Human Detection in Images via Piecewise Linear Support Vector Machines

Qixiang Ye, *Member, IEEE*, Zhenjun Han, *Member, IEEE*, Jianbin Jiao, *Member, IEEE*, and Jianzhuang Liu, *Senior Member, IEEE*

***Abstract*— Human detection in images is challenged by the view and posture variation problem. In this paper, we propose a piecewise linear support vector machine (PL-SVM) method to tackle this problem. The motivation is to exploit the piecewise discriminative function to construct a nonlinear classification boundary that can discriminate multiview and multiposture human bodies from the backgrounds in a high-dimensional feature space. A PL-SVM training is designed as an iterative procedure of feature space division and linear SVM training, aiming at the margin maximization of local linear SVMs. Each piecewise SVM model is responsible for a subspace, corresponding to a human cluster of a special view or posture. In the PL-SVM, a cascaded detector is proposed with block orientation features and a histogram of oriented gradient features. Extensive experiments show that compared with several recent SVM methods, our method reaches the state of the art in both detection accuracy and computational efficiency, and it performs best when dealing with low-resolution human regions in clutter backgrounds.**

***Index Terms*— Classification, object detection, piecewise linear, support vector machine.**

I. Introduction

DETECTION of humans in images and video frames is an important problem in the area of image based sensing with applications such as robotics, entertainment, surveillance, and pedestrian warning for driving assistance [1]–[5]. Although the detection of humans in some common views and in static video background has been greatly put forward in recent years, it is still a challenging problem in the situations of moving cameras, complex backgrounds, and in particularly, large variations of views and postures.

In existing human detection methods, feature representation and classifier design are two main problems being investigated. Visual feature descriptors have been proposed for human detection including Haar-like features [5], HOG [6], v-HOG [7], Gabor filter based cortex features [8], covariance features [9], Local Binary Pattern (LBP) [10], HOG-LBP [11], Edgelet [12], Shapelet [13], Local Receptive Field (LRF) [14], Multi-Scale Orientation (MSO) [15], Adaptive Local Contour [16], Granularity-tunable Gradients Partition (GGP) descriptors [17], pose-invariant descriptors [18]. A most recent research demonstrates the superior performance when using mixture of different kinds of visual features, motion and depth information [19].

The extracted features on labeled samples are usually fed into a classifier for training. Linear SVM is the most popular classifier with several reported landmark works for human detection [6], [8], [11], [19]. However, when we need to detect multi-view and multi-posture humans simultaneously in a video system, the performance of a linear SVM often drops significantly. It is observed in experiments that humans of continuous view and posture variations form a manifold, which is difficult to be linearly classified from the negatives. An algorithm that requires multi-view and multi-posture humans to be correctly classified by a linear SVM in the training process often leads to over-fitting. Some non-linear classification methods such as Kernel SVMs [20] are options to handle this problem, but they are generally much more computationally expensive than linear methods. In addition, the use of the kernel trick in a very high-dimensional feature space, such as a 3780 dimensional HOG feature space, may magnify the curse of dimensionality.

On the other hand, some approaches use a divide-and-conquer strategy to deal with the multi-view and multi-posture problem, by first dividing training positives into sub-classes and then training multiple models for detection [21], [22]. In [21], [22], tree structure and pyramid boosting classifiers are developed to detect multi-view humans in images. These divide-and-conquer strategies can reduce empirical error in training process and improve the detection performance in some cases, but sometimes they also bring higher structural risk and more false positives.

Another solution to the multi-view and multi-posture problem is to segment a human body into parts [2], [23], [24], considering that each part has smaller deformation, lower dimensionality and non-linearity, and therefore can be better detected with a linear classifier. In [2], a deformable part-based model (DPM) is proposed for human detection. Human parts and their spatial bias are modeled with a structure SVM with latent variables (latent SVM). When performing training or detection, a local searching operation is carried out to optimize the location of each part-based model, which is called local deformation. By the local deformation, the detection avoids

Manuscript received March 31, 2012; revised September 8, 2012; accepted September 19, 2012. Date of publication October 5, 2012; date of current version January 10, 2013. This work was supported in part by the National Basic Research Program of China 973 Program under Grant 2011CB706900 and Grant 2010CB731800, and the National Science Foundation of China under Grant 61039003, Grant 61271433, and Grant 61202323. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chun-Shien Lu.

Q. Ye, Z. Han, and J. Jiao are with the School of Electronics and Communication Engineering, Graduate University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: qxye@gucas.ac.cn; hanzhj@gucas.ac.cn; jiaojb@gucas.ac.cn).

J. Liu is with the Media Laboratory, Huawei Technologies Co., Ltd., Shenzhen 518129, China, and also with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: liu.jianzhuang@huawei.com).

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/TIP.2012.2222901

suffering from the view and posture variations. In [24], an extension of the DPM is proposed to allow sharing of object part models among multiple mixture components as well as object classes. This results in more compact models and allows training examples to be shared by multiple components, reducing the effect of a training set of limited size. The DPM methods contribute an elegant framework for object detection, showing state-of-the-art performance on human detection. But they suffer from low resolution images of human objects, on which local model optimization has little significance.

In machine learning research, piecewise and localized SVMs have attracted much attentions [25]–[27], [28], [29], due to their superior performance over global kernel SVMs. In [26], the authors derive the upper bound of the structure risk of a piecewise SVM. However, the problem of how to construct a piecewise decision boundary in a high-dimensional feature space is not well discussed. In [28], cross distance minimization algorithm (CDMA) is designed to compute hard margin of non-kernel SVMs. In [29], multicategory SVMs are proposed to extend the binary SVM to the multicategory case, which is essentially different from our proposed piecewise linear SVM (PL-SVM) method in both the theoretical basis and the training procedure. In terms of the theoretical basis, the multicategory SVMs are developed to approximate the Bayes rule for multicategory classification purpose. Our PL-SVM method exploits the piecewise discriminative function to construct a non-linear classification boundary that can discriminate multiple positive sub-classes from the negative class. In the training of the multicategory SVMs, the method of Lagrange multipliers is employed to solve the objective equation of the dual problem. In the training of PL-SVM, nearest point analysis (NPA) on convex hulls together with an iterative linear SVM solution is used, which guarantees the max-margin of the final classifier. In [27], Cheng et al propose a profile SVM (P-SVM) to reach local linear classification, using the minimal distance to each pre-calculated cluster center to decide which local SVM a sample should belong to. Although profile SVM has the advantages of non-linear discrimination and sample division in a low-dimensional feature space, its sample division strategy suffers from the curse of dimensionality. In addition, the max-margin property of the profile SVM is not fully considered in the classifier training procedure.

In this paper, pedestrian detection is formulated as a non-linear classification problem in a high-dimensional feature space. The piecewise linear SVM (PL-SVM) method is introduced into multi-view and multi-posture human detection for the first time. Our PL-SVM is essentially different from other piecewise SVMs in the feature space division and model training strategy. When training the PL-SVM, with a membership degree maximization criterion, the feature space is divided into subspaces,[1] each of which can be better discriminative for a linear SVM. This approach ensures a lower empirical risk than using only one linear SVM. The training of the PL-SVM is an iterative division of training samples and the feature space. The convergence of the iterations is guaranteed by the monotonically increasing and bounded margins of the PL-SVM, which also guarantees that the PL-SVM is a maximal margin classifier, and thus has a small structural risk. This further ensures the generalization ability and the performance of the PL-SVM, providing a simple and effective way for multi-view and multi-posture human detection. A new kind of feature, called Block Orientation (BO), is proposed as a complement to the popular HOG features. BO and HOG features are incorporated with two cascaded PL-SVMs, improving both the accuracy and efficiency in human detection.

