Unsupervised Dance Motion Patterns Classification from Fused Skeletal Data using Exemplar-based HMMs

by

A. Kitsikidis, N. V. Boulgouris, K. Dimitropoulos and N. Grammalidis

Reprinted from

International Journal of

Heritage in the Digital Era

volume 4 number 2 2015

Unsupervised Dance Motion Patterns Classification from Fused Skeletal Data using Exemplar-based HMMs

A. Kitsikidis[1], N. V. Boulgouris[2], K. Dimitropoulos[1] and N. Grammalidis[1*]

[1]Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece

[2]Brunel University, London, United Kingdom

*ngramm@iti.gr

Unsupervised Dance Motion Patterns Classification from Fused Skeletal Data using Exemplar-based HMMs

A. Kitsikidis, N. V. Boulgouris, K. Dimitropoulos and N. Grammalidis

Abstract:

In this paper, we propose a method for the partitioning of dance sequences into multiple periods and motion patterns. The proposed method deploys features in the form of a skeletal representation of the dancer observed through time using multiple depth sensors. This representation is the fusion of skeletal features captured using multiple sensors and combined into a single, more robust, skeletal representation. Using this information, initially we partition the dance sequence into periods and subsequently into motion patterns. Partitioning into periods is based on observing the horizontal displacement of the dancer while each period is subsequently partitioned into motion patterns by using an exemplar-based Hidden Markov Model that classifies each frame into an exemplar representing a hidden state of the HMM. The proposed method was tested on dance sequences comprising multiple periods and motion patterns providing promising results.

1. INTRODUCTION

Dance is an immaterial art relying on the motion of the performers' body. The capture, analysis and modelling of this motion with the help of ICT technologies could contribute significantly to the preservation and transmission of this intangible cultural heritage [1, 2]. However, the main challenge of this task lies in the accurate recognition of human body movements. Today, the major advantages over earlier systems include the ability to make more precise measurements with a wider array of sensing strategies, the increased availability of processing power to accomplish more sophisticated interpretations of data, and a greatly enhanced flexibility in the area of media rendering [3, 4, 5].

There exist different sensing technologies applied to motion capture, which can be broadly divided onto three categories depending on the degree of precision, the cost and the constraints posed by each technology. Optical motion capture is currently the most accurate motion capture technique, but it is also the most expensive one. It is based on the triangulation of reflective markers taped to the performer's body, which are detected by the surrounding cameras. Inertial motion sensors are less expensive but also less accurate. They are attached to the limbs and can track the angles between the body segments. Finally, markerless motion capture based on real-time depth sensing systems such as Microsoft Kinect [6, 7] can track the volume of a performer and produce skeletal data. These sensors are relatively cheap and offer a balance in usability and cost compared to optical and inertial sensors.

Human action and gesture recognition using markerless motion capture technologies can be coarsely grouped into two classes. Earlier approaches used sequences of depth maps to extract features and model the dynamics of the action explicitly. Bag of Words methods are employed as an intermediate representation with subsequent use of statistical models such as Hidden Markov models (HMM), graphical models (GM) and Conditional Random Fields (CRF) [8]. More recent approaches that use skeletal data, such as 3D positions of human joints, calculated from the depth maps [9]. Joint position trajectories are used in conjunction with Dynamic Time Warping (DTW) variants in order to classify the motion patterns [10]. Another approach is the extraction of features from the whole skeleton in histogram form and the use of statistical models for classification [11].

In this paper we present a method for dance capture and analysis in the form of automatic recognition of motion patterns, which constitute the choreography. More specifically, we use multiple synchronized depth sensors for skeletal data acquisition, and we extract low level features, such as skeletal joint positions in 3D space along with rotation angles. The features extracted from each sensor are combined then by applying a skeletal fusion procedure. The representation of a dance sequence in terms of these features requires a suitable classification procedure that will be able to detect temporal segments of the sequence that correspond to dance patterns. Since, Hidden Markov Models have been shown to be very efficient in representing motion due to their ability to capture dynamics of motion [12], we use exemplar-based HMMs for partitioning of dance sequences into their constituent periods and motion patterns. The proposed method is completely unsupervised, allowing the automatic identification of exemplars, and computationally efficient.

