http://dx.doi.org/10.5392/IJoC.2013.9.3.001

Fast Pedestrian Detection Using Histogram of Oriented Gradients and Principal Components Analysis

Trung Quy Nguyen, Soo Hyung Kim, In Seop Na
School of Electronics and Computer Engineering
Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju, 500-757, Korea

ABSTRACT

In this paper, we propose a fast and accurate system for detecting pedestrians from a static image. Histogram of Oriented Gradients (HOG) is a well-known feature for pedestrian detection systems but extracting HOG is expensive due to its high dimensional vector. It will cause long processing time and large memory consumption in case of making a pedestrian detection system on high resolution image or video. In order to deal with this problem, we use Principal Components Analysis (PCA) technique to reduce the dimensionality of HOG. The output of PCA will be input for a linear SVM classifier for learning and testing. The experiment results showed that our proposed method reduces processing time but still maintains the similar detection rate. We got twenty five times faster than original HOG feature.

***Key words**: Pedestrian detection, Histogram of Oriented Gradients, Principal Components Analysis, linear SVM.*

1. INTRODUCTION

Pedestrian detection has many application in our life, the applications include robotics, entertainment, surveillance and advanced driver assistance systems. Detecting people in images is the difficult task in computer vision because there are many varied situations such as changes of appearances (difference clothes and color, changing size), wide variety articulated human poses, the unconstraint illumination and viewpoint, etc. Due to these important and challenge, pedestrian detection has attracted an extensive amount of interest from the computer vision community over the past few years.

Many techniques have been proposed in terms of features, models, and general architectures. Papageorgiouet al.[2] proposed a pedestrian detection algorithm based on Haar wavelets and polynomial SVM . Viola et al.[3] continues using Haar-like feature and a variant AdaBoost algorithm to select best weak classifier to procedure the strong clasifier, then combine these selected classifiers to build cascade for quickly reject all non-pedestrian windows and only keep pedestrian window through all cascade layers, this method got the near real time human detection system. In [1], Navneel Dalat and Bill Triggs introduced histograms of oriented gradients (HOG) feature which is inspired from SIFT descriptor of D.Love [10], and use linear SVM as a learning method with excellent detection results. In some recent survey about the state of art pedestrian detection systems [12], [13], authors showed that there are not any single feature which has been shown to outperform HOG. After HOG is proposed for human detection task with promised results, there are a lot researches continue improving HOG in term of performance and consumption time or adapt to other tasks such as tracking [6], [19], [20], action recognition [21], human pose estimation [22], [25].

Another approach that recently gets a lot of attentions is part-based method. In the contrary with the holistic techniques, it classifies different parts of human body (e.g., head, arms, legs, body) instead of classifying the entire human. Mohan et al. [16] use Haar wavelets and a quadratic SVM to independently classify four human parts (head, legs, right arm, and left arm), then combine these parts and classify by linear SVM. Felzenszwalb et al. [17] build pictorial structures represent objects by a collection of parts arranged in a deformable configuration. In this case, the authors use latent SVM and HOG.

In this paper, we aim to describe an effective pedestrian detection system. Firstly, we extract HOG from input image, then using Principal Components Analysis to reduce the dimensional of HOG, it can help we reduce computational time and resources consumption, it can help speed up the training phase, especially the classification phase. Next, linear SVM is used as training and classification tool with input is the output of PCA step. Using linear SVM for simplicity and speed, the experiment still is able to get very excellent results. Non-linear SVM can give a slightly better performance, but we have to trace off with the computational resources and processing time. Our system can archive the same performance with original HOG feature of Navneel Dalat and Bill Triggs with both MIT [9] and INRIA pedestrian dataset ([8]) while the processing time is reduced.

