View Invariant Human Action Recognition Using Histograms of 3D Joints

Lu Xia, Chia-Chih Chen, and J. K. Aggarwal
Computer & Vision Research Center / Department of ECE
The University of Texas at Austin
{xialu|ccchen}@utexas.edu, aggarwaljk@mail.utexas.edu

Abstract

In this paper, we present a novel approach for human action recognition with histograms of 3D joint locations (HOJ3D) as a compact representation of postures. We extract the 3D skeletal joint locations from Kinect depth maps using Shotton et al.'s method [6]. The HOJ3D computed from the action depth sequences are reprojected using LDA and then clustered into k posture visual words, which represent the prototypical poses of actions. The temporal evolutions of those visual words are modeled by discrete hidden Markov models (HMMs). In addition, due to the design of our spherical coordinate system and the robust 3D skeleton estimation from Kinect, our method demonstrates significant view invariance on our 3D action dataset. Our dataset is composed of 200 3D sequences of 10 indoor activities performed by 10 individuals in varied views. Our method is real-time and achieves superior results on the challenging 3D action dataset. We also tested our algorithm on the MSR Action3D dataset and our algorithm outperforms Li et al. [25] on most of the cases.

1. Introduction

Human action recognition is a widely studied area in computer vision. Its applications include surveillance systems, video analysis, robotics and a variety of systems that involve interactions between persons and electronic devices such as human-computer interfaces. Its development began in the early 1980s. To date research has mainly focused on learning and recognizing actions from video sequences taken by a single visible light camera. There is extensive literature in action recognition in a number of fields, including computer vision, machine learning, pattern recognition, signal processing, etc. [1, 2]. Among the different types of features for representation, silhouettes and spatio-temporal interest points are most commonly used [3]. Methods proposed in the past for silhouette based action recognition can be divided into two major categories. One is to extract action descriptors from the sequences of silhouettes. Conventional classifiers are frequently used for recognition [13, 14, 15, 16]. The other one is to extract features from each silhouette and model the dynamics of the action explicitly [15, 17, 18, 19, 20].

Here we enumerate three major challenges to vision based human action recognition. First is intra-class variability and inter-class similarity of actions. Individuals can perform an action in different directions with different characteristics of body part movements, and two actions may be only distinguished by very subtle spatio-temporal details. Second, the number of describable action categories is huge; the same action may have different interpretations under different object and scene contexts. Third, occlusions, cluttered background, cast shadows, varying illumination conditions and viewpoint changes can all alter the way actions are perceived.

Particularly, the use of range cameras significantly alleviates the challenges presented in the third category, which are the common low-level difficulties that reduce the recognition performance from 2D imagery. Furthermore, a range camera provides the discerning information of actions with depth changes in certain viewpoints. For example, in a frontal view, it would be much more accurate to distinguish person *pointing* from *reaching* from depth map sequences than in RGB footage. However, earlier range sensors were either too expensive, provided poor estimation, or were difficult to use on human subjects. For example, sonar sensors have poor angular resolution and are susceptible to false echoes and reflections. Infrared and laser range finders can only provide measurements from one point in the scene. LIDAR and radar systems are considerably more expensive and typically have higher power consumption requirements. For the use of low-cost digital cameras, distance has to be inferred either from stereoscopic cameras, or from the motion of objects within the image, e.g. optical flow.

The recent release of the Microsoft Kinect addresses these issues by providing both an RGB image and depth image streams [4]. Although targeted primarily for the entertainment market, the Kinect has excited considerable interest within the vision and robotics community for its broad applications [5]. Shotton et al. [6] proposed a method to quickly and accurately estimate 3D positions of

