An X-T Slice Based Method for Action Recognition

Yanhu Shan, Shiquan Wang, Zhang Zhang, Kaiqi Huang
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, P.R.China
{yanhu.shan, sqwang, zzhang, kqhuang}@nlpr.ia.ac.cn

Abstract

This paper proposes a novel method for human action recognition. Different from many action recognition methods which consider an action sequence along the time axis, the proposed method views an action sequence along the space axis. This brings two advantages: the human body structures in all frames are encoded in the feature; the time information is completely used. The process of feature extraction is as follows: first an action sequence is cut into slices parallel to the X-T plane. Every slice, we call X-T slice, is transformed to a mean histogram and a variance histogram along the T axis. Then all mean histograms and all variance histograms are concatenated separately to two vectors, and finally encoded with Mel Frequency Cepstrum Coefficient (MFCC). MFCC, a feature commonly used in speech recognition, can effectively capture changes of 1-D signals over time. The encoded values are sent to classifier for action recognition. Our system achieves very efficient result: it needs only 0.02 second to deal with a frame on average with Matlab.

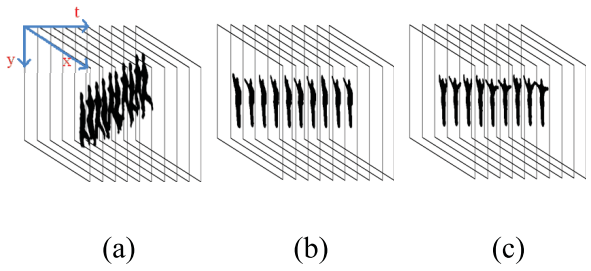

Figure 1. 3D silhouette volumes of three actions. (a) "running"; (b) "wave one hand"; (c) "wave two hands".

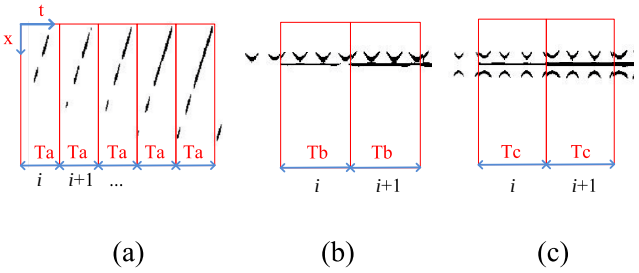

Figure 2. Three action X-T slice sequences corresponding to Figure 1. i is the coordinate on the Y axis. Ta, Tb, Tc represents the length of the three action volumes along T axis.

1. Introduction

Human activity analysis attracts great attention in recent years because of its wide application prospect, e.g., intelligent video surveillance, human-computer interaction, sport and entertainment video analysis, etc. It is also a challenging problem in computer vision. Human activity can be divided into low-level action, such as "run", "walk" and "jump", and high-level activity, such as "fight" and "loiter". We focus on low-level action recognition in this paper. Human action recognition systems usually contain the following procedures: human detection, feature extraction, motion representation and action recognition. In these steps, human detection itself is an independent research area and has gained promising achievements in recent year[25][6], thus in this paper, our work focuses on the rest steps.

There are two categories of methods for human action recognition: space-time approaches and sequential approaches [1]. Space-time approaches view an action as a 3-D volume while sequential approaches treat it as a sequence of observations.

In this paper, we introduce a new action representation method. We follow the space-time approaches to view an action as a 3-D silhouette volume. Our method is motivated by describing an action sequence as a whole and using the fixed structure characteristics of the human body. We take three actions, "run", "wave one hand" and "wave two hands", as shown in Figure 1 for example to illustrate our method.

Three actions are represented as 3-D silhouette volumes. People commonly view the action volumes along the time