* *Corresponding author, Email: ypencil@daum.net*
Manuscript received Feb. 06, 2013; revised Aug 16, 2013; accepted Sep 03, 2013

2. RELATED WORK

Many researchers tried to reduce computational cost of pedestrian detection system. Zhu et al. [6] integrate the cascade of rejectors approach with HOG features to speed up detection system without losing performance. Adaboost is used for feature selection. They reduce processing time in classification time by adapt HOG with a faster classification algorithm. However, in our method, we follow another approach by optimizing HOG feature its self using PCA. Applying PCA to HOG feature in entire image has additional advantages. First, PCA will help reduce significantly the dimension of HOG. HOG is a high dimensional vector (3780 dimension), and hence it causes long processing time for extracting and classification. After PCA steps, time consumption for extracting and classification will be reduced. Second, because our training images are taken from natural scene images, there is variety of backgrounds; it will lead the noise and redundant information in final HOG feature. These noises and redundant information will be gotten rid through PCA steps and keep only the most principal information of human shape.

Other authors tried to apply PCA to HOG but it is different with our approach at some points. Lu et al [15] applied PCA to HOG feature for tracking purpose not for detection. Kobayashi et al [19] extract HOG on edge (human boundary) of training images then use these HOG feature for training PCA. But it cannot preserve all information of human shape in case of contrast between human and background is low, or it can lead redundant information if there is a lot of edges information in background and inside human boundary. In order to get rid these problems, they have to manual normalize all training images by background subtraction. It is not a practical task in case of large training dataset. In our approach, we train PCA by extracting HOG all block locations, and PCA will automatically filter the most principal characteristics of human images.

3. THE SYSTEM ARCHITECTURE

3.1 System Overall

Our pedestrian detection system has two main phases: training phase and detection phase. The overall of system is showed in figure 1. Detailed steps of two phases will be described in the next paragraphs.

3.2 Histogram of oriented gradients

The basic idea of Histogram of Oriented Gradients (HOG) is that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions [1]. Input image is divided into small spatial regions (cells), for each cell accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. In order to obtain a complete descriptor of an image, we have computed local histograms of gradient according to the following steps: Firstly, we compute gradients of the image; secondly, we build histogram of oriented gradient for each cell; thirdly, normalizing histograms within each block of cells; finally, all histograms are concatenated to build the final HOG feature (figure 2 and 3).

The following paragraphs give more details on each of these steps.

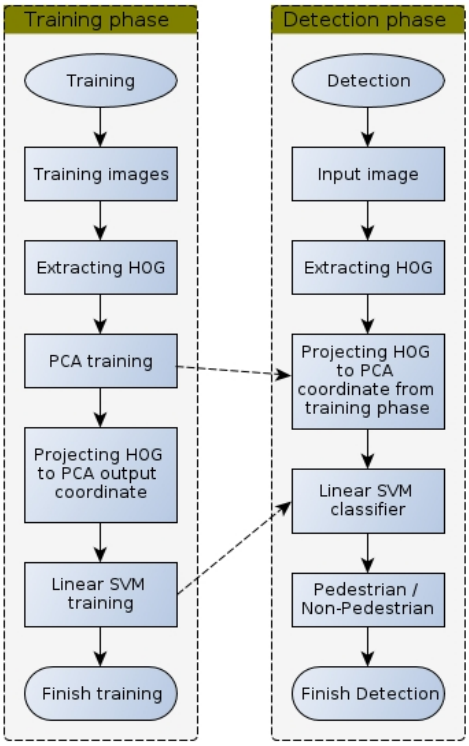

Fig. 1. An overview of our system

3.2.1 Gradient computation

An image gradient is a directional change in the intensity or color in an image. Image gradient are less susceptible to lighting changes. The gradient of an image has been simply obtained by filtering it with two 1-D filters:

- Horizontal : $(-1\ 0\ 1)$
- Vertical : $(-1\ 0\ 1)^T$

In other works, suppose that we have image $\mathrm{I}(\mathrm{x},\mathrm{y})$, the formula to calculate gradient magnitude $\mathrm{m}(\mathrm{x},\mathrm{y})$ and orientation $\theta(\mathrm{x},\mathrm{y})$

$$G_x(x,y) = I(x+1,y) - I(x-1,y)$$
$$G_y(x,y) = I(x,y+1) - I(x,y-1)$$
$$\mathrm{m}(\mathrm{x},\mathrm{y}) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2}$$
$$\theta(\mathrm{x},\mathrm{y}) = tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right)$$

3.2.2 Building histogram of oriented gradient for each cell

Size of cells is 8x8 pixels. For each cell, we compute the histogram of gradient by accumulating votes into bins for each orientation. Votes could be weighted by the magnitude of a gradient, so that histogram takes into account the importance of gradient at a given point. A gradient orientation around an edge

should be more significant than the one of a point in a nearly uniform region.

