Multi-sensor technology and fuzzy logic for dancer's motion analysis and performance evaluation within a 3D virtual environment

Alexandros Kitsikidis[1], Kosmas Dimitropoulos[1], Erdal Yilmaz[2], Stella Douka[3] and Nikos Grammalidis[1]

[1]Informatics and Telematics Institute, ITI-CERTH, 1st Km Thermi-Panorama Rd, Thessaloniki, Greece {ajinchv, dimitrop, ngramm}@iti.gr

[2]KANAVA Tech, Turkey erdal.yilmaz@kanavatech.com

[3]Department of Physical Education and Sport Science, Aristotle University of Thessaloniki, Greece sdouka@phed.auth.gr

Abstract. In this paper, we describe a novel methodology for dance learning and evaluation using multi-sensor and 3D gaming technology. The learners are captured during dancing, while an avatar visualizes their motion using fused input from multiple sensors. Motion analysis and fuzzy-logic are employed for the evaluation of the learners' performance against the performance of an expert. Specifically, a two level Fuzzy Inference System is proposed which uses as input low level skeletal data and high level motion recognition probabilities for the evaluation of dancer's performance. Tests with real dancers, both learners and experts, dancing Tsamiko, a very popular traditional Greek dance, are presented showing the potential of the proposed method.

Keywords: Kinect, fuzzy inference system, dance performance evaluation, Unity

1 Introduction

As traditional dances are forms of intangible cultural heritage, there is always a risk that certain elements of this culture could die out or disappear if they are not safeguarded and transmitted. ICT technologies can play an important role in their preservation, e.g. in the form of virtual learning systems, assisting in the transmission of dancing knowledge. Such systems employ various sensors to capture the movements of the learner, analyse the movement and provide a feedback, thus facilitating the learning procedure [1]. Automatic performance evaluation in the form of scoring and visual feedback through a 3D virtual environment can significantly improve the competency of the learner.

Detection, classification and evaluation of dance gestures and performances are active topics of research [2], while commercial products also exist, such as the Harmonix' Dance Central video game series [3], where a player tries to imitate the motion demonstrated by an animated character. Many research projects have been conducted on the topic of dance assistance and evaluation employing various sensor technologies. Saltate![4] is a wireless prototype system to support beginners of ballroom dancing. It acquires data from force sensors mounted under the dancers' feet, detects steps, and compares their timing to the timing of beats in the music playing, thus detecting mistakes. Sensable project [5] also employs wireless sensor modules, worn at the wrists of dancers, which capture motions in dance ensembles. The VR-

Theater project [6] allows choreographers to enter the desired choreography moves with a user-friendly user interface, or even to record the movements of a specific performer using motion capture techniques. Also, different kinds of augmented feedback (tactile, video, sound) for learning basic dance choreographies are investigated in [7].

Markerless motion capture based on real-time depth sensing systems have recently emerged with the release of Microsoft Kinect [8] and other similar depth cameras like Asus Xtion [9]. These sensors offer a cost-effective alternative to more expensive inertial and optical motion capture systems. In [10], evaluation of dance performance is conducted against the performance of a professional using skeleton tracking data captured by Kinect sensor, visualized within 3D virtual environments.

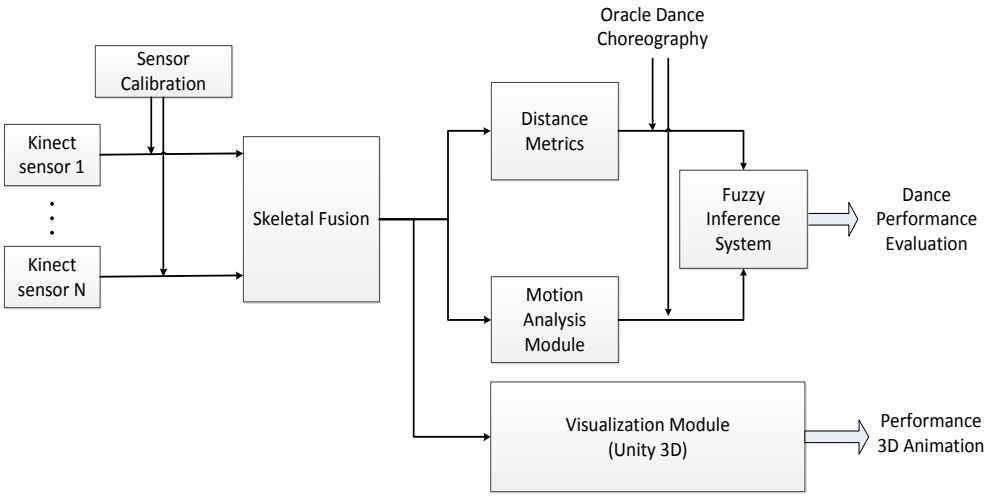
In this paper, we propose a dance evaluation system, which offers two novel features. First, a multi-Kinect acquisition system is used, where synchronized skeletal data from each sensor are fused in order to improve the quality of the final skeletal tracking, based on our previous work [11]. Secondly, we propose a scoring system based on fuzzy inference. The reasons for choosing fuzzy inference are the following: i) the ability to produce realistic and less predictable reactions, ii) the ability to capture a real human knowledge-base and use it extensively with minimal coding and iii) the use of an AI technique that is more suitable to model complex virtual behaviour.

