Remote Physical Therapy: Requirements for a Single RGB Camera Motion Sensing

1st Jindrich Adolf
CIIRC
Czech Technical University in Prague
Prague, Czech Republic
jindrich.adolf@cvut.cz

2nd Jaromir Dolezal
CIIRC
Czech Technical University in Prague
Prague, Czech Republic
jaromir.dolezal@cvut.cz

3rd Martin Macas
CIIRC
Czech Technical University in Prague
Prague, Czech Republic
martin.macas@cvut.cz

4th Lenka Lhotska
CIIRC
Czech Technical University in Prague
Prague, Czech Republic
lenka.lhotska@cvut.cz

***Abstract*—The aim of this work is to assess the minimal technical requirements for using a simple RGB camera for motion sensing in a standard home environment. We experimentally verify the recording requirements for subsequent motion analysis. Our work contributes to the development of physical telerehabilitation without the need to use special HW and thus enable telerehabilitation for the general public, which is especially important during the COVID-19 lockdown. We have found out that such a system can work surprisingly well even with a low-cost camera in poor recording conditions and slow 3G internet connection.**

***Keywords*—Telerehabilitation, Physical therapy, OpenPose, RGB camera, Webcam**

I. Introduction

Computer vision and image recognition using deep neural networks have made a large step forward in recent years. Progress has also been made in body pose estimation from the simple RGB image. Nowadays, it is possible to use a standard webcam to detect the human body, recognize the position of the joints, and then evaluate body movement. This method can be used to analyze the movement of athletes, but it could be also used in health care or physical rehabilitation. For its simplicity, this method can also be used in a home environment. This research experimentally verifies what conditions need to be met for recording with a common webcam at home to use the method for motion analysis.

II. Related work

Conventional MoCap systems capture the human body in a three-dimensional space. Commercial systems such as Vicon or Qualisys are considered as the ground truth for the pose estimation. To summarize, these systems are very accurate, but also costly and difficult to setup up. This limits their use to special clinical workplaces only [1].

Another approach is to use inertial sensors. These systems are less expensive, but they suffer from drift and still require special HW. They could be used in home environments, but their accuracy allows only activity monitoring, but not a quality assessment due to large absolute position errors [2].

Other technologies that are closer to home use are RGB-D systems, such as MS Kinect or Nintendo Wii. Their use in rehabilitation is described by Levene in his review [3]. The review discusses the basic properties of RGB-D systems and their applicability in rehabilitation. These systems are suitable for home rehabilitation, but the need for a dedicated HW limits their practical use and effectively prevents their mass deployment.

In our research, we deal with recordings using only one RGB camera. We only work with a 2D model which brings some limitations which are further discussed in this paper, but such a system can be used anywhere by anyone without additional costs.

We do not aim to compete with the accuracy of 3D systems. We deal with applications where it is necessary to recognize the body pose of the trainee, but there is no need for clinical accuracy.

Such applications include, for example, physical rehabilitation in households after limb injuries, or routine exercises for the elderly to maintain mobility and prevent falls. Nakano examined the accuracy of the RGB based system in his research. He used Vicon as a reference system [4]. His research shows that the accuracy of detecting key points using standard cameras from a distance of several meters is about a few centimeters. For use in home rehabilitation, even a few centimeters accuracy is sufficient.

Single camera systems utilize image recognition using deep neural networks. The first full-body sensing system using one RGB camera was the DeepPose system published by Google in 2014 [5]. Another important publication was the publication of Pfister et al. from Oxford University in 2015 [6]. This was followed by the work of Cao from CMU in 2017 [7] and the work of the same author from 2019 [8]. These two works resulted in the implementation of the OpenPose system.

OpenPose is published with source code, allowing

2021 International Conference on Applied Electronics (AE) | 978-80-261-0973-0/21/$31.00 ©2021 10.23919/AE51540.2021.9542912

ISBN 978-80-261-0973-0, © University of West Bohemia, 2021

academia to apply the algorithms in their research. The system can be also used with several synchronized calibrated cameras to create a 3D model, but this is not a preferred way in a home environment. The OpenPose system is not the only one computer vision pose detection system, there are other alternatives such as PoseNet [9] or wrnchAI [10].

