A Posture Evaluation System for Fitness Videos based on Recurrent Neural Network

An-Lun Liu
National Cheng Kung University
Tainan, Taiwan
liuallen871219s@gmail.com

Wei-Ta Chu
National Cheng Kung University
Tainan, Taiwan
wtchu@gs.ncku.edu.tw

***Abstract*—We present a posture evaluation system especially for fitness. Given a fitness video where a user repetitively performs a movement for fitness, we detect human posture at each video frame, and then characterize the evolution of posture in consecutive frames by a recurrent neural network (RNN). This RNN examines this movement and outputs the degree of goodness (badness). This examination is important for users because prompt inspection of bad movement avoids injury and improves effectiveness of fitness. We demonstrate that the proposed system can accurately evaluate goodness for two types of movements, i.e., Dumbbell Lateral Raise and Biceps Curl. We believe this work is one of the very few studies of using deep neural networks for fitness evaluation.**

***Index Terms*—Thermal face images, facial landmark detection, emotion recognition, multi-task learning**

I. INTRODUCTION

Because of the COVID-19 pandemic, many people are prohibited to go out for exercise. Exercising or working out at home thus becomes popular and important. For working out or fitness, people used to go to a gymnasium and got advice from a fitness coach, in order to make fitness more effective and avoid injury. Therefore, working out at home without professional advice but with the risk of injury becomes a problem. A real-time fitness posture evaluation system is thus demanded.

In this work, we propose a posture evaluation system specially for fitness videos. Two common fitness movements are studied, i.e., Dumbbell Lateral Raise and Biceps Curl. We first detect the 2D human skeleton, and then model the evolution of joints by a recurrent neural network. Based on the domain knowledge from the fitness coach, we design indicators to show if a movement is well performed. For example, for a biceps curl, the angle between the upper arm and the forearm should meet some criterion when the dumbbell is lifted to the highest point. Given a series of 2D skeletons in a move, the proposed recurrent neural network outputs estimated values of the predefined indicators, showing how likely the subject badly or well performs a biceps curl.

Overall, the contributions of this work include:

- This is one of the very earliest works specifically on posture evaluation for fitness videos.
- We propose a recurrent neural network that takes the evolution of 2D human skeletons as the input, and outputs the indicators showing whether a move is performed well or not.

The rest of this paper is organized as follows. Sec. II briefly describes related works. Sec. III provides details of the proposed method. Experimental results are given in Sec. IV, followed by concluding remarks in Sec. V.

II. RELATED WORKS

Human pose estimation and human behavior analysis have been studied for years. They were widely adopted in various video domains, such as movies, news, and sports videos. However, as a special case of sports videos, pose analysis for fitness or personal training just emerged in recent years. Wang et al. [1] proposed a personalized athletic training assistance system, especially for freestyle ski. They proposed a spatio-temporal relation module to detect human pose. The angles between body parts were then calculated, and a classifier based on support vector machines was developed to classify a pose as good or bad. Khurana et al. [2] developed a GymCam system that detects and tracks multiple subjects working out in a gym based on motion information. They detect exercise trajectories, and cluster moving points for each exercise. On the basis of repetitive motion, this system counts how many times an exercise is performed, and also recognizes what the exercise is, such as bicep curl, pushup, and benchpress. Similarly, Alatiah and Chen [3] also worked on recognizing exercises and counting the number of repetitions. Exercise recognition in [2] and [3] was conducted based on videos, while the work in [4] was based on the signals captured by smart watches. Tharatipyakul et al. [5] developed a web-based application that overlays the detected 2D skeleton of a subject on screen, so that the subject can realize how his/her pose is different from the teacher. In this way, the subject immediately gets visual feedback from his/her own pose. Similarly, Xie et al. [6] developed a visual feedback system for users to maintain correct posture. They instead reconstructed 3D human shape from estimated 3D human pose.

Li et al. [7] proposed a spatio-temporal encoding method to represent a sequence of 2D human skeletons, and achieved fitness action recognition and action matching. Different from action recognition in [7], we work on evaluating which part of a movement is performed badly. We relatively target at fine details of a fitness movement.