The remainder of this paper is organized as follows: PL-SVM modeling and training are presented in Section II. Human detection with the proposed PL-SVM is described in Section III. Experimental results are provided in Section IV. Section V concludes the paper.

II. Piecewise Linear Support Vector Machine

In this section, we present the PL-SVM and explain how to train it, given a training sample set $X = \{(x_n, y_n)\}, n = 1, \ldots, N$, where x_n is a sample feature vector, $y_n \in \{-1, +1\}$ denotes the sample label and N denotes the number of samples.

A. PL-SVM Model

A PL-SVM, made up of K linear SVMs, is described as a piecewise linear function

$$f(x) = \underset{f_k(x),\ x \in \Omega_k}{\arg\max} \{C_k(x)\} \tag{1}$$

where $f_k(x) = w_k^T \cdot x + b_k, k = 1, \ldots K$, represents the kth local linear SVM with normal vector w_k^T and threshold b_k. In (1), $\Omega_k = \Omega_k^+ \cup \Omega^-$ denotes the kth subspace occupied by a subset of the training samples as shown in Fig. 1.

In (1), $C_k(x)$ is the membership degree of a sample x to Ω_k. From the viewpoint of probability, the membership degree is defined as

$$C_k(x) = P_k(y = 1|x) \tag{2}$$

where $P_k(y = 1|x)$ is the outputted probability of a sample x being a positive when it is inputted into the kth linear SVM. The probability is defined as the functions of the SVM output as follows

$$P_k(y = 1|x) = \frac{1.0}{1.0 + \exp\left(-(A_k \cdot f_k(x) + B_k)\right)} \tag{3}$$

where A_k and B_k are two parameters calculated with a maximum likelihood estimation on the training subset [30], and $A_k \cdot f_k(x) + B_k$ is called the parameterized sample-to-hyper-plane distance. By (2) and (3) we know that the larger this distance is, the larger the probability, and then the larger the membership degree to the corresponding SVM, as shown in Fig. 2.

With the membership degree maximization criterion in (1) and (2), each linear SVM is responsible for a subspace for classification. The final non-linear classification boundary in the whole feature space consists of linear hyper-planes, as

[1] In this paper, a subspace of a space is a part of the space. They have the same dimensionality.

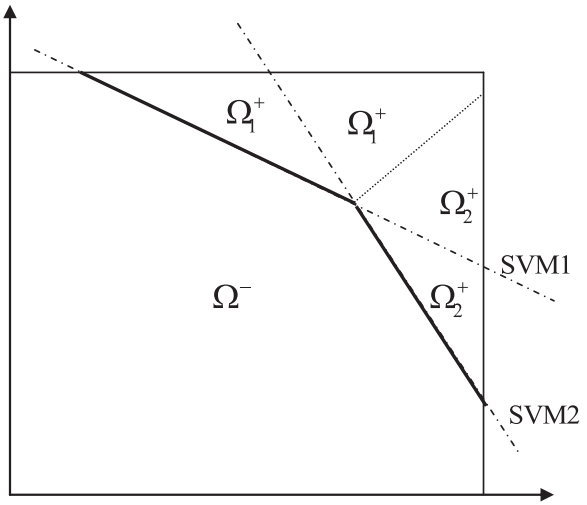


Fig. 1. Illustration of the PL-SVM and feature space division. subspaces are bounded with dotted lines and Ω_1^+, Ω_2^+ denote the positive subspaces corresponding to linear SVMs 1, 2, respectively, with Ω^- denoting negatives. Different positive subspaces are related to samples of different views and postures. Classification boundary of the PL-SVM is marked by bold line segments.

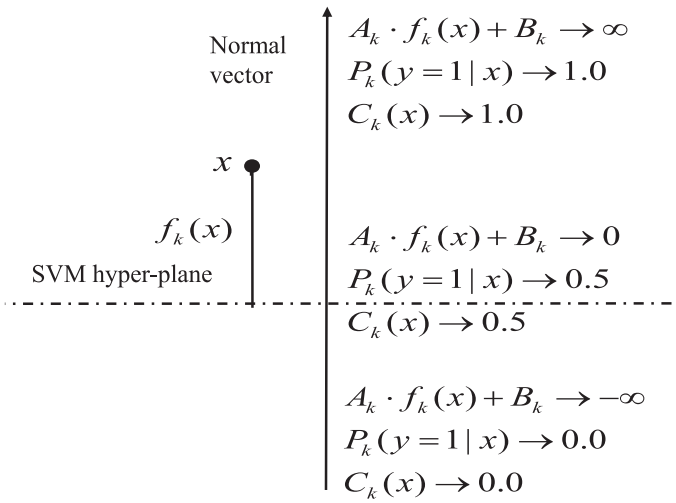


Fig. 2. Relations among parameterized sample-to-hyper-plane distance, classification probability, and membership degree.

illustrated in Fig. 1. When this criterion is used to divide the feature space and assign positive samples in an iterative training, the parameterized sample-to-hyper-plane distance will be enlarged and then the SVM margins will be also enlarged step by step. This is consistent with the maximal margin principle, ensuring that the PL-SVM keeps the essence of the original SVM approach.

When performing classification, (1) can be converted to a PL-SVM discriminative function

$$F(x) = \text{Sign}\ (f(x)) \tag{4}$$

with a sign function for discrimination and detection.

Given a training set $X = \{(x_n, y_n)\}, n = 1, \ldots, N$, we need to solve the following multi-objective programming problem

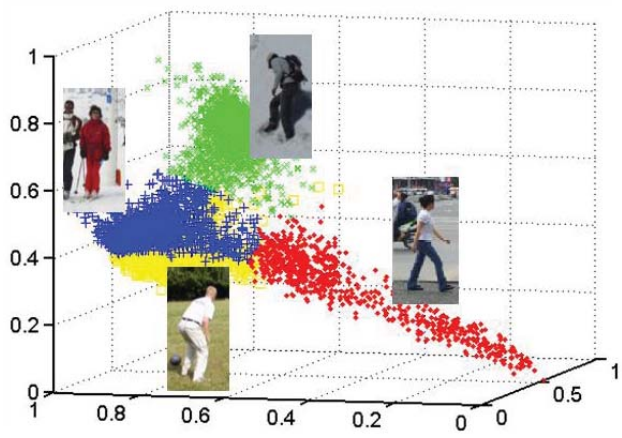


Fig. 3. Initial human sample division in a 3-D manifold embedded space. Points of different colors denote samples of different subsets.

to train a PL-SVM:

$$\min \left(\|w_1\|^2 + \lambda \cdot \sum_{n_1}^{\xi_{n_1}} \right), \ldots, \min \left(\|w_K\|^2 + \lambda \cdot \sum_{n_K}^{\xi_{n_K}} \right)$$
$$\text{s.t. } y_n \cdot F(x_n) - 1.0 + \xi_n \geq 0, \qquad \xi_n \geq 0,$$
$$n = 1, 2, \ldots, N. \tag{5}$$

The above objective function assumes that all of the local SVMs in a PL-SVM are equally important. n_k denotes the sample index in the kth sample subset. λ is a parameter to balance the training error and the SVM margins, ξ is the slack factor and $F(x)$ is the PL-SVM discriminative function defined in (4).

Since the memberships of samples to local SVMs $f_k(x) = w_k^T \cdot x + b_k, k = 1, \ldots, K$ are undermined, (5) is a programming problem with undetermined linear constrains. It can also be regarded as a kind of latent SVM with latent sample memberships. To solve (5), an iterative training procedure is described in the following section.

