An Experimental Evaluation of Local Features for Pedestrian Classification

Sakrapee Paisitkriangkrai[1,2], Chunhua Shen[1,3], Jian Zhang[1,2]
[1]NICTA [2]University of New South Wales [3]Australian National University
email: {*paul.pais*,*chunhua.shen*,*jian.zhang*}*@nicta.com.au*

Abstract

The ability to detect pedestrians is a first important step in many computer vision applications such as video surveillance. This paper presents an experimental study on pedestrian detection using state-of-the-art local feature extraction and support vector machine (SVM) classifiers. The performance of pedestrian detection using region covariance, histogram of oriented gradients (HOG) and local receptive fields (LRF) feature descriptors is experimentally evaluated. The experiments are performed on both the benchmarking dataset used in [1] and the MIT CBCL dataset. Both can be publicly accessed. The experimental results show that region covariance features with radial basis function (RBF) kernel SVM and HOG features with quadratic kernel SVM outperform the combination of LRF features with quadratic kernel SVM reported in [1].

1 Introduction

Detecting pedestrians has attracted a lot of research interests in recent years, due to its key role for several important applications in computer vision, *e.g*., smart vehicles, surveillance systems with intelligent query capabilities, intersection traffic analysis. In particular, there is growing effort in the development of intelligent video surveillance systems. Public spots like airports, train stations and parking area have a large number of security cameras recording at all times. Due to the vast amount of video data being processed, it is very difficult to detect and respond to an abnormal event in real-time. An example of such abnormal events is unusual human activity in a scene. An automated method for finding human in a scene serves as the first important pre-processing step in understanding human activity. The challenges are 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 machine vision perspective.

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 [3], *etc*.

The performance of several pedestrian detection approaches has been evaluated in [1]. Multiple feature-classifier combinations have been examined with respect to their receiver operating characteristic (ROC) performance and efficiency. Different features including principal component analysis coefficients (PCA), local receptive fields (LRF) feature [4], and Haar wavelets [5] are used to train neural networks, support vector machines (SVM) [6, 7] 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 [1]. This may be due to the large variability of pedestrian shapes. Global features like principal component analysis are more powerful modeling objects with stable structures such as frontal faces, rigid car images taken from a fixed view angle.

Although [1] provides some insights on pedestrian detection, it has not compared state-of-the-art techniques in this topic. Very recently, histogram of oriented gradients (HOG) [8] and region covariance features [9] are proffered for pedestrian detection. It has been shown that they outperform those previous approaches. HOG is a grey-level image feature formed by a set of normalized gradient histograms; while region covariance is an appearance based feature, which combines pixel coordinates, intensity, gradients *etc*. 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 our knowledge, these approaches have not been compared yet. It remains unclear whether silhouette or appearance based features are better for pedestrian detection. This paper tries to answer this question. The main purpose of the paper therefore is a systematic comparison of some novel techniques for pedestrian detection.

In this paper, we perform an experimental study on the

0-7695-3067-2/07 $25.00 © 2007 IEEE
DOI 10.1109/DICTA.2007.34

state-of-the-art pedestrian detection techniques: LRF, HOG and region covariance; along with various combination with SVM. The reasons we select these three features along with SVM classifiers are mainly:

- These three local features seem to be the best candidates for this task;
- SVM is one of the advanced classifiers. 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. For the same reason, we do not apply Adaboost to select the most discriminant features.

The paper is organized as follows. Section 2 reviews various existing techniques for pedestrian detection. Sections 3 and 4 describe methods used for feature extraction and a brief introduction to SVM. The experimental setup and experimental results are presented in Section 5. The paper concludes in Section 6.

2 Related work

Many pedestrian classification approaches have been proposed in the literature. These algorithms can be roughly classified into two main categories: (1) approaches which require pre-processing 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); and (2) approaches which detects pedestrian directly without using pre-processing techniques [8, 5, 9, 4].

