 © 2013 ACADEMY PUBLISHER

Human Action Recognition Using APJ3D and Random Forests

Ling Gan
Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts & Telecommunications, Chongqing, China
Email: ganlingcq@yeah.net

Fu Chen
Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts & Telecommunications, Chongqing, China

***Abstract*—Human action recognition is an important yet challenging task. In this paper, a simple and efficient method based on random forests is proposed for human action recognition. First, we extract the 3D skeletal joint locations from depth images. The APJ3D computed from the action depth image sequences by employing the 3D joint position features and the 3D joint angle features, and then clustered into K-means algorithm, which represent the typical postures of actions. By employing the improved Fourier Temporal Pyramid, we recognize actions using random forests. The proposed method is evaluated by using a public video dataset, namely UTKinect-Action dataset. This dataset is constituted of 200 3D sequences of 10 activities performed by 10 individuals in varied views. Experimental results show that the robustness of 3D skeletal joint location estimation display very well results, and the proposed method performs very well on that dataset. In addition, due to the design of our method and the robust 3D skeletal joint locations estimation from RGB-D sensor, our method demonstrates significant reliability against noise on 3D action dataset.**

***Index Terms*—APJ3D, Fourier Temporal Pyramid, random forests**

I. Introduction

Recent progresses on human action recognition have greatly improved multi-media technologies in a widely studied area, including video surveillance, sports video analysis, human computer interaction and video retrieval. Although human action recognition research work has received many encouraging developments, but still is quite a challenging task. Here we make human action recognition are the three major challenges. First is description of human action. Human action in the video sequence is a dynamic process that characterized not only with each frame of the body posture, but also with these the emergence of gesture sequences and continuous time. And even with a type of action, different individuals at the completion of the action of the process will be different due to the different height, shape, agility and so on. Therefore, on human action identification process, how to quickly extract simple but effective features is still facing a great difficulty in human action recognition. Second is representation model of human action, the relatively large changes in human action, but also has a strong combination of structural features, and how to combine these characteristics, design a strong distinction between the ability of the action of the model is an important issue in human action recognition. Third is efficient action classification algorithm design, action recognition has a high data dimension, training data acquisition difficulties characteristics, we hope that the behavioral categories algorithm has the training and classification speed, good effect, generalization ability characteristics.

In order to solve the above challenges, researchers have proposed many solutions by the efforts. In 1975, Johansson's experiment shows that humans can recognize action with highly compact observers [1]. Johansson demonstrated his statement taking a movie of a person walking in a dark room with lights connected to the person's major joints. Even though only light spots could be observed that there was a strong identification of the 3D motion in these movies. In recent studies, Fuijiyoshi and Lipton [2] proposed to use "star" skeleton extracted from silhouettes for motion analysis. Yu and Aggarwal [3] use extremities as semantic posture representation in their application for the detection of fence climbing. Zia et al. [4] propose an action recognition algorithm using body joint-angle features extracted from the RGB images from stereo cameras. Their dataset includes 8 simple actions (e.g., left hand up), and they were all taken from frontal views. Inspired by natural language processing and

Manuscript received January 12, 2013; revised March 17, 2013; accepted April 3, 2013.
National Natural Science Foundation of China under Grant （No.61075019）.

© 2013 ACADEMY PUBLISHER
doi:10.4304/jsw.8.9.2238-2245

information retrieval, bag-of-words approaches are also applied to recognize actions as a form of descriptive action unites. In these approaches, actions are represented as a collection of visual words, which is the codebook of spatio-temporal features. Schuldt et al. [5] integrate space-time interest point's representation with SVM [6] classification scheme. Dollar et al. [7] employ histogram of video cuboids for action representation. Wang et al. [8] represent the frames using the motion descriptor computed from optical flow vectors and represent actions as a bag of coded frames. However, all these features are computed from RGB images and are view dependent. Researchers also explored action recognition algorithms from depth images. However, due to price reasons, research has been limited.

