Recognizing Human Activities by Key Frame in Video Sequences

Hao Zhang*, Zhijing Liu and Haiyong Zhao
School of Computer Science and Technology, Xidian University, Xi'an, P.R.China
Email: zhanghao@mail.xidian.edu.cn, liuzhijing@vip.163.com,
zhaohaiyongym@163.com

Guojian Cheng
School of Computer, Xi'an Shiyou University, Xi'an, P.R.China
Email: gjcheng@xsyu.edu.cn

***Abstract*—This paper presents a new method of human activity recognition, which is based on $\Re$ transform and dynamic time warping (DTW) after the key frame is extracted from a cycle. For a key binary human silhouette, $\Re$ transform is employed to represent low-level features. The advantage of the $\Re$ transform lies in its low computational complexity and geometric invariance. The DTW distance based on the extracted features are calculated and compared similarities to recognize activities. Compared with other methods, ours is superior because the descriptor is robust to frame loss in the video sequence, disjoint silhouettes and holes in the shape, and thus achieves better performance in similar activities recognition, simple representation, computational complexity and template generalization. Sufficient experiments have proved the efficiency.**

***Index Terms*—feature extraction, activity recognition, $\Re$ transform, Dynamic Time Warping (DTW), key frame**

I. INTRODUCTION

Human activity recognition is one of the most appealing, yet challenging problems of computer vision [1]. Reliable and effective solutions to this problem would be highly useful in many areas, such as behavioral biometrics [2], contend-based video analysis [3], security and surveillance [4,5]. However, current methods to these problems are very limited, and understanding human behavior by computers remains unresolved.

Human activity recognition has been a widely studied topic, but the solutions to this problem that have been submitted to date are very premature and still specific to the dataset at hand. Building a general and effective activity recognition and classification system is a challenging task, due to the various parameters from the environment, objects and activities. Variations in the environment can be caused by cluttered or moving background, camera motion, occlusion, weather and illumination. Variations in the objects are caused by differences in appearance, size or posture of the objects or due to self-motion which is not itself part of the activity. Variations in the activity can make it difficult to recognize semantically equivalent activities.

This paper describes a simple and unique representation of activities with $\Re$ transform as the template, since they are detected and extracted based on the whole image from the video sequence, rather than some parts of the image. This is contrary to the work of Gorelick et al [6], where a whole image sequence is represented as a space-time shape. Furthermore, the use of dynamic time warping (DTW) computes the distance between the template and the key frame of the subjects, as opposed to the work [6,7], where features are matched based on the maximum similarity over a whole video sequence. We provide a comparable result and show the advantages with the state of the art methodologies.

The rest of the paper is organized as follows. Section 2 provides a brief literature review on DTW-based human activity recognition. Subsequently, we present the details of our method in Section 3, including key frame detection, feature extraction and modeling, activity recognition. After that, we analyze the experiment results with different data and compare them with state-of-the-art methods. In the last section, we draw the conclusions.

II. RELATED WORKS

Approaches for modeling activities can be categorized into three major classes—nonparametric, volumetric, and parametric time-series approaches. Nonparametric approaches, including 2-D templates, 3-D object models and manifold learning methods, typically extract a set of

Manuscript received September 21, 2009; revised December 5, 2009; accepted February 10, 2010.

Based on "Key-frame based Human Activity Recognition", by Hao Zhang, Zhijing Liu, Haiyong Zhao, and Guojian Cheng, which appeared in the Proceedings of 2009 WASE Global Congress on Science Engineering, Taiyuan, China, December 25-27, 2009. © 2009 WASE.

This research was supported by funds from National Natural Science Foundation of China (NSFC) under Grant 40872087.

*corresponding author.

doi:10.4304/jsw.5.8.818-825

features from each frame of the video. The features are then matched to a stored template.

There are two representative methods in 2-D templates methods. Polana and Nelson [8] proposed one of the earliest attempts at activity recognition. In their work, the moving actor can be segmented, normalized spatially and temporally, and recognized by matching against a spatio-temporal template of motion features. It is essentially a three dimensional template, including X, Y and T. Bobick and Davis [9] proposed "temporal templates" as models for activities. Its basis of the representation is a temporal template — a static vector-image where the vector value at each point is a function of the motion properties at the corresponding spatial location in an image sequence.