2. OVERVIEW OF THE PROPOSED SYSTEM

The proposed system is based on four main processes: feature extraction, period detection, unsupervised HMM training, and dance pattern partitioning (Fig. 1). In the first stage, a sequence of detailed features is captured based on a human body model. The Microsoft Kinect SDK is used for the acquisition of skeletal data, which provides position and rotation information of 20 predefined joints of a human body shown in Fig. 2. For each joint, a confidence value of tracking is also provided, which can be high, medium or low. Subsequently, these skeletal data, captured by each sensor, are combined into a single skeletal representation using a fusion procedure.

The sequence of features provided by the fused skeleton is used then in order to partition the sequence into its periods, i.e., the parts of the sequence that are repeated through time. Once the periods are known, each period is used for the automatic training of an exemplar-based HMM, where the probability that an observation was generated by a state is a function of the distance between the observation and the exemplar that represents the HMM state.

Initially, the HMM exemplars for each period are iteratively calculated and subsequently, the sequence of features can be seen as being generated by a sequence of HMM states. The interpretation

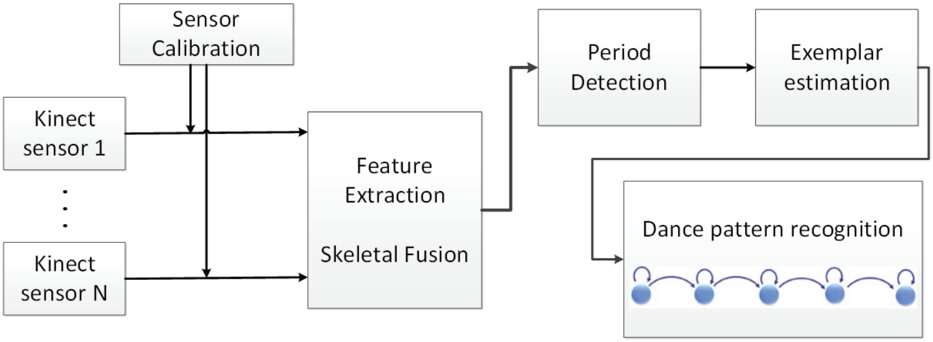
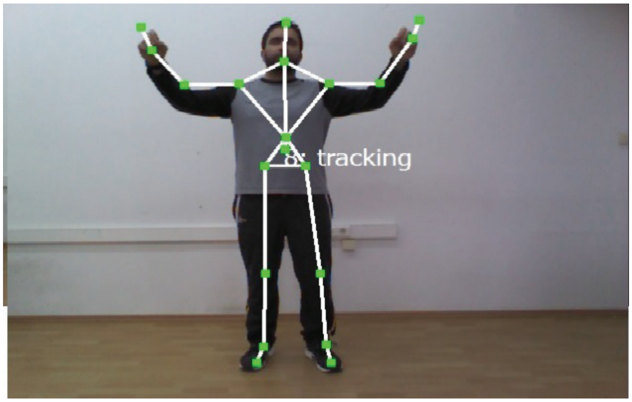