Gradient histograms measure the orientations and strengths of image gradients within an image region

3.2.3 Normalize histograms within each block of cells

Due to the illumination variations and other variability in the images, it is necessary to normalize cells histograms. Cells histograms are locally normalized, according to the values of the neighbored cells histograms. The normalization is done among a group of cells, which is called a block. Dalal and Bill tried with multiple block types with different cell and block sizes in the overall descriptor such as vertical (2x1 cell) and horizontal (1x2 cell) blocks and a combined descriptor including both vertical and horizontal pairs. They concluded that 2x2 blocks give the best accuracy.

A normalization factor is then computed over the block and all histograms within this block are normalized according to this normalization factor. Once this normalization step has been performed, all the histograms can be concatenated in a single feature vector. Different normalization schemes are possible for a vector V containing all histograms of a given block. The normalization factor nf could be obtained by these schemes:

- $\mathrm{L1}-\mathrm{norm}: nf = \frac{\mathrm{V}}{\|V\|_1+\varepsilon}$
- $\mathrm{L2}-\mathrm{norm}: nf = \frac{\mathrm{V}}{\|V\|_2^2+\varepsilon^2}$
- $\mathrm{L1}-\mathrm{sqrt}: nf = \sqrt{\frac{\mathrm{V}}{\|V\|_1+\varepsilon}}$
- $\mathrm{L2}-\mathrm{Hys}$ is $\mathrm{L2}-\mathrm{norm}$ by clipping at 0.2 (limiting the maximum values of v to 0.2)

ε is a small regularization constant. It is needed as we sometime evaluate empty gradients. The value of ε has no influence on the results.

Note that each cell occurrences several times with different normalization in final descriptor.

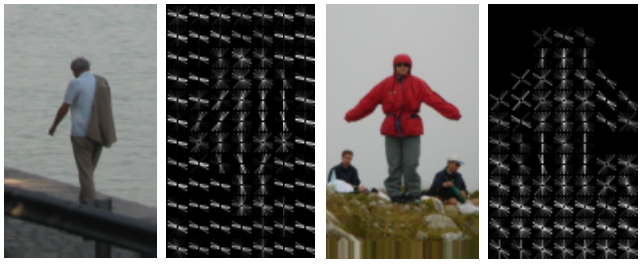

Fig. 2. Examples of HOG image (right) are extracted from input image (left). HOG can capture well human shape information.

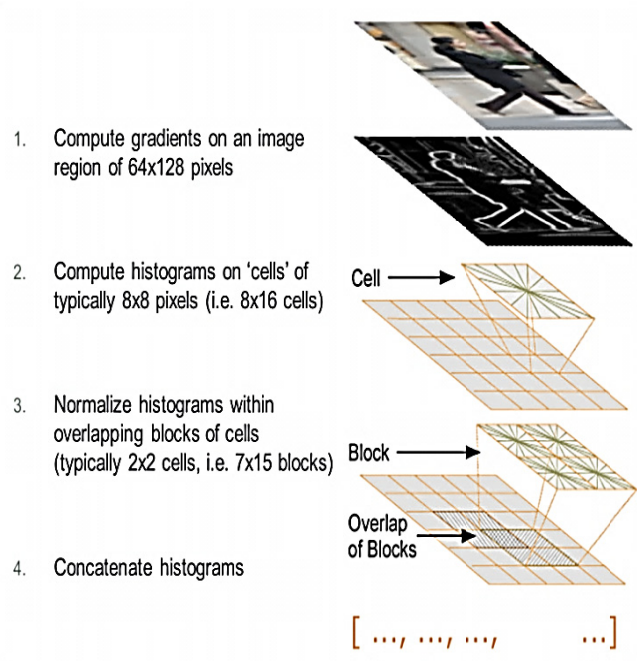

Fig. 3. Step by step to extract HOG from pedestrian image where the typically image size is 64x128 pixels.