B. PL-SVM Training

Before training, human samples are initially divided into subsets with a K-means clustering algorithm in a manifold embedded space, as shown in Fig. 3, while all negatives are assigned to each of the subsets. Having been clustered into initial subsets, human samples assigned to the same subset have smaller differences, leading to a better sample division than a random one.

The local linear embedding (LLE) algorithm [31] is employed to construct the human manifolds. LLE computes the low-dimensional and neighborhood-preserving embeddings of the high-dimensional samples by mapping them into the low-dimensional space. Given a set of human samples in the high-dimensional feature space, LLE starts with finding nearest neighbors based on the Euclidean distance. Then LLE identifies the optimal local convex combinations of the nearest neighbors to represent each original sample. Finally it obtains an embedded space by solving a sparse eigenvector problem. More specifically, the d eigenvectors associated with the d smallest non-zero eigenvalues provide an ordered set of an

orthogonal base [31], as shown in Fig. 3 where d is set to 3 for easy visualization.

The reason of performing clustering in the embedded space instead of in the original high-dimensional feature space is to make the computation tractable. Similar views/postures form a manifold and the spatial topology of them in the embedded space is an approximation to that in the original space. Therefore, we can initially divide the samples by clustering in the embedded space. The reason of performing clustering in the embedded space instead of in the original high-dimensional feature space is to make the computation tractable. Similar views/postures form a manifold and the spatial topology of them in the embedded space is an approximation to that in the original space. Therefore, we can initially divide the samples by clustering in the embedded space. Next we develop the following iterative training procedure to obtain an approximate optimal solution to the problem in (5).

C. Training Convergence Analysis

Each of the linear SVMs of the PL-SVM is trained by sequential minimization optimization, with the following objective function:

$$\begin{aligned} &\min_{w_k} \ \tfrac{1}{2}||w_k||^2 + \lambda \cdot \sum_{n_k}^{\xi_{n_k}} \\ &\text{s.t.} \ \ y_{n_k} \cdot \text{Sign}\left(f_k(x_{n_k})\right) - 1.0 + \xi_{n_k} \geq 0, \\ &\qquad\qquad \xi_{n_k} \geq 0, \qquad n_k = 1, 2, \ldots, N_k \end{aligned} \tag{6}$$

where n_k denotes the sample index in the kth subset and N_k is the number of the samples of the subset. The convergence of the PL-SVM training is analyzed by the nearest point algorithm (NPA) [32]. Let us construct the positive convex hull U_k and the negative convex hull V_k for the kth subset, shown as the polygons in Fig. 4. Also let $\tilde{u}_k \in U_k$ and $\tilde{v}_k \in V_k$ such that

$$||\tilde{u}_k - \tilde{v}_k|| = \min_{u \in U_k, v \in V_k} ||u - v||. \tag{7}$$

Then the problem of finding $\tilde{u}_k$ and $\tilde{v}_k$ is equivalent to finding the solution of kth SVM [32]. If $(\tilde{w}_k, \tilde{b}_k)$ is the solution of the kth linear SVM $f_k(x) = \tilde{w}_k^T \cdot x + b_k$, by using the fact that from the maximum margin $2/||\tilde{w}_k|| = ||\tilde{u}_k - \tilde{v}_k||$ and $\tilde{w}_k = \delta \cdot (\tilde{u}_k - \tilde{v}_k)$ for some δ, the relation between the normal vector and the nearest point pair $(\tilde{u}_k, \tilde{v}_k)$ can be derived as [32]:

$$\tilde{w}_k = \frac{2}{||\tilde{u}_k - \tilde{v}_k||^2}(\tilde{u}_k - \tilde{v}_k), \qquad \tilde{b}_k = \frac{||\tilde{u}_k||^2 - ||\tilde{v}_k||^2}{||\tilde{u}_k - \tilde{v}_k||^2}. \tag{8}$$

By (8), we know that the margin of the kth SVM is equal to the distance between the nearest point pair$\tilde{u}_k$ and $\tilde{v}_k$. When we perform sample re-assignment in the training procedure in Algorithm I, a sample is reassigned to the subset of the SVM, to which the membership degree of the sample is the largest. According to (2)–(3), the parameterized sample-to-hyper-plane distance of this sample to the hyper-plane of this SVM is also the largest. In addition, step 2.3 of Algorithm I makes sure that the distance between the nearest point pair does not decrease. These ensure that the distances between the nearest point pairs, $(\tilde{u}_k, \tilde{v}_k), k = 1, \ldots, K$, increase monotonically (or non-decrease monotonically) in the training procedure. Consequently the margins of the SVMs increase monotonously. Since the margins are bounded, the training algorithm is thus convergent.

Algorithm 1 PL-SVM Training

Definitions: t: Iteration number; $R^{(t)}$: Number of reassigned positive samples in iteration t; $r^{(t)}$: Reassigned sample ratio in iteration t.

1. Initialization

Given a training human object set $X = \{(x_n, y_n)\}, n = 1, \ldots, N$, and K initial subsets $\left\{X_k^{(0)}\right\}, k = 1, \ldots K$, with $X = \bigcup_{k=1}^{K} \left\{X_k^{(0)}\right\}$, train K linear SVMs $\{f_k(x)\}, k = 1, \ldots, K$, as the initial PL-SVM model. Set $t = 0$.

2. Iteration

2.1. Calculate the membership degrees $C_k(x_n), k = 1, \ldots, K$, of every feature vector x_n to the K linear SVMs in the PL-SVM by (2) and (3).

2.2. For a random and unselected positive sample (x_n, y_n), select the k that maximizes the membership degree of x_n as $k = \arg\max_m \{C_m(x)\}, m = 1, \ldots, K$. Set $C_k(x) = 0.0$.

2.3. Check whether the assignment of x_n to the kth subset reduces the distance between the positive and negative convex hulls[2]. If it does, goto 2.2; otherwise assign x_n to the kth subset.

2.4. Train the linear SVMs $\{f_k(x)\}, i = 1, \ldots, K$, using the current subsets $\left\{X_k^{(t)}\right\}, k = 1, \ldots, K$.

2.5. If the reassigned sample ratio $r^{(t)}$ is larger than a predefined threshold τ, then $t \leftarrow t + 1$ and go to step 2.1; otherwise go to step 3.

3. Output

K sample subsets $\left\{X_k^{(t)}\right\}, k = 1, \ldots, K$, and a trained PL-SVM consisting of the K linear SVMs.

Step 2.3 in Algorithm I is used to ensure the monotonous increase of the SVM margins and thus the convergence of the algorithm. See the next section for the detail. The threshold τ is set to 0.02 empirically.

Fig. 4 shows an example of PL-SVM training with two subsets. The samples denoted by filled-in circles belong to subset 1 and the samples denoted by open circles belong to subset 2. The samples labeled by 1, 2, 3, 4 and $\tilde{u}_1$ form the positive convex hull for subset 1. The samples labeled by 5, 6, 7 and 8 form the positive convex hull for subset 2. Suppose that at current iteration, we obtain two nearest point pairs $(\tilde{u}_1, \tilde{v}_1)$ and $(\tilde{u}_2, \tilde{v}_2)$. Then they are used to generate the hyper-planes of SVM1 and SVM2. After steps 2.1, 2.2 and 2.3 in Algorithm I, the samples are reassigned to subset 1 or subset 2 with the maximization of membership degree criterion in (4). It can be seen from Fig. 4(b) that samples 1, 2, 5 and 7 are assigned to subset 2 and samples 3, 4, 6, 8 and $\tilde{u}_1$ are assigned to subset 1. With the new subsets, new positive convex hulls are constructed, as shown in Fig. 4(b). Then with the negative hull and new positive convex hulls, new SVMs

[2]The convex hull for a set of points in a real vector space is the minimal convex set containing the set, following the method in [32].