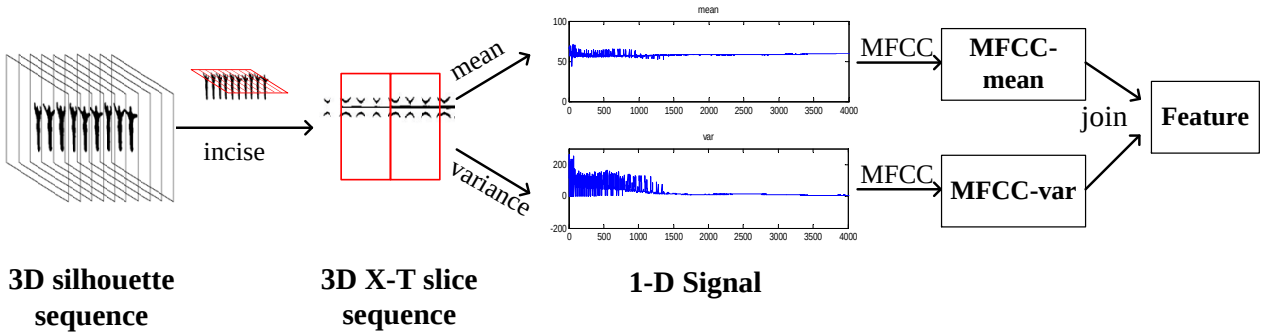

Figure 3. Flowchart of the MFCC feature extraction.

axis, and the information we get is the transition of different poses as shown in Figure 1. It can directly describe an action and follow common sense how people view an action. As 3-D volumes, we can also view them along the X and Y axis to observe the changes of body structure. It is easy to imagine that viewing along Y axis can highlight the characteristics of body structure differences better than viewing along X axis, because it has a larger interval(or a higher resolution) of observation. Moreover, viewing the 3-D volume along X axis reveals less information because of the symmetry characteristic along Y axis of human body. So we cut the silhouette sequence of an action video perpendicular to Y axis, and then, we acquire the images shown in Figure 2 by combining all the cut images one by one. The details of this method to obtain Figure 2 from Figure 1 are introduced in later sections. It is difficult to classify actions if we directly use an image as a feature due to its variable size. Mean and variance can describe the stable and variable information, so we calculate the mean and variance of coordinate position with non-zero values on on X axis to get a pair of one-dimensional signals. We find that the signals we get are similar with voice signals. Inspired by this, we try to classify actions with methods used in speech recognition. MFCC feature widely applied in speech recognition is used here as the representation of action. It can describe the frequency information of a 1-D single. The final feature of an action is produced by concatenating the two MFCC features together.

The feature extracted by our method has three advantages:

- Human has a fixed order from head to feet, and this method utilizes human's body structure feature efficiently;
- the uncertainty of time can just natural handled by MFCC feature. The length of feature is decided by the number of filters in MFCC;
- we need not to extract features in every frame and classify them into different poses.Our method treats an action sequence as a complete unit, so we can grasp the characteristic of an action as a whole.

Multi-class SVM classification is used to classify various actions. We adopt leave-one-out cross validation method to test the recognition ratio of different actions.

The rest of this paper is organized as follows. We review related work on action recognition in Section 2, and introduce the details of our approach in Section 3. Section 4 presents the experimental results and discusses related issues. We conclude our work in this paper and make a future plan in Section 5.

2. Related work

There are a great many methods of action recognition. Aggarwal et al.[1] sort out the typical methods used in action recognition and classify them into two types.

One type is space-time approach. Space-time approaches recognize human actions by analyzing space-time volumes of action videos. Ke et al.[10] use over-segmented valumes, automatically calculating a set of 3-D XYT volume segments corresponding to a moving human. Bobick and Davis[3] represent each action with a motion-energy image (MEI) and a motion-history image (MHI). These two methods are based on space-time volumes. Although 3-D volumes are also used in the experiment of this paper,the angle is quite different. Sheikh et al.[18] describe an action with a set of 13 joint trajectories in a 4-D XYZT space and use an affine projection to obtain normalized XYT trajectories of an action for the purpose of measuring the similarity between two sets of trajectories. Some other approaches[8][15][24] utilize space-time local features to recognize actions, and the most representative one is sparse spatiotemporal interest points[11].

The other type is sequential approach. Sequential approaches recognize human actions by analyzing sequences of features. They consider an input video as a sequence of observations. One representative catego-

ry of this type is exemplar-based approaches. The dynamic time warping (DTW) algorithm has been developed and widely used in matching two sequences in lots of work[7][21][9]. Some other exemplar-based methods are also proposed, such as decomposing signals with singular value decompositions (SVD)[22] and modeling human activities as linear time invariant (LTI)[13]. The other category of sequential approach represented by hidden Markov models (HMMs)[23][4] and dynamic Bayesian networks (DBNs)[16][17] is based on state model. Some extended methods of the category are proposed, such as decomposing an efficient recognition algorithm using coupled hidden semi-Markov models (CHSMMs)[14].

