Fast Pedestrian Detection Using a Cascade of Boosted Covariance Features

Sakrapee Paisitkriangkrai, Chunhua Shen, and Jian Zhang

***Abstract*—Efficiently and accurately detecting pedestrians plays a very important role in many computer vision applications such as video surveillance and smart cars. In order to find the right feature for this task, we first present a comprehensive experimental study on pedestrian detection using state-of-the-art locally extracted features (e.g., local receptive fields, histogram of oriented gradients, and region covariance). Building upon the findings of our experiments, we propose a new, simpler pedestrian detector using the covariance features. Unlike the work in [1], where the feature selection and weak classifier training are performed on the Riemannian manifold, we select features and train weak classifiers in the Euclidean space for faster computation. To this end, AdaBoost with weighted Fisher linear discriminant analysis-based weak classifiers are designed. A cascaded classifier structure is constructed for efficiency in the detection phase. Experiments on different datasets prove that the new pedestrian detector is not only comparable to the state-of-the-art pedestrian detectors but it also performs at a faster speed. To further accelerate the detection, we adopt a faster strategy—multiple layer boosting with heterogeneous features—to exploit the efficiency of the Haar feature and the discriminative power of the covariance feature. Experiments show that, by combining the Haar and covariance features, we speed up the original covariance feature detector [1] by up to an order of magnitude in detection time with a slight drop in detection performance.**

***Index Terms*—AdaBoost, boosting with heterogeneous features, local features, pedestrian detection/classification, support vector machine.**

I. Introduction

EFFICIENTLY and accurately detecting pedestrians is of fundamental importance for many applications in computer vision, e.g., smart vehicles, surveillance systems with intelligent query capabilities, and sports video content analysis. In particular, there is growing effort in the development of intelligent video surveillance systems. An automated method for finding humans in a scene serves as the first important preprocessing step in understanding human activity. Despite the multitude of approaches in the literature, the problem of automatic detection of objects is far from being solved (e.g., [2]–[8]). Pedestrian detection in still images is one of the most difficult examples of generic object detection. The challenges are due to a wide range of poses that humans can adopt, large variations in clothing, as well as cluttered backgrounds and environmental conditions.

Pattern classification approaches have been shown to achieve successful results in many areas of object detections. These approaches can be decomposed into two key components: feature extraction and classifier construction. In feature extraction, dominant features are extracted from a large number of training samples. These features are then used to train a classifier. During testing, the trained classifier scanned the entire input image to look for particular object patterns. This general approach has shown to work very well in detection of many different objects, e.g., face [2] and car number plate [9].

The literature on pedestrian detection is abundant. Mainly, two types of image features are used, motion and shape. Motion approaches, which require preprocessing techniques like background subtraction or image segmentation (e.g.,[10]), segments an image into so-called super pixels and then detects the human body and estimates its pose. Approaches based on shape information typically detect pedestrian directly without using preprocessing techniques [1], [3], [11], [12]. Features can be distinguished into global features and local features depending on how the features are measured. The difference between global and local features is that global features operate on the entire image of datasets whereas local features operate on the subset regions of the images. One of the well-known global feature extraction methods is principal component analysis (PCA). The drawback of global features is that the approach fails to extract meaningful features if there is a large variation in object's appearance, pose and illumination conditions. On the other hand, local features are much less sensitive to these problems since the features are extracted from the subset regions of the images. Some examples of the commonly used local features are wavelet coefficient [2], gradient orientation [11], and region covariance [1]. Local feature approaches can be further divided into whole body detection and body parts detection [13]. In the part-based approach, individual results are combined by a second classifier to form whole body detection. The advantage of using part-based approach is that it can deal with variation in human appearance due to body articulation. However, this approach adds more complexity to the pedestrian detection problem. As pointed out in [14], the classification performances reported in literature are quite different. This may be due to datasets' composition with respect to negative samples. Data sets with negative samples containing large uniform image regions typically lead to much better classification performance.

Manuscript received November 14, 2007; revised March 7, 2008 and May, 22, 2008. First published July 9, 2008; current version published August 29, 2008. NICTA is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the ARC. This paper was recommended by Associate Editor F. Pereira.

S. Paisitkriangkrai and J. Zhang are with NICTA, Neville Roach Laboratory, Kensington, NSW 2052, Australia, and also with the University of New South Wales, Sydney, NSW 2052, Australia (e-mail: paul.pais@nicta.com.au; jian.zhang@nicta.com.au).

C. Shen is with NICTA, Canberra Research Laboratory, Canberra, ACT 2601, Australia, and also with the Australian National University, Canberra, ACT 0200,Australia (e-mail: chunhua.shen@nicta.com.au).

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/TCSVT.2008.928213

The performances of several pedestrian detection approaches have been evaluated in [14]. Multiple feature-classifier combinations have been examined with respect to their receiver operating characteristic (ROC) performances and efficiency. Different features including PCA, local receptive fields (LRF) feature [12], and Haar wavelets [3] are used to train neural networks, support vector machines (SVM) [15] and k-NN classifiers. The authors conclude that the combination of SVM with LRF features performs best. An observation is that local features based detectors significantly outperform those using global features [14]. This may be due to the large variability of pedestrian shapes. Global features like PCA are more powerful for modeling objects with stable structures such as frontal faces, rigid car images taken from a fixed view angle.

Although [14] provides some insights on pedestrian detection, it has not compared state-of-the-art techniques due to the fast progress on this topic. Recently, histogram of oriented gradients (HOG) [11] and region covariance features [1] are proffered for pedestrian detection. It has been shown that they outperform those previous approaches. HOG is a gray-level image feature formed by a set of normalized gradient histograms; while region covariance is an appearance based feature, which combines pixel coordinates, intensity, and gradients, into a covariance matrix. Hence, the type of features employed for detection ranges from purely silhouette based (e.g., HOG) to appearance-based (e.g., region covariance feature). To the best of our knowledge, these approaches have not yet been compare. It remains unclear whether silhouette- or appearance-based features are better for pedestrian detection. The first part of this paper tries to answer this question. Also, in order to find the right feature for human detection, we perform a systematic experimental study on the state-of-the-art pedestrian detection techniques: LRF, HOG, and region covariance. The reasons we select the SVM classifier are: 1) it is one of the advanced classifiers and 2) it is easy to train and, unlike neural networks, the global optimum is guaranteed. Thus, the variance caused by suboptimal training is avoided for fair comparison.

Building upon the results of our experiments, we then propose a new, simpler pedestrian detector using the covariance features. Therefore, the second contribution of our work is that we show how multidimensional covariance features can be integrated with weighted linear discriminant analysis before being trained by the AdaBoost framework. In other words, the AdaBoost framework is adapted to vector-valued covariance features, and a weak classifier is designed according to the weighted linear discriminant analysis. This technique is not only faster but also accurate. In order to support our claim, we compare the performance of our proposed method with the state-of-the-art pedestrian detection techniques mentioned in [14].

The proposed boosted covariance detector achieves a detection speed that is about four times faster than the method in [1], but it is still not fast enough for real-time applications. On the one hand, the Haar feature can be computed rapidly due to its simplicity [2], but it is less powerful for classification [16]. On the other hand, although the covariance feature is a better candidate for representing pedestrians, it requires heavier computation than the Haar feature. Here, to further accelerate our proposed detector, we adopt a faster strategy—two-layer boosting with heterogeneous features—to exploit the efficiency of the Haar feature and the discriminative power of the covariance feature in a single framework. This idea has also been implemented in face detection [17] for combining Haar features with Gaussian features. It is well known that the cascade classification structure decreases the detection time by rejecting at the beginning of the cascade most of the regions in the image that do not contain a target. Thanks to the flexibility of the cascaded classifier, we employ the Haar feature-based classifiers at the beginning of the cascade and use the covariance feature at latter stages. Experiments show that, by combining the Haar and covariance features, we speed up the conventional covariance feature detector [1] by an order of detection time without greatly compromising the detection performance.

