Range of Motion Assessment using a Digital Voice Assistant

Hassam Khan Wazir, Kshitij Gaikwad, and Vikram Kapila

***Abstract*— Range of motion (ROM) is an important indicator of an individual's physical health, and its degradation impacts their ability to perform activities of daily living. The elderly are particularly susceptible to mobility-loss due to muscular decline, neuromuscular disorders, sedentary lifestyle, etc. Thus, they must undergo periodic ROM assessments to track their physical well-being and consult doctors for any decline in ROM. An at-home ROM assessment device can assist the elderly to self-perform ROM assessment and facilitate remote monitoring of and compliance to therapy. The pervasive adoption of digital voice assistants (DVAs), that include a monocular camera, offers an opportunity for at-home ROM assessment. This paper proposes using a DVA for ROM measurement by utilizing 2D pose estimation techniques to estimate 3D limb pose for specific exercises. The system employs the MediaPipe library to perform 2D pose estimation and uses the joint coordinates to find the 3D pose of the limb using a 2D projection method. To validate the system, it is first compared with a 3D human model performing various shoulder and elbow exercises in a virtual environment. Next, for further validation, a neurologically intact individual performs the same exercises and the results of the proposed system are compared with the results from a markerless optical motion capture system (Kinect). The Bland-Altman limits of agreement (LOA) are computed and provided for the two sets of comparisons. The results demonstrate the feasibility of the proposed system in providing reliable ROM measurements using a DVA and suggest possible enhancements.**

***Clinical relevance*— This paper introduces the idea of ROM measurement using digital voice assistants embedded with a monocular camera.**

I. INTRODUCTION

With the advent of smart devices, high-speed internet, and digital services, many hospitals and clinics have embraced telemedicine and digital healthcare as a viable service to the public [1]. In the context of the Coronavirus-2019 pandemic, a large majority of people in the US have started to receive virtual clinical services [2]. Currently, such services entail teleconferencing software, making the interactions a virtual analog of an in-person visit to a doctor's office. However, other aspects of telehealth are yet to make their transition to the digital domain. One such aspect is data-driven telemedicine and therapy compliance monitoring in elderly care. A paradigm shift is waiting to unfold, since the technology to make it happen is already available, albeit used for other purposes. Digital voice assistants (DVAs) are one example of such a technology. The last decade has seen a growth in the adoption of DVAs, with 46% of the US population already using some kind of DVAs as of 2017 [3]. Recently, DVAs have started to ship with built-in cameras, which greatly increases the potential of these devices. Yet, today the DVAs are used in a limited capacity, e.g., video conferencing, information retrieval, smart home control, etc.

As indicated previously, in recent years, discussions about digital healthcare and telehealth have come to the forefront [4] with an emphasis on geriatric health [5]. Even as the use of DVAs as a modality for telehealth and telemedicine remains a nascent concept, efforts are afoot to explore their utility and effectiveness for synchronous and asynchronous healthcare delivery [6], in remote healthcare monitoring [7], and as a conversational agent for the elderly [8]. The recent integration of built-in cameras in the DVAs offers yet another opportunity to embed novel modalities for therapy compliance into the DVA device ecosystem. Thus, this paper proposes a DVA application for the ROM measurement, which is traditionally performed with sensor-based [9] or optical [10] methods. We demonstrate the use of the built-in camera of a DVA device for ROM monitoring. This proof-of-concept system can be explored further to develop more effective telehealth solutions involving the DVA devices, e.g., for remote monitoring of compliance with physical therapy.

The paper is organized as follows. Section 2 elaborates on the design and development of the system used for ROM tracking. Section 3 discusses the results based on the preliminary work and addresses the benefits and shortcomings of the approach used. Finally, section 4 provides concluding remarks and suggests directions for future research.