The evaluation of a dancer is performed based on low level and high level features. Low level features are the fused skeletal tracking data and high level features are the motion recognition probabilities are used by the 3D virtual environment for the evaluation of the dancer's performance against an expert's performance and the generation of visual feedback. The 3D environment is based on Unity 3D engine [12],

which is a popular multiple platform gaming and visualization solution among the graphics and gaming community.

2 Methodology

The architecture of the proposed system is illustrated in **Fig. 1**. Specifically, for capturing we use several Kinect sensors placed around the dancer. Captured skeletal data consists of 3D position and rotation data (relative to a reference coordinate system centred at the origin of each sensor) of 20 predefined skeletal joints of the dancer's body, along with the confidence level of tracking of each joint. A skeletal fusion procedure is proposed to combine the data obtained from multiple sensors onto a single fused skeleton (described in section 2.1), which is then used for motion analysis of the dancer. Subsequently, the evaluation of the dancer's performance takes place. The low level features are extracted to calculate the distance metrics (section 2.2). In addition, the motion analysis module performs motion recognition to extract high level features (section 2.3). Those features are subsequently provided to the Fuzzy Inference System, where the final evaluation of the dancer takes place (section 3). Moreover, the visualization module provides a 3D environment for the learner to examine his performance along with the numerical and textual performance grading. The visu-



alization module is implemented in Unity and takes input from the fused skeletal animation data. The resulting animation screenshots are shown in **Fig. 2**.

Fig. 1. System architecture overview

Fig. 2. Visualisation of an expert dancing the tsamiko dance. A 3D avatar wearing a traditional costume is animated in Unity 3D using the fused skeletal animation data acquired during the recording session.

2.1 Skeletal Fusion

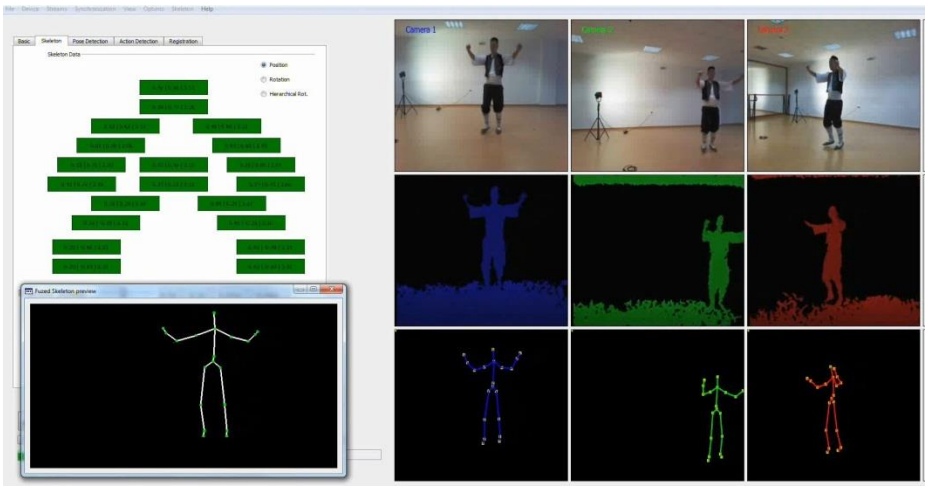
Skeletal fusion is the process of combining skeletal data captured by multiple sensors into a single, more robust skeletal representation. It allows to reduce occlusion and self-occlusion problems and to increase the total area of coverage. Prior to fusion,

sensor calibration procedure must take place in order to estimate the rigid transformation between the coordinate systems of each sensor and the reference sensor. We use Iterative Closest Point algorithm [13] implementation found in the Point Cloud Library (PCL, http://pointclouds.org/) [14] to estimate the rigid transformation (Rotation-Translation) which is subsequently used to register the skeleton captured by each sensor in the reference coordinate system.

