SVM-Based Classification of Moving Objects

Zhanchuang Li[1], Jianmin Jiang[1,2], and Guoqiang Xiao[1]

[1] Faculty of Information and Computing Science, Southwest University, China
[2] Digital Media & Systems Research Institute, University of Bradford, UK

Abstract. In this paper, we propose a single SVM-based algorithm to classify moving objects inside videos and hence extract semantics features for further multimedia processing and content analysis. While standard SVM is a binary classifier and complicated procedures are often required to turn it into a multi-classifier, we introduce a new technique to map the output of a standard SVM directly into posterior probabilities of the moving objects via Sigmoid function. We further add a post-filtering framework to improve its performances of moving object classification by using a weighted mean filter to smooth the classification results. Extensive experiments are carried out and their results demonstrate that the proposed SVM-based algorithm can effectively classify a range of moving objects.

Keywords: SVM(support vector machine); moving object classification and video processing; weighted mean filter.

1 Introduction

There are many different approaches, which have been proposed to automatically classify objects in images and videos. Most existing research on moving object classification requires pre-definition of the moving objects, such as walking human object, running vehicles etc. [1]-[6]. Wender et. al[1] introduced an object classification system for vehicle's monitoring applications, which is applied with a laser scanner. However, common cameras are widely used rather than laser scanners in social life, therefore, the algorithm reported in [1] can only be used for very limited occasions, and the cost of installation is high. In [2][3], some new features were proposed for object classification, such as Bag of Words and SIFT features, which have been popular for large-scale object classification in still images. In general, these methods are not suitable for classifying moving objects in low resolution surveillance videos. In references [4]-[6], several algorithms have been reported for object classification, which apply specialized detectors at each frame to detect object instead of segmenting the object. These methods may not sufficiently consider the temporal correlation among neighboring frames and the detected object regions usually lack shape information, making it difficult for their classification.

Essentially, moving object classification is a multi-class problem, which requires multi-class machine learning techniques to learn from training samples before objects can be classified. Typical examples of such learning machines include neural networks, K nearest neighbours and SVM (support vector machine) [7]-[9]. In the community of

D. Ślęzak et al. (Eds.): MulGrab 2009, CCIS 60, pp. 37–45, 2009.

multimedia and signal processing, recent trend of combining artificial intelligence with video processing indicates that SVM is one of the most popular machine learning tools for video content analysis and processing [7].

In this paper, we describe a multi-class SVM based object classification algorithm towards semantics extraction for multimedia content processing and analysis, in which our contribution can be highlighted as: (i) mapping of SVM output into multiple posterior probability domain for object classification; and (ii) introduction of weighted mean-based filtering for post-processing to improve the classification performances. The rest of the paper is organized into three sections. While section 2 describes the proposed algorithm, Section 3 reports our experimental results and Section 4 provides concluding remarks.

2 Proposed Algorithm Design

Correspondingly, we introduce a new multi-class SVM algorithm, where the output of SVM is mapped into a posterior probability domain for the classified objects via Sigmoid function and post-filtering approaches. By selecting the maximum value of the posterior probability, optimal performances can be achieved in classifying all the moving objects.

For the conciseness and coherence of description, M classes are pre-defined as $C=\{c_1,c_2,\ldots,c_M\}$. To represent each object, we construct L features, which can also be arranged as an input feature vector: $\boldsymbol{X}=(x_1,x_2,\ldots,x_L)$, where x_i is the ith feature extracted to describe the moving object. The proposed SVM-based algorithm consists of three operation phases, which include: (i) segmentation of moving objects; (ii) extraction of multiple features; and (iii) design of classification rules.

To convert the input video into a moving object sequence, we have a range of existing algorithms to select, where detection of changes is exploited to segment those moving objects and background information is removed to pave the way for feature extraction around the moving objects [10][11]. As our work is primarily focused on object classification, we simply adopted the existing segmentation techniques reported in [10] to expose the moving objects. The exposed moving objects are then further processed to produce binary mask sequences, from which corresponding ROI (region of interests) can be generated for feature extraction. For the convenience of evaluating the proposed algorithm without incurring complicated procedures and high computing costs, we selected three features to be extracted for describing the moving objects, which include contour, statistics of intensity and texture. Details of individual feature extraction are given below.