The release of the low-cost RGB-D sensor Kinect has brought excitement to the research in computer vision, gaming, gesture-based control, and virtual reality. Shotton et al. [9] proposed a method to predict 3D positions of body joints from a single depth image from RGB-D sensor. Xia et al. [10] proposed a model based algorithm to detect humans using depth maps generated by RGB-D sensor. There are a few works on the recognition of human actions from depth data in the past two years.

In this paper, we employ a feature based representation of 3D human posture named APJ3D. In this representation, we extract the 3D joint position features and the 3D joint angle features based on the depth data and the estimated 3D skeletal joint locations. We propose the APJ3D feature that extended from two types of features. We manually select 15 informative joints to build a compact representation of human posture. The APJ3D feature can against minor posture variation. The collection of APJ3D vectors from training sequences are first extracted using the joint position estimation and then clustered into K-means [11] algorithm. By employing the improved Fourier Temporal Pyramid, we recognize actions using random forests. Our method only utilizes depth information. Experiments show that this method achieves superior results on a challenging dataset. Our method is shown in Fig. 1.

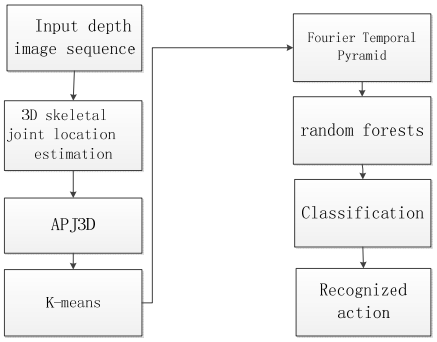

Figure 1. Overview of the method.

Our main contribution consists of two parts. First, we present a new method on human action recognition from depth imagery. Second, we propose a robustness representation of human postures and prove it is effective at action recognition, and the whole system runs at real-time.

The paper is organized as follows. Section 2 describes body part inference and joint position estimation from depth images. Section 3 describes our APJ3D as human posture representation. Section 4 addresses action recognition technique using random forests. Section 5 introduces dataset and discusses the experimental results. Section 6 concludes the paper.

II. Body Part Demarcate and Joint Location Estimation

The human body is an articulated system of rigid segments connected by joints and human action is considered as a continuous evolution of the spatial configuration of these segments [12]. Here, we use joint locations to build a compact representation of postures. The locations estimation of the objects or persons in the 3D scene is an important problem, and RGB-D sensor provides a cheap and real-time solution. Shotton et al. [9] propose to extract 3D body joint locations from a depth image using an object recognition proposal. The human body is marked as body parts based on the per-pixel classification results. The parts include LU/ RU/ LW/ RW head, neck, L/R shoulder, LU/ RU/ LW/ RW arm, L/ R elbow, L/ R wrist, L/ R hand, LU/ RU/ LW/ RW torso, LU/ RU/ LW/ RW leg, L/ R knee, L/ R ankle and L/ R foot (Left, Right, Upper, Lower). They calculate the confidence-scored 3D location estimation of body joints by using a local mode-finding channel based on mean shift with a weighted Gaussian kernel. Their enormous and multiple training set allows the classifier to estimate body parts invariant of posture, body shape, clothing, and so on. Employing their algorithm, we obtain the 3D locations of 20 skeletal joints which include hip center, spine, shoulder center, head, L/ R shoulder, L/ R elbow, L/ R wrist, L/ R hand, L/ R hip, L/ R knee, L/ R angle and L/ R foot. Note that part of the body segmentation results can not be used directly. Fig. 2 shows an example result of 3D skeletal joints and the corresponding depth map.

We apply these skeletal joint locations to express our representation of postures. Among these joints, hand and wrist and foot and ankle are very close to each other and thus superfluous for the characterization of body part constitution. Therefore, we calculate our features based expression of postures from 15 of the 20 joints, including head, neck, L/ R shoulder, L/ R elbow, L/ R hands, L/ R knee, L/ R feet, torso center and L/ R hip.

Note that the estimated joint locations from RGB-D sensor offer information concerning the orientation the person is facing RGB-D sensor, we are able to inform the left limb joints from those of the right limbs.

 © 2013 ACADEMY PUBLISHER

(a) (b)

Figure 2. (a) Skeletal joints locations (b) Depth image.