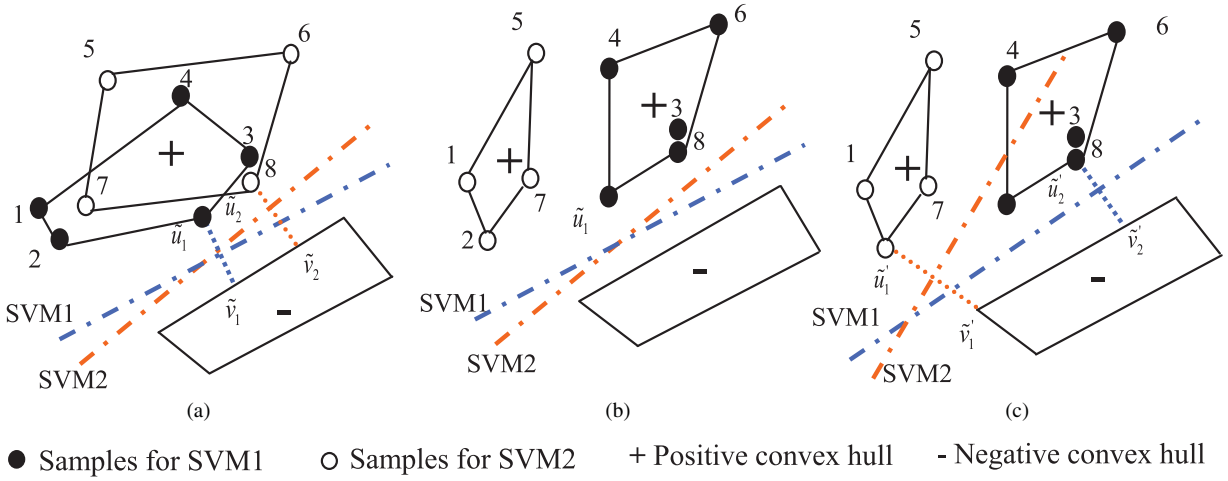


Fig. 4. Illustration of sample reassignment and convex hull changes in the training of two linear SVMs where $(\tilde{u}_1, \tilde{v}_1)$, $(\tilde{u}_2, \tilde{v}_2)$, $(\tilde{u}'_1, \tilde{v}'_1)$, and $(\tilde{u}'_2, \tilde{v}'_2)$ are nearest point pairs. (a) Convex hulls and their corresponding SVMs in the current iteration. (b) subsets after sample re-assignment. (c) Convex hulls and their corresponding SVMs in the next iteration.

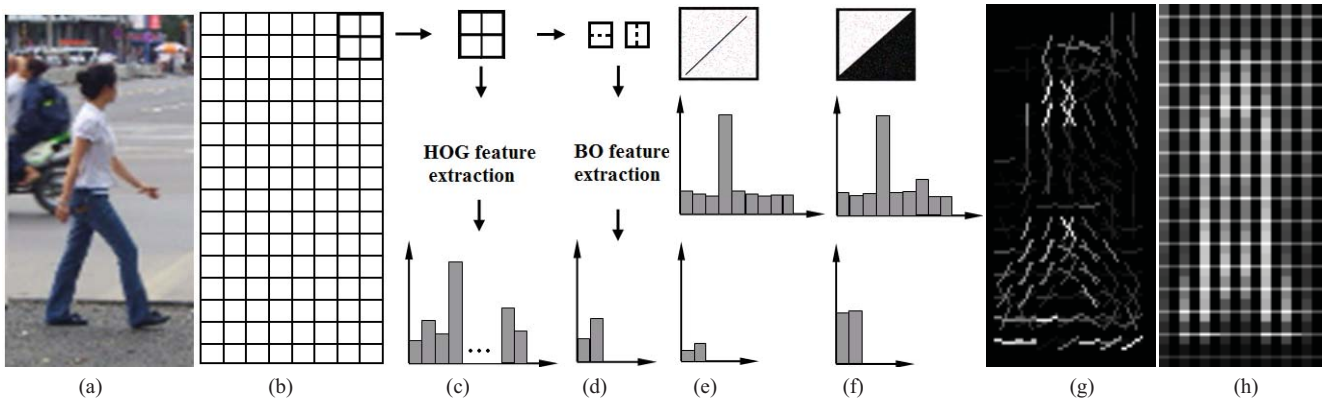


Fig. 5. HOG and BO feature extraction. (a) Human example. (b) HOG cells. (c) HOG feature extraction in a block. (d) BO feature extraction in a cell. (e) Stroke pattern in a cell (enlarged) with noise and its HOG and BO features. (f) Region pattern in a cell with noise and its HOG and BO features. (g) Visualization of the HOG features multiplying with the SVM norm vector. (h) Visualization of the BO features multiplying with the SVM norm vector.

are trained, as shown in Fig. 4(c). It can be seen that after the iteration, the margins of the SVMs increase or remain the same.

III. Human Detection

The proposed PL-SVM is incorporated with two kinds of features for human detection. A cascade detector is designed to improve detection performance.

A. Feature Representation

HOG features, proposed by Dalal and Triggs [6], are adopted for our application. As shown in Figs. 5(a)–(c), a sample of 64×128 pixels is divided into cells of size 8×8 pixels, each group of 2×2 cells is integrated into a block in a sliding fashion, and blocks overlap with each other. To extract HOG features, we firstly calculate the gradient orientations of the pixels in the cells. Then in each cell, we calculate a 9-dimensional histogram of gradient orientations as the features. Each block is represented by a 36-dimensional feature vector, which is normalized by dividing each feature bin with the vector module [6]. Each sample is represented by 105 blocks (420 cells), corresponding to a 3780-dimensional HOG feature vector.

We also propose a new kind of features, called Block Orientation (BO) features, derived from Haar-like features, as a complement to the HOG features for human detection. Each of the 420 cells is first divided into left-right and up-down sub-cells as shown in Fig. 5(d), and then the horizontal and vertical gradients of the cell are calculated by

$$Bh = \max_{c\in\{R,G,B\}} \left\{ \left| \sum_{X\in \text{ left subcell}} I_c(X) - \sum_{X\in \text{ right subcell}} I_c(X) \right| \right\}$$
$$Bv = \max_{c\in\{R,G,B\}} \left\{ \left| \sum_{X\in \text{ up subcell}} I_c(X) - \sum_{X\in \text{ down subcell}} I_c(X) \right| \right\} \quad (9)$$

where $I_c(X)$ is one of the R, G and B color values at pixel X. The BO features are the normalizations of Bh and Bv:

$$BO_h = Bh/\sqrt{Bv^2 + Bh^2 + \varepsilon}$$
$$BO_v = Bv/\sqrt{Bv^2 + Bh^2 + \varepsilon} \quad (10)$$

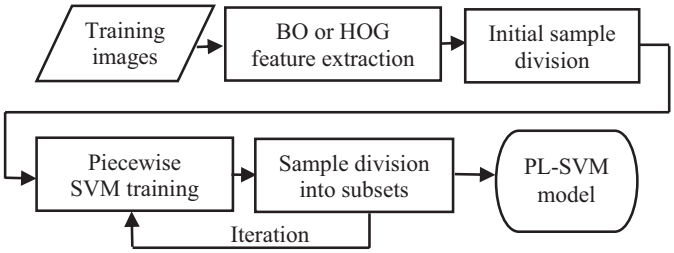


Fig. 6. Flowchart of PL-SVM training.

