International Journal of Control and Automation
Vol.7, No.1 (2014), pp.235-242
http://dx.doi.org/10.14257/ijca.2014.7.1.20

Recursive Coarse-to-Fine Localization for Fast Object Detection

Quy Nguyen Trung, Dung Phan,Soo Hyung Kim, In Seop Na*
and Hyung Jeong Yang

School of Electronics and Computer Engineering Chonnam National University
77 Yongbong-ro, Gwangju, 500-757 South Korea

trungquy@gmail.com, dungptmy@gmail.com,shkim@chonnam.ac.kr,
ypencil@hanmail.net, hyungjeong@gmail.com

Abstract

Sliding window (SW) technique is one of common paradigms employed for object detection. However, the computational cost of this approach is so expensive because the detection window is scanned at all possible positions and scales. To overcome this problem, we propose a compact feature together with fast recursive coarse-to-fine object localization strategy. To build a compact feature, we project the Histograms of Oriented Gradient (HOG) features to linear subspace by Principal Component Analysis (PCA). We call this feature as PCA-HOG feature. The exploitation of the PCA-HOG feature not only helps the classifiers run faster but also still maintains the accuracy. In order to further speeding up the localization, we propose a recursive coarse-to-fine refinement to scan image. We scan image in both scale space and multi-resolution space from coarsest to finest resolutions. Only the best obtained hypothesis from the coarser resolution could be passed to finer resolution. Each resolution has its own linear Support Vector Machine (SVM) classifier and PCA-HOG features. Evaluation with INRIA dataset shows that our method achieves a significant speed-up compared to standard sliding window and original HOG feature, while even get higher detection accuracy.

Keywords: *Object detection; PCA-HOG; coarse-to-fine localization*

1. Introduction

Object detection is one of foundational problems in computer vision. Detecting objects in natural scene is a challenging task due to object variant appearance, the unconstraint illumination and viewpoints, *etc.* There are many approaches have been proposed on object detection in term of improving accuracy and reducing processing time. However, real-time object detection still is an open issue which its optimal goal is satisfaction both the speed and accuracy. To find the objects in images, the common technique is sliding window which the detector window is moved over all possible scales and positions. Features are extracted from windows (image patches) then pass these features to a binary classifier to make the decision whether it is object or not. However, standard sliding window technique is a brute-force approach. There are many researches tried to overcome the burden image scanning by speeding up the detector based on cascade of classifier. Viola and Jones in [6] combined Haar-like wavelet features and Adaboost algorithm to build cascade of weak classifiers which are able to quickly reject negative examples at first layers of the cascade and leave only

* Corresponding author In Seop Na (ypencil@hanmail.net)

ISSN: 2005-4297 IJCA

difficult examples through last layers. However, the drawback of cascade approaches is training time which can last some days or even weeks to complete training. Speeding feature extraction is one way to speed up SW. In [6], Viola et al. used integral image to speed up Haar-like feature. In [1], Navneel Dalat and Bill Triggs introduced histograms of oriented gradients (HOG) feature and combine with linear SVM classifier which give the state of art object detection for a single feature [2-3]. But the classification time is computational due to its high dimensional feature vector. In [7], Zhu et al. used integral histogram to speed up extracting HOG features and combined with cascade approach to further speeding up detection. Instead of improving classifier or feature extraction, other authors in [4-5] tried to reduce number of scanned window by using prior knowledge of human vision behavior which searches objects from coarse resolution to fine resolution. Images are scanned both in scale space and resolution space. The coarser image is resized from the finer one by a factor of two. Zhang *et al.*, in [4] used cascade of classifiers at every resolution. First coarse classifiers quickly reject almost negative windows; only leave fewer windows to finer resolution. These first coarse classifiers should be defined very well which is difficult because the image size in coarsest resolution is usually very small. In our proposed method, to avoid this problem we do not utilize cascade model. In lowest resolution image, we divide image to many blocks of pixels with radius δ. Only the highest response of this classifier level will be passed to higher resolution. Final detection score is the response of the finest classifier. In the contrast, Pedersoli *et al.*, in [5] obtained final scores by getting summation of scores at all resolution, but the scores from coarse resolutions are not reliable, so it can reduce the performance of detection. In our method, purpose of scoring from coarse resolutions is to guide the detector from the initial positions at coarse resolution move to higher confidence position at fine resolution.