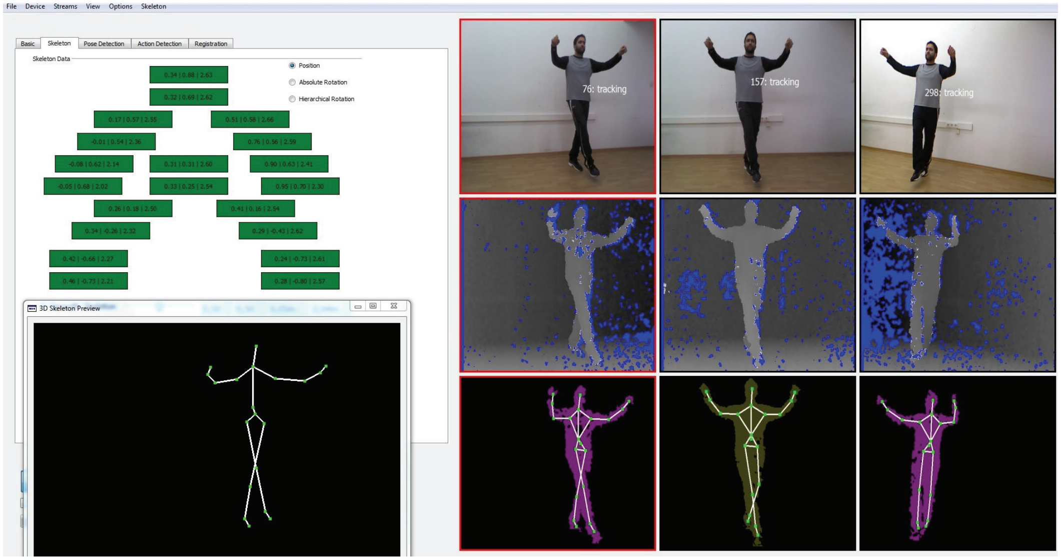
Figure 1. System overview

Figure 2. Skeletal tracking overlaid on top of the color map frame as captured by the sensor. 20 joints connected with bone links can be observed.



of the sequence of states that have been generated by the sequence of features, yields the dance patterns in each dance period. The whole process is considered as entirely unsupervised.

Figure 3. Motion capture with three Kinect sensors placed around the dancer, with skeleton fusion combining the three initial skeletons.



Skeleton Data

● Position
○ Absolute Rotation
○ Hierarchical Rotation

0.34 | 0.88 | 2.63
0.32 | 0.69 | 2.62
0.17 | 0.57 | 2.55
0.51 | 0.58 | 2.66
-0.01 | 0.54 | 2.36
0.76 | 0.56 | 2.59
-0.08 | 0.62 | 2.14
0.31 | 0.31 | 2.60
0.90 | 0.63 | 2.41
-0.05 | 0.68 | 2.02
0.33 | 0.25 | 2.54
0.95 | 0.70 | 2.30
0.26 | 0.18 | 2.50
0.41 | 0.16 | 2.54
0.34 | -0.26 | 2.32
0.29 | -0.43 | 2.62
0.42 | -0.66 | 2.27
0.24 | -0.73 | 2.61
0.46 | -0.73 | 2.21
0.28 | -0.80 | 2.57

3. SKELETON FUSION

Initially, the motion capture is performed and skeletal tracking data are combined by fusing the data. Skeleton fusion is the process of combining skeletal data recorded by multiple sensors into a single, more robust skeletal representation. It helps address problems occurring due to occlusions and self-occlusions of parts of the dancer’s body, for example when one leg is raised in front of the other, thus hiding the other leg from the camera. In addition, combining skeletal tracking data from multiple sensors allows for larger capturing area, since the total field of view can be increased, depending on the placement of the sensors. For our recording sessions, we used three depth sensors placed in arc topology in front of the dancer, thus allowing the dancer for more room to move (Fig. 3). In addition, fusion decreases the noise inherent in skeletal tracking data of the depth sensor. The concurrent use of depth sensors creates some interference due to infrared emission, which depends on the number of sensors used and the topology of their placement [13]. We found that using three sensors in an arc topology provided significant benefits due to skeletal fusion without major degradation of the original skeletal data due to interference.

Prior to fusion, skeletal data from all sensors have to be transformed to a common reference coordinate system of the reference sensor, a process called registration. In order to perform skeleton registration, the sensors have to be calibrated; i.e. a rigid transformation of position and orientation of each sensor pair must be estimated. Our calibration procedure does not require checker boards or similar patterns; instead it uses the point clouds of the skeleton joints as detected by each sensor. We run the Iterative Closest Point (ICP) algorithm [14] on the joint point clouds to estimate the rigid transformation that minimizes the distance between the two point clouds. This transformation is then used to register the skeletons acquired from each sensor in the reference coordinate system. We have used the implementation of ICP algorithm found in the Point Cloud Library (PCL) [15].