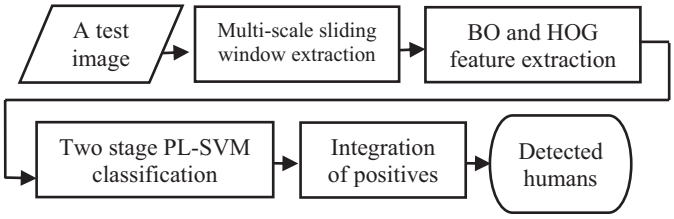


Fig. 7. Flowchart of human detection.

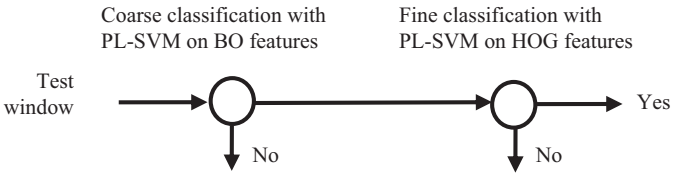


Fig. 8. Cascaded classification with two PL-SVMs on BO and HOG features, respectively.

Fig. 9. Examples of positive and negative samples from the SDL dataset.

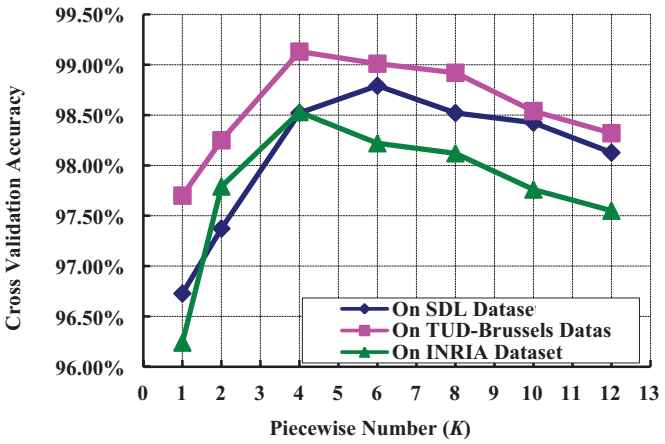


Fig. 10. Determination of the piecewise number K by cross validation.

where ε is a constant to reduce noise effect. Its value is set as $10.0 \times$ (the size of a cell), empirically.

Since the BO features are extracted on a whole cell, it is discriminative between stroke and region patterns and can depress local texture and noise. From Figs. 5(e) and f, it can be seen that for the stroke and region patterns (with noise), the HOG features are indistinctive, while the BO features can distinguish one from the other very well. In human detection, if only HOG features are used, some stroke patterns such as tree branches and railings maybe detected as parts of human bodies. With the BO features as a complement to the HOG features, we can reduce these false detections.

B. Cascade Detector With PL-SVMs

Given a set of training samples, we train two PL-SVM models, one with the BO features and the other with the HOG features, as shown in Fig. 6. In the detection procedure (Fig. 7), we apply a histogram equalization and median filtering of radius equal to 3 pixels on the test image firstly, as the preprocessing. Then the test image is repeatedly reduced in size by a factor of 1.1, resulting in an image pyramid. Sliding windows are extracted from each layer of the pyramid. In each window, the BO features are extracted and tested with the PL-SVM in the first stage. If the window is classified as a human, the HOG features will be extracted and tested with the PL-SVM of the second stage to finally decide whether it is a human or not. If the window is classified as non-human in the first stage, the second stage will not be used, as shown in Fig. 8. This scheme ensures that most of the windows are rejected in the first stage, which therefore leads to high detection efficiency. To ensure that most of the positives can pass to the second stage, we use a small threshold for the PL-SVM in the first stage. Adjusting the threshold in the second stage can balance the detections of false positives and false negatives.

IV. Experiments

In this section, we evaluate the proposed PL-SVM and compare it with the linear SVM and three state-of-the-art SVMs, kernel SVM, profile SVM and latent SVM, for human detection. When training the local linear SVMs in the iterative PL-SVM training of algorithm I, we use LIBLINEAR [33], which is designed for linear classification of a large amount of data. Both BO and HOG features are calculated with integral image methods on color [5] and gradient images [6] to improve the efficiency.

Three datasets are used in the experiments. The first one is the SDL dataset with 7550 human samples and 5769 negatives, which is publicly available [34]. In the dataset there are 258 images for testing [34]. The second one is the TUD-Brussels dataset [35], with 1167 training positives, 6759 negatives, and 508 test video frames. In our experiments, the number of training positives is doubled by flipping the images horizontally. The third one is the INRIA dataset [6], which is widely used for human detection evaluation in recent years. It has 288 test

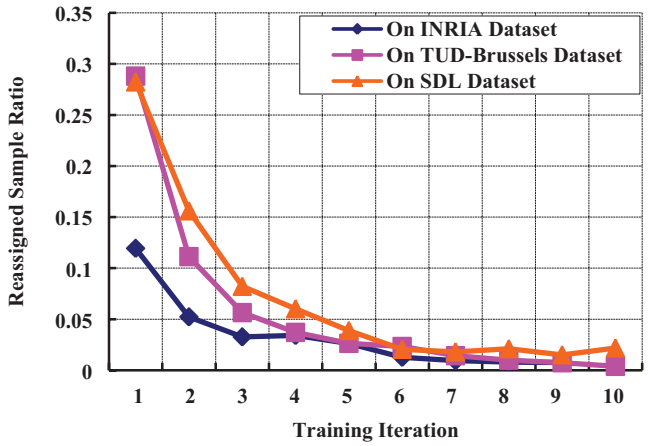


Fig. 11. Training convergence of the PL-SVM with HOG features.

Fig. 12. Human examples of six subsets from the SDL dataset. (a) Frontal or back views with legs apart. (b) Frontal or back views with legs close together. (c) Side views with legs apart. (d) Side views with legs close together. (e) Standing views different from (a)–(d). (f) On bicycles.

images, and its training set includes 2478 positives and 12180 negatives boosted from 1218 person-free photos.

The INRIA and SDL datasets contain human samples of multi-views and multi-postures (such as running, sporting, and bicycling) as shown in Fig. 9. The TUD-Brussels dataset has human samples of multiple views and postures captured from a practical driving platform. Although the samples in these datasets cannot cover all the views and postures, this large amount of samples are able to construct different manifolds. The convincible results in the experiments described later validate our approach.

A. Parameter Setting of PL-SVM

To determine the piecewise number K of a PL-SVM, we design a ten-fold cross validation. Cross validation accuracies with different piecewise numbers are tested and the K with the highest accuracy is selected, as shown in Fig. 10. In the experiments, it is found that the more the training samples, the larger the value of K is. As shown in Fig. 10, the piecewise

TABLE I

CROSS VALIDATION ACCURACY OF FOUR SVM METHODS

Classification Methods	Cross Validation Accuracy on INRIA, SDL, and TUD Datasets	Average Cross Validation Accuracy
Linear SVM [6]	96.72% 97.70% 96.24%	96.88%
Kernel SVM [20]	97.21% 98.25% 97.10%	97.52%
Profile SVM [27]	97.62% 98.51% 97.36%	97.83%
PL-SVM	98.79% 99.13% 98.53%	98.82%

TABLE II

CROSS VALIDATION ACCURACY OF FOUR SVM METHODS

Classification Methods	Cross Validation Accuracy on SDL, TUD, and INRIA Datasets	Average Cross Validation Accuracy
Linear SVM [6]	96.73% 97.31% 96.31%	96.78%
Kernel SVM [20]	97.46% 98.44% 97.71%	97.87%
Profile SVM [27]	98.60% 99.12% 98.37%	98.69%
PL-SVM	99.10% 99.46% 99.01%	99.19%

number for the SDL dataset should be six, and those for the TUD-Brussels and INRIA datasets are both four. In general, it is not true that a larger K is better due to the over-fitting problem.