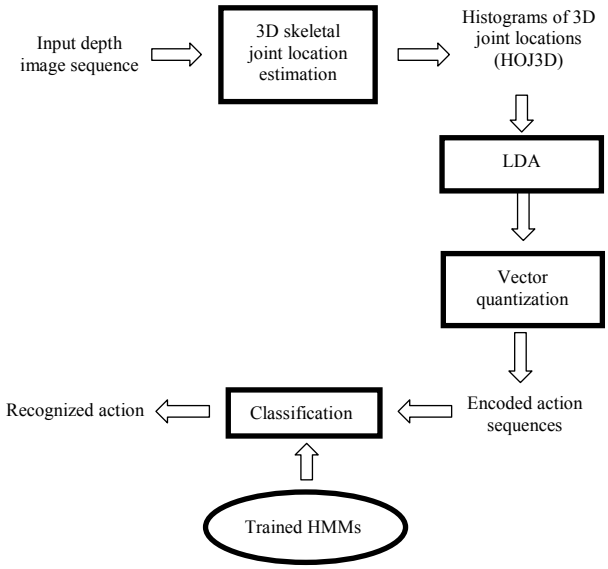


Figure 1: Overview of the method.

skeletal joints from a single depth image from Kinect. They provide accurate estimation of 3D skeletal joint locations at 200 frames per second on the Xbox 360 GPU. This skeletal joint information brings benefits to human centric computer vision tasks. With this information, we can address the problem of human action recognition in a simpler way compare to the use of RGB imagery. Most importantly, it achieves better view invariance and relatively faster speed.

In this paper, we employ a histogram based representation of 3D human posture named HOJ3D. In this representation, 3D space is partitioned into *n* bins using a modified spherical coordinate system. We manually select 12 informative joints to build a compact representation of human posture. To make our representation robust against minor posture variation, votes of 3D skeletal joints are cast into neighboring bins using a Gaussian weight function. The collection of HOJ3D vectors from training sequences are first reprojected using LDA and then clustered into k posture vocabularies. By encoding sequences of depth maps into sequential vocabularies, we recognize actions using HMM classifiers [11]. Our algorithm utilizes depth information only. Experiments show that this algorithm achieves superior results on our challenging dataset and also outperforms the algorithm of Li et al. [25] on most of the cases.

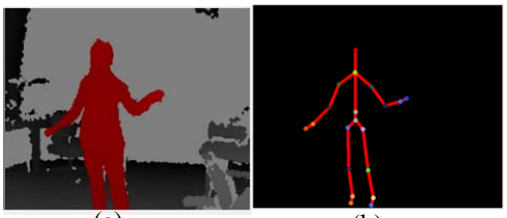
Our main contribution consists of three parts. First, we present a new algorithm on human action recognition from depth imagery. Second, we propose a view-invariant representation of human poses and prove it is effective at action recognition, and the whole system runs at real-time. Moreover, we collected a large 3D dataset of persons performing different kinds of indoor activities with a variety of viewpoints.

The paper is organized as follows. Section 2 presents the related work. Section 3 describes human part inference and joint position estimation from depth images. Section 4 describes our HOJ3D as human pose representation. Section 5 describes feature extraction from the HOJ3D. Section 6 addresses action recognition technique using discrete HMM. Section 7 introduces our dataset and discusses the experimental results. Section 8 concludes the paper.

2. Related Work

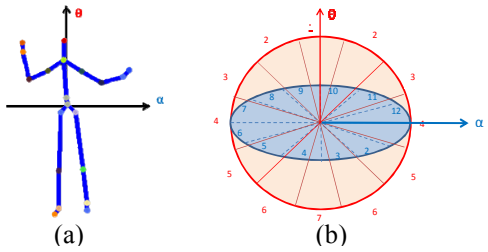
Researchers have explored different compact representations of human actions in the past few decades. In 1975, Johansson's experiment shows that humans can recognize activity with extremely compact observers [8]. Johansson demonstrated his statement using a movie of a person walking in a dark room with lights attached to the person's major joints. Even though only light spots could be observed, there was a strong identification of the 3D motion in these movies. In recent studies, Fuijiyoshi and Lipton [9] proposed to use "star" skeleton extracted from silhouettes for motion analysis. Yu and Aggarwal [10] use extremities as semantic posture representation in their application for the detection of fence climbing. Zia et al. [12] present an action recognition algorithm using body joint-angle features extracted from the RGB images from stereo cameras. Their dataset contains 8 simple actions (e.g., left hand up), and they were all taken from frontal views.

