REAL-TIME MOVING OBJECT CLASSIFICATION WITH AUTOMATIC SCENE DIVISION

Zhaoxiang Zhang, Yinghao Cai, Kaiqi Huang and Tieniu Tan

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{zxzhang, yhcai, kqhuang, tnt}@nlpr.ia.ac.cn

ABSTRACT

We address the problem of moving object classification. Our aim is to classify moving objects of traffic scene videos into pedestrians, bicycles and vehicles. Instead of supervised learning and manual labeling of large training samples, our classifiers are initialized and refined online automatically. With efficient features extracted and organized, the approach can be real-time and achieve high classification accuracy. Once the view or scene changes detected, the algorithm can automatically refine the classifiers and adapt them to new environments. Experimental results demonstrate the effectiveness and robustness of the proposed approach.

Index Terms— Surveillance, Pattern classification, Object recognition, Motion detection, Video signal processing

1. INTRODUCTION

Moving object classification from videos is an important issue in image processing and video analysis. With object type information known, more specific and accurate methods can be developed to recognize high level actions of video objects. Especially in traffic scene videos, classification of moving objects into predefined categories allows the operator to program the monitoring system by specifying events of interest, such as 'alarming when a pedestrian is coming into a forbidden area' or 'alarming when a vehicle is running in a reverse direction', which is very common in smart video surveillance.

Much work has been done in the field of moving object classification in traffic scene videos. In [1, 2], foreground objects are detected using motion information and certain image features, like area, compactness, speed and bounding box aspect ratio are extracted for training and classification. However, most of these features are based on 2D space and cannot avoid projective distortion, which is much more significant in far-field traffic scene videos. For example, nearby objects in images appear to be larger and move faster than those far away. Therefore, simply using these features for classification is unsuitable and limits the accuracy rate of the results. In addition, most of these algorithms can only distinguish pedestrians from vehicles and are not robust to environment changes. In [3], series of algorithms are described to demonstrate the effectiveness of local features for object detection and classification. However, most of these methods are time consuming and not applicable to low resolution videos. Viola and *etc.* give us a good framework for feature selection and object class recognition using boosting [4]. However, it is a tremendous work to collect large samples of training data in all kinds of conditions and label all of them manually. In conclusion, a convenient moving object classification algorithm should have the following desirable properties:

• $Practical$: The algorithm should be real-time and achieve high classification accuracy.
• $Robust$: The algorithm should be robust in all kinds of conditions and perform well in different environments.
• $Automatic$: The algorithm should avoid any supervised learning and manual labeling of large samples of training data.

Most of the existed algorithms only have one or two properties mentioned above and it is necessary to design an algorithm to satisfy all of the three properties. In this paper, we propose such an approach to classify moving objects of traffic scene videos into three categories: pedestrians, bicycles and vehicles. With moving objects detected using motion information, a small number of 2D features are extracted so that the approach can achieve real-time and be suitable for real applications. Using a novel subregion based strategy, we conquer the disadvantages of significant projective distortion of conventional 2D features and improve the performance very much. A novel framework is designed to achieve automatic classification by avoiding supervised learning and manual labeling of training data. In addition, the approach can deal with scene changes and be effective in different conditions.

The remainder of the paper is organized as follows. In Section 2, we introduce the method for motion detection with shadows removed. Classification framework is described in Section 3. Experimental results and analysis are presented in Section 4. Finally, we draw our conclusions in Section 5.

2. MOTION DETECTION

As the focus of the paper is moving object classification, we employ motion information to detect targets of interest in videos. Gaussian Mixture Model (GMM) [5] is a popular method

for motion detection because it can deal with slow lighting changes, periodical motions in clutter background, slow moving objects and long term scene changes. However, this method cannot deal with fast illumination changes and shadows very well. As we know, fast illumination changes are very common in video surveillance and shadows affect the accuracy of detection results greatly. We adopt the method described in [6] to deal with the disadvantages mentioned above and the method can be summarized as follows:
(1) The intensity of each pixel is modeled as the product of irradiance component and reflectance component.
(2) The reflectance value of each pixel is modeled as a mixture of Gaussian.
(3) Every new pixel is matched against each of the existed Gaussian distribution. A match is defined as a pixel value within 2.5 standard deviations of a distribution.
(4) Sort the Gaussians and determine whether it is background.
(5) Adjust the Gaussians and their prior weights.
(6) If there is no match, replace the least probable Gaussian and set mask pixel to background.