III. Angles and Positions of 3D Joint as Posture Representation

As discussed, this section gives a detailed description of two types of features: the 3D joint position features and the 3D joint angle features, and we propose the APJ3D feature that extended from two types of features. We utilize to use the APJ3D feature to represent the actions. These features can characterize the human motions as well as the interactions between the objects and the human. These features are invariant to the translation of the human and robust to noise.

A. 3D Joint Angles

As the above discussed, the following features use the above skeletal joint locations. For each frame t, the skeleton rate is a sequence of graphs with 15 joints, where each joint has its geometric location. For a human subject, 15 joint positions are tracked by the skeleton tracker and each joint i has 3 coordinates $\chi_i(t)=\{x_i(t),y_i(t),z_i(t)\}$ at a frame t. It is represented as a 3D point in a global Cartesian coordinate system. The joints contiguous to the torso are usually called first-degree joints, while joints contiguous to first-degree joints are classified as second-degree joints. First-degree joints include the elbows, the knees and the head, while second-degree joints are the extremities: the hands and feet.

Different body postures are virtually acquired by alternating first and second-degree joints. Note that each joint movement has 2 degrees of freedom: a zenith angle θ and an azimuth angle μ, while the distance between contiguous joints is always invariant (Fig. 3).

In the work of Raptis et al. [13], an oversimplified

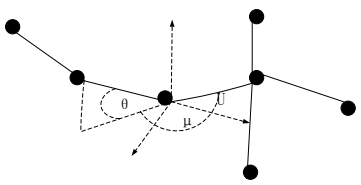

Figure 3. Joint-angles representation: the zenith angles θ and the azimuth angles μ.

joint-angles representation is developed by alternating each joint location $\chi_i \in R^3$ to local globular coordinates. First, a torso basis is estimated by applying a PCA [14] to a 7-3 torso matrix filled with the torso joint positions. Then, the globular coordinates of each first-degree joint are calculated as an interpretation of this torso basis to the joint.

However, this same torso basis is used as consultation to alternate the second-degree joints, bring about a non-local description of the angles. Also, as refered by the authors, some associations of joint positions can result in crumbled projections and consequently inconsistent angles, as in the open arms position [13].

B. 3D Joint Positions

It is inadequate to only use the 3D joint angles to completely model an action. Therefore, it is indispensable to employ a feature to describe the local "depth appearance" for the joints. The 3D joint positions [15] are used to establish the motion of the human body. Since joint positions is invariant features. The pairwise relative positions of the joints results in more discriminative features for representing the human movement is our key suggestion.

Due to the coordinates are normalized, so the motion is invariant to the absolute body position, the initial body orientation and the body size.

For each joint i, we extract the pairwise relative position features by taking the difference between the position of joint i and that of each other joint j:

$$p_{ij}=p_i-p_j \tag{1}$$

The 3D joint feature for joint i is defined as:

$$p_i=\{p_{ij}|\ i\neq j\} \tag{2}$$

Although enumerating all the joint pairs introduces some information that is irrelevant to our classification task, our approach is able to select the joints that are most relevant to our recognition task.

C. APJ3D as Posture Representation

Through the analysis above, we introduce our new feature, namely APJ3D. The APJ3D feature is described as follows.

First of all, we use the same torso basis for first-degree joints, and the representation of second-degree joints by considering rotations of the orthonormal torso basis $\{U,R,T\}$. Let v be the vector defined by the right arm and the right elbow and w the vector between the right elbow and the right hand. To define a local basis for the right hand, we rotate the torso basis $\{U,R,T\}$ by the angle $\beta=\arccos(v\cdot r)$ around the axis $b=v\times r$. Note that if $v\cdot r=1$, no rotation is applied. Also $v\cdot r\neq -1$ since the right arm can never rotate completely left due to body constraints. The rotated basis is translated to the right elbow and the spherical

© 2013 ACADEMY PUBLISHER

coordinates of the right hand are computed as

- θ the angle between v and w
- μ the angle between the rotated t and the projection of w in the plane whose normal is v

If v and w are collinear, we just set $\varphi = 0$, as the azimuth is not defined, and this will not be an issue to us. The other second-degree joints are similarly constructed using variants of the torso basis, such that collapsing issues are avoided by other body constraints.