II. Feature Extraction

Feature extraction is the first step in most object detection and pattern recognition algorithms. The performance of most computer vision algorithms often relies on the extracted features. The ideal feature would be the one that can differentiate objects in the same category from objects in different categories. Commonly used low-level features in computer vision are color, texture, and shape. Here, we evaluate three state-of-the-art local features, namely, LRF, HOG, and region covariance. LRF features are extracted using multilayer perceptrons by means of their hidden layer. The features are tuned to the data during training. The price is heavier computation. HOG uses histogram to describe oriented gradient information while region covariance computes covariance from several low-level image features such as image intensities and gradients.

A. Local Receptive Fields

Multilayer perceptrons provide an adaptive approach for feature extraction by means of their hidden layer [12]. A neuron of a higher layer does not receive input from all neurons of the underlying layer but only from a limited region of it, which is call local receptive fields (LRF). The hidden layer is divided into a number of branches.

B. Histograms of Oriented Gradients

Since the development of scale-invariant feature transformation (SIFT) [18], which uses normalized local spatial histograms as a descriptor, many research groups have been studying the use of orientation histograms in other areas. The work in [11] is one of the successful examples. This work [11] proposes histogram of oriented gradients in the context of human detection. Their method uses a dense grid of histogram of oriented gradients, computed over blocks of various sizes. Each block consists of a number of cells. Blocks can overlap with each other. For each pixel $\mathbf{I}(x, y)$, the gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ is computed from

$$dx = \mathbf{I}(x+1, y) - \mathbf{I}(x-1, y) \tag{1}$$

$$dy = \mathbf{I}(x, y+1) - \mathbf{I}(x, y-1) \tag{2}$$

$$m(x, y) = \sqrt{dx^2 + dy^2} \tag{3}$$

$$\theta(x, y) = \tan^{-1}\left(\frac{dy}{dx}\right). \tag{4}$$

A local 1-D orientation histogram of gradients is formed from the gradient orientations of sample points within a region. Each histogram divides the gradient angle range into a predefined number of bins. The gradient magnitudes vote into the orientation histogram. In [11], the orientation histogram of each cell has nine bins covering the orientation range of [0,180] degrees (unsigned gradients). Hence, each block is represented by a 36-D feature vector (9 bins/cell $\times$ 4 cells/block). The final step is to combine these normalized block descriptors to form a feature vector. The feature vector can then be used to train SVMs.

C. Region Covariance

Tuzel *et al.* [1], [19] have proposed region covariance in the context of object detection. Instead of using joint histograms of the image statistics (b^d dimensions where d is the number of image statistics and b is the number of histogram bins used for each image statistics), covariance is computed from several image statistics inside a region of interest (dimensions). This results in a much smaller dimensionality. For each region, the correlation coefficient is calculated. The correlation coefficient of two random variables X and Y is given by

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)\mathrm{var}(Y)} = \frac{\mathrm{cov}(X,Y)}{\sigma_x^2\sigma_y^2} \tag{5}$$

$$\begin{aligned}\mathrm{cov}(X,Y) &= \mathbf{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \\ &= \frac{1}{n-1}\sum_k (X_k-\mu_X)(Y_k-\mu_Y)\end{aligned} \tag{6}$$

where $\mathrm{cov}(\cdot,\cdot)$ is the covariance of two random variables, μ is the sample mean, and σ^2 is the sample variance. Correlation coefficient is commonly used to describe the information we gain about one random variable by observing another random variable.

Image statistics used in this experiment are similar to the one used in [1]. The 8-D feature image used are pixel location x, pixel location y, first-order partial derivative of the intensity in horizontal direction $|\mathbf{I}_x|$, first-order partial derivative of the intensity in vertical direction $|\mathbf{I}_y|$, the magnitude $\sqrt{\mathbf{I}_x^2+\mathbf{I}_y^2}$, edge orientation $\tan^{-1}(|\mathbf{I}_y|/|\mathbf{I}_x|)$, second-order partial derivative of the intensity in the horizontal direction $|\mathbf{I}_{xx}|$, second-order partial derivative of the intensity in the vertical direction $|\mathbf{I}_{yy}|$. The covariance descriptor of a region is an 8×8 matrix. Due to the symmetry, only the upper triangular part is stacked as a vector and used as covariance descriptors. The descriptors encode information of the correlations of the defined features inside the region. Note that this treatment is different from that in [1] and [19], where the covariance matrix is directly used as the feature and the distance between features is calculated in the Riemannian manifold.[1] However, eigen-decomposition is involved for calculating the distance in the Riemannian manifold. Eigen-decomposition is very computationally expensive ($O(d^3)$ arithmetic operations). We instead vectorize the symmetric matrix and measure the distance in the Euclidean space, which is faster.

[1]Covariance matrices are symmetric and positive semi-definite, hence they reside in the Riemannian manifold.

Preliminary experiments, similar to the experiment described in [19], have been conducted to compare the two different distance measures: distance of the correlation coefficient from two covariance matrices in the Euclidean space and distance of two covariance matrices in the Riemannian manifold. The results indicate that their performance on pedestrian detection are quite similar.

In order to improve the covariance matrices' calculation efficiency, a technique which employs integral image [2] can be applied [19]. By expanding the mean from previous equation, covariance equation can be written as

$$\mathrm{cov}(X,Y) = \frac{1}{n-1}\Big[\sum_k X_kY_k - \frac{1}{n}\sum_k X_k\sum_k Y_k\Big]. \tag{7}$$

Hence, to find the fast covariance in a given rectangular region, the sum of each feature dimension, e.g., $\sum_k X_k$ and $\sum_k Y_k$, and the sum of the multiplication of any two feature dimensions, e.g., $\sum_k X_kY_k$, can be computed using the integral image.

The extracted covariance features assume that the image statistics follow a single Gaussian distribution. Although this assumption may look overly simple, experiments prove the covariance features' efficacy. Jin *et al.* [20] have used an identical idea for network intrusion detection.

III. Classifiers

There exist many classification techniques that can be applied to object detection. Some of the commonly applied classification techniques are SVM [15] and AdaBoost [2], [21].

A. Support Vector Machines

SVM is one of the popular large margin classifiers [15] that has a very promising generalization capacity. Due to space limit, we omit details of SVM. The reader is referred to [15] for details. In our experiments, SVM classifiers with three different kernel functions, linear, quadratic, and RBF kernels, are compared with the features calculated from previous section.