Experimental results of background modeling and motion detection are shown in Fig. 1. As we can see, foreground objects are detected accurately and cast shadows are removed.

(a) One frame of videos (b) Background recovered (c) Detected moving objects

Fig. 1. Motion detection results with shadows removed

3. CLASSIFICATION

A new framework for classification of moving objects in traffic scenes is described in detail in this section. The flowchart of the approach is shown in Fig. 2.

3.1. Scene Division for Subregion Strategy

2D features are efficient for classification in near-field videos, which have been demonstrated in [1, 2]. However, these features are based on image plane and cannot avoid projective distortion so that they are not applicable in far-field videos. A subregion based strategy is applied here to conquer this disadvantage. By dividing the scene of the far-field video into many equal parts, distortion of 2D features can be ignored in each part because one part just covers a narrow field of view. So classification can be realized in each part by using efficient 2D features. The more parts the scene is divided into,

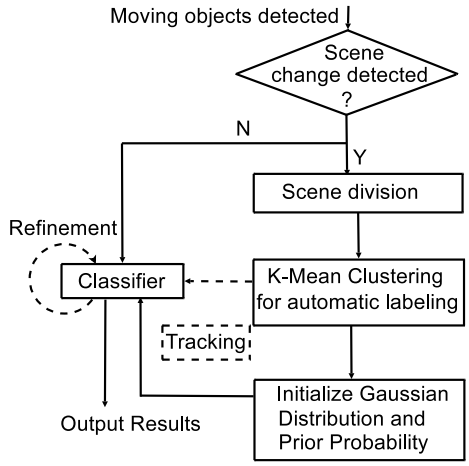


Fig. 2. Flowchart of the classification algorithm

the more effective 2D features will be. However, since small regions need more time to collect enough samples for online learning and cause maintenance of large number of classifiers, there is a trade-off between accuracy and computation. In our approach, the average size of moving objects is estimated and compared with the size of the scene, which is adopted as a criteria for the number of subregions. Taking the scene shown in Fig. 1 as an example, we divide the scene into 16 parts as 4×4 grids, which correspond to 16 classifiers initialized and maintained online later.

3.2. K-Mean Clustering for Automatic Labeling

In order to classify moving objects automatically without supervised learning, we adopt K-Mean clustering and decision level fusion for automatic labeling. There are totally 5 shape and motion features used in our algorithm:
• $area$: size of objects in pixels
• $velocity$: time derivative of centroid of the object
• $compactness$: equals to $\frac{area}{perimeter^2}$
• $area'$: time derivative of $area$
• $angle$: angle between motion direction and direction of major axis of the silhouette
As we know, $area$ and $velocity$ have the most significant projective distortion among all the 5 features. So, we use the additional 3 features for K-Mean clustering and automatic labeling. After videos processed frame by frame for a period of time, K-Mean clustering is adopted to establish 3 clusters. One cluster corresponds to one category, respectively. The decision level fusion based on the following three intuitive rules is adopted to establish the correspondence: (1) $compactness$ has the advantages of distinguishing vehicles from pedestrians and bicycles. (2) $area'$ has the advantages of distinguishing pedestrians from vehicles and bicycles. (3)$angle$ has the advantages of classifying pedestrians and vehicles. Using voting strategy, we can conveniently achieve automatic labeling.

3.3. Gaussian Assumption for Classification

The approach is based on the Gaussian Assumption that in each grid $v = (area, speed, compactness)$ of every category satisfies a multivariate Gaussian distribution, which is tested valid with appropriate subregion size and denoted as:

$$P_i(v) = \eta(v, \mu_i, \Sigma_i) \quad i = 1, 2, 3 \tag{1}$$

Using Bayesian rules, we obtain the derivation as follows:

$$P(category = i|v) \propto P_i(v) \cdot p_i \quad i = 1, 2, 3 \tag{2}$$

where $P(category = i|v)$ and p_i are posterior and prior probability of each category, respectively.