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 like human, vehicle and animal, based on their shape, color, texture, *etc*. One of the main drawbacks of these techniques are 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 features extracted from the image. Features can be distinguished into global features, local features and key-points 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 image. One of the well known global feature extraction method is 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 image. Some examples of the commonly used local features are wavelet coefficient [2], gradient orientation [8], region covariance [9], *etc*. Local features approaches can be further divided into whole body detection and body parts detection [11, 12]. In 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 [1], the classification performances reported in different literature are quite different. This is 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.

3 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. In this paper, we evaluate three 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. Region covariance computes covariance from several low-level image features such as image intensities and gradients.

3.1 Local receptive fields

Multilayer perceptrons provide an adaptive approach for feature extraction by means of their hidden layer [4]. 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.

In [1], the authors further investigate the concept of LRF. In their experiments, they have shown that receptive fields of size 5×5, shifted at a step size of two pixels over the input image of size 18×36 are optimal. In order to further improve the performance of LRF, the authors combine SVM with the output of the hidden layer of a neural network/LRF.

3.2 Histograms of oriented gradients

Since scale invariant feature transformation (SIFT) [13], which uses normalized local spatial histograms as a descrip-

tor, many research groups have been studying the use of orientation histograms in other areas. [8] is one of the successful examples. [8] 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

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

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

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

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

A local 1D 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 [8], the orientation histogram of each cell has 9 bins covering the orientation range of $[0, 180]$ degrees (unsigned gradients). Hence, each block is represented by a 36D feature vector (9 bins/cell $\times 4$ cells/block).

Each of the HOG descriptor blocks is then normalized based on the energy of the histogram contained within it. Normalization introduces better invariance to illumination, shadowing and edge contrast. In order to reduce the effect of non-linear illumination changes due to camera saturation or environmental illumination changes that affect 3D surfaces, ℓ_2-norm is applied followed by clipping (limiting the maximum values of the gradient magnitudes to 0.2) and renormalizing. The value of 0.2 is determined experimentally using images containing different illuminations for the same 3D objects [13]. The final step is to combine these normalized block descriptors to form a feature vector. The feature vector can then be used to train (SVM) classifiers.

3.3 Region covariance

Tuzel, *et al*. [9, 14] 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. Similar to HOG, the image is divided into small overlapped regions. For each region, the correlation coefficient is calculated. The correlation coefficient of two random variables X and Y is given by

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

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

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

A positive correlation coefficient, $\rho_{X,Y} > 0$, suggests that when X is high relative to its expected value, Y also tends to be high and *vice versa*. A negative correlation coefficient, $\theta_{X,Y} < 0$, suggests that a high value of X is likely to be accompanied by a low value of Y and *vice versa*. A linear relationship between X and Y produces the extreme values, $\theta_{X,Y} = \{+1, -1\}$. In other words, correlation coefficient is bounded by -1 and 1.

Image statistics used in this experiment are similar to the one used in [9]. The 8D 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 $\tan^{-1}\left(\frac{|\mathbf{I}_y|}{|\mathbf{I}_x|}\right)$, second order partial derivative of the intensity in horizontal direction $|\mathbf{I}_{xx}|$, second order partial derivative of the intensity in vertical direction $|\mathbf{I}_{yy}|$. The covariance descriptor of a region is an 8×8 matrix. Due to the symmetry, only 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 [14, 9], 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. We instead vectorize the symmetric matrix and measure the distance in the Euclidean space, which is faster.

In order to improve the covariance matrices' calculation time, technique which employs integral image [2] can be applied [14]. By expanding the mean from previous equation, covariance equation can be written as

$$\mathrm{cov}(X, Y) = \frac{1}{n-1}\left[\sum_k X_k Y_k - \frac{1}{n}\sum_k X_k \sum_k Y_k\right]. \quad (7)$$

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

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

The final step is to concatenate these covariance descriptors from all regions into a combined feature vector which can then be used to train SVM classifiers.

4 Support vector machines

There exist several classification techniques which can be applied to object detection problem. Some of the commonly used classification techniques are support vector machine and Adaboost [2].

