Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 207 (2022) 2536–2545

Procedia Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

A comparative analysis of pose estimation models as enablers for a smart-mirror physical rehabilitation system

Cristina Bolaños[a], Jesús Fernández-Bermejo[a], Javier Dorado[a], Henry Agustín[a], Félix Jesús Villanueva[a], María José Santofimia[a]

[a]*School of Computer Science, University of Castilla-La Mancha, Paseo de la universidad 4, Ciudad Real, Spain*

Abstract

Smart mirrors are gaining attention as a smart device that could integrate a set of functionalities intended to assist older adults in their day-to-day life. These devices are seamlessly integrated in the environment, providing a user-friendly interface and naturally fitting into the daily-care routines. People face a mirror several times a day, thus ensuring that any application running on a smart mirror will have several guaranteed interactions per day. It is therefore essential to detect when the user is in front of the mirror and also to interpret what he or she is doing. Very powerful and accurate libraries are currently available, but the limited computational resources and the need to work in real time limit the valid options for smart mirror devices. This paper therefore analyses and evaluates several body pose estimation models in order to determine which one can be deployed in a smart mirror-like device dedicated to supporting older adults in their physical rehabilitation routines.

Keywords: pose estimation; smart mirror; evaluation; rehabilitation support

1. Introduction

Providing digital support to the older adult population group is a task that requires devices that users are familiar with. The lack of digital skills or the cognitive decline associated to age could be barriers when this support is provided with devices such as mobile phones, laptops, tablets, etc. [4]. More natural interaction interfaces could facilitate digital support for older adults (e.g. voice interactions, gestures, etc.) [11]. One piece of home furniture that represents a natural interface is the mirror. Mirrors are present in all areas of our lives, especially in routines of daily care (mainly for beauty and physical habits) or as decoration.

E-mail address: Cristina@Bolanos@uclm.es

10.1016/j.procs.2022.09.312

Smart mirrors have been recently receiving great attention because of the different purposes to which they can be applied. Johri et al. [10] propose a smart mirror device as a newspaper reader assistant. Purohit et al. [14] smart mirror operates as an assistant, focusing on face recognition and user requests solver. Carmona et al. [6] present a smart mirror conceived to promote healthier lifestyles and general well-being. Silapasuphakornwong et al. [18] introduce a smart mirror which monitorize users emotions to identify initial signs of depression in the elderly. Henriquez et al. [8][1] smart mirror aims to detect cardio-metabolic risks from both user face analysis and psycho-somatic status recognition. Erazo et al. [3] propose a smart mirror device for neurorehabilitation, specially for users with upper limb dysfunction, using a Microsoft Kinect.

H2020 SHAPES project, acronym for the Smart & Healthy Aging through People Engaging in Supportive Systems (SHAPES) [7] is devoted to explore digital solutions that support older adults in extending the time they can live independently. In the context of this project, several digital solutions are proposed that could benefit from having a more natural interface. The smart mirror idea is proposed to support, among other digital solutions, one that is intended to assist older adults while performing physical exercises, either in the context of a physical rehabilitation intervention or, just to delay the effects that ageing has on body musculature. To this end, a traditional mirror is equipped with an embedded computer and a display. When the mirror display is off, it offers a normal reflection [2]. When the display is on, the mirror turns into a smart mirror, as shown in Figure 1. Still, even when the the display is on, the mirror offers some reflection. The information and functionality provided in the smart mirror varies from showing calendar reminders to monitoring the user activity, or even working as a home gateway. Mirrors of medium or small size cover part of the body, mainly face and shoulders, so face/gesture recognition is the key application as shown in [2], whereas full-body mirrors capable of estimating the body pose open the door to other applications. Among these applications, this paper focuses in the support for the performance of physical exercises. The idea is that the smart mirror assists older adults during the performance of an exercise routine, correcting and tracking the user movements. To do so, it is therefore necessary to equip the smart mirror with an RGB camera.

Fig. 1. H2020 SHAPES project's smart mirror device

The proposed solution for assisting physical rehabilitation routines is comprised of a list of exercises. These exercises are the most common ones for this purpose (physical activity in older adults). These exercises have been modelled so that the user performance while exercising is compared against master body poses, providing corrections whenever deviations are detected. The proposed solution also provides a management system in which, according to the user needs, the physical therapist prescribes the user with a set of exercises from the available list. These routines also include the number of iterations and the specific number of sessions during which they have to be performed. This is a rehabilitation program setup for a specific user. When the user or therapist starts a session, with the user in front of the mirror, the smart mirror provides first, voice and visual instructions about each exercise and second, by monitoring the body pose, it provides the user with feedback about the correct execution of the exercise, if needed. Finally, the smart mirror also generates periodic reports about the rehabilitation routines performed in front of the mirror (number of sessions, number of executions per exercise, degree of correctness, etc.).