The skeletal fusion is performed on registered skeletons, i.e. the representations of each skeleton transformed to the coordinate system of the reference sensor. This is accomplished by multiplying the skeleton joint positions 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 (**Fig. 3**) according to a specific fusion strategy.

The proposed fusion strategy is applied on joint positional data, which can be easily extended on joint rotations as well. Initially, the sum of all joint confidence levels per 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 dancer's real pose.

We consider the joints of this skeleton as a base and construct the fused skeleton joints in the following manner. We examine the confidence values of each joint of the base skeleton. There are three possible values: high, medium and low. If the confidence of the base joint is high, it is left as is for the fused skeleton. 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 joints are present, their average is used as a position value of the fused joint.

Fig. 3. Fused skeleton from three Kinect sensors. Color maps, depth maps and skeleton previews of each sensor along with the resulting fused skeleton are displayed.

As a last step, a stabilization filtering is applied in order to overcome problems due to the 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. The thresholds are different for each joint, since it is expected that some joints (hands and feet) move more rapidly than others. In our experiments of Tsamiko dance these thresholds were set to 40cm for the feet joints and 20cm for the remaining joints.

2.2 Distance Metrics

To evaluate the performance of a dancer, specific metrics should be defined for measuring the motion similarity between a learner and an expert. Taking into account that in Tsamiko dance leg movements constitute the key element of the choreography,

in this paper we propose two metrics for measuring the motion accuracy of the dancer. Specifically, we define the knee-distance D_K and the ankle distance D_A for each frame as (**Fig. 4 A**):

$$D_K = |K_L - K_R| \tag{1}$$

$$D_A = |A_L - A_R| \tag{2}$$

However, both distances heavily depend on the height of the dancer that is their values change from dancer to dancer. To ensure the invariance of the proposed metrics (in terms of dancer's height), a specific normalization process is proposed. More specifically, we calculate the normalized distances by dividing the proposed metrics by the distance of the path connecting the joints. For the normalized knee-distance $\widehat{D}_K$, the path is computed by dividing the distance between the knee joints by the sum of the distances between the *Left Knee*, *Left Hip*, *Root*, *Right Hip* and *Right Knee* joints:

$$\widehat{D}_K = \frac{D_K}{|K_L - H_L| + |H_L - R| + |R - H_R| + |H_R - K_R|} \tag{3}$$

The normalized ankle distance $\widehat{D}_A$ is calculated in a similar manner:

$$\widehat{D}_A = \frac{D_A}{|A_L - K_L| + |K_L - H_L| + |H_L - R| + |R - H_R| + |H_R - K_R| + |K_R - A_R|} \tag{4}$$

The estimation of the above metrics is repeated in each time instant, i.e. frame, resulting in the creation of time series like the ones presented in **Fig. 4 B**. To compare

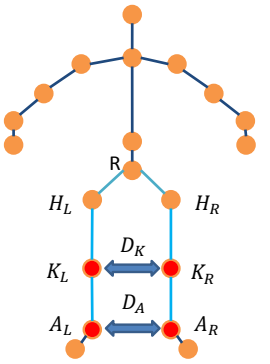
the similarity of these time series we introduce the use of two motion accuracy scores S_K and S_A, which are computed by calculating the maximum correlation coefficient between the testing subject's normalized distances ($\widehat{D}_K$t) and the reference subject's normalized distances ($\widehat{D}_K$r). The maximum correlation coefficient is computed by iteratively shifting the testing signal by one sample at a time, with respect to the reference signal and by computing the maximum correlation coefficient over all these shifts. The shifting step ranges from 1 sample to 250 samples, which is approximately the duration of a single dancing cycle of Tsamiko dance. The correlation coefficient is defined as:

$$R = \frac{\sigma_{x,y}}{\sqrt{\sigma_x * \sigma_y}} \tag{5}$$

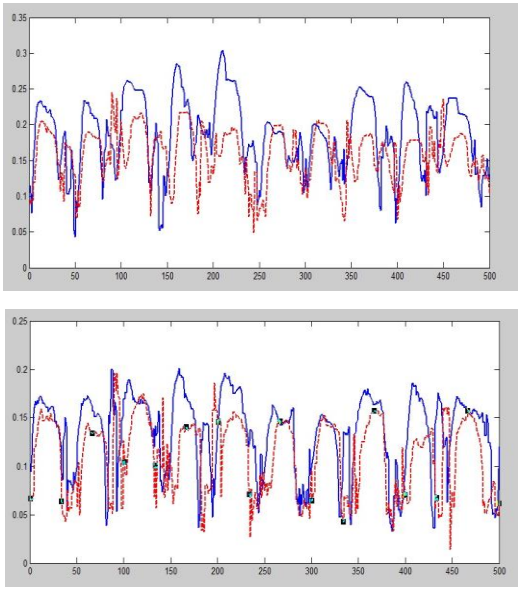
Where:

$$\sigma_{x,y} = E[(x - E[x])(y - E[y])] \tag{6}$$

$$\sigma_x = E[x^2] - {\mu_x}^2 \tag{7}$$



(a)



(b)

$$\sigma_y = E[y^2] - \mu_y{}^2 \quad (8)$$

where $E[\,]$ is the expected value and μ is the mean value.

Fig. 4. a) The knee distance and the ankle distance metric b) Time series of $\widehat{D}_K$ and $\widehat{D}_A$ of two dancers.

2.3 Motion Analysis

The motion analysis subsystem (**Fig. 5**) performs the detection of the basic motion patterns, in our case, the three dance movements of the Tsamiko dance. The correct choreography of this dance consists of sequential repetition of these three moves. Thus, we derive a choreography score S_{Ch} which is the precision of the correct detection of those motion patterns by the motion analysis subsystem.

A pre-processing step for motion analysis is a view-invariance transform of the skeleton, by translating each joint position relative to the root joint and subsequently rotating the skeleton around the y axis so that it is facing towards the positive z direction. Next, the skeleton is split into five parts: *torso*, *left hand*, *right hand*, *left foot* and *right foot*, each consisting from a root and children joints. For each skeleton part we generate a feature vector consisting of positions of joints relative to the root joint of the part. In fact, those feature vectors constitute a representation of a dancer's posture. For each skeleton part, a codebook of k basic postures is defined using k-means clustering in a large set of postures obtained from recorded training sequences. A multiclass SVM classifier is used to classify each incoming feature vector as a specific posture from this posture codebook. Thus, each motion sequence is transformed to a sequence of symbols of this codebook, one sequence per body part. Those sequences

are fed to the final stage of the motion analysis subsystem, which consist of a Hidden-state Conditional Random Fields classifier (HCRF).

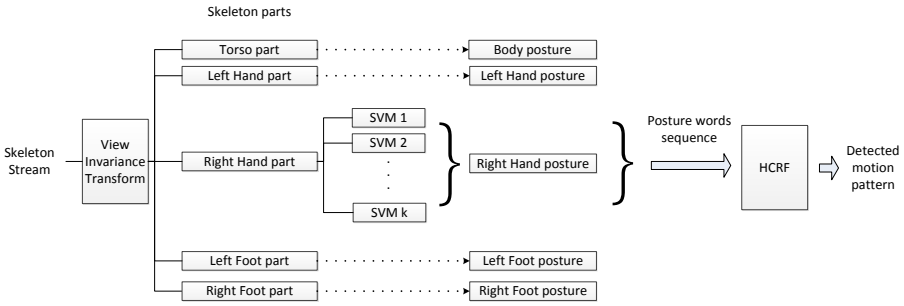


Fig. 5. Motion analysis module

HCRFs [15][16] are a class of statistical modelling method (discriminative undirected probabilistic graphical model) often applied to pattern recognition problems and machine learning in general. HCRFs are a generalization of / is alternative to Hidden Markov Models and are popular in natural language processing, object recognition and motion recognition tasks. We use multi-class HCRF model trained on a set of M basic motion patterns (the three dance moves of Tsamiko). For the training of the HCRFs we use labelled sequences described in the previous paragraph. For the detection phase HCRFs classifier provides a probability of the model of the HCRF fitting the observed sequence, thus it is labelled accordingly.

