Boosting Local Feature Descriptors for Automatic Objects Classification in Traffic Scene Surveillance

Zhaoxiang Zhang, Min Li, Kaiqi Huang and Tieniu Tan
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{*zxzhang, mli, kqhuang, tnt*}*@nlpr.ia.ac.cn*

Abstract

We address the problem of automatic object classification for traffic scene surveillance, which is very challenging for the low resolution videos, large intra-class variations and real-time requirement. In this paper, we propose a new strategy for object classification by boosting different local feature descriptors in motion blobs. We not only evaluate the performance of each local feature descriptor, but also fuse these descriptors to achieve better performance. Numerous experiments are conducted and experimental results demonstrate the effectiveness and efficiency of our approach with robustness to noise and variance of view angles, lighting conditions and environments.

1 Introduction

Automatic object classification in videos is an important issue in the field of traffic scene surveillance with great potential for real applications. However, it is also very challenging for the following aspects. Firstly, regions of interest are of limited size and low resolution due to capacity of conventional surveillance cameras. Secondly, intra-class variance for every category is very large due to different view angles, lighting conditions and environments. Further more, object classification always require real-time performance for applications.

Since its importance and difficulty, much work has been done in this field. In [3, 11], foreground objects are detected using motion information and certain image features, like area, compactness and speed are extracted for training and classification. However, most of these features are based on 2D image plane and cannot avoid projective distortion, which is much more significant in far-field traffic scene videos. Therefore, simply using these features limits the accuracy of object classification. In [4], series of algorithms are described to demonstrate the effectiveness of texture features for object detection and classification. However, most of these methods are time-consuming and not applicable to low resolution videos. Viola et al. give us a good framework for automatic feature selection and object classification by Boosting [10], which has been successfully applied to face and pedestrian detection. In addition, generative models like HMM or Graph Models are also been used for object recognition such as [2].

Local features descriptors have developed quickly during these years. Lowe [8] proposes SIFT descriptor with invariance to scale and rotation changes, which is widely used in image matching and object recognition. Belongie et al. [1] propose shape context which is similar to SIFT but only make use of edge point information. Since there are so many kinds of local feature descriptors, Mikolajczyk et al. [9] have done research on evaluation of local feature descriptors and draw a conclusion that SIFT gives the best performance.

Most of local feature descriptors have good properties to be invariant to image transformations, distinctive for recognition and robust to noise and lighting variance. In this case, local feature descriptors should also be discriminant for object classification in surveillance videos. In this paper, we propose the method of boosting local feature descriptors for object classification. We not only evaluate the performance of each local feature descriptor, but also attempt to fuse these descriptors for better performance. Experimental results demonstrate the effectiveness, efficiency and robustness of our approach, which are described in detail as follows.

2 Data Set Acquisition

Abundant videos are collected from series of outdoor overlooking cameras which are mounted with different location and view angles at different times. The data set is then acquired from these collected videos. As we

only focus on moving objects in traffic scene surveillance, motion information is employed here to detect targets of interest in videos. In our work, we adopt the improved GMM [7] to deal with shadows and fast illumination changes. With regions of interest detected, every detected foreground region is fixed to a square according to its mass center, which is then normalized to the same size 20×20. All of these foreground targets are then labeled manually. In this way, we collect the whole data set including 58958 vehicles, 56942 pedestrians and 3297 other individuals as outliers. Samples of the whole database are illustrated in Figure 1.

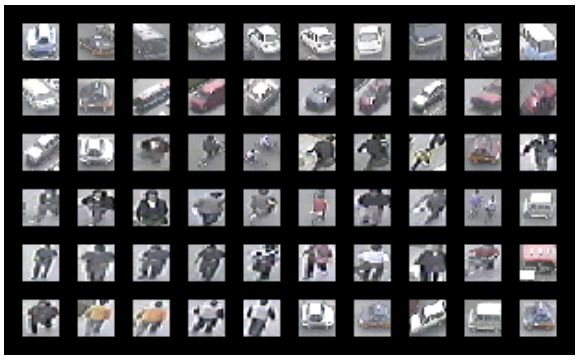


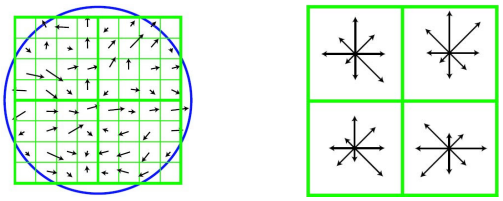
Figure 1. Samples in the database

3 Local Feature Descriptors

Three local feature descriptors: SIFT [8], Spin Image [6] and RIFT [6] are adopted for construction of distinctive feature vectors. HOG feature [6], which is simply calculated as histogram of oriented gradient in image blobs, is chosen for comparison. All of these four feature descriptors are introduced briefly as follows.

(1) **HOG descriptor**: HOG descriptor is formulated by encoding gradient information of all pixels within the motion blob into a 1D histogram. The gradient orientation is divided into N bins and every pixel contributes its gradient magnitude to the nearest two bins weighted by Euclidean distance, which is then normalized to be the N dimensional feature vector. The HOG feature descriptor is robust to illumination changes but discards all spacial information.

