Dance Analysis using Multiple Kinect Sensors

Alexandros Kitsikidis[1], Kosmas Dimitropoulos[1], Stella Douka[2] and Nikos Grammalidis[1]

[1]*Informatics and Telematics Institute, ITI-CERTH, 1st Km Thermi-Panorama Rd, Thessaloniki, Greece*

[2]*Department of Physical Education and Sport Science, Aristotle University of Thessaloniki, Greece*

{ajinchv, dimitrop, ngramm}@iti.gr, sdouka@phed.auth.gr

Keywords: Body Motion Analysis and Recognition, Conditional Random Fields, Skeletal Fusion, Dance Analysis.

Abstract: In this paper we present a method for body motion analysis in dance using multiple Kinect sensors. The proposed method applies fusion to combine the skeletal tracking data of multiple sensors in order to solve occlusion and self-occlusion tracking problems and increase the robustness of skeletal tracking. The fused skeletal data is split into five different body parts (torso, left hand, right hand, left leg and right leg), which are then transformed to allow view invariant posture recognition. For each part, a posture vocabulary is generated by performing *k*-means clustering on a large set of unlabeled postures. Finally, body part postures are combined into body posture sequences and Hidden Conditional Random Fields (HCRF) classifier is used to recognize motion patterns (e.g. dance figures). For the evaluation of the proposed method, Tsamiko dancers are captured using multiple Kinect sensors and experimental results are presented to demonstrate the high recognition accuracy of the proposed method.

1 INTRODUCTION

Dance is an immaterial art as it relies on the motion of the performer's body. Dance can convey different messages according to the context, and focus on aesthetics or artistic aspects (contemporary dance, ballet dance), the cultural and social aspects (folk dances, traditional dances), story telling (symbolic dances), spiritual meanings (whirling dervishes), etc. Especially traditional dances are strongly linked to local identity

and culture. The know-how of these dances survives at the local level through small groups of people who gather to learn, practice and preserve these traditional dances. Therefore, there is always a risk that certain elements of this form of intangible cultural heritage could die out or disappear if they are not safeguarded and transmitted to the next generation.

ICT technologies can play an important role towards this direction. Specifically, the development of a system for the capturing, analysis and modelling of rare dance interactions could significantly contribute to this transfer of knowledge. 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 (Aylward, 2006).

Depending on the degree of precision of the captured motion and the constraints posed, different sensing technologies are used. They can be broadly divided into three main categories: optical motion capture, inertial motion capture and markerless motion capture. Optical motion capture is the most accurate technique but it is also expensive and constraining. Inertial motion capture is less accurate and less stable. Finally, markerless motion capture based on real-time depth sensing systems, such as Microsoft Kinect, is relatively cheap and offer a balance in usability and cost compared to optical and inertial motion capture systems. To this end, this approach is considered as the most promising one and has attracted particular attention recently (Alexiadis et al., 2011).

Existing approaches to human action and gesture recognition using markerless motion capture technologies can be coarsely grouped into two classes. The first uses 3D depth maps / silhouettes which form a continuous evolution of body pose in time. Action descriptors, which capture both spatial and temporal characteristics, are extracted from those sequences and conventional classifiers can be used for recognition. The other category of the methods extracts features from each silhouette and model the dynamics of the action explicitly. Bag of Words (BoW) are often 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) (Li et al., 2010) (Wang et al., 2012)

Another more recent approach is to use the skeletal data acquired from the depth maps (Shotton et al., 2011). The subsequent use of skeletal data for action detection can be divided into two categories. The methods of the first category are based on 3D joints feature trajectories (Waithayanon and Aporntewan, 2011). Those features are either joint position, rotation data, or some transformation of the above. They are mainly based on various Dynamic Time Warping (DTW) variants, like multi-dimensional time warping (MD-DTW) (ten Holt et al., 2007). The recognition is based on the alignment of the movement trajectories compared to the 'oracle' move which is being detected. Another approach is to extract features from the whole skeleton (histograms) and to use statistical models as in the case of silhouette based methods (Xia et al., 2012).

In this paper, we present a method for dance capture, analysis and recognition using a multi-depth sensors set-up and a skeleton fusion technique to address occlusion problems and increase the robustness of the skeletal tracking. Subsequently, we propose the splitting of the skeleton into five different parts, and the automatic generation of a posture vocabulary (codebook) for each part. Finally, a Hidden State Conditional Random Field (HCRF) (Quattoni et al., 2004; Wang et al., 2006) is applied for the recognition of the dance figures (motion patterns). Experimental results with real Tsamiko dancer (Tsamiko is a traditional Greek dance) have shown the great potential of the proposed method.