Classification is realized in each subregion after automatic labeling. If the number of moving objects passing the subregion has not reached a threshold N, classification is simply realized by comparing the distance between the feature vector and every cluster. If it reaches N, the classifier of this subregion is initialized. The prior probability is determined by the number of individuals belonging to each cluster and the Gaussian distribution is estimated in the following way:

$$\hat{\mu}_i = \frac{1}{N} \sum_{r=1}^{N} v_{i,r} \quad i = 1, 2, 3 \tag{3}$$

$$\hat{\sigma}^2_{ij} = \frac{1}{N} \sum_{r=1}^{N} (v_{i,r} - \hat{\mu}_i)(v_{j,r} - \hat{\mu}_j) \quad i, j = 1, 2, 3 \tag{4}$$

The category is determined by the posterior probability and the classifier is refined at the same time to be robust to condition changes:

$$p_{k,new} = (1 - \beta)p_{k,old} + \beta(M_{k,t}) \quad k = 1, 2, 3 \tag{5}$$

$$\hat{\mu}_{new} = (1 - \gamma)\hat{\mu}_{old} + \gamma v_t \tag{6}$$

$$\hat{\sigma}^2_{i,j,new} = (1 - \gamma)\hat{\sigma}^2_{i,j,old} + \gamma(v_{i,t} - \hat{\mu}_t)(v_{j,t} - \hat{\mu}_t) \tag{7}$$

where β and γ are the refinement rate. $M(k, t)$ is 1 if v_t is classified as the category k and 0 otherwise. The prior probability is renormalized after that. In every frame, there is posterior probability output for every moving objects. We can determine the category using the sum of posterior probability of tracked frames to improve robustness of classification.

3.4. Scene Change Detection

Most of scene changes in video surveillance are abrupt transitions caused by zooming or moving of cameras rather than gradual transitions. We can simply detect scene changes when

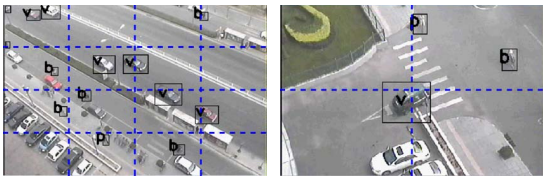
$$\Sigma_{x,y}(|B_t(x, y) - B_{t-1}(x, y)|) > T \tag{8}$$

where T is a threshold and B_t and B_{t-1} are recovered background of the current and previous frames, respectively. As we use reflectance component for background modeling, the detection is robust to fast illumination changes. When scene changes are detected, subregions will be redivided according to the new situation and the classifiers will be refined online effectively.

4. EXPERIMENTAL RESULTS AND ANALYSIS

Experiments are conducted to demonstrate the performance of the proposed algorithm. The videos are captured with the size of 320×240 using the Panasonic WV-CW860A common CCD camera and all the experiments are carried out on computers of P4 3.0G CPU and 512M DDR.

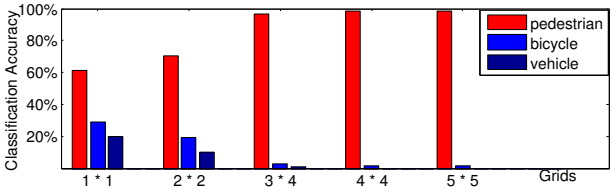
Two scenes are shown in Fig. 3 and classification results of one frame are given. Here, 'p' represents pedestrians, 'b' represents bicycles and 'v' represents vehicles. The field of scene of Video 1 is much farther than that of Video 2, which leads to different number of subregions in these two scenes.



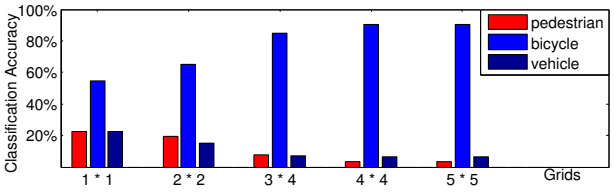
(a) Video 1 (b) Video 2

Fig. 3. The results of classification in two scenes

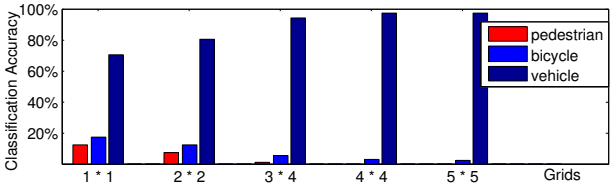
Different size of subregions effects the classification accuracy. Experiments are conducted with different number of grids in Video 1 and the accuracy is shown in Fig. 4. As we can see, the accuracy increase with smaller size of subregions, which demonstrates the advantage of the subregion strategy.