(2) **SIFT descriptor**: Instead of simply encoding gradient orientation as 1D histogram, SIFT [8] divided the square blob into $N \times N$ sub-blobs and the gradient orientation into 8 bins as shown in Figure 2. In this way, a 3D feature space which contains $N \times N \times 8$ sampling points can be constructed . For every pixel in the blob, its gradient magnitude is contributed to the nearest 8 sampling points in the feature space weighted by Euclidean distance. Normalized alignment of these sampling points into 1D vector forms the feature vector.

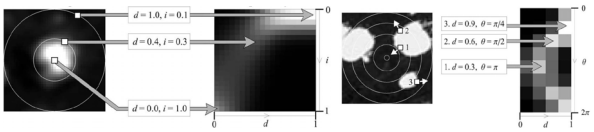


(a) Blob division (b) SIFT descriptor

Figure 2. SIFT descriptor formulation [8]

(3) **Spin Image descriptor**: Spin Image descriptor [6] is based on homocentric squares and completely invariant to rotation transformation. As shown in Figure 3(a), the intensity domain Spin image is a two-dimensional histogram encoding the distribution of intensity value i and distance d from the reference point.

(4) **RIFT descriptor**: As shown in Figure 3(b), RIFT descriptor [6] divides the motion blob into concentric rings of equal width. A two dimensional histogram is formulated. One dimension is the distance from the central pixel while the other is the gradient orientation. Similarly to SIFT descriptor, gradient magnitude of every pixel is contributed to the 4 nearest sampling points in the feature space weighted by distance.



(a) Spin Image (b) RIFT

Figure 3. Formulation of Spin Image and RIFT [6]

In this section, we have introduced four kinds of local feature descriptors to formulate discriminant features for objects classification. All these four descriptors are invariant to illumination changes and robust to all kinds of translations.

4 Classification

We apply Adaboost [5] for feature selection and classifier formulation. The basic principle of this method is to combine a set of weak classifiers to form a strong classifier. During the training stage, training samples are re-weighted according to their training error. Those weak learners which are trained later are focused on the miss-classified samples with higher weights. Finally, the strong classifier is constructed by weighted combination of weak classifiers.

Adaboost also supplies a good framework for feature level fusion. With all kinds of features extracted, Adaboost can automatically select useful candidates from them to improve classification performance. The results of classification by boosting all kinds of local feature descriptors are described in detail as follows.

5 Experimental Results and Analysis

Experiments are conducted to demonstrate the performance of classification, which are all carried out on a PC computer with P4 3.0G CPU and 512M DDR.

As described in Section 2, the whole data set are labeled manually and divided randomly into two equal parts. One part is taken for training while the other is taken to test the performance of classification.

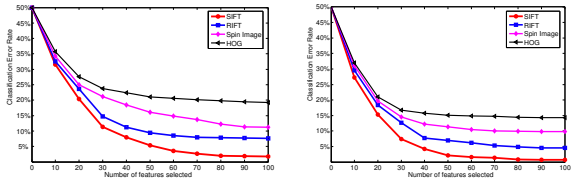
Each of the local feature descriptor is applied for boosting with classification accuracy shown in Table 1. Experimental results demonstrate the effectiveness of

Table 1. Classification accuracy

Category	HOG	SIFT	Spin Image	RIFT
Vehicle	81.5%	**98.2%**	88.7%	92.3%
People	85.7%	**99.3%**	90.2%	95.4%

blob based local feature descriptors for moving object classification. The performance of our approach is comparable to the best performance of the latest progress. Among these four local feature descriptors, SIFT gives the best performance with the highest classification accuracy. Although it is not completely rotation and affine invariant, SIFT descriptor makes use of local gradient information and describe features with spacial information. RIFT descriptor is completely rotation invariant, however, it has only spacial information in the radial direction but lose that of the tangent direction. Spin image makes use of intensity information instead of gradient information while HOG casts off all the spacial information. Comparison of these feature vectors indicates that: (1) Spacial information is quite important to improve the classification accuracy; (2) Gradient information is more distinctive than direct intensity information for object classification.

As we know, small number of features selected may not be distinctive enough for classification while large number of features decrease efficiency. Here, we adopt Boosting to evaluate the effect of feature dimensions. With all kinds of features normalized to the same dimension 100, we select N from them for classification. The curves of classification error rate due to N are shown in Figure 4(a) and 4(b). From these two curves, we can see that more than 60 features from 100 features can basically ensure the distinctiveness.



(a) Vehicle classification error rate due to feature dimension

(b) People classification error rate due to feature dimension

Figure 4. Effect of feature dimension

As described before, feature level fusion help improve the classification performance, which is tested with results shown in Table 2. As we can see, fusion of SIFT and Spin Image gives the best performance. Fusion of HOG and RIFT or SIFT has not quite significant improvement. These phenomena indicate that: (1) Fusion of gradient information based descriptor and intensity information based descriptor give the most significant improvement; (2) SIFT and RIFT has already include the total information of HOG descriptor; (3) Slight improvement of classification accuracy by fusion of the same type of features is owing to the redundancy of features.