The rehabilitation application of the smart mirror employs an RGB camera which gets a video stream from the user and feeds a processing workflow to extract the pose of the user, compares it with the exercise setup previously

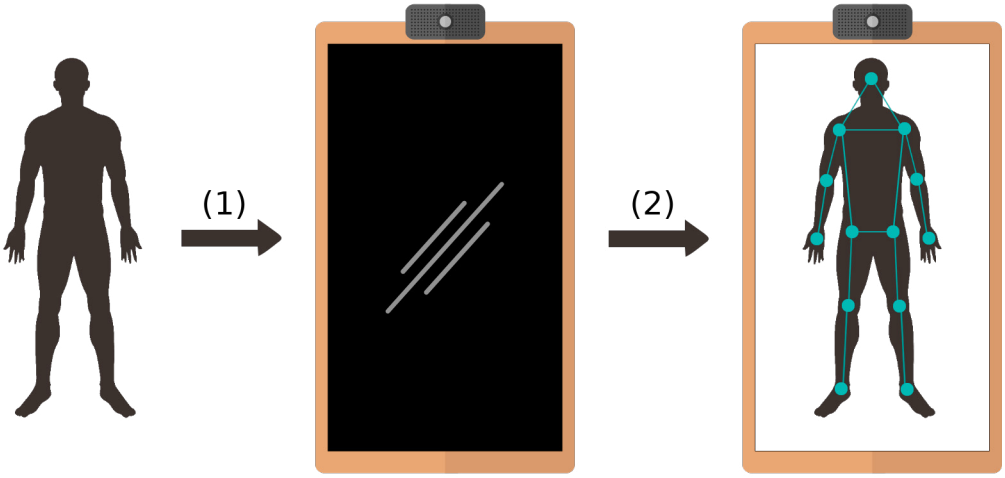


Fig. 2. Image processing carried out in the smart mirror. Two main steps can be identified in the overall process: (1) capture the image of the person in front of the mirror using the connected camera, and (2) process that image to retrieve the position of the joints of the person showing this output through the mirror display. This diagram has been designed using some resources from Freepik website [5]

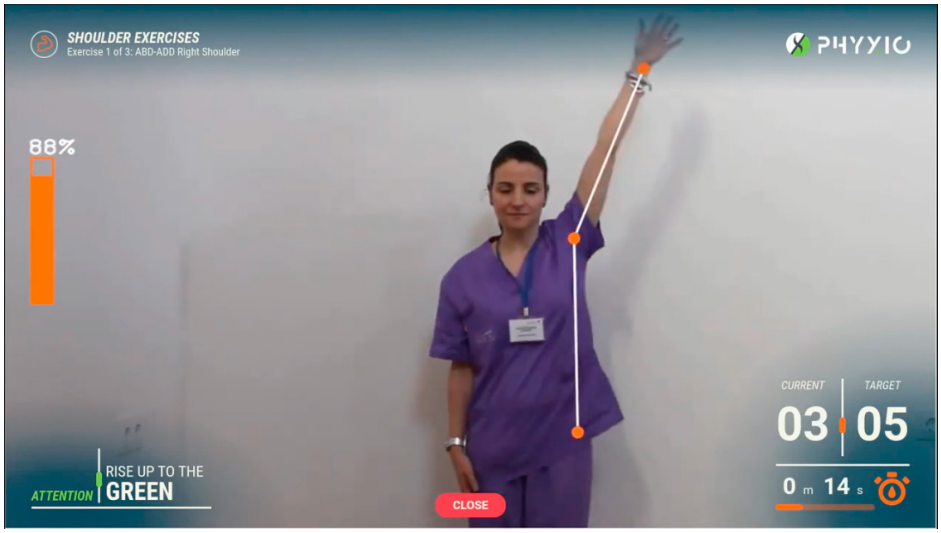


Fig. 3. Smart mirror application feedback

modelled, and show, in the smart mirror, the required corrections and show the count of the performed executions. This workflow is described in Figure 2, as well as the user feedback in Figure 3.

