Using Whole and Part-based HOG Filters in Succession to Detect Cars in Aerial Images

Satish Madhogaria[1], Marek Schikora[1,2] and Wolfgang Koch[1]

[1]Dept. Sensor Data and Information Fusion, Fraunhofer FKIE, Wachtberg, Germany

[2]Department of Computer Science, Technical University of Munich, Munich, Germany

Keywords: Car Detection, Image Analysis, HOG, SVM, LSVM, Part Models, Aerial Images.

Abstract: Vehicle detection in aerial images plays a key role in surveillance, transportation control and traffic monitoring. It forms an important aspect in the deployment of autonomous Unmanned Aerial System (UAS) in rescue and surveillance missions. In this paper, we propose a two-stage algorithm for efficient detection of cars in aerial images. We discuss how sophisticated detection technique may not give the best result when applied to large scale images with complicated backgrounds. We use a relaxed version of HOG (Histogram of Oriented Gradients) and SVM (Support Vector Machine) to extract hypothesis windows in the first stage. The second stage is based on discriminatively trained part-based models. We create a richer model to be used for detection from the hypothesis windows by detecting and locating parts in the root object. Using a two-stage detection procedure not only improves the accuracy of the overall detection but also helps us take complete advantage of the accuracy of sophisticated algorithms ruling out it's incompetence in real scenarios. We analyze the results obtained from Google Earth dataset and also the images taken from a camera mounted beneath a flying aircraft. With our approach we could achieve a recall rate of 90% with a precision of 94%.

1 INTRODUCTION

In this paper we address the task of solving object detection in large-scale aerial images. When we talk about large-scale aerial images, car detection could be termed as one of the most challenging task as car appear very small in large images and vary greatly in shapes and sizes. Besides, the appearance of the object within the observed scene changes quite often depending on the flight altitude and camera orientation. Given the complexity of the problem and the scope for improving the accuracy of detection makes it an important topic of research. This work is inspired from the fact that although the problem of aerial car detection is attempted number of times, still, there is much scope for improving the accuracy and efficiency of the task. Various approaches have been proposed for vehicle detection in aerial images like that of neural network-based hierarchical model for detection in (Ruskone et al., 1996), use of gradient features to create a generic model and Bayesian network for classification as shown in (Zhao and Nevatia, 2001), feature extraction comprising of geometric and radiometric features and detection using top-down matching approach shown in (Hinz, 2003; Nguyen et al., 2007). (Han et al., 2006) proposed a two-stage method to detect people and vehicles by using HOG+SVM as the final verfication stage. HOG-based features (Dalal and Triggs, 2005) have consistently outperformed in various object detection tasks, however, it has its limitation when it comes to small objects like that of cars in aerial images because many details of the cars are not always visible. There are attempts to combine hog features with several other feature extraction technique for performance improvement. The most recent work is the one shown in (Kembhavi et al., 2011), where the authors combine HOG with *Color probability Maps* and *Pairs of pixels* to form a high-dimensional feature set and shows good result. Comparisons of results can easily prove that the performance of the proposed method improves.

The main aim of this work is to build an effective system which can distinguish cars from the background in aerial images with high accuracy. We propose a two-stage method for detecting vehicles in large-scale aerial images. We show that using the standard HOG filters (Dalal and Triggs, 2005) in two steps, one for the root object detection and another for parts detection (Felzenszwalb et al., 2008), in series can greatly improve the detection accuracy. First,

Madhogaria S., Schikora M. and Koch W..
Using Whole and Part-based HOG Filters in Succession to Detect Cars in Aerial Images.
DOI: 10.5220/0004297406810686
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 681-686
ISBN: 978-989-8565-47-1