B. AdaBoost

AdaBoost is the first practical and efficient algorithm for ensemble learning [21]. The training procedure of AdaBoost is a greedy algorithm, which constructs an additive combination of weak classifiers such that the exponential loss $L(y, f(\boldsymbol{x})) = e^{-yf(\boldsymbol{x})}$ is minimized. Here, $\boldsymbol{x}$ is the labeled training examples and y is its label; $f(\boldsymbol{x})$ is the final decision function which outputs the decided class label. AdaBoost iteratively combines a number of weak classifiers to form a strong classifier. A weak classifier is defined as a classifier with accuracy on the training set greater than average. The final strong classifier $H(\cdot)$ can be defined as $H(\boldsymbol{x}) = \mathbf{sign}\big(\sum_{i=1}^{N_w}\alpha_i h_i(\boldsymbol{x})\big)$, where α_i is a weight coefficient, $h_i(\cdot)$ is a weak learner, and N_w is the number of weak classifiers. At each new round, AdaBoost selects a new hypothesis $h(\cdot)$ that best classifies training samples with minimal classification error. Each training sample receives a weight that determines its probability of being selected for a training set. If a training sample is correctly classified, then its probability of being used again in a subsequent component classifier is reduced. Conversely, if the pattern is misclassified, then its

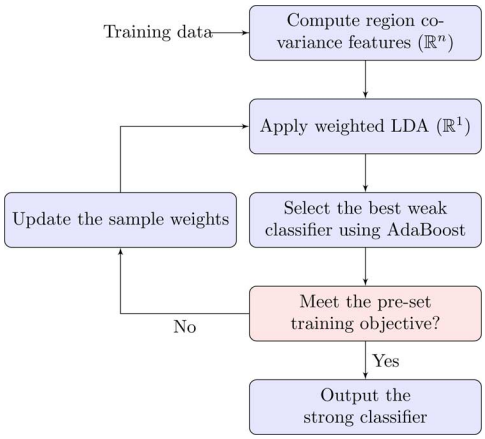


Fig. 1. Architecture of the proposed pedestrian detection system using boosted covariance feature. We set the training objective as detection rate: 99.5%; false positive rate: 50%.

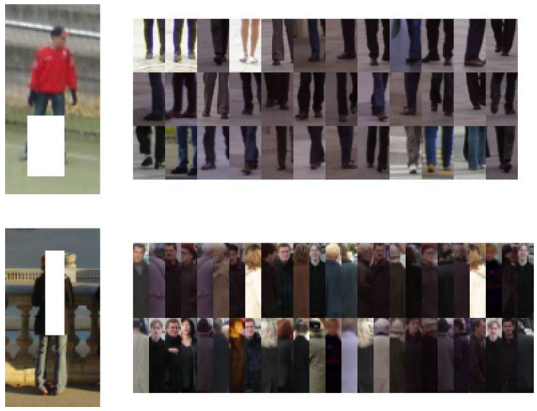


Fig. 2. First and second covariance region selected by AdaBoost. The first two covariance regions overlayed on human training samples are shown in the first column. The second column displays human body parts selected by AdaBoost. The first covariance feature represents human legs (two parallel vertical bars) while the second covariance feature captures the information of the head and the human body.

probability of being used again is increased. In this way, the algorithm focuses more on the misclassified samples after each round of boosting.

Since Viola *et al.* [2] introduced AdaBoost into computer vision for face detection, extensions have been proposed for better classification performance [22], fast training [23], or dealing with imbalanced training data [24]. These techniques can be applied to our problem. We leave this as a future research direction.

IV. Boosted Covariance Features

Here, we describe a new pedestrian detector system. Fig. 1 shows the structure of the new pedestrian detector. We use the covariance feature originally invented in [19]. The reasons why we choose this local feature will be explained in detail in Section IV-A.

Given the training dataset, each training sample is assigned a weight which determines its probability of being selected for a training set. From a set of given rectangular windows, covariance matrix is calculated from several low-level image statistics within a rectangular region. The upper triangular part of computed covariance matrix is stacked as a vector and used as a covariance descriptors. A vector of covariance descriptors is projected onto a 1-D space using the algorithm described in Section IV-A. AdaBoost is then applied to select the best rectangular region w.r.t. the weak learner that best classifies training samples with minimal classification error. The best weak learner is added to a cascade. Weak learners are added until the predefined classification accuracy is met. The process is replicated for the next stage of the cascades.

This section begins with a short explanation of Fisher linear discriminant analysis (LDA) concept. We then extend these methods to varying weighted training samples. Finally, we describe in details how to apply these techniques to train multidimensional covariance features on a cascade of AdaBoost classifiers framework.

A. Weighted Fisher Linear Discriminant Analysis

The objective of the Fisher's criteria is to find a linear combination of the variables that can separate two classes as much as possible. However, the criterion proposed by Fisher assumes uniformly weighted training samples. In AdaBoost training, each data point is associated with a weight which measures how difficult to correctly classify this data. Therefore, we need to apply a weighted version of the standard Fisher linear discriminant analysis (WLDA). Similar to LDA, WLDA finds a linear combination of the variables that can separate two classes as much as possible with emphasis on the training samples with high weights.

It is well known that the choice of weak classifiers is vital to the classification accuracy of boosting techniques. Although effective weak classifiers increase the performance of the final strong classifier, the large amount of potential features make the computation prohibitively heavy with the use of complex classifiers such as SVMs. For scalar features such as Haar features in [2], [4], a very efficient stump can be used. For vector-valued features such as HOG or covariance features, unfortunately, seeking an optimal linear discriminant would require much longer time. As shown in [25], it is possible to use linear SVMs as weak learners, the training procedure is very time-consuming. Here we adopt a more efficient approach. We project the multi-dimensional covariance features onto a 1-D line using WLDA, which finds a linear projection function which guarantees optimal classification of normally distributed samples of two classes.

Each weak learner can then be defined as

$$h(\boldsymbol{x}) = \begin{cases} +1, & \text{if } \boldsymbol{w}^\top \boldsymbol{x} > \theta \\ -1, & \text{otherwise} \end{cases} \tag{8}$$

where $h(\cdot)$ defines a weak learner, $\boldsymbol{x}$ is the calculated covariance features, and θ is an optimal threshold such that the minimum number of examples are misclassified.

B. Cascade of Covariance Descriptors

The covariance feature efficiently captures the relationship between different image statistics. Combining with WLDA, this information can be used to represent a distinct part of the human body. At each AdaBoost iteration, a simple classifier is trained from the collection of region covariance features. The experimental results show that the covariance region selected by AdaBoost are physically meaningful and can be easily interpreted

as shown in Fig. 2. The first selected feature focuses on the bottom part of the human body while the second selected feature focuses on the top part of the body. It turns out that covariance features are well adapted to capture patterns that are invariant to illumination changes and human poses/appearance changes.

Our fast boosted covariance features-based detection framework is summarized in Algorithm 1.

Algorithm 1. The training algorithm for building the cascade of boosted covariance detector.

Input:

- A positive training set and a negative training set;
- $D_{\min}$: minimum acceptable detection rate per cascade level;
- $F_{\max}$: maximum acceptable false positive rate per cascade level;
- F_{target}: target overall false positive rate.

Initialize: $i = 0$; $D_i = 1$; $F_i = 1$;

while $F_{\text{target}} < F_i$ **do**
$i = i + 1$; $f_i = 1$;

while $f_i > F_{\max}$ **do**

(1) Normalize AdaBoost weights;

(2) Calculate the projection vector $\boldsymbol{w}$ with WLDA; and project the covariance features to 1D;

(3) Train decision stumps by finding a optimal threshold θ, using the training set;

(4) Add the best decision stump classifier into the strong classifier;

(5) Update sample weights in the AdaBoost manner;

(6) Lower threshold such that $D_{\min}$ holds;

(7) Update f_i using this threshold.

end

$D_{i+1} = D_i \times D_{\min}$; $F_{i+1} = F_i \times f_i$; and remove correctly classified negative samples from the training set;

if $F_{\text{target}} < F_i$ **then**

Evaluate the current cascaded classifier on the negative images and add misclassified samples into the negative training set.

end

end

Output:

- A cascade of boosted covariance classifiers for each cascade level $i = 1, \ldots$;
- Final training accuracy: F_i and D_i.

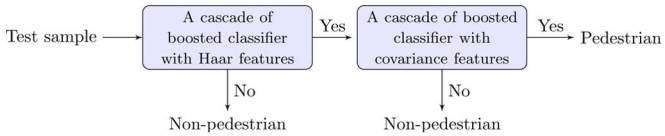


Fig. 3. Structure of our two-layer pedestrian detector.