Table 2. Classification accuracy of fusion

Vehicles	HOG+SIFT	HOG+SPIN	HOG+RIFT
Accuracy	98.3%	89.2%	92.5%
Pedestrians	HOG+SIFT	HOG+SPIN	HOG+RIFT
Accuracy	99.3%	92.2%	95.5%
Vehicles	SIFT+SPIN	SIFT+RIFT	RIFT+SPIN
Accuracy	**98.8%**	98.3%	96.2%
Pedestrians	SIFT+SPIN	SIFT+RIFT	RIFT+SPIN
Accuracy	**99.7%**	99.4%	96.9%

As we know, image gradient is related to the selected scale. For a 20×20 image blob, it is necessary to see which scale is the most distinctive one for classification. Instead of analysis in the scale space as [8], we select 5 scales corresponding to 5 σ ($\sigma = 1, 2, 3, 4, 5$) for Gaussian Smoothing. Then the algorithm is realized for every scale to see each of their contributions. In addition, we apply Adaboost to obtain fusion results of different scales. Experimental results are shown in Figure 5. As we can see, the $\sigma = 1$ scale is most distinctive for classification while other scales contain helpful information for classification.

From the above analysis, it can be concluded that the SIFT descriptor with Boosting gives the outstanding performance. Besides evaluation of this combination in the test data set, we take real-time videos to see the feasibility of our approach for real applications. In this case, foreground moving blobs are first extracted with cast shadow removed. Then feature vectors are calculated for every blob and categories are judged by classi-

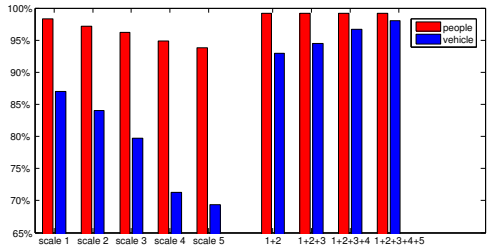


Figure 5. Effect of scale selection

fiers. Here, 4 20-minute videos are taken from 4 scenes with different view angles, view fields and lighting conditions with classification results shown in Figure 6 and average classification accuracy shown in Table 3. As we

(a) Video1 (b) Video2 (c) Video3 (d) Video4

Figure 6. Samples of respective classification results of testing videos

can see, classification performance is a bit worse than that on test data set. This phenomenon is related to: (1) the complicate circumstance which cannot be the same to those when the data set is collected; (2) the construction of different view angles in the data set; (3) the fact that the feature descriptor cannot totally afford all kinds of transformations in real world.

Table 3. Accuracy in real videos

Vehicles	Video1	Video2	Video3	Video4
Accuracy	97.3%	96.2%	97.9%	95.9%
Pedestrians	Video1	Video2	Video3	Video4
Accuracy	98.9%	98.2%	98.8%	97.6%

Further more, we apply conventional tracking to supply temporal information. With decision level fusion of tracking series of the same object, we can obtain more precise and stable classification results which are shown in Table 4. The processing speed is about 18 frames per second, which is enough for real applications and demonstrates the efficiency of the approach.

6 Conclusions

In this paper, we have proposed an approach for automatic object classification in traffic scene surveillance videos based on boosting local feature descriptors. Experimental results demonstrate the effectiveness and efficiency of our approach with robustness to noise, view angle change, lighting conditions and environments, which is very desirable in real applications.

Table 4. Accuracy of Tracked Series

Video1	Tracks	Correct Tracks	Accuracy
Vehicles	132	130	98.4%
Pedestrians	200	198	99.0%
Video2	Tracks	Correct Tracks	Accuracy
Vehicles	512	505	98.6%
Pedestrians	170	169	99.4%
Video3	Tracks	Correct Tracks	Accuracy
Vehicles	372	365	98.1%
Pedestrians	92	91	98.9%
Video4	Tracks	Correct Tracks	Accuracy
Vehicles	208	205	98.5%
Pedestrians	176	174	98.8%

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 authors also thank the anonymous reviewers for their valuable comments.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transaction on PAMI*, 2005.

[2] M. Bicego, U. Castellani, and V. Murino. A hidden markov model approach for appearance-based 3d object recognition. *Pattern Recognition Letters*, 2005.

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

[4] M. Everingham, A. Zisserman, and et al. The 2005 pascal visual object classes challenge. *Lecture Notes in Computer Science*, 2006.

[5] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting, 1998.

[6] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on PAMI*, 2005.

[7] Z. Liu, K. Huang, and T. Tan. Cast shadow removal with gmm for surface reflectance component. In *Proceedings of 18th ICPR*, 2006.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[9] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on PAMI*, 2005.

[10] P. A. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proceedings of 9th IEEE International Conference of Computer Vision*, 2003.

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