For 64x128 input image, the dimensional of extracted HOG is about 3780. Each dimension values is stored in double variable type, it take 8 bytes in memory. For example, in our case, we use about 15000 images for training, the total memory consumes about 450MB, and it is large amount of memory. Total memory consumption will be required much more over calculation operators. Some programming environments such as Matlab have a limitation for one variable. Furthermore, with that high dimensional, the computational will be expensive. Therefore, the demand to compress HOG feature without losing accuracy is necessary.

3.3 Principal Components Analysis:

For dimensionality reduction, there are linear and non-linear methods. Non-linear methods are complex. It can increase the processing time of projection steps which compensate the benefit of dimensionality reduction. Therefore, linear principal components analysis is chosen because of its simplicity and efficiency. Principal Component Analysis (PCA) is a mathematical procedure for dimensionality reduction, has been widely applied in the area of computer vision such as face recognition [11]. All data is projected into its principal components which minimize the lost information. In other words, it will maximize variance of data in new coordinate system.

In training phase, we compute mean HOG vector $\bar{\boldsymbol{H}}$. All HOGs feature extracted from training images will be subtracted to $\bar{\boldsymbol{H}}$. We use Randomized SVD [23], [24] to train PCA. This algorithm is faster and less resources than built-in SVD function in Matlab when we want to get truncated SVD. Let $\mathbf{U} \in R^{n*p}$ denote first p principal components which is computed from HOG descriptors of training images and n is dimensional of HOG. We will project all HOG descriptors **H** to linear subspace spanned the principal components **U**:

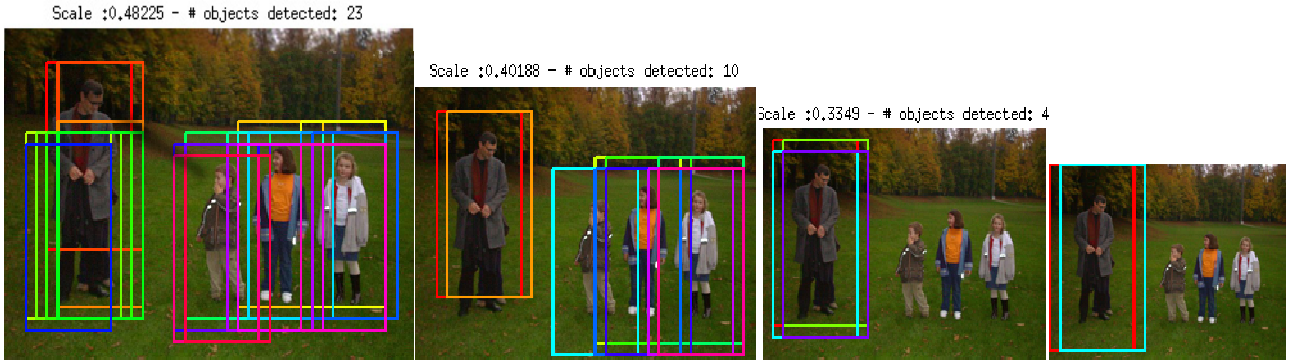

Fig. 4. Examples of detected windows after siding detector window at all image scales and locations

$$\mathbf{Y} = \boldsymbol{U}^{\boldsymbol{T}}(\boldsymbol{H} - \bar{\boldsymbol{H}}) \quad (1)$$

Where $\bar{\boldsymbol{H}}$ is the mean HOG of all training images, **Y** will be input for SVM classifier.

