Human Motion Assessment on Mobile Devices

Hobeom Jeon
Department of Computer Software
University of Science and Technology
Deajeon, Republic of Korea
tiger@etri.re.kr

*Dohyung Kim**
Intelligent Robotics Research Division
Electronics and Telecommunications Research Institute
Deajeon, Republic of Korea
dhkim008@etri.re.kr

Jaehong Kim
Intelligent Robotics Research Division
Electronics and Telecommunications Research Institute
Deajeon, Republic of Korea
jhkim504@etri.re.kr

2021 International Conference on Information and Communication Technology Convergence (ICTC) | 978-1-6654-2383-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICTC52510.2021.9621114

***Abstract*—Human motion analysis is being performed in various places using various devices. However, the complexity of the analysis algorithms required by high-performance devices makes it difficult to run these algorithms anytime and anywhere. We have successfully ported a pose estimation model to mobile devices and applied various post-processing algorithms. Our analysis algorithm, which uses the joint coordinates of the pose estimation model, has both fast calculation speed and versatility. Our lightweight pose models and motion analysis algorithms allow users to analyze fitness actions and receive instant feedback in real time using their smartphones. Our motion assessment system is applied to real mobile devices and tested in the real world. Based on empirical research on real environments, our motion assessment system is shown to be applicable to various fields.**

Keywords—human motion analysis, human pose estimation

I. Introduction

During the global pandemic, the demand for non-face-to-face services has exploded. Accordingly, the number of web and app-based media service users has increased, and the need for home training content is also growing. However, when using video-based home training content, the user follows the motions but cannot receive appropriate feedback. In recent years, significant progress in activity recognition [1] technology using human pose estimation has made AI-based coaching systems [2],[3] (i.e., interactive home training) possible.

This paper introduces a fitness motion assessment system optimized for mobile devices. Our fitness motion assessment system aims to interact with users in real time on their devices. Fig. 1. shows an overview of our system. First, a deep learning-based lightweight human pose estimation model [4] is used to obtain 2D joint coordinates faster and more accurately. Second, our motion analyzer uses static and dynamic comparisons between standard motion and learner's motion. Since the scalability of the standard fitness action database is compromised when using a data-driven learning approach in analyzer, we use analytical techniques. The system analyzes posture similarity and motion similarity, i.e., pose configuration and joint dynamics. Whenever the user completes a movement during an exercise, the system immediately provides the motion analysis results. By giving users instant feedback, users can quickly improve their incorrect form.

Our contributions are summarized below:

- We optimize 2D human pose estimation for mobile devices using effective and lightweight techniques. Furthermore, post-processing techniques are adapted for providing a better user experience with smooth visualization results in a real-time environment.

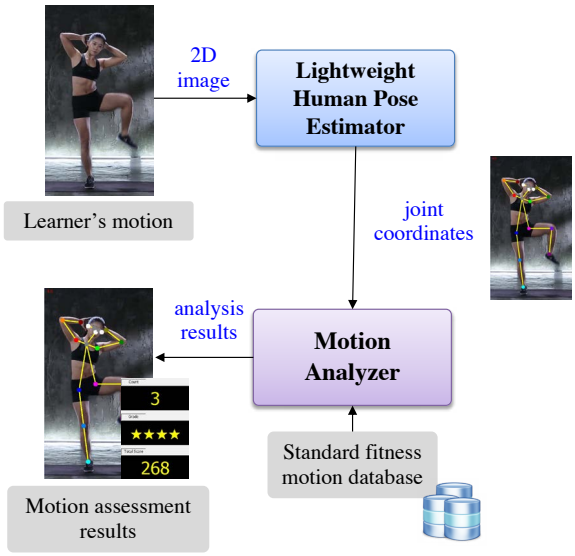

Fig. 1. Overview of proposed fitness motion assessment system

- We propose a simple method for assessing movements. Our approach is highly scalable when adding new exercises and works regardless of human anthropometric ratios such as height and body shape.
- We present a pragmatic pipeline for self-learning fitness movements on mobile platforms. We test our fitness assessment system performance in a real environment.

II. Related Work

A. Fast Human Pose Estimation

Recently, 2D human pose estimation (HPE) has leveraged deep learning technology and achieved usable accuracy in real world [5]. Knowledge distillation (KD) methods that reduce accuracy loss are also accurate in pose estimation. The pose distillation (PD) [6] for human pose estimation uses an output heatmap to mimic a pre-trained larger model's heatmap. The PD performed effective pose knowledge transfer by applying intermediate distillation loss between the hourglass stages. Self-evolutionary pose distillation [7] has proposed a remarkable knowledge distillation technique that reduces the number of steps by initializing parameters through self-replication.