DTW is a flexible and effective approach based on model to recognize human activity. Kale et al [10] extracted different gait features from the width vector and used the DTW approach for matching. Though the width vectors are simple in extraction and representation, they are less distinctive and effective in representing and recognizing complex activities, such as walking, running, bending, jumping, crouching and fainting.

Chen et al [11] proposed an approach for recognition and matching the human behavior sequence. The key postures in each sequence is represented and normalized for matching. A DTW algorithm is used to perform the alignment between two time series. The postures images are defined as model-based template, so the measure of similarity is difficult and the computational complexity is higher.

Oikonomopoulos et al [12] extracted a set of visual descriptor derived from spatiotemporal salient points and provide local space-time description of the visual activity. These descriptors are inherently translation invariant. Subsequently, a clustering algorithm is used in order to cluster our descriptors across the whole dataset and create a codebook of visual verbs. Then they used DTW to align the sequences in the dataset and structure in time. Though this method is extremely effective for image and video retrieval applications, the procedure of creating codebook is inconvenient to build a system.

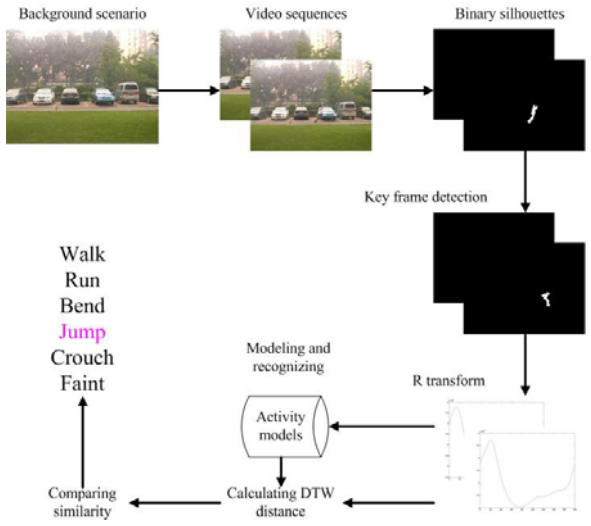

Figure 1. The flowchart of activity recognition.

Ikizler and Duygulu [13] proposed a pose descriptor named Histogram-of-Oriented-Rectangles (HOR) for representing human activities and using DTW to carry the information from the spatial domain in videos. This descriptor is not sufficiently strong for all activities in test database, so a simple velocity descriptor as a prior to the pose based classification step is incorporated necessarily. Therefore, its recognition in performance is not general for human activities.

In this paper, we introduce the simple pose descriptor, $\Re$ transform, and DTW distance to representation and recognition. We extract the key frames from each cycle of video sequences and represent their features. Then some of them, termed "gallery", are used to model. The others, termed "probe", are done to test. We present comprehensive results on datasets, and elaborate on certain value choices of the template. We also compare our method with its counterparts and provide the result evaluations of our method.

III. Our Approach

In [14], Schindler and Gool analyzed the features and rules of human activity and discussed how many frames are necessary in video sequences for effectively recognizing it. They demonstrated that the accuracy of recognition is nearly 90% by single frame over WEIZMANN and KTH datasets. The more information the key frame contains, the higher accurate rate of recognition is obtained. Therefore, we present a method that represents and recognizes human activities based on key frame over Schindler and Gool's work.

Our method contains three parts, including key frame detection, feature extraction and modeling, activity recognition. The flowchart is shown in Fig. 1.

A. Activity Analysis

a. Activity Description

The videos used are taken from activity database shot by Institute of Automation, Chinese Academy of Sciences (CASIA). In this database, there are two sets of activities, including single person and two interactive persons. Each of them is screened in 3 visual angles, i.e. horizontal, vertical and angle. The dataset of single person contains 8 classes per angle, including 11 or 16 flips (320×240, 25fps) respectively. All flips are shot outdoors by stationary camera, including 16 persons in 8 types of activities. In our system, the video flips of single person in horizontal view, i.e. walk, run, bend, jump, crouch and faint, are used for experiment, as shown in Fig. 2.

b. Feature Analysis and Key Frame Extraction

In human motion, we assume that frame number is t, the length of silhouette in horizontal and vertical are $w(t)$ and $h(t)$ respectively, and their ratio is r, so we define it as

$$r = \frac{w(t)}{h(t)}.$$

(a) Walk

(b) Run

(c) Bend

(d) Jump

(e) Crouch

(f) Faint

Figure 2. The sequences of activities.