Contour and shape of the object is widely reported to be important features for describing moving objects [12]-[16], which presented significant discriminating power in characterizing different objects. In our proposed algorithm, we combine various features extracted from the contour of moving objects to ensure that the best possible performances could be achieved in object classification. These features include: (i) ρ: ratio of height to width inside the boundary rectangle; (ii) $\rho_{1/3}$ and $\rho_{2/3}$: ratio of height to width when its height is one third and two thirds of the object; (iii) φ: ratio of squared circumference to the object area; (iv) δ: ratio of areas between the object and its bounding rectangle; (v) α : the rotation angle of the minimum bounding ellipse of the moving object; (vi) γ: the eccentricity of the minimum bounding ellipse of the moving object. Their definitions are given as follows.

$$\rho = \frac{h_o}{w_o} \tag{1}$$

Where h_o and w_o are the height and width of object respectively.

$$\rho_{1/3} = \frac{h_o}{w_{o1/3}}, \tag{2}$$

$$\rho_{2/3} = \frac{h_o}{w_{o2/3}} \tag{3}$$

Where $w_{o1/3}$ and $w_{o2/3}$ are the width of the object when its height are one third and two thirds respectively.

$$\varphi = \frac{\tau_o^{\;2}}{\eta_o} \tag{4}$$

Where τ_o stands for the circumference of the object, and η_o the object area.

$$\delta = \frac{\eta_o}{\eta_r} \tag{5}$$

Where η_r is the area of the bounding rectangle.

Statistics of intensity simply contains the mean and variance of all the pixels inside the moving objects, which can be defined as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} I_i^O \tag{6}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(I_i^o - \mu\right)^2 \tag{7}$$

Where I_i^o stands for the intensity of the ith pixel inside the object and N the total number of the internal pixels.

The texture feature is adopted from the one reported in [17], which can be organized into a feature vector of 140 elements to represent each object. This procedure is detailed in [17].

To ensure the comparability of all the extracted features, we pre-process all the features by the following normalization process:

$$x_i\,' = \left(x_i - x_{i\ \mathrm{min}}\right)\Big/\left(x_{i\ \mathrm{max}} - x_{i\ \mathrm{min}}\right) \tag{8}$$

Where x'_i is the normalized feature x_i, and $x_{i,\min}$ and $x_{i,\max}$ are the minimum and the maximum values of the extracted feature, respectively.

As SVM is essentially a binary classifier, we adopt the modeling technique reported by Platt [18] to train the SVM with known data set and exploit Sigmoid function to map its outputs into posterior probability $P(c_j\!=\!1|\boldsymbol{X})$. Such mapping is described as follows:

$$P(y = 1 \mid f) = 1\big/\left(1 + \exp(Af + B)\right) \tag{9}$$

Where *f* is the output of SVM, *P*(*y*=1|*f*) is the probability of correct classification under the condition of *f*, and *A*, *B* are the controlling parameters, which need to be determined by solving the following problem of maximum likelihood:

$$F(z) = \min_{Z=(A,B)} \left(-\sum_{i=1}^{l} (t_i \log(p_i) + (1-t_i)\log(1-p_i))\right) \tag{10}$$

where, $p_i = 1/\exp(Afi + B)$, $f_i = f(x_i)$, and t_i are determined via:

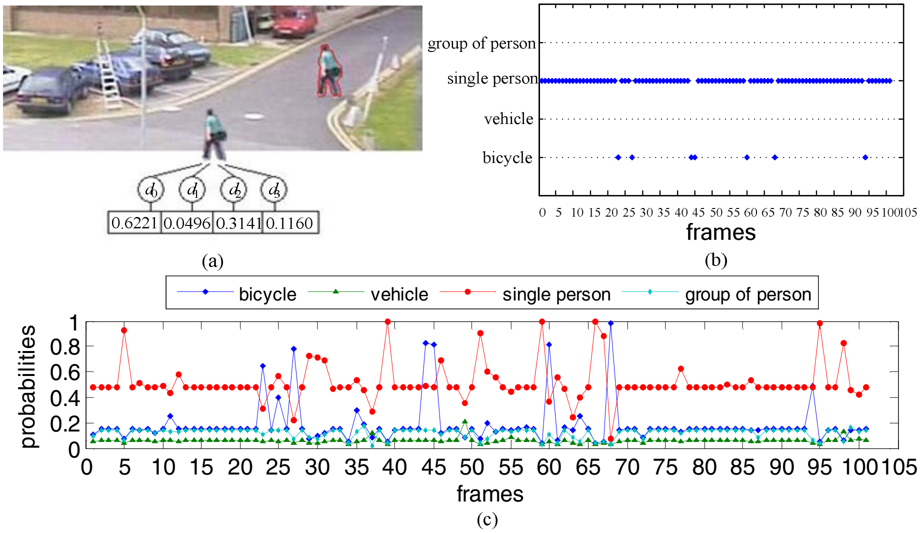
$$t_i = \begin{cases} (N_+ + 1)/(N_+ + 2) & if\ y_i = 1 \\ 1/(N_- + 2) & if\ y_i = -1 \end{cases} \tag{11}$$

and N_+, N_- are the number of positive and negative samples inside the training set.

Given M classes $\{c_1, c_2, \ldots c_j \ldots c_M\}$, the above mapping is exploited to construct a classifier d_j for each c_j and assign the output $P(c_j = 1|\boldsymbol{X})$ as the probability of classifying the input sample into c_j. Therefore, The class of the input sample ***X*** can be determined as the one to make the output probability maximum, i.e.

$$\hat{c} = \arg\max_{c_j} P(c_j = 1 | X) \tag{12}$$

In summary, the proposed algorithm constructs M SVMs and map their outputs into a posterior probability space to complete the M-class classification for input samples. Our experiments suggest that the proposed SVM multi-classification performs better than those with fusion module such as voting [19] and pair-wise coupling [20] when tested on the same dataset.



(a)

(b)

(c)

Fig. 1. Experimental results:(a) posterior probabilities for the object of single person generated in the 23rd frame of the video PETS2001; (b)The classification results of the single person across all frames in the video. (c)The posterior probability outputs of single person across all frames inside the video.

During the mapping procedure, it is observed that correct classification is dependent upon the accuracy of moving object segmentation and feature extraction. Yet such segmentation and extraction is largely affected by noises and limitation of segmentation techniques. As shown in Figure-1, the mapped posterior probabilities in the 23rd frame of the video, PETS2001, are 0.6221, 0.0469, 0.3141 and 0.1160 for bicycle, vehicle, single person, and group of people, respectively. According to (12), the input moving object, single person, will be wrongly classified as bicycle.

The part (c) in Figure-1 illustrates the posterior probability distribution over all the frames. By comparing with the part (b), it can be seen that classification for most frames are actually correct and such trend remains true even for those frames that are neighbours of the wrongly classified frames. To this end, we are encouraged to propose a weighted mean filtering as the post-processing to smooth the classification results and improve the performances of the proposed SVM-based classifier.

Essentially, all the posterior probabilities for the sequence of video frames can be regarded as a time series. As the moving object presents significant correlation within neighboring frames, there exists large extent of stability among the posterior probability values within this neighbourhood, and hence the wrong classification shown in Figure-1 can be regarded as caused by the noise of the high frequency components within the input time series. By applying the principle of low-pass filtering, such wrong classification could be eliminated and thus the classification performances could be further improved. Figure-2 illustrates the structure of our proposed weighted mean filter.