Once the skeletons are registered, they are combined according to a specific fusion strategy. Specifically, we have developed a fusion strategy working on joint positional data, which could be extended to rotational data as well with slight modifications. Initially, the sum of all joint confidence levels of each skeleton is computed and the skeleton with the highest total is selected. This skeleton consists of most successfully tracked joints and it is expected to be the most accurate representation of the real person posture. We consider the joints of this skeleton as base and construct the fused skeleton joints in the following manner, by examining the confidence values of each joint of the base skeleton. There are three possible values of confidence: high, medium and low. If the confidence of the base joint is high, it is left as is for the fused skeleton. Otherwise, if the confidence is medium or low, the joint position is corrected by taking into account the remaining skeletons. If corresponding joints with high confidence are found in any of the remaining skeletons, their average position is used to replace the position value of the joint. If there are no corresponding joints with high confidence, the same procedure is applied for joints tracked with medium confidence. Lastly, if only low confidence joints exist, their average is used as a position value of the fused joint.

Since joint averaging, switching from one base skeleton to another in conjunction with sensor calibration inaccuracies can introduce artifacts in the form of sudden rapid changes in joint position from frame to frame, a filtering stabilization step has been introduced which is applied to the fused skeleton stream. A time window of three frames is used in order to keep the last three high-confidence positions per joint. The centroid of these three positions is calculated and updated at each frame. If the Euclidean distance between a joint position and this centroid is higher than a certain threshold, then the joint position is replaced by the centroid, so as to avoid rapid changes in joint positions. The thresholds are different for each joint, since it is expected that some joints (hands and feet) move more rapidly than others.

4. PARTITIONING INTO PERIODS

The dance sequence is initially partitioned into periods. For the purpose of sequence partitioning, we used the horizontal displacement of the waist of the dancer during dancing. This displacement, shown in Figure 4a, exhibits a specific periodic behavior and its minima indicate the end point of a period and the beginning of another. In this way, the detection of periods can be trivially achieved by detecting all minima. Experimental application of this method showed that it detects periods in a remarkably accurate and consistent way.

5. UNSUPERVISED EXEMPLAR-BASED HMM CLASSIFCATION

Once periods in the dance sequence have been identified, the features extracted from each period will be treated as a sequence of observations to be used to determine the parameters of the HMM. The sequence of such observations, extracted using the process above, will be denoted

$$H = h_1, h_2, h_3, \ldots, h_T$$

where h_t, $t = 1, \ldots, T$ is a vector representing the features extracted from each frame in the dance sequence. The feature vector selected, consists of position and rotation data of eight lower body joints. The position features are the location of the joint in 3D space and the rotation features consist of a quaternion which contains the relative rotation of the bone connecting the joint to its parent in the skeletal hierarchy. We selected the hip, knee, ankle and foot joints of both legs, since the leg movements were considered to have a greater discriminating power in dance pattern detection.

Initially, the HMM parameters and the exemplars associated with each HMM state need to be determined. It must be emphasized, however, that the eventual partitioning of dance periods into patterns will be deduced based on the classification of frames to exemplars. In this sense, the proposed method is unsupervised and does not require a separate set of data for training and testing but, instead, the test data are processed through the training of the HMM.

The calculation of the parameters of the HMM is straightforward and follows that in [16]. These parameters will be denoted $\lambda = \{\pi, A, e\}$, where π denotes the *initial state probabilities*, A denotes the matrix of *state transition probabilities*, and e is the set of *state exemplars*. A left-to-right HMM is considered, the initial state probabilities of which are

$$\pi_n = \begin{cases} 1 \text{ if } n = 1 \\ 0 \text{ otherwise} \end{cases} \qquad (1)$$

where π_n is the initial state probability for the *n*th state, $n = 1,\ldots, N$, of the HMM.