3.4 Linear SVM classifier

The SVM classifier finds a hyperplane which separates two-class data with maximal margin [7]. The margin is defined as the distance of the closest training point to the separating hyperplane. For given observations X, and corresponding labels Y which takes values ±1, one finds a classification function:

$$f(x) = \text{sign}(w^T x + b)$$

Where **w**, **b** represents the parameters of the hyperplane.

Data sets are not always linearly separable. In case of non-linear dataset, we can use kernel function to project dataset into a higher dimensional space in which data are linearly separable. But in case of pedestrian detection problems, typical linear SVM is sufficient to get the high detection rate. Using kernel function, we can get higher detection rate and decrease false positive but it take more computational resources and processing time.

3.5 Detection phase

In order to obtain initial object location hypotheses, we use the siding window technique, where the detector window is sided at various scales and all locations over the image (Figure 5). At one scale and location, we will extract HOG from this window image, and then project into principal component space using formula (1) to obtain the final feature vector. We term this feature is PCA-HOG. PCA-HOG feature will be classified by linear SVM to decide pedestrian or non-pedestrian.

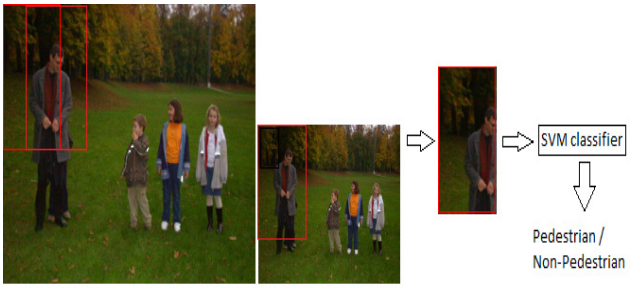

Fig. 5. Detector window is shifted at every scales and locations

After shifting detector window at all scale and all locations of image, we will obtain many detected window candidates which include human (Figure 4). In order to obtain the final detected windows, we have to evaluate all detected window pairs. Without loss of generality, we can denote two detected windows as W_1 and W_2.

$$\frac{area(W_1 \cap W_2)}{\max\left(area(W_1), area(W_2)\right)} > 0.5 \quad (2)$$

If two detected windows satisfy condition (2) we will eliminate smaller window and keep the bigger one for the next evaluating iterations. In many cases, we will have some false detection at one or some few scales. In order to eliminate false detection, we map all top-left points of detected windows into a 2D space. Then a mean-shift algorithm is used to group detected points into clusters. All small clusters, where the number of windows is less than a threshold, are considered as noises. We remove all these noisy windows from the detected window list.

Table 1. System performance comparison between HOG and PCA-HOG (p = 100) in INRIA dataset

	Accuracy	FPPW	Precision	Recall	Processing time (s)
HOG	98.39%	$8.6 * 10^{-3}$	0.96	0.95	35.00
PCA-HOG	98.60%	$5.9 * 10^{-3}$	0.98	0.95	1.34

4. EXPERIMENT RESULTS

In our experiment, we use two well-known pedestrian datasets: MIT CBCL Pedestrian dataset and INRIA Person dataset. MIT CBCL Pedestrian has total 924 positive images with frontal and back views only. MIT dataset doesn't separate into testing and training. In our experiment, we used 700 images for training and 224 images for testing. Our system recognized near-perfectly in this database with 99.55% correct detection rate, only one case is false detection. We also make an experiment with another much more challenge dataset - INRIA person dataset. In INRIA, all positive images is cropped to 64x128 pixels image which human is in center of positive images. Sizes of all negative images are variance. These negative images are taken from natural scene without including people. 'INRIA' contains 2416 positive images and 1218 negative images for training set. For testing set, INRIA include 1126 positive images and 453 negative images. In experiment, each negative image will generate randomly ten 64x128 negative window images for training and testing. It means that we used 12180 negative images for training and 4530 negative images for testing. Because of using 64x123 pixel images for training, our system only can detect pedestrians in the center of windows whose size is equal or bigger that window size. Without PCA steps, our system recognize with accuracy is 98.39% and false positive is $8.6*10^{-3}$. With PCA steps, accuracy is 98.60% and false positive is $5.9*10^{-3}$. The results of with and without PCA are just slightly different, but the processing time reduce significantly. When we apply PCA, it takes 1.34(s) to test entire INRIA data but in case of without PCA steps, it takes 35(s), about 25 times slower (See table 1 and figure 6 for more detail).

