VNU Journal of Science: Comp. Science & Com. Eng., Vol. 30, No. 3 (2014) 22-30

Human Action Recognition Using Dynamic Time Warping and Voting Algorithm[1]

Pham Chinh Huu*, Le Quoc Khanh, Le Thanh Ha

Faculty of Information Technology, VNU University of Engineering and Technology
144 Xuan Thuy, Cau Giay, Hanoi, Vietnam

Abstract

This paper presents a human action recognition method using dynamic time warping and voting algorithms on 3D human skeletal models. In this method human actions, which are the combinations of multiple body part movements, are described by feature matrices concerning both spatial and temporal domains. The feature matrices are created based on the spatial selection of relative angles between body parts in time series. Then, action recognition is done by applying a classifier which is the combination of dynamic time warping (DTW) and a voting algorithm to the feature matrices. Experimental results show that the performance of our action recognition method obtains high recognition accuracy at reliable computation speed and can be applied in real time human action recognition systems.

Manuscript communication: received 10 December 2013, revised 04 March 2014, accepted 26 March 2014
Corresponding author: Pham Chinh Huu, phamchinhhuu@gmail.com

Keywords: Human action recognition, feature extraction, Dynamic time warping.

1. Introduction

Human action recognition has become an interesting computer vision research topic for the last two decades. It is motivated by a wide range of potential applications related to video surveillance, human computer interaction aimed at identifying an individual through their actions. The evaluation of human behavior patterns in different environments has been a problem studied in social and cognitive sciences. However, it is raised as a challenging approach to computer science due to the complexity of data extraction and its analysis. The challenges originate from a number of reasons. Firstly, human body is non-rigid and it has many degrees of freedom. Human body can also generate infinite variations for every basic movement. Secondly, every single person has his own body shape, volume, and gesture style that challenges the recognition process. In addition, the uncertainties such as variation in viewpoint, illumination, shadow, self-occlusion, deformation, noise and clothing make this problem more complex. Over the last few decades, a large number of methods have been proposed to make the problem more tractable.

A common approach to recognize or model sequential data like human motion is the use of Hidden Markov Model (HMM) on both 2D observations [1] and 3D observations [2]. In HMM-based action recognition methods, we

[1] This work was supported by the basic research projects in natural science in 2012 of the National Foundation for Science & Technology Development (Nafosted), Vietnam (102.01-2012.36, Coding and communication of multiview video plus depth for 3D Television Systems).

must determine the number of states in advance for a motion. However, since the human motions can have different time length, it is difficult to set the optimal number of state corresponding to each motion. Recently, there have been increasing interests in using conditional random field (CRF) [3, 4] for learning of sequences. Although the advantage of CRF over HMM is its conditional nature resulting in relaxation of the independence assumption which is required by HMM to ensure tractable inferences, all these methods assume that we know the number of states for every motion. In [5], the author proposed a very intuitive and qualitatively interpretable skeletal motion feature representation called sequence of the most informative joints. This method resulted in high recognition rates in cross-database experiments but remains the limitation when discriminate different planar motions which are around the same joint. Another well-known method for human action classification is to use support vector machines (SVMs) [6, 7]. In [7], because temporal characteristics of human actions are applied indirectly by transforming to scalar values before inputted to SVMs, the loss of information reveals in time domain and deteriorates the recognition accuracy. Recently, an increasing number of researchers are interested in Dynamic Time Warping (DTW) [8] for human action recognition problems [9, 11] because DTW allows aligning two temporal action feature sequences varying in time to be taken into account. DTW in [9] was used with feature vectors constructed from 3D coordinate of human body joints. Even though this algorithm was enhanced from original DTW by improving distance function, this method faced the problem of body size variances, which caused noises for DTW to align two action series. Meanwhile the approach of [10] computed the joint orientation along time series that was invariant to body size to be the feature for DTW. Since the computation of this method required high complexity, it did not adapt to build a real time application. Reference [11] compared the difference between pairs of 2D frames to accomplish a self-similarity matrix for feature extraction. Recognition method included both DTW and K-nearest neighbor clustering. Although the recognition method achieved, as stated in the paper, robustness across camera views, the complexity of feature extraction is high due to comparison on the whole frame and it is difficult to reduce computation time to run in real time as the method in [10].

