Detection and Classification of Vehicles from Omnidirectional Videos using Temporal Average of Silhouettes

Hakki Can Karaimer and Yalin Bastanlar
Computer Vision Research Group, Department of Computer Engineering, Izmir Institute of Technology, 35430, Izmir, Turkey
{cankaraimer, yalinbastanlar}@iyte.edu.tr

Keywords: Omnidirectional Camera, Omnidirectional Video, Object Detection, Vehicle Detection, Vehicle Classification.

Abstract: This paper describes an approach to detect and classify vehicles in omnidirectional videos. The proposed classification method is based on the shape (silhouette) of the detected moving object obtained by background subtraction. Different from other shape based classification techniques, we exploit the information available in multiple frames of the video. The silhouettes extracted from a sequence of frames are combined to create an 'average' silhouette. This approach eliminates most of the wrong decisions which are caused by a poorly extracted silhouette from a single video frame. The vehicle types that we worked on are motorcycle, car (sedan) and van (minibus). The features extracted from the silhouettes are convexity, elongation, rectangularity, and Hu moments. The decision boundaries in the feature space are determined using a training set, whereas the performance of the proposed classification is measured with a test set. To ensure randomization, the procedure is repeated with the whole dataset split differently into training and testing samples. The results indicate that the proposed method of using average silhouettes performs better than using the silhouettes in a single frame.

1 INTRODUCTION

Omnidirectional cameras provide 360 degree horizontal field of view in a single image (vertical field of view varies). If a convex mirror is placed in front of a conventional camera for this purpose, then the imaging system is called a catadioptric omnidirectional camera (Fig. 1). Despite its enlarged view advantage, so far omnidirectional cameras have not been widely used in object detection and also in traffic applications like vehicle classification. This is mainly due to the fact that the objects are warped in omnidirectional images and techniques that are developed for standard cameras cannot be applied directly.

Object detection and classification is an important research area in surveillance applications. Quite a variety of approaches have been proposed for object detection. A major group in these studies uses the sliding window approach in which the detection task is performed via a moving and gradually growing search window. Features based on gradient directions, gradient magnitudes, colors, etc. can be used for classification. A significant performance improvement was obtained with this approach by employing HOG (Histogram of Oriented Gradients) features (Dalal, and Triggs, 2005). Later on, this technique was enhanced with part based models (Felzenszwalb et al., 2008).

In some recent studies, the sliding window approach has been applied to omnidirectional cameras as well. Cinaroglu and Bastanlar (2014) modified HOG computation for omnidirectional camera geometry. Haar-like features are also used with omnidirectional cameras (Dupuis et al., 2011; Amine Iraqui et al., 2010).

Another major group for object detection uses shape based features after background subtraction step. For instance, Morris and Trivedi (2006a, 2006b) created a feature vector consisting of area, breadth, compactness, elongation, perimeter, convex hull perimeter, length, axes of fitted ellipse, centroid and five image moments of the foreground blobs. Linear Discriminant Analysis (LDA) is used to project the data to lower dimensions. Objects are compared by weighted k-nearest neighbor classifier. Training set was made up by clustering prototype measurement vectors with fuzzy-C means algorithm.

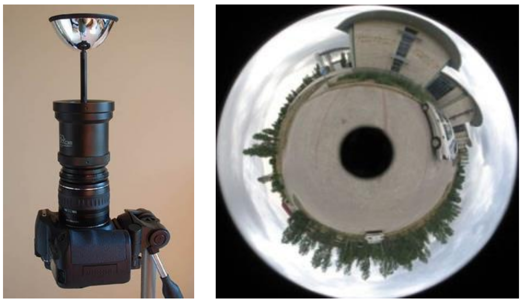


Figure 1: (a) A mirror apparatus is placed in front of a conventional camera to obtain a catadioptric omnidirectional camera. (b) An example image obtained by such a camera.