The training convergence of the PL-SVM is also validated by our experiments. It can be seen from Fig. 11 that at the first iteration, there are large ratios of samples being reassigned. After about ten iterations, the ratios are close to zero (smaller than the threshold 0.02 for stopping the algorithm), showing the convergence of the training procedure.

Fig. 12 contains human examples from subsets of the SDL dataset in different views and postures when $K = 6$. We can see that the samples in each subset have similar appearances, showing that in the PL-SVM training, the division of the samples is significant. Therefore, it is expected that when these subsets are used to train the PL-SVM models, both the training and detection performances can be improved.

B. Comparison of PL-SVM With Other SVMs

To assess the PL-SVM classification method, we design another ten-fold cross validation experiment, as shown in

TABLE III
TRAINING AND DETECTION EFFICIENCY OF FIVE SVM METHODS[3]

Method	Training Time Without Boosting of Negatives (Hours)	Training Time After Five Rounds of Boosting of Negatives (Hours)	Detection Speed (Images/Second)
Linear SVM [6]	0.034	0.19	0.92
Intersection Kernel SVM [20]	0.45	2.79	0.18
Profile SVM [27]	0.17	1.03	1.50
Latent SVM [2]	—	3.2	0.40
PL-SVM	0.15	0.95	0.33 (PL-SVM of HOG) 1.6 (Cascaded PL-SVMs of BO and HOG)

TABLE IV
RECALL AND FALSE POSITIVE RATES WITH DIFFERENT THRESHOLD VALUES IN THE FIRST STAGE ON THE INRIA DATASET

Threshold in the First Stage	−0.4	−0.2	−0.1	0	0.1	0.2	0.4
Recall Rate of the First Stage	98.2%	97.4%	97.1%	95.6%	95.2%	92.3%	87.7%
FPPI of the First Stage	68.2	47.8	32.4	22.5	15.7	13.6	12.1
Detection Speed	0.5	0.8	1.2	1.5	1.6	1.8	2.1

Table 1, to compare three SVM methods with the proposed PL-SVM. It can be seen that on all the datasets the proposed PL-SVM outperforms the linear, kernel and profile SVMs.

The accuracies of the PL-SVM are 98.79%, 99.13% and 98.53% on the three datasets, respectively, on average 1.3% higher than the kernel method, which builds a histogram intersection kernel and reports good human classification performance. When compared with the recent local linear SVM (Profile SVM) [27], the PL-SVM has a 0.5–1.0% higher performance. As it is more and more difficult to improve the cross validation accuracy when it is close to 100%, 1.0–2.0% accuracy improvement is significant. When both HOG features and BO features are used, it can be seen from Table 2 that higher performances are observed in most cases. Again, the PL-SVM obtains the best accuracies. This validates that the BO-HOG features are more effective than HOG alone for human detection.

Given M dimensional features, the time complexity of a linear SVM classification is $O(M)$, as it needs only one inner product operation between the test feature vector and the normal vector. The time complexity of a kernel SVM classification is $O(SM)$, where S is the number of support vectors, as it needs S inner product operations between the test feature vector and the support vectors. The time complexity of the PL-SVM is $O(KM)$, by testing the test feature vector

[3]In the experiments, the program of Intersection Kernel SVM is downloaded from the website http://ttic.uchicago.edu/~smaji/projects/fiksvm/. The program of latent SVM is from the website http://www.cs.brown.edu/~pff/latent/.

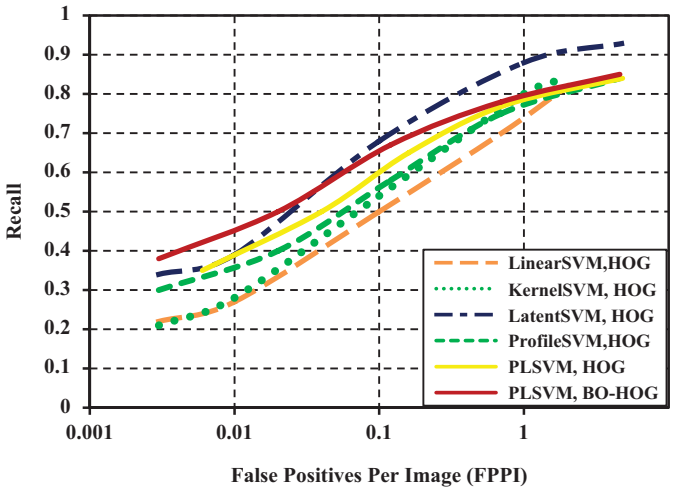


Fig. 13. Detection performance and comparison on the SDL dataset.

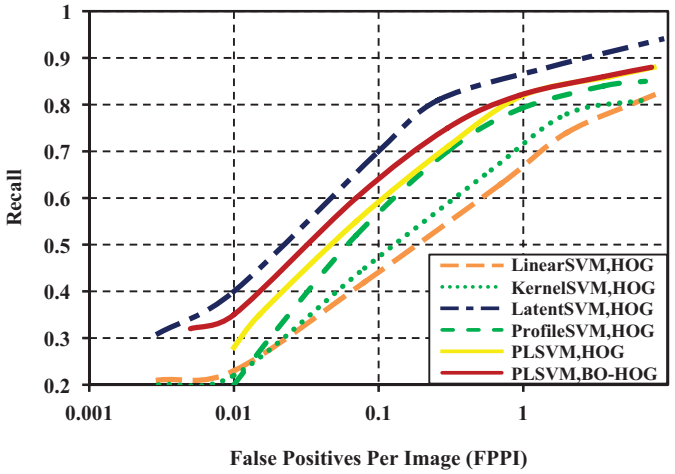


Fig. 14. Detection performance and comparison on the INRIA dataset.

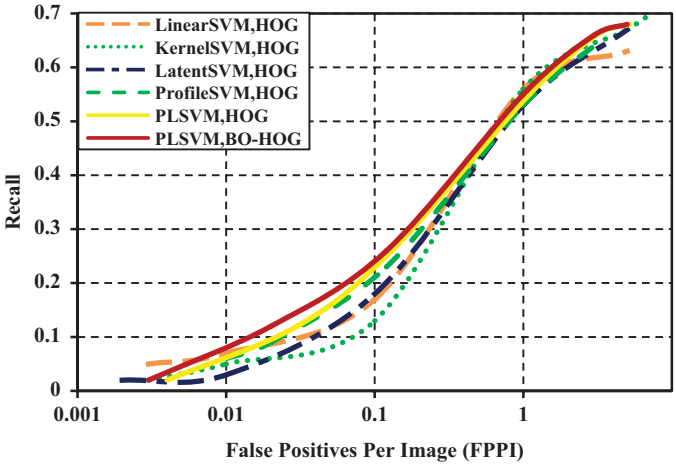


Fig. 15. Detection performance and comparison on the TUD-Brussels dataset.

Fig. 16. Detection examples from the SDL and INRIA datasets obtained by the PL-SVM.

with K linear SVMs. Since $K << S$ in general, the time complexity of the PL-SVM is much lower than that of the kernel SVM. When performing human detection, the first PL-SVM in Fig. 8 uses the BO features, which have a much lower dimensionality than the HOG features.