These systems require a lot of computing power. For our experiments, we use Nvidia RTX2070 running OpenPose with just 18fps. However, the HW required for evaluation can be placed in the server and shared by many clients, while the client will only need a device capable of streaming video such as a smart phone, tablet, computer or smart TV.

III. Methods

The aim of our research is to simulate common situations that may occur during recording at home and to determine what are the requirements for a single RGB camera motion sensing.

The most common problems are caused by the incorrect position of the subject on the camera, poor resolution of the camera, insufficient lighting, and low quality of the transmitted video. To determine the effects of these typical practical problems, we have performed an experimental measurement. Experimental recording occurred in a typical home environment. We have used several cameras at once to record a person performing a typical whole body physical exercise.

The studied effects can be divided in two categories:

Effect of subject to camera position

In the first part, the person was recorded by multiple cameras at the same time. Each camera has a different relative position in the space between the trained person and the imaging camera.

Effect of video signal quality

In the second part, we studied the effect of camera resolution, illumination, and encoding quality. For this, we used a record created by the camera "REF CAM"

We have used the same evaluation procedure for both categories. Please see the block diagram description in Figure 1.

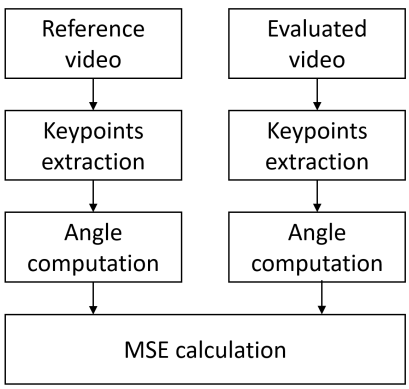

Fig. 1. Evaluation procedure

The videos are processed frame by frame by Openpose to extract the keypoints. The keypoints are used to compute the relevant angles, see Figure 2 with skeleton model.

Each angle is calculated using three keypoints as shown in Table I.
The angle between keypoints A,B,C is computed as follows:

$$\alpha = atan2(C_y - B_y, C_x - B_x) - atan2(A_y - B_y, A_x - B_x))$$

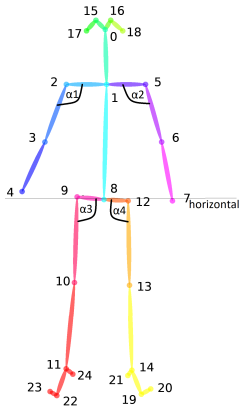

Fig. 2. Skeleton with selected angles

TABLE I
Selected body joints

	Anatomical part of the body	Selected keypoints
α_1	Right elbow	1,2,3
α_2	Left elbow	1,5,6
α_3	Right hip	8,9,10
α_4	Left hip	8,12,13

To obtain comparable values, we have normalized the MSE by the reference signal power as follows:

$$NMSE(x, y) = MSE(x, y)/MSE(x, 0)$$

A. Effect of subject to camera position

To take advantage of recording with only one RGB camera and get correct anatomical angles, the movements must be captured in their corresponding plane. To capture physical exercise with abduction and adduction movements that happen in the anatomical frontal plane, the subject must stand directly on the camera.

To simulate common errors in the subject to camera position, we have used 4 synchronized FLIR Blackfly S cameras. We have placed the reference camera "REF CAM" at the height of the subject chest directly facing the subject. We have placed the second camera directly facing the subject but shifted 10° vertically. The third camera and the fourth camera were placed at the same height as the reference camera but shifted 15° and 30° horizontally, respectively, see Figure 3 and Figure 4.

Fig. 3. Four camera views from different angles

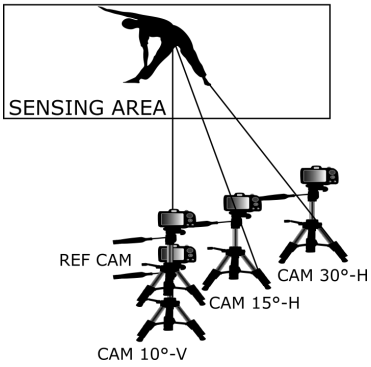

Fig. 4. Four camera setup

B. *Effect of video signal quality*

For this, we used one of the FLIR Blackfly S cameras with Fujinon 3 MP Varifocal lenses, the "REF CAM", please see Figure 4.

