Urban Road User Detection and Classification using 3D Wire Frame Models

N. Buch, J. Orwell and S.A. Velastin

This paper presents a detection and classification system for vehicles and pedestrians in urban traffic scenes. This aims to guide surveillance operators and reduce human resources for observing hundreds of cameras in urban traffic surveillance. We perform per frame vehicle detection and classification using 3D models on calibrated cameras. Motion silhouettes (from background estimation) are extracted and compared to a projected model silhouette to identify the ground plane position and class of vehicles and pedestrians. The system is evaluated with the reference i-LIDS datasets from the UK Home Office. Performance for varying numbers of classes, for three different weather conditions and for different video input filters is evaluated. The full system including detection and classification achieves a recall of 87% at a precision of 85.5% outperforming similar systems in the literature. The i-LIDS dataset is available to other researchers to compare with our results. We conclude with an outlook to use local features for improving the classification and detection performance.

1 Introduction

We are addressing detection and classification of vehicles and pedestrians in urban traffic scenes. The advanced digital data infrastructure of deployed surveillance systems enables the development of automated video analysis tools. This can allow online and post-event detection of events of interest, which is useful for surveillance operators. The current main bottleneck of surveillance is the limitation of human resources for observing hundreds of cameras. Automatic pre- processing allows efficient guidance for the operators to pick cameras to view and accumulate statistics, with the aim to improve traffic flow.

Detection and classification results can be used to detect events and raise alarms. Systems should work independently of camera shake, weather, day or night, rain and so on. This highlights the difficult task of dealing with outdoor scenes.

Please refer to Figure 1 for two examples of typical cases. Note the low angle camera view which is inherent to urban scenes leading to occlusion of vehicles.

The problem we solve is vehicle classification on a per frame basis of a video stream. Every frame is treated independently for classification and no reasoning about the movement of vehicles is performed. Silhouettes extracted by foreground analysis are the input to our classifier. The classification process is based on 3D models for vehicles and can be restricted to an active region of the camera view. This allows the configuration of suitable locations for classification and restrictions to lanes.

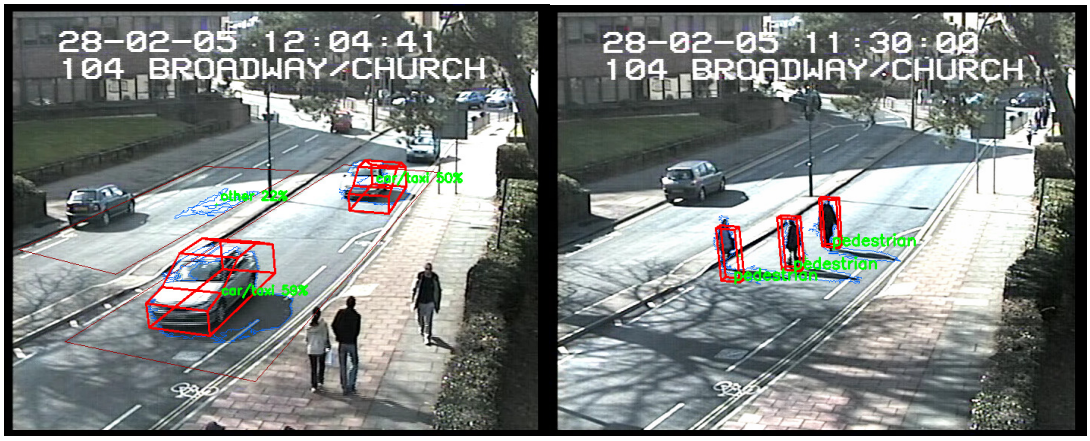

Figure 1 Example views from the i-LIDS data set with active windows and detected vehicles and pedestrians

The following assumptions are made: Every silhouette contains exactly one vehicle being fully visible. This implies no occlusion in the scene and between vehicles. The orientation of the vehicles throughout the scene remains approximately constant, which implies vehicles following a straight road. Every active region in Figure 1 can be processed for a different orientation.

There are five categories used for our classifier:

- Bus / Lorry
- Van
- Car / Taxi
- Motorbike / Bicycle
- Pedestrian

Our novel contribution is twofold. Firstly the use of 3D models for road user classification. In particular, matching those 3D models with closed contours

extracted from motion foreground is novel. All main road users are detected and classified with a single framework. The second contribution is the system evaluation on a public dataset. We present evaluation results on the i-LIDS dataset from the UK Home Office which can be licensed by research institutions and manufacturers [1].

The rest of the paper is organised as follows. Section 2 gives an overview of related work and our solution. The detector is introduced in section 3. Section 4 covers the classifier and models used. The evaluation of the proposed system is given in section 5. Finally, conclusions and future research can be found in section 6.

2 Related work

This section introduces the basic idea of traffic surveillance and discusses the literature dealing with vehicle detection and classification. The last sub-section provides an outline for our new approach.

2.1 Vehicle detection and classification

The most common application for vehicle classification is counting to generate statistics about road usage. This is often applied to highways and free flowing traffic with cameras mounted high above the ground. A growing area of deployment for automatic surveillance systems is the urban environment where traffic control or policing like bus lane enforcement is desirable to allow better attention coverage of existing cameras. Often inductive loops are used to detect vehicles at traffic lights in order to control traffic flow. Those loops are expensive to install and maintain and could be replaced by cameras. The potentially denser urban traffic increases the difficulty for vehicle detection and classification. Path information is required to deal with advanced requirements such as detection of traffic infringements.

2.2 Previous work

This section reviews work on detection and classification of vehicles in particular for urban scenes. The classification of vehicles has lacked attention compared to detection

and is usually performed on a low number of classes. In addition, pedestrians and vehicles are rarely considered within a single framework.

A motion region- based vehicle detection and classification system is proposed in [6]. The main focus of that paper is on detection. Effort is put into a fast background estimation using the 'instantaneous background' to get good segmentation, however a risk of persistent artefacts arises. Tracking is performed using graph correspondence based on blob overlap. The proposed classifier uses only two classes (cars, non cars) with size based features, which are not robust to occlusion common for urban conditions. Full camera calibration is required to normalise those features. On the validation sequence of 20 minutes, only 70% of vehicles are correctly classified.

The paper of Messelodi et *al.* [9] introduces a real time system to track and classify vehicles at intersections in urban areas. 3D models are used to initialise an object list for every fifth frame based on the convex hull overlap of model projection and motion map which requires camera calibration. Homogenous texture of the road surface (no markings) is required, which is very limiting. A feature tracker follows the detected objects along some frames before a new initialisation takes place due to complexity constraints. The objects are classified into 8 classes based on a two stage classifier. Performance is evaluated on 45 minutes of video data from two different sites. The system has a recall of 82.8%. The classification rate for the classifier independently is given with 91.5% for the test data.

