International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.5 (2015), pp.141-154
http://dx.doi.org/10.14257/ijmue.2015.10.5.14

Pedestrian Classification Based on Full-SVM Decision Tree

Hongmin Xue[1,2], Zhijing Liu[1] and Jing Xiong[1]

[1]*School of Computer Science and Technology, Xidian University, Xi'an, P. R. China*
[2]*Department of Computer Science and Technology, ShaanXi XueQian Normal University, Xi'an, P. R. China*
xuehongmin@163.com, liuzhijing@vip.163.com, xiongjing_mail@163.com

Abstract

Visual analysis has potential to be used for recognition, and it is one of the hottest but most difficult subjects in computer vision. In order to identify pedestrian movement in an Intelligent Security Monitoring System, the video activity in the prospect is represented by a series of spatio-temporal interest points. Since human posture has the characteristics of uncertainty and illegibility, the clustering centers of each class are computed by fuzzy clustering techniques. We presented a pedestrian classification method based on improved support vector machines in order to solve non-rigid objects that are difficult to identify in intelligent monitoring systems. The Support Vector Machine technology and the decision tree have combined into one multi-class classifier so as to solve multi-class classification problems. Then a full-SVM (Support Vector Machine) decision tree is constructed based on the conventional decision tree. At last, the method is evaluated on the KTH action dataset and receives a comparatively high correct recognition rate.

Keywords: *action classification; SVM; Decision Tree; Interest Point*

1. Introduction

Human motion analysis is one of the most active research topics in the field of computer vision, which has a number of promising applications, ranging from security and surveillance to augmented reality. With public security coming along the increasing importance of Intelligent Video Surveillance System (IVSS) and its wide installation in the train stations, bus stations, subways, airports, streets, schools, venues, laboratories, and other public places. As an active research topic in computer vision, the IVSS in dynamic scenes attempts to detect, track and recognize walkers from video sequences, and then to describe and understand pedestrian behaviors. The ultimate goal is the substitution of IVSS for the traditional passive video surveillance system which has proved to be ineffective when the number of cameras exceeds the capability of human operators to monitor [1]. In short, the goal of IVSS is not only to put cameras in place of human eyes, but also to accomplish the entire surveillance task as automatically as possible. The prerequisites for effective automatic detection by means of a monocular camera include the following steps: motion detection, objects tracking and behavior understanding.

Though the detection of human abnormal behavior can help to detect crimes, accidents, terrorist attacks, etc., it is hard to define *abnormal* in different scenarios. The most convenient way is to classify human's motion in the beginning, and then the behavior validity can be judged under special scenarios. Motion classification attempts to use gait as biometric information to understand people by the characteristics of their motion. It also includes efforts to classify different types of human activities, such as Walking, jogging, running, boxing, hand waving, bending, jumping and skipping action. Researches in the field have been being actively

ISSN: 1975-0080 IJMUE

conducted, but the current techniques available are too intuitive or computing-costly to be used for real-time applications. Thus, it is desirable to reduce the dimensionality of the motion data space without the loss of its features. Furthermore, existing technologies deal with the human action classification with fixable scene, and the method cannot be used for any other scene. Thus, it is desirable to find a method of recognition human abnormal motion in different scenarios. Most of the researches classify the motion based on posture. The aim of the posture analysis is to determine the state of a person in an image. In contrast, gait classifiers analyze people's movements, that is, changes in posture over a number of frames.

This paper addresses the computing-costly problem with space-time interest points, which reduce the dimensionality of the motion data space without the loss of its features. This paper also addresses the problem of learning and recognizing abnormal actions in different scenarios through double-layer Bag-of-Words method using standard surveillance cameras. These are operator-controlled nonstationary cameras and usually of the pan-tilt-zoom (PTZ) variety, which allows multiple views. They are capable of covering larger outdoor scenes compared with static cameras. What is done in this paper is a model-free method with prior knowledge that denotes the entire body, and it belongs to the work of the behavior understanding. Through behavior classification we can speculate about the abnormal behavior. Our paper therefore makes three main contributions:

- We propose an approach for inferring and generalizing gait classification simply from interest point, even if those particular motions have never been seen before.
- We present a method which can robustly, compactly and effectively represent different behavior sequences. The method is a complex number notation based on centroid.
- We present a Support Vector Machine technology and the decision tree method, which allows efficient classification movement.

