Action Tutor: Real-Time Exemplar-based Sequential Movement Assessment with Kinect Sensor

Chi-Wen Chen[a], Min-Chun Hu[b,c], Wen-Huang Cheng[b],
Che-Han Chang[a], Jui-Hsin Lai[a], and Ja-Ling Wu[a]
[a]Communications and Multimedia Lab, National Taiwan University, Taiwan
[b]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[c]Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan
{euro, trimy, wisley, frank, larrylai, wjl }@cmlab.csie.ntu.edu.tw

ABSTRACT

With the aid of depth camera, such as Microsoft Kinect, the difficulty of vision-based posture estimation is greatly decreased, and human action analysis has achieved a wide range of applications. However, there is still much to do to develop effective movement assessment technique, which bridges the results of human posture estimation and the understanding of human action performance. In this work, we propose an action tutor system which enables the user to interactively retrieve the learning exemplar of the target action movement and to immediately acquire motion instructions while learning it in front of the Kinect. In the retrieval stage, non-linear time warping algorithms are designed to retrieve video segments similar to the query movement roughly performed by the user. In the learning stage, the user learns according to the selected video exemplar, and the motion assessment including both static and dynamic differences is presented to the user in a more effective and organized way, helping him/her to perform the action movement correctly.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Evaluation/methodology*

Keywords

Human action video retrieval, movement assessment

1. INTRODUCTION

For the beginner who wants to learn new dance moves or the patient who needs to do rehabilitation exercises everyday, a motion instruction system would be very useful since it is not possible to hire a tutor or a professional instructor to teach the detailed movements at any time. Moreover, compared to learning with a human tutor/instructor, learning in front of the computing devices would make the user feel less embarrassed especially when he/she performs wrong movements. In recent years, dancing action learning games developed based on Kinect are getting popular, e.g. "LET'S DANCE with Mel B" [1]. These games define several key postures for each action, and the correctness score increases as long as the user performs postures similar to the defined ones in limited time. However, these games are not practical enough in our case due to the following reasons: 1) The temporal correlation among postures is seldom considered while calculating the score and no detailed motion instruction for each human part is given to the user. 2) Sometimes the user is forced to skip a part of movements to follow the speed of the learning exemplar, but skipping movements consequently lowers the learning effectiveness.

MM'12, October 29–November 2, 2012, Nara, Japan.
ACM 978-1-4503-1089-5/12/10.

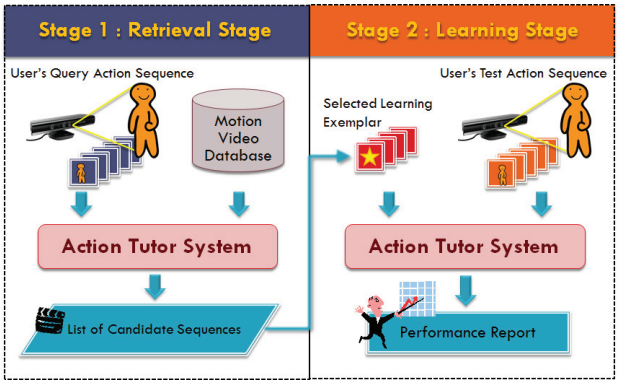

Figure 1: The scenario of the proposed system.

In this work, we develop a real-time action tutor system to analyze the user's action captured by Kinect, and help the user to learn/practice exercises without interacting with a human expert. Our system not only lets the user learn with a given exemplar video, but also facilitates him/her to find the target learning exemplar in a large video database. The user first performs a query action resembling the target movements in his/her mind, and the system recommends a list of similar action videos (also captured by Kinect) in the database by the technique of action video retrieval. The user then selects one video from the list as the learning exemplar and follows it to practice. In the mean time, the system gives detailed movement assessment so that the user can perform exercise movements by himself/herself as if accompanied by a private instructor. For more details, please see the demo video at http://www.youtube.com/ActionTutor.

2. SYSTEM FRAMEWORK

The scenario of proposed system includes two stages as illustrated in Figure 1. The first stage is the *retrieval stage*. In most of the existing motion assessment or instruction systems, before starting learning a specific motion, the user

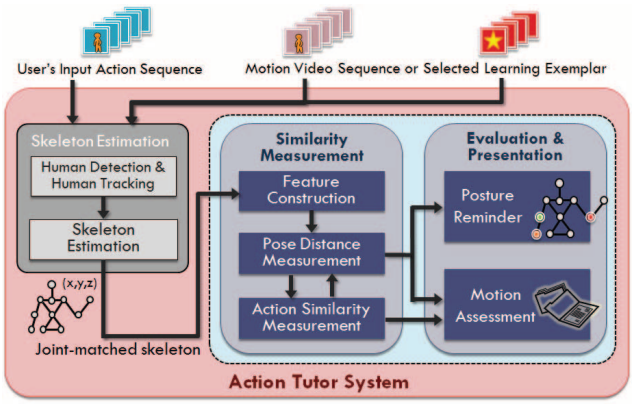

Figure 2: Technical modules of the proposed action tutor system.

