DEEP LEARNING FOR GESTURE RECOGNITION IN GYM TRAINING PERFORMED BY A VISION-BASED AUGMENTED REALITY SMART MIRROR

B. Lanza[1], C. Nuzzi[1], S. Pasinetti[1], M. Lancini[2]

Dep. Of Mechanical and Industrial Engineering, University of Brescia, Italy[1]
Dep. Of Medical and Surgical Specialities, Radiological Sciences and Public Health, University of Brescia, Italy[2]

This paper illustrates the development and the validation of a smart mirror for sport training. The application is based the skeletonization algorithm *MediaPipe* and runs on an embedded device Nvidia Jetson Nano equipped with two fisheye cameras. The software has been evaluated considering the exercise biceps curl. The elbow angle has been measured by both *MediaPipe* and the motion capture system BTS (ground truth), and the resulting values have been compared to determine angle uncertainty, residual errors, and intra-subject and inter-subject repeatability. The uncertainty of the joints' estimation and the quality of the image captured by the cameras reflect on the final uncertainty of the indicator over time, highlighting the areas of improvements for further developments.

KEYWORDS: smart mirror, skeletonization, deep learning, embedded vision.

INTRODUCTION: In the context of free body physical exercise, the variety of human movements make traditional mechanical measurement systems ill-suited to assess human motion (Chan & Liu, 2009). Moreover, such approaches are often calibrated on the athlete to produce reliable results. However, vision-based Deep Learning (DL) techniques allow model-based estimation of the body movement to measure anthropometrics quantities (i) without the need to calibrate the model on the athlete, and (ii) without forcing the athlete to wear measurement devices during the exercise (Cao *et al.*, 2019). The method adopted in this work is a DL skeletonization algorithm named *MediaPipe* (Bazarevsky & Grishchenko, 2020) applied on color images that estimates the position of the athlete's body joints (Figure 1). These joints could be used to produce a smart mirror of the exercise in real-time. Their position along with their estimation uncertainty could also be used to compute intuitive feedback indicators for the athlete. Therefore, this study presents a comparison of embedded hardware to determine the best performing one and a brief study on the feedback indicator computed for an example exercise *biceps curl*.

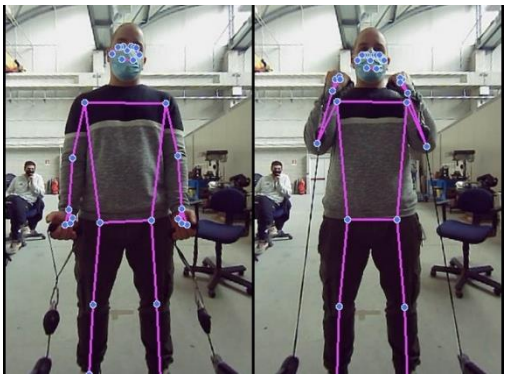

Figure 1. Example of the estimated skeleton computed on a color image.

METHODS: The proposed prototype involves (i) a low-cost embedded device equipped with a dedicated graphic processing unit (GPU) and (ii) two wide-angle fisheye color cameras. The cameras are positioned vertically to acquire the whole-body image. However, given the evident distortion of the fisheye lenses, it is fundamental to first calibrate the cameras in order to remove the effect of lens curvature. This step is performed by a calibration procedure involving a calibration pattern (i. e. a chessboard image printed on quality paper and glued on a planar sturdy surface) (Zhang, 2000). The images acquired by the two cameras are stitched together

to compose a full frame that is both used by the DL model to perform skeletonization and as a reference for the athlete to check if the exercise is executed in the right way, thus creating a smart mirror.
It is known by literature that DL models are often optimized for computers equipped with high-performing GPUs. However, embedded devices have different processor's architectures and component's limitations; hence, they require a lighter model in terms of byte size. Among the plethora of skeletonization algorithms available, *MediaPipe* has been chosen for this work thanks to its accuracy performance estimating the skeleton's joints at a reduced computational cost compared to other models (Cao *et al.*, 2019; Bazarevsky & Grishchenko, 2020; Sarkar *et al.,* 2018). Considering the deployment of the software on an embedded device with limited resources, it is fundamental to determine the right hardware to keep up with the model's computational needs (Mitchell *et al.*, 2021). Among the options available on the market, after a careful testing the Nvidia Jetson Nano device has been selected as the target hardware.
To estimate the gesture recognition repeatability during exercise, we focused on relevant kinematic variables, i. e. elbow angle α in the *biceps curl* exercise. However, the accuracy of the elbow angle measurement is subjected to pixel-related aberration. In addition, every time two body parts overlap, the image features are blurred and mixed. These phenomena affect the repeatability of the proposed measurement system; thus, we need to estimate them. For example, angle α between arm and forearm in the *biceps curl* exercise is measured considering an uncertainty composed of two factors: $\sigma_{tot}^2 = \sigma_{PCK@0.2}^2 + \sigma_{KP\_overlap}^2$, where $\sigma_{PCK@0.2}$ is an accuracy parameter estimated during the training phase of *MediaPipe*. This method considers a joint keypoint to be correctly detected if: (i) its predicted visibility matches the ground truth's one, and (ii) the absolute Euclidean error between the reference and target keypoints, normalized by the torso diameter projection, is smaller than 20% (Mitchell *et al.*, 2021). The second parameter $\sigma_{KP\_overlap}$ is influenced by the distance between two different keypoints. When two joints overlap, the neural network is unable to identify them correctly. Despite the absence of human motion, the overlapped joints coordinates show a scattered trend over consecutive frames. Hence, this uncertainty is strictly dependent on the joints distance $\sigma_{KP\_overlap} = \sigma(\Delta P_{i,j})$, where keypoints P_i and P_j represent the wrist and the shoulder joints respectively.