3. Our method

Out approach includes two parts, i.e, feature extraction and action classification. The core of our work is feature extraction, and it is presented in part one. The other part briefly shows action classification.

3.1. Feature extraction

The flowchart of feature extraction is shown in Figure 3. We divide this part into two phases: the first phase is acquiring the X-T slice sequence from a 3-D silhouette volume as shown in figure 2, and transforming this 2-D slice sequence into two 1-D signals by calculating the means and variances of coordinate position with non-zero values on X axis; the second phase is extracting MFCC features of the two 1-D signals separately and joining them together as one vector. The vector is what we used as the feature of the action.

3.1.1 X-T slice and 1-D Signals

Silhouette of human body is the basic information we used in this paper. One action sequence, "wave two hands", is took here as an example to explain our method. The images of foreground sequence are joined together into a 3-D sequence volume as shown in Figure 4(a). X axis and Y axis are image coordinate, and T axis is time axis. Based on the 3-D volume, we cut it in X-T plane along the direction of Y-axis and call the slices X-T slices. One of the X-T slices responses to the small rectangular area with the length T as shown in Figure 4(b). T indicates the number of frames in an action sequence. The second rectangular area is the slice below the foregoing one. These slices are joined one by one and finally form a long slice sequence. It should be noted that the image showing in Figure 4(b) is just a fragment in the long slice sequence.

Suppose the size of a slice sequence $V_{sequence}$ was m by n. So there are n columns, and every column is m by 1. The mean and variance of every column are calculated and concatenated together separately, and we get two 1 by n

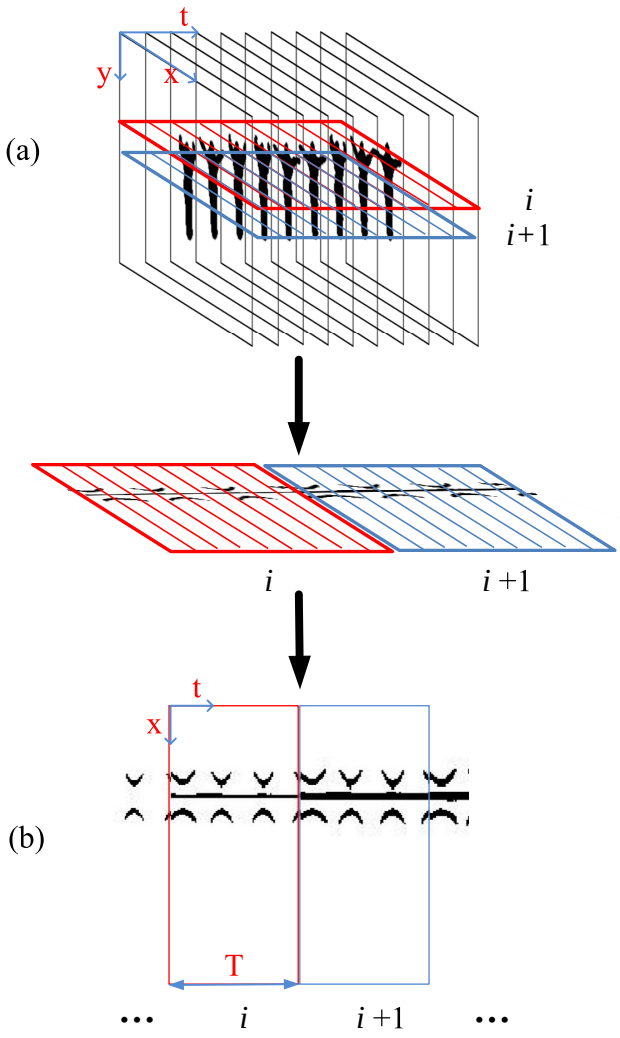

Figure 4. Diagram of the transition from Figure 2(c) to Figure 3(c).

vectors. As lots of the values in the beginning (corresponding to the the area above head) and end (corresponding to the area below feet) of the sequence are zeros, only non-zero columns are kept to reduce the dimension of vectors. The final mean and variance vectors of the foregoing three actions are shown in Figure 5.