Recently, with the release of many low-cost and relatively accurate depth devices, such as the Microsoft Kinect and Asus Xtion, 3D human skeleton extraction have become much easier and gained much interest in skeleton-based action representation [2, 6, 9, 10]. A human skeletal model consists of two main components: body joints and body parts. Body joints connect body parts whose movements express human motions. Even though human performs same actions differently, while generating a variety of joint trajectories, the same set of joints with large amount of movements significantly contributes to that action in comparing with other joints. In this paper, we propose a human action recognition method based on the skeletal model, in which, instead of using joints, relative angles among body parts are used for feature extraction due to their invariance to body part sizes and rotation. Feature descriptor is formed from the relative angles describing the action per each frame sequence. Based on the movements contributing to the action of each angles, we reduce the size of feature descriptor for better representation of each action. For action recognition, we compare test action sequence with a list of defined actions using DTW algorithm. Finally, a voting algorithm is applied to figure out the best action label to the input test action sequence.

The remainder of this paper is organized as follows: the extraction and representation of action feature are introduced in Section 2; in

Section 3, we present the DTW and voting algorithm for action recognition; we demonstrate the datasets used in our experiments, the experiment conditions and the experimental results in Section 4. Finally, Section 5 concludes our proposed method in this paper.

2. Feature Extraction Based on Body Parts

2.1. Action Representation

The human body is an articulated system that can be represented by a hierarchy of joints. They are connected with bones to form the skeleton. Different joint configurations produce different skeletal poses and a time series of these poses yields the skeletal action. The action can be simply described as a collection in time series of 3D joint positions (i.e., 3D trajectories) in the skeleton hierarchy. The conventional skeletal model in Figure 1.a consists of 15 body joints that represent the ends of body bones. Since this skeletal model representation lacks of invariant properties for view point and scale, and 3D coordinate representation of the body joints cannot provide the correlations among body parts in both spatial and temporal domains, it is difficult to derive a sufficient time series recognition algorithm which is invariant to scale and rotation. Another representation for human action is based on relative angles among body parts due to their invariance to rotation and scale. In here, a body part, namely *left hand* shown in Figure 1.b, is a rigid connector between two body joints, namely *left hand* and *left elbow* shown in Figure 1.a.

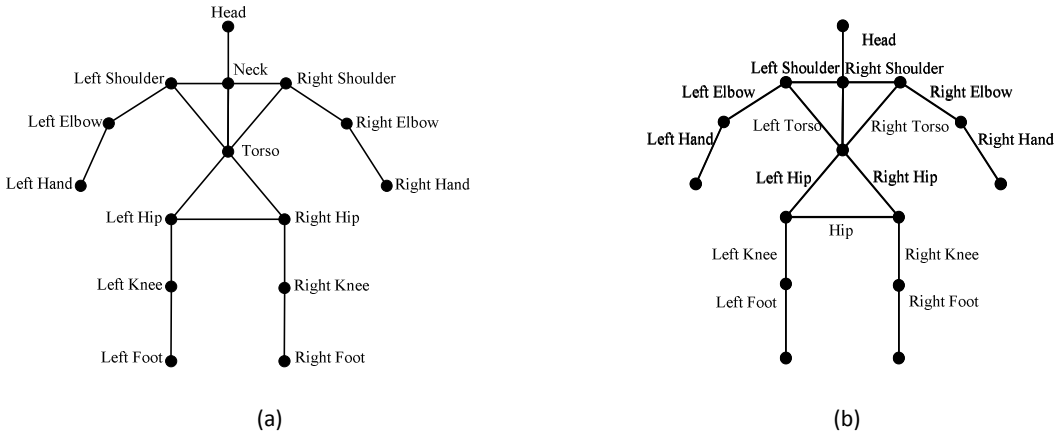

Figure 1: Human representation (a) Human skeletal model (b) 17 human body parts.