Inspired by natural language processing and 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. [21] integrate space-time interest point's representation with SVM classification scheme. Dollar et al. [22] employ histogram of video cuboids for action representation. Wang et al. [23] 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 free viewpoint action recognition algorithms from RGB images. Due to the large variations in motion induced by camera perspective, it is extremely challenging to generalize them to other views even for very simple actions. One way to address the problem is to store templates from several canonical views and interpolate across the stored views [29, 30]. Scalability is a hard problem for this approach. Another way is to map an example from an arbitrary view to a



(a) (b)

Figure 2: (a) Depth image. (b). Skeletal joints locations by Shotton et al.'s method



(a) (b)

Figure 3: (a) Reference coordinates of HOJ3D. (b) Modified spherical coordinate system for joint location binning.

stored model by applying homography. The model is usually captured using multiple cameras [31]. Weinland et al. [32] model action as a sequence of exemplars which are represented in 3D as visual hulls that have been computed using a system of 5 calibrated cameras. Parameswaran et al. [33] define a view-invariant representation of actions based on the theory of 2D and 3D invariants. They assume that there exists at least one key pose in the sequence in which 5 points are aligned on a plane in the 3D world coordinates. Weinland et al. [34] extend the notion of motion-history [35, 29] to 3D. They combine views from multiple cameras to build a 3D binary occupancy volume. Motion history is computed over these 3D volumes and view-invariant features are extracted by computing circular FFT of the volume.

The release of the low-cost RGBD sensor Kinect has brought excitement to the research in computer vision, gaming, gesture-based control, and virtual reality. Shotton et al. [6] proposed a method to predict 3D positions of body joints from a single depth image from Kinect. Xia et al. [24] proposed a model based algorithm to detect humans using depth maps generated by Kinect. There are a few works on the recognition of human actions from depth data in the past two years. Li et al. [25] employ an action graph to model the dynamics of the actions and sample a bag of 3D points from the depth map to characterize a set of salient postures that correspond to the nodes in the action graph. However, the sampling scheme is view dependent. Lalal et al. [27] utilize Radon transformation on depth silhouettes to recognize human home activities. The depth images were captured by a ZCAM [28]. This method is also view dependent. Sung et al. [26] extract features from the skeleton data provided by Prime Sense from RGBD data from Kinect and use a supervised learning approach to infer activities from RGB and depth images from Kinect. Considering they extract features from both types of imageries, the result is interesting but at the same time not as good as one would expect.

In this work, we present an action recognition algorithm using a HOJ3D representation of postures constructed from the skeletal joints' locations extracted from depth images. Taking advantage of Kinect devise and J. Shotton et al.'s algorithm [6], this method improves on the previous ones in that it achieves excellent recognition rates and is also view invariant and real time.

3. Body Part Inference and Joint Position 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 (i.e. body postures) [7]. Here, we use joint locations to build a compact representation of postures. The launch of Kinect offers a low-cost and real-time solution for the estimation of the 3D locations of objects or persons in the scene. Shotton et al. [6] propose to extract 3D body joint locations from a depth image using an object recognition scheme. The human body is labeled 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 compute the confidence-scored 3D position estimation of body joints by employing a local mode-finding approach based on mean shift with a weighted Gaussian kernel. Their gigantic and diverse training set allows the classifier to estimate body parts invariant of pose, body shape, clothing, and so on. Using their algorithm, we acquire the 3D locations of 20 skeletal joints which comprise 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 body part segmentation results are not directly available. Fig. 2 shows an example result of 3D skeletal joints and the corresponding depth map.

We use these skeletal joint locations to form our representation of postures. Among these joints, *hand and wrist* and *foot and ankle* are very close to each other and thus redundant for the description of body part configuration. In addition, spine, neck, and shoulder do not contribute discerning motion while a person is performing indoor activities. Therefore, we compute our histogram based representation of postures from 12 of the

20 joints, including head, L/ R elbow, L/ R hands, L/ R knee, L/ R feet, hip center and L/ R hip. We take the hip center as the center of the reference coordinate system, and define the x-direction according to L/ R hip. The rest 9 joints are used to compute the 3D spatial histogram.