Each joint position χ_i is represented using a pair of spherical angles (θ_i, μ_i) that specifies it in a locally defined spherical coordinate system. We also compute the angle η between the directional vector z from the RGB-D sensor and the inverted vector $-t$ from the torso basis, to detect torso inclinations. Thus, a body posture joint-angles representation is a posture descriptor vector $v = (\theta_1, \mu_2, \cdots, \theta_9, \mu_9, \eta) \in R^{19}$.

Afterward, we select the pairwise relative position features as

- m the relative position between the torso center and the hands
- n the relative position between the torso center and the feet

Thus, we use vector $p = (m, n)$ to act as the features for action. Representing the human motion as the relative joint positions results in more discriminative and intuitive features. This can be better characterized through the pairwise relative positions.

Finally, we extract two types of features from each frame t: the 3D joint position features $p[t]$, and the 3D joint angle features $v[t]$. We combine them through weighted sum method, and superimposed them to form a new feature. So we can represent the posture by the vector $S\{p[t], v[t]\}$.

IV. Action Recognition Using Random Forests

We recognize a variety of human actions by using random forests technique similar to the standard random forests structure [16]. In this section, we introduce how to recognize human action by using random forests.

Although the applied feature is robust to noise, but we using the current random forests to recognize human action, we always experience temporal misalignment so that the recognition results is very sensitive to temporal interval. So we review Fourier Temporal Pyramid [15], and we propose to use the improved Fourier Temporal Pyramid to represent the temporal dynamics of these frame-level features, and to solve the problem of temporal interval.

A. Key Posture Learning

We apply an image sequence or video to represent each action, and using the vector representation of postures to represent an image sequence or video. As the vector representation of postures is in a continuous space, the key process is to alternate each frame into an observation symbol so that each action may be represented by an observation sequence.

In order to reduce the number of observation symbols, we extract APJ3D features from each frame t: the vector $S\{p[t], v[t]\}$. We cluster the feature vectors by using K-means. Then each action is represented as a sequence of the key postures. In this way, each action is a time series of the key postures.

B. Fourier Temporal Pyramid

The Fourier temporal pyramid technique similar to the spatial pyramid method [17]. In order to capture the temporal structure of the action, apart from the global Fourier coefficients, we recursively partition the action into a pyramid, and use the short time Fourier transform for all the segments, as showed in Fig. 4 .The final feature is the concatenation of the Fourier coefficients from all the segments.

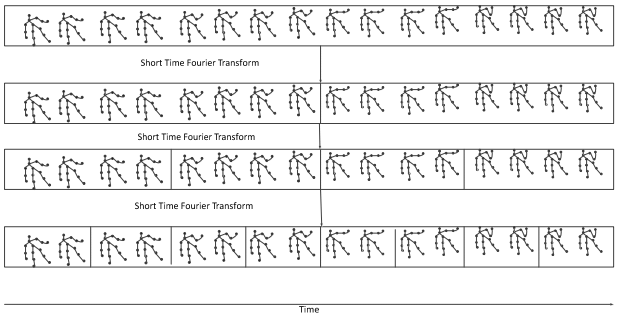

Figure 4. fourier temporal pyramid.

We improve Fourier Temporal Pyramid. The improved Fourier Temporal Pyramid is described as follows. For each key posture s, let $g = (p, v)$ denote its overall feature vector where p is its 3D pairwise position vector and v is its 3D joint angle vector. Note that each element g is a function of time and we can write it as $g[s]$. For each time segment at each pyramid level, we use Short Fourier Transform [18] to element $g[s]$ and acquire its Fourier coefficients, and we utilize its high-frequency and low-frequency coefficients as features. The Fourier Temporal Pyramid feature at key postures is defined as the high-frequency and low-frequency coefficients at all levels of the pyramid, and is denoted as G.