The relative angle between two body parts, a body part pair, J_1J_2 and J_3J_4 is computed by (1):

$$\theta = \arccos\left(\frac{\overrightarrow{J_1J_2} \times \overrightarrow{J_3J_4}}{\left|\overrightarrow{J_1J_2}\right| \times \left|\overrightarrow{J_3J_4}\right|}\right) \qquad (1)$$

Because there are 17 body parts defined in this model, the number of body part pairs is the 2-combination of the 17 body parts which is $C_{17}^2 = 136$. An action performing in a sequence of *N* frames can be represented by the temporal variation of relative angles between each pair of body parts. Let $\theta_{i,j}$ denote the relative angle of the j^{th} body part pair, $1 \le j \le 136$, at frame i^{th}, $1 \le i \le N$. For simplicity, this relative angle of a body part pair is called body part angle (BPA). Let $\theta_j = \{\theta_{1,j}, \theta_{1,j}, ..., \theta_{N,j}\}$ be the time ordered set of the j^{th} body part pair in the N-frame sequence. A complete representation of a human action in the frame sequence is denoted by a matrix $V=\left[\theta_{i,j}\right]_{N\times 136}$. Matrix V stores all BPAs in time sequence and is considered as a complete feature (CF) representation for a single action in terms of both spatial and temporal

information. Although a comprehensive action movement is included in matrix V, large number of elements exposes high computation time and complexity for learning and testing in recognition.

Our observations show that a specific human action may relate to a few number of body parts which have large movements during the action performance and the rest of body parts stay still or take part in another action. Two types of human actions are considered in this work in order to handle the task of action recognition. Actions which are performed by motions of some simple particular body parts, e.g. hand waving action includes hand motion, elbow motion while other body parts stay still, are called Single Motion Actions (SMAs). Other actions which are the combination of many body part motions, i.e. a person makes a signal by raising hand up while still walking, are called Multiple-Motion Actions (MMAs). Notice that an MMA may be the combination of multiple SMAs. For a specific action performance, some body parts which mainly contribute motions to form the meaning of the action, e.g. hands and elbow for hand clapping action, are called active body parts. It can be seen that in SMAs, only active body parts have large movement and others are staying still or becoming noise sources. In MMAs, beside active body parts, many other unexpected body parts also have large movements. Therefore, in order to recognize these actions accurately, only active body parts should be considered. It leads to the reduction number of relative angles for the representation of an action and results in the reduction of the dimension of CF. In this work, we propose two simple yet highly intuitive and interpretable methods which are based on the temporal relative angle variation and based on observation to efficiently reduce the dimension of CF.

2.2. Reduction of CF Based on Time Variance

We observed that a specific action requires human to engage a fixed number of body parts whose movements are at different intensity level and at different times. Therefore, in the first method of CF dimension reduction, we assume that the movements of active body parts are very large which results in the large variation of their corresponding BPAs in temporal domain. Here, the standard derivation can be used to measure the amount of movements for each BPA. For a given CF matrix $V=\left[\theta_{i,j}\right]_{N\times 136}$, a list of standard derivation values $(\delta_1, \delta_1, ..., \delta_{136})$ for all 136 BPAs can be constructed. Each value is calculated as following (2):

$$\delta_j = \sqrt{\frac{\sum_{i=1}^{N}\left(\theta_{i,j} - \overline{\theta}_j\right)}{N}} \quad (2)$$

$$\text{where } \overline{\theta}_j = \frac{\sum_{i=1}^{N}\theta_{i,j}}{N} \quad (3)$$

For a predefined action, only BPA j^{th} with large δ_j, called active BPA, should be involved in training and testing procedures and all others with lower motion activity should be discarded. To this end, a fixed number D of active BPAs is empirically defined for each action as shown in Table 1. Then, the size of CF matrix representing a training sample is reduced from *N×136* to *N×D*. The resulted feature matrix from this dimension reduction is called time variance reduction feature (TVRF) presentation. In testing procedure, testing samples are aligned with training samples by only using BPAs available in training samples.