In our method, we do not only optimize the scanning method but also speed up feature vector by projecting HOG feature into a linear subspace by using PCA. The PCA-HOG feature can help the classifier make decision faster because its low dimensional feature. The experimental results show that both speed and accuracy are improved with an approximate number of principal components. It is a worth thing to do because the number of possible windows in SW are very large. The SVM classifier at the finest resolution is designed by maximum accuracy and classifier at lower resolution is designed for speed purpose.

Contents of this paper are organized as follow: first proposal methods are described, in this section we explain how to extract PCA-HOG feature and how recursive coarse-to-fine localization work. In the next session, we explain about our experimental result, compare our method with HOG in [1]. Final section is our conclusion.

2. Proposed Methods

2.1. Features Extraction

In [1], Navneel Dalat and Bill Triggs introduced HOG feature for human detection with the excellent results. According surveys in [2] and [3], there is no other single feature has been shown to outperform HOG. HOG/linear-SVM framework has been used a base-line for many benchmarks of proposed methods [3]. In our paper, we use this framework as a baseline for evaluation purpose. To building HOG features, image is divided into smaller spatial units called “cells”; typical size is 8x8 pixels. Then the gradient magnitude of every pixel in each cell is computed and voted to the oriented histogram. Next, image is divided into blocks which contain multiple cells; typical size is 2x2 cells. Two adjacent blocks overlap 50% each other by vertical or horizontal. Normalization is done for each block. Final HOG features are

the concatenation vector of all normalized blocks histogram. HOG feature is a high dimensional feature vector (3780 dimensions).

Given an input image patch, HOG feature H is extracted, and then HOG features are projected into a linear subspace by using PCA. Let $U \in R^{n*p}$ denote first p principal components which is computed from HOG descriptors of training images and n is dimensional of HOG. We project all HOG descriptors H to linear subspace spanned the principal components U:

$$Y = U^{T}(H - \overline{H}) \quad (1)$$

where $\overline{H}$ is computed in training process by getting mean of HOG descriptors of training images, Y is the projected vector of HOG feature in new subspace, it is PCA-HOG feature.

By doing feature dimension reduction using PCA, it not only helps the classifier run faster but also still maintains the accuracy with an approximate p value.

2.2. Recursive Coarse-to-Fine Scanning Scheme

In our method, objects are searched over scale space and resolution space. In each scaled image, the detection window is scanned hieratically from the lowest resolution to full resolution of its scale, see Figure 1. The lower resolution image is obtained by getting down sample twice from higher adjacent resolution image. Each resolution level has its own classifier and a fixed size detection window. The ratio of detection window size between two adjacent resolutions is two. Classifiers of all resolution level can be trained separately with different parameters and window size. At lowest resolution level, image is divided into many grid blocks of pixels with radius δ. Only the position that it has highest score response from the classifier at coarsest resolution level is passed to higher resolution. The highest score position of one block have nine associate neighborhood positions at the higher resolution image. The highest score position of these nine positions have another nine associate neighborhood positions at the next higher resolution image. Recursively, we find the best hypothesis position at the full resolution image as a representative for all pixels of one block at lowest resolution. It means one block includes maximum one object and the number of hypotheses depends on the number of divided blocks at the lowest resolution. In contrast with standard SW, the number of hypothesis is the number of all possible position at every scale. Final detection scores are the response score of the classifier at full resolution. Table 1 describes recursive coarse-to-fine scanning algorithm at one scale. x, y, score are the position and score values of the best hypotheses of the input image at one resolution level respectively. The finest resolution is 1, user pre-defined value N is the coarsest level. The initial value for the second parameter level is N because we start to scan from coarsest resolution. After having positions and scores of best hypotheses at the finest levels, the true positive prediction are hypotheses that have scores are higher than a threshold. In our experiment, the threshold value is zero.