The paper is organized as follows. In Section 2 we briefly review some related work, which includes pedestrian classification. Section 3 describes principle theories for Video Representation Support Vector Machine and decision tree. Our algorithm Non-linear decision tree and NSVM decision tree is presented in Section 4. An experimental evaluation of the proposed techniques is presented in Section 5, and the paper concludes a discussion and summary in Section 6.

2. Related Work

2.1. Pedestrian Classification

It is usually a challenge to classify human actions automatically because of its complexity, especially under outdoor circumstances. But a lot of attempts have already been done in this field.

A number of techniques are used for gait classification. Some are limited to one person on the frame while others can track multiple people. Some systems work on separate frames to analyze posture while others follow changes in posture over a sequence of frames.

Some attempts at behavior classification have already been made. One of the pioneers is [2], who has proposed a method for modeling several body parts. The trajectories are used to train hidden Markov Models (HMM), and one HMM denotes each kind of gait.

The importance of human behavior classification is evident with the increasing requirement of machines to interact intelligently and effortlessly with a human

inhabited environment. Research in the field has been being actively conducted but the currently available techniques are extremely intuitive or high compute cost which cannot be used for real-time applications. It is thus desirable to reduce the dimensionality of the data space representing motion without the loss of its features [3]. The original research on measuring human gait was initially for medical purposes. For example, Murray [4] applied gait to classify patients into different groups in order to pursue different types of medical treatment. In the field of computer vision research, human motion has been studied intensively [5], and a significant amount of progress has been achieved on human gait analysis and recognition. The performance of gait recognition can be affected by many factors[6]such as silhouette quality, walking speed, dynamic/static component, elapsed time, shoes, carrying condition [7], physical and medical condition, disguise, indoor/outdoor conditions, etc. Moreover, the effects of different factors may be correlated to each other. For example, a change on walking surface or shoe type may cause a change on speed. Although gait can be affected by numerous factors, from the viewpoint of recognition it is still quite useful [6].

The effective representation of gait is a key issue. Currently, there are several successful representation models such as appearance-based models [6-9], stochastic statistical models [10], articulated biomechanics models [11], in which a set of parameters describes the gait, and other parameter-based models[12]. Several of these models can be combined to further improve the representation of gait. Many appearance-based models have been developed for human gait recognition [13]. Some models use the silhouette of the entire body[10-12], whereas others use the most discriminant parts[14] such as the torso and the thighs. Gait energy image (GEI) [15] is a spatio-temporal method which is proposed to characterize human walking properties for individual recognition by gait. Global representation and conditional model are employed for human motion recognition [16].

Although considerable achievements in this field have been accomplished in recent years, some challenges still remain to be overcome. The optimal gait classification should classify motion with realtime. With this purpose, in this paper the video activity in the prospect is represented by a series of spatio-temporal interest points and then describes and evaluates the SVM technology and the decision tree combined into one multi-class classifier for gait classification.

2.2. Interest Point Methods

Now a popular approach in object recognition is based on spatial-temporal interest points and local feature descriptors. All of the local descriptors are usually vector-quantized to obtain a finite set of visual words before they are fed into any classification algorithms. Laptev[17] propose a spatial-temporal interest point operator, which detects local structures in space-time, and the local structures have large variations both in space and time. Schuldt trains a Support Vector Machine classifier based on spatial-temporal features for recognizing human actions. Dollar proposes a space-time interest point detector based on a set of linear filters, and then uses these local features with the k-nearest neighbor classifier for action recognition. Nowozin first detects local interest points, and learns a set of discriminative subsequences for action classification using the sequence mining techniques from data mining. Niebles [18] combine shape information with local appearance features by building a hierarchical model, which can be characterized as a constellation of bag-of-features. Local descriptors extracted from space-time interest points have also been shown to work well on videos with complex scenes. Laptev [17] learn a boosted classifier based on those local descriptors to do action recognition on movie data.

2.3. Topic Models for Visual Recognition

Recently, Bag-of-Words models have drawn more attention in the field of object recognition. The Bag-of-Words model is originally proposed for analyzing text documents, where a document is represented as a histogram over word counts. Generative topic models are then applied to this Bag-of-Words representation, and the topics of the document are denoted as latent variables in these models. Popular topic models include probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA) and Correlated Topic Models. Sivic et al. perform unsupervised learning of object categories using variants of the pLSA model. In their models, the *words* correspond to local patches extracted by interest point operators, and the *topics* correspond to the different object categories.

One shortage of these methods is that the *visual words* are usually obtained from local patches and are not as discriminative as their counterparts in text analysis. To address this issue, Russell [19] use image segmentation to produce groups of related visual words. For each image, they obtain multiple segmentations by varying the parameters of the segmentation algorithms, and represent the segments from the segmentation algorithms as the visual words in the pLSA model.