However, the aforementioned lightweight networks are not optimized for mobile environments, so we apply knowledge distillation to even smaller networks. We also introduce techniques that appropriately manage lag and jitter in small HPE networks for mobile applications.

* Corresponding author

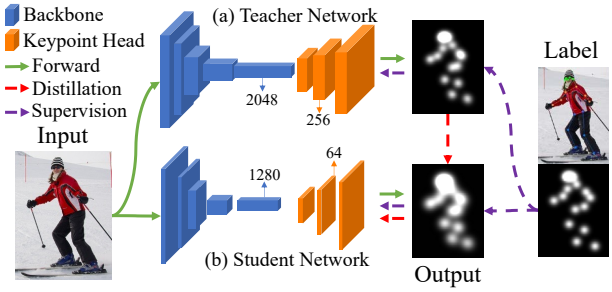

Fig. 2. Overview of lightweight human pose estimation and Online Pose Distillation Path.

B. Motion Assessment System

Motion assessment systems are primarily divided into 3D and 2D camera-based approaches. The accuracy of human joint coordinates is the highest for motion-capture-based 3D coordinates and the lowest in lightweight network-based 2D coordinates on mobile devices.

Dedicated hardware connected to smart mirrors or large monitors uses 3D cameras to analyze behavior with high accuracy [2], [8], [9]. In addition, some products offer personalized services that multiple people can use in public gyms [10], [11]. 2D camera-based systems typically have the advantage of running independently on personal mobile devices, regardless of location or time. 2D camera-based products for sports such as basketball and other fitness activities provide practical exercise guidelines [12], [13], [14].

III. Method

A. Lightweight Human Pose Estimation

It is challenging to deploy previous large HPE models to resource-constrained devices such as smartphones and embedded systems. To address this problem, we use Mobilenetv2 [15] to optimize the existing HPE model [16] for our target systems, which have low-powered processing units and a small amount of memory. The (b) student model in Fig. 2. illustrates the structure of our lightweight pose model. Mobilenetv2 uses depthwise-separable convolution and, at our input size, has a low computational cost of 0.31 GFOPs. Nevertheless, it demonstrated competitive performance by minimizing information loss in nonlinear functions using inverse residual blocks. Visual features extracted using Mobilenetv2 represent the location of the joint using a heatmap; the heat map is constructed using a 3-time deconvolution operation on the keypoint head and depicts the Gaussian distribution centered on the coordinates of the joints located in the image.

To minimize the performance drop caused by lightening the model, we use the KD method. The Online Pose Distillation (OPD) trains a teacher network and a student network simultaneously; the student network is trained with the teacher network, which provides the correct answer. Thus, the student network has fine-grained representation capabilities through the output heatmaps generated by the teacher network during the training process.

The HPE model, trained with a 15% larger box than that for joint coordinates due to cropping of the input image, most accurately estimates the coordinates when the person is in the center. We use bounding box propagation to obtain a person-centered, cropped image. Lighttrack [17] showed that in multi-person frames, fast and accurate estimation is made possible by skipping the human detector using bounding box propagation between adjacent frames. The bounding box propagation technique crops the next frame with the box coordinates created using the joints estimated from the current frame and adjusts the input values to a distribution similar to the trained image.

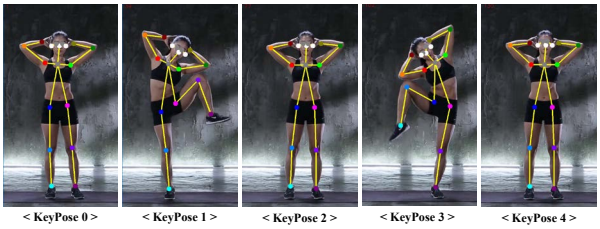

Fig. 3. Sample KeyPose frames in side-bend knee up.

We apply heatmap-smoothing, quarter-shift [15], and the one-euro filter [18] in post-processing. Heatmap-smoothing uses the weighted sum of the heatmaps from the previous frame and the current frame. Our system avoids large fluctuations in body coordinate using heatmap-smoothing to considering joints whose estimations are uncertain due to interference by the background or occlusion. Quarter-shift compares the joint positions x and y with x-1, x+1 and y-1, y+1 at the output resolution and moves the position by 1/4 pixel in the direction with higher probability. The One-Euro filter uses a jitter correction coefficient that dynamically changes according to the lag. We use the values $f\_cmin = 1.7, \beta = 0.4$, which are optimized for the 2D pose model [19]. When estimating poses in real time, delays fluctuate significantly and are applied to provide users with smooth results.