2 SYSTEM OVERVIEW

The flowchart of the data acquisition and motion recognition process is presented in Figure 1. Several Kinect sensors placed around the subject are used to acquire skeletal animation data. Microsoft Kinect SDK (Kinect for Windows, 2013), has been used as a solution for skeletal tracking and acquisition. It provides 3D position and rotation data (relative to a reference coordinate system centred at the origin of the sensor) of 20 predefined skeletal joints of a human body. In addition a confidence level of the joint tracking (low/medium/high) is provided per joint. A skeletal fusion is proposed to combine the data coming from multiple sensors onto a single fused skeleton, which is then provided to the Motion Analysis Module. Specifically, the skeleton is split into five body parts (torso, left/right hand, left/right foot), which are then transformed to allow view invariant posture recognition. The next step is to recognize each body part posture appearing in a frame, based on a predefined vocabulary of postures, obtained from a set of training sequences. Finally, body part postures are

combined into body posture sequences and an HCRF is used to recognize a motion pattern (e.g. a dance move) from a predefined set of motion patterns from which the HCRF was previously trained.

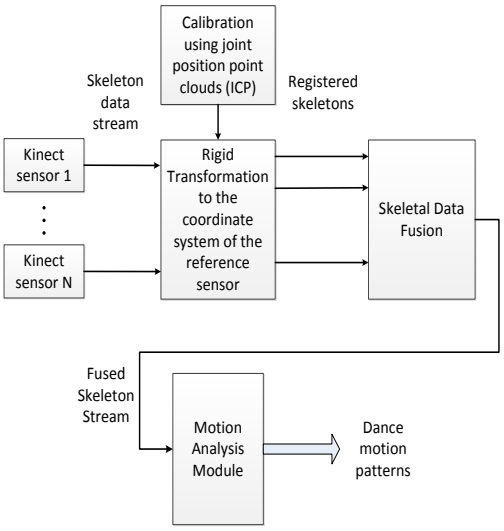

Figure 1: System overview

2.1 Calibration

In order to improve the robustness of skeleton tracking provided by the Microsoft Kinect SDK, to reduce occlusion and self-occlusion problems and to increase the area of coverage, multiple Kinect devices were used. Prior to fusion, skeletal data from all sensors have to be transformed to a common reference coordinate system. One sensor is selected as the reference sensor providing the reference frame. A calibration procedure is then required to estimate the transformations between the coordinate systems of each sensor and the reference sensor. The proposed calibration procedure does not require any checker boards or similar patterns. Instead, the only requirement is that a person needs to be visible from multiple sensors, whose FOV's need to partially overlap. The skeleton joint positions are then fed into the Iterative Closest Point algorithm (Besl and McKay, 1992) to estimate the rigid transformation (Rotation-Translation) that minimizes the distance between the transformed positions in the reference frame. This transformation is then used to register the skeletons acquired

from each sensor in the reference coordinate system. The implementation of ICP algorithm found in the Point Cloud Library (PCL, http://pointclouds.org/), (Rusu, 2011) was used.

Since the skeleton frame data are sparse, each containing at most 20 points, and the estimation of the joint positions can be erroneous, the calibration procedure is iterated until two convergence criteria are both met. The first criterion is that the number of joints tracked with high confidence on both devices needs to be higher that a threshold T_{joints}. The higher this number is the better the expected accuracy of the calibration. The second criterion is that the fitness score of the ICP algorithm needs to be lower than a threshold T_{ICP}. These thresholds can be adjusted to accommodate various setups and recording conditions.

2.2 Skeleton Fusion

Once all sensors are calibrated, skeleton registration is performed, i.e. the representation of each skeleton is transformed to the reference coordinate system. This is accomplished by multiplying the skeleton joint positions obtained from each sensor by the corresponding RT matrix, estimated in the calibration process. Then, a skeletal fusion procedure is used to combine these registered skeletons into a single skeleton representation (Figure 2).

Specifically, the following fusion strategy has been used on joint positional data, but could easily be extended on joint rotations as well. Initially, the sum of all joint confidence levels of each skeleton is computed and the skeleton with the highest total is selected. Since this is the skeleton with the most successfully tracked joints, it is expected to be the most accurate representation of the real person pose.