Topic models have also been applied in understanding images and videos containing human figures. Niebles [18] recently demonstrate some impressive results on unsupervised learning of human action categories using pLSA and LDA models. The *visual words* in their models are based on features extracted from spatial-temporal interest points. Different human actions are captured by the different "topics" discovered by either pLSA or LDA.

Despite the considerable achievements in the field accomplished in recent years, there are still some challenges to be overcome. The optimal abnormal behavior detecting should allow the detection of suspicious events with a minimal description and perform the detection without assumptive scenario, and the most important can recognize more abnormal behavior using the same library. For this goal to be achieved, we develop a double-layer Bag-of-Words model, which is described and evaluated in this paper.

2.4. Codebook Formation

To obtain a descriptor for each spatial-temporal cube, we calculate its brightness gradients in *x, y* and *t* three directions. The spatial-temporal cube is then smoothed at different scales before the image gradients are computed. The computed gradients are concatenated to form a vector. The size of the vector is equal to the number of pixels in the cube *times* the number of smoothing scales *times* the number of gradients directions. This descriptor is then projected to a lower dimensional space using the Principal Component Analysis (PCA) dimensionality reduction technique. In Niebles [18], different descriptors have been used, such as normalized pixel values, brightness gradient and windowed optical flow. Both the gradient descriptor and the optical flow descriptor are equally effective in describing the motion information. In the rest of the paper, we will employ results obtained with gradient descriptors.

As mentioned above, a very simple descriptor based on image gradients is adopted. Such a descriptor does not provide scale invariance neither in the space nor in the time domain. It does not capture relative camera motion. However, more complex descriptors are available with the cost of more computational complexity. In our experiment, we rely on the codebook to handle scale changes and camera motions. All the video sequence words are stored in video codebook.

3. Principle Theories

Similar to, we represent a video sequence as a "bag of words". But our representation is different from in two aspects. Our method represents a frame as a single word, rather than a collection of words from vector quantization of space-time interest points. In other words, a "word" corresponds to a "frame", and a "document" corresponds to a "video sequence" in our representation?

3.1. Video Representation

There are several choices in the selection of good features to describe moving objects. In general, there are three popular types of features: static features based on edges, dynamic features based on optical flow measurements, and spatial-temporal features obtained from local video patches. In particular, spatial-temporal interest points have turned out to be useful in the human motion categorization task, providing a rich description and powerful representation.

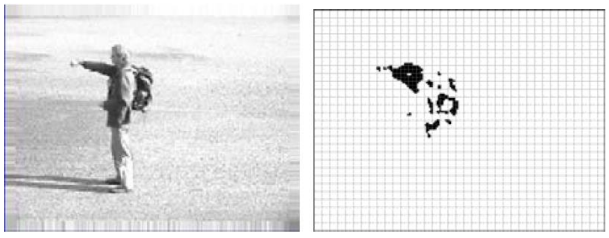
We use the motion descriptor in to represent the video sequences. This motion descriptor has been shown to perform reliably with noisy image sequences, and has been applied in various tasks, such as action classification, motion synthesis, etc.

To calculate the motion descriptor, we first extract moving objects. We make no attempt to track objects. The motion information is computed directly via spatiotemporal filtering of the image frames:

$$M_t(x,y,t) = (I(x,y,t) * g(x,y;\sigma) * h_{ev}(t;\tau,\omega))^2 + (I(x,y,t) * g(x,y;\sigma) * h_{od}(t;\tau,\omega))^2 \quad (1)$$

Where $g(x,y;\sigma) = e^{((\frac{x}{\sigma_x})+(\frac{y}{\sigma_y}))}$ is the 2D Gaussian smoothing kernel, applied only along the spatial dimensions (x,y), and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally, which are defined as $h_{ev} = (t;\tau,\omega) = -\cos(2\pi t\omega) \times e^{-t^2/\tau^2}$ and $h_{od} = (t;\tau,\omega) = -\sin(2\pi t\omega) \times e^{-t^2/\tau^2}$.

The two parameters σ and τ correspond to the spatial and temporal scales of the detector respectively. This convolution is linearly separable in space and time and is fast to compute. To detect moving objects, we threshold the magnitude of the motion filter output to obtain a binary moving object map: $M_t(x,y,t) > T_{h_a}$. The process is demonstrated in Figure 1.



