A Two-Stage Approach to People and Vehicle Detection With HOG-Based SVM

Feng Han, Ying Shan, Ryan Cekander, Harpreet S. Sawhney, and Rakesh Kumar
Sarnoff Corporation
201 Washington Road
Princeton, NJ 08540, USA
{fhan, yshan, rcekander, hsawhney, rkumar}@sarnoff.com

***Abstract*— In this paper, we present a two-stage approach to robustly detect people and vehicles in static images using extended histogram of oriented gradient (HOG) and SVM for classification. The first stage is focus of attention generation, in which possible people and vehicle locations are hypothesized. This step uses stereo cue and generates potential target locations using some prior knowledge about what people and vehicle may look like in the depth map of the whole scene. The second stage is hypothesis verification. In this stage, all the hypothesis are verified by a strong classifier using extended HOG feature and SVM, which is robust to the wide range of variations of poses and viewpoints within people and vehicles. By adaptively combining the two stages, the final system achieves both speeding up and performance improvement. The system has been tested on some challenging datasets and illustrates good performance.**

I. Introduction

Automatic object detection and classification is a key enabler for applications in robotics, navigation, surveillance, or automated personal assistance. On the other hand, automatic object detection is a difficult task. The main challenge is the amount of variation in visual appearance. An object detector must cope with both the variation within the object category and the diversity of visual imagery that exists in the world at large. For example, cars vary in size, shape, color, and in small details such as the headlights, grille, and tires. The lighting, surrounding scenery, and an object's pose affect its appearance. A car detection algorithm must also distinguish cars from all other visual patterns that may occur in the world, such as similar looking rectangular objects.

The common approach to automatic object detection is shifting a search window over an input image and categorizing the object in the window with a classifier. To speed up the system without losing classification performance, one can exploit the following two characteristics common to most vision-based detection tasks: First, the vast majority of the analyzed patterns in an image belong to the background class. For example, the ratio of non-face to face patterns in the tests in [8] is about 50,000 to 1. Second, many of the background patterns can be easily distinguished from the objects. Based on these two observations, object detection is always carried out in a two-stage scheme as illustrated in Figure 1: First, all the regions in the image that potentially contain the target objects are identified. This is what we call "focus of attention's mechanism". Second, the selected regions are verified by a classifier.

Numerous approaches to focus of attention generation have been proposed in the literature. Most of them fall into one of the following three categories: (1) knowledge-based, (2) stereo-based, and (3) motion-based. Knowledge-based methods make use of our knowledge about object shape and color as well as general information about the context. For instance, the prior knowledge that vehicles are symmetric about the vertical axis has been used in vehicle detection approaches using the intensity or edge map in [1], [2]. Stereo-based approaches usually employ the Inverse Perspective Mapping (IMP) [3] to estimate the locations of vehicles, people, and obstacles in images. One specific example is the work by Bertozzi et al. [4], in which the IMPs are computed from the left and right images respectively and compared with each other. Based on the comparison, the objects that were not on the ground plane can be easily found. With this information, the free space in the scene can be determined at the same time. Most motion-based methods detect objects such as vehicles, people, and obstacles using optical flow. However, generating a displacement vector for each pixel is time-consuming and also impractical for a real-time system. To attach this problem, some discrete methods use image features such as color blobs [5] or local intensity minima and maxima [6] as the basic unit and have produced some better results.

A number of different approaches to hypothesis verification that use some form of learning have been proposed in the literature. In these approaches, the characteristics of the object class are learned from a set of training images which should capture the intra-class variabilities. Usually, the variability of the non-object class is also modelled to improve performance. First, each training image is represented by a set of local or global features (e.g. Harr wavelet, SIFT, Shape Context) [8], [9], [16], [17] into some underlying configuration (e.g. "bag of features", constellation model) [10], [11], [12], [13], [14]. Then, the decision boundary between the object and non-object classes is learned either by training a classifier (e.g., Adaboost, Support Vector Machine, Neural Network (NN)) or by modelling the probability distribution of the features in each class (e.g., using the Bayes rule assuming Gaussian distributions) [8], [11], [10]. These methods differ on the details of the features and decision functions, but more fundamentally they differ in how strictly the geometry