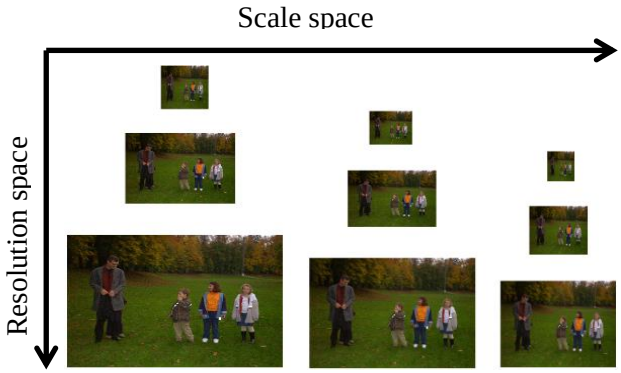

Figure 1. Coarse-to-fine localization framework. At each scale, image is scanned from lowest resolution to full resolution recursively.

Table 1. The pseudo code of recursive coarse-to-fine scanning algorithm at one scale

Function [x, y, score] = OneScaleScan(image, level)

If *(level is coarsest level)*

Divide image to many grid blocks of pixels with radius δ

Compute and return the position (x, y) and score of the best hypothesis among positions of each block.

Else

Down sample image twice: image = imresize(image,0.5)

Scan recursively at higher resolutions:[x_, y_, score_] = OneScaleScan(image, level - 1)

Upscale positions by twice : x = 2. x_; y = 2.*y_;*

Compute score of 9 positions: (x, y) and 8 neighborhoods.

Return the position (x, y) and score of the best hypothesis (has highest score) among above 9 positions.

End

3. Experimential Results

We evaluate our proposed method on INRIA pedestrian dataset which is used as a standard dataset for pedestrian detection [1]. It includes two test set: normalized images for per-window testing and original images for per-image testing. We use per-window methodology for features evaluations and use per-image for detection evaluations which follow evaluation methodology suggestions in [2]. We use linear SVM as a classifier for all experiments.

3.1. Feature Evaluation with Per-Window Methodology

We made experiments with three features: HOG, PCA-HOG-200 (p=200) and PCA-HOG-100 (p=100). Three classifiers are trained separately by linear SVM in full resolution window

(128x64 pixels) and tested with normalized images test set. This test set includes 1133 cropped pedestrian window images and 454 negative images. At the false positive rate 10-4, the miss rate is 10.30%, 8.53% and 11.99% for HOG, PCA-HOG-200 and PCA-HOG-100 respectively. To compare features at every false positive rate level, we plot Detection Error Tradeoff (DET) curves on a log-log scale between miss rate versus False Positive Per Window (FPPW) as in [8] (See Figure 2). The lower values are better. Figure 2 shows that PCA-HOG-200 out-performs both HOG and PCA-HOG-100 at every false positive level. HOG out-performs PCA-HOG-100 at low false positive level but from 2x10-3 false positive rate level PCA-HOG-100 have better performance.