In a previous paper [2], the authors used 3D models matched to motion foreground from Gaussian Mixture Models (GMM). Performance is given for diverse weather conditions on the public i-LIDS [1] dataset. In this paper we provide more input image filters and include pedestrian models to allow the framework to operate on both vehicles and pedestrians at the same time. We have also considered a more balanced population of classes: motorbike/bicycle, car/taxi, van and bus/lorry. In addition, we introduce quantitative pedestrian evaluation and hence report more results which are more representative of operational performance.

A real time monitoring system for intersections is proposed in [17]. A standard Gaussian Mixture Model [14] is used for foreground segmentation similar to our approach. Tracking of foreground regions is done with graph correspondence. The tracked objects are classified into two classes, pedestrians and vehicles based on the

main orientation of the bounding box which is not reliable for low camera angles. Further distinction between vehicle classes would be desirable. No quantitative evaluation is given.

In [4] a vehicle classification algorithm is introduced. Standard foreground segmentation is performed to get bounding boxes of vehicles. The pixel values inside the bounding boxes form a basic feature vector. Independent Component Analysis (ICA) is performed on the training images to reduce the dimension of the feature space. This is similar to the image based features (IB) described in [10], however they were found to be the weaker approach. To assign one of the three class labels to new feature vectors during operation, three one class Support Vector Machines (SVM) are used. The reported performance is roughly 65% recall at 75% precision. A more advanced classifier for vehicles based on local features is presented in [8], where a tracking algorithm produces normalised images of vehicles. Good results are shown for binary classification between vehicle categories, however, a multi class classifier is missing. Car versus Minivan achieves precision of up to 98.5%, whereas Sedan versus Taxi achieves 96.5%.

The use of 3D models for vehicle detection and classification was proposed in [15]. First, a hypothesis for a vehicle position is generated by correlating 1D templates with the image. The hypothesis is verified by correlating the gradient input image with the wire frame image of vehicles. A classification into cars and vans takes place based on the hypothesis. The performance is evaluated in a very short 1 minute video sequence, where 45 out of 46 vehicles were classified correctly. This approach is followed up in [19]. Another approach for urban vehicle tracking with 3D models is presented in [13]. However, only a single model for cars is used to estimate a vehicle constellation. A Bayesian problem is formulated for multiple vehicle positions and solved with a Markov Chain Monte Carlo (MCMC) algorithm to give several good solutions. A Viterbi algorithm is used to find the optimal track through the set of solutions for every frame. The reported detection rates are 96.8% and 88% for two videos, but the system is limited to dealing with vehicles of a single size in the scene.

The paper of Morris and Trivedi [11] presents a combined tracking and classification approach for side views of highways. This combination is an extension to [10]. A single Gaussian background model is used for foreground segmentation. For

every segmented foreground silhouette, a feature vector based on silhouette measurements is calculated. The evaluation on a 24 hour test sequence shows a classification accuracy of 74.4% for independent tracking and classification. The accuracy can be increased by combining tracking and classification to 88.4% which is desirable. An earlier paper [10] focuses on the classification task assuming known segmentation information for the input video data. A comparison between image based features (IB) like pixels and image measurements features (IM) like region size is given. IM with Linear Discriminant Analysis LDA was used for the final algorithm as it gave the best performance, however, it relies on correct motion silhouettes. The features are classified using a weighted *k*- nearest neighbour algorithm. A Kalman filter is used to track the foreground regions based on the centroids. The tracking improves the classification result from 82.9% to 91.4% by rejecting single miss-classification.

2.3 Outline of the proposed approach

In the system reported here, a classifier generates a hypothesis of a vehicle or pedestrian being present in the scene by placing each 3D model onto the scene's ground plane and projecting it to the camera view. A match measure is calculated for every hypothesis by comparing the model with the image silhouette. Every model is placed on a grid of positions on the ground plane to produce the match measure for every silhouette based on superior (IM) features according to [10]. The highest measure indicates the most likely position of the vehicle given the silhouette. The highest match measures of different classes are compared to make a decision about the class of a silhouette. Silhouettes with consistently low match measures for all classes are rejected as ambiguous. To use the 3D models, cameras are calibrated by means of a map and five corresponding points with the image.

The orientation of vehicles is assumed to be constant, as pointed out earlier. One orientation is defined for every detection region of interest. A single object is assumed for every silhouette (i.e. no overlap). With those assumptions, the score is a size and overlap measure between the silhouette and the projected wire frame of the model.

3 Detection

The detector uses background estimation to extract motion silhouettes from a video frame. See Figure 2 for a block diagram of the proposed system. The detector is described in more detail in this section, followed by the classifier in section 4.