3.1.2 MFCC feature

Mel Frequency Cepstrum Coefficient (MFCC) is well known for its application in speech recognition. Terasawa et al.[19] derived MFCC feature according to the flowchart shown in figure 6. MFCC is the Fourier transform of a spectrum on the logarithmical scale. Similar to video, an audio single is also separated into different frames as shown in Figure 7. Every frame is a unit for MFCC, and overlap exists between adjacent two frames. As the size of over-

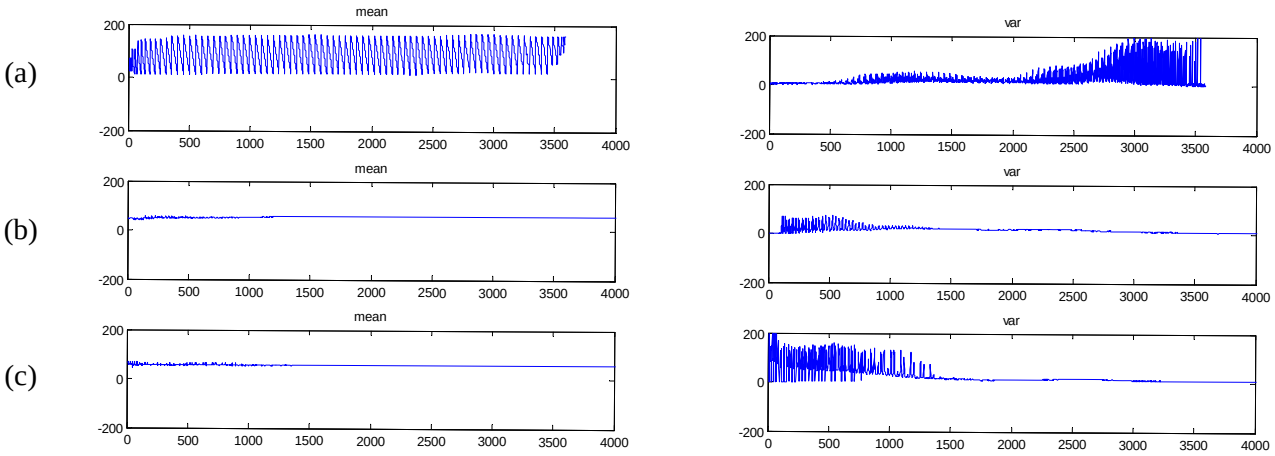

Figure 5. Non-zeros columns of means and variances of the foregoing three X-T slice sequences.

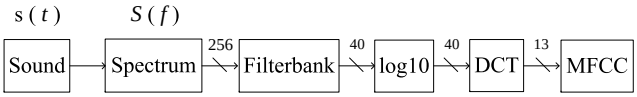

Figure 6. Flowchart of MFCC algorithm. The numbers between the blocks indicate the dimensionality of the data. $S(f)$ is the spectrum of an audio signal $s(t)$.

lap is related to the sample rate (S_r) and the rate of sampling frames(F_s) in the whole signal, we extract the length $L_{audio}(i)$ and sample rate Sr_{audio} from M standard voice units. Sr_{video} , which is used as both V_{mean}'s and V_{var}'s sample rate, is

$$Sr_{video} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{L_{mean}(j)}{L_{audio}(i)}Sr_{audio}(i) \quad (1)$$

where N is the number of video sequences we used here, and L_{mean} is the length of V_{mean}. Sr_{video} is also used as V_{var}'s sample rate by the reason of minor difference between the length of V_{mean} and V_{var}.

We extract MFCC feature of a frame as follows. Assuming $x(n)$ is a signal of one frame, we firstly windows the signal with a time windows $\omega(n)$. We use Hamming windows, where $\omega(n) = 0.54 - 0.46\cos(\pi n/N)$, for convenience. Then we calculate the FFT of discrete-time signal $x(n)$ with length N, given by

$$X(k) = \sum_{k=1}^{N}\omega(n)x(n)\exp(-j2\pi kn/N) \quad (2)$$

where $k = 0, ..., N-1$. The following step is contributing Mel filter bank. The Mel filter band is a collection of triangular filters. The number of filter banks is M. The width of every filters is defined as

$$Bwidth(H_i) = \begin{cases} 133.3 & (i \le 13) \\ 1000 \times 1.072^{i-13} & (i > 13) \end{cases} \quad (3)$$

where $i = 1, ..., M$ is the number of filter banks. The total energy from each filter is

$$E_i = \sum_{k=0}^{N-1}|X(k)| \cdot H_i(k) \quad (4)$$

The MFCC of different filter banks are calculated by taking DCT of the log-scaled filter bank output, given by

$$C_i = DCT\left\{\log_{10}(E_i)\right\} \quad (5)$$

MFCCs of all filter banks are used here.