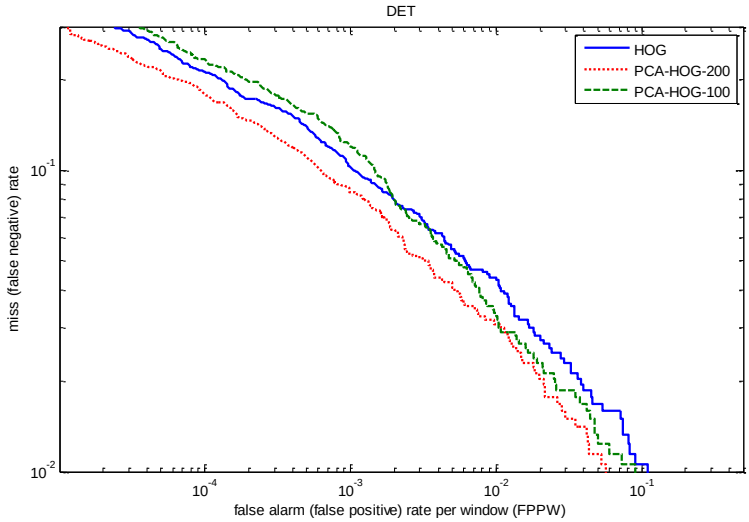

Figure 2. DET curve for comparison between HOG, PCA-HOG-200 (p=200) and PCA-HOG-100 (p=100)

3.2. Feature Design for Resolution Levels

Feature design for each resolution level plays an important role on a multi-resolution approach. The system can execute fast and accurate at the same time with an appropriate features. The design objective is to archive fast speed at low resolution levels and high accuracy at high resolution levels. Therefore, in our experiments, we select the good performance PCA-HOG-200 features for first (finest) resolution level and the faster PCA-HOG-100 features in second resolution level. In the third resolution level, the extracted HOG feature dimension is just 108 dimensions. It is low dimensional feature vector, it is not necessary to project to any subspaces. Further resolution levels are not used because the detection window is too small (8x16 pixels for 4th resolution).

3.3. Detecion Evaluation with Per-Image Methodology

We made three experiments for per-image evaluation. All experiments use linear SVM as a classifier tool. Used features in experiments are shown in Table 2. The first experiment (Dalal's method) is the same as [1] using HOG and linear SVM in full resolution. The second (Our method-1) and third (Our method-2) ones are our methods with difference configurations. To report the experimental results, we use the average precision (AP) score which represent the entire precision-recall curve by a single value. AP score is used in VOC challenges [9]. The experimental results are shown in Table 3.

Table 2. Features and number of resolutions are used in experiments. The experimental results are shown in Table 3. X: no use

Resolution - window size	Dalal's method [1]	Our method-1	Our method-2
1 – 64x128	PCA-HOG-200	HOG	PCA-HOG-200
2 – 32x64	PCA-HOG-100	X	PCA-HOG-100
3 – 16x32	HOG	X	X

Table 3. The computed average-precision on INRIA pedestrian dataset and the average speed-up ratio of experiment 2nd and 3rd

	Dalal's method [1]	Our method-1	Our method-2
Average-Precision	0.402	0.515	0.528
Speed-up	24.3x	1x	14.7x

The experiment shows the efficient of the combination of PCA-HOG and our coarse-to-fine localization scanning. In [5], authors used three resolution levels to get 12.2 times speed up while our method only use two resolution can get 14.7 times speed up and better performance. This is just a relative comparison but it still shows the efficient of our method because of the less resolution levels are used, the less information is lost, and so performance is higher while archive similar acceleration. Using third level of resolution reduces the performance significantly as shown in the third experiment. At this level the detection window is only 16x32 pixels, and it is too small to capture the information of big objects such as human which is used in our experiments. In Figure 3, we show the detection results of some sample images from second experiment.

