An Experimental Study on Pedestrian Classification using Local Features

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*— 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].**

I. 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. 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. 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]. 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. 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. To our knowledge, these approaches have not been compared yet. It remains unclear whether silhouette based (HOG) or appearance based (covariance) 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 state-of-the-art pedestrian detection techniques: LRF, HOG and region covariance; along with various combination with SVM. The reason why we select these three features along with SVM classifiers is because 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.

The paper is organized as follows. Section II reviews various existing techniques for pedestrian detection. Sections III and IV describe methods used for feature extraction and a brief introduction to SVM. The experimental setup and experimental results are presented in Section V. The paper concludes in Section VI.

II. 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. 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. 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

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

978-1-4244-1684-4/08/$25.00 ©2008 IEEE

features approaches can be further divided into whole body detection and body parts detection [11], [12].

III. Feature extraction

Feature extraction is the first step in most object detection and pattern recognition algorithms. 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.

A. 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.

B. Histograms of oriented gradients

HoG was first introduced in the context of human detection by Dalal and Triggs [8]. Their method uses a dense grid of Histogram of Oriented Gradients, computed over blocks of size 16×16 pixels. Each block can be further divided into cells of size 8×8 pixels. Cells are integrated into a block in a sliding fashion. Also, blocks can overlap with each other.

For each region, a local $1D$ histogram of gradients over all the pixels in the cell is accumulated. Each orientation histogram divides the gradient angle range into 9 bins. The gradient magnitudes of the pixels in the cell are used to vote into the orientation histogram. Each block contains a concatenated histogram vector of all its cells. Hence, each block can be represented by a $36D$ feature vector that is normalized to an ℓ_2-norm unit length. Normalization introduces better invariance to illumination, shadowing and edge contrast. The final step is to collect these normalized block descriptors from all blocks of a dense overlapping grid of blocks into a combined feature vector. The feature vector can then be used to train a linear SVM classifier.

C. Region covariance

Tuzel, *et al.* [9], [13] 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

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)\mathrm{var}(Y)} = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y} \qquad (1)$$

$$\begin{aligned}\mathrm{cov}(X,Y) &= \mathbf{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \\ &= \frac{1}{n-1}\sum_k (X_k-\mu_X)(Y_k-\mu_Y), \qquad (2)\end{aligned}$$

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 and vertical direction $|\mathbf{I}_x|$, $|\mathbf{I}_y|$, 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 and vertical direction $|\mathbf{I}_{xx}|$, $|\mathbf{I}_{yy}|$. 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. Note that this treatment is different from [13], [9], where the covariance matrix is directly used as the feature and the distance between features is calculated in the Riemannian manifold.

IV. 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 [6] and Adaboost [2]. Due to space constraints we limit our explanation of SVM classifiers algorithm to an overview. 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. In this experimental work, SVM classifiers with three different kernel functions, linear, quadratic and RBF kernels, are combined with the features calculated from previous section.

V. Experiments

The experimental section is organized as follows. First, the datasets used in this experiment 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.

A. *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 I). 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. All samples are scaled to size 18×36 pixels. Performance on the test sets is analyzed similarly to the techniques described in [1].

1) Parameter optimization: From the preliminary experiments on the HOG features, 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). 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 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. For SVM classifiers, the HOG and region covariance descriptors are trained with linear, quadratic and Gaussian kernel SVM using SVMLight [14]. Preimilarny 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.

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

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

TABLE I: Benchmark dataset of [1].

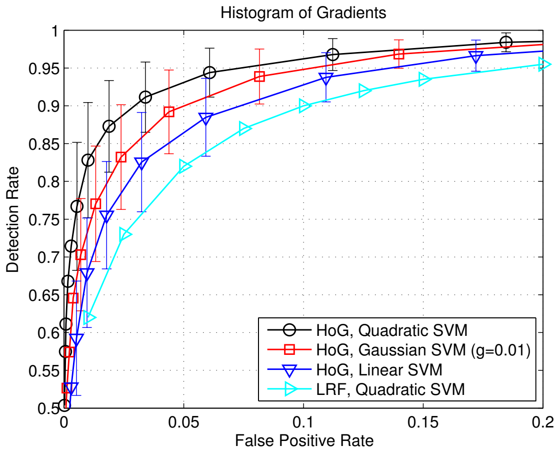

Fig. 1: Performance of different classifiers on histogram of oriented gradients Features.