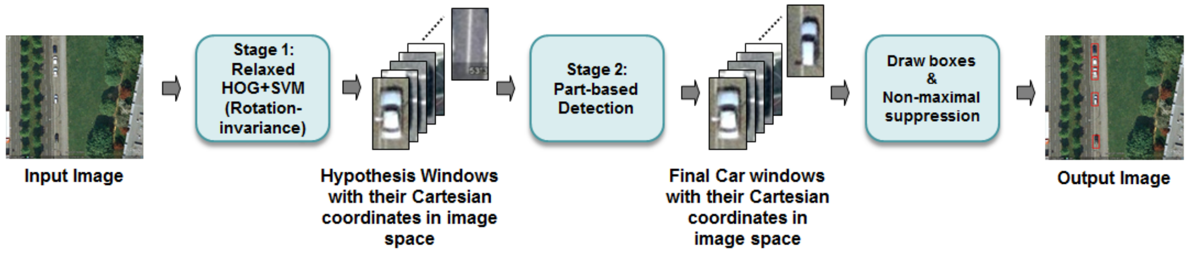

Input Image
Stage 1: Relaxed HOG+SVM (Rotation-invariance)
Hypothesis Windows with their Cartesian coordinates in image space
Stage 2: Part-based Detection
Final Car windows with their Cartesian coordinates in image space
Draw boxes & Non-maximal suppression
Output Image

Figure 1: Proposed vehicle detection method.

we apply the HOG filter to extract hypothesis windows followed by part-based filters on each of the hypothesis windows to detect parts at twice the resolution of original image. Although part-based models have high accuracy rate, it is often avoided in big images because of efficiency issues. In addition, when such sophisticated models are applied to large-scale images with multiple small objects, we show that it misses out objects (see Figure 3). However, when selected windows are given as an input to the part-based model, it can give an impressive performance. In our approach, the second stage is strongly constrained by specific knowledge and the first stage is more general and less constrained.

We test our method using two different data sets. First, we create a library of training and test images from Google Earth. Our second set of testing images consists of high resolution camera images taken from a camera mounted on an aircraft. With our approach, we could achieve a detection rate of more than 90% with a precision of 94%. We also show that our approach achieves higher accuracy when compared to each step applied individually to the test images.

In the next section (2), we describe the methods adapted for use in vehicle detection task and follow it up with the performance analysis and results (Section 3) and conclusion (Section 4)

2 SYSTEM OVERVIEW

Figure 1 shows how the two steps work in series. The overall system is based on HOG filters. In the first stage, a relaxed version of HOG+SVM method is used to generate hypothesis windows. We make several deviations from the standard HOG+SVM (Dalal and Triggs, 2005) in order to have negligible or a very low miss rate. The hypothesis windows are generated at multiple orientations. These subwindows and the cartesian coordinates in image space serves as an input to the second stage. The second stage is highly constrained by using part-based filters to verify the presence of object parts in the hypothesis windows. The part-based filters are applied at double the resolution at which single HOG filters are applied. The part filters give an overall score to each of the hypothesis window and a decision whether it contains a car is made by thresholding the score. Finally, the non-maximal suppression method is used to remove the overlapping windows.

2.1 Relaxed HOG+SVM Detection

To create a less constrained model, we use the "histogram of orientation gradient" feature descriptors (Dalal and Triggs, 2005) to extract features that can resemble a car. Since the time HOG features are introduced to detect people, there has been constant modifications to the standard HOG in order to improve the detection of people as well as other objects in images (Wang and Zhang, 2008; Monzo et al., 2011; Meng et al., 2012). In this work, as we are interested in using HOG features and a Linear SVM classifier (Cortes and Vapnik, 1995; Chang and Lin, 2001) to extract hypothesis windows. HOG features count the occurrences of gradient orientations within overlapping rectangular blocks in the search window. HOG filters are rectangular templates defining weights for features. Let $\mathbf{x}$ be an image subwindow and $\Theta(\mathbf{x})$ denote its extracted feature. $\mathbf{x}$ is labeled as a "hypothesis window", if

$$f(\mathbf{x}) > 0, \quad f(\mathbf{x}) = \mathbf{w} \cdot \Theta(\mathbf{x}) \qquad (1)$$

where $\mathbf{w}$ is the filter. Here, $\mathbf{w}$ is obtained from the linear SVM training of positive and negative training samples.

To Achieve High Detection Rate in First Stage:

1. In the detection step, we keep the window strides low so as to have as many detection as possible around the same object. This in turn increases the probability of detection of the object in the second stage.
2. The threshold, defined by the distance between the feature and SVM classifying plane, is kept lower than usual to improve the detection rate. We conduct several initial experiments with different

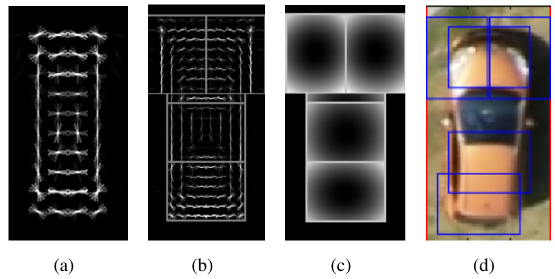


Figure 2: (a) shows the HOG filter. (b) Shows higher resolution part filters and (c) shows the deformation model which defines the cost of placing part filters inside the root filter. (d) shows the parts located in a hypothesis windows.

threshold values and choose the one which can detect nearly all the cars. Despite the fact that it results in high number of false detection, the overall performance of the detector is least affected because of the highly accurate second stage of our algorithm.