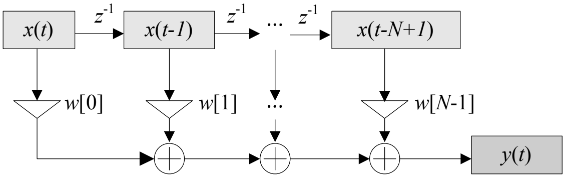


Fig. 2. Illustration of the proposed weighted mean filter

In Figure-2, $x(t)$ represents the value of input sample at the time t, z^{-1} is the delaying unit, $w[k]$ are the parameters of the filter corresponding to sampling at t-k, N represents the length of filter, and $y(t)$ is the output value of the filter at the time t. Hence, the proposed filter [21] in the time domain can be described as:

$$\hat{P}(c_j = 1 \mid X_{t-k}) = \sum_{k=0}^{N-1} \left[w[k] \left(P(c_j = 1 \mid X_{t-k}) \right) \right] \tag{13}$$

Where $\widehat{P}$ $(c_j=1|X_t)$ is the probability of $x_t \in C_j$, following the filtering of the input classification by SVM in the tth frame.

To determine the filter coefficients, we apply the following two principles: (i) the nearer the observation to the tth frame, the larger its influence upon the probability estimation of the tth frame; (ii) all coefficients must satisfy normalization process, i.e. $\sum w[k]=1$, and $0 \le w[k] \le 1$. Under the principles, our extensive experiments indicate that the following weighting produces good filtering performances:

Let k be the distance between the tth frame and the t-kth frame, we define a descending function $F(k)$, $k \in [0, N-1]$, as follows:

$$F(k) = (N-k)^2 \tag{14}$$

Therefore, the filter coefficients can be determined by:

$$w[k] = \frac{F(k)}{\sum_{i=0}^{N-1} F(i)} \tag{15}$$

In general, larger value of N enables the filter to consider more neighbouring values for producing its output, the result of which is that the output becomes more smooth, and hence more storage space would be needed in order to buffer all the relevant values before the moving objects inside the N video frames can be classified. Further, experiments suggest that different moving object classification requires different length N. Figure-4 illustrates the relationship between the length of filter N and the type of moving object, in which the vertical axis represents the performance of the filter, i.e. the normalized ECR (error correcting rate):

$$ECR = (AR' - AR)/AR \tag{16}$$

Where AR is the accuracy rate of classification before the filter, and AR' is the accuracy rate of classification after the filter.

The results in Figure-3 indicate that, when N>20, the classification results after filtering tends to be stabilized, and therefore, we select N=20 for the proposed algorithm and all evaluation experiments in the rest of this paper.