Pose estimation is, indeed, a challenge on its own due to the sensibility of the models to variations of the light, distance of the subject, camera quality, processing requirements and the vast literature about the topic. Nonetheless, the state of the art in video-based body pose estimation is sufficiently matured [20] to serve as support for other functionalities, built upon it. The work presented here is, indeed, built upon a body pose estimation functionality that

Fig. 4. Pose estimation on SHAPE's smart mirror

supports a physical rehabilitation system. The main contribution of this paper is to present the comparative analysis carried out to identify the body pose estimation models capable of performing on a smart mirror device. This analysis compares the estimation processes that each model (or library) propose, their relevant characteristics, the available toolchains and the final results they yield. The robustness of the pose estimation model to changing environments and correctness of the provided results will be determinant for the final decision. These aspects are considered under the optics of the peripherals and the computational capability of the embedded platform that it is the smart mirror device. It has to be noted that the limited physical space inside the smart mirror also impacts on the type of applications that can be deployed in such devices. Based on considerations such as efficiency, size, cost, and computing power the Raspberry Pi platform has been adopted inside the smart mirror device as the computing platform. Despite being resourcefully enough for most of the applications [16], video-based applications at real-time are among the most demanding ones and problems have been detected when performing tasks that involve analysing or showing video streams.

Figure 4 shows the key points that comprise the estimated pose of the person that appears in the image. This is the information that comprises the input for the physical rehabilitation support system. The capability of extracting key points representing body skeleton from human images has been a challenging computer-based vision problem [13]. This functionality is a cornerstone for many other applications, including situational awareness, human activity estimation, fall detection, training robots, etc.

This paper presents the comparative analysis that has been carried out on several vision-based body pose estimation models. This analysis has been undertaken with the endeavour of selecting the one with more advanced features capable of performing, in real time, in a constrained device as it is the Raspberry Pi 4 (8GB RAM version). This paper is organised as follows. First, the specific requirements for performing body pose estimation are described in Section 2. Then Section 3 presents the models that have been considered for this analysis. Section 4 describes the proposed methodology for comparing the considered models. Section 5 presents the obtained results while comparing with existing datasets and Section 6 presents the results while comparing models performing in the wild. Finally, Section 7 summarises the most relevant conclusions regarding the convenience of the different models for the purpose considered here.

2. Smart mirror requirements for the pose estimation model

The smart mirror application considered here is based on performing a continuous pose estimation from the video stream. The obtained poses are then compared to the modelled exercises. The modelled exercises (working as the master poses) represent the correct execution of the rehabilitation routine. In this sense, the estimated pose and the correct pose are then compared to detect deviations so that the user can be accordingly notified to correct the position. The physical rehabilitation system therefore depends on the accuracy and performance of the body-pose estimation

model in which it is built upon. In this sense, in the process of selecting the most appropriate pose estimation model, among the most well-known ones [13], the following aspects should be considered:

- Overall average precision is relevant although it is more important here to achieve a high precision on those parts of the body that are relevant for the rehabilitation routines.
- A trade-off between a fast inference and resource demanded by the model is needed as the system has to process images in real time and the computer is embedded in the smart mirror device, therefore being very limited in the available physical space.
- Robustness and flexibility is required as the proposed solution cannot rely on having to train the model for each different emplacement of the smart mirror device.

The proposed system does not count on a fit-for-purpose trained system. On the contrary, it is based on one-fits-all model capable of estimating the human body pose without having to adjust the model weights in every deployment of the system. It is also a functional requirement for the model to run as fast as possible so that the system can perform in real time.

3. Pose estimation models

In this article, three different models, and their variants, have been evaluated, first, against a well-known dataset for accuracy evaluation, and, second, against a previously recorded dataset of videos, which will conform the custom test for SHAPES, for speed and overall performance on in-the-wild environments. These models, all available in Tensorflow[1] JavaScript (JS) format[2], are: MoveNet, BlazePose, and PoseNet.

3.1. MoveNet

This model, based in the MobileNetV2 architecture [17], predicts human joint 2D locations from an RGB image. It has two different variants:

- **Lightning**[3]: It performs at high speed, perfect for real-time inference while achieving good performance.
- **Thunder**[4]: This variant obtains better results in terms of accuracy, but the speed is slightly penalised.

A skeleton will be obtained as inference output, which is represented by the one specified in Figure 5.a.

3.2. BlazePose

This model is part of the Pose solution of Mediapipe[5], which consists in executing two separate models within a pipeline:

- **BlazePose detector**: It detects the person ROIs, also known as *Region of Interest*.
- **BlazePose GHUM 3D**[19]: With the original image and the detected ROIs by the detector, it calculates 2D and 3D keypoints, with the latter being referenced to the hips center.

