Detection of moving targets from a moving ground platform

Thomas B. Sebastian[a], Christopher M. Wynnyk[a], Peter H. Tu[a], and Sabrina B. Barnes[b]

[a]GE Global Research, One Research Circle, Niskayuna NY, USA;
[b]Lockheed Martin Missiles and Fire Control, Dallas TX, USA

ABSTRACT

Semi-autonomous operation of intelligent vehicles may require that such platforms maintain a basic situational awareness with respect to people, other vehicles and their intent. These vehicles should be able to operate safely among people and other vehicles, and be able to perceive threats and respond accordingly. A key requirement is the ability to detect people and vehicles from a moving platform. We have developed one such algorithm using video cameras mounted on the vehicle. Our person detection algorithms model the shape and appearance of the person instead of modeling the background. This algorithm uses histogram of oriented gradients (HOG), which model shape and appearance using image edge histograms. These HOG descriptors are computed on an exhaustive set of image windows, which are then classified as person/non-person using a support vector machine classifier. The image windows are computed using camera calibration, which provides approximate size of people with respect to their location in the imagery. The algorithm is flexible and has been trained for different domains such as urban, rural and wooded scenes. We have designed a sensor platform that can be mounted on a moving vehicle to collect video data of pedestrians. Using manually annotated ground-truth data we have evaluated the person detection algorithm in terms of true positive and false positive rates. This paper provides a detailed overview of the algorithm, describes the experiments conducted and reports on algorithmic performance.

Keywords: Person detection, moving platforms, machine learning, camera calibration

1. INTRODUCTION

Intelligent vehicles that need to operate in an autonomous or semi-autonomous fashion have to maintain a basic situational awareness with respect to people, other vehicles and their intent. Further, the safe operation of these vehicles require that they have the ability to detect and track people and vehicles. We have developed one such algorithm for detecting dynamic objects using video cameras mounted on the vehicle. In this paper we present how this algorithm is used for detecting people from a moving platform and how it is evaluated.

Person detection is a well-studied area of computer vision and pattern recognition. There are several commercial person detection and tracking systems available for fixed camera applications (e.g the IBM S3,[1] or the ObjectVideo OnBoard[2]), which rely on continuously estimating a statistical model for the background. For example, in,[3] Mixture of Gaussians was used to create a statistical model of the background intensities. Pixels with intensities that do not fit the underlying model are classified as foreground pixels. A shape-based post-processing of the foreground map is used to detect people. These systems have been used for applications such as retail loss prevention and for perimeter surveillance. However, these methods are not applicable for cameras mounted on moving platforms since the background is constantly changing.

Person detection approaches for moving platforms must focus on modeling the shape and appearance of people as opposed to the background, and operate on a frame-by-frame basis.[4–6] These person detection approaches typically use a scanning window approach, where several (often several hundreds) of sub-regions of the image are classified as either a person or a non-person. This typically involves two steps: *(i)* feature extraction and *(ii)* classification. A descriptive feature vector is computed for each sub region, and it has to be invariant to changes

Further author information: (Send correspondence to T.B.S.)
T.B.S.: E-mail: sebastia@crd.ge.com.edu, Telephone: 1 518 387 4413
C.M.W.: E-mail: wynnyk@crd.ge.com, Telephone: 1 518 387 5946
P.H.T.: E-mail: tu@crd.ge.com, Telephone: 1 518 387 5838
S.B.B.: sabrina.barnes@lmco.com, Telephone: 1 972 603 3520

(a) 

(b) 

Figure 1. Typical images taken from a car, of a pedestrian in the middle of the street (a) and of pedestrians walking on a sidewalk (b). In a vision-based collision avoidance application, the focus is on pedestrians who appear in the middle of the street, which has a relatively uniform background and is on the same ground plane as that of the car. The method in this paper focuses on people who are on sidewalks and other areas. This leads to unpredictable terrains and complex backgrounds.

in lighting, pose of the individual etc. Once the feature vector is computed a classifier is used to determine whether a person is present or not. This classifier is trained using machine learning techniques and is based on labeled examples of person and non-person images. The machine learning algorithm automatically selects the features, thresholds and weights to form a person/non-person classifier.