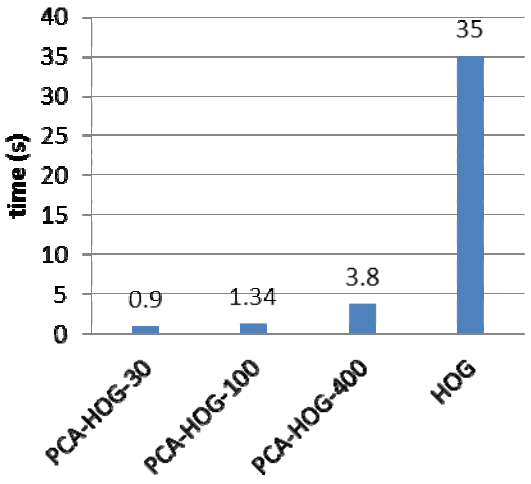

Fig. 6. Total processing time of PCA+SVM for all the samples in INRIA dataset

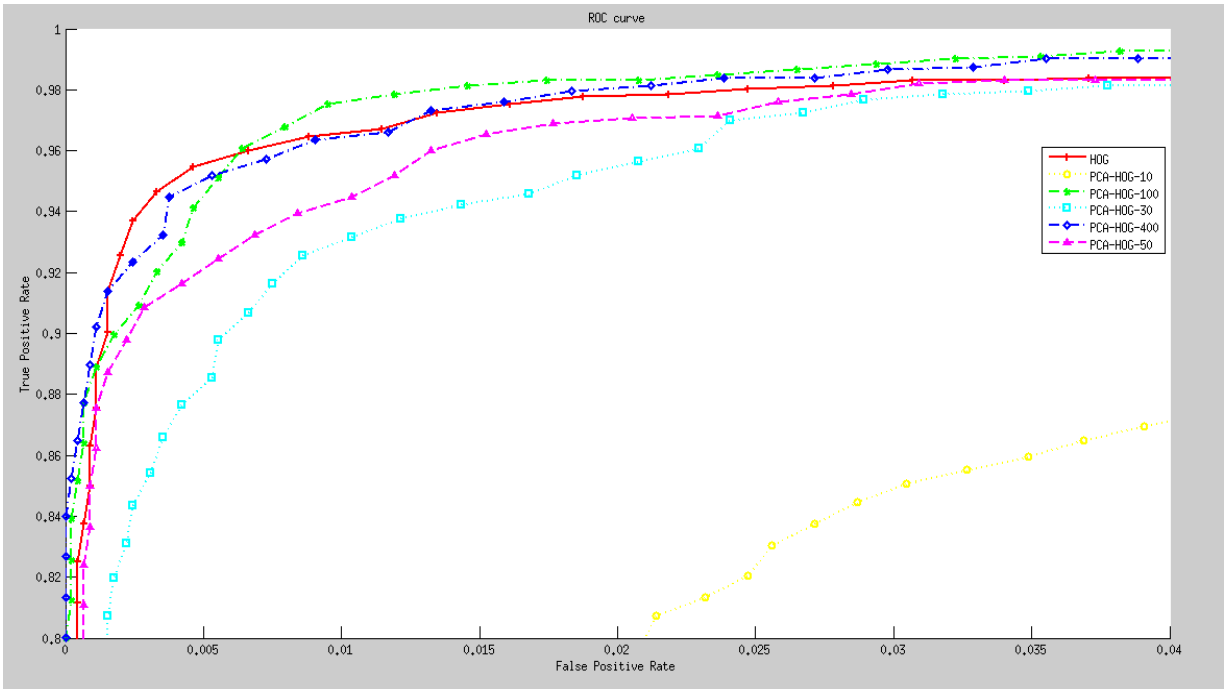

Fig. 7. Evaluation on INRIA pedestrian dataset using different number of principal components and original HOG

Fig. 8. Typical false positive window detections from INRIA dataset testing.