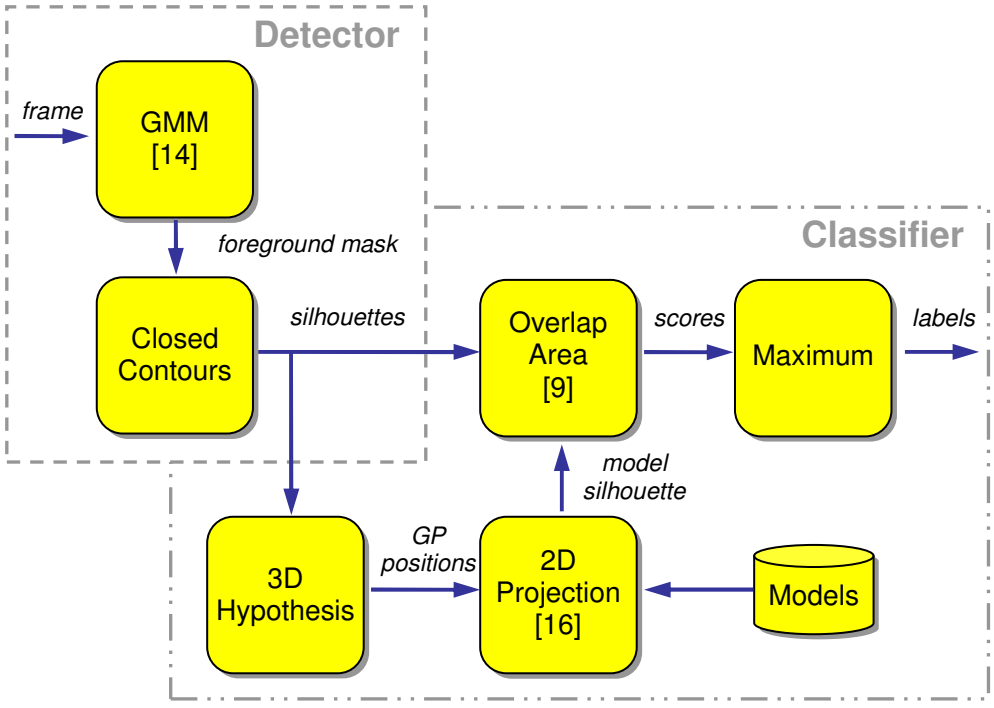

Figure 2 Block diagram of the detection and classification system

Surveillance videos are commonly captured with analogue cameras producing interlaced video signals. To rectify the zigzag boundary artefacts generated for moving objects, a pre-processing step linearly interpolates odd video lines between even lines. Performance will be demonstrated with and without de-interlacing filtering. A Gaussian Mixture Model implementation [7] from the OpenCV library [12] was used. The GMM, first introduced in the seminal paper of Stauffer and Grimson [14], is used to generate an initial foreground mask. Five Gaussians are estimated using a background threshold of 0.7 , which is the default value. The window size which is the inverse of the learning rate is chosen at 50 to allow fast adaptation to illumination changes. Stationary changes are therefore incorporated into the background within 15 seconds. The outdoor scene recorded with an auto iris function of the camera requires

fast learning to accommodate illumination changes. Large objects in the scene can change the overall illumination conditions due to this gain control.

The foreground pixels are post processed with a constant chromaticity shadow removal algorithm like in Cucciara et *al.* [5]. The background image, as well as the current input image, is transformed into the HSV (Hue, Saturation, Value) space. This algorithm assumes that hue and saturation stay constant and only the value changes for shadowed surfaces. This assumption holds for light shadows as seen in overcast condition if the camera is not saturated. Value reductions down to 55% of the input with respect to the background are considered shadows.

Closed contours $\{S_m\}$ are extracted from the final foreground mask. A filter operation is used to produce a smaller set of valid silhouettes $\{S_k\}$ considering size and location with respect to the region of interest. Those silhouettes $\{S_k\}$ will be used as input for classification. The length operator $\mathrm{L}(S_m) \in$ pixels computes the circumference of the contour. Based on the area operator $\mathrm{A}(\ ) \in pixels$, the overlap ratio operator $\mathrm{O}(S_m, R) \in [0,1]$ gives the overlap of a contour with the region of interest mask R (red outline in images *e.g.* Figure 1):

$$\mathrm{O}(S_m, R) = \frac{\mathrm{A}(S_m \cap R)}{\mathrm{A}(S_m)} \tag{1}$$

Valid silhouettes $\{S_k\}$ have to satisfy the length threshold $\tau_L \in$ pixels and the overlap threshold $\tau_O \in [0,1]$. We use $\tau_L = 200$ and $\tau_O = 0.25$ for our experiments:

$$\{S_k\} = \{S_m \,|\, \mathrm{L}(S_m) > \tau_L \wedge \mathrm{O}(S_m, R) > \tau_O\} \tag{2}$$

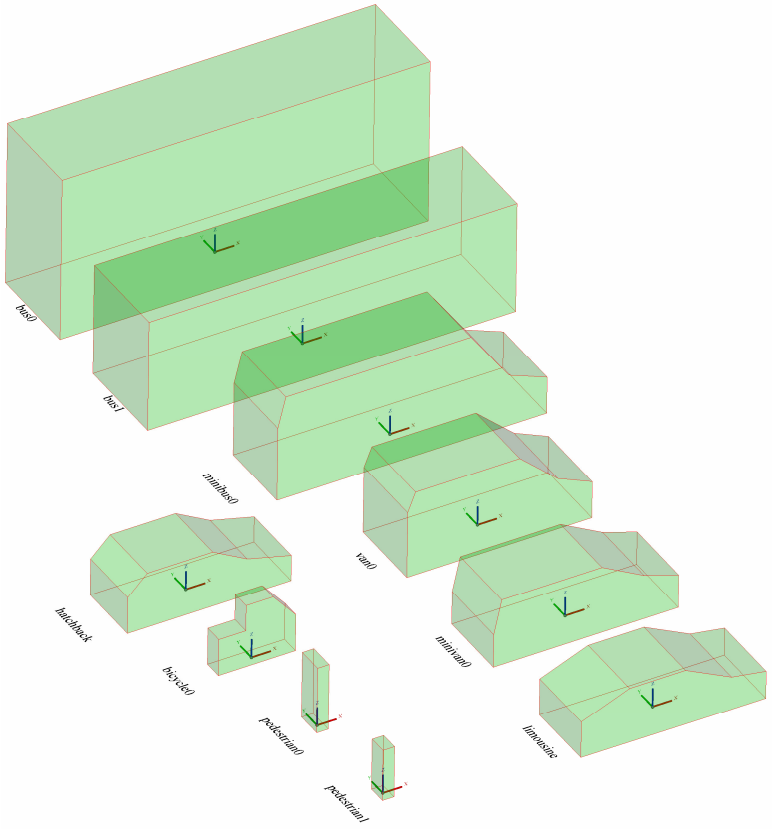

Figure 3 Wire frame models $\{F_i\}$ used for classification

4 Classification

The classifier uses 3D wire frame models to find the corresponding class label j for every silhouette S_k generated by the detector. The camera requires calibration to be able to use 3D models. The algorithm of Tsai [16] is used to obtain the ground plane calibration for the camera using a map of the road and defining at least five corresponding points between map image and camera image. Based on the calibration, ground plane coordinates can be converted to image coordinates $\mathrm{IM}(\ )$. Conversion from image to ground plane $\mathrm{GP}(\ )$ requires definition of the z coordinate of the image point.