Note that the estimated joint locations from Kinect provide information regarding the direction the person is facing, i.e., we are able to tell the left limb joints from those of the right limbs. This enables us to compute the reference direction of a person independent of the viewpoints. More specifically, our representation achieves view invariance by aligning the modified spherical coordinates with the person's specific reference direction, as detailed in the next section.

4. HOJ3D as Posture Representation

The estimation of 3D skeleton from RGB imagery is subject to error and significant computational cost. With the use of Kinect, we can acquire the 3D locations of the body parts in real-time with better accuracy. We propose a compact and viewpoint invariant representation of postures based on 3D skeletal joint locations. And we employ a vocabulary of postures to describe the prototypical poses of actions.

4.1. Spherical Coordinates of Histogram

Our methodology is designed to be view invariant, i.e., descriptors of the same type of pose are similar despite being captured from different viewpoints. We achieve this by aligning our spherical coordinates with the person's specific direction, as shown in Fig. 3(a). We define the center of the spherical coordinates as the hip center joint. Define the horizontal reference vector $\boldsymbol{\alpha}$ to be the vector from the left hip center to the right hip center projected on the horizontal plane (parallel to the ground), and the zenith reference vector $\boldsymbol{\theta}$ as the vector that is perpendicular to the ground plane and passes through the coordinate center.

We partition the 3D space into n bins as shown in Fig. 3(b) (in our experiment, we take n=84). The inclination angle is divided into 7 bins from the zenith vector $\boldsymbol{\theta}$: [0, 15], [15, 45], [45, 75], [105, 135], [165, 180]. Similarly, from the reference vector $\boldsymbol{\alpha}$, the azimuth angle is divided into 12 equal bins with 30 degrees resolution. The radial distance is not used in this representation to make the method scale-invariant. With our spherical coordinate, any 3D joint can be localized at a unique bin.

4.2. Probabilistic Voting

Our HOJ3D descriptor is computed by casting the rest 9 joints into the corresponding spatial histogram bins. For each joint location, weighted votes are contributed to the geometrically surrounding 3D bins. To make the representation robust against minor errors of joint

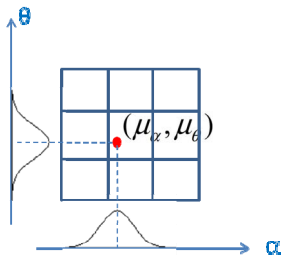


Figure 4: Voting using a Gaussian weight function.

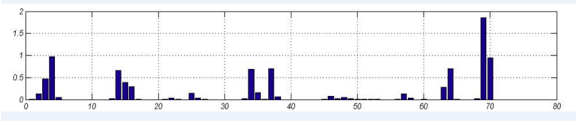


Figure 5: Example of the HOJ3D of a posture.

locations, we vote the 3D bins using a Gaussian weight function:

$$p(X,\mu,\Sigma)=\frac{1}{(2\pi)^{n/2}\,|\Sigma|^{1/2}}e^{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)} \quad (4.1)$$

, where $p(X,\mu,\Sigma)$ is Gaussian probability density function with mean vector μ and covariance matrix Σ (For simplicity we use identity matrix here). For each joint, we only vote over the bin it is in and the 8 neighboring bins. We calculate the probabilistic voting on $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ separately since they are independent (see Fig. 4). The probabilistic voting for each of the 9 bins is the product of the probability on α direction and θ direction. Let the joint location be (μ_α,μ_θ). The vote of a joint location to bin $[\theta_1,\theta_2]$ is

$$p(\theta_1<\theta<\theta_2;\mu_\theta,\sigma)=\Phi(\theta_2;\mu_\theta,\sigma)-\Phi(\theta_1;\mu_\theta,\sigma) \quad (4.2)$$