We choose one flip from each class to analyze human activities. In order to do them accurately, we utilize every single cycle from the flips, which is separated by the dashed, as shown in Fig. 3. From this figure, we can find that the activities i.e. walk, run and bend are periodic obviously, as shown in Fig. 3(a,b,c). The minima in *r*s are determined to the separate point in two cycles, and the maximums to the key frame as below

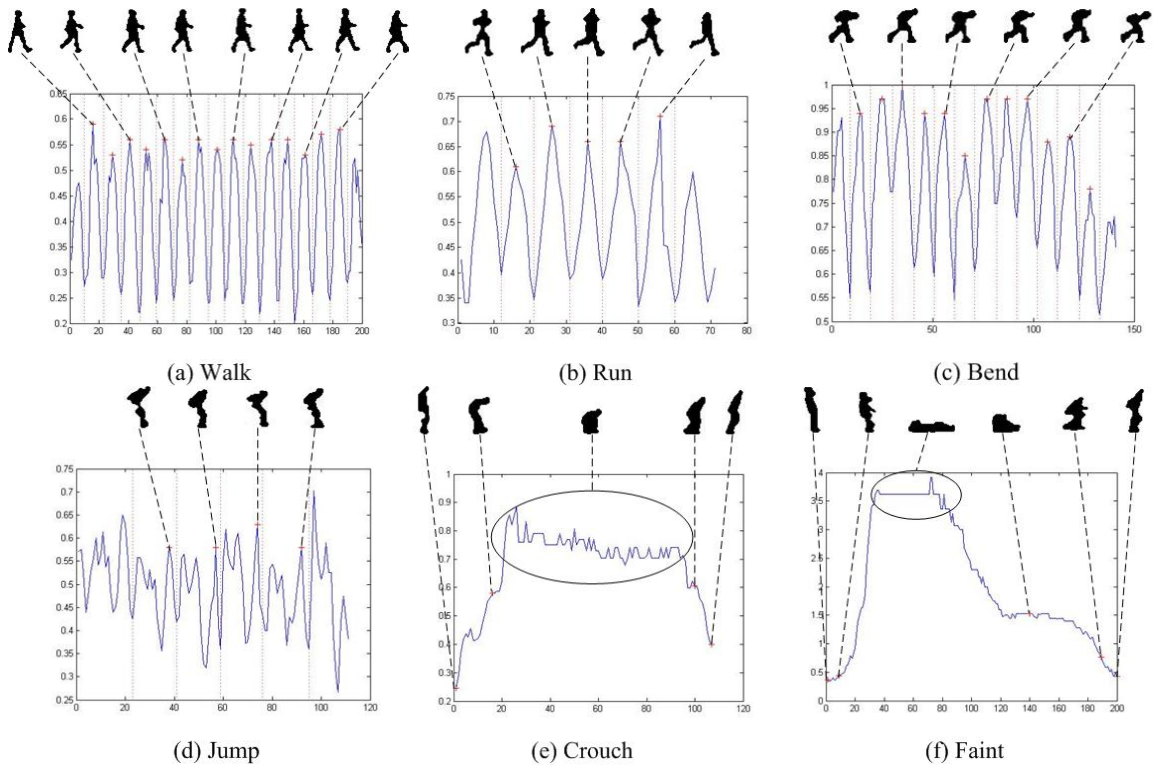

Figure 3. The analysis of activities cycle.

$$keyframe = \arg\max_{t}\{r\}.$$

As it is variable in jumping, as shown in Fig. 3(d), its rule is less distinctive than the former three, but the key frames are determined by the maximum of r in each cycle. Though there is only one cycle in the sequences of crouch and faint, the key frames are also determined by the method above.

B. Feature Extraction

We extract and represent activity features by $\Re$ transform [15,16], which is defined by

$$\begin{aligned} & T_{R^f}(\rho,\theta) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(x\cos\theta + y\sin\theta - \rho)dxdy, \quad (1) \\ &= R\{f(x,y)\} \end{aligned}$$

$$\Re_f(\theta) = \int_{-\infty}^{\infty} T_{R^f}^2(\rho,\theta)d\rho, \quad (2)$$

where $\theta \in [0,\pi]$, $\rho \in [-\infty,\infty]$ and $\delta(.)$ is the Dirac delta-function,

$$\delta(x) = \begin{cases} 1 & if\ x = 0 \\ 0 & otherwise \end{cases}$$

Therefore, $\Re$ transform is robust to geometry transformation, which is appropriate for activity representation. According to [17], $\Re$ transform outperforms other moment based descriptors, such as Invariant moment, Zernike moment and Wavelet moment, on similar but actually different shape sequences, and even in the case of noisy data.