This model, as MoveNet does, has different variants:

- **Lite**: Similar to MoveNet Lightning, this variant is fast without penalising accuracy much.

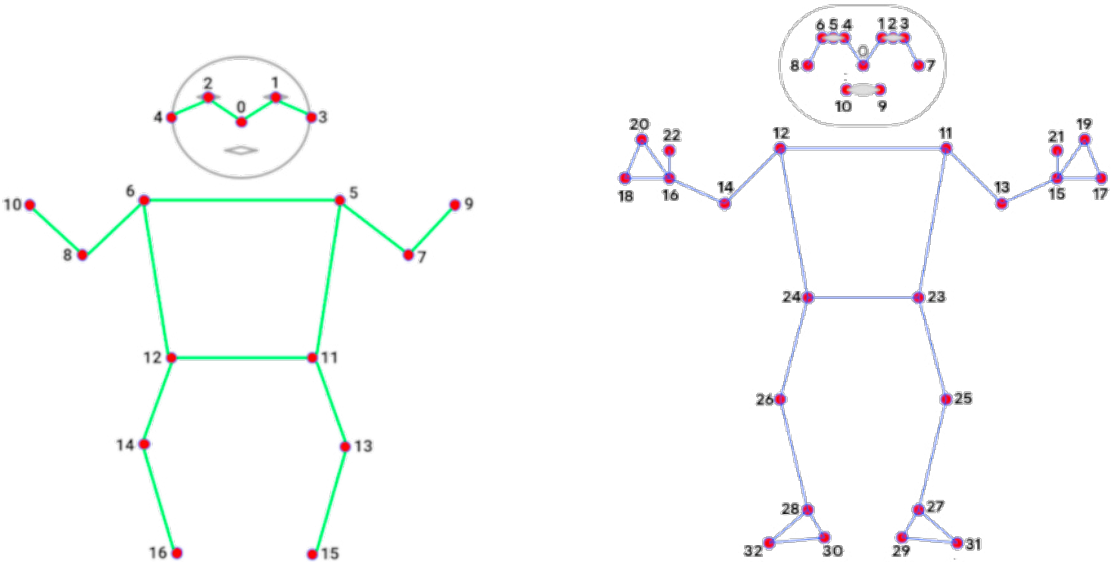
[1] https://www.tensorflow.org/
[2] https://github.com/tensorflow/tfjs-models/tree/master/pose-detection
[3] https://tfhub.dev/google/movenet/singlepose/lightning/4
[4] https://tfhub.dev/google/movenet/singlepose/thunder/4
[5] https://google.github.io/mediapipe/solutions/pose

- **Heavy**: Similar to MoveNet Thunder, it performs slower than the Lite version but with high accuracy.
- **Full**: A middle ground between Lite and Heavy.

This model output is a skeleton with the keypoints seen in the Figure 5.b.

3.3. PoseNet

The PoseNet version chosen for this article, by default in TensorFlow JS demos, is based MobileNetV1 architecture [9]. It can predict human joint 2D locations from up to 5 persons simultaneously, although for this paper only one person detection will be done, for comparing to the other models. Its output is a skeleton with the keypoints seen in Figure 5.a.



(a) MoveNet and PoseNet keypoints. (b) BlazePose keypoints.

Fig. 5. Model keypoint diagrams. Obtained from TensorFlow *tfjs-models* GitHub repository.

4. Comparison description

To compare the previously mentioned models, we used a well-known dataset: COCO. COCO dataset [12], also known as *Common Objects in Context*, is a large group of images which are annotated on separate files, which enables extracting data by computer vision technics. For the evaluation, the keypoint evaluation process[6] will be followed.

First, all COCO images containing a person, at least, must be processed by each one of the models and obtain their corresponding pose data in COCO result format[7]. For the set of images used, in this paper we focused on the keypoints 2017 val dataset, which can be downloaded, both images and their annotations, from the official website[8].

Then, the *upload_video* demo, available in the *tfjs-models* GitHub repository[9], was modified to use COCO images as input instead of a video file. Once all images were processed by the selected model, the detections are exported in a JSON file.

[6] https://cocodataset.org/#keypoints-eval
[7] https://cocodataset.org/#format-results
[8] https://cocodataset.org/#download
[9] https://github.com/tensorflow/tfjs-models

Finally, and with the generated JSON file, we execute *run_analysis* tool, from the *coco-analyze* [15] GitHub repository[10], to obtain some charts representing the models accuracy.