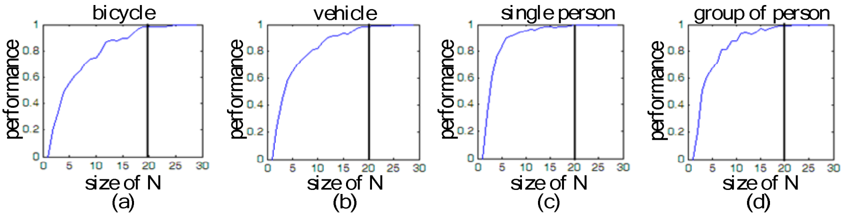


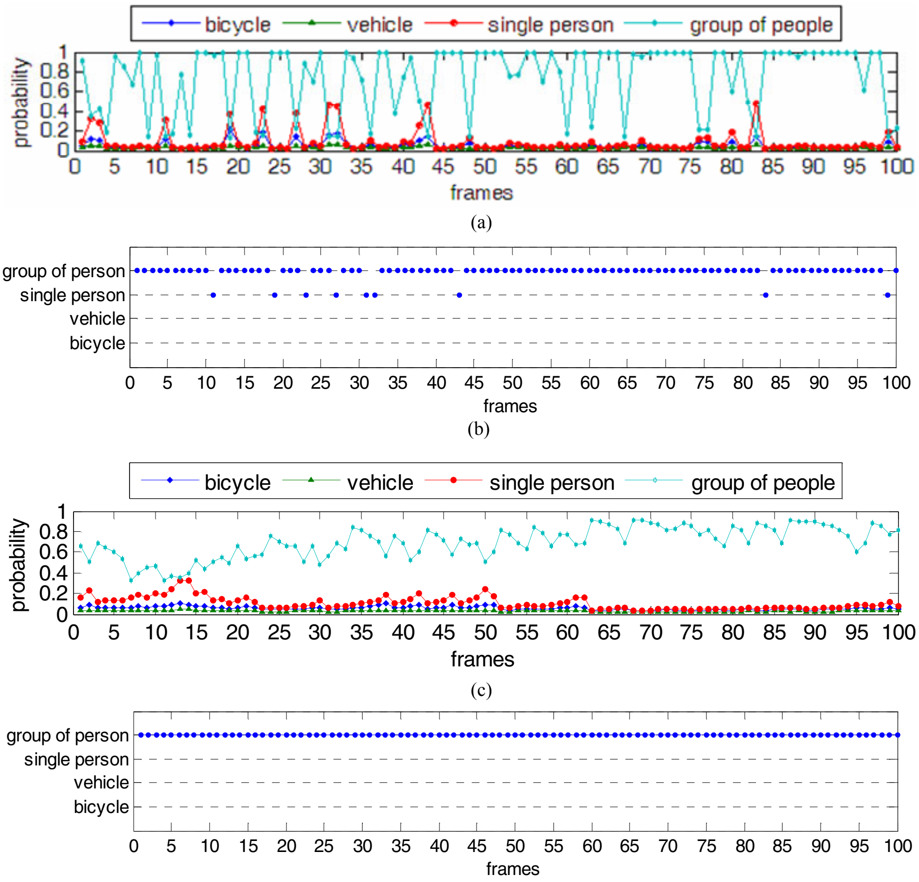
Fig. 3. Illustration of the dependency between the length of filter and the results of classification, where (a)bicycle (b)vehicle (c) single person and (d) group of people

Due to the fact that multiple moving objects could appear in the same frame, it is required to have object tracking module to characterize the relationship among all the moving objects in order to achieve sustainable classification results. As object tracking is a well researched area and many good tracking algorithms have already been reported in the literature [16][17], we select the algorithm reported in [17] as the multiple-object tracking method.

3 Experimental Results

To evaluate the proposed algorithm, we developed its software implementation in C++ and run extensive experiments for a set of surveillance videos, which are publicly available in the Internet [18]. According to the content nature of all the videos, we identified four types of moving objects, which are ‘single person’, ‘group of people’, ‘bicycle’ and ‘vehicle’.

As the PETS2001 video set contains 14 video clips, we used 4 video clips for training and the rest for testing. The size of all the video frames are 768x576. Figure 4 illustrate the classification results for the moving object ‘group of people’, where part-(a) illustrates the value of posterior probabilities generated by the mapping of SVM output without the filtering, part (b) the results of its classification without the filtering, part (c) the smoothed values of posterior probabilities after the filtering,



(a)

(b)

(c)

(d)

Fig. 4. Classification results for the moving object ‘group of people’

part (d) the final results of classification. Via comparison between part (b) and part (d), it can be seen that the introduction of the weighted mean filtering has improved the classification performances.

Table-1 lists all the experimental results for the video set PETS2001, where the performances of all the compared algorithms are measured by the correct classification rate. To benchmark the proposed algorithm, we compare the proposed algorithm with two other relevant versions to ensure their comparability, which are SVM plus voting for post-processing and the proposed SVM without post-processing. From all the results in Table 1, it can be seen that the proposed SVM with filtering outperforms the other two versions in terms of the correct classification rate. Note the number of classified moving objects is dependent on the content of the training and testing video set. In PETS2001, there are only four moving objects, which are listed in Table-1. For other videos, the number and type of the moving objects may vary and hence the proposed algorithm needs to be re-trained.