2.3. Reduction of CF Based on Observations

In this method, instead of automatically creating a list of BPAs for each action based on their movement standard derivation, we definitely create the list by using our own observations to figure out which BPAs movements contribute most to a given action. It

results in the reduction of feature matrix and it is called observational reduction feature (OBRF) presentation. For example, the list of nine active BPAs for action *left hand low waving* in terms of body part pairs is defined as *{(head, left hand), (left shoulder, left hand), (right shoulder, left hand), (left elbow, left hand), (torso, left hand), (head, left elbow), (left shoulder, left elbow), (right shoulder, left elbow), (torso, left elbow)}*. For simplicity of explanation, we only show D, the number of BPAs, for each action in Table 2.

Table 1: Predefined number of BPAs for actions

Index	Action	D
1	Left hand low waving	6
2	Right hand low waving	6
3	Left hand high waving	6
4	Right hand high waving	6
5	Hand clapping	12
6	Greeting	6

Table 2: Predefined active joint angles

Index	Action	*D*
1	Left hand low waving	9
2	Right hand low waving	9
3	Left hand high waving	15
4	Right hand high waving	15
5	Hand clapping	16
6	Greeting	7

3. Action Recognition Using Dynamic Time Warping and Voting Algorithm

Since time-varying performance of human actions causes the feature presentation for each action to be different from sample to sample, many popular classifiers such as neuron networks, support vector machines which require a fixed size of input feature vectors are not capable for solving the action recognition problem. Therefore, in this research, we propose a classification model in which DTW is used to measure the similarity between two action samples and a voting algorithm for matching the testing action to a set of training action samples as shown in Figure 2.

In this model, the training set consists of a number of sample actions for each type in Table 2. The DTW algorithm is for computing the similarity between a testing action and a training action in a sample set. This results a set of similarity values that are used as the input for voting algorithm. Finally, voting algorithm produces the testing action label based on the input similarity values.

3.1. Dynamic Time Warping for Action Recognition

The original DTW was to match temporal distortions between two models and find an alignment warping path between two series. DTW algorithm was applied to find the warping path satisfying the conditions minimizing the warping cost. Here, the warping path reveals the similarity matching between two temporal input series. For the action recognition problem, each BPAs series of testing action should be aligned with those of training action using DTW to result the value of similarity between to actions.

Let $\boldsymbol{T}= \left[t_{i,d} \right]_{M \times D}$ and $\boldsymbol{S}= \left[s_{i,d} \right]_{M \times D}$ be the CF matrix of a testing action and training sample action respectively where *M* and *N* are the number of temporal sampled frames and *D* is the number of BPAs. In case that dimension reduction is not applied to training sample action, *D* equals to 136. Figure 3 shows the pseudo-code for algorithm calculating the similarity between a testing and training sample actions.