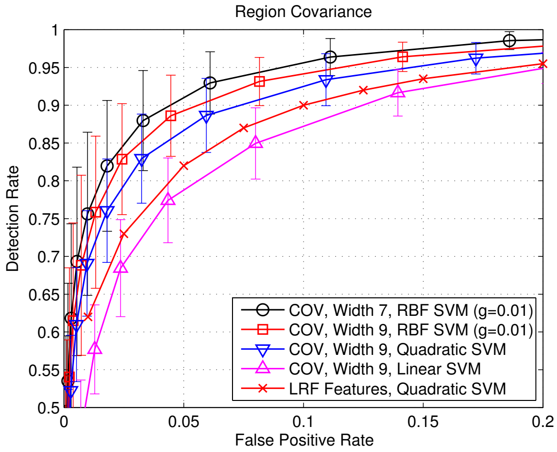

Fig. 2: Performance of different parameters on region covariance features.

Figure 2 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 3. 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 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.

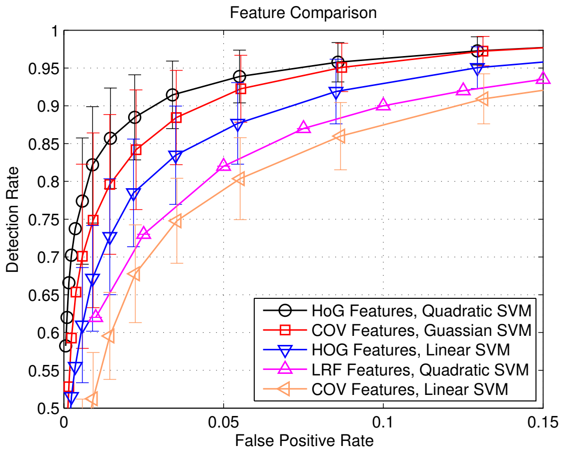

Fig. 3: A performance comparison of the best classifiers for different feature types on the dataset of [1].

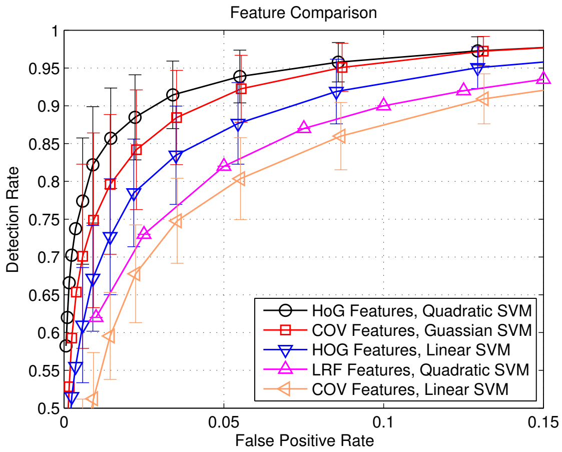

Fig. 4: A performance comparison of the best classifiers for different feature types on the MIT CBCL dataset.

In all experiments, nonlinear SVMs 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).

B. Experiments on the MIT CBCL dataset

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

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

The MIT CBCL Pedestrian Dataset[1] 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 II). Each set consists of 184 pedestrian samples. For MIT CBCL Pedestrian database, the parameters used are the same as the ones used previously in the dataset of [1].

1) Results and analysis: Figure 4 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 3.

VI. 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 and normalized histogram of oriented gradients (HOG) features in dense overlapping grids significantly outperform the adaptive approach like local receptive fields (LRF) feature.

Acknowledgments

The authors thank Dr. Fatih Porikli for helpful advice.

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] 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.

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

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