First, 3D hypothesis are generated from the silhouettes. The centroids $\{\mathbf{c}_k\}$ of $\{S_k\}$ are back projected to the ground plane assuming zero height giving reference points $\{\mathbf{r}_k\}$

$$\mathbf{r}_k = \mathrm{GP}(\mathbf{c}_k). \tag{3}$$

The zero height assumption of the centroids $\{\mathbf{c}_k\}$ introduces position noise, which is dealt with by generating several hypotheses around the reference point. The full set of 3D hypotheses ground plane positions $\{\mathbf{h}_{x,y,k}\}$ is generated by placing square grids $\mathrm{G}(\ )$ of points at the reference points $\{\mathbf{r}_k\}$

$$\mathbf{h}_{x,y,k} = \mathrm{G}(\mathbf{r}_k). \tag{4}$$

For our experiments, the total grid width was 7 metres square containing 7 rows and columns. The 2D projection generates model masks $\{M_{x,y,i,k}\}$ for every 3D hypothesis $\mathbf{h}_{x,y,k}$. Figure 3 shows the full set of wire frame models $\{F_i\}$ used for classification. The model dimensions are based on current manufacturers' information. The function $\mathrm{SIL}(\ )$ projects the wire frame to the image and gives the silhouette

$$M_{x,y,i,k} = \mathrm{SIL}(F_i, \mathbf{h}_{x,y,k}). \tag{5}$$

Every model point is projected independently and the wire frame is drawn in the image between the projected points. The silhouette of this image is returned. The measure of quality of fit between the silhouettes $\{S_k\}$ and model masks $\{M_{x,y,i,k}\}$ is defined by the normalised overlap area $\mathrm{ON}(\ )$:

$$\mathrm{ON}(M_{x,y,i,k}, S_k) = \frac{\mathrm{A}(M_{x,y,i,k} \cap S_k)}{\mathrm{A}(M_{x,y,i,k} \cup S_k)} \tag{6}$$

which is similar to the approach in [9]. Figure 4 gives an illustration of the normalised overlap output. Note the well shaped peak of the overlap function. A maximum operation is performed to find the highest quality of fit $P_K \in [0,1]$ for every silhouette S_k

$$P_k = \max_{x,y,i} \mathrm{ON}(M_{x,y,i,k}, S_k) \tag{7}$$

with corresponding ground plane position (x, y) and model label i. The configuration table T is used to retrieve the class label j for the model i as there can be many models for one class to cope with intra class variability

$$j = \mathrm{T}(i). \tag{8}$$

A final threshold $\tau_P = 0.48$ is applied to $\{P_k\}$ to eliminate silhouettes which do not correspond to any class. The set of detected vehicles $\{D_L\}$ with $L < k$ is given by

$$\{D_L\} = \{P_k \,|\, P_k > \tau_P\}. \tag{9}$$

The intermediate results and internal steps of the algorithm are illustrated in Figure 5 showing mask and silhouette images.

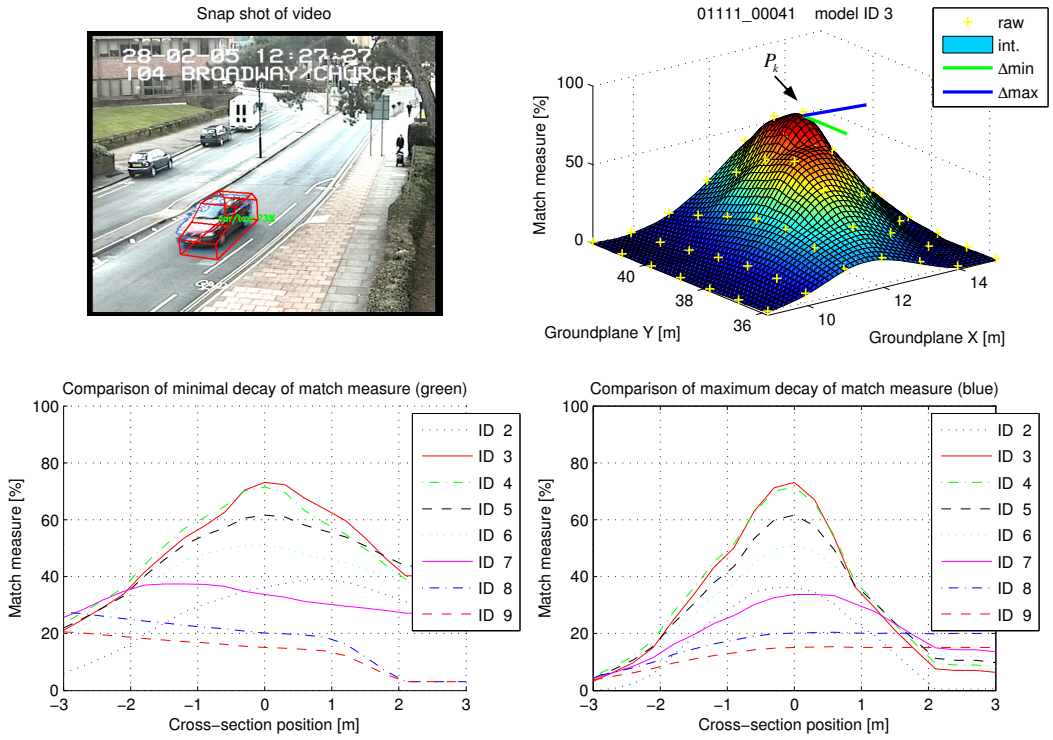

Figure 4 Match measure for one silhouette S_k. The upper left image shows the silhouette and best fitting model i at (x, y). Top right: the winning match surface $\max_i \mathrm{ON}\left(M_{x,y,i,k}, S_k\right)$ with data points. Bottom: Cross- section through every model's match surface $\mathrm{ON}\left(M_{x,y,i,k}, S_k\right)$ along the minimum and maximum decay direction at (x, y).