4.1. Metrics

COCOanalyze class, from *coco-analyze*, outputs a lot of different metrics starting from the base metrics[11], which are the average precision (AP) and the average recall (AR), in relation to a models accuracy.

This paper focuses on the predicted localization errors [15], which are:

Jitter	Small error around the correct keypoint location. The keypoint similarity between the detection and its ground-truth is within range [0.5, 0.85).
Miss	Large localization error, the detected keypoint is not within the proximity of any body part. The keypoint similarity between the detection and its ground-truth is below 0.5.
Inversion	Confusion between semantically similar parts belonging to the same instance. The detection is in the proximity of the true keypoint location of the wrong body part.
Swap	Confusion between semantically similar parts of different instances. The detection is within the proximity of a body part belonging to a different person.

Every keypoint detection having a keypoint similarity with its ground-truth that exceeds 0.85 is considered *good*. Threshold limits can be modified in the COCO framework, being 0.85 the "threshold above which also human annotators have a significant disagreement (around 30%) in estimating the correct position" ([15]).

Also, we mainly focused on the joints present in rehabilitation routines supported by the smart mirror. For example, any localization errors detected on the eyes or ears are not relevant to our research.

5. Results

We obtained, following the evaluation process described in Section 4, the localization errors for the models MoveNet, BlazePose and PoseNet, which could be seen in Figure 6, Figure 7 and Figure 8, respectively. For the first two models, only the light variant of each one were used, as, at least in this comparison, the results were not that different from their heavier siblings.

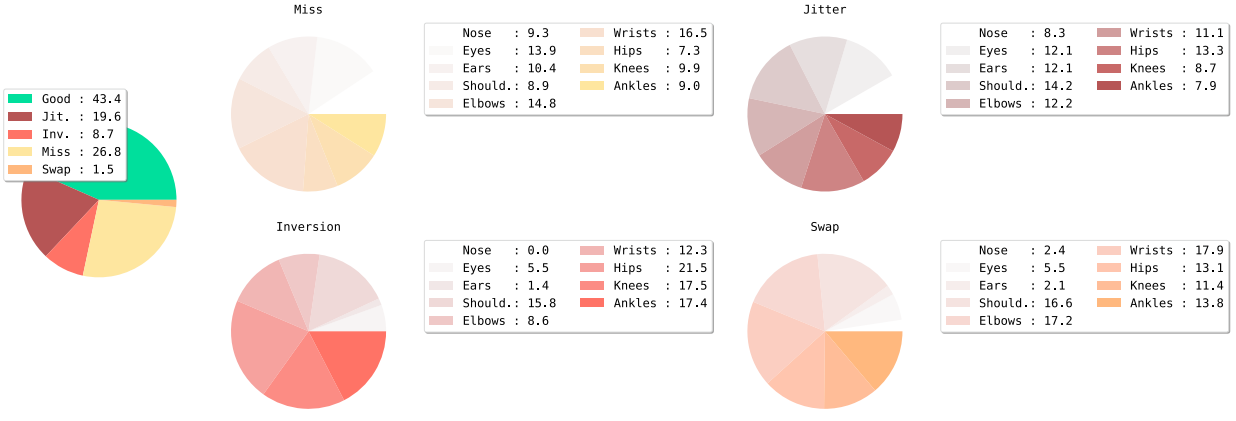


Fig. 6. Localization errors of MoveNet Lightning model.

[10] https://github.com/matteorr/coco-analyze
[11] https://cocodataset.org/#detection-eval

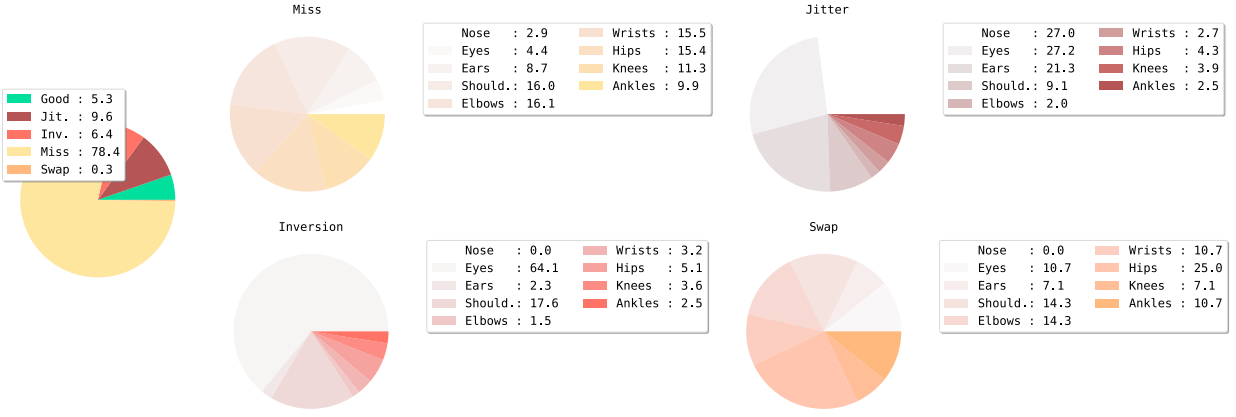


Fig. 7. Localization errors of BlazePose Lite model.

