

# Real-time Pedestrian Detection Using a Boosted Multi-layer Classifier\*

Sakrapee Paisitkriangkrai<sup>1,2</sup>, Chunhua Shen<sup>1,3</sup>, Jian Zhang<sup>1,2</sup>

<sup>1</sup>NICTA    <sup>2</sup>University of New South Wales    <sup>3</sup>Australian National University

## Abstract

*Techniques for detecting pedestrian in still images have attracted considerable research interests due to its wide applications such as video surveillance and intelligent transportation systems. In this paper, we propose a novel simpler pedestrian detector using state-of-the-art locally extracted features, namely, covariance features. Covariance features were originally proposed in [1, 2]. Unlike the work in [2], 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 adopted. Multiple layer boosting with heterogeneous features is constructed to exploit the efficiency of the Haar-like feature and the discriminative power of the covariance feature simultaneously. Extensive experiments show that by combining the Haar-like and covariance features, we speed up the original covariance feature detector [2] by up to an order of magnitude in processing time without compromising the detection performance. For the first time, the proposed work enables covariance feature based pedestrian detection to work real-time.*

## 1 Introduction

Although much effort has been spent recently, the problem of automatic detection of objects is far to be solved (e.g., [2–10]). Pedestrian detection in still images is one of the most difficult examples due to a wide range of poses that human can adopt, large variations in clothing, as well as cluttered backgrounds and environmental conditions. All these issues have made this problem very challenging from a computer vision point of view. Classification based methods have comprised the mainstream of research and have been shown to achieve successful results in object detection. 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 detection, the trained classifier scans 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 [11], human [4, 5, 8, 10], car number plate [12], etc.

In this work, we propose a novel pedestrian detection technique using the covariance features. The main contribution of this work is two-fold. The first contribution is that we show how multi-dimensional covariance features can be integrated with weighted linear discriminant analysis before being trained on 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 [13].

The proposed boosted covariance detector achieves about four times faster detection speed than the method in [2], but it is still not fast enough for real-time applications. On one hand, the Haar-like feature can be computed rapidly due to its simplicity [11] but it is less powerful for classification [14]. On the other hand, although the covariance feature is a better candidate for representing pedestrians, it requires heavier computation than the Haar-like feature. Here, to further accelerate our proposed detector, we proffer a novel strategy—two-layer boosting with heterogeneous features—to exploit the efficiency of the Haar-like feature and the discriminative power of the covariance feature in a single framework. 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 which do not contain a target. Thanks to the flexibility of the cascaded classifier, we employ the Haar-like feature based classifiers at the beginning of the cascade; and use the covariance feature at latter stages. Experiments show that by combining the Haar-like and covariance features, we speed up the conventional covariance

\*NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Center of Excellence program.

feature detector [2] by an order of magnitude in detection time without compromising the detection performance. On a  $360 \times 288$  pixels image, our system can process at around 4 frames per second. To our knowledge, this is the first real-time covariance feature based pedestrian detector. Preliminary version of this work has been published in [15]. The results and analysis presented in this paper is an extended version of the results presented in [15].

The paper is organized as follows. Section 2 reviews various existing techniques for pedestrian detection. Section 3 gives a detailed description of our method. The experimental setup and experimental results are presented in Section 4. The paper concludes in Section 5.

## 2 Related work

The literature on pedestrian detection is abundant. Two of the well known image features often being used are motion and shape. Motion approaches, which require preprocessing techniques like background subtraction or image segmentation (*e.g.* [16]), segment an image into so-called super pixels and then detect the human body and estimate its pose. Approaches based on shape information typically detect pedestrian directly without using preprocessing techniques [2, 4, 6, 7, 17]. Background subtraction and image segmentation techniques can be applied to segment foreground objects from the background. The foreground objects can then be classified into different categories, *e.g.*, human, vehicle and animal, based on their shape, color, texture, *etc.* One of the main drawbacks of these techniques is that they usually assume that the camera is static, background is fixed and the differences are caused only by foreground objects. In addition, the performance of the system is often affected by outdoor light changes.

The second approach is to detect human based on shape features extracted from still images. Features can be distinguished into global features and local features depending on how the features are measured. 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 [11], gradient orientation [4], region covariance [2], edgelet [5], *etc.*

## 3 Proposed method

Classification accuracy of boosting techniques depends greatly on the choice of weak classifiers. Although effective weak classifiers increase the performance of the final strong

classifiers, 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-like features in [8, 11], 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 [18], it is possible to use linear SVMs as weak learners. Here we adopt a more efficient approach. We project the multi-dimensional features onto a 1D line using weighted Fisher linear discriminant analysis (WLDA). WLDA finds a linear projection function which guarantees optimal classification of normally distributed samples of two classes.

