# Contextual Combination of Appearance and Motion for Intersection Videos with Vehicles and Pedestrians

Mohammad Shokrolah Shirazi and Brendan Morris

University of Nevada, Las Vegas
shirazi@unlv.nevada.edu, brendan.morris@unlv.edu

**Abstract.** Object detection and classification is challenging problem for vision-based intersection monitoring since traditional motion-based techniques work poorly when pedestrians or vehicles stop due to traffic signals. In this work, we present a method for vehicle and pedestrian recognition at intersections that benefits from both motion and appearance cues in video surveillance. Vehicle and pedestrian recognition performance is compared using motion, appearance and combined cues in contextually relevant stop areas to improve recognition. Experimental evaluation shows 5% average improvement for vehicle and pedestrian recognition at two Las Vegas intersections.

## 1 Introduction

An important research effort in Intelligent Transportation Systems (ITS) is the development of automated systems that monitor flow of traffic at intersections. Intersections are interesting targets for vision-based monitoring systems since pedestrian and vehicle behaviors and their interactions can be analyzed. In addition, safety is a major concern at intersections and vision-based intersection monitoring systems could address this by detecting or predicting some situations that might lead to an accident [1,2]. The foundational steps for vision-based traffic safety analysis are object detection, classification and finally tracking.

An appropriate object detection and classification method in a video surveillance system should be able to deal with challenging problems like different environment situations, occlusions and low resolution images [3]. Object detection and classification at intersections adds another challenging problem. At traffic conflict points, long-term stationary traffic participants could merge into background leading to missing targets. Most traditional object recognition techniques designed for surveillance use motion cues which are not appropriate for intersections. The traditional surveillance issues are addressed in [3] where motion is used to segment moving objects. Different feature extraction techniques like HOG and PCA with SVM are applied for object classification. In [4] Differences of Histograms of Oriented Gradients (DHOG) are used for distinguishing people from groups of people and vehicles.

Appearance-based object recognition for still images has not been widely used in video surveillance due to the small size of objects and their low resolution

G. Bebis et al. (Eds.): ISVC 2014, Part I, LNCS 8887, pp. 708–717, 2014.

which limits performance. Zhang et al. [5] used a texture feature, the local binary pattern (LBP) [6], along with an Adaboost classifier for vehicles and pedestrian classification. However, it has been shown that texture is not as strong a feature set as compared to other appearance features [7]. This indicates improvements are possible.

In this paper, object detection and classification is addressed at intersections by contextually defining specific areas that benefits from fusion of both appearance and motion cues. Comprehensive datasets from literature were used to train appearance-based classifiers for pedestrians and vehicles. The new UNLV dataset is introduced to improve the performance of vehicle classification at intersection (surveillance) settings.

The remainder of paper is organized as follows. Section 2 presents the vehicle/pedestrian detection and classification system. Section 3 provides contextual definition of areas for fusing appearance and motion-based techniques. Section 4 discusses the datasets used for classifier training and introduces the UNLV vehicle dataset. Section 5 provides a system evaluation and Section 6 concludes the paper.

## 2 System Overview

A three stage cascaded system is proposed for reliable vehicle/pedestrian detection and classification at intersections as presented in Fig. 1. The main advantage of this system is the use of both motion and appearance cues in a contextually meaningful manner for accurate classification. The addition of appearance to the traditional surveillance processing pipeline is motivated by the following:

1. Although motion is used reliably on highways, it is not consistent at intersections due to traffic signals which force participants to stop temporarily.
2. Pedestrian detection using motion is more difficult than vehicles since they have small size and non rigid body which easily is confused with background clutter. In addition, they tend to be more stationary causing them to be incorporated into a background model.
3. Motion-based techniques are not able to reliably distinguish nearby objects which lead to false detections. At intersections, there is occlusion between pedestrians crossing intersections and vehicles that have been recently stopped and pedestrians often walk together in groups at crosswalks both leading to erroneous large object blobs.

### 2.1 Motion-Based Object Detection

As part of a background subtraction technique, a Gaussian mixture model (GMM) [8] is used to create an adaptive background and detect moving objects at scenes.