Due to space constraints we limit our explanation of SVM classifiers algorithm to an overview. Large margin classifiers have demonstrated their advantages in many vision tasks. SVM is one of the popular large margin classifiers [6, 7] which has a very promising generalization capacity. The linear SVM is the best understood and simplest to apply. However, linear separability is a rather strict condition. Kernels are combined into margins for relaxing this restriction. SVM is extended to deal with linearly non-separable problems by mapping the training data from the input space into a high-dimensional, possibly infinite-dimensional, feature space. Using the kernel trick, the mapping function is not necessarily known explicitly. Like other kernel methods, SVM constructs a symmetric and positive definite kernel matrix (Gram matrix) which represents the similarities between all training datum points. Given N training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, the kernel matrix is written: $K_{ij} \equiv K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle, i, j = 1 \cdots N$. When K_{ij} is large, the labels of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, y_i and y_j, are expected to be the same. Here, $y_i, y_j \in \{+1, -1\}$. The decision rule is given by $\mathbf{sign}\,(f(\boldsymbol{x}))$ with

$$f(\boldsymbol{x}) = \sum_{i=1}^{N_S} \hat{\beta}_i K(\hat{\mathbf{x}}_i, \boldsymbol{x}) + b \tag{8}$$

where $\hat{\mathbf{x}}_i, i = 1 \cdots N_S$, are support vectors, N_S is the number of support vectors, $\hat{\beta}_i$ is the weight associated with $\hat{\mathbf{x}}_i$, and b is the bias. The training process of SVM then determines the parameters $\{\hat{\mathbf{x}}_i, \hat{\beta}_i, b, N_S\}$ by solving the optimization problem

$$\begin{aligned} \underset{\boldsymbol{\xi}, \boldsymbol{w}, b}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_r^r + C\sum\nolimits_{i=1}^{N} \xi_i, \\ \text{subject to} \quad & y_i(\boldsymbol{w}^\top \Phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \quad \forall i, \\ & \xi_i \geq 0, \quad \forall i, \end{aligned} \tag{9}$$

where $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^N$ is the slack variable set and the regularization parameter C determines the trade-off between SVM's generalization capability and training error. $r = 1, 2$ corresponds to 1-norm and 2-norm SVM respectively. The solution takes the form $\boldsymbol{w} = \sum_{i=1}^N y_i \alpha_i \Phi(\boldsymbol{x}_i)$. Here, $\alpha_i \geq 0$ and most of them are 0, yielding sparseness. The optimization (9) can be efficiently solved by linear or quadratic programming in its dual. Refer to [7] for details.

Figure 1: Pedestrian and non-pedestrian samples from the benchmark dataset.

In this experimental work, SVM classifiers with three different kernel functions, linear, quadratic and RBF kernels, are compared with the features calculated from previous section.

5 Experiments

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

5.1 Experiments on the dataset of [1]

This dataset consists of three training sets and two test sets. Each training set contains $4,800$ pedestrian examples and $5,000$ non-pedestrian examples (see Table 1). 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. Some examples of pedestrian and non-pedestrian samples

#	data splits	pedestrians/split	non-pedestr./split
Train	3	4800	5000
Test	2	4800	5000

Table 1: Benchmark dataset of [1].

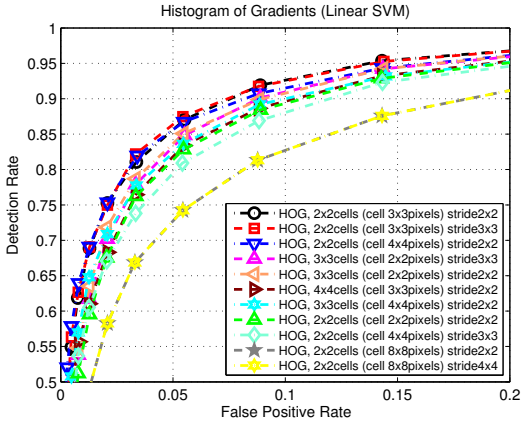

Figure 2: Performance of different descriptor blocks on histogram of oriented gradient (HOG) features.