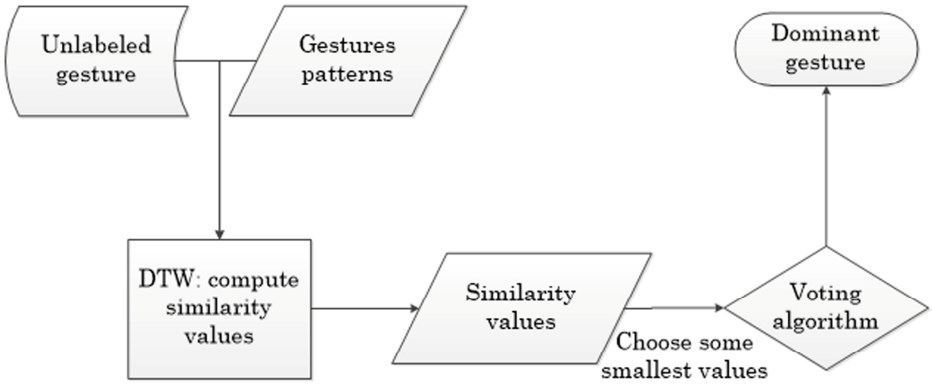

Figure 2: DTW classification model.

Input: $T = \left[t_{i,d} \right]_{M \times D}$ and $S = \left[s_{i,d} \right]_{M \times D}$
Output: similarity between two input action matrices
function matrixsimlilarity(T,S)
    let $\left[w_{i,j} \right]_{M \times N}$ be warping matrix aligning two feature representations
    set $w_{i,j}$ infinity for all *i* and *j*
    for i=1 to M do
        for j=1 to N do
            let T_i be the row vector of matrix T at row i^{th}
            let S_j be the row vector of matrix S at row j^{th}
            distance = EuclidDistance$(T_i, S_j) = \sqrt{\sum_{d=1}^{D} \left(t_{i,d} - s_{j,d} \right)^2}$
            $w_{i,j}$ = distance + min($w_{i-1,j-1}$, $w_{i-1,j}$, $w_{i,j-1}$)
        end
    end
    return $w_{M,N}$
end function

Figure 3: DTW algorithm for action similarity.

3.2. Voting Algorithm for Action Recognition

The distances between a testing action sample and all training action samples are obtained by using DTW. It is clear that the smaller the distance, the more similar the training and testing action samples are. In addition, the distances from testing sample to samples of the same action are somewhat similar while those to samples of different actions are much different. Therefore, in this part, a voting algorithm is proposed in order to find the action label for the current testing action sample.

For a given testing action sample, after calculating the distances from this testing sample to all training samples, these distances are then ascending sorted. Afterward, training

samples corresponding with p first sorted distance values are extracted. The action labels of the extracted training samples are counted and let q be the highest count number. The action label with the highest count number is assigned to the testing action sample if $q \geq p/2$ otherwise 'unknown' label is assigned to the testing action sample. Notice that the condition $q \geq p/2$ is used to get rid of the recognition ambiguity and the value of p should be carefully chosen based on the size of training samples and the number of training samples in each action label.

4. Experimental Results

In this section, we evaluate the recognition performance of our method in terms of both accuracy and complexity.

Two datasets were used for recognition accuracy evaluation in which we recorded one dataset and reference the other from [12]. Three computers with different hardware specifications were used to run the test and evaluate the computational complexity. Three types of action feature presentations including CF, TVRF, and OBRF are involved in the tests. DTW algorithm is applied to calculate 30 similarity values between the feature vectors of testing sample and those of training samples. The voting algorithm figures out which action type dominates the others and assigns action label to the testing action sample. An "unknown" label is assigned if there is no dominating action type as stated in previous section. Finally, comparison and discussion about the effectiveness of these feature presentations were made based on the experimental results.

4.1. Data Collection

4.1.1 Dataset #1

The first action dataset, dataset #1, is collected using OpenNI library [13] to generate skeleton structure from depth images captured from a depth sensor. Depth frames with resolution of 640x480 are recorded at 30 frames per seconds (FPS). It has been considered 13 different actors, 5 different backgrounds and about 325 sequences per action type for collecting the dataset. There are 6 different types of actions in this dataset as shown in Table 3. These action types describe common human actions using two hands. For each of the 6 actions, we collect three different subsets: sample set, single-motion action set (SMA set), multiple-motion action set (MMA set). Sample set was recorded from 5 different actors for training phase in recognition model. SMA set consists of 872 samples of 5 actors. Actors are required to perform the action accurately without any irrelevant movements. MMA set contains 1052 samples of 3 actors. Different from SMA, to collect MMA set, the 3 actors are asked to perform the actions while keeping other body parts moving. An example action of MMA set is that an actor may both wave hands and take a walk at the same time.