, where Φ is the CDF of Gaussian distribution. Similarly, the vote of joint location (μ_α,μ_θ) to the bin $[\alpha_1,\alpha_2]$ is

$$p(\alpha_1<\alpha<\alpha_2;\mu_\alpha,\sigma)=\Phi(\alpha_2;\mu_\alpha,\sigma)-\Phi(\alpha_1;\mu_\alpha,\sigma). \quad (4.3)$$

Then, the probability voting to bin $\alpha_1<\alpha<\alpha_2$, $\theta_1<\theta<\theta_2$ is:

$$p(\theta_1<\theta<\theta_2,\alpha_1<\alpha<\alpha_2;\mu,\Sigma)= p(\theta_1<\theta<\theta_2,\mu_\theta,\sigma)\cdot p(\alpha_1<\alpha<\alpha_2,\mu_\alpha,\sigma)$$

Figure 6: Sample images from videos of the 10 activities in our database. We show RGB image frames as well as the corresponding depth maps. Note only depth images are used in the algorithm. Action type from left to right, top to bottom: *walk, stand up, sit down, pick up, carry, throw, push, pull, wave hands, clap hands.*

(4.4)

The votes are accumulated over the 9 joints. As a result, a posture is represented by an *n*-bin histogram. Fig. 5 shows an instance of the computed histogram.

4.3. Feature Extraction

Linear discriminant analysis (LDA) is performed to extract the dominant features. LDA is based on the class specific information which maximizes the ratio of between-class scatter and the within-class scatter matrix. The LDA algorithm looks for the vectors in the underlying space to create the best discrimination between different classes. In this way, a more robust feature space can be obtained that separates the feature vectors of each class. In our experiment, we reduce the dimension of the HOJ3D feature from *n* dimensions to *nClass*-1 dimensions.

5. Vector Quantization

As each action is represented by an image sequence or video, the key procedure is to convert each frame into an observation symbol so that each action may be represented by an observation sequence. Note that the vector representation of postures is in a continuous space. In order to reduce the number of observation symbols, we perform vector quantization by clustering the feature vectors. We collect a large collection of indoor postures and calculate their HOJ3D vectors. We cluster the vectors into *K* clusters (a *K*-word vocabulary) using *K*-means. Then each posture is represented as a single number of the visual word. In this way, each action is a time series of the visual words.

6. Action Recognition Using Discrete HMM

We recognize a variety of human actions by the discrete HMM technique similar to what Rabiner did in speech recognition [11]. In discrete HMM, discrete time sequences are treated as the output of a Markov process whose states cannot be directly observed. In Section 4, we have encoded each action sequence as a vector of posture vocabularies, and we input this vector to learn the HMM model and use this model to predict for the unknown sequence.

A HMM that has N states $S=\{s_1,s_2,...,s_N\}$ and M output symbols $Y=\{y_1, y_2, ..., y_M\}$ is fully specified by the triplet $\lambda = \{A, B, \pi\}$. Let the state at time step t be S_t. The $N\times N$ state transition matrix A is,

$$A = \{a_{ji} \mid a_{ij} = P(s_{t+1} = q_j \mid s_t = q_i)\} \tag{6.1}$$

The N×M output probability matrix B is,

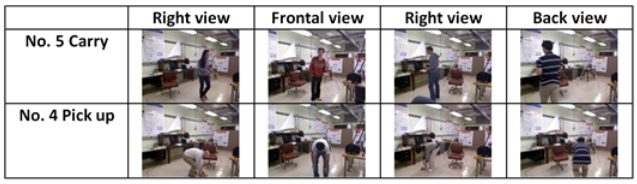


Figure 7. Different views of the actions.

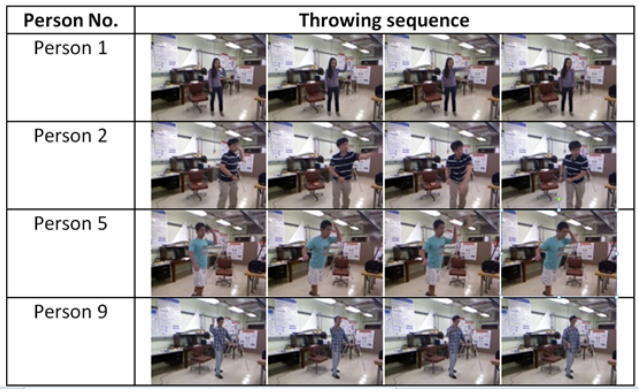


Figure 8: The variations of subjects performing the same action.