3. We do not suppress the overlapping windows in the first stage since the detection rate improves slightly when the second classifier is given multiple windows around the same object.

Besides, the window size also plays an important role in improving the performance of the detector. We chose 48x96 size window to represent the object. Experiments show that having a size smaller than 48x96 reduces the detection. To detect objects at multiple scales, the given input image is upscaled or downscaled depending on the altitude.

Rotation-invariance Detection in the First Stage. As the first stage is relatively faster because of parallel implementation, during the first stage, we detect cars at all possible orientations. Since the HOG fe.atures provide slight invariance in rotation (depending on the number of orientation bins), instead of smaller angle we rotate the image in steps of 30° each time up to 150°. The detected window coordinates are rotated and translated back with respect to the input image. Saved window coordinates and the patches representing subwindows in the input image serves as an input to the second stage. In our first stage, since the image is rotated 6 times and then precessed to detect cars at each rotation angle, it adds to the overall time taken to evaluate the image. Currently, it takes less than 1 second on a 2.8 GHz intel processor with NVIDIA GeForce GT 430 graphics card to extract hypothesis windows at all rotations from a 1000x1000 image.

2.2 Part-based Detection

We now build a model which is strongly constrained by part locations in the whole object. For this purpose we adapt a sophisticated approach described in (Felzenszwalb et al., 2010) to use as the second stage detection model. Part-based models are built on the pictorial structural framework, first introduced in (Fischler and Elschlager, 1973). The main concept introduced in (Felzenszwalb et al., 2008) was that of ”Latent SVM”, which enables the use of part positions as latent variables. The latent SVM formulation of Equation (1) would be:

$$f_{\beta}(\mathbf{x}) = \max_{z\in Z(\mathbf{x})} \beta\cdot\Theta(\mathbf{x},z) \qquad (2)$$

where β is the concatenation of whole filter, part filters and deformation cost weights, z are latent values, in this case part placements and $\Theta(\mathbf{x},z)$ is the concatenation of subwindows and part deformation features. Part filters are defined at double the resolution of root filter which means that they represent finer edges compared to the root filter. The model for an object with n parts is defined by a root filter and a set of part models $(P_1,...,P_n)$. To make a decision on whether the hypothesis window contains car or not, we score the window according to the best possible placement of the parts and threshold this score. A placement of a model in HOG feature space is defined by $z = (p_0,...,p_1)$, where p_0 is the location of root filter and $p_1,..p_n$ are the location of part filters. The score of placement z is expressed by Equation (2). For further details about how the model is trained using latent variables we recommend reading (Felzenszwalb et al., 2008).

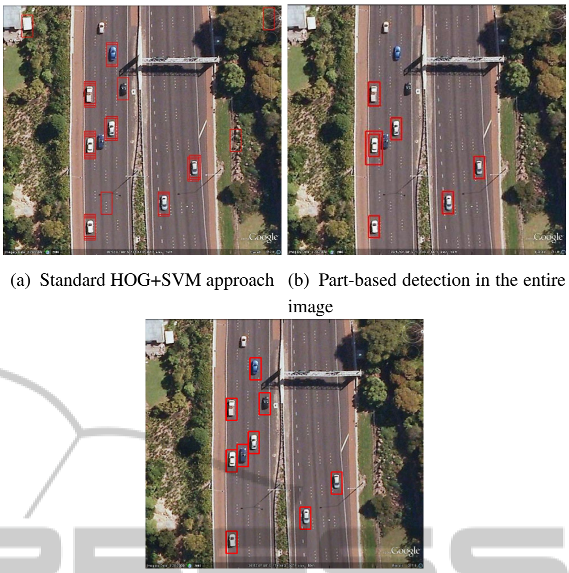
Improving the Accuracy of the Part-based Detector:

1. When given a small search area, in this case ”hypothesis windows” the object detector automatically becomes more efficient, given the fact that the detection need not be done at multiple scales and rotations. In this case we, have fixed size windows on which parts are located using the part filters and a confidence value is generated based on the location of parts in the whole object.
2. We use 6 part filters as it shows slight improvement in the detection rate in comparison to 4 or 5 parts.
3. Since the part models are used as the final deciding model, we could increase the threshold (the distance between the classifying plane and the feature vector) slightly to be able to reduce false alarms keeping the recall rate constant, thereby having greater precision in overall detection.