Table 3: Dataset #1

Index	Action type	Training sample set	SMA set	MMA set
1	Left hand low waving	5	145	180
2	Right hand low waving	5	145	181
3	Left hand high waving	5	145	179
4	Right hand high waving	5	146	181
5	Hand clapping	5	146	180
6	Greeting	5	145	151
	Total	**30**	**872**	**1052**

4.1.2 Dataset #2

The second dataset, dataset #2, is referenced from MSR Action3D [13] which consists of the skeleton data obtained from a depth sensor similar to Microsoft Kinect at 15 FPS. For the purpose of appropriate comparison, dataset #2 should include same types of actions relevant to dataset #1. Therefore, we select actions which are high arm waving, horizontal arm waving, hand low clapping, and hand high clapping for testing the performance of our recognition model. The actions are performed by 8 actors, with 3 repetitions of each action. The subset consisted of 85 samples in total. We used 49 samples of 5 actors for training and 36 samples of 3 actors for testing.

4.2. Experimental Results

The recognition accuracy for each action is summarized in Table 4 for both datasets. The recognition accuracy is the proportion between the correct label assigned for actions and their ground truths. It can be seen in the column of dataset #1 that the accuracy of CF presentation about 93.91% and 86.98% is highest for SMA and MMA sets respectively. The accuracies of OBRF presentation about 92.53% and 85.82% are much higher than those of TVRF presentation about 68.43% and 36.15% for SMA and MMA sets respectively. The reason of these gap is the experimental actor do some actions at the same time, then the TVRF presentation of each action is not only focused on the related joints. From these results, we can conclude that active BPAs empirically selected from observations are more efficient for action recognition than those automatically calculated by using time variance. The column of MMA set in Table 4 shows that the number of actions whose OBRF accuracy is higher than CF accuracy is 4 while this number in column of SMA set is 0. This observation can be used to prove the effectiveness of the feature reduction method OBRF in comparing with complete feature representation CF. The same conclusions can also be made when concerning with the experimental results of dataset #2.

Table 4: Accuracy (%) evaluation with Dataset #1 and Dataset #2; (*) low clapping; (**) high clapping

Test set		Dataset #1						Dataset #2		
		SMA set			MMA set					
Feature		CF	TVRF	OBRF	CF	TVRF	OBRF	CF	TVRF	OBRF
Action	1	93.79	63.34	88.27	98.33	17.77	94.44	55.56	33.33	55.56
	2	92.41	61.37	88.27	95.02	42.54	63.53	77.78	22.22	88.89
	3	89.65	56.55	89.65	67.59	46.36	82.12	N/A	N/A	N/A
	4	88.35	45.2	89.72	79.55	39.22	87.29	N/A	N/A	N/A
	5	99.31	95.89	99.31	82.77	30.00	88.88	(*)77.78 (**)33.33	(*)11.11 (**)11.11	(*)44.44 (**)66.67
	6	100	88.27	100	98.67	41.05	98.67	N/A	N/A	N/A
Average		**93.91**	**68.43**	**92.53**	**86.98**	**36.15**	**85.82**	**61.11**	**19.44**	**63.89**

Three computers with different specifications were used to run the tests and the average computation times for action training and testing each action of each dataset are presented in Table 5. It can be seen in the Table that the computation times of using TVRF and OBRF are shorter than those of using CF. It is resulted from the small size of feature matrix in TVRF and OBRF in comparing with the complete large size of feature matrix in CF. For all cases, OBRF shows the best time efficient method and, as discussed above, OBRF produces recognition accuracies comparative to CF in both SMA set and MMA set. Therefore, OBRF is a recommended candidate to build a real time application of human action recognition.