In order to reduce computation time, a cascade of classifiers is built [2]. The key insight is that efficient boosted classifiers, which can reject many of the simple nonpedestrian samples while detecting almost all pedestrian samples, are constructed and placed at the early stages of the cascades. Time-consuming and complex boosted classifiers, which can remove more complex nonpedestrian samples, are placed in the later stages of the cascades. By constructing classifiers in this way, we are able to quickly discard simple background regions of the image, e.g., sky, building, or road, while spending more time on pedestrian-like regions. Only samples that can pass through all stages of the cascades are classified as pedestrians.

V. Two-Layer Boosting With Heterogeneous Features

In order to further accelerate our proposed detector, an approach which consists of a two-layer cascade of classifiers is built [17]. The objective of designing a two-layer approach is to achieve high detection speed and accuracy. The idea is to place simple and fast-to-compute features in the first layer while putting a more accurate but slower-to-compute features in the second layer of the cascade. The simple features filter out most simple nonpedestrian patterns in the early stage of the cascade.

Haar wavelet features have proved to be extremely fast and highly powerful in the application of face detections [2]. However, the Haar feature performs poorly in the context of human detection as reported in [4]. In order to improve the overall accuracy, we apply boosted covariance features in the second layer. In other words, Haar features are used in the first cascade while boosted covariance features are used in the second cascade. This way we utilize the efficiency of the Haar feature and the discriminative power of the covariance feature in a single framework. Fig. 3 shows the detector architecture of the two-layer approach.

We experimentally evaluate covariance features and Haar features by training two different classifiers on the same training set using AdaBoost. The positive training set is extracted from INRIA dataset [11] which consists of 2416 human samples (mirrored). The negative training set comes from random patches extracted from negative images. The classifiers are evaluated on the INRIA test set. Fig. 4 gives a comparison of the performances of different feature types. The following observation can be made from the figure. The test error decreases quickly with the number of AdaBoost iterations for all features. The test error of covariance features run into saturation after about 100 iterations while the test error rate of Haar feature continues to decrease slowly. The results can also be interpreted in terms of the number of selected features and test error rate. For example, it is possible to achieve a 5% test error rate using either 25 covariance features or 100 Haar features. Table I shows the computation time for different feature types (including computation

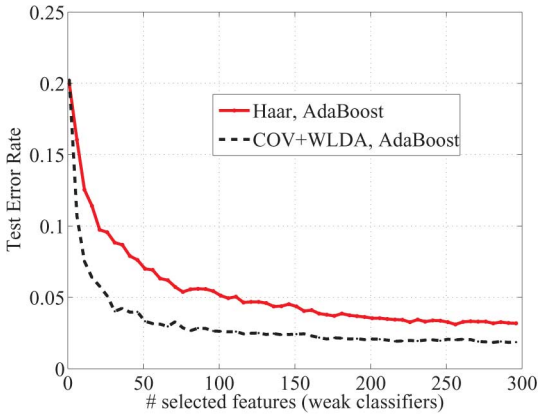


Fig. 4. Performance comparison of covariance and Haar features on INRIA test set [11].

TABLE I
AVERAGE TIME REQUIRED TO EVALUATE COVARIANCE AND HAAR FEATURES

# features	COV (μ-seconds)	Haar (μ-seconds)
20	71	5
50	137	7
100	250	11
200	490	20
300	715	29

overhead of integral images). The computation of Haar features is much faster than the computation of covariance features.

Due to the flexibility of the cascaded structure, it is easy to integrate multiple heterogeneous features. Although we use Haar and covariance features here, some combination of various features may lead to better performance. It remains a future study topic on how to find the best combination.

VI. EXPERIMENTS

We evaluate the performance of our techniques on two publicly available datasets, dataset of [14] and INRIA dataset [11]. The first dataset [14] contains a set of extracted pedestrian and non-pedestrian samples which are scaled to size 18×36 pixels. We conduct three experiments on the dataset of [14] using covariance features trained with SVM and AdaBoost. The second dataset [11] contains 1176 pedestrian samples from 288 images. We conduct two experiments using covariance features trained with AdaBoost. To our knowledge, [11] and [1] are the state-of-the-art on human detection in the literature. Hence, we mainly compare our algorithm with these two techniques.

The experimental section is organized as follows. First, the datasets used in this experiment, including how the performance is analyzed, are described. Experiments and the parameters used to achieve optimal results are then discussed. Finally, experimental results and analysis of different techniques are compared. In all experiments, associated parameters are optimized via cross-validation.

A. Experiments on DaimlerChrysler Dataset With SVM

We first use the dataset in [14]. This dataset consists of three training sets and two test sets. Each training set contains 4800 pedestrian examples and 5000 nonpedestrian examples The pedestrian examples were obtained from manually labeling and extracting pedestrians in video images at various time and locations with no particular constraints on pedestrian pose or clothing, except that pedestrians are standing in an upright position. Pedestrian images are mirrored and the pedestrian bounding boxes are shifted randomly by a few pixels in horizontal and vertical directions. A border of 2 pixels is added to the sample in order to preserve contour information. All samples are scaled to size 18×36 pixels. Performance on the test sets is analyzed similarly to the techniques described in [14]. For each experiment, three different classifiers are generated. Testing all three classifiers on two test sets yields six different ROC curves. A 95% confidence interval of the true mean detection rate is given by the t-distribution.

1) Experiment Setup: We train covariance features with various combination of SVMs. For this method, we concatenate the covariance descriptors for all regions into a combined feature vector. An SVM classifier is trained using this feature vector. Our preliminary experiments show that training Gaussian kernel SVM with region of size 7×7 pixels, shifted at a step size of 2 pixels over the entire input image of size 18×36 gives optimal results. Increasing the region width and step size decreases the performance slightly. The reason is that increasing the region width and step size decreases the feature length of covariance descriptors to be trained by SVM.

In contrast, training a linear SVM with region of size 7×7 pixels gives a very poor performance (all positive samples are misclassified). We suspect that the region size is too small. As a result, calculated covariance features of positive and negative samples can not be separated by linear hyperplane. In our experiments, the feature length of covariance descriptors per training samples is between 1,000–2,000 features. The length is proportional to the number of image statistics used and the total number of regions used for calculating covariance.

For the HOG features, the configurations reported in [11] are tested on the benchmark datasets. However, our preliminary results show a poor performance. This is due to the fact that the resolution of benchmark datasets used (18×36 pixels) is much smaller than the resolution of the original datasets (64×128 pixels). In order to achieve a better result, HOG descriptors are experimented with various spatial/orientation binning and descriptor blocks (cell size ranging from 3 to 8 pixels and block size of 2×2–4×4 cells). From our experimental results, we have decided to use a cell size of 3×3 pixels with a block size of 2×2 cells, descriptor stride of 2 pixels, and 18 orientation bins of signed gradients (total feature length is 8064) to train SVM classifiers.

2) Results Based on SVM on the Dataset of [14]: LRF features with quadratic SVM is the best approach among the features compared in [14]. For completeness, we compare it with our results.

Fig. 5 shows detection results of covariance features trained with different SVM classifiers. When trained with the RBF SVM, a region of size 7×7 pixels turns out to perform best compared with other region sizes. From the figure, region covariance features perform better than LRF features when trained with the same SVM kernel (quadratic SVM).