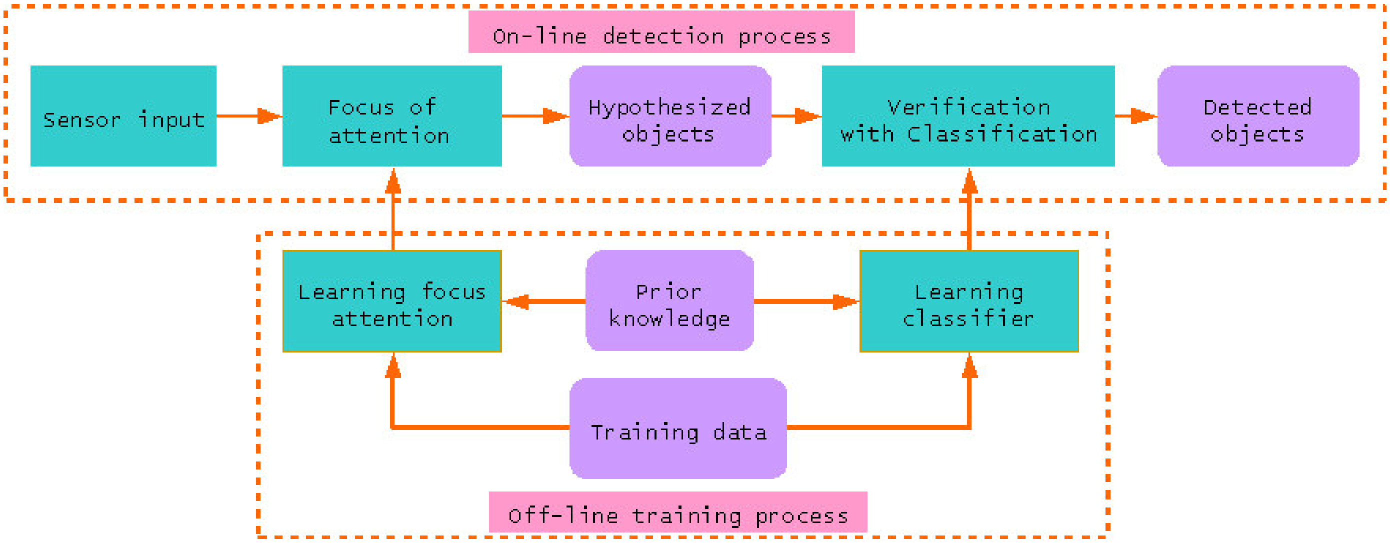


Fig. 1. Overview of the two-stage system for object detection.

of the configuration of parts constituting an object class is constrained.

Following the two-stage paradigm discussed above, we have built a system as illustrated in Figure 1 to stably detect standing people and vehicles over a wide range of viewpoints. The rest of paper is organized as follows: Section II presents focus of attention generation using stereo cue; Section III details the hypothesis verification with HOG-Based SVM; Section IV describes some implementation issues and a series of experiments; Section V concludes the paper.

II. Focus of Attention Using Stereo Cue

To generate focus of attention for the target objects in the scene, we use a stereo matching algorithm [7] to get the depth map of the scene. One example pair of left image and right image, and the depth map computed from this stereo pair are shown in Figure 2. In the depth map, red color implies closer points and blue to green implies further points.

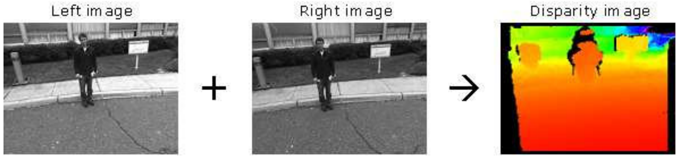


Fig. 2. Compute depth map from the stereo images.

After getting the depth map of the scene, we can further align it with the ground plane with the help of the IMU attached with the stereo system, which gives us the pitch angle. Then we can remove the ground plane from the depth map. For the remaining depth map, we project it to the XZ plane and represent it with a uniform grid. For each cell in this grid, we compute the hight and pixel density to get the ‘height map” and ‘occupancy map” as illustrated in Figure 3 (a) and (b) respectively. Then we compute the response of a predefined adaptive Gaussian kernel on the ‘occupant map”. Finally, we choose peaks with local maximum response as the target object position hypothesis. An example result of this process is shown in Figure 3. Note that the algorithm has discovered relatively compact vertical objects.

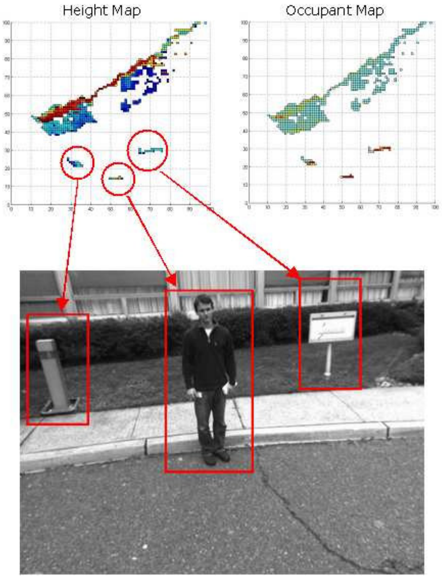


Fig. 3. Generate target hypothesis based on the height map and occupancy map.

III. Classification by HOG-Based SVM

We develop separate classifiers for each object class that are each specialized to one specific aspect or pose. For example, we have one classifier specialized to front/rear view of people and one that is specialized to side view of people. We apply these view-pose-based classifiers in parallel and then combine their results. If there are multiple detections at the same or adjacent locations, the system selects the most likely one through non-maximum suppression.

We empirically determined the number of views/poses to model for each object. For people we use two view-based detectors: front/rear and side view, as shown in Figure 4. For cars we use eight detectors, which are specialized to each of the eight aspects shown in Figure 5.

Each of these detectors is not only specialized in orientation, but is trained to find the object only at a specified size within a rectangular image window. Therefore, to be able to detect the object at any position within an image, we re-apply the detectors for all possible positions of this rectangular window. Then to be able to detect the object at any size we iteratively resize the input image and re-apply the detectors in the same fashion to each resized image.