After extracting the key frames, we subtract the median background and remove with a median filter by a 3×3 template. A predetermined threshold is used to obtain the binary images. Fig. 4 shows the key frame silhouettes of six activities in the first column, and their respective $\Re$ transform. The figures in the second column show the difference of each activity. In third column, they are the $\Re$ transform curves of the key frames in the same sequence. Comparing with six activities, we can conclude that there are many differences. Though the type of walk is similar to run, their gap is more than twice. It can be seen from the third column that the key frame in the same activity is more similar. Moreover, we can take the unique model based on the key frame in each activity for recognition.

After extracting the key frames from the videos, we represent their features with $\Re$ transform and modeling by the average of $\Re$ transform. The red curve in the third column is shown in Fig. 4. The activities in dataset, i.e. walk, run, bend, jump, crouch and faint, are defined by 1 to 6 respectively.

C. Activity Recognition

After modeling, we use DTW as a classifier to measure the similarity. The DTW was proposed by Bellman in the 1950s. Its advantage is that the optimal result can be obtained in low computational complexity, in other words, it is dynamic time warping in matching

time sequences [18]. Therefore, we define two time sequences as

$$A = a[1], a[2], \cdots, a[m]$$
$$B = b[1], b[2], \cdots, b[n]$$

The distance between a[i] and b[j] is defined by $d(i, j) = a[i] - b[j]$, so the total distance between A and B is $DIST(A,B)$, which can be obtained by (3) below.

$$\begin{cases} D(1,1) = d(1,1) \\ D(i,j) = d(i,j) + \min \begin{Bmatrix} D(i-1,j) \\ D(i-1,j-1) \\ D(i,j-1) \end{Bmatrix}, \\ DIST(A,B) = D(m,n) \end{cases} \quad (3)$$

where $D(i,j)$ is the distance between $A(i)$ and $B(j)$, $A(i)$ is $a[1],a[2],\cdots,a[i]$ and $B(j)$ is $b[1],b[2],\cdots,b[j]$. The smaller the value of $DIST(A,B)$ is, the more similar two sequences are.

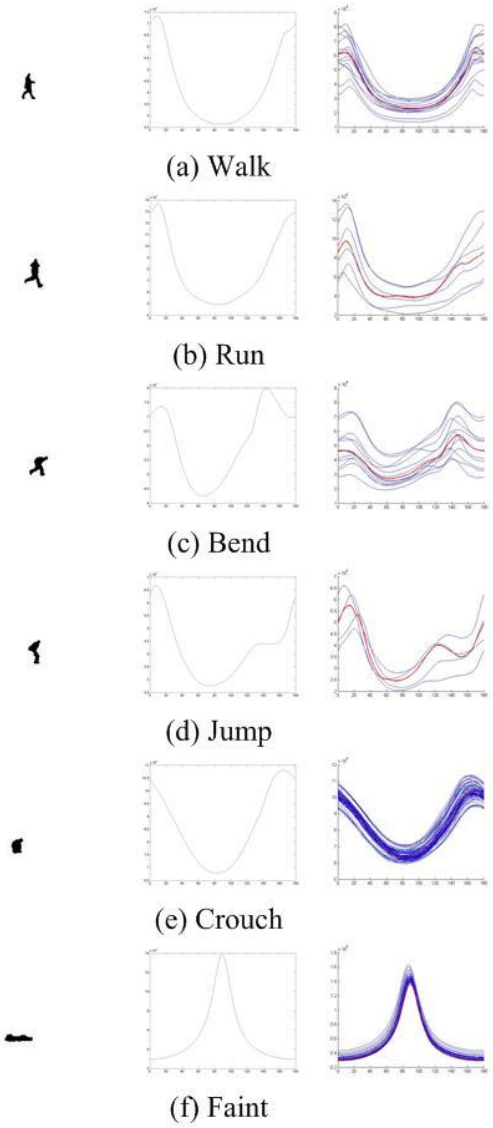

Figure 4. The $\Re$ transform of single and multiple key frames for different activities

The shape of $\Re$ transform is one dimensionality consisting of 180 values. In order to reduce computational cost, we select the points with equal interval as feature to calculate DTW distance. The data is divided into v groups, each of which includes u points, so

$$v = \frac{180}{u}$$

where $u,v \in Z^+$. We assume that p is the sequence number of feature points, and R_s and T are the $\Re$ transform values of gallery and probe respectively. Therefore, as $p = w \cdot u$, $0 \leqslant w < v$, the algorithm of recognition is shown below.