The applied Fourier Temporal Pyramid way has several benefits. First, by obtaining the low-frequency Fourier coefficients, the applied way is robust to noise. Second, by obtaining the high-frequency Fourier coefficients, the applied way can reflect the action mutation. Finally, this way is insensitive to temporal interval, because time series with temporal translation have the same Fourier coefficient magnitude, and the temporal structure of the actions can be characterized by

the pyramid structure.

In this section, we use four-level Fourier Temporal Pyramid, with 1/4 length of each segment as the high-frequency coefficients and low-frequency coefficients.

C. Randomized Tree Training

The training process is constructed according to the standard random forests structure [16]. We extract features from the training sets are trained with the random forests classifier, and assembled by a set of randomized decision trees. In each decision tree, W segment features are randomly selected from the training sets and put at a root node, and mapped to a set of termination leaf nodes by the interior binary splitting joints. At each interior joint, f variables are randomly selected out of the F feature dimension and the decision threshold T is correspondingly chosen in the range $\left\{T|\min_q f(v_q) \le T \le \max_q f(v_q)\right\}$. The splitting function is defined as:

$$f_{l,r}(v_q) = \begin{Bmatrix} 1, if\left\{q \in I_n | f(v_q) > T\right\} \\ 0, otherwise \end{Bmatrix} \quad (3)$$

After training, supposing K leaf nodes are generated in a decision tree, each segment $w \in W$ must fall into a leaf node $k \in K$. As illustrated in Fig. 5, the class label at a leaf node kP_c^k, refers to the proportion of segments within each action class that reaches this leaf node after training $\sum kP_c^k = 1$.

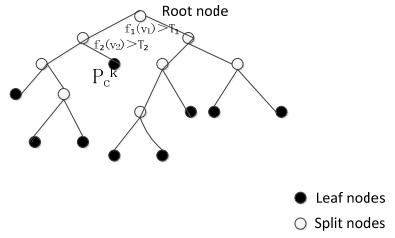

Figure 5. Decision tree growing.

To measure the training quality of each leaf node, the proportion of segments from sequences of a same action falling into the same leaf node, the information gain is defined at each split node:

$$\Delta E = -\frac{|I_l|}{|I_n|}E(I_l) - \frac{|I_r|}{|I_n|}E(I_r) \quad (4)$$

ΔE refers to the information gain, $E(\ )$ denotes entropy, I_l, I_r and I_n indicate the left splitting features, the right splitting features and the total input features at the splitting node respectively. This equation is given with respect to the splitting function at each splitting node. Since the f variables are randomly selected from the feature vector and the decision threshold T cannot be predefined to maximize the information gain, a set of combinations of 'f's and 't's are recursively tried to boost training quality.

D. Learning Forests

The training set is equally divided into a number of subsets, and then dispatched to different decision trees. Normally, in order to boost the general performance, the subsets are set to have overlaps with each other. Supposing the total training feature number is N and there are N^1 decision trees within the random forests classifier, and the features that are dispatched to each decision tree are more than the number N / N^1. In the testing stage, each segment feature is pushed to the root node of each decision tree in the random forests classifier, and eventually forwarded to a terminating leaf node. The path between a root node and a terminating leaf node consists of a set of split nodes, and each split node contains a binary splitting function. When the segment feature drops into a terminating leaf node, a histogram P_n refers to the proportion of segments per class label that fall into this leaf node during training stage, which is the soft voting result at the decision tree $n \in N^1$. Finally, the prediction histogram of the whole forests is acquired by summing up the voting histograms from all the decision trees:

$$P_f = \sum_{n=1}^{N^1} P_n \quad (5)$$

V. Experiments Results

In this section, we first give the experimental environment in Section A. Second, we give the description of experimental data sets in Section B. Last, we give the experimental results and analysis in Section C.

A. Experimental Environment

The experiment has been implemented on a personal computer with a 2. 3-GHz Intel i5 processor CPU using Microsoft VS2010. The experiment selected some representative standard test environments to test the robustness of joint location estimation.

B. Data Sets

We choose UTKinect-Action dataset [19] to evaluate the proposed action recognition method. In the dataset, there are about containing 10 types of human actions in indoor settings. The RGB images and depth maps were captured at 30 frames per second. The 10 actions include: walk, sit down, stand up, pick up, carry, throw, push, pull, wave and clap hands. As shown in table.1.