Fig. 4. Examples poses for people.

To build each view-pose-based classifier, we extend the histogram of oriented gradient (HOG) [15] representation and use support vector machines (SVM) as the classifier [20], [21]. Unlike some commonly used representations, the extended histogram of oriented gradient gives good generalization by grouping only perceptually similar images together. With a support vector machine, this gives rise to a decision function that discriminates object and non-object patterns reliably in images under different kinds of conditions and results good performance on some challenging datasets.

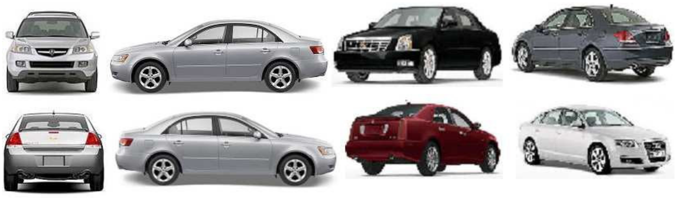


Fig. 5. Example viewpoints for vehicles.

A. Object Class Representation

1) HOG feature: Histogram of oriented gradient (HOG) is an adaptation of Lowe's Scale Invariant Feature Transformation (SIFT) approach to wide baseline image matching [16] with local spatial histogramming and normalization. In this work, HOG is used to provide the underlying image patch descriptor for matching scale invariant key points. SIFT-style approaches perform remarkably well in this application.

A HOG feature is created by first computing the gradient magnitude and orientation at each image sample point in a region around an anchor point. The region is split into NxN subregions. An orientation histogram for each subregion is then formed by accumulating samples within the subregion, weighted by gradient magnitudes. Concatenating the histograms from all the subregions gives the final HOG feature vector as illustrated in Figure 6.

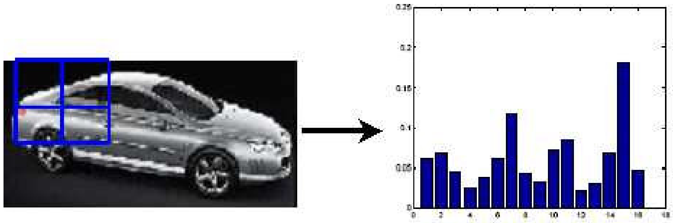


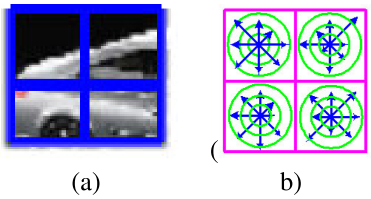
Fig. 6. HOG feature computation and structure.

2) Extend HOG by Incorporating Spatial Locality: The standard HOG feature only encodes the gradient orientation of one image patch, no matter where this orientation is from in this patch. Therefore, it is not discriminative enough if the spatial property of the underlying structure of the image patch is crucial. This is especially true for highly structured objects like vehicles. To incorporate the spatial property in HOG feature, we add one distance dimension to the angle dimension in the binning of all the pixels within each subregion. The distance is relative to the center of each subregion. The new binning process is illustrated in Figure 7.

3) Dense Grid Representation: Following [15], we divide the image window into small spatial regions, which consists of a number of subregions (or cells). For each cell we accumulate a local 1-D histogram of gradient directions over the pixels of the cell. The combined histogram entries form the representation. For better invariance to illumination, shadowing, etc., it is also useful to contrast-normalize the local responses before using them. This can be done by accumulating a measure of local histogram over somewhat larger spatial regions (or blocks) and using the results to normalize all of the cells in the block.

B. SVM Classifier

In this paper we choose the support vector machine [20], [21] as the classifying function. The Support Vector Machine



(a) (b)

Fig. 7. Binning both distance and gradient direction for the pixels in each sub-region to compute the extended HOG feature. (a) sample image patch. (b) distance and direction ranges to do the binning.

(SVM) is a statistical learning method based on the structure risk minimization principle. It's efficiency has been proved in many pattern recognition applications [20], [22], [23]. In the binary classification case, the objective of the SVM is to find a best separating hyperplane with a maximum margin.

The form of a SVM classifier is:

$$y = sign(\sum_{i=1}^{N} y_i \alpha_i K(x, x_i) + b),$$

where x is the feature vector of an observation example, $y \in \{+1, -1\}$ is a class label, x_i is the feature vector of the i^{th} training sample, N is the number of training samples, and $K(x, x_i)$ is the kernel function. Through the learning process, $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_N\}$ is computed.

One distinct advantage this type of classifiers has over traditional neural networks is that support vector machines achieve better generalization performance. While neural networks such as multiple layer perceptrons (MLPs) can produce low error rate on training data, there is no guarantee that this will translate into good performance on test data. Multiple layer perceptrons minimize the mean squared error over the training data (empirical risk minimization) where support vector machines use an additional principal called structural risk minimization [21]. The purpose of structural risk minimization is to give an upper bound on the expected generalization error.

Compared with the popular Adaboost classifiers, SVM is slower in test stage. However, the training of SVM is much faster than that of Adaboost classifiers.