The feature extraction step involves the computation of image features such as edges, blobs, histogram of oriented gradients, region covariances etc. A subset of these features are used to form a descriptor for each image sub-region. Region covariances of spatial extent and intensity gradients and orientations have been used to create the descriptor.[7] In this approach, the boosting framework[8] is used to train the classifier. Another approach to person detection based on boosting uses grouped edgelets to create the feature vector.[5] A third approach to boosting-based person detection[6] uses a collection of edge and blob features for creating the descriptor. GE in conjunction with LMC is conducting active research in the area of person detection for moving platforms. The focus of this paper is the experimental analysis of one of these approaches.

Person detection systems have also been developed by the automobile industry to augment their collision avoidance suite. The Honda Intelligent Night Vision System[9] uses infrared cameras to detect oncoming pedestrians who have entered the road. Daimler Chrysler Research's PROTECTOR system[10] uses two cameras and a shape-based algorithm for the same purpose. These systems have focused on detecting pedestrians on the road in front of the vehicle, where the road provides a clean ground-plane, as well as a uniform background. In contrast, our system will focus on detecting people on the sidewalk. Hence, it has to be robust to a variety of terrains and backgrounds. See Figure 1.

The overview of the paper is as follows. In the next section, we review person detection algorithm that is evaluated in this paper. In Section 3 we describe how camera calibration is used for nominating plausible image sub-regions. Section 4 discusses some issues with training a person classifier, and Section 5 describes the camera and mobile platform used in the experiments. Section 6 describes the results of our approach.

2. PERSON DETECTION FROM MOVING PLATFORMS

Person detection and tracking from moving platforms is challenging due to a variety of reasons: (i) variable pose of people, (ii) dynamic backgrounds, (iii) variable vehicle speeds, (iv) variable terrain and (v) partial occlusion. The shape and appearance of people is widely variable due to articulation of limbs and variation in viewing angles. Since the vehicle is moving, the background observed by the system is constantly changing. Person detection approaches for moving platforms typically focus on modeling the shape and appearance of people as opposed to modeling a static background. Further, the system has to function in both urban and rural environments. The terrain observed by the face cataloging system is variable. Another challenge for person detection is that people

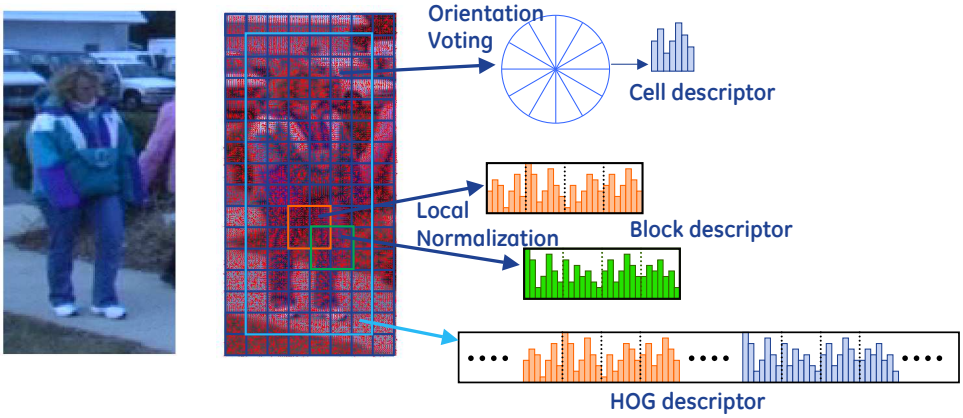


Figure 2. Schematic of HOG descriptor computation.

can be partially occluded by other vehicles and trees, shrubs, garbage cans and miscellaneous obstacles on the side of the road.