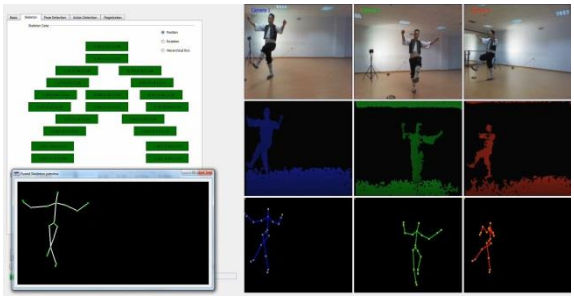

Figure 2: Colour maps, depth maps and skeleton frames from 3 Kinect sensors and a fused skeleton result.

We use this skeleton as a base, and enrich it with data provided from the remaining skeletons. Specifically, the confidence of each joint of the base skeleton is examined. If the confidence is medium or low, the joint position is corrected by taking into account the position of this joint in 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. Otherwise, the same procedure is applied for joints containing medium confidence values. Finally, if only low confidence values exist, the same procedure is applied using the available skeleton data for the joint.

As a last step, a stabilization filtering step is applied in order to overcome problems due to rapid changes in joint position from frame to frame which may occur because of the use of joint position averaging in our fusion strategy. We use a time window of three frames, to keep the last three high-confidence positions for each joint. The centroid of these three previous positions is calculated and updated for each frame. If the Euclidean distance between a joint position and this centroid is higher than a certain threshold, then we replace the joint position with the value of the centroid, so as to avoid rapid changes in joint positions. We have used different thresholds for each joint since hands and feet are expected to move more rapidly.

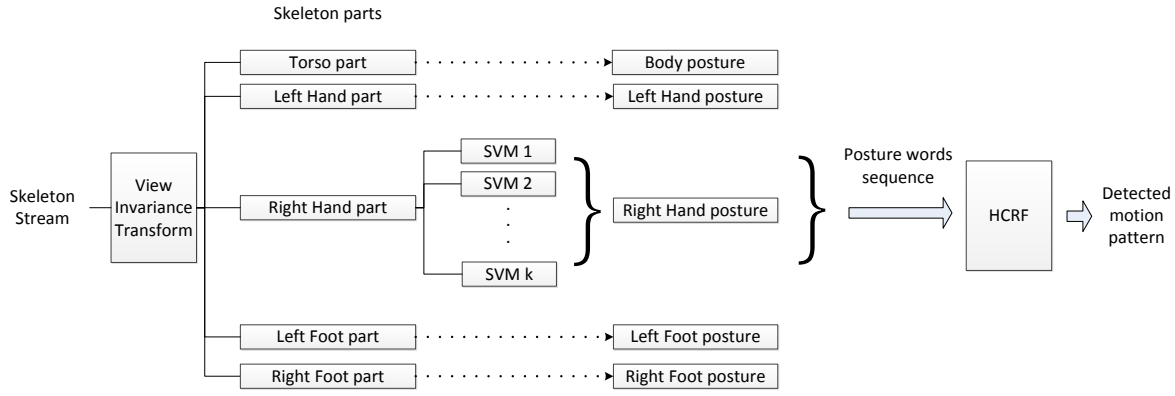

Figure 3: Motion Analysis subsystem

3 MOTION ANALYSIS

The motion analysis subsystem (Figure 3) can use as input a skeleton animation stream, either provided from a single Kinect device, or from multiple Kinect devices, after using the skeleton fusion procedure described in Section 2.2.

Initially, to achieve view invariance of motion recognition, the skeleton joint positions are translated relative to the root of the skeleton (*Hip Center*) and rotated around the y axis so that the skeleton is facing towards the y axis. Next, the skeleton is divided into five parts, shown in Figure 4 (torso, left hand, right hand, left foot, right foot). Each part has a root joint and children joints. For each skeleton part we generate a feature vector consisting of positions of each joint relative to the root of the part (also shown in Figure 4). Specifically, the root of the torso part is the *Hip Center* and the children joints are: *Spine*, *Shoulder Center* and *Head*. The root of the left hand part is the *Shoulder Center* and the children are: *Left Shoulder*, *Left Elbow*, *Left Wrist* and *Left Hand*. The root of the left foot part is the *Hip Center* and the children are: *Left Hip*, *Left Knee*, *Left Ankle* and *Left Foot*. The right hand and right foot parts consist of the symmetrical joints of their left counterparts.

3.1 Posture codebook

For each of the 5 skeleton parts described above, we construct a codebook of basic postures of a predefined size k. The identification of these basic postures is performed automatically by using *k*-means clustering of a large set of postures obtained from one or more recorded training sequences. Clustering essentially divides the 'posture space' into *k* discrete posture subspaces. After building a posture codebook for each body part, we train a multiclass SVM classifier to classify each incoming feature vector as a specific posture from this posture codebook. Thus we obtain five posture classifiers, one per body part.