IV. Experiments

A. Training

Our people training database contains images of 2000 standing people with various aspects, poses, and illumination conditions. Some of these images are from the public downloadable MIT people dataset and INRIA people dataset, while the rest are taken by ourselves. The resolution of each image is 64x128. For the vehicle training data, we collected 1000 images with 128x64 resolution and containing four types of vehicles (sedan, minivan/SUV, pick-up truck and U-Haul type truck) across a wide range of viewpoints. We also generate some rendered vehicle images, some of which are shown in Figure 8, by 3D vehicle models and use them as training data. Using this type of virtual training data is crucial since sometimes it is too time consuming or even impossible to get normal training data covering all possible pose-view variations for some object classes. The performance of vehicle classifier trained using these rendered images is tested in Section IV-D.

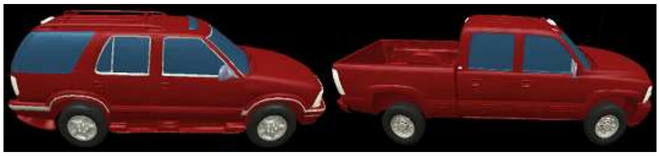


Fig. 8. Samples of rendered vehicle images.

One very important issue in the classifier training for one object class is how to select effective negative training samples. As negative training samples include all kinds of images, a prohibitively large set is needed in order to be representative, which would also require infeasible amount of computation in training. To alleviate this problem, a bootstrapping method, proposed by Sung and Poggio [24], is used to incrementally train the classifier as illustrated in Figure 9.

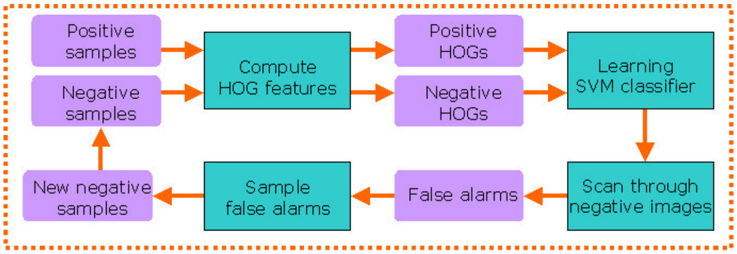


Fig. 9. The bootstrap training diagram.

All the view-pose-based SVM classifiers for each object class are trained separately, but with the same negative training samples. In this way, their outputs can compete with each other to remove multiple detections through non-maximum suppression.

B. Detection

As each specific view-pose-based classifier for every object class is designed on an image window with specific size (64x128 for people, 128x64 for vehicle), it implicitly requires that the to-be-detected objects lie roughly within a specific window in the testing images. To detect all the objects appearing at different scales in the test image, we build an image pyramid by successively up sampling and/or down sampling the test image by a factor of 1.2 till all the objects in the test image are scaled to the image window size at some layer in the pyramid.

C. Evaluation of Detection Results

Evaluation of detection results was performed using ROC curve analysis. The output required to generate such curves is a set of bounding boxes with corresponding "confidence" values, with large values indicating high confidence that the detection corresponds to an instance of the object class of interest. Figure 10 shows some example ROC curves, obtained by applying a set of thresholds to the confidence output by the SVM classifier. On the x-axis is plotted the average number of false alarms on one image; on the y-axis is detection rate. The ROC curve makes it easy to observe the tradeoff between the two; some thresholds may have high detection rate but more false alarms, while other thresholds may give more balanced performance.

To generate the ROC curves, we also need a criteria to evaluate the detection output. Judging each detection output by a method as either a true positive (object) or false positive (non-object) requires comparing the corresponding bounding box predicted by the method with ground truth bounding boxes of objects in the test set. To be considered a correct detection, the area of overlap α_{ovlp} between the predicted bounding box B_p and ground truth bounding box B_{gt} was required to exceed 50% by the formula used in [25],

$$\alpha_{ovlp} = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}.$$

In [25], the threshold of 50% was set low to account for inaccuracies in bounding boxes in the ground truth data. This inaccuracy of ground truth is due to some ambiguities, for example defining the bounding box for a highly non-convex object, e.g. a side view of a motorbike or a car with an extended radio aerial.

D. Performance of People and Vehicle Classifier

To test the performance of the trained classifier for people, we apply it to one people dataset selected from INRIA people database and PSACAL database [26]. This people dataset consists of 800 images and most people in the images are standing or walking. The performance curve of the people classifier is shown in Figure 10 (a).

To test the performance of the trained classifier for vehicle, we apply it to two vehicle datasets. The first one is the UIUC dataset [11], which consists of 278 images of vehicles in side-view. The second one consists of 600 images selected from PSACAL database [26] and vehicles appear in any poses in the images. The performance curves of the vehicle classifier on these two datasets are shown in and Figure 10 (b) and (c) respectively.

In all the above testing, we search the whole image without using any focus of attention. Some typical results for these two classifiers on the three datasets are shown in Figure 11 and Figure 12 respectively.

We also test the performance of classifiers using the rendered images as training data. To do this, we use the rendered vehicle images to train a vehicle classifier and apply it to the selected PASCAL dataset. The performance curves of this classifier and then one using normal images as training data are shown in Figure 10 (d) together for comparison, from which we can see that the classifier using rendered images as training data can achieve compatible performance.