These two major approaches are compared in a study by Morris and Trivedi (2006b). HOG or Haar-like features are named as image based features and the features of the shape based approach are called image measurement based features. It was stated that using simple measurements extracted from the shapes is computationally cheaper. Extracting image based features for each position of sliding window requires a considerable amount of time. Also the storage requirement is much less with shape features. Regarding omnidirectional images, an extra load of converting original image to panoramic image (or conversion of features) is required. To decrease the computational load for image based features approach, one can extract features only for the region where the moving object exists. Even in that case, fitting a single window to the object is not possible. To give an example, in the study of Ghandi and Trivedi (2007), where HOG features are computed on virtual perspective views generated from omnidirectional images, the windows are located manually. These facts make the image based features unsuitable for real-time applications in most cases.

We are also able to compare the performances of the mentioned two approaches on standard images. The accuracy of the HOG based method, by Ghandi and Trivedi (2007), is lower than the accuracy of shape based classification in their previous work (Morris and Trivedi, 2006a). The classification accuracy is 64.3% for HOG based approach (accuracy is 34/36 for sedan, 17/34 for minivan and 5/17for pickup) and 88.4% for shape based approach (accuracy is 94% for sedan, 87% for truck, 75% for SUV, 100% for semi, 90% for van, 0% for TSV and 85% for MT).

Motivated by the facts given above, we decided to develop a shape based method for omnidirectional cameras. Before giving the details of our method, let us briefly present more related work on shape based methods for vehicle classification.

In one of the earliest studies on vehicle classification with shape based features, authors first apply adaptive background subtraction on the image to obtain foreground objects (Gupte et al., 2002). Location, length, width and velocity of vehicle fragments are used to classify vehicles into two categories; cars and non-cars. In another study, (Kumar et al., 2005), authors use position and velocity in 2D, the major and minor axis of the ellipse modelling the target and the aspect ratio of the ellipse as features in a Bayesian Network.

In a 3-D vehicle detection and classification study which is based on shape based features, Buch et al. (2008) use the overlap of the object silhouette with region of interest mask which corresponds to the region occupied by the projection of the 3D object model on the image plane. Although features like area, convex area, bounding ellipse axes or bounding box size are not used, the accuracy of the method is high.

In a ship classification study, researchers use MPEG-7 region-based shape descriptor which applies a complex angular radial transform to a shape represented by a binary image and classified ships to 6 types with k-nearest neighbor algorithm (Luo et al., 2006).

Instead of standard video frames, some researchers employed time-spatial images, which are formed by using a virtual detection line in a video sequence. Rashid et al. (2010) construct a feature vector obtained from the foreground mask. Employed features are width, area, compactness, length-width ratio, major and minor axis ratio of fitted ellipse, rectangularity and solidity. The training set is clustered in desired number of vehicle classes by fuzzy C-means algorithm. The samples are classified by k-nearest neighbor algorithm. Later, they improved their work using multiple time spatial images (Mithun et al., 2012).

Although not applied to vehicle classification, a radically different method that uses silhouettes was proposed by (Dedeoglu et al., 2006). They define ‘silhouette distance signal’ which is the sum of distances between center of a silhouette and contour points. They create a database of sample object silhouettes with manually labelling object types. An object is classified by comparing its silhouette distance signal with the ones in the template database.

Regarding the shape based classification studies with omnidirectional cameras, the only work that we found in the literature (Khoshabeh et al., 2007) uses only the area of the blobs and classifies them into two classes; small and large vehicles. In our study, we detect each vehicle type separately using a higher number of features.

The main contribution in our study can be considered as exploiting the information available in multiple frames of the video. The silhouettes extracted from a sequence of frames are combined to create an ‘average silhouette’. This process is known as ‘temporal averaging of images’ in image processing community and usually used to eliminate noise. To our knowledge, the proposed method is the first that combines several silhouettes for object detection/classification.

Another contribution in this paper is that we use a portable image acquisition platform which is more practical than fixing the cameras to building facades. Previous work, that employ cameras fixed to buildings, use “area” as a feature to classify vehicles (Morris and Trivedi (2006a, 2006b), Khoshabeh et al. (2007), Buch et al. (2008), Rashid et al. (2010)). Since that feature becomes invalid when the distance between the camera and the scene objects change, those methods are not versatile. As a consequence, in our method area of the silhouette is not a feature.