© 2013 ACADEMY PUBLISHER

TABLE I.
THE NUMBENR OF ACTION TYPE

Number	Action	Number	Action
1	Walk	6	Throw
2	Sit down	7	Push
3	Stand up	8	Pull
4	Pick up	9	Wave
5	Carry	10	Chap hands

Each action was collected from 10 different persons for 2 times: 9 males and 1 female. One of the persons is left handed. Sample RGB images from the dataset are shown in Fig. 6.

Figure 6. Sample images from videos of the 10 activities in the dataset.

C. Experimental Results and Analysis

To verify the robustness of skeletal joint location estimation, we test and compare the skeletal joint location estimation algorithm in different environments.

Figure 7. In a fluorescent light environment.

Figure 8. In the dark environment.

Fig. 7, Fig. 8 and Fig. 9given a few representative frames in the joint location estimation results. As is shown in Fig. 7 and Fig. 8 when human in a fluorescent

Figure 9. In the complex background environment.

light and the dark environment the skeletal joint location estimation algorithm can recognize the human target timely. And Fig. 9 shows that when the complex background environment algorithm can eliminate the target joint location in time. The experimental results indicated that the real-time and the robustness are high for the skeletal joint location estimation algorithm in different environments.

In our experiments, we using UTKinect-Action dataset because those actions were chosen to cover various movement of arms, legs, torso and their combinations, and the subjects were advised to use their right arm or leg if an action is performed by a single arm or leg. Although the background of this dataset is clean, this dataset is challenging because many of the actions in the dataset are highly similar to each other.

We extract the 3D skeletal joint locations from the depth sequence by using the real time joint location estimation. Since there is no human-object interaction in this dataset, we only extract the 3D joint location features.

In this dataset, we take the actions of five people for training and use those of the remaining one for testing. For each action sequence of a person in this dataset, we extract APJ3D features from each frame. In order to reduce the training time, the [20] give an experiment on KTH data sets. The experimental results show that snippets of 5-7 frames (0.3-0.5 seconds of video) are enough to achieve a performance similar to the one obtainable with the entire video sequence. As shown in Fig. 10. So we take K-means algorithm, and made action sequence of a person clustering key postures. Then obtaining the high-frequency and low-frequency coefficients by the improved Fourier Temporal Pyramid. We train and test by random forests. Training set or testing set is composed by those actions.

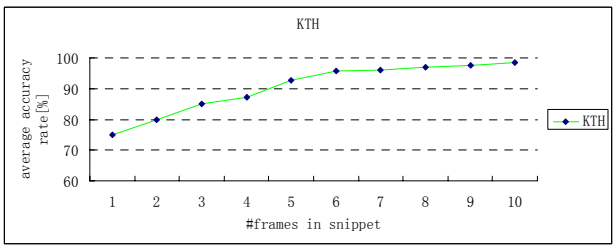

Figure 10. Recognition accuracy rate in the case of different number of frames on KTH date sets.

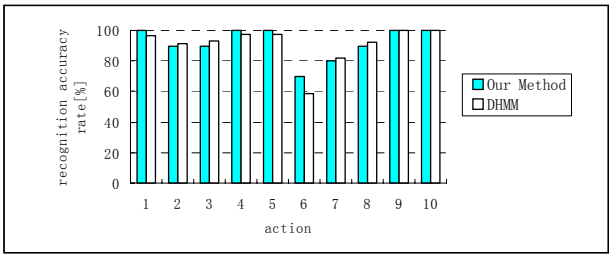

Figure 11. Recognition performance of our method and DHMM.

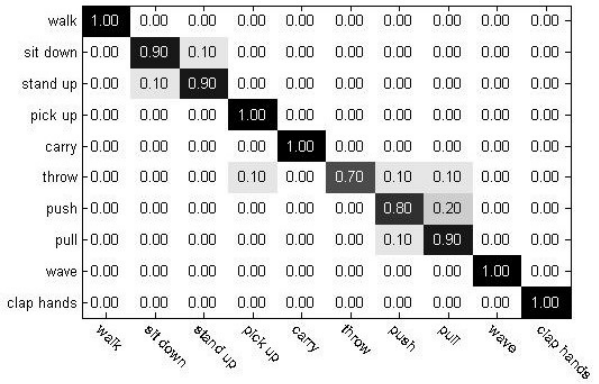

Figure 12. Recognition performance of our method measured using confusion matrices.