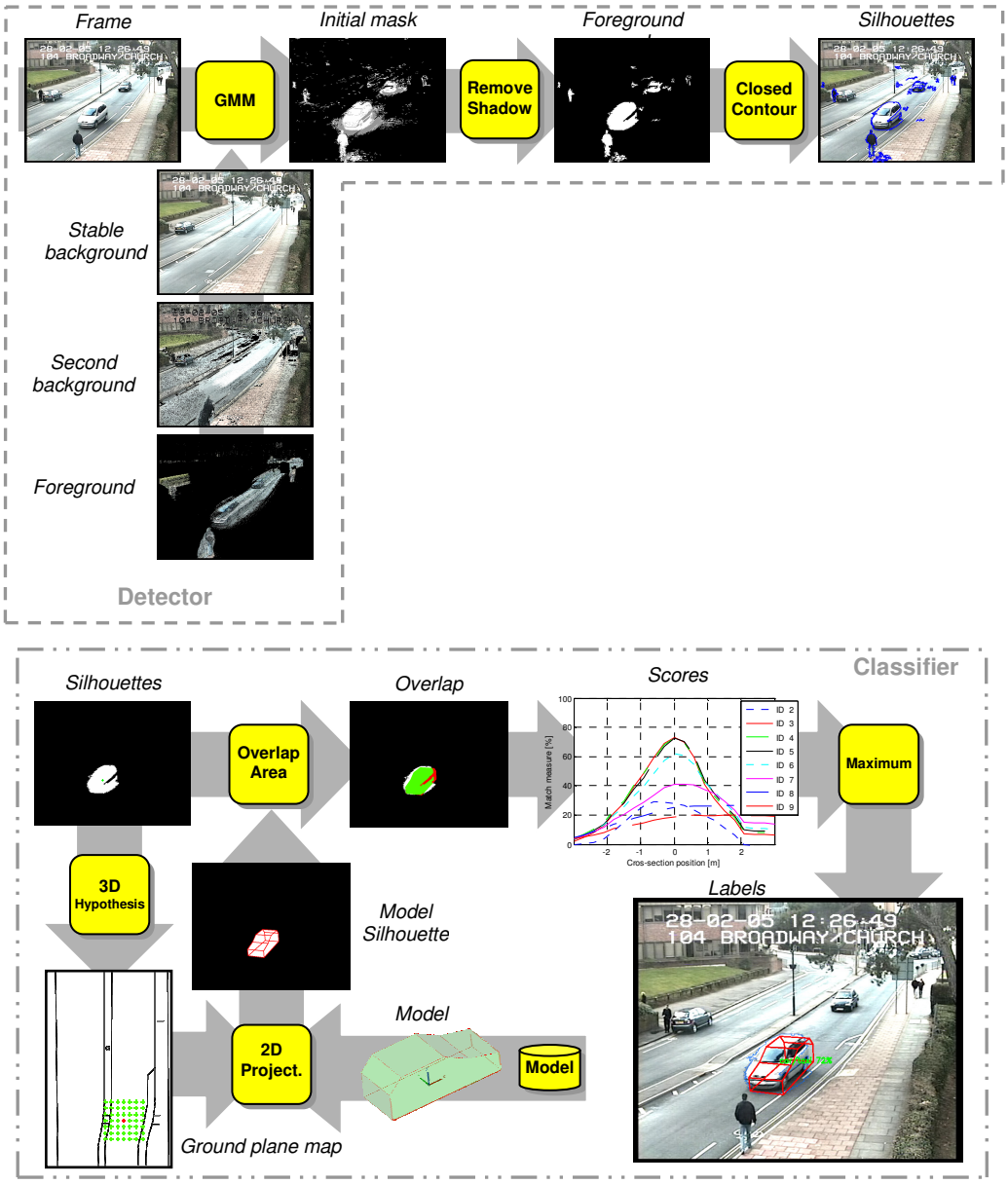

Figure 5 Illustration of data flow for the detection and classification framework

5 Evaluation

The proposed system was evaluated on video from the i-LIDS datasets [1]. Ground truth $\{GT_n\}$ was provided in Viper format [18] consisting of bounding boxes and class labels for vehicles. The classifier produces bounding boxes and class labels for detected vehicles $\{D_L\}$ also in Viper format. For every detected vehicle D_L, the best

matching bounding box out of the ground truth $\{GT_n\}$ is considered as matching vehicle GT_n. The entry in the confusion matrix depends on the class labels. If no overlapping vehicle GT_n is found in the ground truth, the detected vehicle D_L is entered in column FP (false positive). Finally, all non matched vehicles in the ground truth $\{GT_n\}$ within the region of interest are entered in row FN (false negatives). The next section introduces the dataset used. Section 5.2 gives results for vehicle detection and classification. Joint operation of vehicles and pedestrians is evaluated in section 5.3. The influence of weather conditions is demonstrated in section 5.4.

5.1 Data set

The i-LIDS datasets [1] are licensed by the UK Home Office for image research institutions and manufacturers. Each dataset comprises 24 hours of video sequences under a range of realistic operational conditions. They are used by the UK government to benchmark video analysis products. They are ideal for evaluating and comparing algorithms in the computer vision community and there is a gradual increase in take-up. Out of the Parked Car dataset, scenario 1 was chosen. Refer to Figure 1, Figure 6 and Figure 7 for example views. There is no public dataset commonly used for urban traffic analysis. This makes direct comparison of reported results difficult. Our contribution is the use of this public data set to allow quick future comparison of systems in the same environment. Approximately one hour of video for sunny, overcast and changing conditions is selected for our evaluation: (PVTRA10xxxx) 1a03, 1a07, 1a13, 1a19, 1a20, 1a21, 2a04, 2a05, 2a06, 2a08, 2a09, 2a10, 2a11 and 2a15. The recordings use a camera with an auto iris function which keeps the average illumination of the view constant. Large vehicles with a predominant colour can cause adjustments in the iris and a changed background. In addition, the overcast videos contain saturated areas in the middle and far end of the view.

Some ground truth usable for our tests (the data is normally used for event detection tests) was provided with the dataset, however it had to be converted and extended for our use. This limited the total length of video used for the evaluation. The total number of vehicle and pedestrian appearances is 782 as in Table 3. The total

is divided between classes as follows: 46% car/taxi, 30% pedestrian and 8% each for motorbike/bicycle, van and bus/lorry.

5.2 Detection and classification of vehicles

This section provides results for the proposed system using vehicle models only. Good results are demonstrated outperforming base line solutions in the literature for vehicle detection and classification. Using the shadow removal filter a without de-interlacing filter gives the best performance. Table 1 shows an extended confusion matrix including FP (false positives) and FN (false negatives) for the evaluation of detector and classifier and Table 2 show results for the classifier only.

ground truth / *detection*	bike	car/taxi	van	bus/lorry	FP
bike	.87	.03	.02	0	.36
car/taxi	0	.86	.1	0	.05
van	0	.02	.84	.02	.1
bus/lorry	0	.02	.03	.98	.05
FN	.13	.07	.02	0	0
count	45	370	63	62	
overlap	.64	.66	.69	.72	

Symbol	*Value*
Recall *R*	87.0%
Precision *P*	85.5%
Classifier P_C	92.9%
Detector R_D	93.7%
Detector P_D	92.0%
GT Overlap	0.67

Table 1 Confusion matrix and overall performance for vehicle operation using shadow removal

ground truth / *detection*	bike	car/taxi	van	bus/lorry
bike	1	.03	.02	0
car/taxi	0	.92	.1	0
van	0	.02	.85	.02
bus/lorry	0	.03	.03	.98
count	39	343	62	62

Table 2 Confusion matrix for classifier using shadow removal

All values are normalised to the ground truth count per class displayed at a bottom row. The overlap indicates the overlap between ground truth bounding box and

detection bounding box, which is obtained as the bounding box of the detected wire frame model. Precision P and recall R can be calculated from the confusion matrix. Separate analysis of classifier (subscript C), detector (subscript D) and the whole system is available. FP and FN can be read from the matrix. TP (true positives) are the diagonal elements for the whole system and the classifier. For the detector, the columns excluding FN are summed to give the TP. With the general definitions

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$P = \frac{TP}{TP + FP} \tag{11}$$

the whole system evaluates to a recall R of 87% at a precision P of 85.5%. The classifier achieves a high precision P_C of **92.9%**. By definition, the classifier recall $R_C = P_C$ when considering all classes jointly. The detector has a recall R_D of **93.7%** at a precision P_D of 92%. For qualitative results, refer to Figure 6 for true positive examples and Figure 7 for wrong classification. The higher number of false positives for the class bike is due to pedestrians being classified as bikes. At this stage, no pedestrian model was used and all the motion silhouettes resulting from pedestrians in the scene should have been rejected. Individual performance of the classifier can be found in Table 2.