Recently, the work most similar to ours may be the pose trainer proposed by Chen and Yang [8]. They detected 2D human pose by the OpenPose library, and then proposed two methods to evaluate whether a fitness movement is performed well. One is a heuristic-based approach checking the angles between the upper arm and the torso. Another is a method based on dynamic time warping to measure the similarity between a query move and a template move. By adopting the domain knowledge of fitness, their methods are designed to evaluate whether a move is good or not. However, in our work, not only the global goodness, we also want to clearly indicate where the incorrect posture happens.

III. PROPOSED METHOD

A. Data Collection

We collect the evaluation fitness videos by asking a user to perform the assigned moves. This user is semi-professional and has good knowledge about what a good/bad posture is. He is asked to alternately perform good moves and bad moves when recording. When recording, the shot angle is selected so that the joints of the whole body are clearly visible. This design makes performance of the 2D pose estimator reliable. In our experiments, the resolution of the video is at least 720P, and the frame rate is 30 frames per second. Please notice that the device for recording videos is not particularly limited. Users can simply use a general smartphone and record his/her moves, and evaluate their posture by our system. This flexibility makes our system more practical and can be widely adopted.

B. Pose Estimator

We adopt OpenPose [9] to estimate 2D human pose from fitness videos, which is reliable to detect joints and can work in real-time. Given a fitness video, for each video frame the OpenPose system outputs 2D coordinates of 25 joints of a human body, as shown in Fig. 1. Based on the relative positions between joints and how they move temporally, we develop the proposed posture evaluation system.

C. Segmentation

When recording the fitness videos, the subject was asked to repetitively perform the same move multiple times. To evaluate each move, we need to segment the input video into clips, each of which contains one complete move. To do this, for both dumbbell lateral raise and biceps curl, we especially check the y coordinate of the joint at the right wrist. In a dumbbell lateral raise, the right arm holds up and down, yielding the y coordinate decreases (up) and then increases (down). We know it is a start of next move when the y coordinate decreases again. The same way is also used to segment a biceps curl. By this simple method, we segment a video into several separate dumbbell lateral raises (biceps curls), and take each of them as the basic unit for analysis.

D. Sampling

Although the subject repetitively performs the same move, different moves do not necessarily have the same length. To ease the succeeding analysis, we would like to represent each move as the same representation and thus can use the same model setting to evaluate different moves. In this work, no matter how long a move is, we uniformly sample skeletons of 17 instants from each move. At each instant, xy coordinates of the 25 joints are stored as a 50-dimensional vector $\boldsymbol{j}$. Therefore, each move can be represented as a series of seventeen 50-dimensional vectors $\boldsymbol{J} = (\boldsymbol{j}_1, \boldsymbol{j}_2, ..., \boldsymbol{j}_{17})$.

E. Posture Evaluation

1) Label: According to the domain knowledge of fitness, we design a few indicators to show if a move is well performed. Notice that these indicators are designed according to fitness domain knowledge. Given a move, these indicators will be evaluated and fed to the constructed model to do posture evaluation.

Taking a biceps curl as the example, three indicators are designed:

- The angle between the forearm and the upper arm should be less than 70 degrees when the dumbbell is lifted to the highest point.
- The upper arm must be close to the body without shaking.
- The upper body should be kept upright without shaking.

Every biceps curl in our training data is annotated as a 3-dimensional binary vector according to whether the aforementioned indicators are met. For example, if the subject's biceps curl move is annotated as $\boldsymbol{a} = (0, 1, 0)$, it means that his upper arm is not close to the body.

Overall, the training data to build the posture evaluation model are the collection of joints and annotations, i.e., $(\boldsymbol{J}_i, \boldsymbol{a}_i)$, $i = 1, 2, ..., N$, where $\boldsymbol{J}_i$ is the set of skeletons of the ith move, and $\boldsymbol{a}_i$ is the indicator vector of the ith move.