E. Performance of the final two-stage system

To test the performance of the two-stage system, we apply it on 100 images that contain both standing people and vehicles spanning a variety of viewpoints. To show the performance improvement achieved by incorporating the first stage of focus of attention generation by stereo cue, we compare the performance of the system by turning on and off the first stage. In Figure 13 (a), we show the two ROC curves corresponding to turning on and off the first stage in the system for people detection. From these two comparisons, we can clearly see that the focus of attention generation stage helps a lot to reduce false alarms. Figure 13 (b) shows the same case for vehicle detection. Some typical detection results in this testing are shown in Figure 14 and Figure 15 for people detection and vehicle detection respectively.

V. Conclusion and Future Work

In this paper, we present a two-stage approach to robustly detect standing people and vehicles over a wide range of viewpoints. The first stage is focus of attention generation, in which possible people and vehicle locations are hypothesized. This step uses stereo cue and generates potential target locations using some prior knowledge about what people and vehicle may appear in the depth map of the whole scene. The second stage is hypothesis verification. In this stage, all the hypothesis are verified by a strong classifier using extended HOG feature and SVM, which is robust to the wide range of variations of poses and viewpoints within people and vehicles.

Although the current two-stage system works reasonably well, the process to manually separate the training samples into pre-defined intra-class categories based on their view/pose is too time consuming and inherently ambiguous. In addition, the errors caused by improperly defined categories and incorrectly assigned labels will eventually be propagated into the final classifier and deter the object detection performance. Recently, we have proposed a novel computational framework that unifies automatic categorization, through training of a classifier for each intra-class exemplar, and the training of a strong classifier combining the individual exemplar-based classifiers with a single objective function [27]. We are working to incorporate the current classifiers into the unified framework to dramatically reduce the training time and improve the performance. Furthermore, we are also working on incorporating motion and video constraints in classification. Increasing the number of detectable object classes to a large set is also a goal.

References

[1] A. Kuehnle, ‘Symmetry-based recognition for vehicle rears,” *Pattern Recognition Letters*, vol. 12, pp. 249258, 1991.

[2] T. Zielke, M. Brauckmann and W. V. Seelen, ‘Intensity and edge-based symmetry detection with an application to carfollowing,” *CVGIP:Image Understanding*, vol. 58, pp. 177190, 1993.

[3] H. Mallot, H. Bulthoff, J. Little, and S. Bohrer, ‘Inverse perspective mapping simplifies optical flow computation and obstacle detection,” *Biological Cybernetics*, vol. 64, no. 3, pp. 177185, 1991.

[4] M. Bertozzi and A. Broggi, ‘Gold: A parallel real-time stereo vision system for generic obstacle and lane detection,” *IEEE Trans. on Image Processing*, vol. 7, pp. 6281, 1998.

[5] B. Heisele and W. Ritter, ‘Obstacle detection based on color blob flow,” *IEEE Intelligent Vehicles Symposium*, pp. 282286, 1995.

[6] D. Koller, N. Heinze and H. Nagel, ‘Algorithm characterization of vehicle trajectories from image sequences by motion verbs,” *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 9095, 1991.

[7] G. van der Wal and M. Hansen and M. Piacentino, “The Acadia Vision Processor”, *International Workshop on Computer Architectures for Machine Perception*, September 2000.

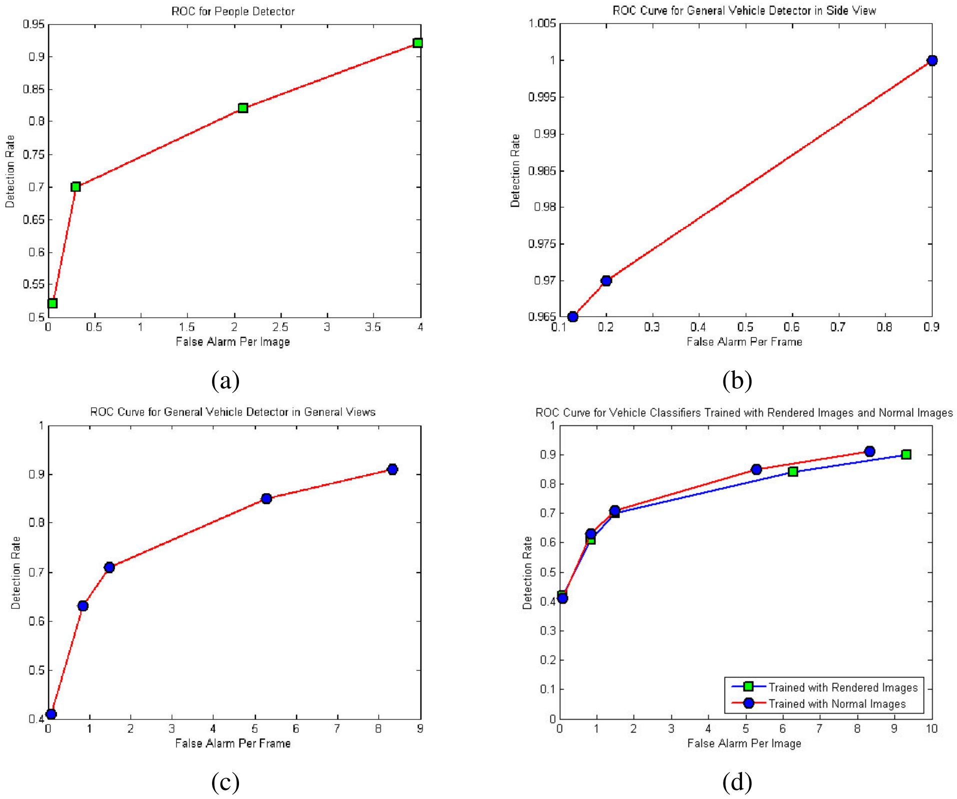
[8] P. Viola and M. Jones. ‘Rapid object detection using a boosted cascade of simple features”. *In Proc. CVPR*, pages 511-518, 2001.