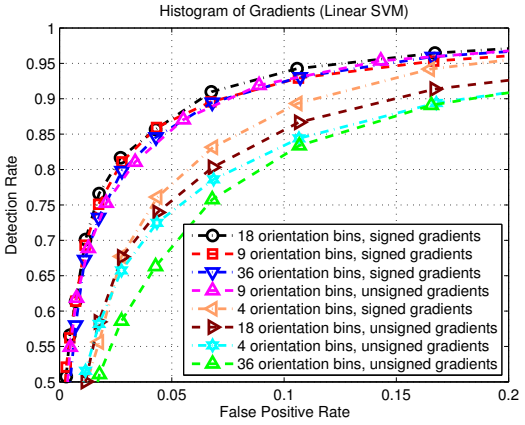

Figure 3: Performance of different orientation binning and gradient signs on histogram of oriented gradient (HOG) features.

are shown in Figure 1. Performance on the test sets is analyzed similarly to the techniques described in [1]. 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.

5.1.1 Parameter optimization

For the HOG features, the configurations reported in [8] 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 $3 - 8$ pixels and block size of $2 \times 2 - 4 \times 4$ cells).

Figure 2 shows our experimental results for various descriptor blocks trained on training set $\#1$, $\#2$ and tested on test set $\#1$ using the linear SVM. The number of orientation bins is set to 9 and the gradient vector is set to unsigned (unsigned gradients are when a gradient vector and its negative vote into the same bin). The following conclusions may be drawn from the figure:

- At datasets' resolution of 18×36 pixels, 2×2 cell blocks of 3×3 pixel cells and a descriptor stride of $2 - 3$ pixels, the perform is best.

- Increasing the number of cells in a block beyond 3×3 cells decreases the performance proportionally. The explanation for this might be that by increasing the number of cells, we are decreasing the feature length of HOG descriptors to be trained by SVM and, therefore, decreases the overall performance.

- Increasing the number of pixels in a cell (increasing cell width) decreases the performance. The reason may be due to the fact that by increasing cell width, the HOG descriptors fail to capture the informative spatial information.

- The size of the descriptor strides should be similar to the number of pixels in a cell for optimal performance.

- The HOG feature length per training sample in this experiment is between $2,000 - 4,000$. It seems that there exists a correlation between feature length and the overall performance *i.e.*, the longer the feature length, the better the performance.

Figure 3 shows the results for different orientation binning and gradient signs. The classifiers are again trained on training set $\#1$, $\#2$ and tested on test set $\#1$ using the linear SVM. The following observations can be made. Increasing the orientation bins increases the detection rate up to about 18 bins (unsigned gradients) and 9 bins (signed gradients). For small resolution human datasets, the gradient sign becomes relevant. The performance of signed gradients significantly outperforms the performance of unsigned gradients. This is in contrast to large resolution human datasets as reported in [8].

From the results in Figures 2 and 3, 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 unsigned gradients (total feature length is 8064) to train SVM classifiers. For the region covariance features, our preliminary experiments have shown a region of size 7×7 pixels, shifted at a step size of 2 pixels over the entire

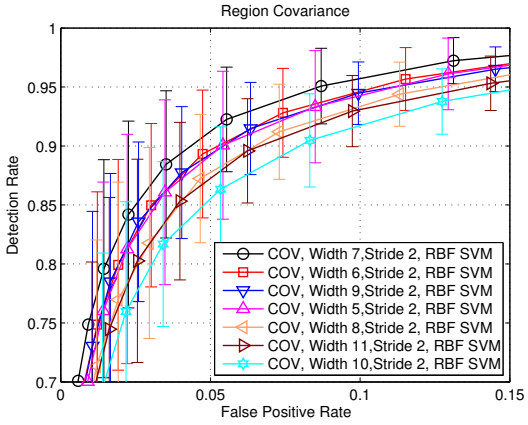

Figure 4: Performance of different parameters on region covariance features.

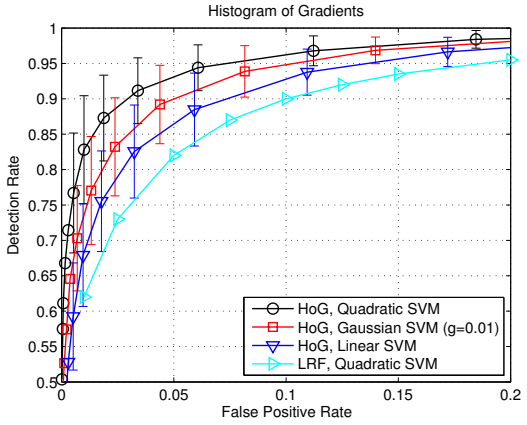

Figure 5: Performance of different classifiers on histogram of oriented gradients Features.

input image of size 18×36 to be optimal for our benchmark datasets. 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.