2) Model Construction: Since fitness moves are related to time, we choose long short-term memory networks (LSTM) as the model to describe the evolution of joints. We construct a 13-layer LSTM model, where each layer is a bidirectional LSTM. Given a sequence of skeletons (joint sets) $\boldsymbol{J} = (\boldsymbol{j}_1, \boldsymbol{j}_2, ..., \boldsymbol{j}_{17})$, the target output of the LSTM model is the annotation vector $\boldsymbol{a}$, i.e., $\hat{\boldsymbol{a}} = \mathcal{M}(\boldsymbol{J})$. The LSTM model $\mathcal{M}$ takes one skeleton at each time instant (from $\boldsymbol{j}_1$ to $\boldsymbol{j}_{17}$), processes to get 13 hidden vectors at each time instant, and then propagates to the end where a fully-connected layer with the sigmoid activation function outputs the predicted annotation vector $\hat{\boldsymbol{a}}$.

Batch normalization is added between layers to avoid overfitting and to speed up model convergence, and the Adam algorithm is used as the optimizer. We calculate binary cross entropy between the predicted annotation vector $\hat{\boldsymbol{a}}$ and the ground truth $\boldsymbol{a}$ as the loss function.

According to the predicted annotation vector $\hat{\boldsymbol{a}}$, we know where the problem is in the move. For example, for a biceps curl if $\hat{\boldsymbol{a}} = (0.2, 0.8, 0.1)$, which second dimension is larger

than a predefined threshold, we then can know that the subject's upper arm is not close to the body, or with clear shaking. In the evaluation, we will show a red circle that clearly displays where the bad posture is. This would be a direct and clear indicator for an amateur to realize where he/she acts badly.

IV. EXPERIMENT

A. Dataset and Experimental Settings

To simulate personal use of the evaluation system, we used a usual smartphone, i.e., iPhone 8, to capture the evaluation videos. The video resolution is 720P and the frame rate is 30 fps. The subject performing the fitness exercise is the author himself.

1) Biceps Curl: We have 983 biceps curl moves in total, consisting of 706 good moves and 277 bad moves. Good moves meet all criteria mentioned in the indicators, while a move is classified as bad when more than one criterion is not met. The biceps curl moves were captured from a slightly side view, as shown in Fig. 1.

2) Dumbbell Lateral Raise: We have 884 dumbbell lateral raise moves in total, consisting of 584 good moves and 300 bad moves. Fig. 2 shows two samples. They were captured from a frontal view.

In the evaluation, we evaluate the proposed LSTM model based on a 10-fold cross-validation scheme. For biceps curls, for example, we randomly select 90% of the data for training, and the remaining 10% is for testing. The training and testing processes are conducted 10 times, and the average test accuracy is reported.

B. Baseline

According to the designed indicators, we implement a baseline method for each considered fitness posture. Taking a biceps curl as the example again, we calculate the angle between the forearm and the upper arm based on the detected joints. If the angle is larger than 70 degrees, a posture failure (on the first indicator) is reported. Similarly, for the other indicator, the other specific implementation is implemented as the baseline method. The "baseline method" in the following is thus a set of basic implementation for each indicator. Notice that the baseline method just checks individual samples $\boldsymbol{j}_i$ without considering the temporal evolution in the entire movie.

C. Evaluation of Biceps Curl

When doing a bicep curl, we focus on the forearm. We lift the dumbbells from the lowest point, bending the forearm toward the direction of the upper arm, causing the biceps to contract, as shown in the left of Fig. 1. During the process of lifting, the upper arm must keep still, with the elbow fixed on the side of the body. The forearm bends toward the direction of the upper arm with the elbow as the fulcrum. During the process, the body is upright without shaking. Common mistakes include: 1) the upper arm is not fixed at the side of the torso, moving when the forearm bends; 2) shaking the body back and forth when the dumbbell is lifted; and 3) the dumbbell is not raised high enough to fully contract the biceps. The right of Fig. 1 shows a sample where the subject's upper arm is not at the side of the torso, and his body shakes too much. Based on this domain knowledge, we set three indicators to judge the quality of a bicep curl. Notice that conceptually these indicators are the same as that mentioned in Sec. III-E1, but with clearer operational definition.