To evaluate *MediaPipe*'s performance to correctly predict skeleton joints, the resulting angle α is compared to the angle obtained from a motion capture BTS system (ground truth). For the exercise *biceps curl* the measured angle spans from 0 deg (arms folded state) to 180 deg (arms extended state). Three subjects have been recruited for the experiment and were asked to perform the exercise repeatedly from 5 to 10 times according to their strength and physical condition. A single repetition starts from the extended state and ends after reaching a folded state, right before the transition to reach another extended state begins. Therefore, for each repetition the extended and folded states have been isolated, and the mean value of the angle has been extracted for both *MediaPipe* and BTS exercise's data. This allows for a comparison between the mean values of the same repetition's state by considering the root mean square error (RMSE) ε computed as the difference between the two angles:
$\varepsilon_{RMSE} = \sqrt[2]{\frac{\sum_{i=1}^{n} (\alpha_{MP\_i} - \alpha_{BTS\_i})^2}{n}}$).
Intra-subject and inter-subject repeatability have been computed as well by extracting the standard deviation of angle values for BTS systems.
RESULTS: *MediaPipe* can be adopted in three operational ways according to the complexity of the model, which reflects on the joint's prediction accuracy. The processing time is expressed in frames per second (fps) and is also called *inference time*. It refers to the time needed to extract the skeleton from a single frame. The skeleton estimated by *MediaPipe* is composed of 33 human body's joints. Hence, to evaluate the athlete's performance during training it is sufficient to monitor a subset of joints of interest according to the exercise and extract an intuitive feedback indicator for the user. Considering the simple exercise *biceps curl*, it is possible to extract the value of angle α and its estimation uncertainty as explained in the

previous Section. Therefore, computing the same analysis frame by frame during the execution of the exercise allows for a simple visualization of the angle's trend over time, as shown in Figure 2.

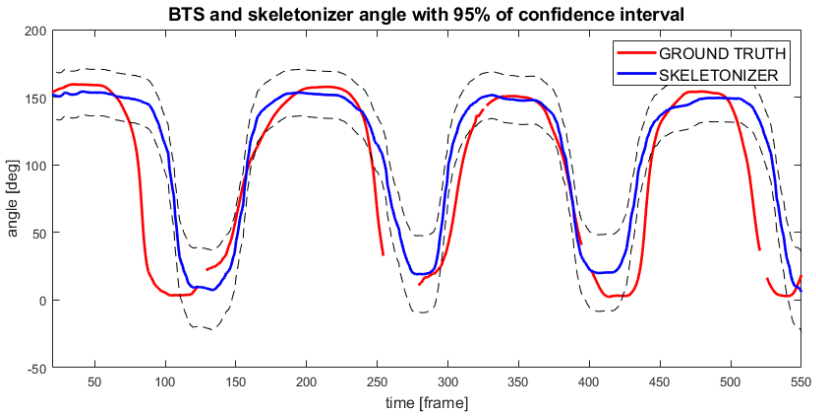

Figure 2. Elbow angle time evolution with uncertainty during a *biceps curl* exercise

Repeatability values have been calculated considering the mean values of the repetition state, as described in the previous Section. Figure 3 shows the root mean square errors for each subject. Figure 4 confronts the repeatability values of the three subjects (intra-subject) and the inter-subject repeatability.

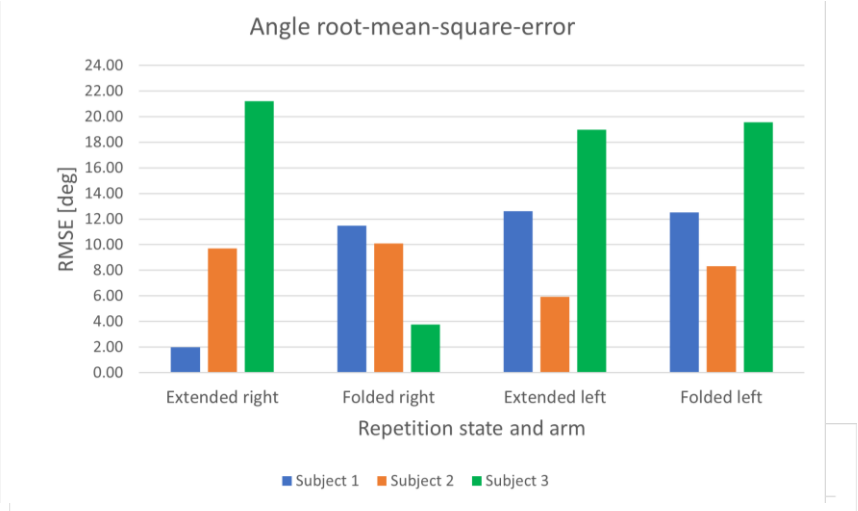

Figure 3. Angle root-mean-square-error of each subject.