(a) (b) (c) (d)

Fig. 9. Detected windows from testing data, human is covered by rectangle box with its index number, these results prove that our system can detect very well on many different situations: different view(back view (a, c) and frontal view), objects with different scale (d) or crowded image (c).

In order to compare performance with original HOG feature, we create the ROC curve as showing in figure 7. For each observation, we have one classification score which is output by SVM process. If the score is larger or equal a specified threshold, we classify it to positive class, otherwise it is classified into negative class. To create the ROC curve the threshold value for the classification score is adjusted from $+\infty$ to $-\infty$. The true positive and false positive rates are 0 when threshold is adjusted to $+\infty$ and 1 when threshold is adjusted to $-\infty$. We find out that with 100 principal components (p=100), we get even better performance compare to original HOG. If p is 10, 30 or 50, the detection rate is lower than original HOG. But the detection rate is not increased if we increase p from 100 to 400. Processing time does not reduce much when we decrease number of principal components from 100 to 30 but the true positive rates reduce more clearly. Therefore, p = 100 is the optimal option. HOG can get higher performance at low false positive rate but when false positive rate is larger than $5*10^{-3}$ HOG have lower performance than PCA-HOG with p=100.

5. CONCLUSIONS AND DISCUSSIONS

In this paper, we demonstrate a completely pedestrian detection system. We proposed a method can detect pedestrians which are fast and accurate at the same time by using principal components analysis to reduce dimensional of HOG therefore speeding up the classification time and training time as well. Our system speed up pedestrian detection system about 25 times compare to original HOG/linear SVM system. We also find out that using 100 principal components can give a best trade off option between detection rate and processing time. Our method can execute require less resources than original HOG feature; it will have advantages when apply to low resource devices such as mobile phone. In figure 8, we show some typical false positive detection of our system, the most errors occur in window images which have strong vertical structure. We also show some our true positive samples in figure 9 with different situations such as different views, different detected scaled windows or crowed image. For future works, we are going to apply cascade approach and boosting for feature selection to further speeding up our system.

ACKNOWLEDGEMENT

"This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST)(2012-047759 and 2013-022495)."

REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, June 2005, pp. 886-893.

[2] C.P. Papageorgiou, M. Oren and T. Poggio, "A general framework for object detection," Sixth International Conference on Computer Vision, Jan. 1998, pp. 555-562.

[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1, 2001, p. I-511,p. I-518.

[4] V.D. Shet, J. Neumann, V. Ramesh and L.S. Davis, "Bilattice-based Logical Reasoning for Human Detection," IEEE Conference on Computer Vision and Pattern Recognition, June 2007, pp. 1-8.

[5] Li Zhang, Bo Wu and R. Nevatia, "Detection and Tracking of Multiple Humans with Extensive Pose Articulation," Computer Vision, ICCV 2007. IEEE 11th International Conference on, Oct. 2007, p. 1, p. 8, pp. 14-21.

[6] Q. Zhu, S. Avidan, M. Yeh, and K. Cheng, "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients,"Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2006, pp. 1491-1498.

[7] V. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, New York, 1998.

[8] INRIA Person Dataset, http://pascal.inrialpes.fr/data/human/ , 2007.

[9] MIT CBCL Pedestrian Database, http://cbcl.mit.edu/cbcl/software-datasets/PedestrianData.html, 2008.

[10] Lowe and D. G., "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, no. 2, 2004, pp. 91-110.

[11] M.A. Turk and A.P. Pentland, "Face Recognition Using Eigenfaces," IEEE Conf. on Computer Vision and Pattern Recognition, 1991, pp. 586-591.

[12] Piotr Dollar, Christian Wojek, Bernt Schiele and Pietro Perona, "Pedestrian Detection: An Evaluation of the State of the Art," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, April 2012, pp. 743-761.