Function ActivityRecognition()
Input:
R(s)：The gallery data of 6 activities with $\Re$ transform
T：The probe data of the key frame with $\Re$ transform
Output:
activitynumber: the number of recognizing result
{
  for activity from run to faint
    DIST[activity]=DTW(R[activity], T);
      //calculate DTW distance
  end;
  $activitynumber = \arg\min_{activity} \{DIST[activity]\}$
  return activitynumber;
}

IV. Experiment Analysis and Discussion

A. Experiment Data

In experiments, the number of key frame in 6 types of activities is shown in Table 1. Each category is divided into two parts, including modeling and test.

The resultant silhouettes contain holes and intrusions due to imperfect subtraction, shadows and color similarities with the background. To train the activity models, holes, shadows and other noise are removed manually. The synthetic data are taken as ground truth

TABLE I. The Statistics of Experimental Data

Number of frame	Walk	Run	Bend	Jump	Crouch	Faint
Total key frame	125	85	126	38	8	8
Gallery frame	16	7	13	6	1	1
Grobe frame	109	78	113	32	7	7

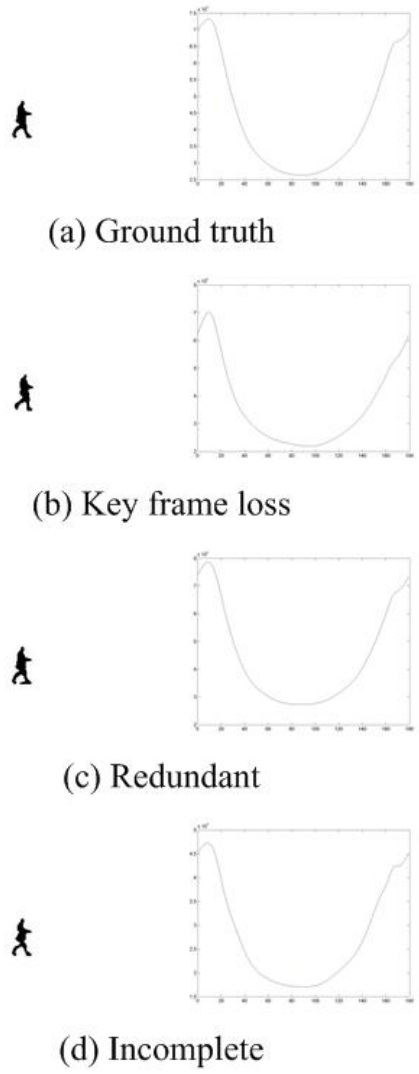

Figure 6. The $\Re$ transform for data of different qualities

data.

The raw data include such cases as disjoint silhouettes, silhouettes with holes and silhouettes with missing parts. Compared with the ground truth data, they are incomplete data.

Shadow and other noises may add an extra part to the human silhouette, and thus induce redundant data.

The incomplete data and redundant data are of low quality, and thus they are used for testing the performance of the $\Re$ transform. Fig. 6 shows some such examples.

In order to test the robustness of $\Re$ transform, we select the frame with less feature information rather than the frames shown in Fig. 4. In other words, we use the frames near the key one to substitute. These artificially generated data are defined as key frame loss data.

Fig. 6 shows the $\Re$ transform of the walking shape in different data case. For the cases of incomplete data and ground truth data, the $\Re$ transform is similar, but the transform of redundant data varies significantly in the peak of the curve. In fact, $\Re$ transform is sensitive to the redundant data, which will have negative effect on activity recognition.