Table 5: Average computation time (in second) for data set #1

Computer Specification	SMA set			MMA set		
	CF	TVRF	OBRF	CF	TVRF	OBRF
Core i5, Ram 4GB	6.53	5.28	4.34	6.84	5.84	5.13
Core i3, Ram 2GB	14.24	12.03	11.28	13.23	12.45	12.11
Core 2 Duo, Ram 1GB	22.84	17.34	16.51	23.09	17.26	16.76

5. Conclusion

In this paper, we have proposed an approach to recognize human actions using 3D skeletal models. To represent motions, we constructed a very intuitive, yet interpretable features based on relative angles among body parts in skeletal model called CF and further be refined by OBRF and TVRF. The feature is computed from relative angles of every body parts in skeletal model which are invariant to body size and rotation. For classification purposes, DTW and voting algorithm were applied respectively. The evaluation of our method has been performed on a novel depth dataset of actions and a set from a Microsoft research. The results show that using OBRF obtains performance improvements in comparing with CF and TVRF in both recognition accuracy and computational complexity.

References

[1] M. Brand, N. Oliver and A. Pentland, "Coupled Hidden Markov models for complex action recognition," IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR), 1997, pp. 994 - 999.

[2] X. Lu, C. Chia-Chih and J.K. Aggarwal, "View Invariant Human Action Recognition Using Histograms of 3D Joints," IEEE Computer Society Workshops on Computer Vision and Pattern Recognition (CVPRW), 2012, pp. 20-27.

[3] J. Laffey, A. McCallum and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Int. Conf. on Machine Learning, 2001, pp. 282 - 289.

[4] C. Sminchisescu, A. Kanaujia, L. Zhiguo and D. Metaxas, "Conditional models for contextual human motion recognition," IEEE Int. Conf. on Computer Vision (ICCV), 2005, pp. 1808 - 1815.

[5] Sha Huang and Liqing Zhang, "Hidden Markov Model for Action Recognition Using Joint Angle Acceleration", Neural Information Processing, Lecture Notes in Computer Science Volume 8228, 2013, pp 333-340.

[6] Q.K. Le, C.H. Pham and T.H. Le, "Road traffic control gesture recognition using depth images", IEEK Trans. on Smart Processing & Computing, 2012, vol. 1, page 1 - 7.

[7] F. Ofli, F. Chaudhry, G. Kurillo, R. Vidal and R. Bajcsy, "Sequence of the Most Informative Joints (SMIJ): A New Representation for Human Skeletal Action Recognition," IEEE Computer Society Workshops on Computer Vision and Pattern Recognition (CVPRW), 2012, pp. 8-13.

[8] S. Salvado, and P. Chan, "Fast DTW: Toward Accurate Dynamic Time Warping in Linear Time and Space", KDD Workshop in Mining Temporal and Sequential Data, 2004, pp. 70-80.

[9] M. Reyes, G. Dominguez and S. Escalera, "Feature Weighting in Dynamic Time Warping for Gesture Recognition in Depth Data", IEEE Int. Conf. on Computer Vision (ICCVW), 2011, pp. 1182-1188.

[10] S. Sempena, N.U. Maulidevi and P.R Aryan, "Human action recognition using Dynamic Time Warping," Int. Conf. on Electrical Engineering and Informatics (ICEEI), 2011, pp. 1-5.

[11] J. Wang, H. Zheng, "View-robust action recognition based on temporal self-similarities and dynamic time warping," IEEE Int. Conf. on Computer Science and Automation Engineering (CSAE), 2012, pp. 498-502.

[12] W. Li, Z. Zhang and Z. Liu, "Action recognition based on a bag of 3D points," IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, 2010, pp.9 -14.

[13] PrimeSense, "Open NI". Available online at: http://openni.org