No.	1	2	3	4	5
Mean	43.60	34.15	25.60	35.50	58.15
Standard Deviation	8.89	9.40	6.44	11.89	27.04
No.	6	7	8	9	10
Mean	11.95	10.30	15.05	45.70	31.00
Standard Deviation	4.10	4.24	7.72	16.30	20.14

Table 1: The mean and standard deviation of the sequence lengths measured by number of frames at 30 fps.

$$B = \{b_i(k) \mid b_i(k) = P(v_k \mid s_t = q_i)\} \tag{6.2}$$

And the initial state distribution vector π is

$$\pi = \{\pi_i \mid \pi_i = P(s_1 = q_i)\} \tag{6.3}$$

We use a HMM to construct a model for each of the actions we want to recognize: the HMM gives a state based representation for each action. After forming the models for each activity, we take an action sequence $V=\{v_1, v_2, \ldots v_T\}$ and calculate its probability of a model λ for the observation sequence, $P(V|\lambda)$ for every model, which can be solved by using the forward algorithm. Then we classify the action as the one which has the largest posterior probability.

$$\text{decision} = \underset{i=1,2,\ldots,M}{\operatorname{argmax}} \{L_i\} \tag{6.4}$$

Action	ACC	Action	ACC
Walk	96.5%	Throw	59.0%
Sit down	91.5%	Push	81.5%
Stand up	93.5%	pull	92.5%
Pick up	97.5%	wave	100%
Carry	97.5%	Chap hands	100%
Overall: 90.92%			

Table 2: Recognition rate of each action type

Action Set 1 (AS1)	Action Set 2 (AS2)	Action Set 3 (AS3)
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup & throw	Side boxing	Pickup & throw

Table 3: The three subsets of actions used for the MSR Action3D dataset.

$$L_i = \Pr(O \mid H_i) \tag{6.5}$$

Where L_i indicates the likelihood of *i*-th HMM H_i and *M* number of activities. This model can compensate for the temporal variation of the actions caused by differences in the duration of performing the actions.

7. Experiments

We tested our algorithm on a challenging dataset we collected ourselves. In addition, we evaluated it on the public MSR Action3D dataset and compared our results with [25].

7.1. Data

To test the robustness of the algorithm, we collected a dataset containing 10 types of human actions in indoor settings. We take the sequence using a single stationary Kinect. Kinect hardware has a practical range of about 4 to 11 feet. The RGB images and depth maps were captured at 30 frames per second (FPS). The resolution of the depth map is 320×240 and resolution of the RGB image is 640×480. The 10 actions include: *walk, sit down, stand up, pick up, carry, throw, push, pull, wave and clap hands*. Each action was collected from 10 different persons for 2 times: 9 males and 1 female. One of the persons is left-handed. Altogether, the dataset contains 6220 frames of 200 action samples. The length of sample actions ranges from 5 to 120 frames. Sample RGB images from the dataset are shown in Fig. 6. Note that we only use the information from the depth image for action recognition in

	Test One		Test Two		Cross Subject Test	
	Li et al.	**Ours**	Li et al.	**Ours**	Li et al.	**Ours**
AS1	89.5%	**98.47%**	93.4%	**98.61%**	72.9%	**87.98%**
AS2	89.0%	**96.67%**	92.9%	**97.92%**	71.9%	**85.48%**
AS3	96.3%	**93.47%**	96.3%	**94.93%**	79.2%	**63.46%**
Overall	91.6%	**96.20%**	94.2%	**97.15%**	74.7%	**78.97%**

Table 4. Recognition results of our algorithm on the MSR Action3D dataset. We compared our result with Li et al. [25]. In test one, 1/3 of the samples were used as training samples and the rest as testing samples. In test two, 2/3 samples were used as training samples. In the cross subject test, half of the subjects were used as training and the rest of the subjects were used as testing.

our algorithm; the RGB sequences are just for illustration.