II. DESIGN AND DEVELOPMENT

Our prototype is based on the *Google Nest Hub Max* DVA device. A computer acts as a local server and uses the Nest API to access the live stream of the device's camera through a Real-Time Streaming Protocol. An *if-this-then-that* applet accesses the Google Assistant in the DVA to respond to a user's voice commands. When the user issues a command, such as, "*OK Google, start ROM app*", the applet sends a GET request using the Flask web framework to the server. Once the command is acknowledged by the server, it starts receiving the live stream from the DVA camera, and the ROM measurement application starts on the server. A confirmation is sent to the user as a voice response by the Google Assistant and the user can start performing the exercise. Fig. 1 shows the overview of the proposed system and various alternative ways in which an overall system can be constructed.

We consider a 3D pose estimation problem with a subject standing at a fixed distance from the DVA camera. It is assumed that the subject's frontal plane (see Fig. 2) is parallel to the camera image plane (see Fig. 1). To determine the 3D pose, 33 image coordinates (i.e., landmarks) in $\mathbb{R}^2$ corresponding to the body joints are detected using the MediaPipe

Authors are with the Mechanical and Aerospace Engineering Department, NYU Tandon School of Engineering, Brooklyn, NY 11201, USA (corresponding author: 646-997-3161; e-mail: vkapila@nyu.edu).

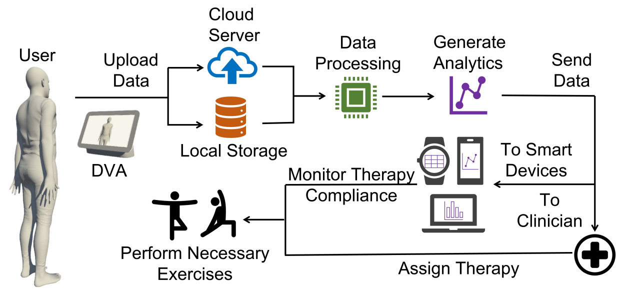


Fig. 1: Overview of the proposed system.

library that utilizes the BlazePose [11] pose detection model. For studying the upper extremity ROM, we use eight of the 33 landmarks from the MediaPipe. Although the model provides the landmarks in $\mathbb{R}^3$, the z-coordinate is discarded as it has a high variance in error when consecutive tests are carried out for the same pose.

A. Shoulder ROM measurement using camera image

Glenohumeral joint motion is modelled using spherical coordinates [12] where the joint is at the center. We consider that the wrist skirts the periphery of a virtual sphere. The three angles of interest are the thoracohumeral angle, i.e., the shoulder elevation angle θ, the clavicular-humeral angle projected on the transverse plane, i.e., the shoulder plane angle ϕ, and the humeral rotation around the long axis, i.e., the internal-external rotation angle ψ. The notation used and measurements performed for the right arm in this paper can be analogously extended to the left arm. Fig. 2 depicts the angles and their relation to the cardinal planes.

1) Calibration Step: First, the user stands in front of the DVA camera with arms down to the side in a "rest pose". The system calculates the length of the trunk vector $\vec{T}_N \triangleq \hat{S}_B - \hat{S}_M$, the maximum length of the right arm vector $\vec{A}_R \triangleq \hat{W}_R - \hat{S}_R$, and the maximum length of the right forearm vector $\vec{F}_R \triangleq \hat{W}_R - \hat{E}_R$ in pixel space. These calculations are performed for $N = 50$ samples and the length averages $T_{N_{\text{avg}}}$, $A_{R_{\text{avg}}}$, and $F_{R_{\text{avg}}}$ are computed to mitigate any errors due to noise. The angle θ (or ψ) calculation is performed using the ratio of the length of the arm (or forearm) to the length of the trunk for the rest pose in which it is maximum and denoted as the *rest ratio* a_r (or f_r) shown below

$$a_r = \frac{A_{R_{\text{avg}}}}{T_{N_{\text{avg}}}} \quad \text{and} \quad f_r = \frac{F_{R_{\text{avg}}}}{T_{N_{\text{avg}}}}. \tag{1}$$

This step helps overcome any changes in the apparent arm (or forearm) length due to a shift in perspective. Next, for calculating ϕ, we compute the *stretch ratio* a_s with the user arm stretched to the side, yielding $a_s = a_r|_{\theta=90°,\phi=0°}$.