Apart from improving the detection rate (see Table 1), there are several advantages of using the two-stage approach: First, for effective detection in a sliding window approach, part-based decision model must be applied at all positions (orientations, if we want to have rotation-invariance detections). Considering only the positions, the decision model would have to make decisions for more than 900,000 windows for a 1000x1000 image. In the current scenario, these models are not fast enough to be used for such large images. With our approach, we generate hypothesis windows using the parallel implementation of HOG+SVM. The number of windows, given as an input to the second stage is reduced to a few hundreds as against close to a million if we were to evaluate directly with part-based detection method. Second, the rotation-invariance and the scale factor is taken care of in the much faster stage 1 of our algorithm. Therefore, in the second stage, the need of evaluation at multiple scales and orientations is averted which, therefore, makes it more efficient apart from being highly accurate.

2.3 Using Two Detectors in Sequence

In many cases, we have seen that a number of weak classifiers are used in series and the decision is passed from left-to-right. Normally, different sets of training samples are used in order to generate weak classifiers and the combination of weak classifiers gives the final decision. However, in this case, we use two strong classifiers using the same set of training samples. To use two classifiers in series, we should try not to miss objects in the first stage, which is why, we relax the detection parameters of the first stage. Given the range of our test images, we deduce an optimal threshold for detection, by which we make sure that the minimal number of cars are missed. In Figure 5, we show one example where we reduce the threshold value (from (a) to (c)), so that all cars are detected. This however, generates many false windows. Altogether, we call them ”hypothesis windows”. Depending on the size and complexity of the image, number of such windows can be anywhere between 50 and 500 (note that the rotation-invariance detection increases this number considerably). In this example, we show that with a threshold of 0.8, all the cars are detected. Likewise, we use the same threshold value for evaluating all our test images. Also, using the hypothesis windows from first stage allows us to increase the threshold of the part-based detector to reduce the false alarms in the second stage. We also compare the results obtained separately from standard HOG+SVM classifier (Dalal and Triggs, 2005), part-based classifier (Felzenszwalb et al., 2010) and our approach (see Figure 3 and Table 1). For comparison, we evaluate the images at fixed orientation as Felzenszwalb’s part-based model is not rotation-invariant. The Table 1 shows that our approach outperforms the part-based detection method by 9% and the HOG+SVM method by about 26%. This clearly proves the superiority of using a whole and part filter in succession as against the part-based detection alone in a large-scale image.



(a) Standard HOG+SVM approach (b) Part-based detection in the entire image

(c) Our approach

Figure 3: **Example comparing 3 different approaches -** We see that the sophisticated approach such as HOG part-based models, when applied to a large image misses objects. However, with our approach where we give hypothesis windows as an input to the part-based approach, the detection is improved to a great extent.

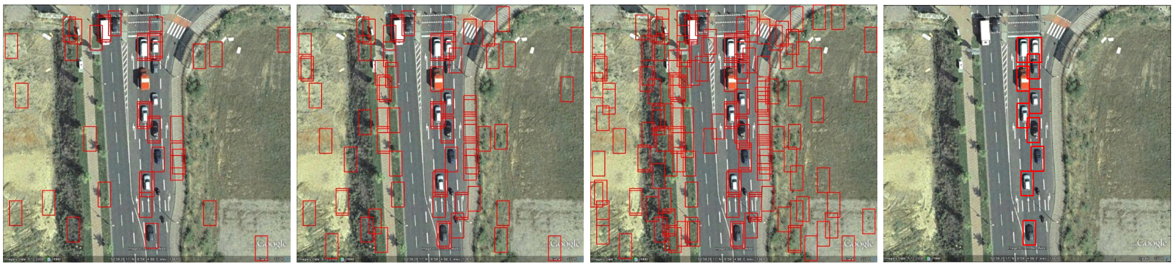
Table 1: Performance Comparison.