As shown in Fig. 7, we took action sequences from different views to highlight the advantages of our representation. In addition to the varied views, our dataset features 3 other challenges which are summarized as follows. First, there is significant variation among different realizations of the same action. For example, in our dataset, some actors pick up objects with one hand while others prefer to pick up the objects with both hands. Fig. 8 is another example, individuals can toss an object with either their right or left arm and producing different trajectories. Second, the durations of the action clips vary dramatically. Table 1 shows the mean and standard deviation of individual action length. In this table, the standard deviation of the carry sequence lengths is 27 frames, while the mean duration of *carry* is 48 frames longer than that of *push*. Third, object-person occlusions and body part out of field of view (FOV) also add to the difficulty of this dataset.

7.2. Experimental Results

We evaluate our algorithm on our 200 sequences dataset using leave one sequence out cross validation (LOOCV). As there is randomness in the initialization of the cluster centroids and the HMM algorithm, we run the experiment 20 times and report the mean performance, as shown in Table 2. We take the set of clusters to be K=125, and number of states N=6. By experiments, the overall mean accuracy is **90.92%,** the best accuracy is **95.0%** and the standard deviation is 1.74%. On a 2.93GHz Intel Core i7 CPU machine, the estimation of 3D skeletal joints and the calculation of HOJ3D is real-time using C implementation. The average testing time of one sequence is 12.5ms using Matlab.

We also test our algorithm on the public MSR Action3D database that contains 20 actions: *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing and pickup & throw.* We divide the actions into 3 subsets the same as in [25], each comprising 8 actions (see table 3). We use the same parameter settings as previously. Each test is repeated 20 times, and the average performance is shown in Table 4. We compare our performance with Li et al. [25]. We can see that our algorithm achieves considerably higher recognition rates than Li et al. [25] in all the testing setups on AS1 and AS2. On AS3, our recognition rate is slightly lower. As we have noticed in [25] that the goal of AS3 was intended to group complex actions together. However, Li et al.'s algorithm actually achieves much higher recognition accuracy on this complex dataset while ours have higher accuracy on the other two dataset. We conjecture the reason to be that the complex actions effects adversely the HMM classification when the number of training samples is small. Note that our algorithm performs better on MSR Action3D dataset than on our own dataset, partially because of the following reasons: 1) the subjects were facing the camera; 2) the whole body is in view all the times; 3) if the action is performed by a single arm or leg, the subjects were advised to use their right arm or leg.

8. Conclusions

This paper presents a methodology to recognize human action as time series of representative 3D poses. We take as input 3D skeletal joints locations inferred from depth maps as input. We proposed a compact representation of postures named HOJ3D that characterizes human postures as histograms of 3D joint locations within a modified spherical coordinate system. We build posture vocabularies by clustering HOJ3D vectors calculated from a large collection of postures. We train discrete HMMs to classify sequential postures into action types. The major components of our algorithm are real-time, which include the extraction of 3D skeletal joint locations, computation of HOJ3D, and classification. Experimental results show the salient advantage of our view invariant representation.

This work also suggests the advantage of using 3D data to recognize human actions and points out a promising direction of performing recognition tasks using depth information. Traditional RGB information can also be combined with the depth data to provide more data and produce algorithms with better recognition rates and robustness.

References

[1] P. Turaga, R. Chellapa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008.

[2] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. In *ACM Computing Surveys*, 2011.

[3] W. Li, Z. Zhang, and Z. Liu. Expandable data-driven graphical modeling of human actions based on salient postures. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1499–1510, 2008.

[4] The teardown, *Engineering Technology*, vol. 6, no.3, pp. 94-95, April 2011.

[5] J. Goles, Inside the race to hack the Kinect, *New Scientist*, vol. 208, no. 2789, p.22, December 2010.