In this paper we evaluate the person detection approach based on Histogram of Oriented Gradients (HOG).[4] In this approach the image sub-regions are partitioned into smaller cells and a histogram based on orientations of image gradients is computed. To achieve invariance to lighting and contrast variations, the information from neighboring cells is used to normalize the histogram of each cell. The concatenation of the normalized histograms of all the cells produces the final HOG descriptor. See Figure 2 for a schematic of how the HOG descriptor is computed. These descriptors are classified using Support Vector Machines (SVM)).[11] The SVM classifier is trained using labeled dataset of person and non-person images. SVM-based approaches have been shown to have good generalization properties for unseen data. Traditional SVM training does not support feature selection, so a modified version that uses recursive feature elimination[12] has been developed.

3. MODELING SCENE GEOMETRY

When performing person detection in static images without any prior knowledge of scene geometry, it is typical to assume that people are equally likely to appear at all possible locations, orientations and scales.[4] However, prior knowledge about the scene can significantly reduce this search space. For instance, a terrain model for the scene can be used to adjust the probability of finding a pedestrian at a given image location. For example, it is reasonable to assume that all pedestrians will have their foot locations on the ground. In addition, camera calibration (height of the camera and its orientation with respect to the ground) provides the size of a person at a given image location. The terrain model and the camera calibration information allows us to limit the use of a person detector to specific locations. Further, pedestrian detectors are typically tuned to look for people of certain heights in upright positions. Therefore, this knowledge of the scene makes person detection computationally efficient. It also reduces the false positive rate since potential missed detections are ruled out based on calibration information.

The camera in our system is rigidly mounted on the roof of the vehicle, and operates differently from a static camera system. For a static camera there is no relative motion between the camera and the scene, allowing us to estimate the ground plane. In our case, the camera is moving because of the motion of the vehicle. The motion of the vehicle can be viewed as a simple translation, but a fair amount of nuisance motion due to car vibrations and the presence of bumps and potholes increases the randomness of the motion.

We use an autocalibration algorithm to automatically estimate camera calibration with respect to a ground plane. This is described in detail in.[13,14] This approach utilizes person detections to perform the calibration. The advantage of this method over existing approaches is that it uses a Bayesian analysis of foot and head locations extracted from both person detection and tracking. This approach is able to tackle large amounts of noise and errors in the measurements (person detections), while still being able to obtain accurate calibration estimates. Since our system is mounted on a moving vehicle, and the ground plane assumption may not be valid,

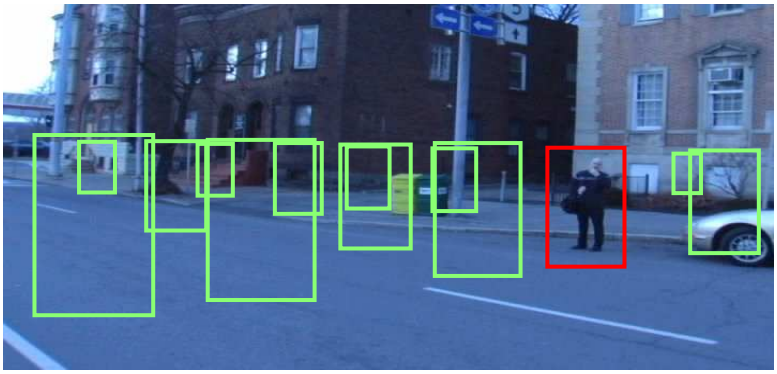


Figure 3. A representative sample of image sub-regions derived from camera calibration, which are classified as person/non-person.

we use sub-regions corresponding to multiple ground planes in our approach. We also use a few scales at each location to obtain robustness to varying heights of people. See Figure 3 to see how sub regions in the image are nominated.

4. TRAINING THE PERSON CLASSIFIER

The approach to person detection presented in this paper is based on a data-driven machine learning approach. By necessity these mechanisms require training. Specifically these algorithms will require sets of positive and negative training examples. The positive examples need to span the appearance space of upright individuals and the negative examples must be representative of "every thing else". For these types of applications 500 or more positive samples are usually sufficient. In order to represent the range of appearance of the non-human space, much larger sets of negative samples are required. One strategy for collecting negative samples is to assemble a large set of diverse images that do not have people in them and then randomly sample these images to produce the negative training set.