Fig. 6 shows detection results of HOG features trained with different SVM classifiers. From the figure, it clearly indicates that a combination of HOG features with quadratic SVM performs best. Obviously, the nonlinear SVM outperforms the linear SVM. It is also interesting to note that the linear SVM trained using HOG features performs better than the

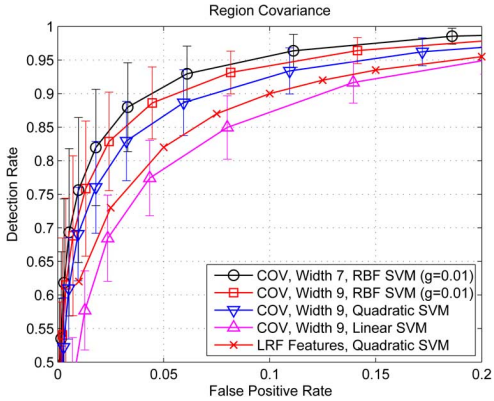


Fig. 5. Performance of different parameters on region covariance features.

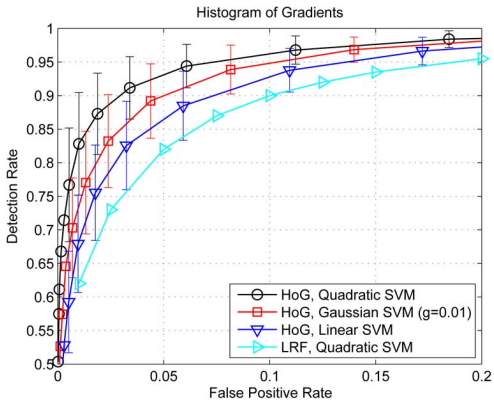


Fig. 6. Performance of different classifiers on histogram of oriented gradients features.

nonlinear SVM trained using LRF features. This means that HOG features are much better at describing spatial information in the context of human detection than LRF features.

From these two experiments, we know that although LRF is considered the best local feature for human detection in [14], it cannot compete with region covariance and HOG.

We have also compared the covariance and HOG features on the MIT CBCL datasets.[2] Both HOG and covariance features perform extremely well on this MIT dataset. This is not too surprising knowing that the MIT dataset contain only a frontal view and rear view of human. Less variation in human poses makes the classification problem much easier for SVM classifiers. It is also interesting to note that the performance of covariance features (with Gaussian RBF SVM) is very similar to HOG features trained using Gaussian RBF and quadratic SVM. It even outperforms HOG features at a low false positive rate. We may conclude that in terms of classification performance, covariance features are the best among the three local features we have compared.

B. Experiments on DaimlerChrylser Dataset With a Cascade of Boosted Covariance Features

1) Experiment Setup: For a boosted cascade of covariance features, we generate a set of overcomplete rectangular covariance filters and subsample the overcomplete set in order to keep

[2][Online]. Available: http:// cbcl.mit.edu/ software-datasets/ PedestrianData.html

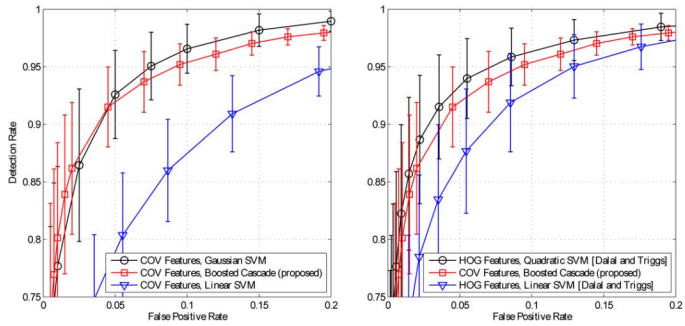


Fig. 7. Performance comparison of our cascade of boosted covariance features with covariance features trained using SVM (left) and histogram of oriented gradients (HOG) features trained using SVM (right).

a manageable set for the training phase. The set contains approximately 1120 covariance filters. Each filter (weak classifier) consists of four parameters, e.g., x-coordinate, y-coordinate, width, and height. A strong classifier consisting of several weak classifiers is built in each stage of the cascade. At each stage, weak classifiers are added until the predefined objective is met. In this experiment, we set the minimum detection rate to be 99.5% and the maximum false positive rate to be 50% in each stage. The negative samples used in each stage of the cascade are collected from false positives of the previous stage of the cascade.

Since the resolution of the test samples is quite small, we extend the border of each test sample by one pixel. The extra margin helps shifting the pedestrian in the test sample to the center. Doing so increases a flexibility of our boosted classifier. During classification, we count the number of the positively classified subwindows and use this number to test whether the test sample is pedestrian or non-pedestrian.

2) Results Based on Boosted Covariance Features on the Dataset of [14]: Fig. 7 shows detection results of covariance features trained with AdaBoost. The performance of our proposed method is very similar to the best performance of covariance features with Gaussian SVM. It also performs better than HOG features with linear SVM. However, the performance is slightly worse compared with the performance of HOG features with quadratic SVM.

We have also applied bootstrapping technique to HOG [11] and covariance features. Bootstrapping is applied iteratively, generating 10 000 new nonpedestrian samples at each iteration. It is observed that collecting the first 10 000 new nonpedestrian samples did not take long, but the second iteration took a long time. This is exactly to be expected since the new classifier has better accuracy than the previous classifier. We observe that the improvement of training HOG feature using bootstrapping technique over initial classifier is up to 7% increase in detection rate at 2.5% false positives rate while the improvement is slightly lower in covariance features (about 3% increases at 2.5% false positives rate). However, this performance gain comes at a higher computation cost for training.

Finally, a comparison of the best performing results for different feature types are shown in Fig. 8. The following observations can be made. Out of the three features, both HOG and covariance features perform much better than LRF. HOG features is slightly better than covariance features. [1] concludes that the covariance descriptor outperforms the HOG descriptor (using

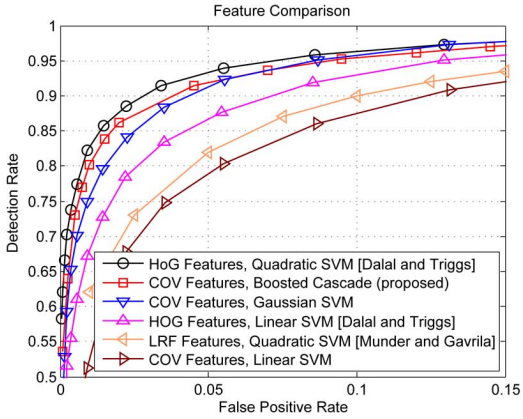


Fig. 8. Performance comparison of the best classifiers for different feature types on the dataset of [14].

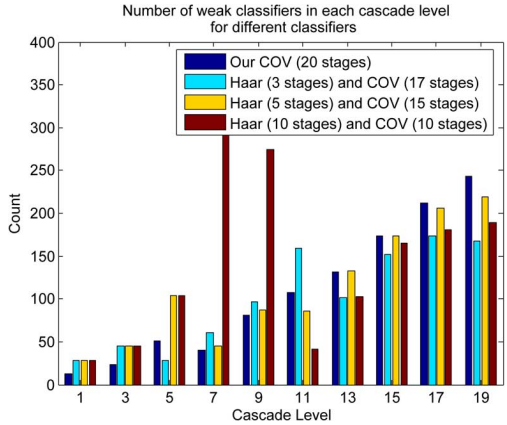


Fig. 9. Number of weak classifiers in different cascade levels on the dataset of [14]. Note that adding Haar features as a preprocessing step does not vary the number of covariance features in later stages of cascade much.