Note that this treatment is different from [1, 2], where the covariance matrix is directly used as the feature and the distance between features is calculated in the Riemannian manifold<sup>1</sup>. 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 correlation coefficient and measure the distance in the Euclidean space, which is faster.

We conducted an experiment similar to the one described in [1] between the linear version (Euclidean space) and manifold version (Riemannian space) of covariance features. The experiment compares the two different distance measures:- distance based on the correlation coefficient from two covariance matrices in the Euclidean space and distance of two covariance matrices in the Riemannian manifold. Figure 1 shows some of the experimental results. From the figure, we can roughly conclude that their performance on pedestrian detection is quite similar.

This section begins with a short explanation of Fisher linear discriminant analysis (LDA) concept. We then extend these methods to varying weighted training samples. Next, we describe in details how to apply these techniques to train multi-dimensional covariance features on a cascade of AdaBoost classifiers framework. Finally, we introduce a new two-layer pedestrian detector which utilizes the efficiency of Haar-like features and the discriminative power of the covariance features.

### 3.1 Weighted Fisher discriminant analysis

The linear discriminant analysis is a representative of the supervised learning method which yields the linear projection. The goal of the Fisher's criteria is to find a linear combination of the variables that leads to the best separation between the two projected sets. The criterion proposed by Fisher assumes uniformly weighted training samples. In AdaBoost training, each data point is associated

<sup>1</sup>Covariance matrices are symmetric and positive semi-definite, hence they reside in the Riemannian manifold.

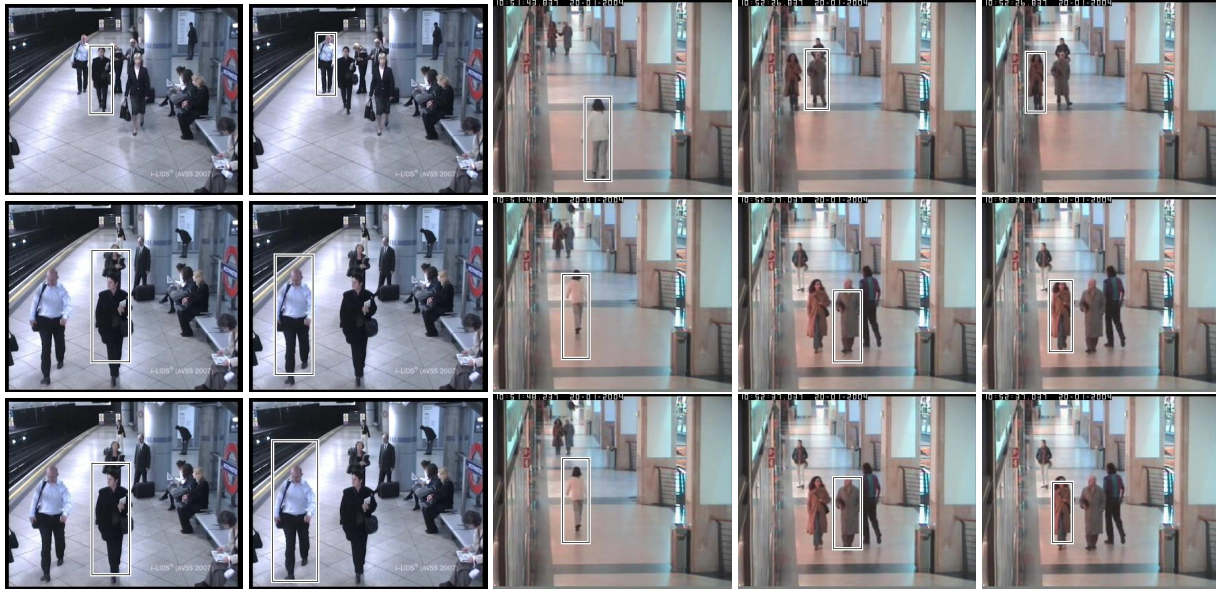


Figure 1: Detection examples on AVSS 2007 and CAVIAR dataset. *Top*: Input region. *Middle*: Best matching region found using covariance features based on distance in the Riemannian manifold [1]. *Bottom*: Best matching region found using covariance features based on distance in the Euclidean space.