Another concern that must be addressed is the issue of over-learning which results in a failure to generalize. Over-learning or over fitting is the problem where by an overly complex classifier can have a very low classification error with respect to the training data but when applied to previously unseen data, performance diminishes dramatically. In general there are two main reasons for over-learning, the classifier is overly complex and there is insufficient training data. A brute force strategy would be to acquire an enormous supply of training data such that the law of large numbers takes effect. Another approach is to use cross-validation, where a set of positive and negative samples are sequestered from the training process and used to evaluate/validate the generalization properties of the trained classifier. One can then vary the dimensionality or complexity of the classifier and choose the complexity level that results in the best performance with respect to the validation data.

Figure 4. Left: Images taken from surveillance cameras suffer from interlacing artifacts and are not suitable for person detection from moving platforms. Right: Images taken from a 1 CCD color camera has poor spatial resolution due to Bayer interpolation.

5. DATA COLLECTION

As part of our research, we conducted experiments with several cameras to select the appropriate one. We report on our analysis on a representative set of cameras in this section. First set of cameras are one that are used for fixed surveillance applications. Some examples of these cameras are GE Security KTC-2000DN and Sony EVI-D70. These cameras output NTSC video format and suffer from interlacing artifacts. Hence, the images from these are not suited for moving platform applications. Second, we tested a progressive scan, color camera with an Gigabit Ethernet output. The images from this camera did not suffer from any interlacing effects, but did have artifacts due to the Bayer interpolation used in creating the color image using a single CCD. Third, we used a progressive scan, 3 CCD camera (Toshiba IK-TF5). This camera retains its native resolution in color mode, as it uses separate CCDs for each color channel. This camera supports fast shutter speeds to minimize the motion blur when data is acquired at vehicle speeds of up to 30mph.

Figure 5. Left: Vibration-damped camera module which is mounted on the moving platform. Right: Camera module mounted on a pickup truck for collecting data in urban streets. This allows for both video capture and real-time evaluation of the person detection algorithm.

The video camera used in the experiments reported in this paper is Toshiba IK-TF5. The camera is mounted to the top of the vehicle using vibration-damping springs to minimize image artifacts due to camera jitter (Figure 5). The camera is connected to an on-board computer to facilitate both real-time processing and also video recording. We have used two vehicles for collecting the data used in this paper: a pickup truck for urban scenes and wooded scenes, and a go-cart for other scenes.

The data that we collected was split into three categories based on environment, which results in different challenges for detecting people: (i) flat terrain, (ii) wooded scenes, and (iii) urban scenes. The flat terrain category is collected in relatively flat regions such as parking lots, where there are few objects that can be confused with people. The people in these examples are volunteers, who are walking around the scene in an arbitrary manner. They come in and out of the field of view of the camera. The occlusion in these scenes are primarily due to parked cars and other vehicles (Figure 6). The data from the wooded scenes is more challenging due to the presence of trees, shrubs and other underbrush (Figure 7). The people in this scene are also volunteers. The urban data was collected in a completely unconstrained manner. The subjects in the video were not coached in any way. The vehicle was driven at the speed of regular traffic and pedestrians on sidewalks were recorded in an opportunistic fashion. The urban scene is the most challenging as people may appear in only a small number for frames due to the vehicle speed. There is also some motion blur. Further, these scenes contain several objects such as garbage cans, lamp posts, tree skirts that tend to be confused with people (Figure 8).

6. PERFORMANCE EVALUATION

Algorithm is evaluated using manually annotated videos. The videos are annotated on a per-frame basis by a person for both the presence and location of people. In addition to per-frame locations of people, the manual annotation groups locations of each person in the video to form person trajectories. The person detection algorithm produces the location of people on each frame, which are then compared to the manual annotations.