TABLE II
AVERAGE TIME REQUIRED TO EVALUATE 10 FRAMES OF A SEQUENCE OF 384 × 288 PIXELS IMAGES. EACH IMAGE CONSISTS OF 17 280 WINDOWS (SCALE FACTOR OF 0.8 AND STEP-SIZE OF 4 PIXELS)

	windows per sec	seconds per frame
HOG, Quadratic SVM	25	714
HOG, Linear SVM	4800	3.6
Our COV approach	6000	2.9

TABLE III
AVERAGE EVALUATION TIME IN WINDOWS PER SECOND FOR DIFFERENT PARAMETERS OF THE TWO-LAYER BOOSTING APPROACHES

	windows per sec
Our COV (20 stages)	6,000
Haar (3 stages) and COV (17 stages)	30,000
Haar (5 stages) and COV (15 stages)	50,000
Haar (10 stages) and COV (10 stages)	100,000
Haar (20 stages)	200,000

human datasets of size 64 × 128 pixels with LogitBoost classification). We suspect the difference would be in the resolution of datasets and the classifiers used. Small resolution datasets give less number of covariance features than large resolution data sets. To support our findings, we conduct experiments on INRIA dataset [14] with a resolution of 18 × 36 pixels and include the results at the end of Section VI-E.

We can see that gradient information is very helpful in human detection problems. In all experiments, nonlinear SVMs (quadratic or Gaussian RBF SVM) improves performance significantly over the linear one. However, this comes at the cost of a much higher computation time (approximately 50 times slower in building SVM models).

Experiments show that most false negatives are due to the subject's pose deformation, occlusions, or the very difficult illumination environments. False positives usually contain gradient information which looks like human body boundaries.

The advantages of our proposed method over features trained using SVM are ease of parameter tuning and much faster detection speed. SVM has more parameters compared to the boosted cascade, e.g., tradeoff between training error and margin or parameters of the nonlinear kernel. These parameters need to be manually optimized for the specific classification task using cross validation. In the next experiment, we compare the processing speed in windows per second of the two best classifiers: HOG with quadratic SVM and 20 stages of boosted covariance features. We apply the two classifiers to a sequence of 10 images with resolution of 384 × 288 pixels in width and height. Table II shows the average detection speed for the two classifiers. As expected, the detection speed of 20 stages of boosted covariance features is much faster than the detection speed of the nonlinear SVM classifier.

C. Experiments on DaimlerChrysler Dataset With Two-Layer Boosting

1) Experiment Setup: We generate a set of overcomplete Haar wavelet filters and subsample the overcomplete set. The set of Haar features that we use to train the cascade contained 20 547 filters: 5540 vertical two-rectangle features, 5395 horizontal two-rectangle features, 3592 vertical three-rectangle features, 3396 horizontal three-rectangle features, and 2624 four-rectangle features. From the preliminary experiments on signed and unsigned wavelets, we observe that the performance of signed wavelets outperform unsigned wavelets. Hence, we preserve the sign of intensity gradients in this experiment. For covariance features, we use a set of rectangular covariance features generated from previous section. Fig. 9 gives some details about our two-layer boosting cascade.

2) Results Based on Multilayer Boosting: Table III shows the evaluation time in windows per second for different hybrid configurations. Adding more stages of Haar wavelet features as a preprocessing step increases the detection speed approximately *exponentially*. Fig. 10 shows the performance of our two-layer boosting. The curve of our method is generated by adding one cascade level at a time. The boosted covariance features outperforms all other approaches. The performance of hybrid classifiers is quite poor at high false positive rate due to Haar-like features in the initial stages of the cascade. Nonetheless, the performance improves as more covariance features have been added to the later stages of the cascade.

D. Experiments on INRIA Human Dataset With AdaBoost

The dataset consists of one training set and one test set. The training set contains 1208 pedestrian samples (2416 mirrored

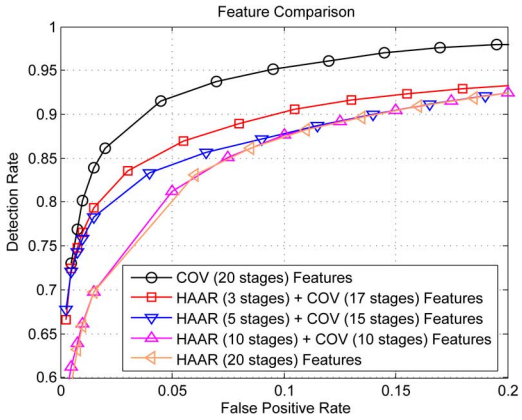


Fig. 10. Performance comparison of the two-layer boosting approach and a cascade of the boosted covariance features on the dataset of [14]. The two-layer boosting approach performs comparable to the cascade of boosted covariance features at low false positive rate (< 0.01), which is the range of interest.

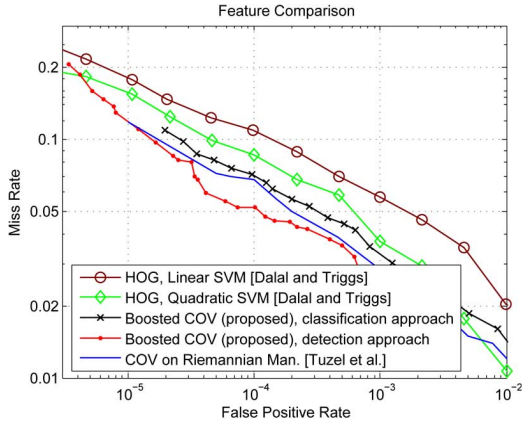


Fig. 11. Performance comparison of our cascade of boosted covariance features with HOG with linear SVM [11] and covariance features on Riemannian manifold [1]. The curve of covariance on Riemannian manifold is reproduced from [1].

samples) and 1200 nonpedestrian images. The pedestrian samples were obtained from manually labeling images taken from a digital camera at various time of the day and various location. The pedestrian samples are mostly in standing position. A border of 8 pixels is added to the sample in order to preserve contour information. All samples are scaled to size 64×128 pixels. The test set contains 1176 pedestrian samples (mirrored) extracted from 288 images.

We evaluate the performance of our classifiers on the given test set using classification approach and detection approach. For human classification, we used cropped human samples taken from the test images. During classification, the number of the positively classified windows is used to determine if the test sample is human or nonhuman. For human detection, a fixed size window is used to scan the test images with a scale factor of 0.95 and a step size of 4 pixels. As in [1], mean shift clustering [26] is used to cluster multiple overlapping detection windows. Simple rules as in [2] are also applied on the clustering results to merge those close detection windows.

The criteria similar to the one used in PASCAL VOC Challenge [27] is adopted here. Detections are considered true or false positives based on the area of overlap with ground truth bounding boxes. To be considered a correct detection, the area of overlap between the predicted bounding box B_p and ground truth bounding box B_{gt} must exceed 40% by

$$a_0 = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})} > 40\%.$$

Multiple detections of the same object in an image are considered false detections. For quantitative analysis, we plot miss rate versus false positive per window tested (false positive rate) curves on a log–log scale. The experiments are conducted using a standard desktop with 2.8-GHz Intel Pentium-D CPU and 2-GB RAM.

1) Experiment Setup: Similar to the previous experiments, we generate a set of overcomplete rectangular covariance filters and subsample the overcomplete set in order to keep a manageable set for the training phase. The set contains approximately 15 225 covariance filters. In each stage, weak classifiers are added until the predefined objective is met. In this experiment, we set the minimum detection rate to be 99.5% and the maximum false positive rate to be 50% in each stage. Each stage is trained with 2416 human samples and 5000 nonhuman samples. The negative samples used in each stage of the cascade are collected from false positives of the previous stages of the cascade. The final cascade consists of 29 stages.