B. Motion Assessment

1) KeyPose

We built a standard fitness action database to evaluate fitness movements. The standard fitness motion database consists of fitness videos, joint coordinates extracted from each frame, and KeyPoses. The KeyPose stores the frame number, which consists of the posture positions that can distinguish fitness action. Fig. 3. shows the KeyPoses for a side-bend knee-up action. The side-bend knee-up repeats a 0-4 KeyPose by pulling the knee toward the elbow from side to side. Our motion assessment system uses joint coordinates corresponding to KeyPoses for evaluation. Using KeyPoses, we can efficiently and quickly assess movement using only the motion at a critical time point-in-time action.

2) KeyPose Matching

The KeyPose matching process compares the similarity between the current user's posture P^T and the KeyPose posture P^K stored in the database. We analyze posture similarity using the Pose Configuration Score (PCS) and motion similarity using the Joint Dynamics Score (JDS). The PCS, which analyzes a user's static posture, uses the angle information between limb vectors. We use a joint pair set $\mathbb{P}$ consisting of limb vectors $\vec{l} = (x_i, y_i) \rightarrow (x_j, y_j)$; the number of vectors is $n(\mathbb{P}) = 13$. The PCS obtains cosine similarity for the user's limb vector $\overrightarrow{l^T}$ and KeyPose limb $\overrightarrow{l^K}$ and measures this similarity as a weighted sum for each limb as follows:

$$PCS\left(\overrightarrow{l^K}, \overrightarrow{l^T}\right) = \sum_{n=1}^{n(\mathbb{P})} w_n \frac{\overrightarrow{l^{K_n}} \cdot \overrightarrow{l^{T_n}}}{\left\|\overrightarrow{l^{K_n}}\right\| \left\|\overrightarrow{l^{T_n}}\right\|} \quad (1)$$

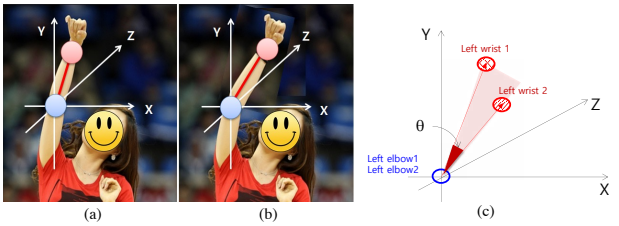

Fig. 4. Visualization of pose confidence score

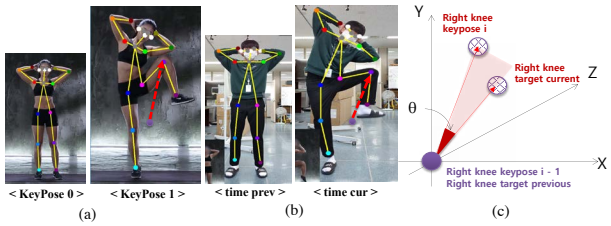

Fig. 5. Visualization of joint dynamics score

Fig. 4. shows limb vectors (left elbow, left wrist) in images (a) and (b). As the angle θ of the vectors for each limb shown in (c) decreases, the similarity of the posture increases.

JDS uses dynamic vectors that have moving information between KeyPoses listed in order. For example, Fig. 5. (a) shows KeyPose 0-1 in the trainer video, and then (b) is the user's joint information obtained in real-time from the pose estimator. We use the time difference between the trainer's KeyPose to find the user's previous posture frame; Fig. 5. (b) depicts that frame. We use a set $\mathbb{D}$, which consists of Joint dynamic vectors $\vec{p} = (x_i, y_i)_{k-1} \rightarrow (x_j, y_j)_k$. The number of vectors is $n(\mathbb{D}) = 9$. At the computation time t, JDS measures the similarity using the dot product between the joint dynamic vectors as follows:

$$JDS\left(p(P_k^K, P_{k-1}^K), p(P_t^T, P_{t-f}^T)\right) = \sum_{n=1}^{n(\mathbb{D})} w_n \frac{\overrightarrow{p^{K_n}}}{r^K} \cdot \frac{\overrightarrow{p^{T_n}}}{r^T} \quad (2)$$

The user's KeyPose P_{t-f}^T is defined in real time using the frame difference f_{k-1} between the k-th KeyPose P_k^K and the previous KeyPose P_{k-1}^K. A vital difference between JDS and PCS is that JDS involves the distance over which the joints moved. To compensate for the vector size, which depends on the length of human limbs and camera viewpoint, we normalize it to the torso vector norm $r = \|\text{Nect} \rightarrow \text{HipLeft}\|$. Therefore, JDS works robustly with human anthropometric ratios calculated using normalization.