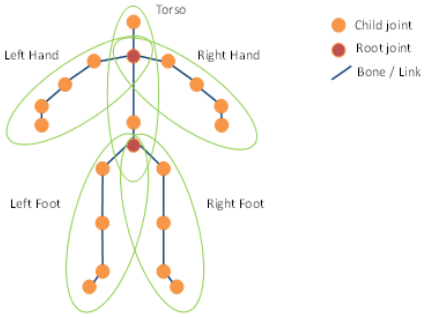

Figure 4: Skeleton Parts

3.2 Motion pattern recognition using a HCRF model

For the motion detection step, we have selected the Hidden-state Conditional Random Fields (HCRF) classifier (Quattoni et al., 2004). A set of *M* basic motion patterns, i.e. sequences of frames of skeleton data describing a specific movement, is first identified. We then train an HCRF multi-class model for each of these basic motion patterns. Specifically, for the training phase, we use labelled sequences of the basic motion patterns. Each sequence consists of a sequence of labelled skeleton part posture vectors, i.e. vectors of five elements, each being the index of a basic posture from the codebook corresponding to the specific skeleton part. For the testing phase, a similar vector is initially estimated for each frame of the input skeleton data sequence and is then used as input to the HCRF classifier. Then, the identification of each motion pattern is based on the probability/likelihood of the model of the HCRF for each observation sequence. For the implementation of HCRF, the Hidden-state Conditional Random Fields Library v2 was used (http://sourceforge.net/projects/hcrf/).

4 EXPERIMENTAL RESULTS

To evaluate our methodology, a data recording session took place, in which several dancers were recorded performing the Tsamiko dance (Figure 5).

Tsamiko is a popular traditional folk dance of Greece, done to music of ¾ meter. It is a masculine (mostly) circular dance with more smoothly steps danced by women. It is danced in an open circle where the first dancer performs variations while the others follow the basic steps. Tsamiko is danced in various areas of Greece such as: Peloponnese, Central Greece, Thessaly, W. Macedonia, with variations in kinesiological structure (10, 12,

8, 16 steps). The dance follows a strict and slow tempo with emphasis on the "attitude, style and grace" of the dancer. The steps are relatively easy but have to be precise and strictly on beat. Its variations consist of both smooth and leaping steps, which give the dance a triumphant air. The handle is hand elbows in position W. The dance is accompanied by various songs.

Figure 5: Recording session

For the evaluation of our methodology we recorded three male dancers, dancing the single step version of the Tsamiko dance (Figure 6). The main dance pattern in Tsamiko can be split in 3 basic dance moves, which were used as the basic motion patterns that we tried to detect. The recordings were manually annotated to mark the beginning and end of each move. Each dancer was recorded separately and was required to perform the basic moves of the dance several times.

Figure 6: Tsamiko dance steps

4.1 Sensors setup

For the final recording we used three Kinect sensors placed in front of the dancer in an arc topology (one in front and two at the sides), as seen in Figure 7(b). One additional setup was tested, but was rejected for the final recording. We tried placing four Kinect sensors all around the dancer, at 90 degree angle between them, as seen in Figure 7(a). This setup allowed for approximately 2x2m active space for the dancer to move. The interference due to infrared emission from the sensors was minimal, but only the two frontal sensors provided useful skeletal data, since skeletal tracking of Microsoft SDK is designed to work on people facing towards the sensor. Since the dancers were moving on a small arc, they were always facing in the same direction. Thus, our final setup proved to be more effective since we had skeletal tracking data from three sensors. In addition, having a smaller angle between adjacent sensor FOVs allowed for increased precision of calibration. Adding

more sensors proved to be problematic since interference caused by the emission of infrared pattern by each sensor increased significantly, which had a negative impact on skeletal tracking.

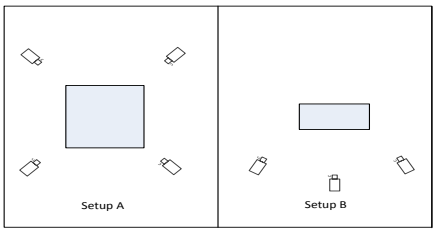

Figure 7: Sensor setups (a) Setup A (b) Setup B (final setup)

4.2 Evaluation results

The recorded data consisted of eight repetitions of basic Tsamiko dance pattern (three dance moves per repetition) executed by each of the three dancers. We split the recorded data into train and test sets by using half repetitions of the basic dance pattern of each dancer (12 repetitions per move) for training and the remaining for testing. Initially the posture codebook was created with a codebook of k=20 basic postures for each body part using the training motion sequences. Then, we trained an HCRF using the train sequences to be able to distinguish between the three basic Tsamiko dance moves. CRFs with a varying number of hidden states were trained as can be seen from Table 1, in which the dance move detection accuracies of the test set are presented, per dancer and overall. The best overall detection accuracy that was achieved is 93,9% using an HCRF with 11 hidden states. In Table 2, detection accuracies are presented for each dance move.