We use the algorithm mentioned above to acquire the MFCC features of all frames in both V_{mean} and V_{var}, and calculate the mean of the MFCC features, F_{mean} and F_{var}, separately.

$$F_{mean} = \frac{1}{N^f_{mean}}\sum_{k=1}^{N^f_{mean}} MFCC\{frame_{mean}(k)\} \quad (6)$$

$$F_{var} = \frac{1}{N^f_{var}}\sum_{k=1}^{N^f_{var}} MFCC\{frame_{var}(k)\} \quad (7)$$

Where N^f_{mean} is the number of the frames in V_{mean}, and $MFCC\{frame_{mean}(k)\}$ is the MFCC feature of the kth frame in V_{mean}. Other parameters can be defined in the same way. The size of F_{mean} and F_{var} is equal to the size of frame.

The final feature, F of an action is derived by concatenating F_{mean} and F_{var} together.

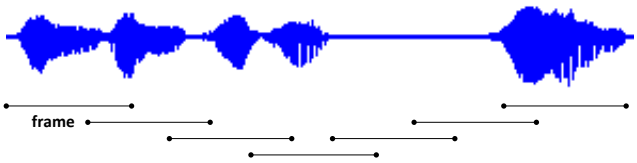

Figure 7. Frames of an audio signal.

3.2. Action classification

LIBSVM is largely used in classification. Chih-Chung Chang and Chih-Jen Lin[5] introduce a multi-class classification method based on "one-against-one" approach. Every two classes construct a classifier. Then $[\frac{n}{2}] = n(n-1)/2$ classifiers are constructed. The voting strategy for test data points is that each binary classification is considered to be a voting and a point of test data is classified as the class with the maximum number of votes. If two or more classes have the same number of votes, the first class is chosen.

LIBSVM (vision 1.51)[5] is applied directly in our work to classify various action classes. The strategy of dividing train data and test data is leave-one-out cross validation method, which is commonly used in classification on little sample size. Human action datasets are commonly constructed by collecting action videos from some persons, and everyone performs fixed classes of actions. On the basis of these characteristics, we leave one person's all actions out as testing data, and the rest as training data. The classification results of the person are calculated using multi-class classification method. We can achieve the other persons' classification result in the same way. The recognition rate is derived by comparing the classification results to ground truth.

4. Experiments and discussions

We test our method on Weizmann[2] and UIUC[20] action datasets. Silhouettes of all videos have been extracted. Here we directly use the silhouette sequences in the datasets as the low-level representation.

The UIUC dataset contains two parts. The first one consists of 532 high resolution (1024×768) sequences of 14 activities performed by eight actors, and the other one has three badminton sequences downloaded from Youtube. The forward one is used here to evaluate our method. Actually, low-resolution sequences are enough for us, so the image of every frame is subsampled before used. Section 1 of UIUC action dataset, with 271 sequences, is used in our experiment to test our method.

Weizmann dataset contains 93 videos of nine actors, and every actor performs ten different actions including "bend", "jack", "jump", "pjump", "run", "side", "skip", "walk", "wave one hand" and "wave two hands". The resolution is 180×144.

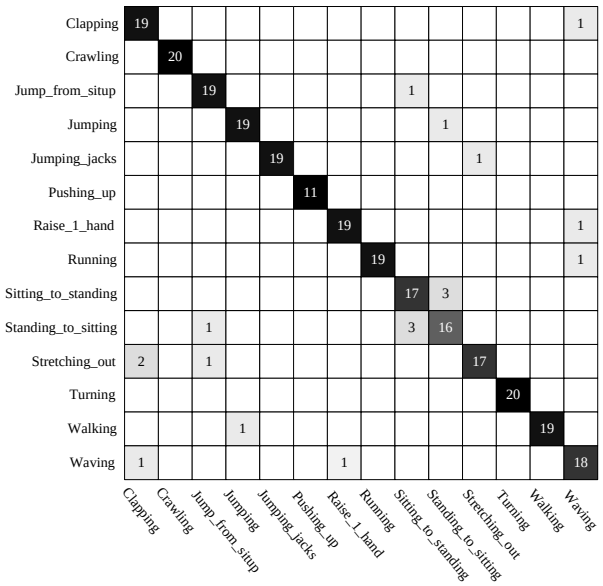

Figure 8. Confusion matrix for UIUC dataset setting the length of frame size to 60 and frame rate to $200Hz$. Overall accuracy is 92.99%.