3) Motion Counting

Using PCS and JDS, the motion assessment system counts each iteration of the fitness actions. The user interface repeatedly plays the video and calculates a KeyPose score by weighting the sum of the PCS and JDS when each KeyPose appears. In all KeyPoses, if the score exceeds the threshold, it counts as the number of iterations. Thus, each fitness assessment score is the average of all KeyPose scores. Our motion assessment algorithm dynamically and statically analyzes joint coordinates in real time to measure motion similarity, track action progress, and monitor the number of iterations.

IV. Experiments

A. Lightweight Human Pose Estimation

We use the COCO dataset [20], a public benchmark dataset, and OKS metric to verify our lightweight pose model. In Table I, the giga floating point operations (GFLOPs) result of the lightweight model is 0.44, which shows that the computational efficiency is nine times greater than the 4.10 of the teacher model. In addition, performance degradation is minimized by applying OPD, denoted by + in the table, and shows an improvement of 1.2 AP.

In Table II, the lightweight model pre-trained with OPD on the COCO dataset achieved 65.2 AP on our fitness validation set. Using the COCO pre-trained model, we achieve 89.6 AP by fine-tuning on fitness datasets. Our pose estimation model is intended to estimate fitness actions, so we further fine-tune using the fitness dataset and achieve greater performance on deformable fitness postures. Moreover, these results show that OPD works effectively on other datasets as well as COCO.

To examine whether a fast estimation speed is possible in a mobile environment, we use the TensorFlow [21] toolkit to measure the performance on Android phones. Table III shows the fastest speed, 13ms, on a Samsung s10. iPhone 11 achieved a speed of 14 ms when using a neural core and demonstrated a surprisingly fast speed of 17 ms using just the CPU.

TABLE I. Pose Estimation Model Performance on the COCO Keypoints 2017 Validation Set.

Method	GFLOPs	COCO			
		AP	AP^{50}	AP^{75}	AP^L
Teacher	4.10	72.5	92.5	80.2	77.0
Ours	0.44	64.2	**89.4**	71.7	68.6
Ours+	**0.44**	**65.4**	**89.4**	**72.9**	**70.0**

+ denote online pose distillation learning

TABLE II. Pose Estimation Model Performance on the Fitness Keypoints Validation Set.

Method	COCO pre-trained				fitness fine-tuned			
	AP	AP^{50}	AP^{75}	AP^L	AP	AP^{50}	AP^{75}	AP^L
Teacher	76.4	95.8	83.1	76.7	93.5	99.0	97.9	93.8
Ours	63.0	90.4	67.6	63.4	88.2	**99.0**	95.5	88.5
Ours+	**65.2**	**91.4**	**70.0**	**65.5**	**89.6**	**99.0**	**96.7**	**89.9**

+ denote online pose distillation learning

TABLE III. Pose Estimation Model Inference Speed Comparison on Various Smartphones.

Type	Samsung S10	Samsung Note9	Google Pixel 3	IPhone 11
CPU	100 ms	108 ms	46 ms	17 ms
GPU	13 ms	17 ms	17 ms	14 ms

TABLE IV. Fitness Action Assesment Test in a Real-World Environment.

Fitness Action	Count iteration / Total iteration	Accuracy
Chest-stretch	115 / 115	100 %
Squat shoulder press	115 / 115	100 %
Tuck Jump	110 / 115	95.65 %
Side-bend knee-up	115 / 115	100 %
Barbell power clean	105 / 115	91.30 %
Total	560 / 575	97.39 %

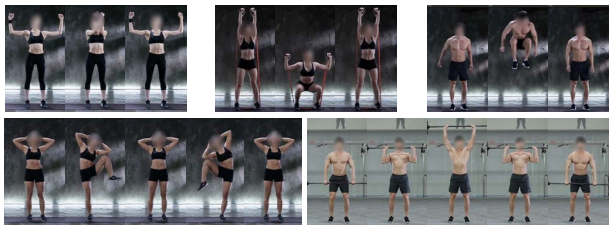

Fig. 6. Example of the standard fitness action database. From the top left are a *chest-stretch, squat shoulder press,* and *tuck jump*. From the bottom left are a *side-bend knee-up* and *barbell power clean*.