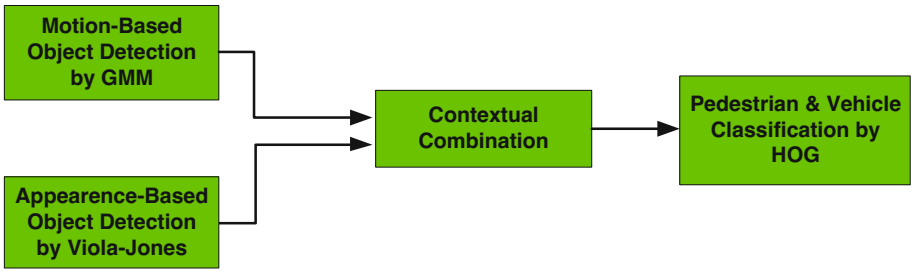



**Fig. 1.** Vehicle & pedestrian detection and classification system

In this method each pixel is modeled by a mixture of $K$ adaptive Gaussian intensity distributions to address lighting changes and slow moving objects. Moving objects (pedestrians or vehicles) are detected as pixels that do not fit any of the $K$ background Gaussian models.

### 2.2 Appearance-Based Object Detection

Haar-like features are rectangular based features that are well known for object detection due to efficient calculation and high detection performance. HAAR like features are used to construct small, efficient and boosted classifiers that can be cascaded to detect almost all objects of interest while rejecting a certain fraction of the non-object patterns in a computationally efficient manner [7].

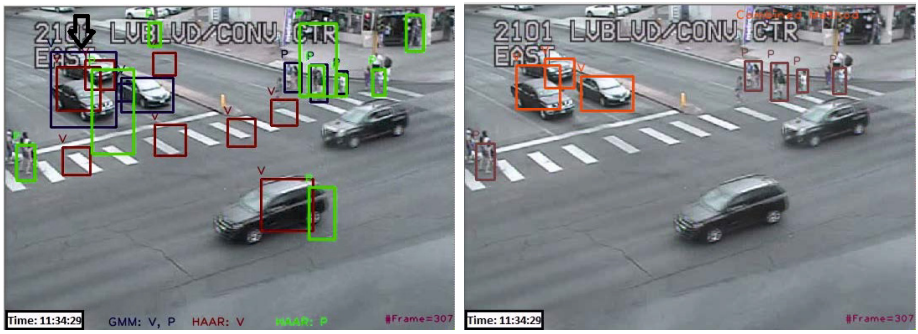
### 2.3 Contextual Combination

The key contribution in the proposed system is the contextual combination or pooling of several positive detection responses. Contextual combination provides fusion at the decision level to combine the outputs from the GMM and Haar detections in mix areas (Fig. 3) where both detectors are active. In this way, appearance detection is limited to smaller processing regions for speed and reliability.

The contextual combination has been defined to be able to:

1. Reject many false appearance-based pedestrian/vehicle detections outside mix areas since detection by motion is more reliable.
2. Perform pooling of detection responses that have overlapping bounding boxes.
3. Select the most reliable detection from either GMM or Haar in each detection pool cluster based on a cumulative score.

The full contextual combination process operates in mix areas where both the GMM and Haar detectors are active. The detections from each are pooled into detection clusters based on bounding box overlap. E.g. all bounding boxes with




(a) (b)

**Fig. 2.** Contextual Combination a) vehicle & pedestrian recognition using each GMM and HAAR, V and P correspond to vehicle and pedestrian respectively, b) Recognition results using contextual combination, vehicles are shown with orange color while pedestrians are shown with brown color, two black vehicles in the motion area are not recognized as result of miss classification

50% overlap are considered part of the same cluster. For each bounding box in a cluster, its cumulative score ($CS$) is computed.

$$CS_d = N \sum_{i=1}^{N} Area_i \tag{1}$$

where $N$ is number of detected objects in a cluster and $Area_i$ indicates bounding box area of $i$th detection from detector $d$ ={GMM, Haar}. The $CS$ is computed separately over each cluster providing a measure of reliability of the detection since only the detector type $d$ that has highest cumulative score is retained for a cluster. The $CS$ is designed to favor smaller appearance detections inside a larger GMM bounding box (case of motion grouping and occlusion).

One example of contextual combination is shown in Fig. 2. A cluster of interest is noted by black arrow and GMM detections are in blue, Haar vehicles in brown, and Haar pedestrians in green. In b) the occlusion merge from GMM is abandoned in favor of the correct detection from multiple Haar boxes leading to higher $CS_{\text{HAAR}}$ than $CS_{\text{GMM}}$.

### 2.4 Pedestrian and Vehicle Classification

Object recognition using HOG features with SVM is quite popular for vehicles and pedestrians [9]. This feature counts the occurrences of gradient orientation computed on a dense grid of uniformly spaced cells to characterize edge-like appearance. Since HOG has high performance for object detection and classification, it is used in this work as well.