(a) (b)

Figure 1. Feature Extraction from Video Frames. (a) Original Video Frame. (b) Binary Map of Objects

The feature we use is the spatial histogram of the detected objects. Let $H_t(i,j)$ be an $m \times m$ spatial histogram, with m typically equal to 10.

$$H_t(i,j) = \sum_{x,y} (x,y,t) \cdot \delta(b_i^x \le x \le b_{i+1}^x) \cdot \delta(b_i^y \le y \le b_{i+1}^y) \quad (2)$$

where $b_i^x, b_i^y (i, j = 1, \cdots, m)$ are the boundaries of the spatial bins. The spatial histograms, shown in Figure 2, indicate the rough area of object movement.

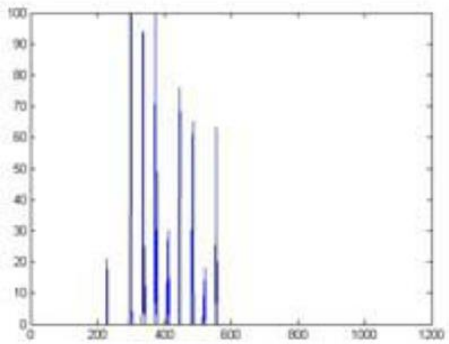


Figure 2. Feature Extraction from Video Frames Binary Map of Objects Spatial Histogram of Figure 1(b)

To construct the codebook, we apply vector quantization to the histogram feature vectors classifying them into a dictionary of W words, $\{\omega_1, \cdots \omega_W\}$ using K-means. The input video is uniformly segmented into one-second clips, and the input to our model at second t is the bag of all visual words occurring in video clip t, denoted as v_t. The video clip v_t is represented as word sequences w_t, given as

$$w_t = [w_{n_1}, \cdots, w_{n_2}, \cdots, w_{n_{t_t}}] \quad (3)$$

where n_t is the length of the t th video clip. w_{n_i} corresponds to the i th image frame of v_n, where $w_{n_i} = w_k$ indicates that a word of the k th word class has occurred in the frame.

3.2. SVM

The theory of Support Vector Machine (SVM) is from statistics which is proposed by Vapnik. The basic principle of SVM is finding the optimal linear hyperplane in the feature space that maximally separates the two target classes. For linearly separable and non-separable data, it can be translated into quadratic programming (QP) and can get an only limit point. In the case of non-linear, SVM can map the input to a high-dimensional feature space by using non-linear mapping and then the linear hyperplane can be found [20].

The basic principle of SVM is finding the optimal linear hyperplane in the feature space that maximally separates the two target classes, which shows in Figure 3.

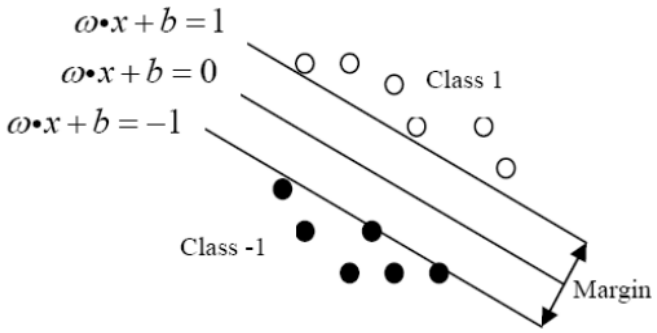


Figure 3. The Optimal Linear Hyperplan

The hyperplane which separates the two classes can be defined as:

$$wx+b=0 \quad (4)$$

Here ω and b are nonzero constants, x_k is a group of samples: $\{(x_1, y_1),(x_2, y_2),\cdots,(x_k, y_k)\}$, $x_k \in R^n, y_k \in \{-1,1\}$, and k is the number of styles.

3.3. Decision Tree

Decision tree comes from the Concept Learning System in 1960s. Decision tree shows decision set with intuitionistic tree-structure, and it is also an effective classifier and easy to get the classification rules. In a decision tree, each internal node splits the instance space into two or more subspaces according to a certain discrete function of the input attributes values. To classify an unknown instance, it is routed down the tree according to the values of the attributes tested in successive nodes, and when a leaf is reached the instance is classified according to the class assigned to the leaf [21].