[9] P. Viola, M. Jones, and D. Snow. ‘Detecting pedestrians using patterns of motion and appearance”. *In Proc. Interational Conference on Computer Vision*, volume 2, pages 734-741, 2003.

[10] M. Weber, M. Welling, and P. Perona. ‘Unsupervised learning of models for recognition”. *In Proc. ECCV*, pages 18-32, 2000.

[11] S. Agarwal, A. Awan, and D. Roth. ‘Learning to detect objects in images via a sparse, part-based representation”. *IEEE PAMI*, 26(11):1475-1490, Nov. 2004.

[12] S. Agarwal and D. Roth. ‘Learning a sparse representation for object detection”. *In Proc. ECCV*, volume 4, pages 113-130, 2002.



(a)

(b)

(c)

(d)

Fig. 10. (a) ROC curve for people detection without using stereo cue to generate focus of attention. (b), (c) ROC curves for vehicle detection on UIUC dataset and select dataset from PSACAL database respectively without using stereo cue to generate focus of attention. (d) ROC curves for vehicle detection on selected dataset from PASCAL database with two classifiers using normal vehicle images and rendered vehicle images as training data respectively.

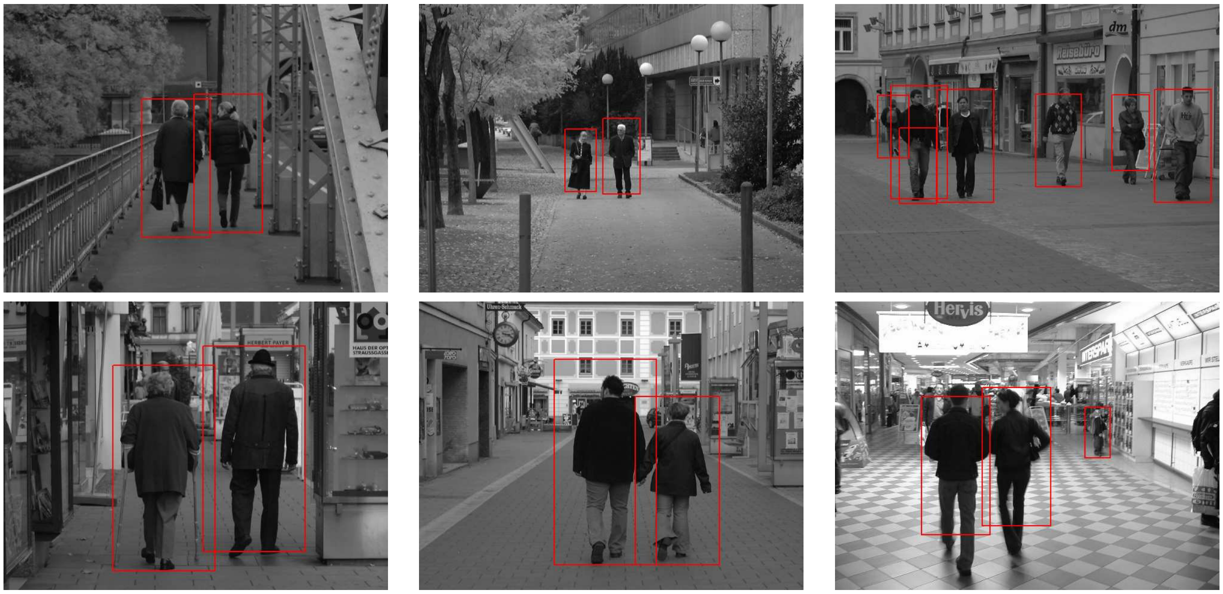


Fig. 11. People Detection Results Without focus of attention Stage

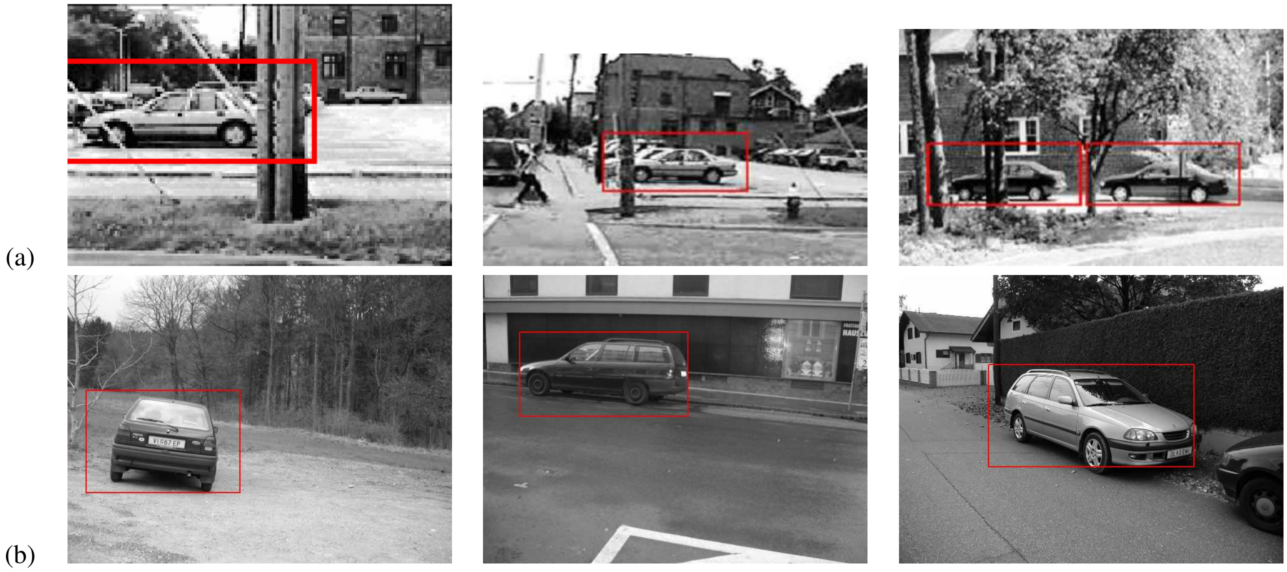
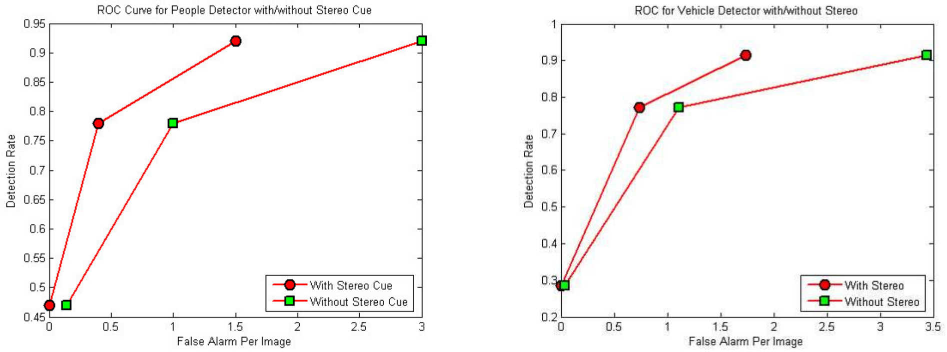


Fig. 12. Vehicle Detection Results Without focus of attention Stage. (a) Results on UIUC dataset. (b) Results on selected dataset from PASCAL database.



(a) (b)

Fig. 13. (a) ROC curve for people detection with and without using stereo cue to generate focus of attention. (b) ROC curve for vehicle detection with and without using stereo cue to generate focus of attention.