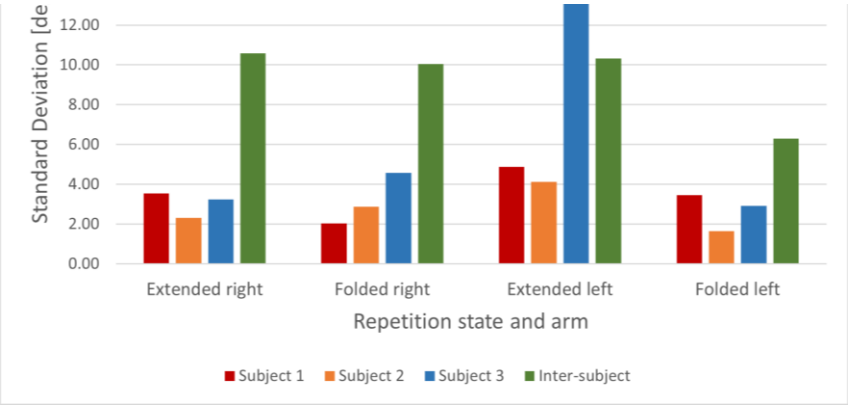

Figure 4. Intra-subject repeatability (red, orange, blue bars); inter-subject repeatability (green).

DISCUSSION: The core idea of the smart mirror application is to (i) visualize the athlete's skeleton during exercise, and (ii) produce valid information to support its training. Therefore, the indicators feedback must be reliable. Considering the biomechanical analysis conducted on the *biceps curl* exercise and shown in Figure 2**Error! Reference source not found.**, the uncertainty produced by the model's joints estimation is relatively small when angle α is high (arms extended, joints fully visible), while is higher when α is close to zero (arms folded, joints overlapping). It is also evident from Figure 4, where the accuracy while in folded state is typically worse. Furthermore, among the participants subject 3 is the worst performing, resulting in higher *RMSE* values (Figure 3).
This is due to several effects affecting the result: (i) the DL nature of the approach, which produces an estimation of the joints location according to the image quality and is thus affected by prediction uncertainty, (ii) the cameras' distortion that could not be completely removed after the calibration procedure, (iii) image resolution and quality, depending on the cameras' sensor, (iv) hardware limitations affecting *MediaPipe* overall performance in terms of both fps and prediction accuracy according to the type of model adopted (i. e. heavy, full, lite). However, even considering these limitations, the approach shown promising results and should be further studied to improve it. For example, a deeper and exhaustive approach using several kinematic variables could help produce more feedback indicators to better help the athlete during training.

CONCLUSION: The idea of this work is to develop a smart mirror application to help the athlete during physical training. A real-time video feedback of its body movements and some helpful biomechanical indicators are produced as output to improve its training performance. Albeit the application is in its embryonal stage, at the present time the investigation conducted allowed to analyse *MediaPipe*'s skeleton performance during the *biceps curl* exercise in comparison with a ground truth motion capture system BTS (Topley & Richards 2020). It is evident from the results that the higher uncertainty values occur when the arms joints overlap, corresponding to small values of angle α. To improve the estimation, better quality cameras should be adopted also considering their compatibility with the hardware, and a thorough calibration procedure must be performed to remove the effect of lens distortion and perspective deformations. Therefore, further developments involve testing several calibration procedures to improve the image quality even on low-cost cameras.

REFERENCES

Bazarevsky V. and Grishchenko I. (2020), "*On-device, Real-time Body Pose Tracking with MediaPipe BlazePose*".
Cao Z. *et al.* (2019*)*, "*Trends in OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields*".
Chan C. S. and Liu H. (2009), "*Fuzzy Qualitative Human Motion Analysis*" in *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 4, pp. 851-862, doi: 10.1109/TFUZZ.2009.2016553.
Mitchell M. *et al.* (2021), "Model Card BlazePose GHUM 3D".
Sarkar D. *et al.* (2018) "*Hands-On Transfer Learning with Python: Implement advanced deep learning and neural network models using TensorFlow and Keras"*. Packt Publishing Ltd.
Topley M. & Richards J. G. (2020*), "A comparison of currently available optoelectronic motion capture systems"* University of Delaware, Newark, DE 19716, USA.
Zhang Z. (2000), *"A flexible new technique for camera calibration."* IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11):1330-1334.

ACKNOWLEDGEMENTS: The authors would like to thank *ABHorizon*'s staff for their professional help and support during the development of this project.