4. Non-linear SVM Decision Tree

4.1. Non-linear Decision Tree

Typically, SVM is designed for two-class problems both positive and negative objects exist, but practical classification applications are multi-class problems. Now there are two common ways to construct SVM multi-class classifier. One is reconstructing multi-class classification models based on traditional SVM algorithmic, so as to satisfy the need of classification; the other is combining several binary classifiers into a multi-class one. Traditional SVM trains Support Vector Machines classifier directly under linear condition, so the training is difficult and the amount of computing is big and the speed of training is slow. In this paper, SVM is extended to non-linear by using kernel functions. Then non-linear SVM combines with decision tree to solve multi-class classification problems. We call this method NSVM Decision Tree. Before we describe the algorithm in detail, the distance between classes should be calculated firstly and then the relativity separability measure between them can be gotten.

In the case of non-linear, we can calculate the relativity separability measure between class i and class j after non-linear mapping [20].

1) For input sample z_1 and z_2, non-linear mapping Φ map them into the feature space H, then the Euclidean distance between z_1 and z_2 in H is：

$$d^H(z_1, z_2) = \sqrt{K(z_1, z_1) - 2K(z_1, z_2) + K(z_2, z_2)} \quad (5)$$

Where K(·,·) is the kernel function. In the space H, suppose m_Φ is the class center and $m_\Phi = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)$， here n is the number of samples within class.

2) For $\{x_1, x_2, \cdots, x_{n_1}\}$ and $\{x_1', x_2', \cdots, x_{n}'\}$ are the training samples for two classes, Φ map them into feature space H, m_Φ and m_Φ' are the class centers in feature space H, then the distance between m_Φ and m_Φ' in H is:

$$d^H(m_\Phi, m_\Phi') = \sqrt{\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}K(x_i, x_j) - \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}K(x_i, x_j') + \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}K(x_i', x_j')} \quad (6)$$

3) For the training samples $\{x_1, x_2, \cdots, x_n\}$ of a given class, the mapping Φ map them into feature space H, the distance between training sample x and class center m_Φ is:

$$d^H(x, m_\Phi) = \sqrt{K(x, x_i) - \frac{2}{n}\sum_{i=1}^{n_i} K(x, x_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K(x_i, x_j)} \quad (7)$$

Then in feature space the class variance can be get $\sigma^H = \frac{1}{n-1}\sum_{i=1}^{n} d^H(x_i, m_\Phi)$. Therefore, the relativity separability measure between class *i* and class *j* in feature space H is: $rsm_{ij}^H = \frac{d^H(m_\Phi^i, m_\Phi^j)}{(\sigma_i^H + \sigma_j^H)}$.

If $rsm_{ij}^H \geq 1$, then there is no overlap between class i and class j, and if $rsm_{ij}^H < 1$, there is overlap between them. The bigger the rsm_{ij}^H, the more easily separated between class *i* and class *j*.

We then define rsm_i^H as the relativity separability measure of class i from the others, and we have the formula $rsm_i^H = \min_{\substack{j=1,\cdots,k \\ j \neq i}} rsm_{ij}^H$. From the formula we can know that the relativity separability measure of class *i* is the minimum one between class *i* and the others.

4.2. NSVM Decision Tree (NSVMDT)

The closer to the root node the occurrence of the classification error is, the greater influence on the classifier. If the most separable classes can be separated firstly, the ability of the classifier will be better [21-22].

For a problem with k-classes, suppose the training set $\Phi = \{X_1, X_2, \cdots, X_k\}$.

Step1: In feature space H, calculate the relativity separability measure rsm_i^H, *i*=1, 2, …,k;

Step2: Array the relativity separability measures of classes by descendent order. Suppose $rsm_{n1} \geq rsm_{n2} \geq \cdots \geq rsm_{nK}$;

Step3: Set counter t=1;

Step4: Construct sub-classifier SVM_t as a two-class problem. Where the training set is $\Phi t = \Sigma_1 + \Sigma_2$, and $\Sigma_1 = \{(X_{nt}, +1)\}, \Sigma_2 = \{(Y, -1) \mid y \in \{\Phi - \{X_{nt}\}\}$;

Step5: Update the training set and the counter. $\Phi = \Phi - \{X_{nt}\}$, t=t+1;

Step6: Repeat Step4 and Step5 till t=k-1 and the classifier SVM_{K-1} is constructed.

The decision tree constructed by this way is hierarchical. There is only one sub-SVM at most in every level, and the importance of these sub-SVMs differs. The closer to the root node, the more important the sub-SVM is and the bigger the number of elements in the training set is.

5. Experimental Results and Discussion

In order to test the feasibility of our approach to pedestrian classification, we experiment with KTH human motion dataset [23]. The KTH dataset is one of the largest video datasets of human actions. It contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. Representative frames of this dataset are shown in Figure 4.