The vehicle types that we worked on are motorcycle, car (sedan) and van (minibus). The features extracted from the silhouettes are convexity, elongation, rectangularity, and Hu moments. The convexity is used to eliminate poor silhouette extraction, the elongation is used to distinguish motorcycles from other vehicles, and the remaining two features (rectangularity and a distance based on Hu moments) are used for labelling an object as a car or a van. The decision boundary is obtained by applying Support Vector Machines (SVM) on the training dataset. The performance of the proposed approach is compared with the results of using silhouettes in a single frame. Using the average silhouette rather than using a single frame (not averaging) improved the rate of correct classification from 80% to 95% for motorcycle, from 78% to 98% for car, and from 81% to 83% for van.

Our omnidirectional video dataset, together with annotations and binary videos after background subtraction, can be downloaded from our website (http://cvrg.iyte.edu.tr/). The organization of the paper is as follows. In Section 2, we introduce the details of silhouette averaging process. Vehicle detector is described in Section 3 and classifier is presented in Section 4. Experiments, given in Section 5, demonstrate that the proposed method of averaging silhouettes outperforms using a single silhouette. Conclusions are given in Section 6.

2 SILHOUETTE AVERAGING

The silhouettes are obtained after a background subtraction step and a morphological operation step. For background subtraction, the algorithm proposed by Yao and Odobez (2007) is used, which was one of the best performing algorithms in the review of Sobral and Vacavant (2014). The final binary mask is obtained by an opening operation with a disk, after which the largest blob is assigned as the silhouette belong to the moving object.

To obtain an ‘average silhouette’ we need to define how many frames are used and the silhouettes from these frames should coincide spatially. If a silhouette is in range of a previously specified angle (which we set as [30°,-30°], and 0° is assigned to the direction that camera is closest to the road), then the silhouette is rotated with respect to the center of omnidirectional image so that the center of the silhouette is at the level of the image center. This operation, also described in Figure 2, is repeated until the object leaves the angle range.

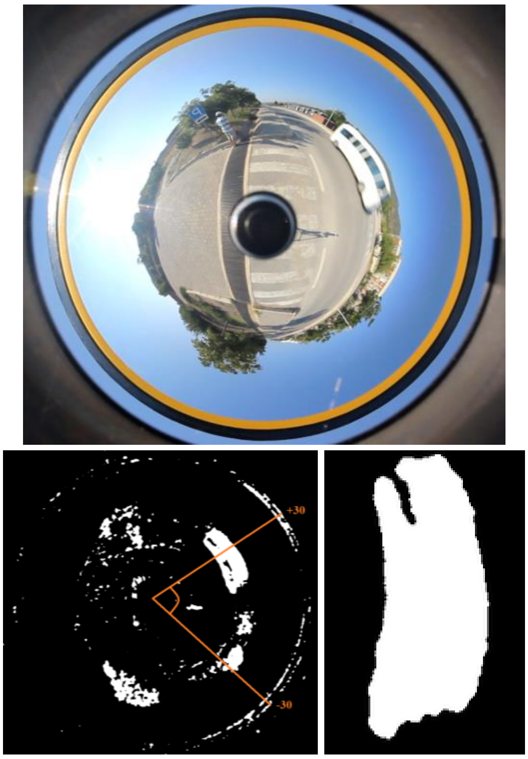


Figure 2: Top: An example omnidirectional video frame containing a van while passing a road. Bottom-left: The same frame after background subtraction. Also the angle range that we used, namely [30°,-30°], is superimposed on the image. Centroid of the largest blob is at 29°. Bottom-right: Rotated blob after morphological operations.

Silhouettes obtained in the previous step are added to each other so that the center of gravity of each blob coincides with others. The cumulative image is divided by the number of frames which results in 'average silhouette' (Figure 3). We then apply an intensity threshold to convert average silhouette to a binary image and also to eliminate less significant parts which were supported by a lower number of frames. Thus we can work with more common part rather than taking into account every detail around a silhouette. The threshold we select here eliminates the lowest 25% of grayscale levels.

