity of ECOC+SVM is much lower than that of kernel SVMs because $M \ll S$ and higher than that of linear SVMs.

In our proposed approach, the detection speed linearly decreases with an increasing number of base classifier. For example, when there are six subclasses and 15 base classifiers, the detection speed is averagely 0.13 images/s on images of 640×480 pixels with an Intel Core-2 2.8-GHz CPU. When the image resolution is reduced to 320×240 pixels, the detection speed rises to about 1.0 image/s. It should be mentioned that many of the state-of-the-art methods with linear/kernel SVMs are also far from real time. The speed of HOG+SVM on images of 640×480 pixels is about 1.2 images/s. In addition, the reported speed of the HIKSVM method is five to six times slower than that of linear SVM [31].

There are two ways to speed up our current detection approach. One is to speed up the base classifiers by a cascaded classification, and the other is to use parallel processing (see Fig. 8).

As the overall detection speed is proportional to the speed of the base classifiers, we propose to use a coarse-to-fine base classifier scheme to speed up the detection. The well-known cascaded Adaboost classifiers on absolute Haar-like features [5] are employed as coarse classifiers, and the SVM on HOG is employed as a fine classifier, which is to construct coarse-to-fine base classifiers. Using this scheme, most of the image windows are rejected in the coarse classification. It is reported in Table IV that when five cascades of Adaboost classifiers are used, the detection speed raises from 0.13 to 1.4 images/s, but the recall rate drops to about 2.4%. It can also be seen in Table IV that a 4.7-images/s detection speed is reported at the cost of 3.1% detection rate. This speed can be improved to real time with a further drop in recall rate.



Fig. 9. Detection examples. Detected false positives are marked with rectangle of white dash line, and missed positives are marked with rectangle composed of a black dashed line.

On the other hand, since each of the base classifiers can run separately in each classification, the parallel-processing-based code optimization can be used to improve the detection speed in practical application systems. Details of parallel processing and code optimization are not within the scope of this paper.

D. Detection Examples

In Fig. 9, we show some detection examples. From Fig. 9(a)–(m), most pedestrians are correctly located with few false positives. In Fig. 9(e), there is a missed positive (marked with dash black rectangle) since the pedestrian is too close to the image boundary. In Fig. 9(m), a squatting pedestrian is missed. The current method can cope with postures of near-standing pedestrians and not the seated or squatting pedestrians. In Fig. 9(g), (k), and (l), there is one false positive in each image, which is caused by a tree trunk, buildings, and cloth

texture, respectively. In our experiments, we found that objects with complex texture may falsely be detected because of their similarity to pedestrians.

In Fig. 10, we show detection examples of successive video images. The video is captured from a moving platform with dynamic background. It could be seen that the detection result is quite stable. The pedestrians with views and postures are correctly detected in successive video images with few false positives from the cluttered background, showing the potential application of the proposed approach to driver warning systems.

IV. CONCLUSION

Although view/posture robustness of object detection in image and video frames is very important to practical applications, particularly for driving warning systems, it is still an open problem. In this paper, we have proposed a new approach to this

