Interactive Dance Performance Evaluation using Timing and Accuracy Similarity

Yeonho Kim
Department of Computer Science and Engineering
Pohang University of Science and Technology
beast@postech.ac.kr

Daijin Kim
Department of Computer Science and Engineering
Pohang University of Science and Technology
dkim@postech.ac.kr

ABSTRACT

This paper presents a dance performance evaluation how well a learner mimics the teacher's dance as follows. We estimate the human skeletons, then extract dance features such as torso and first and second-degree feature, and compute the similarity score between the teacher and the learner dance sequence in terms of timing and pose accuracies. To validate the proposed dance evaluation method, we conducted several experiments on a large K-Pop dance database. The proposed methods achieved 98% concordance with experts' evaluation on dance performance.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*;

KEYWORDS

pose estimation, dance feature, similarity score

ACM Reference Format:
Yeonho Kim and Daijin Kim. 2018. Interactive Dance Performance Evaluation using Timing and Accuracy Similarity. In *Proceedings of SIGGRAPH '18 Posters*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3230744.3230798

1 INTRODUCTION

Since pose estimation methods have become reliable, many real-world applications are now realistic. One typical application is to recognize human actions using the spatial differences between detected joints [Yang and Tian 2014] [Xia et al. 2012], a hidden Markov model (HMM) [Sung et al. 2011], and dynamic time warping (DTW) [Reyes et al. 2011]. Most applications have focused on classifying human activities, but few methods have been introduced to evaluate the precision of human actions, such as dance gestures [Raptis et al. 2011], music-conducting gestures [Schramm et al. 2015], or performance-evaluation in health care [Ofli et al. 2016].

We propose a dance performance evaluation using an accurate human pose estimation. We extract the dance feature to measure the similarity of timing and pose accuracies between dance performances of teacher and learner. To the best of our knowledge, this is the first time to evaluate the qualitative and quantitative aspect of human dance performance by using a large K-Pop dance database.

SIGGRAPH '18 Posters, August 12-16, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5817-0/18/08.
https://doi.org/10.1145/3230744.3230798

2 DANCE PERFORMANCE EVALUATION

To evaluate dance performance, we propose a dance teacher program (Fig. 1) that can help people to learn dances from examples in five steps: (1) It shows the dance sequence contained in teacher's dance database, where the human joints are obtained by a commercial motion capture system. (2) The human pose estimator extracts the joint positions of the learner [Kim and Kim 2015]. (3) The dance feature generator makes the dance features from the joint positions of learner and teacher. (4) It performs dynamic time warping of the learner's dance features to the teacher's dance features. (5) It evaluates the learner's dance performance by matching the learner's dance features to the teacher's dance features, where bright (or dark) skeletons denote higher (or lower) matching.

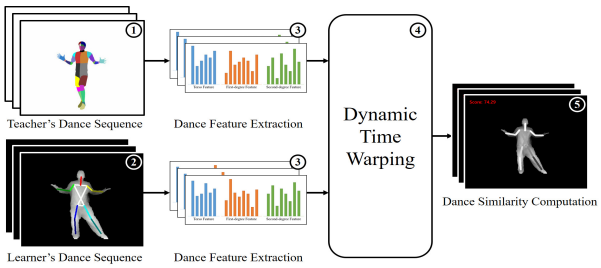


Figure 1: Overall process of the dance teacher.

3 DANCE FEATURE EXTRACTION

Human skeletons are commonly represented by 15 joint positions with the 3D coordinate system. Because of the variation in camera position and orientation, or in human body shape and size, the traditional representation is not suitable to compare dance performances of teacher and learner. We expand a previous method [Raptis et al. 2011] to design a new 22-dimensional feature: it is composed of a six-dimensional torso feature, an eight-dimensional first-degree feature, and an eight-dimensional second-degree feature.

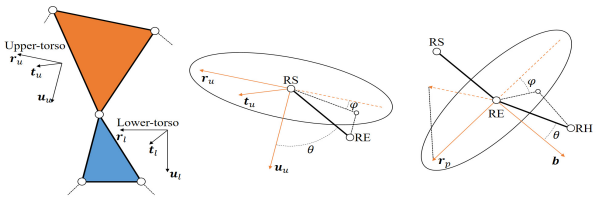
3.1 Torso Feature

In [Raptis et al. 2011], the human torso is represented as just one component, and the torso angles are encoded with respect to the world coordinate. However, K-Pop dance includes very complex poses, so we design a six-dimensional torso feature of an upper-torso joints and a lower-torso joints. The upper-torso joints form a plane that contains the torso center and the left/right shoulders; the lower-torso joints form another plane that contains the torso center and the left/right hips. We obtain first three-dimensional angles by comparing the three-dimensional upper and lower base

axes, and obtain last three-dimensional angles by comparing the upper base axes of the previous and current frames (Fig. 2(a)). The proposed torso feature can represent the bend, twist, and lean and full-body rotation.

3.2 First- and Second-degree Feature

The first- and second- degree feature has eight-dimensional angles that represent the movement of the upper limbs (elbows, knees), and the lower limbs (hands, feet), respectively. Each joint has an inclination and an azimuth with respect to the adjacent parent joint [Raptis et al. 2011] (Fig. 2(b) and (c)).



(a) Torso Feature (b) First-degree Feature (c) Second-degree Feature

Figure 2: Proposed dance features.