[13] B. Leibe, E. Seemann, and B. Schiele. "Pedestrian detection in crowded scenes". *In CVPR*, pages 878-885, 2005

[14] B. Leibe, A. Leonardis, and B. Schiele. "Combined object categorization and segmentation with an implicit shape model". *In ECCV'04 Works. on Stats Learning in Comp. Vision*, pages 17-32, May 2004.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection". *In Proc. CVPR*, volume 1, pages 886-893, 2005.

[16] D. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, 60(2):91-110, Nov. 2004.

[17] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts". *IEEE PAMI*, 24(4):509-522, 1998.

[18] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition". *Intl. Workshop on Automatic Faceand Gesture Recognition*, IEEE Computer Society, Zurich, Switzerland, pages 296-301, June 1995.

[19] W. T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma, "Computer vision for computer games". *2nd International Conference on Automatic Face and Gesture Recognition*, Killington, VT, USA, pages 100-105, October 1996.

[20] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: an application to face detection," *in Proceedings of the IEEE Conference on Computer Vision and Patter Recognition*, 1997, pp. 130-136.

[21] V. Vapnik, "The nature of statistical learning theory". *New York: Springer-Verlag*, 1995.

[22] B. Heisele, T. Serre, S. Prentice, and T. Poggio. "Hierarchical classification and feature reduction for fast face detection with support vector machines". *Pattern Recognition*, 36(9):20072017, Sep 2003.