Table 1: Recognition accuracies of Tsamiko dance moves per person and overall recognition accuracies for varying number of hidden states in the HCRF classifier.

Hidden States	5	8	11	12	15	20
Dancer A	38,4	61,5	84,6	76,9	76,9	69,2
Dancer B	90,9	90,9	100	100	90,9	72,7
Dancer C	66,6	88,8	100	100	100	77,7
Overall	63,6	78,7	**93,9**	90,9	87,8	72,7

Table 2: Recognition accuracies of Tsamiko dance moves for varying number of hidden states in the HCRF classifier.

Hidden States	5	8	11	12	15	20
Dance move 1	83,3	66,6	91,6	100	83,8	100
Dance move 2	27,2	81,8	90,9	81,8	90,9	36,3
Dance move 3	80	90	100	90	90	80
Overall	63,6	78,7	**93,9**	90,9	87,8	72,7

5 CONCLUSIONS AND FUTURE WORK

This paper presents a study on recognizing predefined dance motion patterns from skeletal animation data captured by multiple Kinect sensors. As can be seen from the experimental results, our method gave quite promising results providing high recognition accuracies of the three Tsamiko dance moves. In future work we aim to experiment on recognition of different styles of these dance moves and adding more complex dance patterns and variations. In addition we plan to extend our skeleton fusion algorithm on joint rotation data (both absolute and hierarchical) which will allow the construction of posture codebooks based on both position and rotation data.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework 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

Aylward, R., "Sensemble: A Wireless Inertial Sensor System for InteractiveDance and Collective Motion Analysis", Masters of Science in Media Arts and Sciences, Massachusetts Institute of Technology, 2006

Alexiadis, D., Kelly, P., Daras, P., O'Connor, N., Boubekeur, T., and Moussa, M., Evaluating a dancer's performance using kinect-based skeleton tracking. In *Proceedings of the 19th ACM international conference on Multimedia (MM '11)*. ACM, New York, NY, USA, pp. 659-662, 2011.

Li, W., Zhang, Z., Liu, Z., **"**Action Recognition Based on A Bag of 3D Points", *IEEE International Workshop on CVPR for Human Communicative Behavior Analysis* (in conjunction with CVPR2010), San Francisco, CA, June, 2010.

Wang, J., Liu, Z., Wu, Y., and Yuan, J., "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR'12*, 2012.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A., "Real-time human pose recognition in parts from single depth images,", in *CVPR*, pp. 1297 – 1304, June *2011*.

Waithayanon, C. and Aporntewan, C., "A motion classifier for Microsoft Kinect," in *Computer Sciences and Convergence Information Technology (ICCIT), 2011 6th International Conference on*, 2011.

ten Holt, G. A., Reinders, MJT., Hendricks, EA., *Multi-Dimensional Dynamic Time Warping for Gesture Recognition*. Conference Paper. 2007.

Xia, L., Chen, C.-C., and Aggarwal, J., "View invariant human action recognition using histograms of 3D joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, 2012.

Wang, S. Quattoni, A., Morency, L.-P., Demirdjian, D., and Trevor Darrell, Hidden Conditional Random Fields for Gesture Recognition, *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, June 2006

Kinect for Windows | Voice, Movement & Gesture Recognition Technology. 2013. [ONLINE] Available at: http://www.microsoft.com/en-us/kinectforwindows/.

Besl, Paul J.; N.D. McKay (1992)."A Method for Registration of 3-D Shapes". IEEE Trans. on Pattern Analysis and Machine Intelligence (Los Alamitos, CA, USA: IEEE Computer Society) 14 (2): 239–256.

Rusu, B., Cousins, S., "3D is here: Point Cloud Library (PCL)," Robotics and Automation (ICRA), 2011 IEEE International Conference on , vol., no., pp.1,4, 9-13 May 2011

Quattoni, A., Collins, M., Darrell, T., Conditional Random Fields for Object Recognition, In Neural Information Processing Systems, 2004.

The original publication is: Alexandros Kitsikidis, Kosmas Dimitropoulos, Stella Douka and Nikos Grammalidis, " Dance Analysis using Multiple Kinect Sensors", in *VISAPP2014, Lisbon, Portugal, 5-8 January 2014.*