TABLE II.
RECOGNITION ACCURACY COMPARISON FOR UTKINECT-ACTION DATASET

Method	overall accuracy
Discrete Hidden Markov Model [19]	90.92%
Proposed Method	92%

We can observe Fig. 11 and the confusion matrix of Fig. 12. Although some actions is performed not well, because of the interference of some external factors, or the classification errors may occur if two actions are too similar to each other, such as "sit down" and "stand up", or if the occlusion is so large that the skeleton tracker fails frequently, such as the action "push" and "throw". But for most of the actions, our method works very well.

Table. Ⅱ shows the accuracies of different methods. The accuracy of Discrete Hidden Markov Model is 90.92%, this method taking histograms of 3D joint locations as the feature, and using Discrete Hidden Markov Model to train and recognition. By employing the APJ3D and the improved Fourier Temporal Pyramid, our method can obtain a recognition accuracy of 92%. This is a relatively good result considering the difficulties in this dataset. From the table, we can see that: a higher recognition rate can be achieved with our method.

In order to test our method against noise stability, we select action "walk" as subjects. Our method and that of Discrete Hidden Markov Model is shown in Fig. 13. In this experiment, we add white Gaussian noise to the 3D joint locations of action "walk", and compare the relative accuracies of the two methods. For each method, its relative accuracy is defined as the accuracy under the noisy environment divided by the accuracy under the environment without noise. We can see that our method is much more robust to noise than the Discrete Hidden Markov Model.

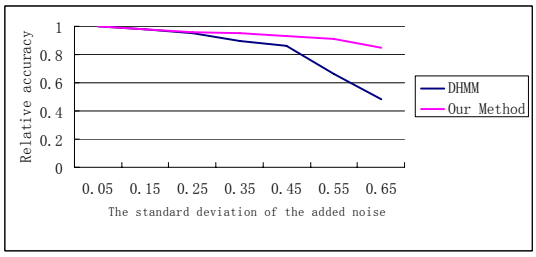

Figure 13. The relative accuracy between our method and DHMM.

In this experiment, the robustness of the proposed method and only random forests to temporal shift is also compared. We circularly shift all the training data, and keep the test data unchanged. The relative accuracy is shown in Fig. 14. It can be seen that both methods are robust to the temporal shift of the depth sequences, but only random forests is more sensitive to temporal shift than the proposed method.

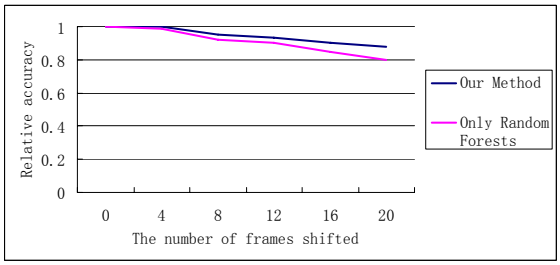

Figure 14. The relative accuracy between our method and only random forests.

VI. CONCLUSION

We have proposed novel features and used random forests model for human action recognition with depth images. The proposed features are discriminative enough to classify human actions with subtle differences as well as human object interactions and robust to noise. In order to solve the sensitive issue of temporal misalignment, we use random forests with the improved Fourier Temporal Pyramid. The improved Fourier Temporal Pyramid is capable of better capturing the intra-class variations and is more robust to the noises and errors in the depth

images and the skeletal joint locations. The experiments demonstrated the superior performance of the proposed method to the state of the art methods. In the future, we aim to exploit the effectiveness of the proposed technique for the understanding of more complex activities.

ACKNOWLEDGMENT

This paper is supported by the National Natural Science Foundation of China under Grant No.61075019.