[13] Enzweiler and D. M. Gavrila, "Monocular Pedestrian Detection: Survey and Experiments," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 12, pp. 2179-2195.

[14] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 7, July 2002, pp. 971-987.

[15] W. Lu and J. Little, "Simultaneous tracking and action recognition using the PCA-HOG descriptor," Proc. 3rd Can. Conf. Comput. Robot Vis., 2006, p. 6.

[16] C. Mohan, Papageorgiou and T. Poggio, "Example-Based Object Detection in Images by Components," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 4, Apr. 2001, pp. 349-361.

[17] P. Felzenszwalb, D. McAllester and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," Computer Vision and Pattern Recognition, CVPR 2008, IEEE Conference on , June 2008, p. 1, p. 8, pp. 23-28.

[18] Takuya Kobayashi, Akinori Hidaka and Takio Kurita. "Selection of Histograms of Oriented Gradients Features for Pedestrian Detection," ICONIP 2007, Part II, LNCS 4985, 2008, pp. 598-607.

[19] M.B. Kaaniche and F. Bremond, "Tracking HoG Descriptors for Gesture Recognition," Advanced Video and Signal Based Surveillance, AVSS '09. Sixth IEEE International Conference on, 2009, pp. 140-145.

[20] P. Bilinski, F. Bremond, Kaaniche and Mohamed Becha, "Multiple object tracking with occlusions using HOG descriptors and multi resolution images," Crime Detection and Prevention (ICDP 2009), 3rd International Conference on, Dec. 2009, p. 1, p. 3, p. 6.

[21] Jin Wang, Ping Liu, M.F.H., A. Kouzani and S. Nahavandi, "Human action recognition based on Pyramid Histogram of Oriented Gradients", Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, 2011, pp. 2449- 2454.

[22] K. Onishi, T. Takiguchi and Y. Ariki, "3D human posture estimation using the HOG features from monocular image," Pattern Recognition, 2008. ICPR 2008. 19th International Conference on , Dec. 2008, pp. 1-4, pp. 8-11.

[23] N. Halko, P. G. Martinsson, Y. Shkolnisky and M. Tygert, (2010). An algorithm for the principal component analysis of large data sets, Arxiv preprint arXiv:1007.5510, 0526. Retrieved April 1, 2011, from http://arxiv.org/abs/1007.5510.

[24] N. Halko, P. G. Martinsson and J. A. Tropp, (2009), Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. Arxiv preprint arXiv:0909.4061. Retrieved April 1, 2011, from http://arxiv.org/abs/0909.4061 .

[25] A. Agarwal and B. Triggs, "A local basis representation for estimating human pose from cluttered images", In: Proc. of ACCV, vol. 1, 2006, pp. 50-59.

Trung Quy Nguyen

He received his B.S. degree in Faculty of Mathematics and Computer Science from University of Science, Vietnam National University - Ho Chi Minh City in 2008. From 2008 to 2012, he was a software engineer at eSilicon Vietnam. Since 2012, he has been taking the M.S. course in Electronics & Computer Engineering at Chonnam National University, Korea. His research interests are pattern recognition, machine learning and web technologies.

Soo Hyung Kim

He received his B.S. degree in Computer Engineering from Seoul National University in 1986, and his M.S. and Ph.D degrees in Computer Science from Korea Advanced Institute of Science and Technology in 1988 and 1993, respectively. From 1990 to 1996, he was a senior member of research staff in Multimedia Research Center of Samsung Electronics Co., Korea. Since 1997, he has been a professor in the Department of Computer Science, Chonnam National University, Korea. His research interests are pattern recognition, document image processing, medical image processing, and ubiquitous computing.

In Seop Na

He received his B.S., M.S. and Ph.D. degree in Computer Science from Chonnam National University, Korea in 1997, 1999 and 2008, respectively. Since 2012, he has been a contract professor in Department of Computer Science, Chonnam National University, Korea. His research interests are image processing, pattern recognition, character recognition and digital library.