Figure 6. Example results from video collected in flat terrain, such as parking lots.

	True Positive Rate (%)	False Positive Rate (per 100 frames)
Flat terrain	98.8	1.4
Rural	88.8	4.3
Urban	66.7	8.3

Table 1. Summary of true positive and false positive rates

.

The detection rate of persons is computed at the person trajectory level. In particular, a person is considered to be detected if there are detections on more than a fixed number of frames (set to 5). The true positive rate is calculated as the ratio of the total number of people detected to total number of people in the videos. The false positive rate is calculated on a per-frame basis, and is the total number of false positives in every 100 frames. These numbers are reported in Table 1.

In order to reduce false positives, temporal grouping of person detections is done by piping the detections through a constant velocity Kalman filter.[15] A detection is not recorded until a certain number of detections have been associated together in the tracked object. This temporal grouping of detections to form tracks reduces the false positive rate, as we eliminate the occasional false detection that do not persist over time. However, there is also a small reduction in the true positive rate. Figure 9 shows the effect of increasing temporal tracking.

ACKNOWLEDGMENTS

Portions of this report were prepared by GE CR&D as an account of work sponsored by Lockheed Martin Corporation. Information contained in this report constitutes confidential technical information which is the property of Lockheed Martin Corporation. Neither GE nor Lockheed Martin Corporation, nor any person acting on behalf of either; "a. Makes any warranty or representation, expressed or implied, with respect to the use

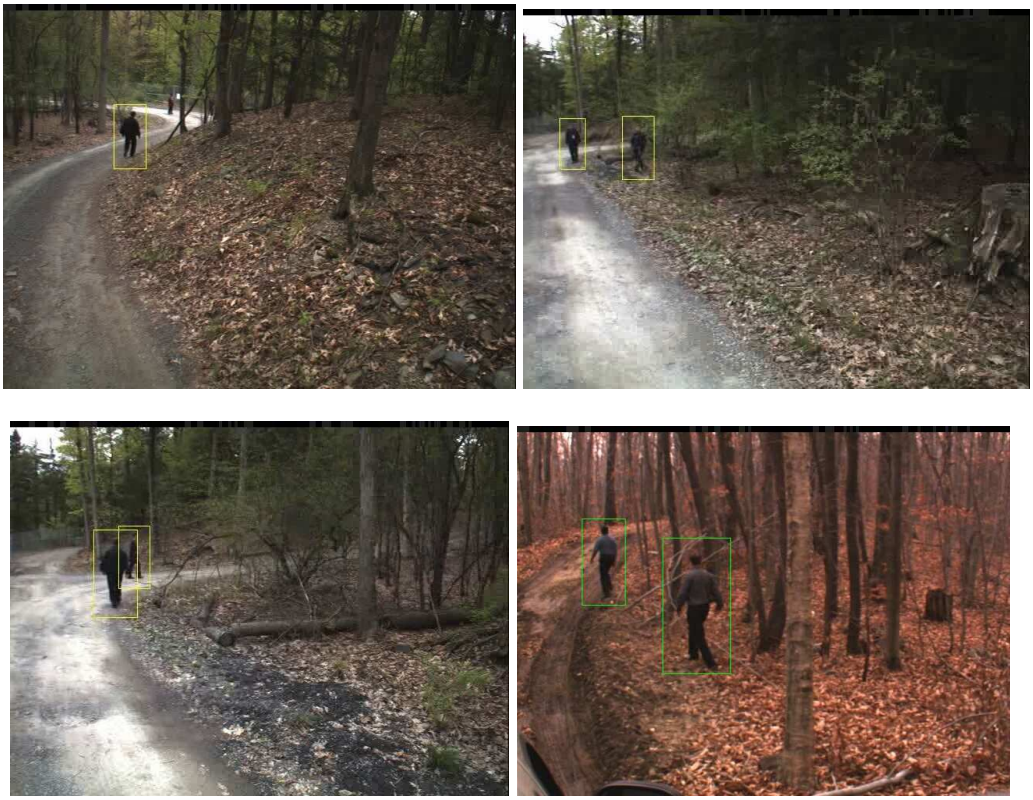


Figure 7. Example results on video collected from rural scenes.

of any information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or b. Assume any liabilities with respect to the use of, or for damages resulting from the use of, any information, apparatus, method, or process disclosed in-this report.