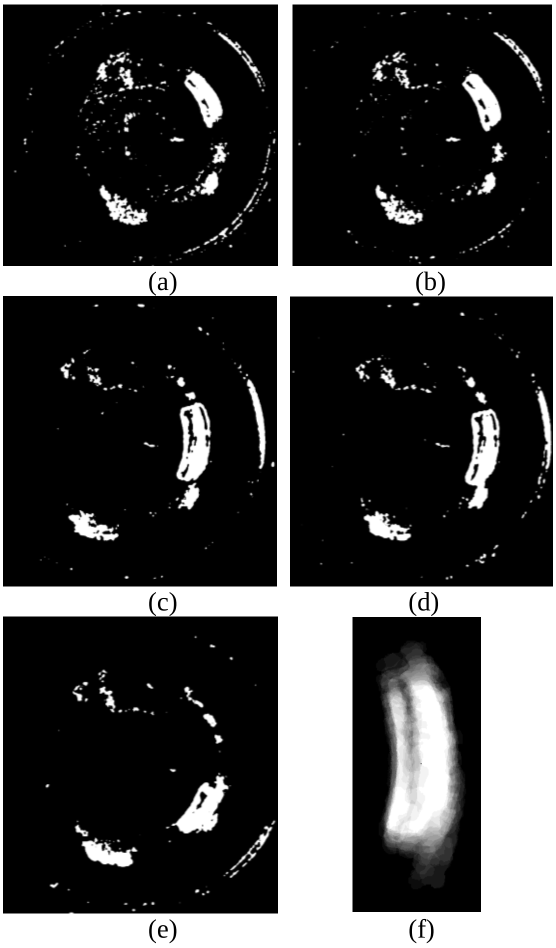

(a) (b) (c) (d) (e) (f)

Figure 3: Example binary images when the centroid of the object is at (a) 29° (b) 26° (c) 0° (d) -11° (e) -29°. (f) Resultant 'average silhouette' obtained by the largest blobs in the binary images.

3 DETECTION

The convexity (1) is used to eliminate detections that may not belong to a vehicle class or poorly extracted silhouettes from vehicles.

$$Convexity = \frac{O_{Convexhull}}{O} \tag{1}$$

where $O_{Convexhull}$ is the perimeter of the convex hull and O is the perimeter of the original contour (Yang et al., 2008). Since we do not look for a jagged silhouette, the set of detected silhouettes $\{D_s\}$ is filtered to obtain a set of valid detections $\{D_v\}$ (2) using the convexity threshold ρ.

$$\{D_v\} = \{D_s | Convexity_{D_s} < \rho\} \tag{2}$$

We set $\rho = 0.75$ for our experiments. The set of valid detections $\{D_v\}$ is passed to the classification step. An example is shown for an eliminated silhouette using convexity threshold in Figure 4.

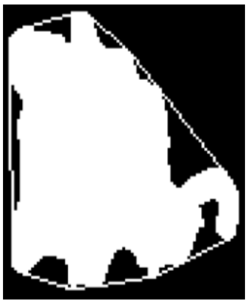


Figure 4: An example of an extracted silhouette and its convex hull. It is extracted from a motorcycle example using a single frame and its convexity is computed as 0.73 which is lower than the threshold $\rho = 0.75$.

Block diagram in Figure 5 summarizes the detection step together with the classification step which is described in Section 4. Please note that with the proposed multi-frame method, morphological operations are carried out for multiple frames and thresholded average silhouette is given as an input to the detection and classification steps. For the single frame method, however, the silhouette from the frame where the object is closest to 0° is used.