(a) (b)

**Fig. 3.** Mix areas a) INT 1 b) INT2. Red indicates areas that vehicles might stop and purple is for pedestrians.

## 3 Intersection Context Definition

Intersections are more challenging monitoring environments due to non-continuous motion characteristics caused by signal phases. Motion detection through background subtraction has poor performance because vehicles and pedestrians are forced to wait until the appropriate phase to cross an intersection resulting in times of inactivity. In order to account for these various phases, contextual mix areas are defined to account for regions in the image where pedestrians or vehicles stop and it is required to use appearance cues for detection.

The GMM is used across the entire scene to account for any visible motion. Mix areas, where Haar appearance detection is also performed, are defined in regions where stopped objects are expected; e.g. the areas before the stop bars, around signals, and in crosswalks. By leveraging appearance, difficult background subtraction occlusion scenarios can be adequately handled. Pedestrian crosswalks are defined as a mix area, even though motion is present, because of the followings:

1. Pedestrians tend to move in (phase directed) groups on a crosswalk. Motion cues result in a single large group-occlusion detection rather than individuals.
2. Recently stopped vehicles at the stop bar are not distinguishable from pedestrians crossing using just background subtraction because of occlusion.

Two examples of mix areas is presented in Fig. 3 showing vehicles areas in red and pedestrians in purple. Pedestrians have a box area showing wait area around the signal.

## 4 Intersection Datasets and Classifier Training

A variety of datasets were needed to effectively train the Haar appearance detector of the first stage and the third stage HOG classifier of the system.

| Dataset | Number of samples |
|---|---|
| Caltech [10] | 946 |
| Graz [11] | 127 |
| MIT [12] | 143 |
| Tripod [13] | 2162 |
| UIUC [14] | 550 |
| VOC [15] | 645 |
| UNLV | 16035 |
| Total | 20608 |

(a)

(b)

| Dataset | Number of samples |
|---|---|
| Daimler [16] | 14401 |
| ETH [17] | 1243 |
| INRIA [9] | 3542 |
| MIT [18] | 924 |
| NICTA [19] | 37344 |
| TUD Brussel [20] | 3272 |
| Total | 60726 |

(c)

(d)

**Fig. 4.** Positive image samples for training detection classifiers a) Number of collected samples from each vehicles dataset b) Typical samples, last row refers to UNLV dataset c) Number of collected samples from each pedestrians dataset d) Typical samples

### 4.1 Datasets

A large image database was constructed from public datasets and a new UNLV image dataset was created for vehicle datasets. The UNLV dataset contains vehicle samples taken from various traffic monitoring cameras at different camera views and orientation overlooking highways and intersections in the Las Vegas Valley. The new dataset was required because:

1. There are only a few publicly available datasets for vehicles and they generally contain high resolution images of a single vehicle with few samples. They do not adequately represent the challenging illumination, clutter, and noise typical in surveillance traffic video.
2. Although intersections require vehicles at different scales and orientations, most existing datasets only consider a few views {side, front, or rear}.

Collecting negative samples was easier since it can be any picture that does not contain a vehicle or a pedestrian. 23985 and 89798 negative samples were collected for the vehicle and pedestrian classifiers respectively.