2) Results Based on Boosted Covariance Features: Fig. 11 shows a comparison of our experimental results with different methods. The curve of our method is generated by adding one cascade level at a time. From the figure, it can be seen that our system's performance is much better than HOG with linear SVM [11] while achieving a comparable detection rate to the technique described in [1]. [1] calculates distance between covariance matrix on the Riemannian manifold. An eigen-decomposition is required which slows down the computation speed [1]. In contrast, our approach avoids the eigen-decomposition and therefore it is much faster. It is also easier to implement. The figure also shows the performance of our system on human detection problem. In order to achieve the results at low false positive rate i.e., $< 10^{-5}$, we manually adjust the minimum neighbor threshold (a number of merged detections). From Fig. 11, our covariance technique with detection approach outperforms the same technique with classification approach. The reason is due to the clustering and merging techniques we used. By clustering and merging multiple overlapping detection windows, we are able to further reduce the number of false detections. As a result, the curve is slightly shifted to the left. As for the processing time, on average our unoptimized implementation in C++ can search about 12 000 detection windows per second. Due to the cascade structure, the search time is faster when human is against plain backgrounds and slower when human is against more complex backgrounds. Table IV shows the average detection speed for three different classifiers. Compared with [11] and [1], our search time is faster than both techniques (2.2 times faster than [11] and 4 times faster than [1]). Note that the system in [1] is implemented in C++ on a Pentium-D 2.8-GHz processor with 2-GB RAM, which is the same as ours.[3]

In the next experiment, we show how adding a cascade of Haar wavelet features as a preprocessing to a cascade of

[3]Personal communication with the author of [1].

TABLE IV
AVERAGE TIME REQUIRED TO EVALUATE A 240 × 320IMAGE (12 800 WINDOWS PER IMAGE) FOR DIFFERENT DETECTORS

	windows per sec
HOG, Quadratic SVM [11]	60
COV, Riemannian Manifold [1]	3,000
HOG, Linear SVM [11]	5,500
Our COV approach (proposed)	12,000

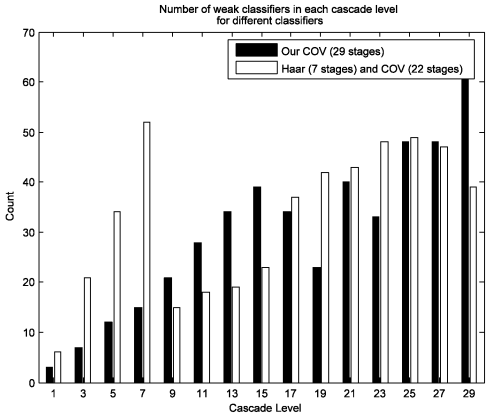


Fig. 12. Number of weak classifiers in different cascade levels on the INRIA dataset [11].

TABLE V
AVERAGE EVALUATION TIME IN WINDOWS PER SECOND FOR DIFFERENT PARAMETERS OF THE TWO-LAYER BOOSTING APPROACHES

	windows per sec
Our COV (29 stages)	12,000
Haar (7 stages) and COV (22 stages)	35,000
Haar (9 stages) and COV (20 stages)	40,000
Haar (15 stages) and COV (12 stages)	52,000
Haar (27 stages)	200,000

boosted covariance features could help improve the detection speed while maintaining a high detection rate.

E. Experiments on the INRIA Human Dataset With Two-Layer Boosting

1) Experiment Setup: Similar to the experiments on the dataset of [14], we subsample the overcomplete set of Haar features to 54 779 filters: 11 446 vertical two-rectangle features, 14 094 horizontal two-rectangle features, 8088 vertical three-rectangle features, 10 400 horizontal three-rectangle features, and 10 751 four-rectangle features. Unlike the previous experiment, the performance of unsigned wavelets seems to outperform the performance of signed wavelets. We think that, when the human resolution is large, clothing and background details can be easily observed and intensity gradient sign becomes irrelevant. In other words, the wide range of clothing and background colors make the gradient sign uninformative, e.g., a person with a black shirt in front of a white background should have the same information as a person with a white shirt in front of a black background. Hence, we used the absolute values of the wavelet responses in this experiment. For covariance features, we use a set of rectangular covariance features generated from previous section. Fig. 12 gives some details about our two-layer boosting cascade.

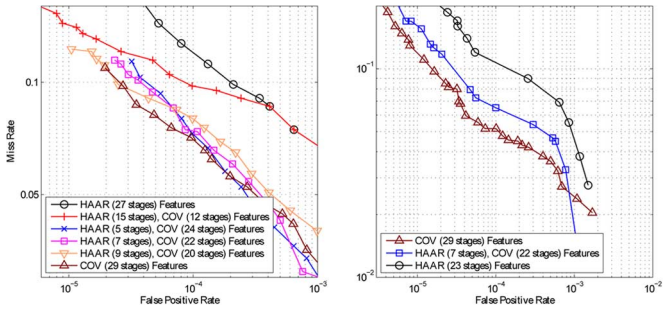


Fig. 13. Performance comparison between different configurations of the two-layer boosting approach based on classification (left) and detection (right) on INRIA dataset. Overlapping amongst the ROC curves of different configurations of two-layer boosting techniques indicates the performance similarity.

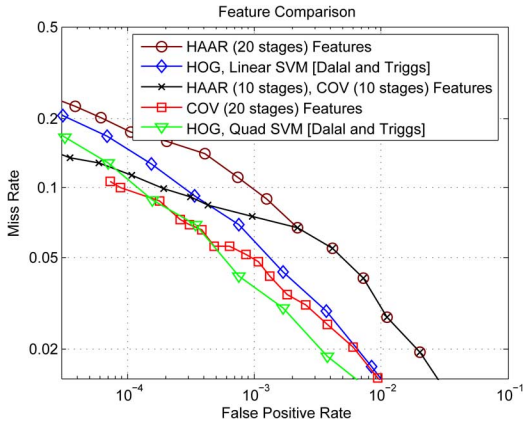


Fig. 14. Performance comparison between the two-layer boosting approach (Haar features plus covariance features) and HOG features on INRIA dataset with resolution of 18 × 36.

2) Results Based on Multilayer Boosting: The evaluation time in windows per second for different hybrid configurations is shown in Table V. Similar to previous results, adding Haar wavelet features as a preprocessing step increases the detection speed significantly. Compared with the original covariance detector in [1], the two-layer boosting approach is ten times faster.

Fig. 13 shows the performance of two-layer boosting approach using the *classification* and *detection* approaches. For the classification approach, the overall performance of different hybrid configurations is very similar to the performance of a cascade of boosted covariance features. A hybrid classifier with 15 levels of Haar features and 12 levels of covariance features might seem to perform poorly at high false positive rate. However, at a low false positive rate, i.e., 2×10^{-5}, its performance is very similar to performance of a cascade of boosted covariance features. For the detection approach, the two-layer boosting approach performs slightly inferior to the cascade of boosted covariance features. This is not surprising since INRIA human datasets contain human with various poses which Haar features are less capable to capture. Nonetheless, applying boosted covariance features in the second cascade greatly improves the overall accuracy of a boosted cascade of Haar features.

We have also compared the two-layer boosting approach and HOG features on the INRIA dataset [11] with a resolution of 18 × 36. Note that the experiment setup used in this experiment

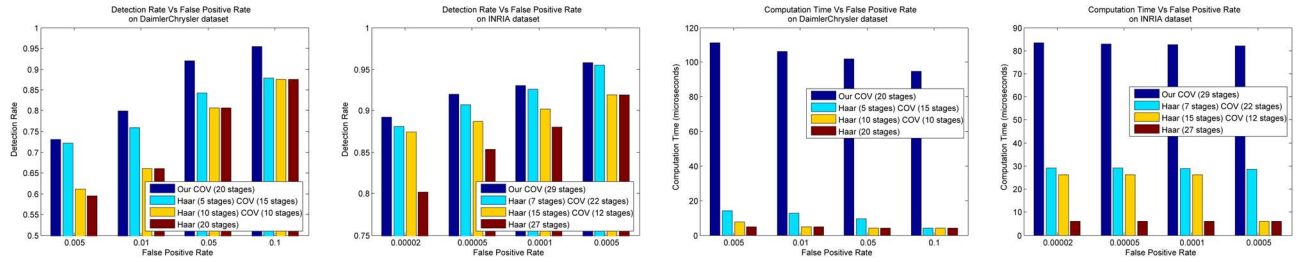


Fig. 15. Detection rate and speed tradeoff for different configurations of two-layer boosting. The first two figures show detection rate versus false positive rate on the dataset [14] and the INRIA dataset [11]. The last two figures show computation time versus false positive rate on the dataset [14] and the INRIA dataset [11]. Clearly, covariance features have the highest detection rate across all false positive rates while Haar features have the lowest detection rate. On the other hand, Haar features are the fastest to compute while covariance features are the slowest.

is similar to the one used in previous experiment (Sections VI-B and C). Fig. 14 shows the experimental results of different approaches. The results look slightly different from experimental results in Section VI-C due to the different datasets used. However, the overall results seem to be consistent with results shown in Figs. 8 and 10.