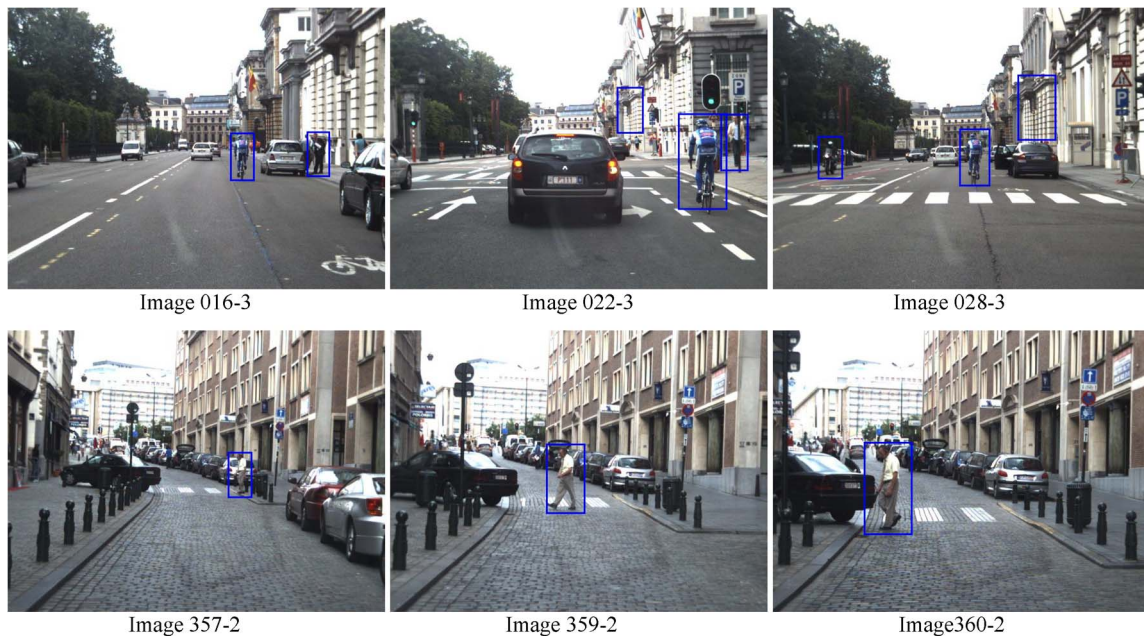


Fig. 10. Detection examples of successive video frames from TUD-Brussels data set.

problem through a manifold-based clustering strategy and an ECOC classification method. The new concepts and techniques proposed in this paper include the pedestrian manifold, the detailed classification and analysis of multiview and multiposture pedestrian patterns, and the ECOC classifier for pedestrian detection. Detailed experiments and comparisons are reported, confirming that our method is capable of handling multiview and multiposture pedestrians effectively.

ECOC-based classification is a framework to multiclassifier combination. Other leading classification methods, such as Kernel or L_1 -norm SVMs [33], can be integrated into the framework as base classifiers. It is also reasonable to extend the proposed approach to multiview objects like faces and vehicles.

A limitation of the proposed approach is that only image cue is used. Other cues, such as infrared [34] or laser [40], should be investigated in the future work. Another limitation is that the efficiency of the proposed approach cannot feed the requirement of some practical applications. At present, we propose to use cascaded base classifiers to improve the detection speed at the cost of some drop of detection rate. More sophisticated techniques should be employed to further improve the detection speed in the future work.

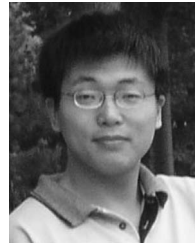
ACKNOWLEDGMENT

The authors would like to thank Dr. J. Chen and Dr. G. Zhu for their discussion of the manuscript and the associate editor and reviewers for their constructive comments.

REFERENCES

- [1] X. B. Cao, H. Qiao, and J. Keane, "A low cost pedestrian detection system with a single optical camera," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 1, pp. 58–67, Mar. 2008.
- [2] L. Li and M. K. H. Leung, "Unsupervised learning of human perspective context using me-dt for efficient human detection in surveillance," in *Proc. IEEE Int. Conf. CVPR*, 2008, pp. 1–8.
- [3] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.
- [4] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *Int. J. Comput. Vis.*, vol. 73, no. 1, pp. 41–59, Jun. 2007.
- [5] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, Jul. 2005.
- [6] Q. X. Ye and J. B. Jiao, "Multi-posture pedestrian detection in video frames by motion contour matching," in *Proc. IEEE Int. Conf. ACCV*, 2007, pp. 896–904.
- [7] Z. Lin, L. Davis, and D. Doermann, "Hierarchical part-template matching for pedestrian detection and segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
- [8] Z. Lin and L. S. Davis, "A pose-invariant descriptor for human detection and segmentation," in *Proc. Eur. Comput. Vis. Conf.*, 2008, pp. 423–436.
- [9] B. Wu and R. Nevatia, "Cluster boosted tree classifier for multi-view, multi-pose object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
- [10] O. Pujol, P. Radeva, and J. Vitria, "Discriminant ECOC: A heuristic method for application dependent design of error correcting output codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 1007–1012, Jun. 2006.
- [11] E. B. Kong and T. G. Dietterich, "Error-correcting output coding corrects bias and variance," in *Proc. Int. Conf. Mach. Learn.*, 1995, pp. 313–321.
- [12] H. M. Zhang, W. Gao, X. L. Chen, S. G. Shan, and D. B. Zhao, "Robust multi-view face detection using error correcting output codes," in *Proc. Eur. Comput. Vis. Conf.*, 2006, pp. 1–12.
- [13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. CVPR*, 2005, pp. 886–893.
- [14] Q. Zhu, S. Avidan, M. Yeh, and K. Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Int. Conf. CVPR*, 2006, pp. 1491–1498.
- [15] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.
- [16] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008.
- [17] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, "Discriminative local binary patterns for human detection in personal album," in *Proc. IEEE Int. Conf. CVPR*, 2008, pp. 1–8.
- [18] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 32–39.