3 Fuzzy Inference System

For the evaluation of the dancer's performance against an expert's performance a two level Fuzzy Inference System (FIS) was designed. FIS is a way of mapping an

input space to an output space using a collection of fuzzy membership functions and rules i.e. linguistic statements in the form of *if….then* that describe how the FIS should make a decision. The proposed FIS system is based on Mamdani method [17], which is widely accepted for capturing expert's knowledge and allows the description of the domain knowledge in a more intuitive, human like manner.

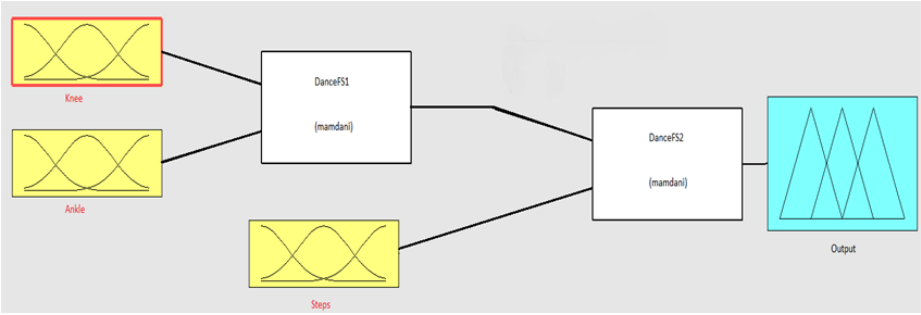


Fig. 6. The structure of the two-level fuzzy inference system

Low level features obtained from raw skeletal tracking data and high level motion recognition probabilities are used as input to the two-level FIS for the evaluation of the dancer's performance. The proposed FIS architecture is illustrated in **Fig. 6**. The estimated maximum correlation coefficients between the normalized joint distances (knee distance and ankle distance) S_K and S_A of the expert and the learner dancer are fed as input to the first FIS to generate the motion accuracy index. While this index contains meaningful information about the motion of the dancer, little information is provided regarding the proper execution of the choreography e.g. it is difficult to dis-

criminate whether the dancer cannot follow the choreography or he/she cannot be synchronized with the music. To address this issue, besides the output of the first FIS i.e. the motion accuracy index, the percentage of the correct identified motion patterns of the learner dancer (the choreography score S_{Ch} provided by the motion analysis module) is also fed as input to the second FIS. The final output is converted into human understandable messages (defuzzification), such as low, medium and high performance score.

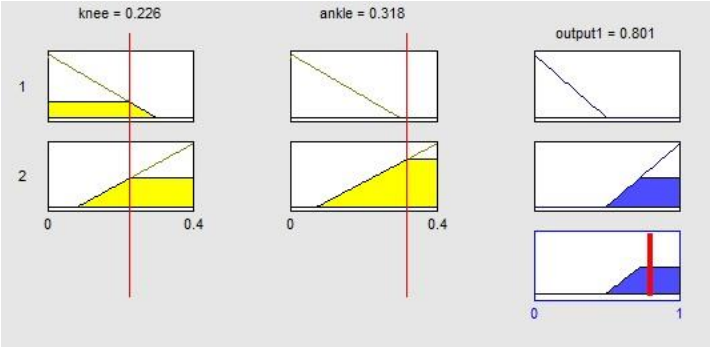


Fig. 7. The output of the first FIS in the case of high motion accuracy scores leading to high motion accuracy index.

The set of rules used for building the fuzzy inference of the first FIS are described below together with the example of an output **Fig. 7**:

- If S_K is High and S_A is High then 'motion accuracy index' is High
- If S_K is Low and S_A is Low then 'motion accuracy index' is Low

Similarly, the set of rules used for building the fuzzy inference of the second FIS are described below:

- If ‘motion accuracy index’ is High and ‘choreography score’ is High then ‘dancing performance’ is High
- If ‘motion accuracy index’ is Low and ‘choreography score’ is Low then ‘dancing performance’ is Low
- If ‘motion accuracy index’ is Low and ‘choreography score’ is High then ‘dancing performance’ is Medium

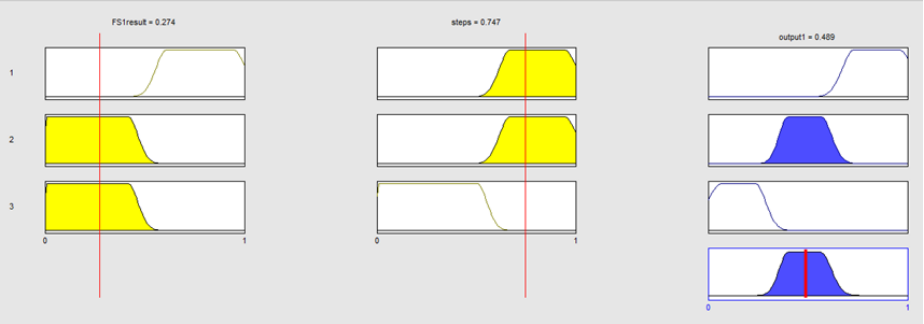


Fig. 8. The output of the second FIS in the case of low motion accuracy index and high choreography accuracy.