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) [19]. Similar to LDA [20], WLDA finds a linear combination of the variables that can separate the two classes as much as possible with emphasis on the training samples with high weights.

### 3.2 A cascade of covariance descriptors

The covariance descriptors encode the relationship information between different image statistics inside the region. Combining with WLDA, the descriptors can be used to represent various parts of the human body. The experimental results show that the covariance regions selected by AdaBoost are very meaningful and can be easily interpreted as shown in Figure 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.

In order to reduce the computation time, a cascade of classifiers is built [11]. The key insight is that efficient boosted classifiers, which can reject many of the simple non-pedestrian 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 non-pedestrian 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, road, *etc.* while spending more time on pedestrian-like regions. Only samples that can pass through all stages of the cascades are classified as pedestrians.



Figure 2: The first and second covariance region selected by AdaBoost. The first two covariance regions overlaid 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.

**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;
- $F_{\text{target}}$ : 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  $\theta$ , 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  $i = 1, \dots$ ;
- Final training accuracy:  $F_i$  and  $D_i$ .

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

### 3.3 Two-layer boosting with heterogeneous features

A two-layer cascade of classifiers is adopted here in order to speed up the proposed detector. The goals of designing the two-layer approach are mainly the 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 non-pedestrian patterns in the early stage of the cascade.

Haar-like wavelet features have proved to be extremely fast and highly powerful in the application of face detections [11]. However, the Haar-like feature performs poorly in the context of human detection as reported in [8]. In order to improve the overall accuracy, we apply boosted covariance features in the second layer. This way we utilize the efficiency of the Haar-like feature and the discriminative power of the covariance feature in a single framework.

Due to the flexibility of the cascaded structure, it is easy to integrate multiple heterogeneous features. Although we use Haar-like 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.

## 4 Experiments

The experimental section is organized as follows. First, the datasets used in this experiment are described. Parameters used to achieve optimal results are then discussed. Finally, experimental results of different techniques are compared and analyzed.

### 4.1 Experiments on DaimlerChrysler dataset [13] with boosted covariance features

The dataset [13] consists of three training sets and two test sets. Each training set contains 4,800 pedestrian examples and 5,000 non-pedestrian examples. All samples are scaled to size  $18 \times 36$  pixels. For boosted cascade of covariance features, 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 1,120 covariance filters.

We also train covariance features with various combination of SVM using SVMLight [21]. For this method, we concatenate the covariance descriptors for all regions into a combined feature vector. SVM classifier is trained using this feature vector. Our preliminary experiments show region of size  $7 \times 7$  pixels, shifted at a step size of 2 pixels over the entire input image of size  $18 \times 36$  to be optimal for our benchmark datasets (total feature length of 2,520). For the HOG features, we have decided to use a cell size of  $3 \times 3$  pixels with a block size of  $2 \times 2$  cells, descriptor stride of 2 pixels and 18 orientation bins of signed gradients (total feature length is 8,064) to train SVM classifiers.

A comparison of the best performing results for different feature types are shown in Figure 4(a). The performance of our proposed method is very similar to the best performance of HOG features and covariance features. From the figure, we can see that gradient information is very helpful in human classification problems. In all experiments, nonlinear SVMs improve performance significantly over the linear one. However, this comes at the cost of a much higher computation time.

It might not be fair to perform a direct comparison between the three detectors since the boosted cascade is trained with more non-pedestrian samples, *i.e.*, by making use of cascade structure, we have manually increased the non-pedestrian training sample size from a set of non-pedestrian images. In order to compare the performance of three detectors, we apply bootstrapping technique to HOG [4] and covariance features. Bootstrapping is applied iteratively, generating 10,000 new non-pedestrian samples at each iteration. The result is shown in figure 4(b). We observed that collecting the first 10,000 new non-pedestrian

samples did not take long but the second iteration took a long time. This is exactly what we expected since the new classifier had better accuracy than the previous classifier. From this figure, 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 during training phase as training samples are now much more complex.



Figure 3: Examples of mistakes made by our boosted covariance detector on the dataset [13]. The first row shows false negative examples and the last row shows false positive examples.

Figure 3 presents a qualitative assessment of the errors made by our detectors, showing some false negative (non-pedestrian-like pedestrians) and false positive (pedestrian-like non-pedestrian) examples from our detectors point of view. The results reveal 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. It is interesting to see that many false positives are road signs which have shoulder-arm and head shaped contours.

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

Table 1: Average time required to evaluate 10 frames of a sequence of  $384 \times 288$  pixels images. Each image consists of 17,280 windows (scale factor of 0.8 and step-size of 4 pixels).