The training and detection efficiency of our proposed method is tested and compared with other four SVM methods. As shown in Table 3, except the linear SVM, PL-SVM is more efficient in both training and testing than the others. PL-SVM is about three times as fast as Intersection Kernel SVM and latent SVM in training. When performing detection, it runs at a speed about 1.6 images per second on average on a PC with an Intel CORE i5 CPU (fastest among all). It can be seen from the last column that the usage of BO features in the cascade detection boosts the detection speed from 0.33 images per second to 1.6 images per second. This speed is about four times as fast as the state-of-the-art latent SVM [2].

C. Human Detection Performance

In our implementation of PL-SVM, all the b_k (threshold) of the local linear SVMs are set the same in each stage. The threshold in the first stage controls the positives passed to the second stage. To ensure that most of the positives can be passed to the second stage, we use a small threshold value for PL-SVM in the first stage. It is not true that a smaller value always leads to better performance since it increases negatives passed to the second stage as well as decreases the detection speed. Table 4 shows some examples how the threshold affects the results. When it is 0.1, the first detection stage has a 95.2% recall rate with a 15.7 false positives per image (FPPI). When it is set to 0.2, both the recall rate and FPPI are reduced.

On the SDL and INRIA datasets, we evaluate the PL-SVM method with recall rate vs. false positives per image (FPPI). Adjusting the threshold in the second stage can balance the detections of false positives and false negatives. Comparisons with the linear SVM, kernel SVM, profile SVM and latent SVM [2] are also reported. It can be seen from Figs. 13 and 14 that our PL-SVM ($K = 4$ for INRIA and $K = 6$ for SDL) with HOG features outperforms the linear, kernel and profile SVMs and is comparable to the latent SVM.

In Fig. 15 we compare our method ($K = 4$) on the TUD-Brussels dataset with the other SVMs. The PL-SVM obtains

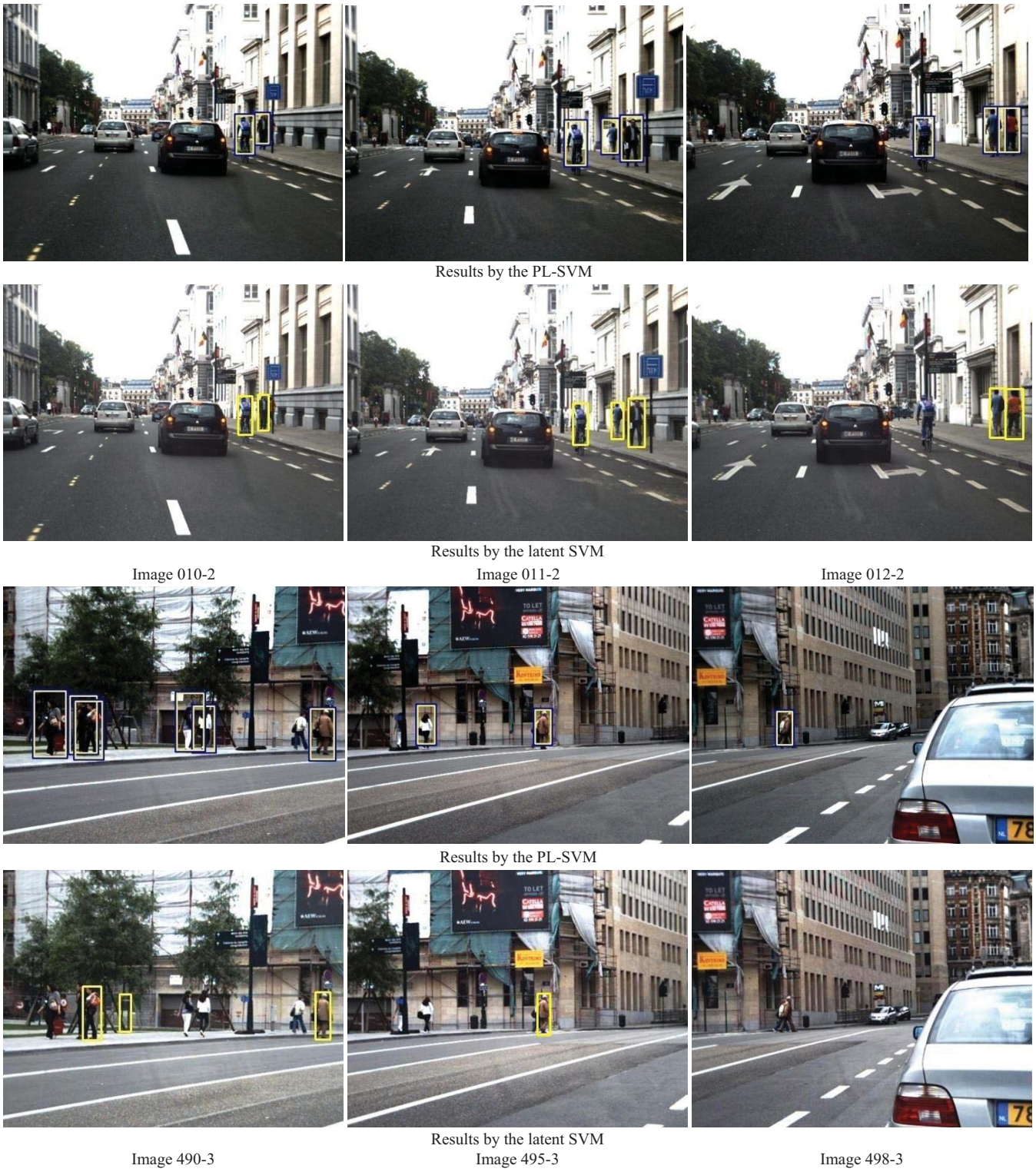


Fig. 17. Detection examples from the TUD-Brussels dataset obtained by the PL-SVM and the latent SVM.

the best result overall. When the FPPI is close to 0.1, the recall rate of our method exceeds the kernel SVM about 10%, and it also exceeds the latent SVM about 6%, showing significant performance improvement. It should be mentioned that the image scenes of this dataset are most complex, and the human regions are of low resolutions with serious occlusions. On this dataset, even the state-of-the-art method (a deformable model with a latent SVM) [2] loses its advantage for the reason that in such low resolution context, the local details of the objects are lost and therefore the deformable model cannot work well. Our method depends on pre-learned piecewise linear models to capture human objects of multi-views and multi-postures. It does not need a local deformation operation and is affected little by low resolution.

When both BO and HOG features are used, it can be seen from Figs. 13–15 that even higher performance is obtained, indicating that BO-HOG features are more effective than HOG features only for human detection.

It can be seen from Figs. 13–15 that overall PL-SVM performs better than latent SVM on the TUD Brussels dataset and the SDL dataset when FPPI is lower than 0.1. This shows the advantage of PL-SVM when a low false detection rate is required or when images contain low-resolution human regions in clutter backgrounds. PL-SVM has lower performance than latent SVM on the INRIA dataset since most of the human objects in this dataset are of high resolution, which provides what latent SVM exactly needs in its part-based detection strategy. However, in many practical applications, such as visual surveillance or driving assistant systems where humans are usually far from the camera, the resolution of captured human regions is often low. In these practical applications, our approach is more competitive.

Fig. 16 shows some detection examples, where most of the humans are correctly located with few false positives. The human objects are in multi-views with posture variations. Fig. 16(c) has humans on bicycles and Fig. 16(g) contains humans of multiple standing postures. Almost all of them are correctly located, which shows that the proposed PL-SVM can correctly capture object patterns of large variations simultaneously with the strategy of piecewise linear SVM models combined. Fig. 16(c) and (f) each has a missing positive. The missing positive in Fig. 16(c) is due to too much occlusion. In our experiments, it is found that when nearly half of an object is occluded, especially when the head-shoulder part is occluded, the object may be missed. The missing positive in Fig. 16(f) is due to the similar colors between the human and the background. In Figs. 16(h)–(j), there are some false positives, which contain animal legs, statues, and clothes in a shop window. These objects with very similar contours to humans can be falsely detected.