REFERENCES

[1] G. Johansson, "Visual Motion Perception", *Sci. Am*, vol.232,no.6, pp. 76–88, 1975.
[2] H. Fujiyoshi and A. Lipton, "Real-time Human Motion Analysis by Image Skeletoniation", *IEEE Workshop on Applications of Computer Vision*, pp. 15-21, 1998.
[3] E. Yu and J. K. Aggarwal, "Human Action Recognition with Extremities as Semantic Posture Representation", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp.1-8, 2009.
[4] M. Z. Uddin, N. D. Thang, J.T. Kim and T.S. Kim, "Human Activity Recognition Using Body Joint-Angle Features and Hidden Markov Model", *ETRI Journal*, vol.33, no.4, pp. 569-579, 2011.
[5] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach", *International Conference on Pattern Recognition*, pp.32–36, 2004.
[6] B. Yu, H. F. Li and C. Y. Fang, "Speech Emotion Recognition Base on Optimized Support Vector Machine", *Journal of Software*, vol.7,no.12, pp.2726-2733,2012.
[7] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition Via Sparse Spatio-temporal Features", *IEEE Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp.65-72, 2005.
[8] Y. Wang, P. Sabzmeydani, and G. Mori, "Semi-latent Dirichlet Allocation: A Hierarchical Model For Human Action Recognition", *Human Motion Workshop in International Conference on Computer Vision*, pp.240-254, 2007.
[9] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-Time Human Pose Recognition in Parts from a Single Depth Image", *Communications of the ACM*, vol. 56, No.1, pp.116-124, 2013.
[10] L. Xia, C. C. Chen, and J. K. Aggarwal, "Human Detection Using Depth Information by Kinect", *Computer Vision and Pattern Recognition*, pp.15-22, 2011.
[11] J. Zhang, G.Q. Wu, X.G. Hu, S.Y.Li and S.L.Hao, "A Parallel Clustering Algorithm with MPI-MKmeans", *Journal of Computers*, vol.8,no.1, pp.10-17,2013.
[12] V. M. Zatsiorsky, "Kinematics of Human Motion", *Human Kinetics Publishers*, 1997.
[13] M. Raptis, D. Kirovski, and H. Hoppe, "Real-time Classification of Dance Gestures from Skeleton Animation", *SCA*, pp.147–156, 2011.
[14] S. Rahman, S.Motahar, A. A.Farooq and M.M. Islam, "Performance of PCA Based Semi-supervised Learning in Face Recognition Using MPEG-7 Edge Histogram Descriptor", *Journal of Multimedia*, vol.6,no.5, pp.404-415, 2011.
[15] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining Actionlet Ensemble for Action Recognition with Depth Cameras", *IEEE Computer Vision and Pattern Recognition*, pp. 2929–2936, 2012.
[16] X.D. Zhang, Q.S.Tang, H.Jin, Y.Qiu and Y.Guo, "Eye Location Based on Adaboost and Random Forests", *Journal of Software*, vol.7, no.10, pp.2357-2364 2012.
[17] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", *IEEE Computer Vision and Pattern Recognition* , pp. 2169-2178, 2006.
[18] A. V. Oppenheim, R. W. Schafer, and J. R. Buck,"Discrete Time Signal Processing", *Prentice Hall*, 1999.
[19] L. Xia, C. C. Chen, and J. K. Aggarwal, "View Invariant Human Action Recognition Using Histograms of 3D Joints", *The 2nd International Workshop on Computer Vision and Pattern Recognition*, pp.20-27, 2012.
[20] K. Schindler and L. van Gool, "Action Snippets: How Many Frames Does Human Action Recognition Require?", IEEE *Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.

Ling Gan, born in 1964. She is currently a professor of chongqing key laboratory of computational intelligence at chongqing university of posts & telecommunications. She received her B.S. degree and M.S. degree from southwest university. Her main research interests are digital image processing.

Fu Chen, born in 1987. He is currently a graduate student of chongqing key laboratory of computational intelligence at chongqing university of posts & telecommunications. He received his B.S. degree in electronic and information engineering from physical and electronic engineering department, Hunan Institute of Science and Technology, china. His main research interests are human action recognition.