Next, 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  $384 \times 288$  pixels in width and height. Table 1 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 non-linear SVM classifier.

In the next experiment, we show how adding a cascade

of Haar-like wavelet features as a preprocessing to a cascade of boosted covariance features could help improve the detection speed while maintaining a high detection rate.

## 4.2 Experiments on DaimlerChrysler dataset [13] with two-layer boosting

We generate a set of overcomplete Haar-like wavelet filters and subsample the overcomplete set. The set of Haar-like features that we use to train the cascade contained 20,547 filters: 5,540 vertical two-rectangle features, 5,395 horizontal two-rectangle features, 3,592 vertical three-rectangle features, 3,396 horizontal three-rectangle features and 2,624 four-rectangle features. From the preliminary experiments on signed and unsigned wavelets, the authors observed the performance of signed wavelets to outperform the performance of 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.

Table 2 shows the evaluation time in windows per second for different hybrid configurations. Adding more stages of Haar-like wavelet features as a preprocessing step increases the detection speed approximately *exponentially*. In terms of performance, the new technique performs very similar to the boosted covariance features experimented earlier at low false positive rate (figure 4(c)). At high false positive rate, the system might seem to perform poorly due to the use of Haar-like features. Nevertheless, most real-world applications often focus on low false detections.

	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

Table 2: Average evaluation time in windows per second for different parameters of the two-layer boosting approaches.

## 4.3 Experiments on INRIA dataset [4] with boosted covariance features

INRIA dataset [4] consists of one training set and one test set. The training set consists of 2,416 mirrored human samples and 1,200 non-human images. The human samples are mostly in standing position. A border of 16 pixels is added to the sample in order to preserve contour information. All samples are scaled to size  $64 \times 128$  pixels. The test set contains 1,176 human samples (mirrored) extracted from 288 images.

Similar to the previous experiments, we generate a set of overcomplete rectangular covariance filters and subsam-