	Dalal & Triggs	Felzenszwalb et al.	Our approach
No of images processed (Fixed orientation)	32	32	32
No of cars present	240	240	240
Detection Rate	65.1%	82.2%	91.1%
False Alarm Rate	42%	5.2%	6%

Shows overall comparison of 3 methods in terms of ”detection rate” and ”false alarm rate”.

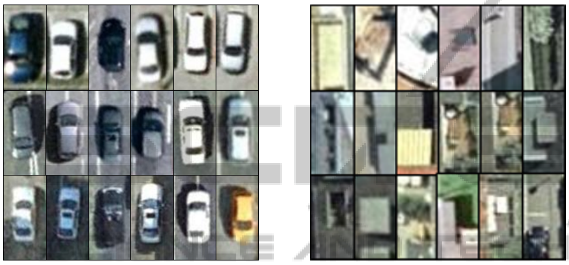
3 RESULTS

We verify the performance of our method using the images taken from Google Earth. The data set consists of 35 images with varying urban background



(a) Threshold = 1.2 (b) Threshold = 1.0 (c) Threshold = 0.8 (d) Part-based detection applied on (c)

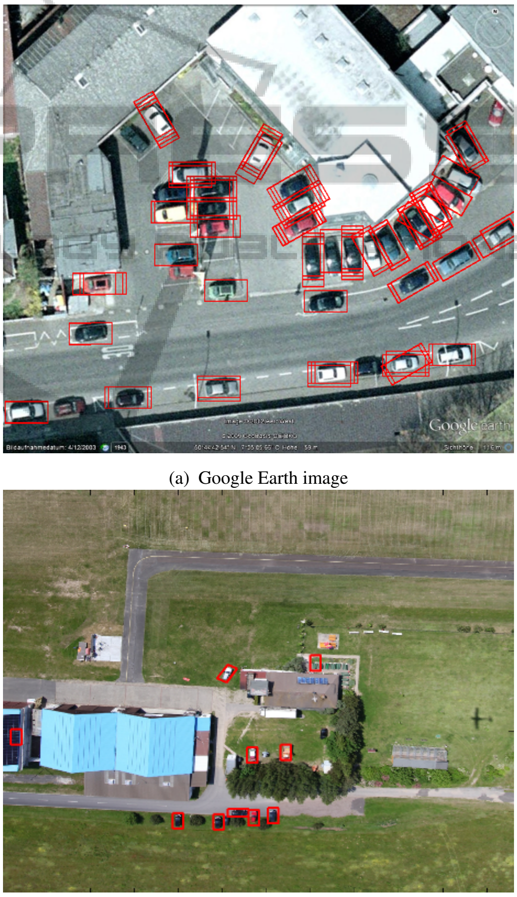
Figure 5: The first stage is designed is such a way that it detects nearly all the cars in our test data set. From (a) to (c), we can see that lowering the threshold results in all the cars being detected. We call these detections as the hypothesis windows which are given as an input to the part-based detection method. (d) shows the final result of the classifier.



(a) Positive Samples (b) Negative Samples

Figure 4: Training data samples - ©Google 2011.

and multiple cars present in each image with image size ranging from 700x700 to 1200x1200 (approximately). Training data (Figure 4) consists of about 200 cars and 600 non-car images. In these experiments, we have kept the window size fixed to 48x96 because the size of cars is more or less within a confined window size for a given altitude. For varying altitudes, the input image should be upscaled or downscaled depending on the height at which the image is taken. In Figure 6, we see sample results, each from Google Earth and an image from flight experiment. In the first stage, the hypothesis windows are generated at multiple orientations. Each of these subwindows is validated by the part-based models in the second stage. Figure 8 displays few more results obtained from our approach. The performance of our system is analyzed by means of the precision-recall curve shown in Figure 7. We see that the precision rate and the recall rate remains above 85% for all our test images. Table 2 gives a clearer picture of the overall performance. With this method, we could detect 90% of the total cars with a precision of 94%. It is worth mentioning that out of 374 total cars present in the test data set, 21 were missed in the first stage itself because of occlusion or shadows, which means that the actual recall rate of the second stage stands at 95%.



(a) Google Earth image

(b) Image from flight experiment

Figure 6: Shows sample results from our two-stage approach. The rotation-invariant method is able to detect cars at all orientation with high recall rate and good precision.