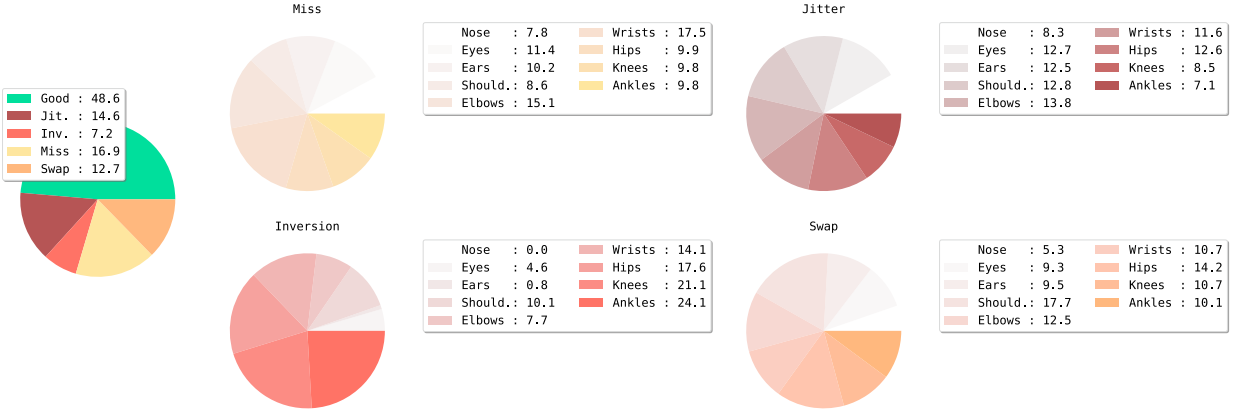


Fig. 8. Localization errors of PoseNet model.

As we can see in the charts, PoseNet offers the less amount of localization errors, with a **48.6%** of good predictions. There is an equal proportion of jitter, miss and swap errors, around 15%, that, for our relevant joints, mostly affects wrists and elbows.

Secondly, MoveNet has great results as well, with a **43.4%** of good predictions. Miss is the most common error, followed by the jitter. The wrists and elbows are also the joints more affected by the errors.

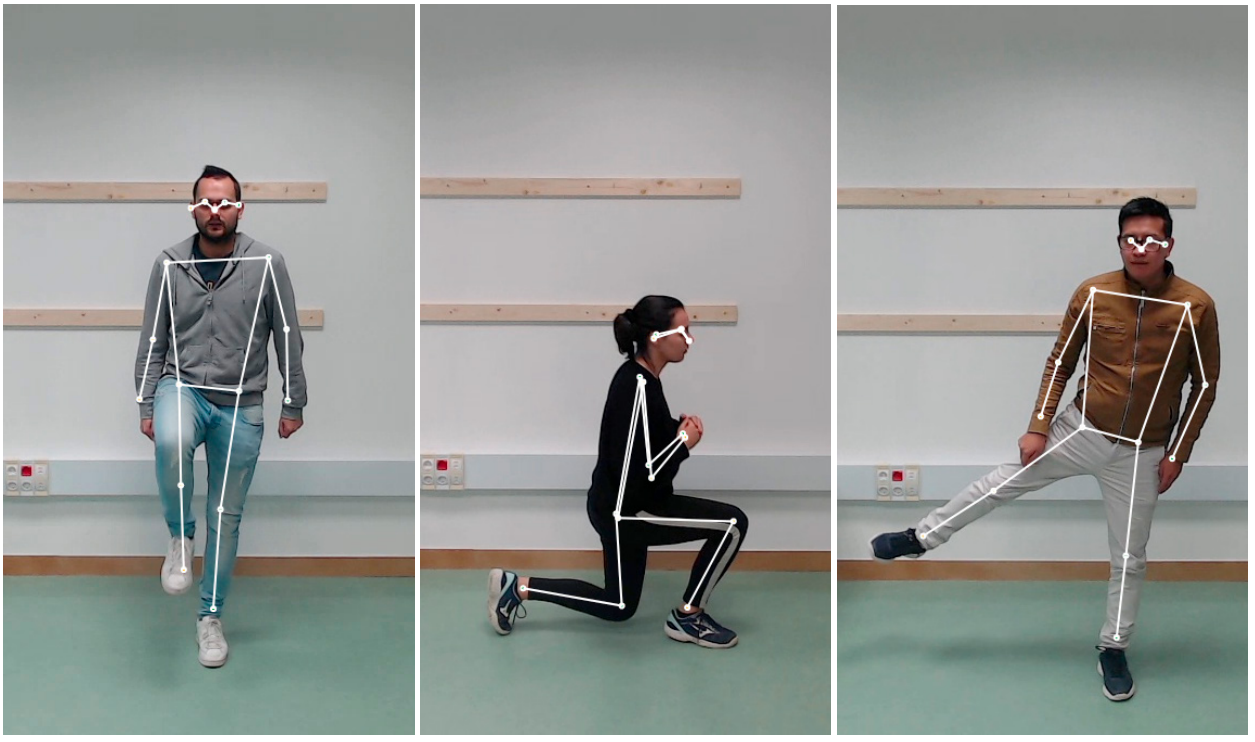
Last, but not least, BlazePose gave bad results, with only a **5.3%** of good predictions. Miss is the main component in its detections, with the wrists, elbows, hips and shoulders affected.