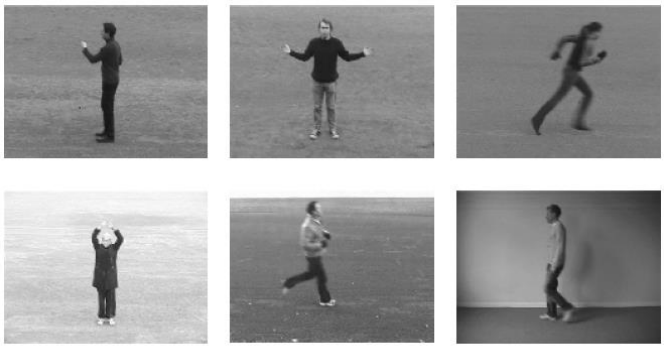
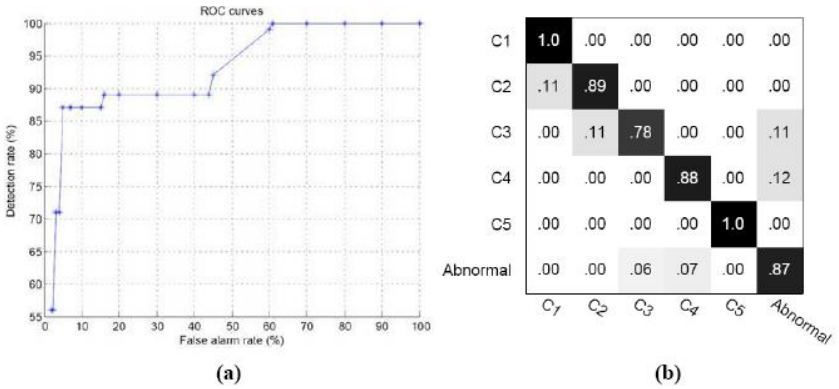


Figure 4. Representative Frames in KTH Dataset

We perform leave-one-out cross-validation on this dataset. For each run, we choose the video sequences of one subject as the test set, and build our model on the rest of the video sequences. We run the same process on each of the video sequence. For each run, we take the features obtained from Sect. 3.1 as the feature vectors. Then these feature vectors are quantized by *K*-medoid clustering to form the visual words. Since the number of feature vectors is huge, we randomly select 60 frames from each training video sequence for the *K*-mediod clustering. We extract interest points and describe the corresponding spatial-temporal patches with the procedure described in Section 3. The detector parameters are set to σ =1.5 and τ =2. Each spatial-temporal patch is described with the concatenated vector of its space-time gradients. Then, the descriptors are projected to a lower dimensional space.

The confusion matrix for the KTH dataset using 610 codewords is shown in Figure 5(b). We can see that the algorithm correctly classifies most of actions. We also test the effect of the codebook size on the overall accuracy. The result is shown in Figure 5(a). The best accuracy is achieved with 610 codewords, but is relatively stable.



(a) (b)

Figure 5. (a) The Effect of the Codebook Size (b) The Confusion Matrix for the KTH Dataset Using 610 Codewords

We compare our results with previous approaches on the same dataset, as shown in Table 1. We would like to point out that the numbers in Table 1 are not directly comparable, since different approaches use different split of training and test data. Our method achieves better performance by a large margin.

Table 1. Comparison of Different Methods in Terms of Recognition Accuracy on the KTH Dataset

Methods	Accuracy(%)
LDMC	92.26
SEMI-LDA[24]	90.43
HTMM[25]	84.46
LDA[22]	81.50

In order to identify 6 kinds of behavior, for each sequence using fuzzy C means clustering technology has 6 clustering center, constructing SVM decision tree multivalued separator. As shown in Figure 6, and each cluster center corresponds to a behavior.

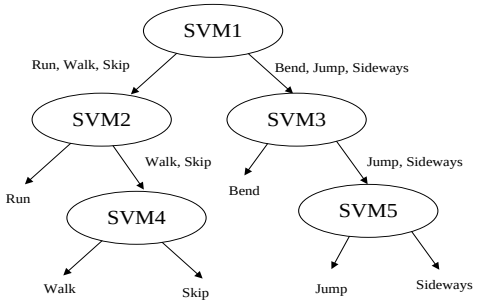


Figure 6. Support Vector Machine Multi-class Classifier

Using the RBF kernel function as kernel function of support vector machine, radial basis function defined as follows:

$$K(x_i, x_j) = \exp(\frac{\left\| x_i - x_j \right\|^2}{\sigma^2}) \quad (8)$$

Table 2 gives 6 kinds of behavior recognition results, overall, this method has a high recognition rate for simple everyday actions, table for bending the recognition rate is the highest, and of walking, running and jumping one leg low recognition rate. This is mainly because the recognition by this method is in the shape of human body characteristics of action sequences as the foundation, if the body position for two kinds of behaviors is similar in shape, it is easy to cause the error recognition.