2) Calculation step: We now determine the shoulder and elbow angles corresponding to various ROM exercises. We begin by noting that for two arbitrary vectors $\vec{v}_i = \hat{y}_i - \hat{x}_i$, $i = 1, 2$, the point of projection $\hat{p}$ of the vector $\vec{v}_2$ on the vector $\vec{v}_1$ is given by

$$\hat{p} = \text{proj}_{\vec{v}_1} \vec{v}_2 + \hat{x}_1 = \left(\frac{\vec{v}_1 \cdot \vec{v}_2}{\|\vec{v}_1\|^2} \right) \vec{v}_1 + \hat{x}_1. \tag{2}$$

Using (2), we can now determine the following quantities: (*i*) the vertical and horizontal projections of the elbow $\hat{E}_R$ on the base vector $\vec{S}_B$ and the trunk vector $\vec{T}_N$, denoted as $\hat{E}_x$ and $\hat{E}_y$, respectively; (*ii*) the vertical and horizontal projections of the wrist $\hat{W}_R$ on $\vec{S}_B$ and $\vec{T}_N$, denoted as $\hat{W}_x$ and $\hat{W}_y$, respectively; and (*iii*) the vertical projection of the shoulder $\hat{S}_R$ on $\vec{S}_B$, denoted as $\hat{S}_x$. See Fig. 2 that shows the projections $\hat{E}_x$, $\hat{E}_y$, $\hat{W}_x$, $\hat{W}_y$, and $\hat{S}_x$ for the right arm.

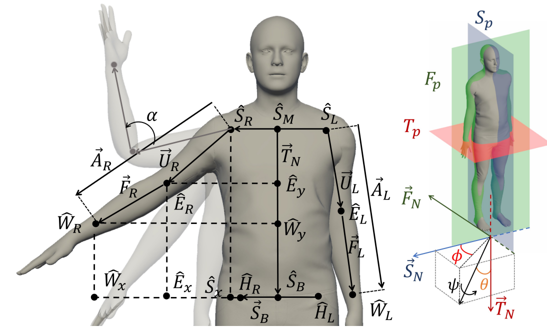


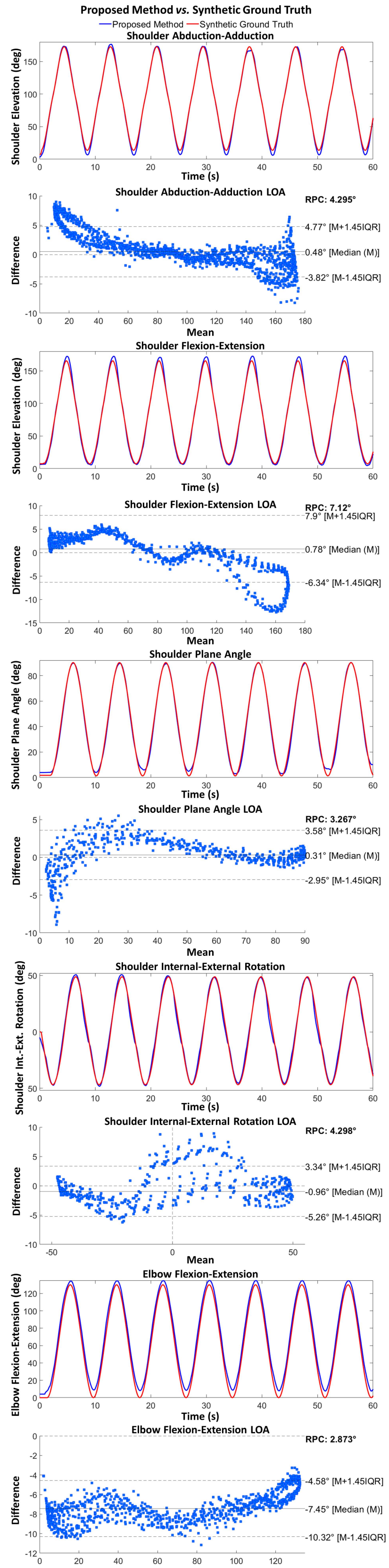
Fig. 2: Joint positions obtained from the MediaPipe and the vectors used to compute the joint angles: $\hat{\cdot}$ and $\vec{\cdot}$ denote points and vectors, respectively.