1) Effect of resolution: To investigate the effect of camera resolution, we have repeatably halved the resolution and computed the MSE between the reference and resized video. The resolution of the original record was 1280x720, and the resolutions of the resized video were 640x360, 320x180, 160x90, and 80x45.

2) Effect of illumination: The recording condition in the home environment can vary. We have simulated the effect of illumination using gamma correction. We have evaluated gamma correction of 0.1, 0.25, 0.5, 2, 4 and 8.

3) Effect of bitrate: To simulate problems with low internet speed, we have performed a bitrate reduction. The original bitrate was 643kb/s and we have reduced it to 500, 300, 100, 50 and 10kb/s.

IV. Results

All results are quantified by Mean Square Error (MSE) normalized by the reference signal power. Normalized MSE values close to 0 means that the signals are almost identical and values close to 1 in our case usually reflect that there is no body pose detection at all. Based on our experience, normalized MSE values of less than 0.1 are acceptable for telerehabilitation. The reference for the MSE calculation is a record from the camera "REF CAM" placed in an axis perpendicular to the person at half of the subject's chest. This unmodified record was used as an reference for all comparisons.

A. *Effect of subject to camera position*

Recordings from all cameras were synchronized and taken at the same time. As can be seen in Table II, the position of the subject on the camera significantly affects the NMSE. Minor deviations from the optimal angle will still allow to provide feedback, but it will not be possible to evaluate the anatomical angle with clinical precision. Therefore, the subject should face the camera as directly as possible.

TABLE II
NMSE DEPENDING ON CAMERA POSITION

	CAM 10° − V	**CAM 15° − H**	**CAM 30° − H**
α_1	0.1017	0.136	0.1563
α_2	0.0559	0.125	0.1661
α_3	0.0053	0.009	0.0175
α_4	0.0069	0.0222	0.046

B. *Effect of resolution*

Our results indicate that a standard-definition 240p video quality is acceptable. The system can work even with lower resolution, see Figure 5. In practice, there is no reason to use such a low resolution and any streaming service will use a better one.

TABLE III
NMSE DEPENDING ON RESOLUTION

RESOLUTION	**640x360**	**320x180**	**160x90**	**80x45**
α_1	0.0021	0.0054	0.0624	0.7768
α_2	0.0036	0.0099	0.127	0.88
α_3	0.0013	0.0019	0.0032	0.783
α_4	0.002	0.005	0.0039	0.773

For example, resolution images, please see Figure 4.

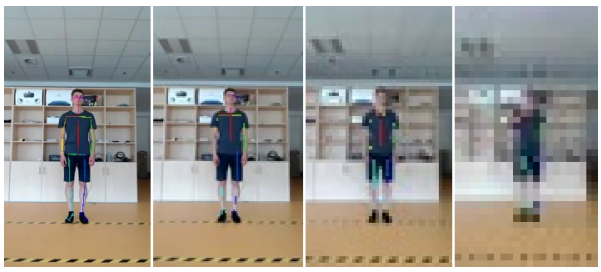

Fig. 5. Four resolution 640x360, 320x180, 160x90, 80x45

C. *Effect of illumination*

Based on our simulation, we can say that the OpenPose algorithm is not sensitive to illumination. As long as the subject can see himself/herself in the video, the detection will work. The gamma value of 0.1 is not shown in the Figure 6 as it would appear just black.