Figure 6 Examples of correct detections and classification of vehicles using the shadow removal filter

Figure 7 Two examples for false positives due to pedestrians being detected as bike and as car due to occlusion in a group. The last image shows a missed car due to its similar colour compared to saturated road area

Direct comparison of quantitative results with the literature is difficult due to the lack of a common dataset for vehicle classification. The total recall R and precision P of the proposed system outperforms the following systems in term of their reported results on their own datasets. The most relevant reference is [9], as a system performance of 82.8% for detection and classification of urban road users into 8 classes is reported. All the following systems use highway imagery, which highlights

the relevance of presenting results on a public urban dataset like i-LIDS [1] for vehicle classification. Total system R 65% at P 75% for classifying 150 car samples into 3 classes after detection and tracking is achieved in [4]. On 20 minutes test video, 70% of vehicles are classified (cars / non cars) after detection and tracking in [6]. A classifier accuracy of 74.4% is reported for a 24 hour test sequence in [10] using 3 classes. The same authors extended the system to 7 classes with classification accuracy of 88.4% in [11].

Results of the proposed algorithm are compared for four different scenarios using input filters for shadow removal and de-interlacing. The effect of using different combinations of those filters is shown in Figure 8.

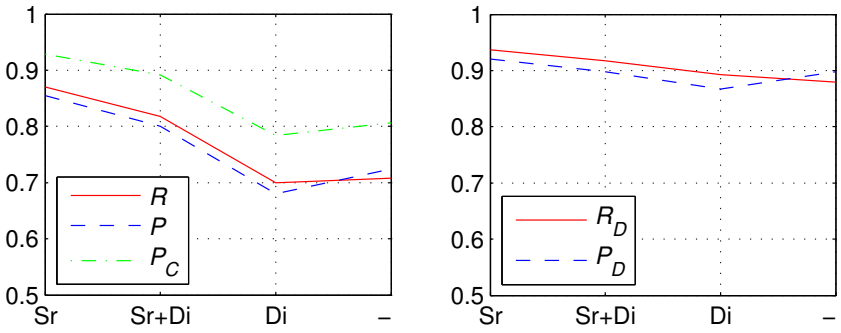

Figure 8 Performance comparison for vehicle framework using 4 different filter algorithms: shadow removal (Sr), shadow removal with de-interlacing (Sr+Di), de-interlacing (Di) and no filter (-). The left diagram shows system recall R, precision P and classifier precision P_C. The right diagram indicates the detector recall R_D and precision P_D

5.3 Simultaneous operation for vehicles and pedestrians

This section demonstrates results of the proposed algorithm, when vehicles and pedestrians are classified with the same framework. Results are given for the same four filter configurations introduced in the last section with a qualitative comparison in Figure 9. Best performance is shown for shadow removal without de-interlacing filter.

Figure 9 Two example views of the combined vehicle and pedestrian framework for 4 different filter configurations from top to bottom: shadow removal (Sr), shadow removal with de-interlacing (Sr+Di), de-interlacing (Di) and no filter (-). Too large silhouettes can be observed without shadow removal causing missed vehicles and wrong classifications (last two columns).

5.3.1. Shadow removal filter

The framework used with the shadow removal filter gives the best performance for vehicle and pedestrian detection. Refer to Table 3 for an extended confusion matrix with overall performance figures and to Table 4 for class wise results. Very good classification performance is observed for the vehicles classes, whereas confusion occurs between bikes and pedestrians. This is due to very similar motion silhouettes of both road users, especially in the far region of the camera view when bicycles are seen front on (see Figure 6). A higher false positive rate for the bike class observed earlier for the vehicle classifier (Table 1) does not appear here, as a pedestrian model was used. The low detection performance of pedestrians is due to their non-rigid nature. The basic cube-like models do not match motion silhouettes of pedestrian as well as they do cars, which required the detection threshold to be halved for pedestrians. In addition, the interlacing of the cameras does affect smaller object more, which explains the performance increase of pedestrians when using a de-interlacing filter in the next section. However, using a single algorithm for all road users is beneficial in terms of system complexity.

ground truth / *detection*	pedestrian	bike	car/taxi	van	bus/lorry	FP
pedestrian	.71	.49	.01	.02	0	.1
bike	.02	.47	.03	0	0	.02
car/taxi	.02	0	.85	.1	0	.04
van	0	0	.02	.84	.02	.08
bus/lorry	0	0	.02	.03	.98	.03
FN	.24	.04	.07	.02	0	0
count	241	45	371	63	62	
overlap	.57	.6	.66	.69	.72	

Symbol	*Value*
Recall R	79.5%
Precision P	83.9%
Classifier P_C	89.8%
Detector R_D	88.6%
Detector P_D	93.5%
GT Overlap	0.64

Table 3 Confusion matrix of full system for using shadow removal filter including overall performance figures

	pedestrian	bike	car/taxi	van	bus/lorry
R_j	71.0%	46.7%	85.2%	84.1%	98.4%
P_j	77.4%	58.3%	92.4%	79.1%	81.3%
R_{Cj}	94.0%	48.8%	91.9%	85.5%	98.4%
P_{Cj}	87.2%	60.0%	96.6%	85.5%	83.6%
R_{Dj}	75.5%	95.6%	92.7%	98.4%	100.0%
P_{Dj}	88.7%	97.2%	95.6%	92.5%	97.3%

Table 4 Class wise performance figures

5.3.2. Shadow removal and de-interlacing filter

The framework with both input filters indicates best performance for pedestrians. The confusion matrix in Table 5 shows system recall 82% for pedestrians, which is an improvement of 11% compared to shadow removal filtering only. The additional de-interlacing filter allows a better match of motion silhouettes compared to last section. However, the classification performance for vehicles degraded, particularly the recall of vans from 84% to 63%.

ground truth / *detection*	pedestrian	bike	car/taxi	van	bus/lorry	FP
pedestrian	.82	.55	.02	.05	0	.12
bike	.02	.45	.04	0	.01	.02
car/taxi	.01	0	.81	.26	0	.06
van	.01	0	.02	.63	.04	.09
bus/lorry	0	0	.02	.02	.94	.06
FN	.14	0	.09	.05	0	0
count	277	51	373	65	68	
overlap	.57	.58	.66	.73	.7	