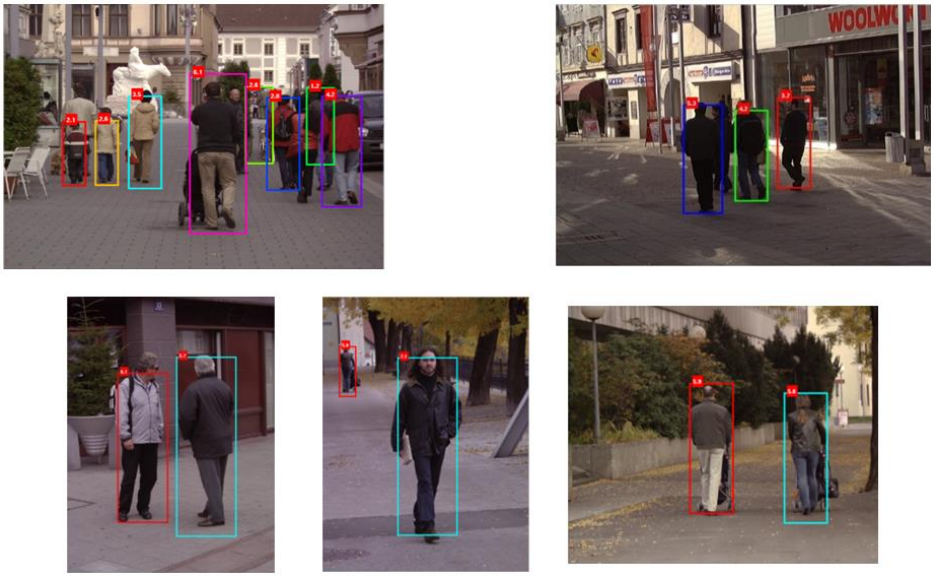

Figure 3. Detection results on sample images from INRIA dataset. Top-left values are detection score

4. Conclusions

In this paper, we propose object detection system using a compact feature PCA-HOG and an efficient recursive coarse-to-fine localization to both speeding up processing time and performance improvement at the same time. PCA-HOG features help the classifier make decision faster. The coarse-to-fine localization reduces the computation of scanning process. This recursive scanning reduces number of hypothesis windows significantly compare to standard sliding window technique. The speed-up factor is significant up to 14.7 times by using only 2 resolution levels. By using more resolution levels can reduce the detection performance.

Acknowledgements

This research was supported by the MSIP (Ministry of Science, ICT&Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2013-H0301-13-3005) supervised by the NIPA (National IT Industry Promotion Agency). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (2013-056480).

References

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, vol. 1, **(2005)**, pp. 886-893.
[2] P. Dollar, C. Wojek, B. Schiele and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, **(2012)**, pp. 743-761.
[3] D. Geronimo, A. M. Lopez, A. D. Sappa and T. Graf, "Survey of Pedestrian Detection for Advanced Driver Assistance Systems", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 7, **(2010)**, pp. 1239-1258.
[4] W. Zhang, G. Zelinsky and D. Samaras, "Real-time Accurate Object Detection using Multiple Resolutions", IEEE 11th International Conference on Computer Vision, **(2007)**, pp. 1-8.
[5] M. Pedersoli, J. Gonzàlez, A. D. Bagdanov and J. J. Villanueva, "Recursive coarse-to-fine localization for fast object detection", in Proceedings of the 11th European conference on Computer vision, **(2010)**, pp. 280-293.
[6] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, **(2001)**, pp. 511-518.
[7] Q. Zhu, S. Avidan, M. Yeh and K. Cheng, "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients", in Proceedings of the IEEE Conferences Computer Vision and Pattern Recognition, **(2006)**, pp. 1491-1498.
[8] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, "The DET curve in assessment of detection task performance", In Proceedings of the Conference on Speech Communication and Technology, **(1997)**, pp. 1895-1898.
[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge", International Journal of Computer Vision, vol. 88, no. 2, **(2010)**, pp. 303-338.

Authors

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.

Dung Phan

She received her B.S. degree in Computer Science of University of Science - Ho Chi Minh City, Vietnam in 2009. After that she has worked at eSilicon and Iritech Company for 2 years. Since 2012, she has been a master process student in the Department of Computer Science, Chonnam National University, Korea. Her main research interests include image processing, pattern recognition and object matching.

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 research professor in Department of Computer Science, Chonnam National University, Korea. His research interests are image processing, pattern recognition, character recognition and digital library.

Hyung-Jeong Yang

She received her B.S., M.S. and Ph. D from Chonbuk National University, Korea. She is currently an associate professor at Dept. of Electronics and Computer Engineering, Chonnam National University, Gwangju, Korea. Her main research interests include multimedia data mining, pattern recognition, artificial intelligence, e-Learning, and e-Design