The probability that the feature observation h_t at time t is generated by state q_t is denoted $b_{qt}(\mathrm{h}_t)$. For this reason, the output probability distribution is calculated, similar to [16, 17, 18], based on the distance of observation h_t from the state exemplar e_{qt}

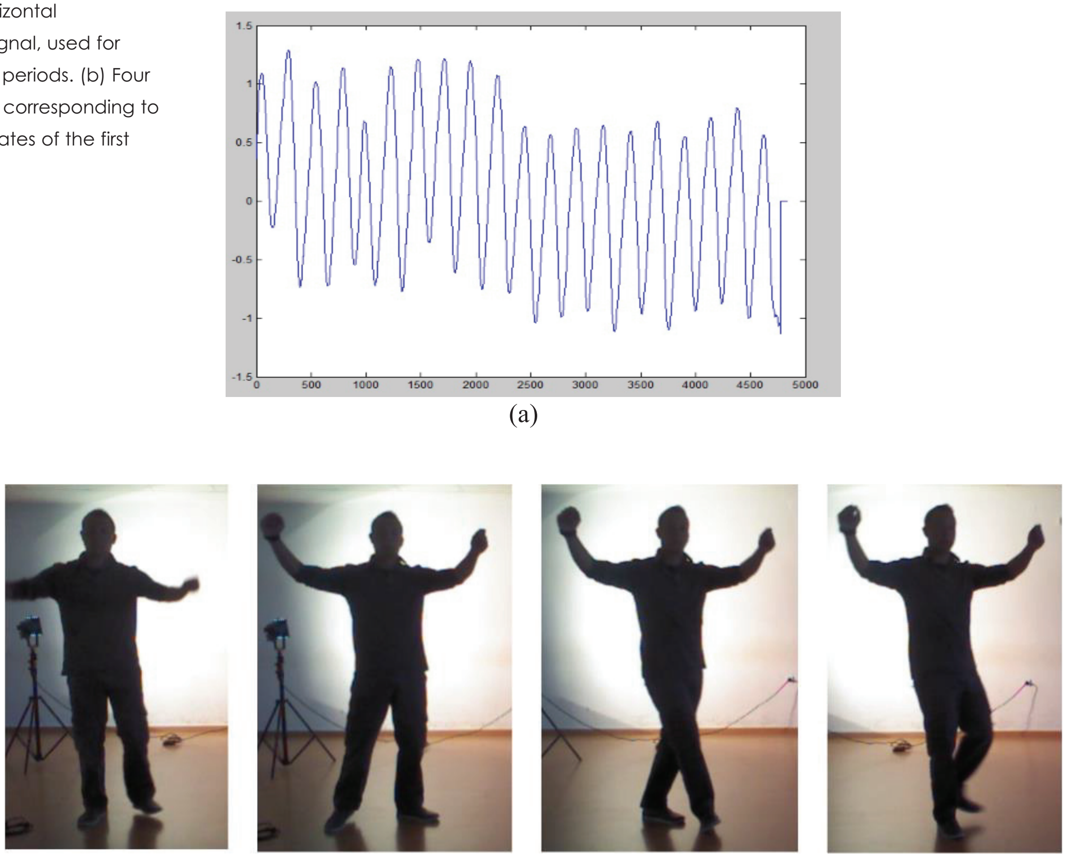
$$b_{qt}(\mathrm{h}_t) = P(\mathrm{h}_t \,|\, e_{qt}) = Ve^{-\upsilon D(\mathrm{h}_t, e_{qt})} \qquad (2)$$

where V and υ are normalization parameters. The parameters of the HMM are calculated using expectation-maximization. The similarity between frames and exemplars is assessed based on a Euclidean distance and the probability that the features of a frame have been generated by a state is given by eq. (2) above. A trellis is formed and the sequence of features in a period is classified to HMM states by calculating the trellis path that corresponds to the highest probability. Exemplars are updated as the average of the features that appear to have been generated by the respective states. This process is iterated until convergence. Usually conversion is reached within a few iterations. The patterns in a period are deduced from the frames that have been allocated to each HMM state.