Table 1. Experimental results in terms of correct classification rate

	SVM+voting method	presented method	
		probability SVM	probability SVM+filtering
bicycle	75.2%	84.9%	90.5%
vehicle	83.3%	89.5%	96.7%
single person	86.3%	91.3%	98.8%
group of person	85.6%	92.4%	97.3%
Total-average	83.9%	90.3%	97.2%

4 Conclusions

In this paper, we described a SVM-based classification algorithm to automatically process input videos and identify moving objects. To overcome the binary nature of SVM, we introduced a post-processing method to map the SVM output into a posterior probability space and thus enable SVM to produce multi-classification results. To improve its performances, we further proposed a weighted mean filtering as a final post-processing to smooth the posterior probability among neighbouring frames and exploit the correlation property within a neighbourhood of the input videos. As a result, such filtering is capable of removing the negative effect of noises and correct those random wrong classifications to make the proposed algorithm more robust. Experimental results support that the proposed algorithm achieves excellent classification results for the video set PETS2001, which is publicly available in the Internet and make it convenient for comparative evaluations with any further research.

References

[1] Wender, S., Dietmayer, K.: An Adaptable Object Classification Framework. In: IEEE Intelligent Vehicles Symposium, pp. 150–155 (2006)

[2] Lowe, D.: Distinctive Image Features from Scale-invariant Keypoints. International Journal of Computer Vision 60(1), 91–110 (2004)

[3] Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: CVPR, vol. 1, pp. 886–893 (2005)

[4] Viola, P., Jones, M., Snow, D.: Detecting Pedestrians Using Patterns of Motion and Appearance. In: ICCV, vol. 2, pp. 734–741 (2003)

[5] Schneiderman, H., Kanade, T.: A Statistical Method for 3D Object Detection Applied for Faces and Cars. In: CVPR, vol. 1, pp. 746–751 (2000)

[6] Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2(2), 121–167 (1998)

[7] Zhang, H., Berg, A., Maire, M., Malik, J.: SVM-KNN: Discriminative Nearest Neighbor Classification for visual Category Recognition. In: CVPR, vol. 2, pp. 2126–2136 (2006)

[8] Ravichandran, B., Gandhe, A., Smith, R., Mehra, R.: Robust Automatic Target Recognition Using Learning Classifier Systems. Information Fusion 8(3), 252–265 (2007)

[9] Yang, T., Li, S.Z., Pan, Q., Li, J.: Real-time and Accurate Segmentation of Moving Objects in Dynamic Scene. In: Proceeding of the ACM 2nd international workshop on Video surveillance & sensor networks, pp. 136–143 (2004)

[10] Dewan, M., Hossain, M.: Background Independent Moving Object Segmentation for Video Surveillance. IEICE Transactions on Communications E92-B(2), 585–598 (2009)

[11] Hota, R.N., Venkoparao, V., Rajagopal, A.: Shape Based Object Classification for Automated Video Surveillance with Feature Selection. In: ICIT 2007, pp. 97–99 (2007)

[12] Zhang, L., Li, S.Z., Yuan, X.T., Xiang, S.M.: Real-time Object Classification in Video Surveillance Based on Appearance Learning. In: CVPR 2007, pp. 1–8 (2007)

[13] Lipton, A.J., Fujiyoshi, H., Patil, R.S.: Moving Target Classification and Tracking from Real-time Video. In: WACV 1998, pp. 8–14 (1998)

[14] Berg, A., Berg, T., Malik, J.: Shape Matching and Object Recognition Using Low Distortion Correspondences. In: CVPR 2005, vol. 1, pp. 26–33 (2005)

[15] Brown, L.M.: View Independent Vehicle/Person Classification. In: Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks, pp. 114–123 (2004)

[16] Nguyen, H.T., Qiang, J., Smeulders, A.W.M.: Spatio-temporal Context for Robust Multitarget Tracking. IEEE Transaction on Pattern Analysis and Machine Intelligence 29(1), 52–64 (2007)

[17] `ftp://ftp.pets.rdg.ac.uk/pub/PETS2001`