It can be shown that for a 2D projection of a vector rotating inside a sphere with radius equal to the length of the rotating vector, the change in the distance from the vertical projection point of a vector ($\hat{E}_y$ or $\hat{W}_y$) to $\hat{S}_M$ is directly proportional to $\cos(\theta)$. Thus, θ can be calculated for the shoulder abduction-adduction and flexion-extension exercises as shown in (3). With the user at a constant distance from the DVA camera and when $\theta = 90°$, ϕ can similarly be calculated as shown in (3). For the shoulder internal-external rotation exercise, let the upper arm point downwards (i.e., $\theta = 0°$), and flex the elbow to 90° (i.e., elbow angle $\alpha = 90°$). Then the internal-external rotation (i.e., shoulder angle ψ) can be computed as shown in (3). Finally, for the elbow flexion-extension exercise, the arm is kept parallel to the camera image plane (i.e., $\phi = 0°$) and the elbow angle α, between the upper arm $\vec{U}_R$ and forearm $\vec{F}_R$ is calculated as shown in (3).

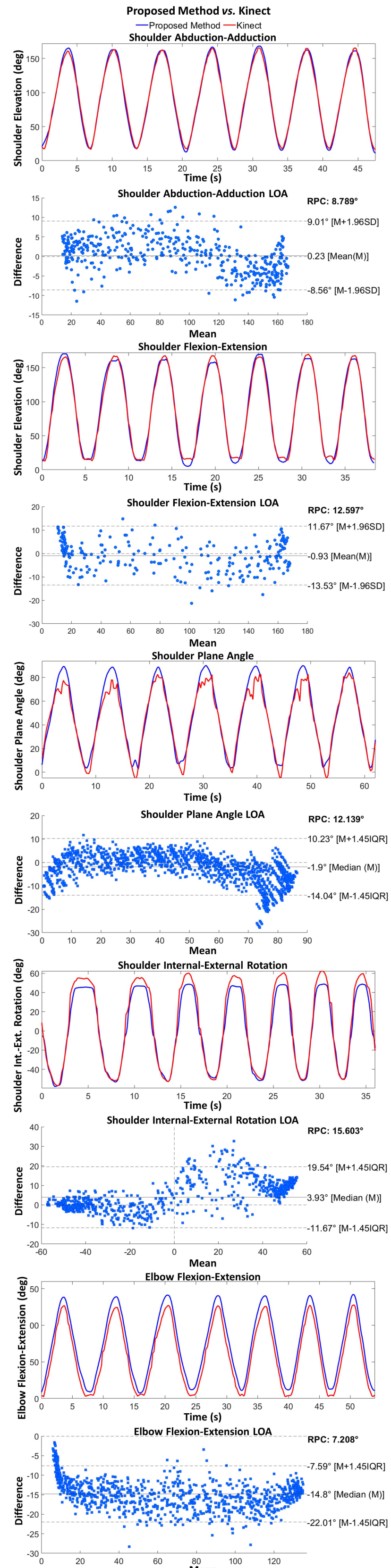
$$\theta = \cos^{-1}\left(\frac{\hat{W}_y - \hat{S}_M}{T_{N_{\text{avg}}} a_r} \right), \; \phi = \cos^{-1}\left(\frac{\hat{W}_x - \hat{S}_x}{T_{N_{\text{avg}}} a_s} \right),$$
$$\psi = \cos^{-1}\left(\frac{\hat{W}_x - \hat{E}_x}{T_{N_{\text{avg}}} f_r} \right), \; \alpha = \cos^{-1}\left(\frac{\vec{U}_R \cdot \vec{F}_R}{\left|\vec{U}_R\right|\left|\vec{F}_R\right|} \right). \tag{3}$$