6. EXPERIMENTAL RESULTS

The proposed method was applied on a dance sequence of Greek traditional Tsamiko dance using a setup of three depth sensors, placed in front of the dancers. The dancers were instructed to dance on a straight line, facing the sensors. No complex figures were performed, due to the inherent limitations of the Kinect skeleton tracker. In order to assess the efficiency of the proposed method, we captured 3 experts and 9 students dancing Tsamiko. The recorded sequences were manually annotated and the start and end frames of each period and

Figure 4. (a) Horizontal displacement signal, used for partitioning into periods. (b) Four exemplar poses corresponding to the four HMM states of the first motion pattern.



(a)

(b)

each pattern were marked. Spatial coordinate and orientation features were extracted from the sequences, as described in Section 2, and were subsequently used as input to our system.

First, the sequences were automatically partitioned into periods as described in Section 4, by tracking the local minima of waist joint horizontal displacement. Since, one period of Tsamiko dance consists of several moves in the right direction followed by movement to the left, the detection of horizontal body displacement works remarkably well for the segmentation of a dance period. Subsequently each period was further partitioned into its constituent patterns. Each period of the single step version of Tsamiko dance can be split into three basic motion patterns. We used a 15-state HMM, with the first 4 states corresponding to the first motion pattern (Fig. 4b). The remaining 5 and 6 states were assigned to the second and third motion pattern respectively, from which a dance period consists.

Recording	Segments 1/2	Segments 2/3	Periods
Expert A	0.4%	2.83%	0.3%
Expert B	2.66%	2.66%	0.1%
Expert C	0.21%	1.69%	0.42%
Average	1.09%	2.39%	0.27%

Table 1. Segmentation error rates of the proposed method applied to dance recordings

Recording	Segments 1/2	Segments 2/3	Periods
Student A	2.64%	6.41%	1.06%
Student B	0.2%	4.14%	0.01%
Student C	0.79%	3.97%	0.4%
Student D	2.24%	3.86%	1.38%
Student E	1.38%	4.56%	0.01%
Student F	0.01%	3.99%	0.01%
Student G	3.76%	8.78%	3.10%
Student H	0.19%	3.78%	1.07%
Student I	1.77%	4.73%	0.01%
Average	1.44%	4.91%	0.78%

Table 2. Segmentation error rates of the proposed method applied to dance recording of nine students.

To measure the accuracy of our automatic segmentation, we calculate the segmentation error as the percentage of the distance between ground truth segment and the detected segment normalized to the length of the whole dance period. Also we consider a deviation of 3 frames from the ground truth should not be regarded as error since the manual annotation of patterns has a small margin for error by a few frames. So, all the error percentages are calculated after a subtraction of 3 frames from them. As can be seen from the tables below, the automatic partitioning into periods we applied came within 0.27% for the experts and 0.78% for the students respectively, of the manually annotated sequences (Table 1). Also, the determined pattern segmentation came within 1.09% - 2.39% for the experts and 1.44% - 4.91% for the students, of the manually annotated. As expected, the accuracy of segmentation of the expert recordings is greater since the movements are 'cleaner' and devoid of mistakes in dance execution. Also we observe that the detection of dance periods has increased accuracy relative to the detection of smaller dance segments.

7. CONCLUSIONS

We proposed a method for the partitioning of dance sequences into periods and patterns. The proposed method deployed features in the form of a skeletal representation of the dancer observed through time. This representation was the fusion of skeletal features captured using multiple sensors and combined into a single, more robust, skeletal representation. Using this information, initially we partitioned the dance sequence into periods and subsequently into patterns. Partitioning into periods was based on observing the horizontal displacement of the dancer while each period was subsequently partitioned into patterns by means of training an exemplar-based Hidden Markov Model that classified frames to exemplars representing HMM states. The proposed method was tested on a multiple dance sequences comprising multiple periods and patterns and was seen to have excellent performance.

ACKNOWLEDGEMENTS

Part of the research leading to these results has received funding from the European Community’s Seventh Frame-work Programme (FP7-ICT-2011-9) under grant agreement no FP7-ICT-600676 ‘’i-Treasures: Intangible Treasures - Capturing the Intangible Cultural Heritage and Learning the Rare Know-How of Living Human Treasures’’.