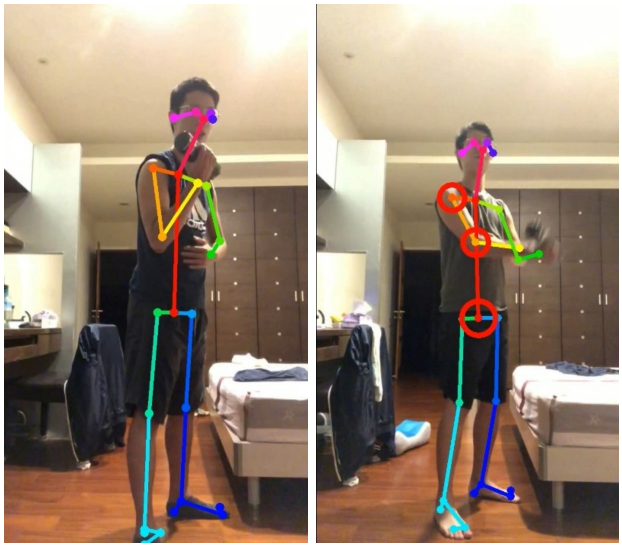

Fig. 1. Two sample bicep curl moves.

1) Indicator 1: The angle between the upper arm and the forearm should be less than 70 degrees when the dumbbell is lifted to the highest point.
2) Indicator 2: The angle between the forearm and the torso should be less than 20 degrees.
3) Indicator 3: The displacement of the hip position should not be greater than 15% of the length of the lower limb during lifting the dumbbell.

Table I shows classification accuracy for three indicators of bicep curls, based on the baseline method and the proposed LSTM, respectively. As can be seen, the proposed LSTM outperforms the baseline by a large margin. The reasons for this superiority may be twofold. First, the LSTM method considers the temporal evolution of joints, and is able to more accurately estimate the status of posture. Second, as shown in Fig. 1, the bicep curl videos were captured from a nearly side view. In this case, positions of joints are more probably missing or mis-detected because of body occlusion, and the baseline method is more sensitive to these noises.

TABLE I
CLASSIFICATION ACCURACY OF THREE INDICATORS OF BICEP CURL, BASED ON THE BASELINE METHOD AND THE PROPOSED LSTM.

Methods	indicator 1	indicator 2	indicator 3
Baseline	81.0%	76.0%	72.0%
LSTM	98.3%	95.6%	98.1%

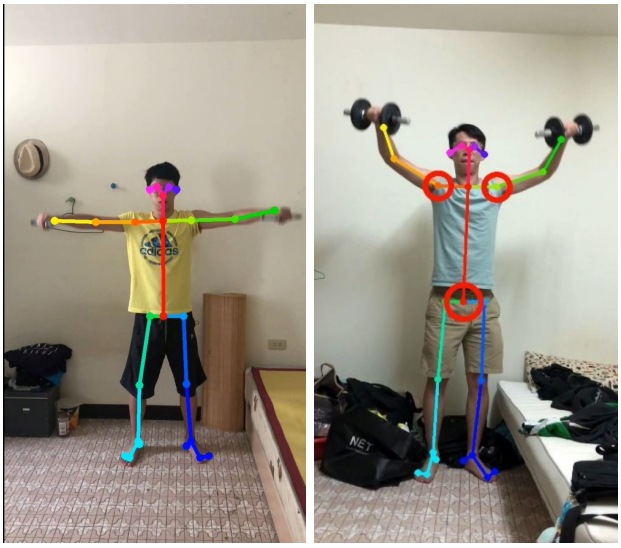

Fig. 2. Two sample dumbbell lateral raise moves.

D. Dumbbell Lateral Raise

When we do a dumbbell lateral raise, we focus on the shoulders. Both upper arms are raised to the height parallel to the ground, with the shoulder as the rotation center, as shown in Fig. 2. If the upper arms are raised too high, the trapezius is largely contracted, and we would feel uncomfortable. The arms should be naturally relaxed and don't bend or straighten excessively. The upper body should remain upright without shaking. Therefore, common mistakes include: 1) the upper arms are raised too high; 2) the arms bend too much, or stiffen too much; and 3) the body excessively shakes. The right of Fig. 2 shows a sample where the subject's upper arms are too high, and he shakes too much. Accordingly we design four indicators to judge the quality of a dumbbell lateral raise.