B. Experiment design

A virtual environment is created using the Unity Engine (Unity Software Inc., San Francisco, CA, USA) to compare the proposed system with the synthetic ground truth data. The virtual environment comprises of an SMPL-X [13] 3D human model standing in front of a virtual camera. The SMPL-X model undergoes various exercises for the right arm: (*i*) shoulder abduction-adduction ($\phi = 0°$, $\theta \in [0, 170]°$); (*ii*) shoulder flexion-extension ($\phi = 90°$, $\theta \in [0, 170]°$); (*iii*) shoulder plane angle ($\theta = 90°$, $\phi \in [0, 90]°$); (*iv*) shoulder internal-external rotation ($\theta = 0°, \alpha = 90°, \psi \in [-50, 50]°$); and (*v*) elbow flexion-extension ($\phi = 0, \alpha = [0, 130]°$). To further validate the system, the exercises are repeated by a neurologically intact individual with nominal upper extremity ROM while the data is collected and results are compared for the proposed approach *vs.* the Kinect. See Fig. 3 for details.



(a)



(b)

Fig. 3: Comparison between the data from (a) the synthetic ground truth *vs.* the proposed method and (b) Kinect *vs.* the proposed method.

For the ground truth data, measurements are timestamped and recorded in a text file. Simultaneously, the video feed from the virtual camera is processed by the MediaPipe and the resulting 2D landmarks are also timestamped and recorded. Fig. 3 shows the graphs of the computed ROM angles for the ground truth *vs.* the proposed method for various exercises. A similar data collection method is used when a user performs the exercises for comparing the proposed method *vs.* the Kinect-based measurements and these results are also shown in Fig. 3. To mitigate noise effects, the data from the MediaPipe is filtered using a moving average filter with a window length of 8, which time-shifts the data. In post-processing, the data is time-aligned with the ground truth and Kinect data for analysis.

III. RESULTS AND DISCUSSION

The results from the proposed method are compared to the ground truth using the Bland-Altman test [15] to find the limits of agreement (LOA). Fig. 3 and Table I show the LOA for various exercises. The LOA for exercise 2 is seen to be higher than for other exercises. This is explained by the arm motion out of the frontal plane for exercise 2, where the lack of depth information causes the proposed approach to be degraded especially near the extreme angles ($\theta = 0°$ and $\theta = 180°$). In exercise 3, a similar effect is seen at $\phi = 0°$, but since the angle never reaches $\phi = 180°$, the LOA is lower in this case. In exercise 4, since only the forearm moves out of the frontal plane, the lack of depth estimation does not degrade the corresponding ROM estimate excessively.

Fig. 3 and Table I also show the LOA between the proposed method *vs.* Kinect. The LOA data has larger values for comparison with Kinect in contrast to the ground truth. One reason for this may be that the camera sensor suffers from noise. Next, for exercise 1, the large LOA against Kinect may arise from the discrepancy in the data for the low values of shoulder elevation angles, which are normally the resting position of the arm and may not be significant for one's ability to perform activities of daily living. For the larger shoulder elevation angles, the data from the proposed system closely follows the data from the Kinect. For exercise 2, the large value of LOA can be attributed to the discrepancy in the data for the low and high values of shoulder elevation angles, where the Kinect measurements may suffer due to only small changes in depth. For exercise 3, the large value of LOA can be attributed to data discrepancy for $\phi = 90°$ where the Kinect sensor suffers from occlusion of the elbow joint. For exercise 4, the large LOA value may arise from the moving average process used in the proposed method that reduces the slope of the motion trajectory as the arm switches from internal to external rotation. While this does not affect the peak values, a reduction in the moving average window size is found to improve the LOA. Finally, for exercise 5, the large LOA against Kinect may be caused by the pose prediction uncertainty from the MediaPipe and the Kinect.