F. Detection Performance and Speed Tradeoff for the Two-Layer Boosting

From the previous experiments, the results show that the speed of Haar features classifier is much faster than the speed of the covariance features classifier. Therefore, it is best to place as many stages of Haar features in the first layer of the classifier. However, having too many stages of Haar features will degrade the overall performance. In this section, we try to find the best combination that will give the best overall results.

To study the tradeoff between the detection performance and speed of our classifiers, we perform a test on different false positive rates. For example, to achieve a 5×10^{-4} false positive rate for a boosted covariance classifier on INRIA dataset, we only use the first 19 stages of covariance features (instead of the full 29 stages). We then calculate the average computation time by evaluating the 19 stages classifier on a test sequence of images. Fig. 15 shows the detection rate and computation time for different configurations of multiple-layer boosting on the dataset of [14] and INRIA dataset [11]. From the figure, it can be concluded that there is a tradeoff between the detection performance and speed. In order to achieve a high detection rate, only a small number of Haar stages should be placed in the first layer of the classifier. For a small-resolution dataset (18×36 pixels), a configuration of Harr (5 stages) covariance (15 stages) seems to perform best at a reasonable computation time. For a larger resolution dataset (64×128 pixels), a configuration of Haar (7 stages) covariance (22 stages) seems to perform best.

VII. Conclusion

This paper has presented a fast and robust pedestrian detection technique. We use weighted Fisher linear discriminant analysis as the weak classifier for AdaBoost training. In order to speed up the computation time, a cascaded classifier architecture is adopted [2].

From the experimental results on datasets used in [14], our system has shown to give high detection performance at a low false positive rate. Comparing with techniques using linear SVM classifier, the proposed system outperforms all the systems evaluated. When compared with nonlinear SVM systems, the system is shown to perform very similar to the covariance features with Gaussian SVM and slightly inferior compared to HOG with quadratic SVM. However, the computation time of HOG with quadratic SVM is much higher than our proposed technique.

The performance of the proposed approach is also evaluated on the INRIA pedestrian dataset [11]. On this dataset, previous methods reported have significantly higher miss rates at almost all the false positive rates per window. Our algorithm's performance is comparable to the state-of-the-art [1] while is almost four times faster for detection due to its new design.

To further accelerate the detection, we have also introduced a faster strategy—two-layer boosting with heterogeneous features—to exploit the efficiency of the Haar feature and the discriminative power of the covariance feature. This way our detector runs ten times faster than the original covariance feature detector [1].

Ongoing work includes the search of new features for human detection. How to optimally design a cascaded classifier may also be a future topic.

References

[1] O. Tuzel, F. Porikli, and P. Meer, "Human detection via classification on Riemannian manifolds," in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit.*, Minneapolis, MN, 2007, pp. 1–8.

[2] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[3] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 15–33, 2000.

[4] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 734–741.

[5] 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, 2007.

[6] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit.*, San Diego, CA, 2005, vol. 1, pp. 878–885.

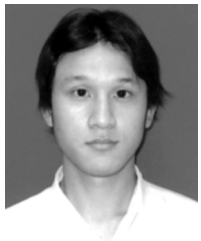
[7] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Proc. IEEE Int. Conf. Comput. Vis.*, Beijing, China, 2005, vol. 1, pp. 90–97.

[8] V. Sharma and J. Davis, "Integrating appearance and motion cues for simultaneous detection and segmentation of pedestrians," in *Proc. IEEE Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[9] Y. Amit, D. Geman, and X. Fan, "A coarse-to-fine strategy for multiclass shape detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 12, pp. 1606–1621, Dec. 2004.

[10] G. Mori, X. Ren, A. Efros, and J. Malik, "Recovering human body configurations: combining segmentation and recognition," in *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, Washington, DC, 2004, vol. 2, pp. 326–333.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, San Diego, CA, 2005, vol. 1, pp. 886–893.
[12] C. Wöhler and J. Anlauf, "An adaptable time-delay neural-network algorithm for image sequence analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 6, pp. 1531–1536, Dec. 1999.
[13] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Proc. Eur. Conf. Comput. Vis.*, Prague, Czech Republic, May 2004, vol. 1, pp. 69–81.
[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] J. Shawe-Taylor and N. Cristianini, *Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[16] K. Levi and Y. Weiss, "Learning object detection from a small number of examples: The importance of good features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Washington, DC, 2004, vol. 2, pp. 53–60.
[17] J. Meynet, V. Popovici, and J.-P. Thiran, "Face detection with boosted Gaussian features," *Pattern Recognit.*, vol. 40, no. 8, pp. 2283–2291, 2007.
[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[19] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Proc. Eur. Conf. Comput. Vis.*, Graz, Austria, May 2006, vol. 2, pp. 589–600.
[20] S. Jin, D. S. Yeung, and X. Wang, "Network instrusion detection in covariance feature space," *Pattern Recognit.*, vol. 40, pp. 2185–2197, 2007.
[21] R. E. Schapire, "Theoretical views of boosting and applications," in *Proc. Int. Conf. Algorithmic Learn. Theory*, London, U.K., 1999, pp. 13–25.
[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] M. T. Pham and T. J. Cham, "Fast training and selection of haar features using statistics in boosting-based face detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, 2007, pp. 1–7.
[24] J. Wu, M. D. Mullin, and J. M. Rehg, "Linear asymmetric classifier for cascade detectors," in *Proc. Int. Conf. Mach. Learn.*, Bonn, Germany, 2005, pp. 988–995.
[25] Q. Zhu, S. Avidan, M. Yeh, and K.-T. Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, New York, 2006, vol. 2, pp. 1491–1498.
[26] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.
[27] The PASCAL Visual Object Classes Challenge VOC (2007). [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2007/index.html



Sakrapee Paisitkriangkrai received the B.E. degree in computer engineering and the M.E. degree in biomedical engineering from the University of New South Wales, Sydney, Australia, where he is currently working toward the Ph.D. degree.

His research interests include pattern recognition, image processing, and machine learning.

Chunhua Shen received the Ph.D. degree from the University of Adelaide, Australia, in 2005.

He is currently a Researcher with the Computer Vision Program, NICTA, Canberra, Australia. He is also an Adjunct Research Fellow with the Australian National University and an Adjunct Lecturer with the University of Adelaide. His main research interests include statistical pattern analysis and its application in computer vision.

Jian Zhang (M'98–SM'04) received the Ph.D. degree in electrical engineering from the University College, University of New South Wales, Australian Defence Force Academy, Australia, in 1997.

He is a Principal Researcher with NICTA, Sydney, Australia. He is also a Conjoint Associate Professor with University of New South Wales, Sydney, Australia. He is currently an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and the *EURASIP Journal on Image and Video Processing*.