(a) Classification accuracy of the category of pedestrian



(b) Classification accuracy of the category of bicycle



(c) Classification accuracy of the category of vehicle

Fig. 4. Different accuracy with different extent of division

We divide the scene of Video 1 into 4×4 grids and the

classification accuracy using our approach is shown in Table 1. As we know, with clustering center known, we can simply realize the classification by comparing the distance between the feature vector and each of the cluster. The results using this simple strategy are shown in Table 2. It is evident that our approach is much better. That is because we use the prior information of the Gaussian distribution. Comparison results demonstrate the validity of the Gaussian Assumption. But with inappropriate scene division as shown in Fig. 4, the accuracy is even worse than clustering center based strategy. That is because the Gaussian Assumption is not satisfied in inappropriate region size and decreases the accuracy.

Table 1. Classification accuracy using Gaussian Assumption

	Pedestrians	Bicycles	Vehicles
Pedestrians	98.2%	1.8%	0.0%
Bicycles	3.4%	90.4%	6.2%
Vehicles	0.0 %	2.7%	97.3 %

Table 2. Classification accuracy using distance from cluster

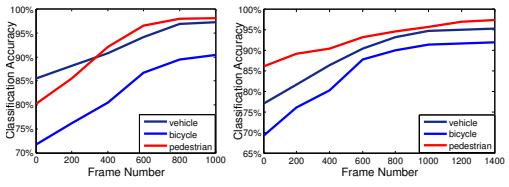
	Pedestrians	Bicycles	Vehicles
Pedestrians	87.2%	9.3%	3.5%
Bicycles	11.3%	79.4%	9.3%
Vehicles	8.4%	9.5%	82.1%

In our approach, initialization and refinement of classifiers are carried out online. With classifiers refined more accurately, the algorithm can achieve higher classification accuracy. The change trend of classification accuracy of Video 1 and Video 2 are shown in Fig. 5(a, b). We also test the performance of the approach dealing with environment changes. A new video connected by Video 1 and Video 2 is processed and the change trend of accuracy is shown in Fig. 5(c). As we can see, the algorithm can detect the scene change and classifiers can be refined to adapt the new scene.

From all of the above, we can see our algorithm is real-time, effective without manual labeling and supervised learning, and can deal with environment changes very well, which can be conveniently applied to many real systems.

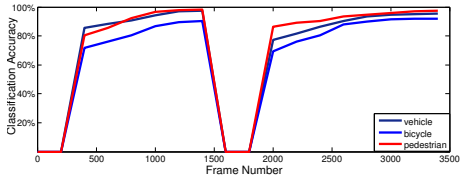
5. CONCLUSIONS

In this paper, we have proposed an approach for classification of moving objects in traffic scenes. With the subregion strategy and Gaussian Assumption, 2D features are efficiently organized and the approach achieves good performance. Using a novel classification framework, classifiers are initialized and refined automatically. The algorithm is effective and robust to condition changes, which can be applied to many systems.



(a) The change trend of Video 1

(b) The change trend of Video 2



(c) The change trend of the connected video

Fig. 5. The change trend of classification accuracy

Acknowledgement

This work is funded by research grants from the National Basic Research Program of China (2004CB318110), the National Science Foundation (60605014, 60332010, 60335010 and 2004DFA06900), the CAS Graduate Student Science and Society Practice Fund (IM07N01), and the CASIA Innovation Fund for Young Scientists. The authors also thank the anonymous reviewers for their valuable comments.

6. REFERENCES

[1] Lisa M Brown, "View independent vehicle/person classification," in *Proc. of the ACM 2nd international workshop on Video Surveillance and Sensor Networks*, 2004.

[2] Quming Zhou and J.K.Aggarwal, "Tracking and classifying moving objects from video," in *Proc. of 2nd IEEE International Workshop on PETS*, 2001.

[3] Mark Everingham, Andrew Zisserman, and so on, "The 2005 pascal visual object classes challenge," *Lecture Notes in Computer Science*, 2006.

[4] Paul A. Viola, Michael J. Jones, and Daniel Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proceedings of 9th IEEE International Conference of Computer Vision*, 2003.

[5] Chris Stauffer and W.E.L Grimson, "Adaptive backgound mixture models for real-time tracking," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1999.

[6] Zhou Liu, Kaiqi Huang, and Tieniu Tan, "Cast shadow removal with gmm for surface reflectance component," in *Proceedings of 18th IEEE International Conference of Pattern Recognition*, 2006.