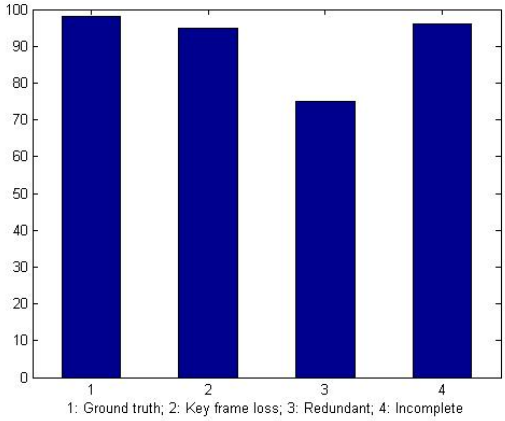

Figure 5. The recognition rates of $\Re$ transform with different data.

TABLE II. THE RECOGNITION RATE OF SIX ACTIVITIES

Activity Type	Walk	Run	Bend	Jump	Crouch	Faint
Recognition rate (%)	97.25	96.15	100	93.75	100	100

B. Result Analysis

Fig. 5 illustrates the recognition results with $\Re$ transform and DTW distance in different data. Note that there are more activity sequences in each probe class, which are different with gallery data. Table 2 concludes that, in spite of the high similarities between running and walking, misclassifications never occur in the case of key frame loss data and incomplete data. The $\Re$ transform descriptor captures both boundary and internal content of the shape, so they are more robust to noise, such as internal holes and separated shape. While in the case of silhouette with shadow, the performance of $\Re$ transform is slightly worse than other cases. This shows that $\Re$ transform is suitable for the background segmentation methods with low false positive rate but keeping some false negative rate [19]. Generally speaking, low level features based on $\Re$ transform are effective for recognizing similar activity even in the case of noisy data.

C. Comparison

As shown in Table 3, our method not only outperforms other three methods, but also has three advantages as follow. Firstly, it has lower computational complexity. The key frames extracted from video sequences are utilized in feature matching, so that it costs less time compared with Chen's work [11], which matches features in two sequences. Secondly, its representation is simple. Though Oikonomopoulos' work [12] represents activities with a codebook in details, it is more complex and ambiguous than our method using $\Re$ transform descriptor. Finally, our method is general for

TABLE III. COMPARISON OF OUR METHOD TO OTHER METHODS THAT HAVE REPORTED RESULTS

Method	Average accuracy (%)
Chen et al [11]	85.9
Oikonomopoulos et al [12]	93.0
Ikizler and Duygulu [13]	94.7
Our method	97.8

different activities. We only use one descriptor, $\Re$ transform, to represent six different activities and extract their features, while Ikizler's work [13] uses HOR and velocity to implement it. In general, our method has made significant improvement in many aspects.

V. CONCLUSION

In this paper, we have approached the problem of human activity recognition and used a new pose descriptor based on $\Re$ transform and a similarity measurement based on DTW. Our pose descriptor is simple and effective, since we extract activity features from a human silhouette in the whole image and form a temporal template. We show that, by effective classification of such templates, reliable human activity recognition is possible. We demonstrate the effectiveness of our method over the state-of-the-art datasets in activity recognition literature, which is CASIA dataset. Our results are intuitively comparable and even superior to the results presented by the pioneers.

Our experiments show that the human pose encapsulates many useful pieces of information about the activity itself, and therefore one can start with a good pose estimator on a key frame, before going into the details of dynamics. We show how we can obtain efficient activity recognition with the minimum of dynamics information and complex modeling. Result is an intuitive and fast activity recognition system with high accuracy rates, even in challenging conditions.

ACKNOWLEDGMENT

The authors would like to thank CASIA to provide activity database and the anonymous reviewers for their constructive comments.

REFERENCES

[1] Pavan Turaga and Rama Chellappa, V. S. Subrahmanian, Octavian Udrea, "Machine Recognition of Human Activities: A Survey," IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 11, pp. 1473-1488, November 2008.

[2] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The Human ID gait challenge problem: Data sets, performance, and analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 2, pp. 162–177, February 2005.

[3] Y. Rui, T. S. Huang, and S. F. Chang, "Image retrieval: Current techniques, promising directions and open issues," J. Visual Commun. Image Represent., vol. 10, no. 1, pp. 39–62, March 1999.

[4] N. Vaswani, A. K. Roy-Chowdhury, and R. Chellappa, "Shape activity: A continuous-state HMM for moving/deforming shapes with application to abnormal activity detection," IEEE Trans. Image Process., vol. 14, no. 10, pp. 1603–1616, October 2005.

[5] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2004, pp. 819–826.

[6] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," IEEE Trans. Pattern Anal. Mach Intell., vol. 29, no. 12, pp. 2247-2253, December 2007.