B. Motion Assessment

We evaluate our motion analysis system performance on 23 people in the real world. Testers follow the fitness movements presented on the screen and evaluate that our system accurately tracks the number of iterations. The fitness videos consist of a total of five movement types: the chest-stretch, side-bend knee-up, squat shoulder press, barbell power clean, and tuck jump. These movements are repeated five times. Table IV presents the experimental results. Our analysis system tracks 560 movements out of 575 executions and calculates the motion analysis score, achieving 97.39% accuracy.

V. Conclusion

This paper proposes a Pose Configuration Score and Joint Dynamics Score to measure posture similarity using joint coordinates for static and dynamic motion assessment. The proposed analysis metric is robust to human body proportions, so it is highly scalable and versatile. The lightweight pose model is successfully implanted to mobile devices and obtains joint coordinates in real time. Using the Online Pose Distillation technique, more accurate coordinates are estimated, and precise motion analysis is possible. Furthermore, the system can analyze users' fitness actions on-device in real time. This strategy eliminates unnecessary operations because it uses only the KeyPoses, which provide essential frame information. However, the proposed motion analysis system has a limitation in that it requires an action database to compare with users. This obstacle can be easily overcome and implies scalability using a verified motion database for dance motion, rehabilitation exercise, and body gestures.

Acknowledgment

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No.2017-0-00162, Development of Human-care Robot Technology for Aging Society)

References

[1] İnce, Ö. F., et al. "Human activity recognition with analysis of angles between skeletal joints using a RGB-depth sensor." *ETRI journal* 42.1, 2020, pp. 78-89.

[2] *Tempo, Inc*. (2021) The award-winning ai-powered home gym. [Online]. Available: https://tempo.fit/

[3] *Kakao VX, Corp.* (2021) Enjoy the exercise that fits you with the exercise service Smart Home operated by Kakao VX. [Online]. Available: https://www.kakaoht.com/

[4] H. Jeon, Y. Yoon, and D. Kim, "Lightweight 2D human pose estimation for fitness coaching system," presented at the 2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Jun. 2021.

[5] Yun, K., Kwon, Y., Oh, S., Moon, J., and Park, J., "Vision-based garbage dumping action detection for real-world surveillance platform." *ETRI Journal* 41.4, 2019, pp. 494-505.

[6] F. Zhang, X. Zhu, and M. Ye, "Fast human pose estimation," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019.

[7] F. Zhang, H. Hu, H. Dai, L. Zhou, and M. Ye, "Self-evolutionary pose distillation," in *2019 16th Int. Computer Conf. on Wavelet Active Media Technology and Information Processing*. IEEE, pp. 240–244.

[8] *Lululemon Athletica, Inc*. (2021) The nearly invisible home gym, MIRROR. [Online]. Available: https://www.mirror.co/

[9] *Interactive Strength, Inc*. (2021) FORME Life studio. [Online]. Available: https://formelife.com/

[10] *JoyFun, Co.,Ltd*. (2021) Joy Fitness, smart mixed reality fitness device. [Online]. Available: https://joyfun.co.kr/

[11] *MyBenefit, Inc*. (2021) Virtual mate, empower your fitness. [Online]. Available: http://www.mybenefit.co/

[12] *VAY, Inc*. (2021) Vay sports, the 1 human motion analysis for fitness health. [Online]. Available: https://www.vay.ai/

[13] *NEX Team, Inc*. (2021) HomeCourt, upgrade your game. [Online]. Available: https://www.homecourt.ai/

[14] *WeHealed, Inc*. (2021) Exercise gets better, LIKEFIT. [Online]. Available: https://www.likefit.me/

[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[16] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. of the European Conference on Computer Vision* (ECCV), 2018, pp. 466–481.

[17] G. Ning, J. Pei, and H. Huang, "Lighttrack: A generic framework for online top-down human pose tracking," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 1034–1035.

[18] G. Casiez, N. Roussel, and D. Vogel, "1€ filter: a simple speed-based low-pass filter for noisy input in interactive systems," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, 2012, pp. 2527–2530.

[19] D. Mehta *et al.*, "XNect: Real-time multi-person 3d motion capture with a single rgb camera," *ACM Transactions on Graphics*, vol. 39, no. 4, pp. 82–1, 2020.

[20] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European Conf. on Computer Vision*. Springer, 2014, pp. 740–755.

[21] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/