VALIDATION OF A SMART MIRROR FOR GESTURE RECOGNITION IN GYM TRAINING PERFORMED BY A VISION-BASED DEEP LEARNING SYSTEM

Bernardo Lanza [(1)], Cristina Nuzzi [(1)], Simone Pasinetti [(1)], Luca Foletti [(1)], Matteo Lancini [(2)], Giovanna Sansoni [(1)]

[(1)] Dep. of Mechanical and Industrial Engineering, University of Brescia, 25123 - Brescia

[(2)] Dep. of Medical and Surgical Specialties, Radiological Sciences, and Public Health, University of Brescia, 25123 - Brescia

Corresponding author's email: b.lanza003@unibs.it

1. INTRODUCTION

This paper illustrates the development and the validation of a smart mirror for sport training. The application is based on 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.

2. METHODS

The method adopted in this work is a Deep Learning skeletonization algorithm named *MediaPipe* [1] applied on colour images that estimates the position of the athlete's body joints. 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. To estimate the gesture recognition repeatability during exercise, we focused on relevant kinematic variables, i. e. elbow angle $\boldsymbol{\alpha}$ 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. When two joints overlap, the neural network is unable to identify them correctly.

To evaluate *MediaPipe*'s performance to correctly predict skeleton joints, the resulting angle $\boldsymbol{\alpha}$ 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) $\boldsymbol{\varepsilon}$ 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.

3. RESULTS

Repeatability values have been calculated considering the mean values of the repetition state, as described in the previous Section. *Figure 1* shows the root mean square errors for each subject. *Figure 2* confronts the repeatability values of the three subjects (intra-subject) and the inter-subject repeatability. 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 1*, where the accuracy while in folded state is typically worse.

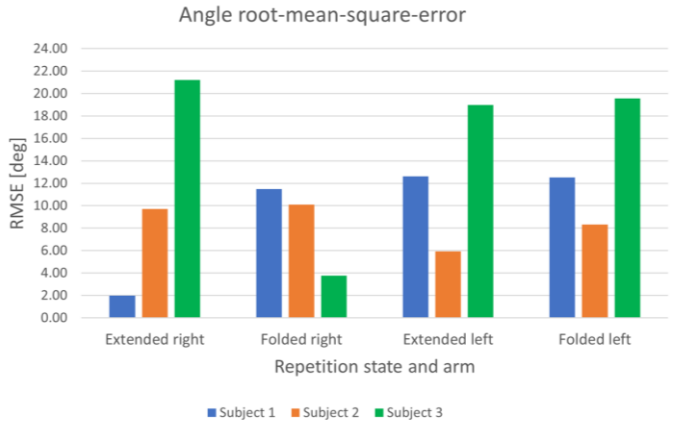

Figure 1 - Angle root-mean-square-error of each subject.

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 [2]).

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.

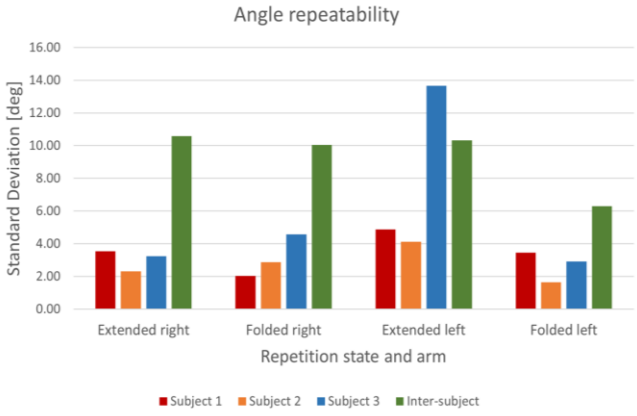

Figure 2 - Intra-subject repeatability (first, second, third column); inter-subject repeatability (fourth).

REFERENCES

[1] Valentin Bazarevsky and Ivan Grishchenko, Research Engineers, Google Research,"On-device, Real-time Body Pose Tracking with MediaPipe BlazePose", 2020.

[2] Model Cards for Model Reporting Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru "Model Card BlazePose GHUM 3D" 14 Jan 2019.