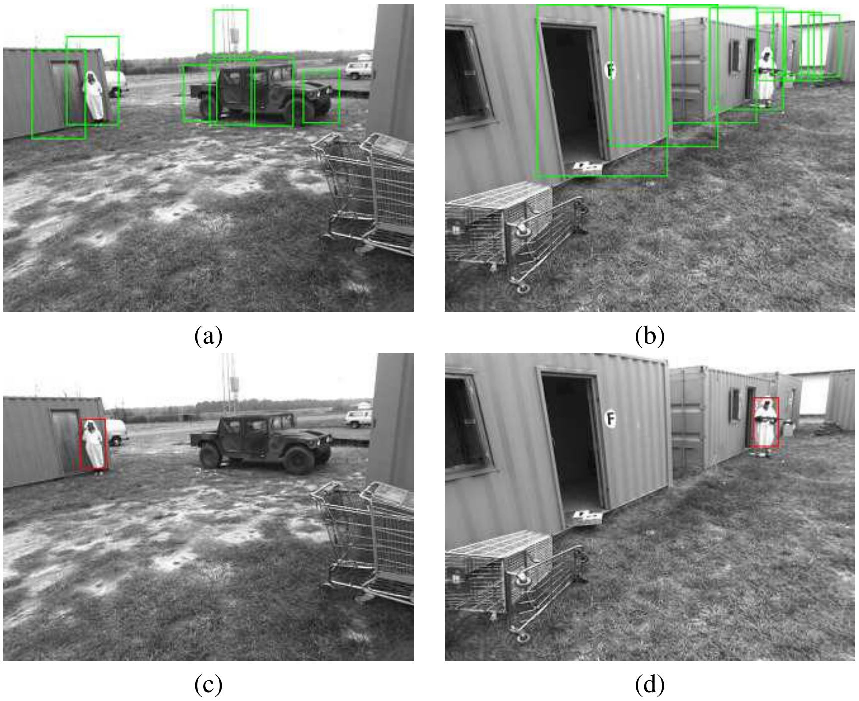
[23] S. Romdhani, P. H. S. Torr, B. Schlkopf, and A. Blake. "Computationally efficient face detection". *In Proc. ICCV*, volume 1, pages 695700, Jul 2001.

[24] K. K. Sung and T. Poggio, "Example-based learning for view-based human face detection", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 39-51, 1998.

[25] M. Everingham, A. Zisserman, C. Williams, L. Van Gool, M. Allan, C. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorko, S. Duner, J. Eichhorn, J. Farquhar, M. Fritz, C. Garcia, T. Griths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A. Storkey, S. Szedmak, B. Triggs, I. Ulusoy, V. Viitaniemi, and J. Zhang. "The 2005 pascal visual object classes challenge". *In In Selected Proceedings of the First PASCAL Challenges Workshop*, LNAI, Springer-Verlag.

[26] http://www.pascal network.org/challenges/VOC/databases.html.

[27] Y. Shan, F. Han, H. S. Sawhney, and Rakesh Kumar, "Learning Exemplar-Based Categorization for the Detection of Multi-View Multi-Pose Objects", *IEEE Conference on Computer Vision and Patter Recognition*, NYC, 2006.



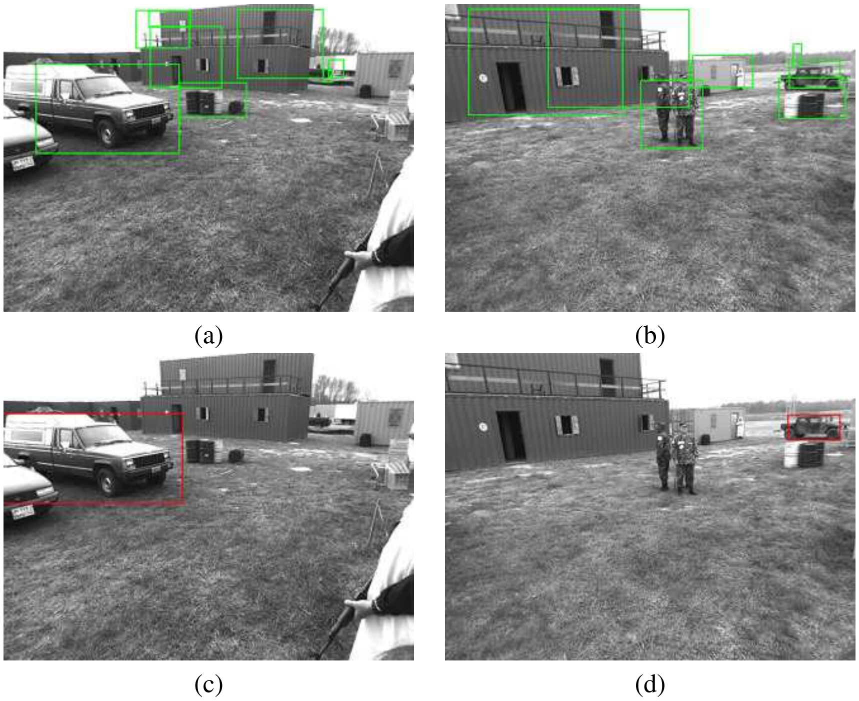
(a)

(b)

(c)

(d)

Fig. 14. (a),(b) Focus of attention for people by stereo cue. (c),(d) Final People Detection Results.



(a)

(b)

(c)

(d)

Fig. 15. (a), (b) Focus of attention for vehicle by stereo cue. (c), (d) Final Vehicle Detection Results.