In Fig. 17, we show some detection examples from the TUD-Brussels dataset. The images are captured from a moving platform with dynamic backgrounds. The humans of different views, with low resolution and under clutter backgrounds are correctly detected in the video images with few missing/false positives, showing the potential of the proposed approach in video based applications, such as intelligent surveillance systems and driving warning systems. The detection results of the latent SVM are also given in Fig. 17 for comparison. It can be seen that the state-of-the-art latent SVM cannot find many humans that can be detected by the PL-SVM.

V. Conclusion

Robustness to view and posture variations is very important in human detection in practical applications, whereas it is still an open problem. In this paper, we propose a solution to this problem by developing a novel classification method called PL-SVM. The PL-SVM consists of multiple linear SVMs and has the ability to do non-linear classification. In the application of the PL-SVM to human detection, each linear SVM of the PL-SVM is responsible for one cluster of humans in a specific view or posture. All the linear SVMs combined can well tackle the multi-view and multi-posture detection problem. We have proposed a PL-SVM training algorithm that can automatically divide the feature space and train the PL-SVM with the margins of the linear SVMs increased iteratively. We have also presented the BO features as a complement to the HOG features for human detection.

Extensive experiments have been carried out to examine the performance of our method. Compared with several recent SVM methods including the linear SVM, kernel SVM, profile SVM, and latent SVM, our method reaches the state-of-the-art in both detection accuracy and its computational efficiency is even higher. Especially, it performs best when dealing with the detection of humans of low-resolutions in clutter backgrounds.

Future work includes the extension of this method to human detection from videos where not only static visual cues but also other information such as motion [36] or context information [37] is available.

Acknowledgment

The authors would like to thank Z. Liu for his valuable discussion of the manuscript, and the editor and anonymous reviewers for their constructive comments.

References

[1] Y. Xu, D. Xu, S. Lin, T. X. Han, X. Cao, and X. Li, "Detection of sudden pedestrian crossings for driving assistance systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 42, no. 3, pp. 729–739, Jun. 2008.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[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] R. Xu, B. Zhang, Q. Ye, and J. Jiao, "Cascaded L1-norm minimization learning (CLML) classifier for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 89–96.

[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, 2005.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[7] 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. Comput. Vis. Pattern Recognit.*, Jul. 2006, pp. 1491–1498.

[8] 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.

[9] 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.

[10] 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. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

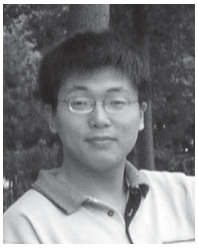
[11] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2009, pp. 32–39.

[12] 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.*, vol. 1. Oct. 2005, pp. 90–97.

[13] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[14] S. Munder and D. M. Gavrila, "An experimental study on pedestrian classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1863–1868, Nov. 2006.

[15] Q. Ye, J. Jiao, and B. Zhang, "Fast pedestrian detection with multi-scale orientation features and two-stage classifiers," in *Proc. IEEE 17th Int. Conf. Image Process.*, Sep. 2010, pp. 881–884.
[16] W. Gao, H. Ai, and S. Lao, "Adaptive contour features in oriented granular space for human detection and segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1786–1793.
[17] Y. Liu, S. Shan, W. Zhang, X. Chen, and W. Gao, "Granularity-tunable gradients partition (GGP) descriptors for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1255–1262.
[18] Z. Lin, L. Davis, and D. Doermann, "Hierarchical part-template matching for pedestrian detection and segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[19] M. Enzweiler and D. M. Gavrila, "Multilevel mixture-of-experts framework for pedestrian classification," *IEEE Trans. Image Process.*, vol. 20, no. 10, pp. 2967–2979, Oct. 2011.
[20] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[21] B. Wu and R. Nevatia, "Cluster boosted tree classifier for multi-view, multi-pose object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[22] S. Z. Li and Z. Zhang, "Floatboost learning and statistical face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1112–1123, Sep. 2004.
[23] C. H. Lampert, "An efficient divide-and-conquer cascade for nonlinear object detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1022–1029.
[24] P. Ott and M. Everingham, "Shared parts for deformable part-based models," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1513–1520.
[25] O. Oladunni and G. Singhal, "Piecewise multi-classification support vector machines," in *Proc. Int. Joint Conf. Neural Netw.*, Jun. 2009, pp. 2323–2330.
[26] S. Q. Ren, D. Yang, X. Li, and Z. W. Zhuang, "Piecewise support vector machines," *Chin. J. Comput.*, vol. 32, no. 1, pp. 77–85, 2009.
[27] H. B. Cheng, P.-N. Tan, and R. Jin, "Efficient algorithm for localized support vector machine," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 4, pp. 537–549, Apr. 2010.
[28] Y. Li, B. Liu, X. Yang, Y. Fu, and H. Li, "Multiconlitron: A general piecewise linear classifier," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 276–289, Feb. 2011.
[29] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines," Dept. Stat., Univ. Wisconsin-Madison, Madison, Tech. Rep. 1063, 2001.
[30] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularization likelihood methods," in *Proc. Adv. Large Marg. Classifiers*, 1999, pp. 61–74.
[31] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.
[32] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "A fast iterative nearest point algorithm for support vector machine classifier design," *IEEE Trans. Neural Netw.*, vol. 11, no. 1, pp. 124–136, Jan. 2000.
[33] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.
[34] Available: http://coe.gucas.ac.cn/SDL-HomePage/resource.asp
[35] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 990–997.
[36] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[37] J. Pang, Q. Huang, S. Yan, S. Jiang, and L. Qin, "Transferring boosted detectors toward viewpoint and scene adaptiveness," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1388–1400, May 2011.



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 2001, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2006.

He has been an Associate Professor with the Graduate University of the Chinese Academy of Sciences, Beijing, since 2009. His current research interests include image processing, pattern recognition, and intelligent systems.

Dr. Ye was a recipient of the Sony Outstanding Paper Award in 2005.

Zhenjun Han (M'10) received the B.S. degree in software engineering from Tianjin University, Tianjin, China, and the Ph.D. degree from the Graduate University of Chinese Academy of Sciences, Beijing, China, in 2006 and 2011, respectively.

He has been a Post-Doctoral Fellow of the Graduate University of Chinese Academy of Sciences, since 2012. His current research interests include image processing and visual surveillance.

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 of China (HIT), Harbin, China, in 1989, 1992, and 1995, respectively.

He was an Associate Professor with HIT from 1997 to 2005. Since 2006, he has been a Professor with the Graduate University of Chinese Academy of Sciences, Beijing, China. His current research interests include image processing, pattern recognition, and intelligent surveillance.

Jianzhuang Liu (M'02–SM'02) received the Ph.D. degree in computer vision from The Chinese University of Hong Kong, Hong Kong, in 1997.

He was a Research Fellow with Nanyang Technological University, Singapore, from 1998 to 2000. From 2000 to 2012, he was a Post-Doctoral Fellow, an Assistant Professor, and an Adjunct Associate Professor with The Chinese University of Hong Kong. He joined the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Beijing, China, as a Professor, in 2011. He is currently a Chief Scientist with Huawei Technologies Co., Ltd., Shenzhen, China. He has authored or co-authored more than 100 papers in journals and conferences. His current research interests include computer vision, image processing, machine learning, multimedia, and graphics.