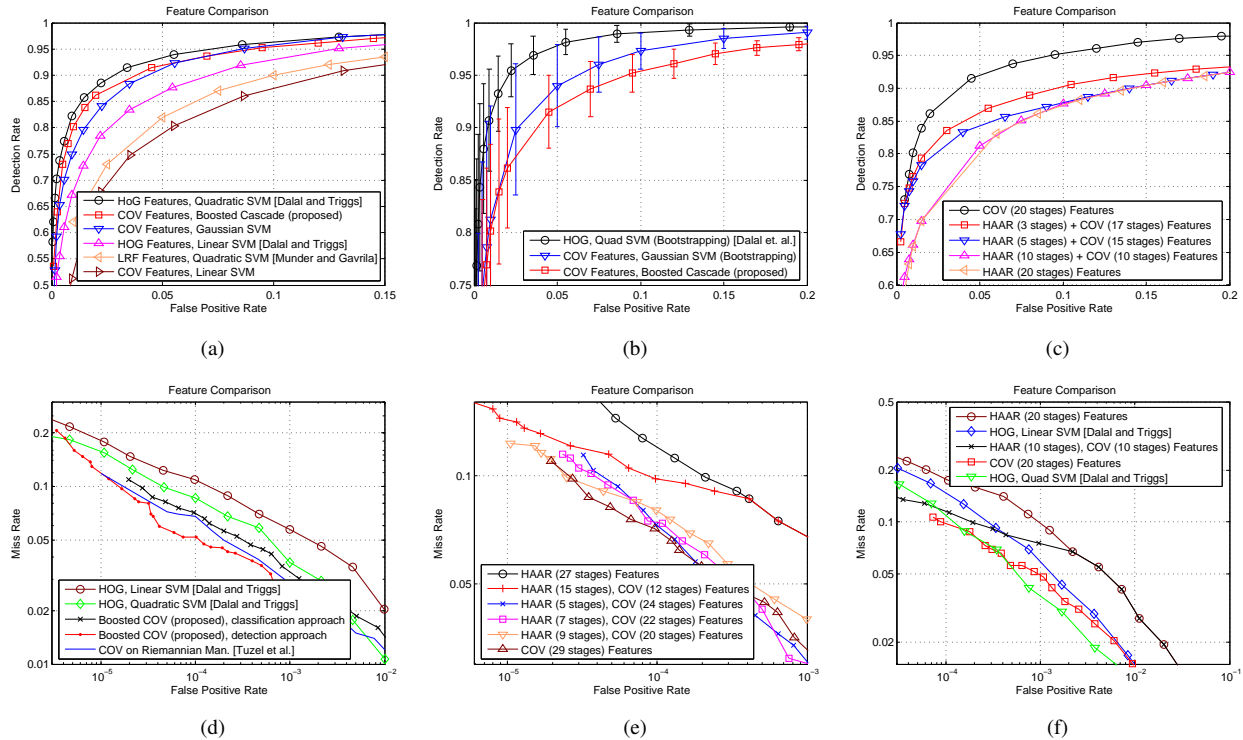


Figure 4: See text for details. Performance comparison of our boosted covariance features with (a) other feature types on DaimlerChrysler dataset [13]. (b) HOG and covariance features trained using SVM with bootstrapping technique. (c) our hybrid approach on DaimlerChrysler dataset [13]. (d) HOG with linear SVM [4] and covariance features on Riemannian manifold [2]. The curve of covariance on Riemannian manifold is reproduced from [2]. (e) our hybrid approach on INRIA dataset [4]. (f) HOG features on INRIA dataset with resolution of  $18 \times 36$  pixels.

ple 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 (detection rate of 99.5% and false positive rate of 50%). Each stage is trained with 2, 416 human samples and 5, 000 non-human samples. The final cascade consists of 29 stages. 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 non-human. 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 [2], mean shift clustering [22] is used to cluster multiple overlapping detection windows. Simple rules as in [11] are also applied on the clustering results to merge those close detection windows. The experiments are conducted using a current standard desktop with 2.80 GHz Intel Pentium-D CPU and 2 GB memory.

Figure 4(d) 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 [4] while achieving a comparable detection rate to the technique described in [2]. [2] calculates distance between covariance matrix on the Riemannian manifold. An eigen-decomposition is required which slows down the computation speed [2]. 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 neighbour threshold (a number of merged detections). As for the processing time, on average our unoptimized implementation in C++ can search around 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 3 shows the average detection speed for three different classifiers. Compared to [4] and [2], our search time is faster than both techniques (2.2 times faster than [4] and 4 times faster than [2]).

	windows per sec
COV, Riemannian Manifold [2]	3,000
HOG, Linear SVM [4]	5,500
Our COV approach (proposed)	12,000

Table 3: Average computation time in windows per second for different detectors.

#### 4.4 Experiments on INRIA dataset [4] with the two-layer boosting

Similar to the experiments on the DaimlerChrysler dataset [13], we subsample the overcomplete set of Haar wavelet features to 54,779 filters: 11,446 vertical two-rectangle features, 14,094 horizontal two-rectangle features, 8,088 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 slightly outperforms the performance of signed wavelets. The authors 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.

Table 4 shows the average evaluation time in windows per second for different hybrid configurations. Similar to previous results, adding Haar-like wavelet features as a pre-processing step increases the detection speed significantly. Compared with the original covariance detector in [2], the two-layer boosting approach is 10 – 15 times faster. Figure 4(e) shows the performance of two-layer boosting approach. The overall performance of different hybrid configurations is very similar to the performance of a cascade of boosted covariance features. Figure 5 demonstrates some detection examples using our hybrid detector on INRIA test dataset and Advanced Video and Signal based Surveillance (AVSS) 2007 dataset<sup>2</sup>. The AVSS detection results are provided as supplementary materials.

In the next experiment, we downsample the INRIA dataset to size  $18 \times 36$  pixels and compare the two-layer boosting approach and HOG features. The experiment setup used in this experiment is similar to the one used in previous experiment. We test the classifier on the INRIA test set and the experimental results are shown in figure 4(f). These results seem to be consistent with results reported in the previous experiment (section 4.1). HOG features with non-linear SVM performs slightly better than our boosted covariance features. However, this contradicts the results

<sup>2</sup>[http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007\\_d.html](http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007_d.html)

reported earlier (figure 4(d)) where boosted covariance features outperform HOG features. We suspect the difference is due to the resolution of the datasets and the classifiers used. Small resolution datasets give less number of meaningful covariance features than large resolution datasets. As a result, covariance descriptors are not powerful enough to capture large pedestrian variations.

	windows per sec
Our COV (29 stages)	12,000
Haar (5 stages) and COV (24 stages)	29,000
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

Table 4: Average evaluation time in windows per second for different parameters of the two-layer boosting approaches.