REFERENCES

[1] "IBM smart surveillance system (S3)," http://www.research.ibm.com/peoplevision/.

[2] "Objectvideo onboard," http://www.dbvision.net/products/onboard/.

[3] Stauffer, C. and Grimson, W., "Adaptive background mixture models for real-time tracking," in [*IEEE Computer Vision and Pattern Recognition*], **2**, 246–252 (1998).

[4] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," in [*IEEE Computer Vision and Pattern Recognition*], 886–893 (2005).

[5] Sabzmeydani, P. and Mori, G., "Detecting pedestrians by learning shapelet features," in [*CVPR*], (2007).

[6] Tu, P., Krahstoever, N., and Rittscher, J., "View adaptive detection and distributed site wide tracking," in [*AVSS*], (2007).

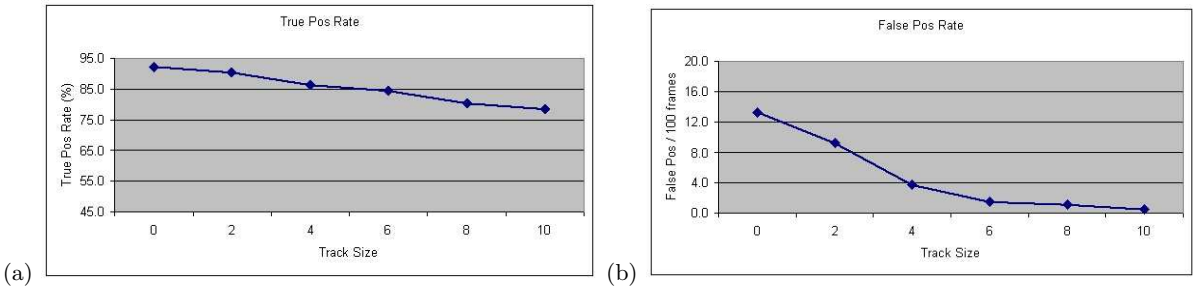
[7] Tuzel, O., Porikli, F., and Meer, P., "Pedestrian detection via classification on riemannian manifolds," in [*IEEE Computer Vision and Pattern Recognition*], (2007).

[8] Freund, Y. and Schapire, R. E., "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences* **55**, 119–139 (1997).

[9] http://world.honda.com/HDTV/IntelligentNightVision/200408/.

[10] Gavrila, D. M., Giebel, J., and Munder, S., "Vision-based pedestrian detection: The PROTECTOR system," in [*Proc. of the IEEE Intelligent Vehicles Symposium*], (2004).

Figure 8. Example results on video collected from urban scenes.



(a) (b)

Figure 9. The effect of temporal tracking is shown. Temporal grouping of detections results in a small reduction in true positive rate (a) and a larger reduction in the false positive rate (b).

[11] Vapnik, V., [*The Nature of Statistical Learning Theory*], Springer-Verlag (1995).

[12] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V., "Gene selection for cancer classification using support vector machines," *Machine Learning* **46**(1-3), 389–422 (2002).

[13] Krahnstoever, N. and Mendonça, P., "Bayesian autocalibration for surveillance," in [*Proc. of IEEE International Conference on Computer Vision (ICCV'05), Beijing, China*], (October 2005).

[14] Krahnstoever, N. and Mendonça, P., "Autocalibration from tracks of walking people.," in [*Proc. British Machine Vision Conference (BMVC), Edinburgh, UK, 4-7 September*], (2006).

[15] Kalman, R. E., "New approach to linear filtering and prediction problems," *Transaction of the ASME Journal of Basic Engineering* , 35–45 (1960).