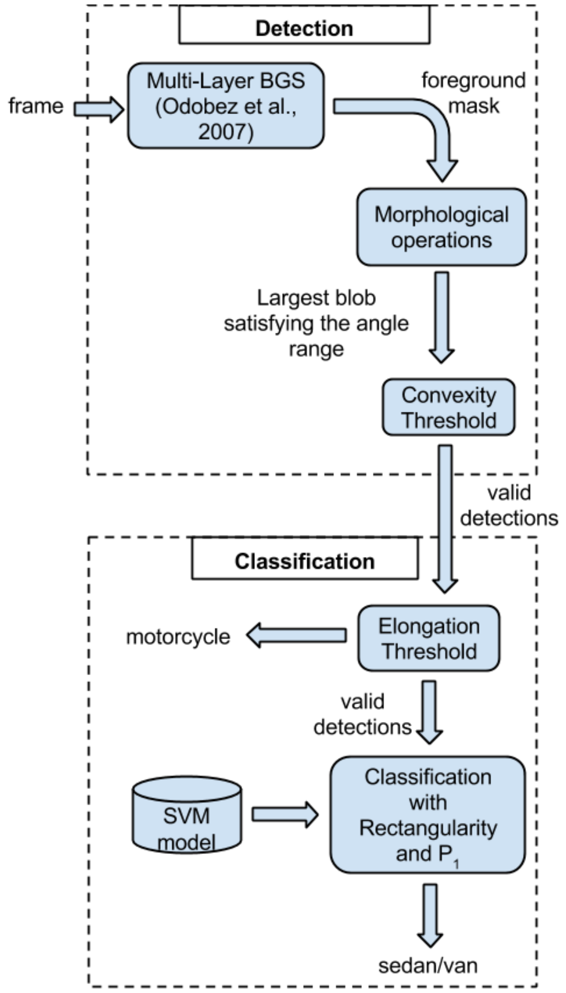


Figure 5: Block diagram of the detection and classification system. With the proposed method, multiple frames are processed and the extracted average silhouette is used instead of a silhouette from a single frame.

4 CLASSIFICATION

Next, the valid detections determined by the detection step are classified (cf. Figure 5). The features we employ for classification are; elongation, rectangularity, and Hu moments. Elongation (3) is computed as follows

$$Elongation = 1 - W/L \qquad (3)$$

where W is the short and L is the long edge of the minimum bounding rectangle (Figure 6) which is the smallest rectangle that contains every point in the shape (Yang et al., 2008).

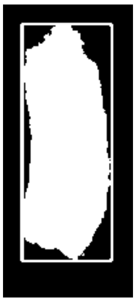


Figure 6: Thresholded silhouette and the minimum bounding rectangle.

Rectangularity (4) measures how much a shape fills its minimum bounding rectangle (Yang et al., 2008):

$$Rectangularity = A_S / A_L \qquad (4)$$

where A_S represents area of a shape and A_L represents area of the bounding rectangle.

We observed that the elongation is able to discriminate motorcycles from other vehicle types with a threshold. Then, the set of detected motorcycles $\{D_m\}$ (5) is given by

$$\{D_m\} \ = \{D_m | Elongation_{D_v} < \tau\} \qquad (5)$$

where τ is the elongation threshold. τ is determined using the samples in the training set.

Rectangularity is a meaningful feature to distinguish between sedan cars and vans since the silhouette of a van has a tendency to fill its minimum bounding box. In our trials, however, we observed that setting a threshold for rectangularity alone is not effective enough to discriminate cars from vans. To discriminate the cars and vans better, we defined an extra feature, named P_1 (8), which is based on Hu moments and measures if an extracted silhouette resembles the car silhouettes in the training set more than it resembles the van silhouettes. P_1 is an exemplar-based feature rather than a rule-based one and it is computed as follows:

$$C_1 = \frac{1}{\#cars} \sum_{i=0}^{\#cars} I_2(\, D_s, Car_i) \qquad (6)$$

$$V_1 = \frac{1}{\#vans} \sum_{i=0}^{\#vans} I_2(\, D_s, Van_i) \qquad (7)$$

$$P_1 = C_1 - V_1 \qquad (8)$$

For a new sample, P_1 corresponds to the difference between the average I_2 (10) distance to the cars in the training set and the average I_2 distance to the vans in the training set. The mentioned I_2 distance is one of the three possible distances, based on 7 Hu moments (Hu, 1962), used for computing the similarity of two silhouettes:

$$I_1\,(A,B)= \sum_{i=1\ldots7}\left|\frac{1}{m_i^A}-\frac{1}{m_i^B}\right| \tag{9}$$

$$I_2\,(A,B)= \sum_{i=1\ldots7}\left|m_i^A-m_i^B\right| \tag{10}$$

$$I_3\,(A,B)= \sum_{i=1\ldots7}\left|\frac{m_i^A-m_i^B}{m_i^A}\right| \tag{11}$$

$$m_i^A = sign(h_i^A)\cdot\log h_i^A \tag{12}$$

$$m_i^B = sign(h_i^B)\cdot\log h_i^B \tag{13}$$

where h_i^A and h_i^B are the Hu moments of shapes A and B respectively (Bradski and Kaehler, 2008).