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. 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. Preliminary experimental results for region covariance are shown in Figure 4. For SVM classifiers, the HOG and region covariance descriptors are trained with linear, quadratic and Gaussian kernel SVM using SVMLight [15]. These results show that setting parameter γ in Gaussian RBF kernel to 0.01 gives the optimal performance. Results of different kernel functions are shown in the next section.

5.1.2 Results and analysis

This section provides experimental results and analysis of the techniques described in previous section. We compare our results with local receptive fields features as experimented in [1].

Figure 5 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 non-linear SVM outperforms the linear SVM. It is also interesting to note that the linear SVM trained using HOG features performs better than the non-linear SVM trained using LRF features. This

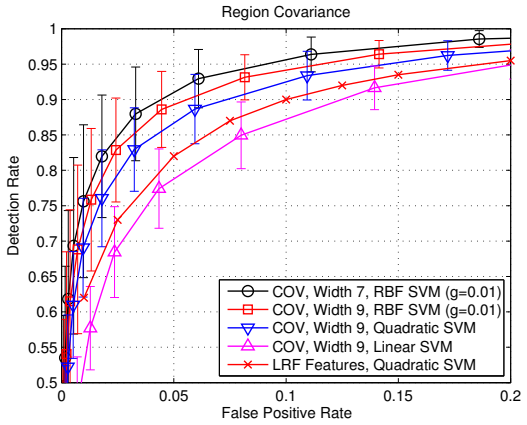

Figure 6: Performance of different parameters on region covariance features.

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

Figure 6 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 to other region sizes. From the figure, region covariance features perform better than LRF features when trained with the same SVM kernel (quadratic SVM). The RBF SVM performs best.

A comparison of the best performing results for different feature types are shown in Figure 7. 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. [9] concludes that the covariance descriptor outperforms the HOG descriptor (using human datasets of size 64×128 pixels

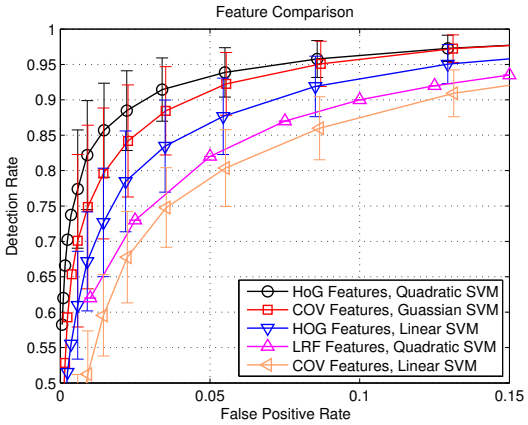

Figure 7: A performance comparison of the best classifiers for different feature types on the dataset of [1].

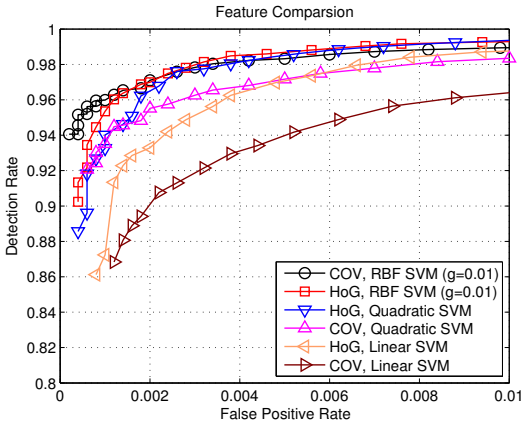

Figure 8: A performance comparison of the best classifiers for different feature types on the MIT CBCL dataset.

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. From the figure, 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 model).