Table 2. Confusion Matrix for Six Human Actions

	Walk	Run	Jump	Bend	Sideways	Skip
Walk	0.86	0.12				0.02
Run	0.10	0.87				0.03
Jump			0.94			0.06
Bend				1.00		
Sideways	0.07	0.02			0.91	
Skip	0.05	0.07				0.88

6. Conclusions

We conduct a gait classification, which aims to understand the effects of the real-time classification based on multiple clues. In our method, a description is given to classify gait using a multi-class classifier, and on the basis of the results, an intelligent security real-time system is built. Our analysis shows that through this way the expense of computation can be reduced and a better recognition rate can be reached, compared with other methods.

In this paper, SVM is extended to non-linear SVM by using kernel functions, and the most easily separated classes are separated firstly, so the iterative error is lower

and the correct classification rate is higher. For pedestrian classification we have proposed a novel and effective method for gait-based human recognition by non-linear SVM (Support Vector Machine) decision tree. The experimental results and analysis indicate that our method is effective in automated gait recognition. The key features include low cost of computation and unnecessary learning by large data, which suggests this method to be real-time and efficient for video surveillance. Whereas it requires clear-cut images, pre-processed image could be implemented for intact human silhouette before feature extraction and representation. Then recognition rate depends on the pre-procession in the following steps of feature extraction, representation and computation. Otherwise, it could suffer from errors of extraction and representation to some extends. The results are promising, though we acknowledge the lack of large and challenging video datasets to thoroughly test our algorithm, which poses an interesting topic for future investigation. Currently, we have been working to demonstrate the applicability and reliability of our abnormal recognition scheme, unconstrained real world settings. In addition, we plan to further investigate the possibilities of using a unified framework by combining generative and discriminative models for pedestrian classification.

Acknowledgements

The project is funded in part by Shaanxi Provincial Department of education scientific research project (12JK0749).

Reference to this paper should be made as follows: Hongmin Xue and Zhijing Liu, `Pedestrian Classification Based on Improved Support Vector Machines', *Proceedings - 5th International Conference on Intelligent Networking and Collaborative Systems, INCoS2013*, p726-730.

References

[1] W. Hu, T. Tan, L. Wang and S. Maybank, "A survey on visual surveillance of object motion and behaviors", IEEE Trans. System, Man, and Cybernetics, vol. 34, no. 3, **(2004),** pp. 334-352.
[2] D. Meyer, J. and H. Niemann, "Gait classification with HMMs for Trajectories of body parts extracted by mixture densities", British Machine Vision Conference, **(1998)**, pp. 459-468.
[3] A. Sharma, D. K. Kumar, S. Kumar and N. McLachlan, "Wavelet Directional Histograms for Classification of Human Gestures Represented by Spatio-Temporal Templates", Proceedings of the 10th International Multimedia Modelling Conference,**(2004)**, pp. 57-63.
[4] M. Murray, A. Drought, and R. Kory, "Walking Pattern of Normal Men", J. Bone and Joint Surgery, vol. 46-A, no. 2, **(1964),** pp. 335-360.
[5] T. Moeslund and E. Granum, "A Survey of Computer Vision-Based Human Motion Capture", Computer Vision and Image Understanding,vol. 81, no. 3, **(2001),** pp. 231-268.
[6] S. Sarkar, P. Phillips, Z. Liu, I. Vega, P. Grother, and K. Bowyer, "The HumanID Gait Challenge Problem: Data Sets, Performance and Analysis", IEEE Trans. Pattern Analysis and Machine Intelligence,vol. 27, no. 2, **(2005),** pp. 162-177.
[7] I. Haritaoglu, R. Cutler, D. Harwood, and L. Davis, "Backpack: Detection of People Carrying Objects Using Silhouettes", Computer Vision and Image Understanding, vol. 6, no. 3, **(2001),** pp. 385-397.
[8] R. Tanawongsuwan and A. Bobick, " Modelling the Effects of Walking Speed on Appearance-Based Gait Recognition", *Proc.* IEEE Intl Conf. Computer Vision and Pattern Recognition,vol. 2, **(2004),** pp. 783-790.
[9] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette Analysis-Based Gait Recognition for Human Identification", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 25, no. 12, **(2003),** pp. 1505-1518.
[10] A. Kale, A. Sundaresan, A. N. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa, "Identification of Humans Using Gait", IEEE Trans. Image Processing, vol. 13, no. 9, **(2004),** pp. 1163-1173.
[11] X. J. Yang, "Human Recognition Using Multi-Frame Gait Silhouette Matching", *IJACT*, vol. 4, no. 22, **(2012),** pp. 788-795.
[12] R. Cutler and L. Davis, "Robust Periodic Motion and Motion Symmetry Detection", Proc. IEEE Intl Conf. Computer Vision and Pattern Recognition, **(2000),** pp. 615-622.