We select I_2 since it achieved better discrimination in our experiments than I_1(9) and I_3(11).

If a detection is not classified as a motorcycle, in other words Elongation $> \tau$, then it can be either a car or a van. To determine the decision boundary between car and van classes we trained a SVM with linear kernel. The boundaries obtained using the training set are depicted in the following section.

5 EXPERIMENTS

Using a Canon 600D SLR camera and a mirror apparatus (www.gopano.com) we obtained a catadioptric omnidirectional camera. We constructed a dataset of 49 motorcycles, 124 cars and 104 vans totaling 277 vehicle instances. Dataset is divided into training and test sets. Training set contains approximately 60% percent of the total dataset corresponding to 29 motorcycles, 74 cars and 62 vans. The rest is used as test set.

We set $\rho = 0.75$ and SVM's parameter $C = 0.2$ for our training set. The elongation threshold is determined by choosing the maximum convexity value of motorcycles in the training set since this value discriminates motorcycles from other vehicles.

Regarding the training of car-van classifier, Figures 7 and 9 show the SVM's linear decision boundary, trained with the average silhouette and single frame silhouette respectively. Training the single frame method with the extracted single frame silhouettes would not be fair since they contain poorly extracted silhouettes. Therefore, the boundaries of the vehicles are manually annotated and used for the training of single frame method. Test results with and without averaging silhouettes are shown in Figures 8 and 10 respectively.

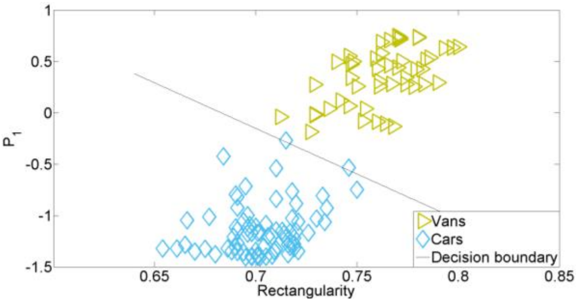


Figure 7: Training result of SVM using the average silhouette method.

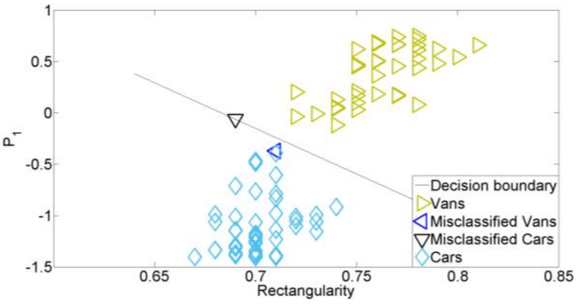


Figure 8: Test result with the average silhouette method.

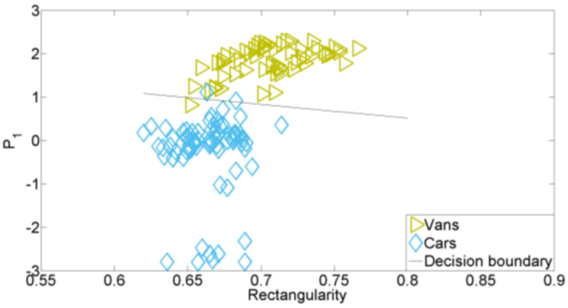


Figure 9: Training result of SVM without averaging silhouettes (single frame method).

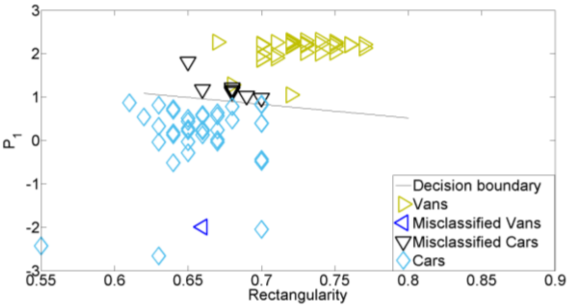


Figure 10: Test result without averaging silhouettes, i.e. using single frame silhouettes.

Table 1: Average classification accuracies for each class when $\rho = 0.75$ and $C = 0.2$ for the average silhouette method and for the single frame method.