4 DANCE SIMILARITY

The timing accuracy is measured by comparing the middle time of matched dance sequences. The dance teacher program (1) obtains the learner's dance sequence by accumulating it for a given period (2 s in this work), (2) finds the corresponding teacher's dance sequence by dynamic time warping. The timing accuracy is computed as

$$S_t = 1 - \min\left(1, \exp\left(\frac{|T_t - T_l| - \tau}{\alpha\tau}\right)\right), \tag{1}$$

where $T_t = \frac{T_t^s + T_t^e}{2}$ and $T_l = \frac{T_l^s + T_l^e}{2}$ are the middle time of the teacher's and learner's dance sequence, respectively. The superscripts s and e denote the start and end point of dance sequence, respectively. α is a parameter that governs the slope of Eq. (1) and τ is a parameter that controls maximum deviation of timing.

The pose accuracy is computed as

$$S_p = \frac{1}{T_t^e - T_t^s + 1} \sum_{i=T_t^s}^{T_t^e} \exp\left(-\frac{\left\|\mathbf{f}_l^i - \mathbf{f}_t^i\right\|}{\beta}\right), \tag{2}$$

where $\mathbf{f}_l^i$ and $\mathbf{f}_t^i$ denote the dance features of the learner and teacher respectively at the ith frame, and β is a parameter that controls the amount of deviation from the teacher's dance.

The dance similarity between learner and teacher is defined by the sum of partial scores of timing and pose accuracies as

$$S = \frac{1}{N} \sum_{j=1}^{N} S_t^j S_p^j, \tag{3}$$

where N is the number of sequences of the learner's dance performance and S_t^j and S_p^j denote the timing and pose accuracies at the jth sequence, respectively.

5 EXPERIMENTAL RESULTS

To validate the dance performance evaluation, we constructed a large dance dataset that consists of 100 popular K-Pop dances. For each dance, we used a Microsoft Kinect 2 to capture the learners' dance performances. Learner's dance sequences were labeled subjectively by a group of dance experts as *best*, *good*, *bad*, and *worst*. Evaluation scores of *best*, *good*, *bad*, and *worst* learners are denoted by S_1, S_2, S_3, and S_4, respectively. We consider that the dance performance evaluation is correct if and only if $S_1 > S_2 > S_3 > S_4$. Table 1 shows the evaluation scores of five dance sequences (totally 100 sequences) using (a) a conventional feature [Raptis et al. 2011], (b) the proposed dance feature with the hierarchical human pose estimation (HPE) method [Kim and Kim 2015], and (c) the proposed feature with the CNN-based HPE method [Huang and Altamar 2016], where the bold scores disagree with expert evaluation.

Table 1: Comparison of evaluation scores.

Seq \ Feature	(a)				(b)				(c)			
	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4
2	**58.9**	**60.1**	53.3	51.5	62.7	62.5	54.2	51.4	58.2	57.8	49.5	46.8
15	59.2	53.1	**48.4**	**50.4**	62.6	55.2	**49.3**	**50.3**	**40.5**	**37.9**	**35.9**	**48.0**
23	63.4	**55.4**	**56.6**	52.5	66.5	57.4	57.3	52.3	58.7	51.5	**45.6**	**46.2**
74	64.4	**54.6**	**55.8**	54.3	68.1	**56.8**	**56.9**	54.1	60.8	54.9	51.5	45.8
89	**41.2**	**39.9**	**38.5**	**53.3**	**43.4**	**41.0**	**39.0**	**52.9**	62.0	57.2	51.1	48.3
Total	86%				97%				98%			

We found that the scores from the proposed dance feature with the CNN-based HPE method agreed with 98 times with the dance experts'; the exceptions were the 15th and 23rd sequences, whereas [Raptis et al. 2011] agreed 86 times and the proposed dance feature with the hierarchical HPE method agreed 97 times.

6 CONCLUSION

We proposed a dance performance evaluation using human pose estimation. The proposed dance teacher helps to learn dances and to evaluate timing and pose. The main contributions are (1) We constructed a large K-Pop dance database, which contains 100 experts' sequences and 400 learners' sequences. (2) The proposed dance teacher program achieved 98% agreement with experts' evaluation.

7 ACKNOWLEDGMENT

This research was partially supported by the MSIT (Ministry of Science, ICT), Korea, under either the SW Starlab support program (IITP-2017-0-00897) or the development of predictive visual intelligence technology (IITP-2014-0-00059).

REFERENCES

J. Huang and D. Altamar. 2016. Pose Estimation on Depth Images with Convolutional Neural Network. (2016).
Y. Kim and D. Kim. 2015. Efficient body part tracking using ridge data and data pruning. In *15th International Conference on Humanoid Robots*. IEEE, 114–120.
F. Ofli, G. Kurillo, Š. Obdržálek, R. Bajcsy, H. Jimison, and M. Pavel. 2016. Design and evaluation of an interactive exercise coaching system for older adults: lessons learned. *Journal of biomedical and health informatics* 20, 1 (2016), 201–212.
M. Raptis, D. Kirovski, and H. Hoppe. 2011. Real-time classification of dance gestures from skeleton animation. In *SIGGRAPH/Eurographics symposium on computer*.
M. Reyes, G. Dominguez, and S. Escalera. 2011. Featureweighting in dynamic time-warping for gesture recognition in depth data. In *ICCVW*. IEEE, 1182–1188.
R. Schramm, C. Jung, and E. Miranda. 2015. Dynamic time warping for music conducting gestures evaluation. *Transactions on Multimedia* 17, 2 (2015), 243–255.
J. Sung, C. Ponce, B. Selman, and A. Saxena. 2011. Human Activity Detection from RGBD Images. *plan, activity, and intent recognition* 64 (2011).
L. Xia, C. Chen, and J. Aggarwal. 2012. View invariant human action recognition using histograms of 3d joints. In *Computer vision and pattern recognition workshops*.
X. Yang and Y. Tian. 2014. Effective 3d action recognition using eigenjoints. *Journal of Visual Communication and Image Representation* 25, 1 (2014), 2–11.