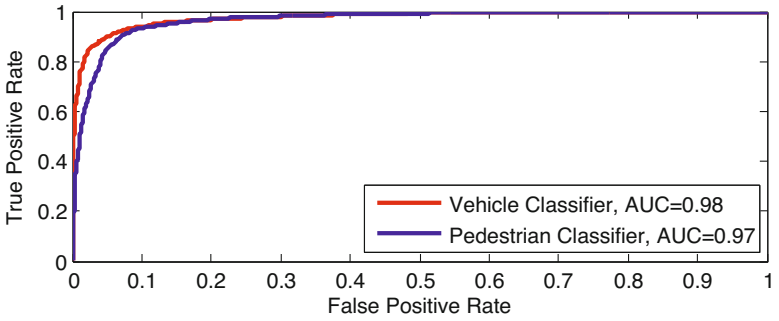



**Fig. 5.** ROC curve for Vehicle & Pedestrian Classifiers

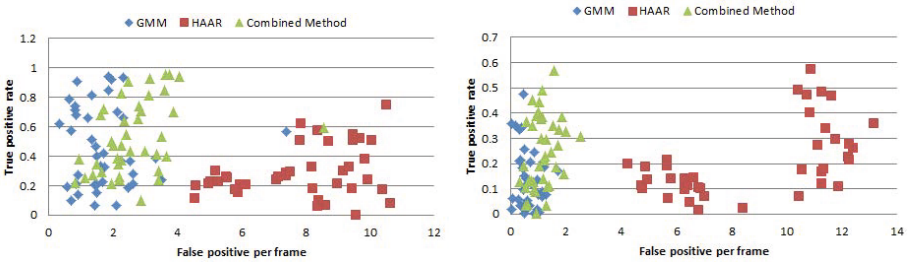



**Fig. 6.** Performance of each GMM, HAAR and Combined method for detection-classification a) Vehicles 1 b) Pedestrians

### 4.2 Classifier Training

**HAAR Like Features with Adaboost Classifier.** The OpenCV implementation of the Haar feature-based cascade classifier was used for pedestrian and vehicle detection. Since same-size positive samples are required, {16,16} and {10,20} were chosen as width and height of vehicles and pedestrians respectively. Given a training set of positive and negative sample images, the Adaboost procedure learns number of weak classifiers which are combined to form a strong classifier.

**HOG Features with SVM Classifier.** HOG features with an SVM classifier, that are used in third stage of the cascaded system, verify detected objects as either vehicles or pedestrians. Using HOG with the SVM classifier as a verification step has some benefits like reducing false positives and system speed up [21].

The HOG classifiers were trained using a linear kernel with LIBSVM [22] to distinguish both vehicles and pedestrians from other objects. To improve the classifier performance, positive vehicle samples were used as negatives during pedestrian training and vice versa. Fig. 5 highlights the HOG performance with

**Table 1.** Vehicle-Pedestrian Detection-Classification Performance During Traffic Phases

| Object | Classifier | Green | | Red | | Total | |
|---|---|---|---|---|---|---|---|
| | | TPR | FPPF | TPR | FPPF | TPR | FPPF |
| Vehicle | GMM | 0.6 | 1.95 | 0.33 | 2.15 | 0.46 | 1.75 |
| | Haar | 0.34 | 8.34 | 0.23 | 7.39 | 0.28 | 7.82 |
| | Combined | 0.64 | 3.06 | 0.42 | 2.70 | 0.52 | 2.69 |
| Pedestrian | GMM | 0.12 | 0.74 | 0.13 | 0.74 | 0.13 | 0.65 |
| | Haar | 0.20 | 8.51 | 0.21 | 8.72 | 0.20 | 8.76 |
| | Combined | 0.24 | 1.13 | 0.23 | 1.25 | 0.24 | 1.15 |

an ROC curve after using a 75-25% training-validation split. It shows that the vehicle classifier performs slightly better, as expected, since there is less deformation in the rigid vehicle body.

## 5 System Evaluation

The vehicle and pedestrian detection-classification system is implemented in C++ using OpenCV 2.3 operating on Intel i7 Quad core with 2GHz clock speed. The performance of each detection-classification method is evaluated for two Las Vegas intersections. Positions of pedestrians and vehicles are manually marked for 1000 frames of each intersection video. GMM, HAAR-like features and Combined methods are separately used at the detection step and each method's performance is evaluated by comparing recognition results with manually annotated text files.

Fig. 6 shows the performance for both intersections. The average of true positive rate (TPR) versus false positives per frame (FPPF) are calculated for each 50 frames leading to 40 points (2000 frames) for each method. Since true negatives lead to large number, false positives per frame is used instead of false positive rate (FPR). As it is shown in Fig. 6 (a), using GMM motion has lower FPPF than appearance methods for vehicles. The contextual combination method provides higher TPR than GMM with only slightly higher FPPF. However, there are still two points from GMM and Combined methods that have large FPPF value (around 8) during drastic lighting change which caused motion noise and resulted in many wrongly detected moving objects. This has direct impact on the Combined method since it uses GMM for all intersection areas. Fig. 6 (b) shows that appearance-based pedestrian detection-classification has higher detection rate than motion at the cost of many false positives. The Combined method has higher true positive rate with a low false positive per frame. It is interesting to note that motion-based techniques work well for detecting vehicles but appearance is required for detecting pedestrians.

Table 1 shows each method's performance for different traffic signal phases. The Total column shows the average over all frames regardless of traffic signals. The Table shows again that the GMM outperforms Haar for vehicles. The Combined method outperforms GMM since there is 6% improvement in TPR with increase of less than 1 in FPPF value. The Combined method has around 9% higher TPR value than GMM during the red signal phase when motion cues are ineffective. However, Haar-based pedestrian detection-classification outperforms GMM in TPR value and the Combined method has higher TPR value than Haar for all signal phases. FPPF for the Combined method is slightly higher than GMM and less than Haar.