4 CONCLUSIONS

We presented a two-stage approach to detect cars in aerial images. Instead of choosing several classifiers in series (which is a more usual practice), we select

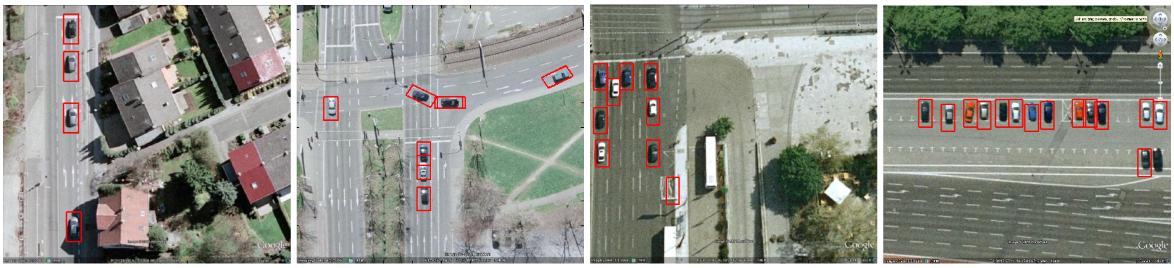


Figure 8: Shows some more results from Google Earth images.

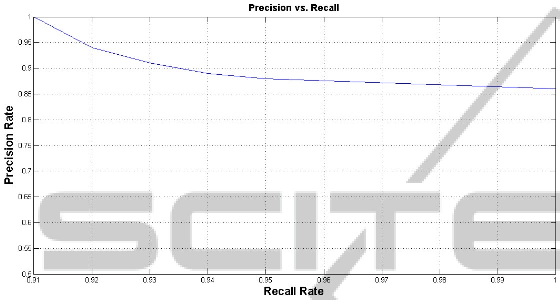


Figure 7: Illustrates the performance of the two-stage algorithm on Google Earth data set.

Table 2: Performance of our two-stage approach

No. of images processed	No. of cars	Overall RR	Overall PR
36	374	90%	94%

The overall recall and precision rate gives a clearer picture of an impressive performance obtained through our approach.

two strong classifiers one after the other. In the process, we improve the detection rate of the first classifier in order not to miss objects in the first stage and improve the precision of the second classifier. Hence, we were able to achieve a high recall rate and with very high precision rate. We have achieved very good results in terms of accuracy, however, to make it a robust system, more work in this direction is required. Knowing that the proposed system performs well, we would be interested in a faster implementation of sophisticated approach such as part-based detection methods so that we are able to detect objects in large images in real time. Besides, we expect to develop a more efficient rotation-invariance scheme to be used in the first stage.

REFERENCES

Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*.

Cortes, C. and Vapnik, V. (1995). Support vector networks. In *Machine Learning*, volume 20, pages 273–297.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893.

Felzenszwalb, P., McAllester, D., and Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9).

Fischler, M. A. and Elschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Trans. Comput.*, 22.

Han, F., Shan, Y., Cekander, R., Sawhney, H., and Kumar, R. (2006). A two-stage approach to people and vehicle detection with HOG-based SVM. In *The 2006 Performance Metrics for Intelligent Systems Workshop*.

Hinz, S. (2003). Detection and counting of cars in aerial images. In *International Conference on Image Processing*.

Kembhavi, A., Harwood, D., and Davis, L. (2011). Vehicle detection using partial least squares. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(6):1250 –1265.

Meng, X., Lin, J., and Ding, Y. (2012). An extended HOG model: SCHOG for human hand detection. In *Systems and Informatics (ICSAI), 2012 International Conference on*.

Monzo, D., Albiol, A., Albiol, A., and Mossi, J. (2011). Color HOG-EBGM for face recognition. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*.

Nguyen, T., Grabner, H., Bischof, H., and Gruber, B. (2007). On-line boosting for car detection from aerial images. In *International Conference on Research, Innovation and Vision for the Future*.

Ruskone, R., Guigues, L., Airault, S., and Jamet, O. (1996). Vehicle detection on aerial images: A structural approach. In *International Conference on Pattern Recognition*, pages 900–904.

Wang, Q. J. and Zhang, R. B. (2008). LPP-HOG: A new local image descriptor for fast human detection. In *Knowledge Acquisition and Modeling Workshop, 2008. KAM Workshop 2008. IEEE International Symposium on*.

Zhao, T. and Nevatia, R. (2001). Car detection in low resolution aerial image. In *International Conference on Computer Vision*.