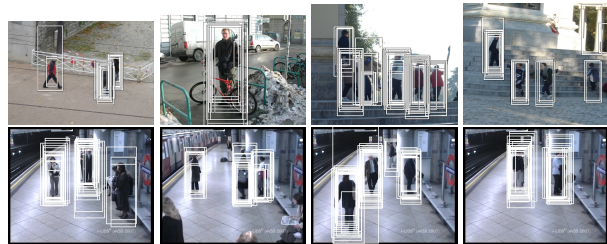


Figure 5: Detection examples. The boxes show the detection results of our hybrid classifier (9 levels of Haar-like features and 22 levels of covariance features). *Top*: INRIA dataset. *Bottom*: AVSS 2007 dataset. Note that no post-processing has been applied to the detection results.

#### 4.5 Limitations

In this section, we test our trained classifier (classifier trained on INRIA dataset) on random internet images with pedestrians having variable illumination, appearance, pose and occlusion. Some of the results are shown in figure 6. The top row shows the raw detection results. The bottom row shows the merged detection results using mean shift clustering. Based on our observations, the system works well on the images where there is a small gap between pedestrians *i.e.*, no occlusion between pedestrians. When humans stand in a group or occlude one another, the human contour is quite complex and different from what the classifier was trained with. In addition, there exist a lot of multiple overlapping detection windows when human occludes one another. Mean shift clustering fails to merge the detection windows correctly when there is a lot of overlapping windows. As a result, the system fails to detect most of the pedestrians. We also note that a lot of false detections came from the human body parts *e.g.*, human limbs. This



Figure 6: Detection examples on random internet images. *Top*: raw detection results. *Bottom*: merged detection results using mean shift clustering technique.

is not surprising since our negative training samples do not contain any of the human body parts.

## 5 Conclusion

This paper presents a new technique for pedestrian detection that combines covariance features with multi-layer boosting. The first contribution of our work lies in the integration of multi-dimensional covariance features with weighted linear discriminant analysis as the weak classifier for AdaBoost training. Weak classifiers are trained in the Euclidean space for faster computation. As our second contribution, we presented an architecture that uses two-layer boosting with heterogeneous features, namely a first layer with Haar-like features, and a second layer with covariance features, to exploit the efficiency of the Haar-like feature and the discriminative power of the covariance feature. This way our detector can perform up to 15 times faster than the original covariance detector. As a result, our system can process at around 4 frames per second on a  $360 \times 288$  pixels image.

## References

- [1] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Proc. Eur. Conf. Comp. Vis.*, Graz, Austria, May 2006, vol. 2, pp. 589–600.
- [2] O. Tuzel, F. Porikli, and P. Meer, "Human detection via classification on Riemannian manifolds," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, Minneapolis, MN, 2007.
- [3] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Proc. Eur. Conf. Comp. Vis.*, Prague, Czech Republic, May 2004, vol. 1, pp. 69–81.
- [4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, San Diego, CA, 2005, vol. 1, pp. 886–893.
- [5] 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. Comp. Vis.*, Beijing, China, 2005, vol. 1, pp. 90–97.
- [6] 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, 1999.
- [7] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comp. Vis.*, vol. 38, no. 1, pp. 15–33, 2000.
- [8] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2003.
- [9] Y. Wu and T. Yu, "A field model for human detection and tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 753–765, 2006.
- [10] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *Int. J. Comp. Vis.*, vol. 73, no. 1, pp. 41–59, 2007.
- [11] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comp. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
- [12] 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.
- [13] 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, 2006.
- [14] K. Levi and Y. Weiss, "Learning object detection from a small number of examples: The importance of good features," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, Washington, DC, 2004, vol. 2.
- [15] S. Paisitkriangkrai, C. Shen, and J. Zhang, "Fast pedestrian detection using a cascade of boosted covariance features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, 2008.
- [16] G. Mori, X. Ren, A. Efros, and J. Malik, "Recovering human body configurations: combining segmentation and recognition," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, Washington, DC, 2004, vol. 2, pp. 326–333.
- [17] P. Szabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2007.
- [18] 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. Comp. Vis. Patt. Recogn.*, New York, 2006, vol. 2, pp. 1491–1498.
- [19] I. Laptev, "Improvements of object detection using boosted histograms," in *Proc. BMVC.*, Edinburgh, UK, 2006, pp. 949–958.
- [20] R. Duda, P. Hart, and D. Stork, *Pattern Classification (2nd ed.)*, John Wiley and Sons, 2001.
- [21] T. Joachims, *Making large-Scale SVM Learning Practical*, Advances in Kernel Methods - Support Vector Learning. MIT Press, 1999.
- [22] 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, 2002.