[6] 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, in *CVPR*, *IEEE*, June 2011

[7] V. M. Zatsiorsky. Kinematics of Human Motion. *Human Kinetics Publishers*, 1997.

[8] G. Johansson: Visual motion perception. *Sci*. Am. 232(6), 76–88 (1975)

[9] H. Fujiyoshi and A. Lipton. Real-time human motion analysis by image skeletoniation. In *IEEE Workshp on Applications of Computer Vision*, pages 15-21, Princeton, 1998.

[10] E. Yu and J. K. Aggarwal, Human Action Recognition with Extremities as Semantic Posture Representation, *International Workshop on Semantic Learning and Applications in Multimedia (SLAM, in conjunction with CVPR)*, Miami, FL, June 2009.

[11] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE*, 77(2), February, 1989. pp. 257-285.

[12] 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, Aug. 2011, pp.569-579.

[13] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Trans, PAMI*, 23(3):257-267, 2001.

[14] H. Meng, N. Pears, and C. Bailey. A human action recognition system for embedded computer vision application. In *Proc*. *CVPR*, 2007.

[15] J. W. Davis and A. Tyagi. Minimal-latency human action recognition using reliable-inference. *Image and Vision Computing*, 24(5):455-473, 2006.

[16] D.–Y. Chen, H.–Y. M. Liao, and S.–W. Shih. Human action recognition using 2-D spatio-temporal templates. In *Proc ICME*, pages 667-670, 2007.

[17] V. Kellokumpu, M. Pietikainen, and J. Heikkila. Human activity recognition using sequences of postures. In *Proc IAPR Conf. Machine Vision Applications*, pages 570-573, 2005.

[18] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. In *Proc ICCV*, volum 2, pages 808-815, 2005.

[19] J. Zhang and S. Gong. Action categoration with modified hidden conditional random field. *Pattern Recognitoin*, 43: 197-203, 2010.

[20] W. Li, Z. Zhang, and Z. Liu. Expandable data-driven graphical modeling of human actions based on salient postures. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1499-1510, 2008.

[21] C. Sch¨uldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, pages III: 32–36, 2004.

[22] P. Doll´ar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior
recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

[23] Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *Human Motion Workshop, (with ICCV),* 2007.

[24] L. Xia, C.-C. Chen, and J. K. Aggarwal, "Human Detection Using Depth Information by Kinect", *International Workshop on Human Activity Understanding from 3D Data in conjunction with CVPR (HAU3D)*, Colorado Springs, CO, June 2011.

[25] W. Li, Z. Zhang, Z. Liu, "Action recognition based on a bag of 3D points", *CVPRW*, 2010

[26] J. Sung, C. Ponce, B. Selman and A. Saxena. Human Activity Detection from RGBD Images, In *AAAI workshop on Pattern, Activity and Intent Recognition (PAIR)*, 2011

[27] A. Jalal, M. Z. Uddin, J.T. Kim and T.S. Kim, Recognition of Human Home Activities via Depth Silhouettes and R Transformation for Smart Homes, *Indoor and Built Environment*, DOI: 10.1177/1420326X11423163, 1-7, 23 Sep 2011

[28] http://www.3dvsystems.com

[29] A. F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 257–267, 2001.

[30] T. J. Darrell, I. A. Essa, and A. P. Pentland, "Task-specific gesture analysis in real-time using interpolated views," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 18, no. 12, pp. 1236– 1242, 1996.

[31] T. F. Syeda-Mahmood, M. Vasilescu, and S. Sethi, "Recognizing action events from multiple viewpoints," *IEEE Workshop on Detection and Recognition of Events in Video*, pp. 64–72, 2001.

[32] D. Weinland, E. Boyer and R. Ronfard. Action recognition from arbitrary views using 3D examplars, *ICCV* 2007.

[33] V. Parameswaran and R. Chellappa, "View invariants for human action recognition," Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 613–619, 2003.

[34] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding,* vol. 104, no. 2, pp. 249–257, 2006.

[35] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 928–934, 1997.