5.2 Experiments on the MIT CBCL dataset

#	data splits	pedestrians/split	non-pedestr./split
Train	3	1840	5000
Test	2	1840	5000

Table 2: MIT CBCL pedestrian dataset. The non-pedestrian examples are randomly sampled from [1].

The MIT CBCL Pedestrian Dataset[2] consists of 924 non-mirrored pedestrian samples. Each sample has a resolution of 64×128. The database contains a combination of frontal and rear view human. We have applied the same techniques as described in [1] by dividing the pedestrian samples into five sets (Table 2). Each set consists of 184 pedestrian samples. Each sample is mirrored and shifted randomly by a few pixels in horizontal and vertical directions before being cropped and resized to a resolution of 18×36. Each sample contains approximately $2 - 3$ pixels of margin around the person on all four sides.

For MIT CBCL Pedestrian database, the parameters used are the same as the ones used previously in the dataset of [1].

5.2.1 Results and analysis

Figure 8 shows a comparison of experimental results on different feature types using the MIT CBCL pedestrian dataset. 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. As a result, there is a noticeable improvement in the experimental results compared to Figure 7.

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. Also nonlinear SVMs are always better the linear SVM.

5.3 Testing on INRIA images

We combine all the training sets from the dataset of [1] ($14,400$ positive samples and $15,000$ negative samples) and trained another set of classifiers using the parameters discussed in the previous section (non-linear SVMs). The classifiers are then tested on INRIA human datasets[3]. In order to speed up the detection time, the test image is scaled down to approximately $300 - 400$ pixels in width and height. Figure 9 shows some of the detection results on test image using HOG features and covariance features. Note that no post-processing has been applied to the detection results.

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

[3] http://pascal.inrialpes.fr/data/human/

Figure 9: Detection results on testing images from INRIA. The top row shows the detection results of Gaussian RBF SVM using Covariance features. The bottom row shows the detection results of quadratic SVM using HOG features. Again we see that HOG and covariance features perform very similarly.

Again we see that HOG and covariance features perform very similarly.

6 Conclusion

This paper presented an in-depth experimental study on pedestrian detection using three of the state-of-the-art local features extraction techniques. Our experimental results show that region covariance (correlation coefficient between image statistics) and normalized histogram of oriented gradients (HOG) features in dense overlapping grids significantly outperform the adaptive approach like local receptive fields (LRF) feature. In [1] the authors show that LRF is the best one among the features they have compared. Also we show that the covariance features' performance is very similar to HOG's, on both the datasets we have used.

Ongoing work includes the use of Adaboost to select the most discriminative features and construction of cascaded classifier to make the detection real-time.

Acknowledgments

The authors thank Dr. Fatih Porikli for helpful advice.

NICTA is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

References

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

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

[3] Y. Amit, D. Geman, and X. Fan. A coarse-to-fine strategy for multiclass shape detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(12):1606–1621, Dec 2004.

[4] C. Wöhler and J. Anlauf. An adaptable time-delay neural-network algorithm for image sequence analysis. *IEEE Trans. Neural Netw.*, 10(6):1531–1536, 1999.

[5] C. Papageorgiou and T. Poggio. A trainable system for object detection. *Int. J. Comp. Vis.*, 38(1):15–33, 2000.

[6] V. Vapnik. *The nature of statistical learning theory*. Statistics for Engineering and Information Science. Springer Verlag, Berlin, 2000.

[7] J. Shawe-Taylor and N. Cristianini. *Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, volume 1, pages 886–893, San Diego, CA, 2005.

[9] 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.

[10] 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.*, volume 2, pages 326–333, Washington, DC, 2004.

[11] Y. Wu and T. Yu. A field model for human detection and tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(5):753–765, 2006.

[12] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proc. Eur. Conf. Comp. Vis.*, volume 1, pages 69–81, Prague, Czech Republic, May 2004.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comp. Vis.*, 60(2):91–110, 2004.

[14] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. Eur. Conf. Comp. Vis.*, volume 2, pages 589–600, Graz, Austria, May 2006.

[15] T. Joachims. *Making large-Scale SVM Learning Practical*. Advances in Kernel Methods - Support Vector Learning. MIT Press, 1999.