The sample rate (S_r) is set as 16000Hz. The frame size of a 1-D signal decides the smallest unit of MFCC, and the frame rate decides the overlap between two adjacent frames. Recognition ratio is affected by these two parameters. We collect recognition ratios with different value combinations of the two parameters in UIUC dataset, and show the result in Table 1. We can get the information that it has good performance when the value of frame size is set from 50 to 70. The influence of frame rate is unnoticeable in the interval from 200 to 800. Based on an overall consideration of various factors, including calculate speed and accuracy of recognition, we set values on feature size and frame rate to 60 and 200.

The confusion matrix as shown in Figure 8 shows the recognition results in UIUC dataset. The length of feature is 60 and the value of frame rate is set 200. Horizontal rows are ground truth, and vertical columns are predictions. The overall accuracy is 92.99%. We can see from the result that "sitting to standing" and "standing to sitting" are not classified well. Figure 9 shows the X-T slice sequences of two actions. The most difference between the two images in Figure 9 is the shape of every unit.

We do not compare the efficiency of feature on UIUC action dataset, because UIUC action dataset is relatively new, and few work are tested on it.

By utilizing the algorithm in the paper, the experiments made on Weizmann action database have achieved good recognition ratio. The confusion matrix in Figure 10 shows the recognition result of every action. Parameters are the

rate \ size	20	30	40	50	60	70	80	90	100
100	72.69	79.70	81.92	86.72	87.08	86.72	88.19	88.93	88.93
200	81.18	84.50	85.98	88.93	**92.99**	92.62	91.88	91.88	90.04
300	85.61	87.08	90.41	92.25	92.62	91.88	92.25	91.88	90.04
400	81.55	87.82	87.08	90.04	92.25	92.25	91.88	91.88	90.04
500	85.98	88.19	91.88	91.51	92.62	**92.99**	92.25	91.51	90.77
600	90.41	87.82	90.04	91.88	92.25	92.62	92.25	91.51	90.41
700	88.56	87.82	90.77	91.88	92.25	91.88	91.88	91.14	90.04
800	87.82	89.67	89.30	91.88	92.25	92.25	91.88	92.62	90.04
900	90.41	88.93	91.14	92.25	91.88	91.88	92.25	91.51	90.04
1000	88.19	88.56	90.77	91.88	92.25	92.25	91.88	91.51	90.41

Table 1. The results of recognition ratios with different value combinations of frame size and frame rate. The numbers in bold type are the hightest recognition ratio. “rate” denotes the value of frame rate , and “size” denotes the value of frame size.

Figure 9. Comparison is made between the X-T slice sequences of two actions. (a) corresponds to ’sitting to standing’; (b) corresponds to ’standing to sitting’.

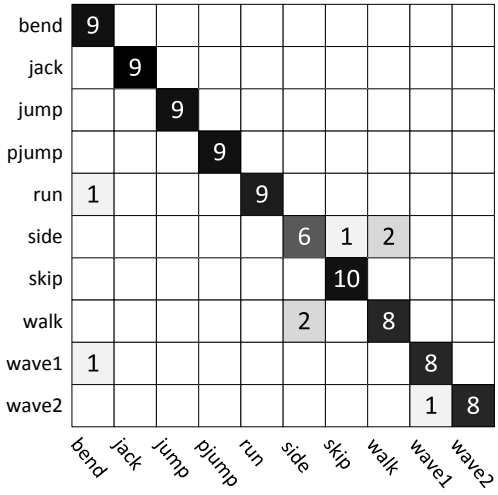

Figure 10. Confusion matrix for Weizmann dataset setting the length of frame size to 60 and frame rate to $200Hz$. Overall accuracy is 91.40%.

same to UIUC dataset’s.

Weizmann dataset is widely used in testing the recognition algorithm. The recognition accuracy of many methods can reach 100%. As our work is to test the efficiency of MFCC feature, we compare our method in terms of feature. Comparison of our feature with the ones mentioned in the work from Jingen Liu et al.[12] is showing in Table 2.

feature	recognition ratio
Original Bag of words[12]	84.2%
Weighted bag of words[12]	90.4%
ST features[12]	64.4%
Spin-Image features[12]	74.2%
ST + Spin-Image features[12]	89.3%
MFCC feature	91.4%

Table 2. The comparison of classification results on Weizmann dataset.