6. In the wild detections

The term ”in-the-wild” refers to the detection, or attempt, in an uncontrolled environment, as the model is not trained with this environment data. This is an important test for our smart mirror, as it could be located anywhere in a house or facility.

For this approach, we recorded videos of fitness exercises in our laboratory. The person is located between 1 and 1.5 meters away from the smart mirror camera. Some examples can be seen in the Figure 9.

In terms of speed, MoveNet Lightning and BlazePose Lite models are the fastest. They run at 2-5 FPS, also known as *frames per Second*, on the smart mirror and at 10-15 FPS on a usual PC, all using the web browser as we are running a JS demo.



(a) High knee (b) Lunge (c) Side leg raise

Fig. 9. MoveNet detections using our custom set of videos.

7. Conclusions and future work

Despite PoseNet having the best accuracy, compared against the COCO dataset, we choose MoveNet as the best of the three models, in terms of both speed and precision, as it offers a reliable and fast form of pose estimation.

The evaluated models are, at the moment, among the most advanced models for 2D body pose estimation. Nonetheless, the continuous advances on video-based 3D body pose estimation suggests that in a near future, new pre-trained models will be available that will improve not only the obtained accuracy but also the required computing capabilities. Having depth information will open the functionality provided by a physical rehabilitation systems as the corrections made by the system will not only be limited to a 2D plane, but it will also consider depth information. Despite the fact that BlazePose already provides this functionality, a the moment it accuracy is very limited, at least against the COCO dataset. In this sense, further improvements are needed for this model to be sucessfully integrated in the proposed system for physical rehabilitation.

Future work also considers evaluating the perception of the end user in terms of the perceived accuracy. Users will be questioned about the correctness of the received corrections from the system. These corrections are related to the body pose estimation made by the system.

Finally, creating a custom 3D dataset from our captured fitness data would allow us to test models accuracy on in-the-wild environments. This can be achieved in future work by annotating each image, and considering it as ground-truth, using a motion capture technique, such as marker/sensor based or multi-view.

Acknowledgements

This research was funded by H2020 European Union program under grant agreement No. 857159 (SHAPES project) and by MCIN/ AEI /10.13039/501100011033 under grant TALENT-BELIEF (PID2020-116417RB-C44).

References

[1] Andreu, Y., Chiarugi, F., Colantonio, S., Giannakakis, G., Giorgi, D., Henriquez, P., Kazantzaki, E., Manousos, D., Marias, K., Matuszewski, B.J., Pascali, M.A., Pediaditis, M., Raccichini, G., Tsiknakis, M., 2016. Wize mirror - a smart, multisensory cardio-metabolic risk monitoring system. Computer Vision and Image Understanding 148, 3–22. URL: https://www.sciencedirect.com/science/article/pii/S1077314216300224, doi:https://doi.org/10.1016/j.cviu.2016.03.018. special issue on Assistive Computer Vision and Robotics - "Assistive Solutions for Mobility, Communication and HMI".