1) Indicator 1: The height of upper arms should not be higher than the shoulder.
2) Indicator 2: The angle between the upper arm and the forearm should be smaller than 170 degrees.
3) Indicator 3: The angle between the upper arm and the forearm should be larger than 120 degrees.
4) Indicator 4: The displacement of the hip should not be larger than 15% of the length of the lower limb during lifting dumbbell.

Table II shows classification accuracy for four indicators of dumbbell lateral raises, based on the baseline method and the proposed LSTM, respectively. We see that the proposed LSTM slightly outperforms the baseline method. Comparing with the results for bicep curls, the performance gain for dumbbell lateral raises is less. This would be because the dumbbell lateral raise videos were captured from the frontal view, and less joint detection miss or errors make the baseline method more reliable.

TABLE II
CLASSIFICATION ACCURACY OF FOUR INDICATORS OF DUMBBELL LATERAL RAISE, BASED ON THE BASELINE METHOD AND THE PROPOSED LSTM.

method	indicator 1	indicator 2	indicator 3	indicator 4
Baseline method	89.0%	94.0%	89.0%	91.0%
LSTM method	93.4%	91.6%	92.2%	90.8%

V. CONCLUSION

We have presented a posture evaluation system specially for fitness videos. We focus on two common fitness moves: bicep curl and dumbbell lateral raise. Given a fitness video, we first detect the joints and skeletons by the OpenPose library. Based on the evolution of positions of joints, the video is segmented into clips so that each clip contains one move. For each move, a series of positions of joints are fed to an LSTM model, which finally outputs the evaluation results of the predefined indicators. The indicators are designed according to the domain knowledge of fitness, in order to show what part of a move is bad. The evaluation results show that, by considering the temporal evolution of joints, the proposed LSTM method clearly outperforms the baseline method. As one of the very few studies on fitness video analysis, these results are very encouraging. In the future, more fitness moves can be evaluated, based on videos captured from various views. The designs of indicators can also be investigated further.

ACKNOWLEDGMENT

This work was partially supported by Qualcomm Technologies, Inc. under the grant number B109-K027D, and by the Ministry of Science and Technology, Taiwan, under the grant 108-2221-E-006-227-MY3, 107-2923-E-194-003-MY3, and 109-2218-E-002-015.

REFERENCES

[1] J. Wang, K. Qiu, H. Peng, J. Fu, and J. Zhu, "Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance," in *Proceedings of ACM International Conference on Multimedia*, 2019, pp. 374–382.
[2] R. Khurana, K. Ahuja, Z. Yu, J. Mankoff, C. Harrison, and M. Goel, "Gymcam: Detecting, recognizing and tracking simultaneous exercises in unconstrained scenes," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018.
[3] T. Alatiah and C. Chen, "Recognizing exercises and counting repetitions in real time," in *https://arxiv.org/abs/2005.03194*, 2020.
[4] A. Soro, G. Brunner, S. Tanner, and R. Wattenhofer, "Recognition and repetition counting for complex physical exercises with deep learning," *Sensors*, vol. 19, no. 3, 2019.
[5] A. Tharatipyakul, K. Choo, and S.T. Perrault, "Pose estimation for facilitating movement learning from online videos," in *Proceedings of International Conference on Advanced Visual Interfaces*, 2020.
[6] H. Xie, A. Watatani, and K. Miyata, "Visual feedback for core training with 3d human shape and pose," in *Proceedings of Nicograph International*, 2019.
[7] J. Li, H. Cui, T. Guo, Q. Hu, and Y. Shen, "Efficient fitness action analysis based on saptio-temporal feature encoding," in *Proceedings of IEEE International Conference on Multimedia & Expo Workshop*, 2020.
[8] S. Chen and R.R. Yang, "Pose trainer: Correcting exercise posture using pose estimation," in *https://arxiv.org/abs/2006.11718*, 2020.
[9] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.