[13] J. Han and B. Bhanu. Statistical Feature Fusion for Gait-Based Human Recognition, Proc. IEEE Intl Conf. Computer Vision and Pattern Recognition, **(2004),** vol. 2, pp. 842-847.
[14] L. Lee, G. Dalley, and K. Tieu, "Learning Pedestrian Models for Silhouette Refinement," Proc. IEEE Intl Conf. Computer Vision, **(2003)**, vol. 1, pp. 663-670.
[15] J. Han and B. Bhanu, "Individual Recognition Using Gait Energy Image", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 2, **(2006)** February**,** pp. 316-322**.**
[16] H. Zhang, Z. Liu, H. Zhao and Q. Wei, "Human Motion Classification Based on Global Representation and Conditional Model", *Springer* **(2011).**
[17] I. Laptev and T. Lindeberg, "Space-time interest points", Proceedings, Ninth IEEE International Conference, **(2003)**, Nice, France, pp. 432-439.
[18] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **(2007)** , pp. 1-8.
[19] B. C. Russell, A. A. Efros and J. Sivic, "Sing multiple segmentations to discover objects and their extent in image collections", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **(2006)**, pp. 1605-1614.
[20] V. Vapnik, "The Nature of Statistical Learning Theory, *Springer*, New York, (**1995).**
[21] C. Scott, and R. D. Nowak, "Minimax-Optimal Classification With Dyadic Decision Trees", IEEE Transactions on Information Theory, **(2006),** pp. 1335-1353.
[22] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words", International Journal of Computer Vision, vol. 79, no. 3, **(2008),** pp. 299-318.
[23] R. Tanawongsuwan and A. Bobick, "Modelling the Effects of Walking Speed on Appearance-Based Gait Recognition", Proc. IEEE Intl Conf. Computer Vision and Pattern Recognition, **(2004)**, vol. 2, pp. 783-790.
[24] Y. Wang and G. Mori, "Human action recognition by semi-latent topic models", Pattern Analysis And Machine Intelligence, vol. 31, no. 10, **(2009),** pp. 1762-1774.
[25] A. Gruber, M. Rosen-Zvi, and Y. Weiss, "Hidden Topic Markov Models", Artificical Intelligence and Statistics, **(2007).**

Authors

Hongmin Xue, he is currently a Ph.D. candidate at Xidian University. His research interest focuses on the fields of vision computing technologies and network ultimedia technologies.

Zhijing Liu, he is a professor and advisor for doctoral students. He currently serves as head of the Research Center of Computer Information Research Application, and is director of the China and America Associated Laboratory of key Technologies of Mobile Electronic Commerce. His research work focuses on the fields of vision computing technologies, network multimedia technologies, technologies of virtual reality, and key technologies of E-government and E-commerce. He has been in charge of or taken part in more than 40 scientific researches at national, provincial and ministerial levels, and won two National Science and Technology Production Awards. Besides, he obtained eight Science and Technology Production Awards at provincial and ministerial levels. He has published more than 110 technological papers on internal and external kernel publications, including 30 ones searched by SCI, EI and ISTP respectively. Presently, he acts as a member of the Expert Advisory Committee of the leading group of the informatization

of Shaanxi province, and is a fellow of the Committee of Experts of Manufacturing Informatization of Shaanxi province, committeeman of Xi'an manufacturing informatization Expert Committee, and appraisal expert of the Committee of Awards of Enterprise Technology Innovation of Shaanxi province.

Jing Xiong, she is a Ph.D. candidate at School of Computer Science and Technology in Xidian University. She received her B.S degree from Xidian University in 2005, the M.Sc degree in computer science from Xidian University in 2008. Her research interest is image processing and visual computing.