IV. CONCLUSION

We proposed a ROM assessment system using a commercially available DVA device with a built-in camera. The built-in camera of the DVA devices creates future research opportunities such as long-term ROM assessment and therapy compliance applications related to telehealth. Our future work will explore the possibility of training deep learning models with the DVA-based measurements to perform ROM assessment and determine if such a model can detect and track changes in the ROM of an individual over time. Such a system will be useful for elderly and it will help them prevent or, at the very least, slow down the aging-related loss of ROM by ensuring adherence to therapy. The healthcare providers can also use the data provided by the system to assign therapies tailored to specific patients, which can lead to a more effective therapy outcome.

TABLE I: Limits of agreement for shoulder (S) and elbow (E) exercises.

No.	Exercise	*vs.* Synthetic data	*vs.* Kinect
1	S: abduction-adduction	$\pm 4.3°$	$\pm 8.8°$
2	S: flexion-extension	$\pm 7.1°$	$\pm 12.6°$
3	S: plane angle	$\pm 3.3°$	$\pm 12.1°$
4	S: internal-external rotation	$\pm 4.3°$	$\pm 15.6°$
5	E: flexion-extension	$\pm 2.9°$	$\pm 7.2°$

REFERENCES

[1] N. M. Lacktman, D. L. Rosen, M. R. Chmielewski, and N. A. Beaver, *2017 Telemedicine and Digital Health Survey*, Foley & Lardner LLP, [Online]. Available: https://www.foley.com/en/files/uploads/2017-Telemedicine-Survey-Report-11-8-17.pdf [Accessed: 13-Jan-2022].

[2] L. M. Koonin et al., "Trends in the use of telehealth during the emergence of the COVID-19 pandemic–United States, January–March 2020," *Morb. Mortal. Wkly. Rep.*, vol. 69, no. 43, pp. 1595–1599, 2020.

[3] K. Olmstead, "Nearly half of Americans use digital voice assistants, mostly on their smartphones," Pew Research Center, 2017. [Online]. Available: http://pewrsr.ch/2kquZ8H [Accessed: 11-Dec-2021].

[4] M. Duque, S. Pink, Y. Strengers, R. Martin, and L. Nicholls, "Automation, wellbeing, and digital voice assistants: Older people and Google devices," *Converg.: Int. J. Res. into New Media Technol.*, vol. 27, no. 5, pp. 1189–1206, 2021.

[5] A.L. Neves et al., "Digital health in an ageing world," *The Role of Family Physicians in Older People Care*, J. Demurtas and N. Veronese, Eds. Springer, Cham, 2022, pp. 107–118.

[6] E. Sezgin, Y. Huang, U. Ramtekkar, and S. Lin, "Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–4, 2020.

[7] O'Brien, K. et al. "Voice-controlled intelligent personal assistants to support aging in place", *J. Am. Geriatr. Soc.*, 2020, vol. 68, no. 1, pp. 176–179.

[8] N. Bott et al., "A protocol-driven, bedside digital conversational agent to support nurse teams and mitigate risks of hospitalization in older adults: Case control pre-post study," *J. Med. Internet Res.*, vol. 21, no. 10, p. e13440, 2019.

[9] A. Rajkumar et al., "Wearable inertial sensors for range of motion assessment," *IEEE Sens. J.*, vol. 20, no. 7, pp. 3777–3787, 2020.

[10] D. Xiang, H. Joo, and Y. Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," *IEEE Conf. Comp. Vision and Patt. Recog.*, 2019, pp. 10957-10966.

[11] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking,", ArXiv, vol. abs/2006.10204, 2020.

[12] D.H. Gates et al., "Range of motion requirements for upper-limb activities of daily living," *Am. J. Occup. Ther.*, vol. 70, no. 1, pp. 7001350010p1-7001350010p10, 2016.

[13] G. Pavlakos et al., "Expressive body capture: 3D hands, face, and body from a single image," *IEEE Conf. Comp. Vision and Patt. Recog.*, 2019, pp. 10967–10977.

[14] J. M. Soucie et al., "Range of motion measurements: Reference values and a database for comparison studies", *Haemophilia*, vol. 17, no. 3, pp. 500-507, 2011.

[15] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Int. J. Nurs. Stud.*, vol. 47, no. 8, pp. 931–936, 2010.