[2] Chaparro, J.D., Ruiz, J.F.B., Romero, M.J.S., Peño, C.B., Irurtia, L.U., Perea, M.G., Garcia, X.d.T., Molina, F.J.V., Grigoleit, S., Lopez, J.C., 2021. The shapes smart mirror approach for independent living, healthy and active ageing. Sensors 21. URL: https://www.mdpi.com/1424-8220/21/23/7938, doi:10.3390/s21237938.

[3] Erazo, O., Pino, J.A., Pino, R., Asenjo, A., Fernández, C., 2014. Magic mirror for neurorehabilitation of people with upper limb dysfunction using kinect, in: 2014 47th Hawaii International Conference on System Sciences, pp. 2607–2615. doi:10.1109/HICSS.2014.329.

[4] Fischer, S.H., David, D., Crotty, B.H., Dierks, M., Safran, C., 2014. Acceptance and use of health information technology by community-dwelling elders. International Journal of Medical Informatics 83, 624–635. URL: https://www.sciencedirect.com/science/article/pii/S1386505614001063, doi:https://doi.org/10.1016/j.ijmedinf.2014.06.005.

[5] Freepik, 2022. Freepik website. https://www.freepik.es/. Accessed: 2022-04-25.

[6] Gomez-Carmona, O., Casado-Mansilla, D., 2017. Smiwork: An interactive smart mirror platform for workplace health promotion, in: 2017 2nd International Multidisciplinary Conference on Computer and Energy Science (SpliTech), pp. 1–6.

[7] H2020 European Union program, 2022. Shapes project website. https://shapes2020.eu/. Accessed: 2022-04-22.

[8] Henriquez, P., Matuszewski, B.J., Andreu-Cabedo, Y., Bastiani, L., Colantonio, S., Coppini, G., D'Acunto, M., Favilla, R., Germanese, D., Giorgi, D., Marraccini, P., Martinelli, M., Morales, M.A., Pascali, M.A., Righi, M., Salvetti, O., Larsson, M., Strömberg, T., Randeberg, L., Bjorgan, A., Giannakakis, G., Pediaditis, M., Chiarugi, F., Christinaki, E., Marias, K., Tsiknakis, M., 2017. Mirror mirror on the wall... an unobtrusive intelligent multisensory mirror for well-being status self-assessment and visualization. IEEE Transactions on Multimedia 19, 1467–1481. doi:10.1109/TMM.2017.2666545.

[9] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 .

[10] Johri, A., Jafri, S., Wahi, R.N., Pandey, D., 2018. Smart mirror: A time-saving and affordable assistant. 2018 4th International Conference on Computing Communication and Automation (ICCCA) , 1–4.

[11] Kleinberger, T., Becker, M., Ras, E., Holzinger, A., Müller, P., 2007. Ambient intelligence in assisted living: Enable elderly people to handle future interfaces, in: Stephanidis, C. (Ed.), Universal Access in Human-Computer Interaction. Ambient Interaction, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 103–112.

[12] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European conference on computer vision, Springer. pp. 740–755.

[13] Munea, T.L., Jembre, Y.Z., Weldegebriel, H.T., Chen, L., Huang, C., Yang, C., 2020. The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation. IEEE Access 8, 133330–133348. doi:10.1109/ACCESS.2020.3010248.

[14] Purohit, N., Mane, S., Soni, T., Bhogle, Y., Chauhan, G., 2019. A computer vision based smart mirror with virtual assistant, in: 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 151–156. doi:10.1109/ICCS45141.2019.9065793.

[15] Ronchi, M.R., Perona, P., 2017. Benchmarking and error diagnosis in multi-instance pose estimation, in: The IEEE International Conference on Computer Vision (ICCV).

[16] Sahana, S., Shraddha, M., Phalguni, M.P., Shashank, R.K., Aditya, C.R., Lavanya, M.C., 2021. Smart mirror using raspberry pi: A survey, in: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 634–637. doi:10.1109/ICCMC51019.2021.9418408.

[17] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520.

[18] Silapasuphakornwong, P., Uehira, K., 2021. Smart mirror for elderly emotion monitoring, in: 2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech), pp. 356–359. doi:10.1109/LifeTech52111.2021.9391829.

[19] Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C., 2020. Ghum & ghuml: Generative 3d human shape and articulated pose models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6184–6193.

[20] Zheng, C., Wu, W., Yang, T., Zhu, S., Chen, C., Liu, R., Shen, J., Kehtarnavaz, N., Shah, M., 2020. Deep learning-based human pose estimation: A survey. arXiv preprint arXiv:2012.13392 .