- [19] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 90–97.
- [20] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *Proc. IEEE Int. Conf. CVPR*, 2007, pp. 1–8.
- [21] S. Munder and D. M. Gavrilu, "An experimental study on pedestrian classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1863–1868, Nov. 2006.
- [22] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 24–31.
- [23] V. D. Shet, J. Neumann, V. Ramesh, and L. S. Davis, "Bilattice-based logical reasoning for human detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
- [24] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 4, pp. 349–361, Apr. 2001.
- [25] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrilu, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. IEEE Int. Conf. CVPR*, 2010, pp. 990–997.
- [26] J. Marin, D. Vazquez, D. Geronimo, and A. M. Lopez, "Learning appearance in virtual scenarios for pedestrian detection," in *Proc. IEEE Int. Conf. CVPR*, 2010, pp. 137–144.
- [27] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.
- [28] T. G. Dieteerich and G. Bakiri, "Solving multi-class learning problems via error correcting output codes," *J. Artif. Intell. Res.*, vol. 2, no. 1, pp. 263–286, Aug. 1995.
- [29] [Online]. Available: <http://coe.gucas.ac.cn/SDL-HomePage/resource.asp>
- [30] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *Proc. IEEE Int. Conf. CVPR*, 2009, pp. 794–801.
- [31] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection SVM is efficient," in *Proc. IEEE Int. Conf. CVPR*, 2008, pp. 1–8.
- [32] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods-Support Vector Learning*, B. Scholkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999.
- [33] R. Xu, B. Zhang, Q. Ye, and J. Jiao, "Cascaded L1-norm minimization learning (CLML) classifier for human detection," in *Proc. IEEE Int. Conf. CVPR*, 2010, pp. 89–96.
- [34] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *IEEE Trans. Circuits Syst. Vid. Technol.*, vol. 6, no. 1, pp. 63–71, Mar. 2005.
- [35] Q. Ye, J. Jiao, and B. Zhang, "Fast Pedestrian detection with multi-scale orientation features and two-stage classifiers," in *Proc. IEEE Int. Conf. Image Process.*, 2010, pp. 881–884.
- [36] O. Sidla and M. Rosner, "HOG pedestrian detection applied to scenes with heavy occlusion," in *Proc. SPIE*, 2007, vol. 6764, p. 676 408.
- [37] C. Hou, H. Z. Ai, and S. H. Lao, "Multiview pedestrian detection based on vector boosting," in *Proc. IEEE Int. Conf. ACCV*, 2007, pp. 210–219.
- [38] Y. X. Hu, Z. H. Zeng, L. J. Yin, X. Z. Wei, X. Zhou, and T. S. Huang, "Multi-view facial expression recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2008, pp. 1–6.
- [39] S. Kamijo, K. Fujimura, and Y. Shibayama, "Pedestrian detection algorithm for on-board cameras of multi view angles," in *Proc. IEEE Intell. Veh. Symp.*, 2010, pp. 973–980.
- [40] S. Gidel, P. Checchin, C. Blanc, T. Chateau, and L. Trassoudaine, "Pedestrian detection and tracking in an urban environment using a multilayer laser scanner," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 579–588, Sep. 2010.



Dr. Ye won the "Sony Outstanding Paper Award" in 2005.



Qixiang Ye (M'10) received the B.S. and M.S. degrees in mechanical and electronic engineering from the Harbin Institute of Technology, Harbin, China, in 1999 and in 2001, respectively, and the Ph.D. degree from the Chinese Academy of Sciences, Beijing, China, in 2006.

Since 2009, he has been an Associate Professor with the Graduate University of the Chinese Academy of Sciences, Beijing. His research interests include image processing, pattern recognition, intelligent systems, etc.

Jixiang Liang received the B.S. degree in computer science from the University of Science and Technology of China, Hefei, China, in 2009. He is currently working toward the M.S. degree with the Graduate University of the Chinese Academy of Sciences, Beijing, China.

His research interests include image processing, pattern recognition, etc.



Jianbin Jiao (M'10) received the B.S., M.S., and Ph.D. degrees in mechanical and electronic engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 1989, 1992, and 1995, respectively.

From 1997 to 2005, he was an Associate Professor with HIT. He had been a Visiting Professor with Kyushu Institute of Technology, Fukuoka, Japan, in 2009 and a Visiting Scholar with the University of Nevada, Las Vegas, from 2002 to 2005. Since 2006, he has been a Professor with the Graduate University of the Chinese Academy of Sciences, Beijing, China.

He is also a "Science-100 professor" with the Chinese Academy of Sciences. His research interests include image processing, pattern recognition, intelligent systems, etc.