Fig. 8 illustrates an example of low motion accuracy index and high choreography accuracy. In this case, while the dancer can follow the choreography i.e. the expected motion patterns are identified accurately, he/she cannot be synchronized with the music and, therefore, a low motion accuracy index is produced. Since, the above case is satisfied by the third fuzzy rule, the performance of the dancer is considered as medium.

4 Experimental Results

For the evaluation of our methodology we recorded a performance of an expert along with performances of experienced dancers and learners of the Greek traditional Tsamiko dance. Tsamiko is a popular traditional folk dance which follows a strict and slow tempo. The steps are relatively easy to execute but must be precise and strictly on beat. We captured the movements of the dancers using a setup consisting of three Kinect sensors, placed in front of a dancer. The reference sensor was placed directly in front, and the other two at the sides, creating an arc topology (**Fig. 9**). This setup allowed for the dancers to have a freedom of movement of about 2,5 meters along a straight line.

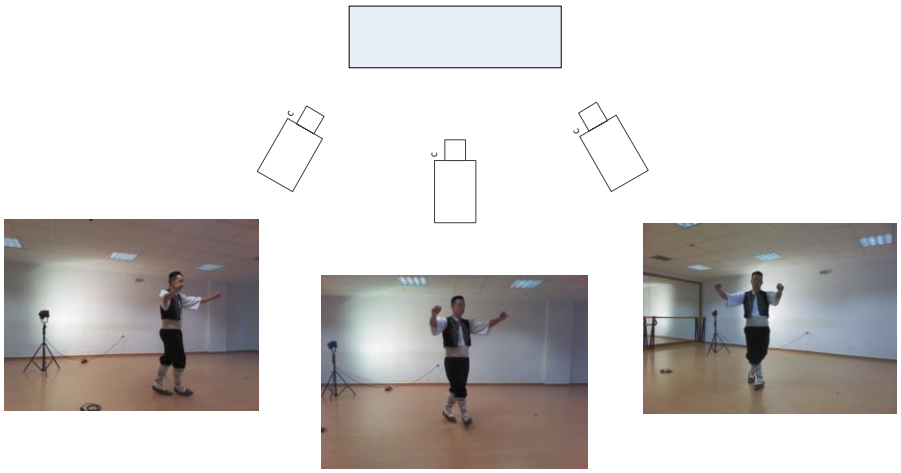


Fig. 9. Sensor setup with three Kinect devices. The recording of an expert dancing Tsamiko as captured by each sensor.

Table 1. The evaluation scores of the dancers together with the outputs of the two fuzzy systems. Learners marked with asterisk were interrupted during the dance because they made mistakes.

	Ankle distance score	Knee distance score	1st fuzzy Motion Accuracy index	Choreography precision score	2nd fuzzy Dancing Performance index	Performance Evaluation Low: <0.4 Medium:0.4-0.6 High: >0.6
Experienced dancer 1	0,28	0,26	**0.7244**	**100%**	**0.8500**	**High**
Experienced dancer 2	0,33	0,33	**0.8270**	**100%**	**0.8608**	**High**
Learner 1	0,19	0,29	**0.7243**	65%	0.5000	Medium

Learner 2	0,22	0,13	0.4221	68.42%	0.3411	Low
Learner 3 *	0,13	0,11	0.2821	70.58%	0.3146	Low
Learner 4	0,18	0,13	0.3788	**77.78%**	0.500	Medium
Learner 5	0,14	0,08	0.1969	40%	0.1414	Low
Learner 6	0,13	0,13	0.3358	55.56%	0.1575	Low
Learner 7	0,17	0,15	0.4044	66.67%	0.2319	Low
Learner 8 *	0,17	0,15	0.4031	66.67%	0.2258	Low

There were two recording sessions performed using the same setup. During the first session the expert and two experienced dancers were captured, and during the second session eight students of different level of experience participated in the experiments. Each person was recorded for the duration of a single dance (approx. 4 minutes). The performance of all dancers was compared against that of an expert, who is considered as a reference. First, the motion accuracy scores S_K and S_A were computed by comparison to the expert. Then, the recorded sequences were manually annotated and fed to the motion classifier, which detected the three motion patterns of Tsamiko dance. The choreography score S_{Ch} was then computed as the percentage of the motion patterns that are identified correctly, i.e. the precision of the recognition. Those were provided as input to the FIS to obtain the final performance evaluation score.