REFERENCES

[1] Miura, T., Kaiga, T., Katsura, H., Shibata, T., Tajima, K., Tamamoto, H.: Quantitative Motion Analysis of the Japanese Folk dance “Hitoichi Bon Odori,” *IPSJ Symposium Series*, 2013, 4, 167—174.

[2] Usui, Y., Sato, K. & Watabe, S. (2013). The Effect of Motion Capture on Learning Japanese Traditional Folk Dance. In J. Herrington, A. Couros & V. Irvine (Eds.), Proceedings of EdMedia: World Conference on Educational Media and Technology 2013 (pp. 2320-2325). Association for the Advancement of Computing in Education (AACE).

[3] Essid, S., Alexiadis, D.S., Tournemenne, R., Gowing, M., Kelly, P., Monaghan, D.S., Daras, P., Dremeau, A., O’Connor, N.E.: An advanced virtual dance performance evaluator. I*n: ICASSP*. pp. 2269-2272. IEEE (2012)

[4] Aristidou, A., Stavrakis, E., Chrysanthou, Y.: “LMA-Based Motion Retrieval for Folk Dance Cultural Heritage”, In Proceedings of the 5th International Conference on Cultural Heritage (EuroMed 2014), Limassol, Cyprus, November 3-8, 2014.

[5] Aristidou, A., Stavrakis, E., Chrysanthou, Y.: Motion analysis for folk dance evaluation. In: EG Workshop on Graphics and Cultural Heritage, GCH 2014. Eurographics (2014)

[6] Kyan, M., Sun, G., Li, H., Zhong, L., Muneesawang, P., Dong, N., Elder, B., Guan L.: An Approach to Ballet Dance Training through MS Kinect and Visualization in a CAVE Virtual Reality Environment. *ACM TIST* 6(2): 23 (2015)

[7] Kinect for windows, voice, movement & gesture recognition technology (2013), http://www.microsoft.com/en-us/kinectforwindows/

[8] Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 *IEEE Computer Society Conference on*. pp. 9-14 (June 2010)

[9] Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: Computer Vision and Pattern Recognition (CVPR), 2012 *IEEE Conference on*. pp. 1290-1297 (June 2012)

[10] Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A.W., Finocchio, M., Blake, A.,Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. *Commun. ACM* 56(1), 116-124 (2013)

[11] Waithayanon, C., Aporntewan, C.: A motion classier for microsoft kinect. In: Computer Sciences and Convergence Information Technology (ICCIT), 2011 *6th International Conference on*. pp. 727-731 (Nov 2011)

[12] Boulgouris, N.V., Huang, X.: Gait recognition using hmms and dual discriminative observations for sub-dynamics analysis. *IEEE Trans. Image Processing* 22(9), 3636 - 3647 (Sep 2013)

[13] Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3d joints. In: Workshop on Human Activity Understanding from 3D Data in conjunction with CVPR (HAU3D). Rhode Island, USA (2012)

[14] Caon, M., Yong, Y., Tscherrig, J., Mugellini, E., Abou Khaled, O.: Context-aware 3D gesture interaction based on multiple kinects. In: The First International Conference on Ambient Computing, Applications, Services and Technologies. p. 712. Barcelona, Spain (Oct.)

[15] Besl, P.J., McKay, N.D.: A method for registration of 3-d shapes. Pattern Analysis and Machine Intelligence, *IEEE Transactions on* 14(2), 239-256 (Feb 1992)

[16] Kale, A., Sundaresan, A., Rajagopalan, A.N., Cuntoor, N., Roy-Chowdhury, A.K., Krueger, V., Chellappa, R.: Identication of humans using gait. *IEEE Trans. Image Processing* 13(9), 1163-6173 (Sep 2004)

[17] Chen, C., J.-Liang, H.-Zhao, H.-Hu, J.-Tian: Factorial hmm and parallel hmm for gait recognition. IEEE Trans. Systems, Man, and Cybernetics-Part C: applications and reviews 39, 114-123 (2009)

[18] Liu, Z., Sarkar, S.: Improved gait recognition by gait dynamics normalization. *IEEE Trans. on Pattern Anal. and Machine Intell.* 28(6), 863-876 (2006)