has to manually search the entire video database for the target movement. As the database grows abundantly, this search task causes the user much burden. In contrast, using our system, the user could retrieve the movement they want to practice by simply performing it. A target movement is roughly performed by the user and captured by Kinect or other 3D sensors, resulting in a query action sequence. The system then takes this query sequence to search for similar video clips in the motion video database (each element in the database is also a human action video captured by Kinect) and returns a ranked list consisting of segmented candidate sequences for the user to select the exact learning exemplar. The second stage is the learning stage. After choosing the exact learning exemplar, the user could follow and imitate it with real-time feedback pointing out body joints which are not posed correctly. A detailed performance report will be presented to the user in the end of the learning stage.

Figure 2 introduces the proposed action tutor system in terms of the three technical modules, namely skeleton estimation, similarity measurement, and evaluation and presentation modules. The system takes user's action sequence captured by Kinect as system input of the skeleton estimation module, which directly adopts the joint-position prediction algorithm proposed in the OpenNI framework [2, 3] and obtains a sequence of joint-matched skeletons. The similarity measurement module then extracts representative features from the joint-matched skeleton sequence to measure the pose distance and action similarity between the user's query and each sequence in the motion video database (at the retrieval stage) or between the test action and the selected learning exemplar (at the learning stage). Dynamic time warping (DTW) and approximate string matching (ASM) are applied in the similarity measurement process. In the evaluation and presentation module, the overall posture difference and the action similarity at each time instance are presented, where the action similarity at the time instance i is calculated by applying DTW and ASM to the video sequences composed of the corresponding past N frames. When the action similarity at the time instance i is smaller than TH_{sim}, the system will automatically stop playing the learning exemplar and play back from the $(i-N)^{th}$ frame. The playing back mechanism prevents the user from skipping learning important movements while trying to follow the learning exemplar.

We summarize our technical contributions as follows: 1) We design a discriminative pose descriptor which can represent the body-joint configuration invariant to static transformations (i.e. viewpoint transformations and anthropometric transformations) and reflect the difference between various postures. 2) A novel way to retrieve and learn target exercise movements is proposed with the aid of Kinect, and the similarity measurement method can deal with the temporal transformation by non-linear time warping approaches, i.e. DTW and ASM, which map the time alignment problem between two multi-variant action sequences to the substring finding problem. In most of the existing systems [4], a pre-segmented step is required to indicate the exact start/end time positions of the actions before calculating pairwise action similarity. By contrast, our system removes this constraint through applying the substring finding technique.

3. EXPERIMENTAL RESULTS

We conducted objective and subjective tests to evaluate the performance of the proposed action tutor system with the action movements of the Chinese Qigong exercise promoted by Meimen Qigong Culture Center. The accuracy of the video retrieval performance is evaluated by the mean average precision (MAP) and our system has an overall retrieval performance of MAP=0.71. Fourteen users were invited to join the user study, and all participants were requested to complete the entire process of our system from the retrieval stage to the learning stage and then answered a questionnaire aiming to evaluate the proposed system in terms of three aspects, i.e. effectiveness, efficiency, and acceptance. The results show that our system is effective for learning exercise movements and is efficient for finding target movements. Moreover, most users are willing to use the proposed system as a long-term tutor because it is interesting and will not bring embarrassment while learning movements compared to learning with a human instructor.

4. CONCLUSIONS

We propose an action tutor system which achieves high-level evaluation of human action movements with the aid of Kinect. The system is operated in two stages: At the retrieval stage, the user can search the video database for the target action movements by different action matching methods. A list of video candidates are returned to the user for choosing the learning exemplar. At the learning stage, the user follows the movements in the learning exemplar, and the system evaluates the detailed pose difference and the accumulated action similarity between the user and the exemplar in real-time. The experiments reveal that the proposed system is effective, efficient, and acceptable to be used for learning exercise movements.

5. REFERENCES

[1] Black Bean Games. *LET'S DANCE with Mel B*, June 2011. http://www.letsdancewithmelb.com/EN-gb/.

[2] OpenNI organization. *OpenNI User Guide*, November 2010. http://www.openni.org/documentation.

[3] PrimeSense Inc. *Prime Sensor ^{TM} NITE 1.3 Algorithms notes*, 2010. http://www.primesense.com.

[4] M. Raptis, D. Kirovski, and H. Hoppe. Real-time classification of dance gestures from skeleton animation. In *ACM SCA'11*, pages 147–156, 2011.