[7] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A Biologically Inspired System for Action Recognition," in Proc. IEEE Int. Conf. Computer Vision, 2007.

[8] R. Polana, and R. C. Nelson, "Detection and recognition of periodic, nonrigid motion," Int. J. Comput. Vis., vol. 23, no. 3, pp. 261–282, 1997.

[9] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 3, pp. 257–267, March 2001.

[10] A. Kale, N. Cuntoor, B. Yegnanarayana, A.N. Rajagopalan, and R. Chellappa, "Gait analysis for human identification," Lecture Notes in Computer Science, vol. 2688, pp. 706-714, 2003.

[11] Yan Chen, Qiang Wu, and Xiangjian He, "Using Dynamic Programming to Match Human Behavior Sequences," in International Conference on Control, Automation, Robotics and Vision, ICARCV, 2008, 1498-1503.

[12] A. Oikonomopoulos, M. Pantic, and I. Patras, "B-spline Polynomial Descriptors for Human Activity Recognition," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008.

[13] Nazli Ikizler and Pınar Duygulu, "Histogram of oriented rectangles: A new pose descriptor for human action recognition," Image and Vision Computing, vol. 27, no. 10, pp. 1515-1526, September 2009.

[14] Konrad Schindler and Luc van Gool, "Action Snippets: How many frames does human action recognition require," in IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[15] Deans, S. R.: Applications of the Radon Transform. Wiley Interscience Publications, Chichester, 1983.

[16] S. Tabbone, L. Wendling, and J.-P. Salmon, "A new shape descriptor defined on the Radon transform," Comput. Vision Image Understanding, vol. 102, no. 1, April 2006.

[17] Ying Wang, Kaiqi Huang, and Tieniu Tan, "Human Activity Recognition Based on $\Re$ Transform," in Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit., pp. 3722-3729, 2007.

[18] Sakoe H and Chiba S, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans Acoust. Speech Signal Process., vol. 26, no. 1, pp. 43-48, 1978.

[19] M Karaman, L Goldmann, and TS Da Yu, "Comparison of Static Background Segmentation Methods", Proceedings of SPIE, 2005.

Hao Zhang was born in Qingdao, Shandong, P.R.China, 1980. He received the B.Eng. degree in computer science and technology from Qingdao Institute of Architecture and Engineering, Qingdao, Shandong, P.R.China, in 2003. Then he received the M.Eng. degree in computer application from Xi'an Shiyou University, Xi'an, Shaanxi, P.R.China, in 2007. He is currently working toward the Ph.D. degree in the School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China. His current research interests include computer vision, pattern recognition, image processing, and their application in activity recognition and classification.

Zhijing Liu was born in Xi'an, Shaanxi, Province, P.R.China, in 1957. He received the B.Eng. degree in computer engineering from Institute of Northwestern Telecommunications Engineering, Xi'an, Shaanxi, P.R.China, in 1982. His major field of study are computer vision and data mining.

He has been teaching and researching at Xidian University since 1982. He is a professor and a supervisor for Ph.D students in Xidian University. He has published more than 100 papers.

Prof. Liu has acted as 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 the city of Xi'an manufacturing informatization Expert Committee, and appraisal expert of the Committee of Awards of enterprise technology innovation of Shaanxi province.

Haiyong Zhao was born in Liaocheng, Shandong, P.R.China, in 1981. He received the B.Eng. degree in computer science and technology from Shandong University of Technology, Zibo, Shandong, P.R.China, in 2004. Then he received the M.Eng. degree in software engineering from Shandong University, Shandong, P.R.China, in 2007. He is currently working toward the Ph.D. degree in the School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, P.R.China. His current research interests include computer vision, object detection and image processing, and their application.

Guojian Cheng was born in Fufeng, Shaanxi, P.R.China, in 1964. He received the B.S. degree in computer application from China University of Petroleum, Dongying, Shandong, P.R.China, in 1990. Then he received the M.S. degree in computer application from Xidian University, Xi'an, Shaanxi, P.R.China, in 1996. He received the Ph.D degree in computer application from Tuebingen University, Tuebingen, Germany, in 2001. His current research interests include neural network, genetic algorithm, pattern recognition, reservoir simulation with soft computing method.

He worked in Research Center of DaimlerChrysler AG from March 2002 to March 2003. He has been teaching and researching in Xi'an Shiyou University since September 2004.