Table 5 Confusion matrix of full system for shadow removal and de- interlacing filter

5.3.3. De-interlacing filter and no filter

For those experiments, only the de-interlacing filter or no filters were used. In both cases, the performance is significantly worse than the experiment including the

shadow removal filter, which can be observed in Table 6. Compared to the best performance in section 5.3.1, recall drops by 11.7% to 67.8% and precision drops by 10.5% to 73.4%. This is due to oversized motion silhouettes, which can be seen in Figure 9. Therefore, shadow removal is essential for this framework to perform well.

ground truth / *detection*	pedestrian	bike	car/taxi	van	bus/lorry	FP
pedestrian	.63	.43	.03	.31	0	.24
bike	.12	.35	.01	.01	0	.08
car/taxi	.01	.02	.7	.09	0	.08
van	.01	0	.09	.45	0	.06
bus/lorry	0	0	.07	.14	1	.1
FN	.22	.2	.1	0	0	0
count	242	51	382	109	52	
overlap	.57	.42	.57	.59	.53	

ground truth / *detection*	pedestrian	bike	car/taxi	van	bus/lorry	FP
pedestrian	.63	.47	0	.08	0	.08
bike	.07	.31	.01	0	0	.06
car/taxi	.02	.02	.72	.11	0	.06
van	0	0	.09	.63	0	.13
bus/lorry	0	0	.05	.19	1	.11
FN	.27	.2	.13	0	0	0
count	208	51	370	80	53	
overlap	.58	.44	.57	.57	.54	

Table 6 Confusion matrix for de-interlacing filtering and no filtering

5.4 Influence of weather conditions

Robust operation under realistic weather conditions is important. This section compares the performance of the vehicle classification and detection for sunny, overcast and changing conditions. Direct comparison is given in Figure 10 indicating that sunny conditions perform best. This can be contributed to the high contrast in the videos and therefore good foreground estimation. The following sections give more details about each condition.

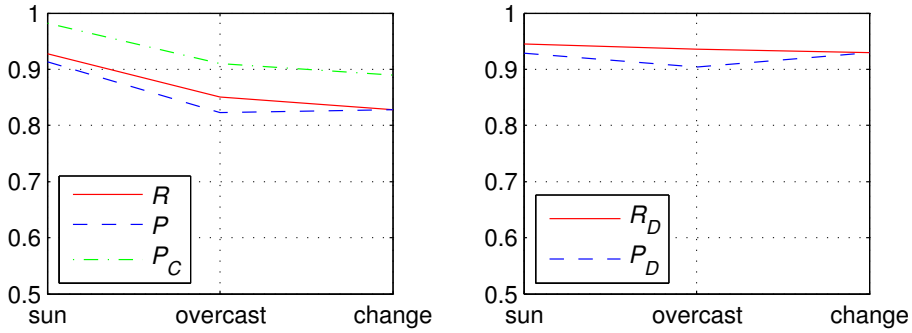

Figure 10 Performance comparison for three different weather conditions

5.4.1. Sunny conditions

The best individual performance is demonstrated for sunny conditions, with the confusion matrix shown in Table 7. This unexpected case of sunny conditions outperforming overcast conditions for classification can be explained by the dynamic range of the images. The high contrast and the deep shadow can be seen in examples of Figure 11. The sun allows a precise detection of the outline of vehicles, however it includes a deep shadow. The classifier can deal with that shadow as the silhouette is only extended in a single direction which reduces the overlap match measure for all models but keeps the ordering. In contrast, the lower dynamic range and tendency of saturation for overcast conditions introduces more noise to the vehicle silhouette. This noise changes the size of the silhouette in general which enables the match of a wrong model. However, due to the shadow, the mean overlap measure of the winning class in sunny conditions is 0.65, lower than the corresponding figure in overcast conditions (0.69).

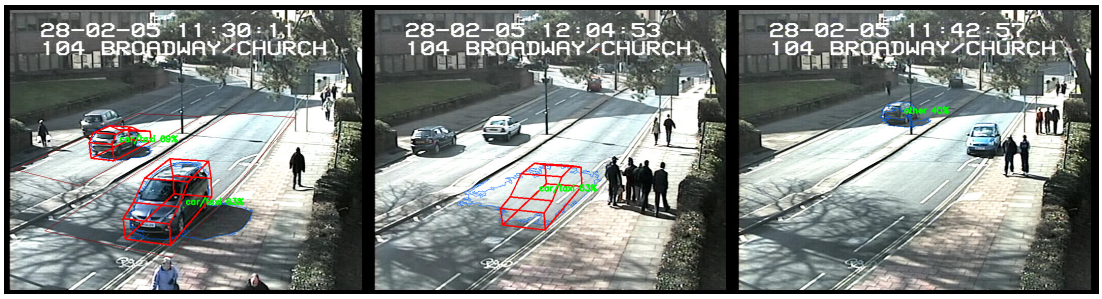

Figure 11 Sunny examples: true positive, false positive car and missed car

ground truth / *detection*	bike	car/taxi	van	bus/lorry	FP
bike	.71	0	0	0	.18
car/taxi	0	.95	0	0	.04
van	0	.01	.95	0	.25
bus/lorry	0	0	.05	1	0
FN	.29	.04	0	0	0
count	17	135	20	9	
overlap	.44	.66	.68	.67	

Table 7 Confusion matrix for sunny conditions

5.4.2. *Overcast condition*

The performance for overcast conditions is second best after sunny. The confusion matrix in Table 8 shows many false positives for. The false positives are observations of pedestrians, which are not rejected as ambiguous. The miss classifications are mainly due to missed foreground areas due to saturation and low dynamic range of the scene. Refer to Figure 12 for examples.

Figure 12 Overcast examples: two correct and one misclassified frame

ground truth / *detection*	bike	car/taxi	van	bus/lorry	FP
bike	.96	.03	0	0	.39
car/taxi	0	.78	.05	0	.05
van	0	.03	.91	.03	.05
bus/lorry	0	.06	.05	.97	.06
FN	.04	.1	0	0	0
count	28	119	22	33	
overlap	.72	.66	.76	.69	

Table 8 Confusion matrix for overcast conditions

5.4.3. *Overcast changing to sunny*

The worst performance can be observed for changing conditions. During those sequences, the sun appears several times which causes the auto iris of the camera to adjust. This produces ambiguous foreground silhouettes for short periods of time resulting in lower performance. Refer to Table 9 for the extended confusion matrix for this case with example views in Figure 13. The low performance of vans is due to their predominant white colour, which causes reduced foreground areas during times of saturation. This problem can be dealt with by exploiting the constraint that the same

vehicles are present in the scene for many frames. Temporal filters and tracking could be used, but this is outside the scope of what is being reported here.