TABLE IV
NMSE DEPENDING ON ILLUMINATION

GAMA	8	4	2	0.5	0.25	0.1
α_1	0.054	0.0021	0.0014	0.0043	0.035	1
α_2	0.0224	0.0056	0.0034	0.0473	0.344	0.994
α_3	0.0021	0.00019	0.0017	0.0014	0.0017	1
α_4	0.0031	0.0025	0.0020	0.0031	0.0023	1

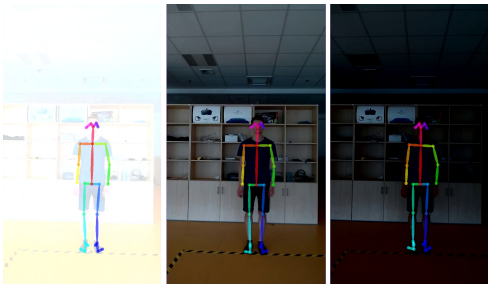

Fig. 6. Gama correction 8, 0.5, 0.25

D. Effect of bitrate

We have used the h264 codec for the test. Due to the fact that the camera captures the movement of the trainee on a static background, it could work quite well using a very low bitrate value. The system would arguably work even on a dial-up connection.

TABLE V
NMSE DEPENDING ON BITRATE

BITRATE	500kb/s	300kb/s	100kb/s	50kb/s	10kb/s
α_1	0.0007	0.001	0.0019	0.0033	0.007
α_2	0.0013	0.0091	0.037	0.0212	0.014
α_3	0.00086	0.0011	0.0015	0.0017	0.0308
α_4	0.0010	0.0015	0.0023	0.0027	0.0580

V. DISCUSSION AND CONCLUSION

The aim of the work was to experimentally verify what conditions need to be met for the practical use of the OpenPose system in home-based telerehabilitation.

As mentioned in the introduction, if we want to record people with only one camera, we have to accept lower accuracy compared to professional MoCap systems.

On the other hand, the OpenPose based telerehabilitation system has sufficient accuracy to evaluate whole body exercise and can provide valuable feedback to the patient as well as to the physiotherapist.

A telerehabilitation system using only one RGB camera can be a very valuable tool for all patients who have no choice but to rehabilitate only in a home environment.

We can summarize that OpenPose is a very robust algorithm and even in very poor recording conditions there are only minimal differences in the calculation of the relevant anatomical angles.

The only conditions that must be met are that the whole body must be in view of the camera, the subject must be directly facing the camera, and the body must be recognizable with the naked eye.

ACKNOWLEDGMENT

Research has been supported by the Grant Agency of the Czech Technical University in Prague, grant $No.SGS20/214/OHK3/3T/37$ "Active learning of time series models" and by the project $No.FV-20696$ "Personal health monitoring and assistive systems", funded by the TRIO program of the Ministry of Industry and Trade of the Czech Republic.

REFERENCES

[1] F. Schlagenhauf, S. Sreeram, and W. Singhose, "Comparison of kinect and vicon motion capture of upper-body joint angle tracking," in *2018 IEEE 14th International Conference on Control and Automation (ICCA)*. IEEE, Jun. 2018. [Online]. Available: https://doi.org/10.1109/icca.2018.8444349

[2] T. Maruyama, M. Tada, A. Sawatome, and Y. Endo, "Constraint-based real-time full-body motion-capture using inertial measurement units," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018, pp. 4298–4303.

[3] T. Levene and R. Steele, "The quantified self and physical therapy," in *Proceedings of the International Conference on Compute and Data Analysis - ICCDA '17*. ACM Press, 2017. [Online]. Available: https://doi.org/10.11452F3093241.3093272

[4] N. Nakano, T. Sakura, K. Ueda, L. Omura, A. Kimura, Y. Iino, S. Fukashiro, and S. Yoshioka, "Evaluation of 3d markerless motion capture accuracy using OpenPose with multiple video cameras," Nov. 2019. [Online]. Available: https://doi.org/10.1101/842492

[5] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2014. [Online]. Available: https://doi.org/10.1109/cvpr.2014.214

[6] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1913–1921.

[7] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1302–1310.

[8] 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.

[9] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial PoseNet: A structure-aware convolutional network for human pose estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017. [Online]. Available: https://doi.org/10.1109/iccv.2017.137

[10] P. Kruszewski and T. J. Mahamad, "The AI powered magic mirror," in *ACM SIGGRAPH 2018 Virtual, Augmented, and Mixed Reality*. ACM, Aug. 2018. [Online]. Available: https://doi.org/10.1145/3226552.3226569