The performance results imply some interesting points. Motion-based techniques work well for detecting vehicles but Haar-like appearance features surprisingly are inefficient. The Combined classifier leverages times when motion is an ineffective cue during the red phase for 9% increase in vehicle detection rate. However, the Haar detector significantly outperforms GMM motion for pedestrians. This is because pedestrians are small and tend to remain still on sidewalks. The Combined method is able to utilize appearance-based detection by dramatically lower false detections. Note that the performance improvement is consistent in both red and green signal phases since the crosswalk is a contextual mix zone.

## 6 Conclusion

This paper address miss detection of temporarily stopped vehicles and pedestrians at intersections by using both appearance and motion cues at predefined areas. The proposed three-stage cascaded detection-classification system has a contextual combination stage responsible for collecting the best detection results of each method and reducing false positives. Detected objects are finally given to HOG-SVM classifier to perform vehicle and pedestrian classification. Experimental results show the success of proposed method in comparison with traditional methods that use only motion at detection step. The proposed system can be used in detection-based trackers for effective intersection monitoring.

## References

1. Veeraraghavan, H., Masoud, O., Papanikolopoulos, N.P.: Computer vision algorithms for intersection monitoring. IEEE Transaction on Intelligent Transportation Systems 4, 78–89 (2003)
2. Kamijo, S., Matsushita, Y., Ikeuchi, K., Sakauchi, M.: Traffic monitoring and accident detection at intersections. IEEE Transaction on Intelligent Transportation Systems 1, 108–118 (2000)
3. Elhoseiny, M., Bakry, A., Elgammal, A.: Multiclass object classication in video surveillance systems experimental study. In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, pp. 788–793 (2013)
4. Chen, L., Feris, R., Zhai, Y., Brown, L., Hampapur, A.: An integrated system for moving object classification in surveillance videos. In: Proceeding of IEEE International Conference on Advanced Video and Signal Based Surveillance, vol. 4, pp. 52–59 (2003)

5. Zhang, L., Li, S.Z., Yuan, X., Xiang, S.: Real-time object classification in video surveillance based on appearance learning. In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
6. Ojala, T., Pietikainen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. Transactions on Pattern Recognition 29, 51–59 (1996)
7. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (2001)
8. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2 (1999)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
10. Philip, B., Updike, P., Weber, M.: Car dataset from the rear, california institute of technology, `http://www.vision.caltech.edu/archive.html`
11. Marszałek, M., Schmid, C.: Accurate object localization with shape masks. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
12. Papageorgiou, C., Poggio, T.: A trainable object detection system: Car detection in static images. Technical Report 1673, CBCL Memo 180 (1999)
13. Ozuysal, M., Lepetit, V.: P.Fua: Pose estimation for category specific multiview object localization. In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, pp. 778–785 (2004)
14. Agarwal, S., Awan, A., Roth, D.: Learning to detect objects in images via a sparse, part-based representation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1475–1490 (2009)
15. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC 2012) Results, `http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html`
16. Enzweiler, M., Gavrila, D.: Monocular pedestrian detection: Survey and experiments. IEEE Transactions on Pattern Analysis and Machine intelligence, 2179–2195 (2009)
17. Ess, A., Leibe, B., Gool, L.V.: Depth and appearance for mobile scene analysis. In: Proceeding of IEEE International Conference on Computer Vision, pp. 1–8 (2007)
18. Papageorgio, C., Poggio, T.: A trainable system for object detection. International Journal of Computer Vision 38, 15–33 (2000)
19. Overett, G., Petersson, L., Brewer, N., Andersson, L., Pettersson, N.: A new pedestrian dataset for supervised learning. In: Proceeding of IEEE Conference on Intelligent Vehicle Symposium, pp. 373–378 (2008)
20. Wojek, C., Walk, S., Schiele, B., Perona, P.: Multi cue on board pedestrian detection. In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, pp. 794–801 (2009)
21. Mogelmose, A., Prioletti, A., Trivedi, M., Broggi, A., Moeslund, T.B.: Two-stage part-based pedestrian detection. In: Proceeding of IEEE International Conference on Intelligent Transportation Systems, vol. 4, pp. 73–77 (2012)
22. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Transaction on Intelligent Systems and Technology 2, 1–27 (2011)