Figure 13 Changing weather examples: Two correct and one misclassified frame

ground truth / *detection*	bike	car/taxi	van	bus/lorry	FP
bike	0	.05	.05	0	2
car/taxi	0	.83	.24	0	.07
van	0	.02	.67	0	0
bus/lorry	0	.02	0	1	.05
FN	0	.09	.05	0	0
count	0	116	21	20	
overlap	1	.66	.61	.8	

Table 9 Confusion matrix for changing conditions

6 Conclusions

We presented a novel solution to road user detection and classification using 3D models. The target application is urban traffic analysis which has different requirement compared to highway surveillance. 3D models based on car manufactures dimensions are projected onto the image plane to generate a silhouette match measure. This match measure produces a distinctive peak at the right ground plane position and distinguishes different classes.

Evaluation was performed on the public i-LIDS datasets [1]. We provide results with balanced numbers of vehicle and pedestrian appearances for several input filters and weather conditions. Good overall performance of a recall of 87% at a precision of 85.5% is demonstrated for our algorithm using vehicle models. The classifier achieves a high precision of 92.9% which outperforms other reported

results. The model set is extended to incorporate vehicles and pedestrians into the same framework. This gives comparable performance to the vehicle classifier showing a raised confusion rate between bicycles and pedestrians. This is due to their similar size and motion silhouette. The evaluation of input filters indicates, that shadow removal filtering gives best performance increase with de-interlacing improving pedestrian detection. Regarding weather conditions, the best classification performance of **98.2%** is achieved for sunny conditions outperforming overcast conditions and changing conditions. This result that contradicts general perception is due to the higher contrast and therefore less noise of the silhouettes in sunshine. As our classifier can deal with deep shadows, this condition gives the best results.

6.1 Future work

Efforts for future research should be directed towards more robust detection and classification of vehicles. There is little work on urban scenes which requires robustness against occlusions. We will focus on incorporating local feature information into the classifier to have a second source of information apart from the motion foreground. Local features indicated good performance in [8] and [3] when used in 2D. Additional cues will be vital to resolve occlusions at low angle camera views. Tracking is required to provide path information of vehicles and can increase classification performance as demonstrated in [10] and [11].

Acknowledgements

We are grateful to the Directorate of Traffic Operations at Transport for London for funding this project on Classification of Vehicles and Pedestrians for Urban Traffic Scenes.

The i-LIDS dataset provided by the UK Home Office is used for evaluation complying with the academic license.

Affiliations

Norbert Buch, James Orwell and Sergio A. Velastin are with the Digital Imaging Research Centre, Kingston University London, Penrhyn Road, Kingston upon Thames, Surrey, KT1 2EE, United Kingdom.

E- mail: n.buch@theiet.org

References

[1] Home Office Scientific Development Branch. Imagery library for intelligent detection systems i-lids. http://scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/vi% deo-based-detection-systems/i-lids/ [accessed 19 December 2008].

[2] Norbert Buch, James Orwell, and Sergio A. Velastin. Detection and classification of vehicles for urban traffic scenes. In *International Conference on Visual Information Engineering VIE08*, pages 182–187. IET, July 2008. ISBN 978-0-86341-914-0.

[3] Norbert Buch and Sergio A. Velastin. Human intrusion detection using texture classification in real- time. In J. Gonzàlez, T.B. Moeslund, and L. Wang, editors, *First International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences THEMIS 2008*, pages 1–6, September 2008.

[4] X. Chen and C. C. Zhang. Vehicle classification from traffic surveillance videos at a finer granularity. *Advances In Multimedia Modeling, Pt 1*, 4351:772–781, 2007.

[5] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, and S. Sirotti. Improving shadow suppression in moving object detection with hsv color information. In *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*, pages 334–339, 2001.

[6] S. Gupte, O. Masoud, R.F.K. Martin, and N.P. Papanikolopoulos. Detection and classification of vehicles. *Intelligent Transportation Systems, IEEE Transactions on*, 3(1):37–47, 2002.

[7] P. KadewTraKuPong and R. Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *Proceedings of 2nd European Workshop on Advanced Video-Based Surveillance Systems*, 2001.

[8] Xiaoxu Ma and W.E.L. Grimson. Edge-based rich representation for vehicle classification. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1185–1192, 2005.

[9] Stefano Messelodi, Carla Maria Modena, and Michele Zanin. A computer vision system for the detection and classification of vehicles at urban road intersections. *Pattern Analysis & Applications*, 8(1-2):17–31, September 2005.

[10] B. Morris and M. Trivedi. Robust classification and tracking of vehicles in traffic video streams. In *Intelligent Transportation Systems Conference. ITSC '06. IEEE*, pages 1078–1083, 2006.

[11] Brendan Morris and Mohan Trivedi. Improved vehicle classification in long traffic video by cooperating tracker and classifier modules. In *AVSS '06: Proceedings of the IEEE International Conference on Video and Signal Based Surveillance*, page 9, Washington, DC, USA, 2006. IEEE Computer Society.

[12] OpenCV. Open source computer vision library. http://sourceforge.net/projects/opencvlibrary [accessed 19 December 2008].

[13] Xuefeng Song and R. Nevatia. Detection and tracking of moving vehicles in crowded scenes. In *Motion and Video Computing. WMVC '07. IEEE W. on*, pages 4–4, 2007.

[14] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2, pages 246–252, June 1999.

[15] G D Sullivan, K D Baker, A D Worrall, C I Attwood, and P R Remagnino. Model-based vehicle detection and classification using orthographic approximations. In *Proceedings of 7th British Machine Vision Conference*, volume 2, pages 695–704, September 1996.

[16] Roger Y. Tsai. An efficient and accurate camera calibration technique for 3d machine vision. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR), (1986)*, pages 364–374, 1986.

[17] H. Veeraraghavan, O. Masoud, and N. Papanikolopoulos. Vision-based monitoring of intersections. In *IEEE 5th International Conference On Intelligent Transportation Systems, Proceedings*, pages 7–12, 2002.

[18] Viper. ground truth schema. http://viper-toolkit.sourceforge.net/ [accessed 19 December 2008].
[19] Zhaoxiang Zhang, Min Li, Kaiqi Huang, and Tieniu Tan. Boosting local feature descriptors for automatic objects classification in traffic scene surveillance. *International Conference on Pattern Recognition (ICPR) 2008*, 2008.