The results obtained are illustrated in Table 1. As expected, both experts received high score in their performance evaluation, since they had high motion accuracy scores and also perfect choreography precision. The learners, on the other hand received medium and low scores. Learner 1 had high motion accuracy but relatively

low choreography precision, while for student 4 the opposite is true. They both were graded as medium by the system. The rest of the learners received low rating, with varying performance indexes (they were not equally bad), since they had both motion accuracy and choreography scores low.

5 Conclusions and future work

This paper presents a methodology for automatic dance evaluation, intended to be used in dance learning systems. The learners are captured during dancing, while an 3-D avatar is used to visualize their motion. The main contributions are the use of a multi-Kinect motion capture and the definition of a new scoring system based on fuzzy inference. Based on the obtained experimental results, the system seems to properly distinguish between the varying levels of dance expertise, so it is suitable to be used as a tool to assess the learners dancing performance. In the future, a feedback can be provided to the user, based on this performance evaluation, together with a visualization of both user and expert performances, which could significantly assist dance learning.

6 Acknowledgement

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

1. Essid, S., Alexiadis, D., Tournemenne, R., Gowing, M., Kelly, P., Monaghan, D., Daras, P., Drumeau, A., O'connor N., An Advanced Virtual Dance Performance Evaluator, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), Kyoto, Japan, 25-30 March 2012
2. Raptis, M., Kirovski, D., Hoppe, H., Real-time classification of dance gestures from skeleton animation, Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, August 05-07, 2011, Vancouver, British Columbia, Canada
3. Dance central http://www.dancecentral.com/
4. Drobny, D., Weiss, M. and Borchers, J., Saltate! - A Sensor-Based System to Support Dance Beginners. In CHI '09: Extended Abstracts on Human Factors in Computing Systems, pages 3943-3948, New York, NY, USA, 2009. ACM.
5. 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
6. VR-Theater project http://avrlab.iti.gr/HTML/Projects/current/VRTHEATER.htm
7. Drobny D. and Borchers, J., Learning Basic Dance Choreographies with different Augmented Feedback Modalities. In CHI '10: Extended Abstracts on Human Factors in Computing Systems, New York, NY, USA, 2010. ACM Press
8. Kinect for Windows | Voice, Movement & Gesture Recognition Technology. 2013. [ONLINE] Available at: http://www.microsoft.com/en-us/kinectforwindows/.
9. Asus Xtion PRO http://www.asus.com/Multimedia/Xtion_PRO/
10. 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.

11. Kitsikidis, A., Dimitropoulos, K., Douka, S., Grammalidis, N., "Dance Analysis using Multiple Kinect Sensors", VISAPP2014, Lisbon, Portugal, 5-8 January 2014
12. Unity. http://unity3d.com.
13. Besl, P., McKay, N., 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, 1992
14. 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
15. Wang, S. Quattoni, A., Morency, L.-P., Demirdjian, D., and Darrell, T., Hidden Conditional Random Fields for Gesture Recognition, *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, June 2006
16. Quattoni, A., Collins, M., Darrell, T., Conditional Random Fields for Object Recognition, In Neural Information Processing Systems, 2004.
17. Mamdani, E.H. and Assilian, S., "An experiment in linguistic synthesis with a fuzzy logic controller," International Journal of Man-Machine Studies, Vol. 7, No. 1, pp. 1-13, 1975.

The original publication is: A. Kitsikidis, K. Dimitropoulos, E. Yilmaz, S. Douka, N. Grammalidis, "Multi-sensor technology and fuzzy logic for dancer's motion analysis and performance evaluation within a 3D virtual environment", in *Universal Access in Human-Computer Interaction. Design and Development Methods for Universal Access (8th International Conference, UAHCI 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part I), Lecture Notes in Computer Science*, Volume 8513, 2014, pp 379-390»

and is available from: http://link.springer.com/chapter/10.1007/978-3-319-07437-5_36