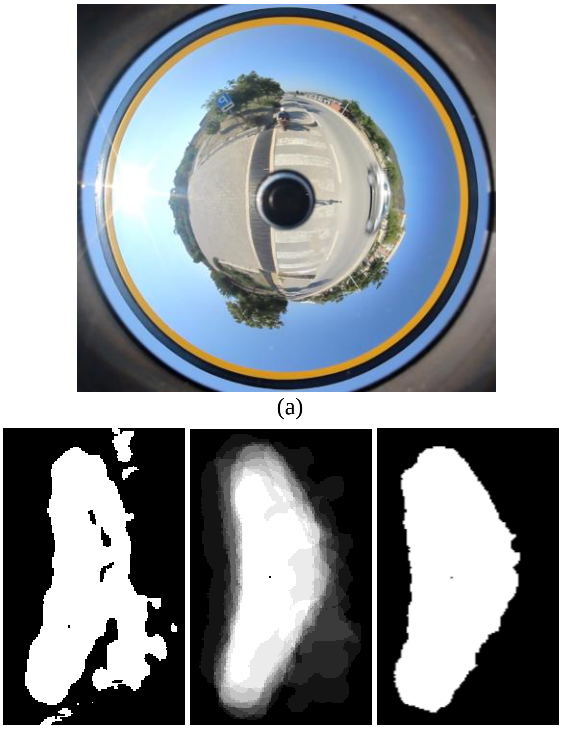
	Motorcycle	Car	Van	Overall
Average silhouette method	95%	98%	83%	92%
Single frame method	80%	78%	81%	79%

Table 2: Confusion matrix for the proposed method of using average silhouettes.

Ground truth		Motorcycle	Car	Van
Detection	Motorcycle	19	0	0
	Car	0	49	1
	Van	1	1	35
	FN	0	0	6

Table 3: Confusion matrix for single frame method.

Ground truth		Motorcycle	Car	Van
Detection	Motorcycle	16	3	4
	Car	0	39	1
	Van	1	7	34
	FN	3	1	3


(a)

(b) (c) (d)

Figure 11: Example car silhouettes (a) original frame, (b) result of using a single silhouette which is misclassified with rectangularity = 0.56 and $P_1 = 3.381$, (c) average silhouette, (d) thresholded average silhouette classified as car rectangularity = 0.68 and $P_1 = -1.602$.

To ensure the randomization of data samples, the procedure is repeated three times with the dataset split randomly into training and testing samples. We report the average results of the two compared methods in Table 1. Values in the table correspond to what percentage of the instances of a vehicle type is classified correctly. Not surprisingly, exploiting the information from multiple frames by averaging the silhouettes has a greater performance than using the silhouette in a single frame.

Tables 2 and 3 depict the number of correctly classified and misclassified samples for each class with the average silhouette and single frame silhouette methods respectively. False negatives are missed samples which are eliminated by convexity threshold ρ , i.e. non-valid detections.

Figure 11 shows an example where a car is correctly classified with using average silhouette, whereas it is misclassified with using a single silhouette. Figure 12 shows an example where a van has passed the detection phase with average silhouette method but failed with the single frame method. Such cases constitute the main performance difference between the two compared methods.

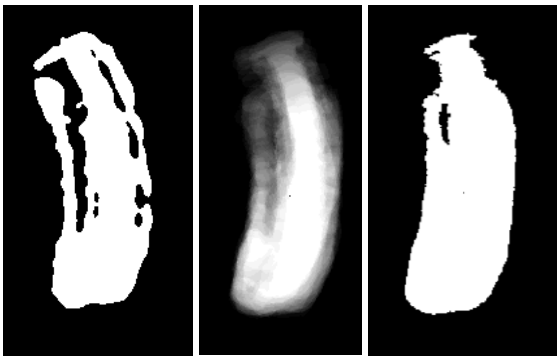

(a) (b) (c)

Figure 12: Example van silhouettes (a) silhouette from a single frame which is eliminated since $\rho = 0.548$.(b) average silhouette (c) thresholded average silhouette which is not eliminated since.$\rho = 0.823$.