It can be seen from the result of Figure 8 and Figure 10 that our method can achieve the purpose of action recognition, but the recognition ratio can not outperform the state-of-the-art methods. We think that there are three possible reasons:

- the problem of image quality. Silhouettes of some videos are not extracted well;
- MFCC feature should be reformulated as the one fitting the characteristic of action to improve the recognition ratio. The filters used in MFCC fit audio signal better, so designing new filters may be a good way to improve recognition ratio;
- the problem of the direction in time mentioned before. Temporal direction information should be embedded more obviously in the feature.

5. Conclusions and future work

In this paper, we have presented a method for action classification from a new angle. X-T slice sequence is utilized

here to describe an action and MFCC is used to extract the feature. We treated an action in global level and found advantages over image based methods. Our method can efficiently make use of moving velocity and body structure.

Its accuracy is not as good as the best algorithms because we only adopt very simple features, i.e., mean and variance, which are not enough to describe the variations of X-T slice sequences. In future, we will focus on characterizing the X-T slice with more appropriate statistics to enhance the accuracy. Finally, it is worthy noting that the proposed system is very fast (50 frames per second). It is very possible to integrate it with other techniques and put it into real-time applications.

6. Acknowledgements

This work is supported by National Natural Science Foundation of China (Grant No.61135002,61175007).

References

[1] J. Aggarwal and M. Ryoo. Human activity analysis: A review. In *ACM Computing Surveys (to appear)*, 2011.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1395 –1402 Vol. 2, 2005.

[3] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257 –267, Mar. 2001.

[4] A. Bobick and A. Wilson. A state-based approach to the representation and recognition of gesture. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(12):1325 – 1337, Dec. 1997.

[5] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[6] N. Dalal and B. Triggers. Histograms of Oriented Gradients for Human Detection. *In Computer Vision and Pattern Recognition*, 2005.

[7] T. Darrell and A. Pentland. Space-time gestures. In *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR '93., 1993 IEEE Computer Society Conference on*, pages 335 –340, June 1993.

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65 – 72, 2005.

[9] D. Gavrila and L. Davis. 3-d model-based tracking of human upper body movement: a multi-view approach. In *Computer Vision, 1995. Proceedings., International Symposium on*, pages 253 –258, Nov. 1995.

[10] Y. Ke, R. Sukthankar, and M. Hebert. Spatio-temporal shape and flow correlation for action recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, 2007.

[11] I. Laptev and T. Lindeberg. Space-time interest points. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 432 –439 vol.1, 2003.

[12] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, 2008.

[13] R. Lublinerman, N. Ozay, D. Zarpalas, and O. Camps. Activity recognition from silhouettes using linear systems and model (in)validation techniques. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 347 –350, 2006.

[14] P. Natarajan and R. Nevatia. Coupled hidden semi markov models for activity recognition. In *Motion and Video Computing, 2007. WMVC '07. IEEE Workshop on*, page 10, 2007.

[15] J. Niebles, H. Wang, and L. Fei-fei. Unsupervised learning of human action categories using spatial-temporal words. *In British Machine Vision Conference*, 2006.

[16] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):831 –843, Aug. 2000.

[17] S. Park and J. Aggarwal. A hierarchical byayesian network for event recognition of human actions and interactions. *In Multimedia Systems*, 10(2):164–179, 2004.

[18] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the space of a human action. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 144 – 149 Vol. 1, 2005.

[19] H. Terasawa, M. Slaney, and J. Berger. Perceptual distance in timbre space. In *In International Conference on Auditory Display*, 2005.

[20] D. Tran and A. Sorokin. Human activity recognition with metric learning. *In Europeon Conference on Computer Vision*, 2005.

[21] A. Veeraraghavan, R. Chellappa, and A. Roy-Chowdhury. The function space of an activity. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006.

[22] Y. Yacoob and M. Black. Parameterized modeling and recognition of activities. In *Computer Vision, 1998. Sixth International Conference on*, pages 120 –127, Jan. 1998.

[23] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, pages 379 –385, June 1992.

[24] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 984 – 989 vol. 1, 2005.

[25] J. Zhang, K. Huang, Y. Yu, and T. Tan. Boosted local structured hog-lbp for object localization. In *Computer Vision and Pattern Recognition, 2011. CVPR 2011. IEEE Conference on*, pages 1 –8, 2011.