6 CONCLUSIONS

We proposed a method for vehicle detection and classification based on a set of features extracted from object silhouettes. We applied our method by using a silhouette from a single frame and also by using temporal average of silhouettes in multiple frames. Our hypothesis was that the classification with average silhouettes of multiple frames is more successful than using a silhouette from a single frame. Results of the experiments indicate a significant improvement in classification performance using multiple frames.

Although we applied the proposed method for vehicles, in essence the advantage of averaging silhouettes is utilizing the information available in a longer time interval rather than a single frame. Therefore the improvement can be expected for other objects types and domains other than traffic applications.

We use a portable image acquisition platform and our method is independent of the distance between the camera and the objects which is more practical than the previously proposed methods that fix the cameras to buildings and use the object's area as a feature since the distance to objects stays same.

ACKNOWLEDGEMENTS

This work was supported in part by the TUBITAK project 113E107.

REFERENCES

Amine Iraqui, H., Dupuis, Y., Boutteau, R., Ertaud, J., and Savatier, X. (2010). Fusion of omnidirectional and ptz cameras for face detection and tracking. In *Emerging. Security Technologies (EST), 2010 International Conference on*, pages 18–23.

Bradski, G. and Kaehler, A. (2008). *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media.

Buch, N., Orwell, J., and Velastin, S. (2008). Detection and classification of vehicles for urban traffic scenes. In *Visual Information Engineering, 2008. VIE 2008. 5th International Conference on*, pages 182–187.

Cinaroglu, I. and Bastanlar, Y. (2014). A direct approach for human detection with catadioptric omnidirectional cameras. *22nd Signal Processing and Communications Applications Conference (SIU)*, pages 2275–2279.

Dalal, N. and Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection, *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR).

Dedeoglu, Y., Toreyin, B., Gudukbay, U., and Cetin, E. (2006). Silhouette-based method for object classification and human action recognition in video. In *Computer Vision in Human-Computer Interaction*, of *Lecture Notes in Computer Science*, vol. 3979 p.64–77.

Dupuis, Y., Savatier, X., Ertaud, J., and Vasseur, P. (2011). A direct approach for face detection on omnidirectional images. In *Robotic and Sensors Environments (ROSE), 2011 IEEE International Symposium on*, pages 243–248.

Felzenszwalb, P., McAllester, D. and Ramanan, D. (2008). A Discriminatively Trained, Multiscale, Deformable Part Model, *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR).

Gandhi, T. and Trivedi, M. (2007). Video based surround vehicle detection, classification and logging from moving platforms: Issues and approaches. In *IEEE Intelligent Vehicles Symposium*, pages 1067–1071.

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

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8(2):179–187.

Khoshabeh, R., Gandhi, T., and Trivedi, M. (2007). Multicamera based traffic flow characterization and classification. In *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE*, pages 259–264.

Kumar, P., Ranganath, S., Weimin, H., and Sengupta, K. (2005). Framework for real-time behavior interpretation from traffic video. *Intelligent Transportation Systems, IEEE Trans. on*, 6(1):43–53.

Luo, Q., Khoshgoftaar, T., and Folleco, A. (2006). Classification of ships in surveillance video. In *Information Reuse and Integration, 2006 IEEE International Conference on*, pages 432–437.

Mithun, N., Rashid, N., and Rahman, S. (2012). Detection and classification of vehicles from video using multiple time-spatial images. *Intelligent Transportation Systems, IEEE Transactions on*, 13(3):1215–1225.

Morris, B. and Trivedi, M. (2006a). Improved vehicle classification in long traffic video by cooperating tracker and classifier modules. In *Video and Signal Based Surveillance (AVSS), 2006. IEEE International Conference on*, pages 9–9.

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

Rashid, N., Mithun, N., Joy, B., and Rahman, S. (2010). Detection and classification of vehicles from a video using time-spatial image. In *Electrical and Computer Engineering (ICECE), 2010 International Conference on*, pages 502–505.

Sobral, A. and Vacavant, A. (2014). A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. *Computer Vision and Image Understanding*, 122(0):4 – 21.

Yang, M., Kpalma, K., Ronsin, J., et al. (2008). A survey of shape feature extraction techniques. *Pattern recognition*, pages 43–90.

Yao, J. and Odobez, J. (2007). Multi-layer background subtraction based on color and texture. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007*. CVPR'07, pages 1–8.