Article:
Li, R, St George, RJ, Wang, X et al. (8 more authors) (2022) Moving towards intelligent telemedicine: Computer vision measurement of human movement. Computers in Biology and Medicine, 147. 105776. ISSN 0010-4825

https://doi.org/10.1016/j.compbiomed.2022.105776

MOVING TOWARDS INTELLIGENT TELEMEDICINE: COMPUTER VISION MEASUREMENT OF HUMAN MOVEMENT

Renjie Li
Discipline of Information and Communication Technology
Wicking Dementia Research and Education Centre
University of Tasmania
renjie.li@utas.edu.au

Rebecca J. St George
Sensorimotor Neuroscience and Aging Group
School of Psychological Sciences
University of Tasmania
rebecca.stgeorge@utas.edu.au

Xinyi Wang
Wicking Dementia Research and Education Centre
Discipline of Information and Communication Technology
University of Tasmania
xinyi.wang@utas.edu.au

Katherine Lawler
Wicking Dementia Research and Education Centre
University of Tasmania
katherine.lawler@utas.edu.au

Edward Hill
Wicking Dementia Research and Education Centre
University of Tasmania
edward.hill@utas.edu.au

Saurabh Garg
Discipline of Information and Communication Technology
University of Tasmania
saurabh.garg@utas.edu.au

Stefan Williams
School of Medicine
University of Leeds
umswi@leeds.ac.uk

Samuel Relton
School of Medicine
University of Leeds
S.D.Relton@leeds.ac.uk

David Hogg
School of Computing
University of Leeds
D.C.Hogg@leeds.ac.uk

Quan Bai*
Discipline of Information and Communication Technology
University of Tasmania
quan.bai@utas.edu.au

Jane Alty†
Wicking Dementia Research and Education Centre
University of Tasmania
jane.alty@utas.edu.au

ABSTRACT

Background

Telemedicine video consultations are rapidly increasing globally, accelerated by the COVID-19 pandemic. This presents opportunities to use computer vision technologies to augment clinician visual judgement because video cameras are so ubiquitous in personal devices and new techniques, such as DeepLabCut (DLC) can precisely measure human movement from smartphone videos. However, the accuracy of DLC to track human movements in videos obtained from laptop cameras, which have a much lower FPS, has never been investigated; this is a critical gap because patients use laptops for most telemedicine consultations.

Objectives

To determine the validity and reliability of DLC applied to laptop videos to measure finger tapping, a validated test of human movement.

*joint senior author
†joint senior author

Method

Sixteen adults completed finger-tapping tests at 0.5Hz, 1Hz, 2Hz, 3Hz and at maximal speed. Hand movements were recorded simultaneously by a laptop camera at 30 frames per second (FPS) and by Optotrak, a 3D motion analysis system at 250 FPS. Eight DLC neural network architectures (ResNet50, ResNet101, ResNet152, MobileNetV1, MobileNetV2, EfficientNetB0, EfficientNetB3, EfficientNetB6) were applied to the laptop video and extracted movement features were compared to the ground truth Optotrak motion tracking.

Results

Over 96% (529/552) of DLC measures were within +/-0.5Hz of the Optotrak measures. At tapping frequencies >4Hz, there was progressive decline in accuracy, attributed to motion blur associated with the laptop camera's low FPS. Computer vision methods hold potential for moving us towards intelligent telemedicine by providing human movement analysis during consultations. However, further developments are required to accurately measure the fastest movements.

***K*eywords** Telemedicine · DeepLabCut · finger tapping · motor control · computer vision

1 Introduction

The assessment of human movement by visual observation is a fundamental part of clinical assessments in all areas of medicine. The clinician's visual judgement of patient movement plays a key role in diagnosis and assessment across a multitude of conditions throughout the life course of their patients. For example, clinicians evaluate how a baby grasps an object for child development assessments, the range of eye movements after surgery, the amplitude of arm movements after rehabilitation, the speed of walking after a stroke, and to track benefits (and side effects) of medications for hand tremor. These are just a few specific examples of medicine consultations taking place all over the world every hour of every day.

However, the accuracy of clinicians, even those expert in movement analysis, are constrained by the limits of human perception, which cannot accurately measure subtle changes. Numerous publications have proposed technological methods to objectively measure human movement [1, 2, 3]. These have the theoretical benefit of allowing remote monitoring and assessment, but a requirement for specialist equipment, wearable sensors or patient engagement with specific apps likely explains why none have entered routine clinical practice.

Wearable sensors are commonly used to extract finger tapping features in the laboratory. In prior studies, machine learning methods have been applied to determine finger tapping pattern classification. For instance, Shima et al. [4] applied Log-linearized Gaussian Mixture Networks on sensor data to extract finger tapping movements. Wissel et al. [5] applied Hidden Markov model and support-vector machine (SVM) to classify finger movement patterns into different groups using electroencephalogram data. Khan et al. [6] applied a random forests model to classify individual finger movements using functional near-infrared spectroscopy data. These wearable sensors or neurophysiological signal based methods can extract accurate finger tapping features; however, they require participants to have sensors or other equipment physically attached on to them which limits implementation in any environment outside of a specialised laboratory. In contrast, camera-based methods are non-touch and can be used remotely, thus providing wider accessibility.

The COVID-19 pandemic has highlighted the reach but also the challenges of remote medicine delivery [7]. Telemedicine consultations remain severely limited compared to standard face-to-face consultations, by the fact that clinicians cannot accurately examine patients' movements remotely. This was highlighted in a study of neurology clinics during the COVID-19 pandemic, that found telemedicine consultations were much better suited to conditions that were based on describing symptoms (e.g. headaches, epilepsy) than those that required doctors to observe abnormalities in movement (e.g. Parkinson's, Multiple Sclerosis [8]). This inability of clinicians to accurately evaluate human movements remotely, combined with the swift uptake of telemedicine, results in a significant risk to patient safety.

Thus, there is a growing and urgent need to extend the capabilities of telemedicine so that intelligent video technologies can provide accurate measurement of movement that is clinically relevant. Recent developments in computer vision deep learning methods open up the opportunity for remote healthcare assessments to include precise measures of human movement. DeepLabCut (DLC) is a new artificial intelligence software that was originally designed to perform marker-less tracking of research mice and insects [9] and has since been used in research-related fields in human movement. The coming age of video consultation presents opportunities to use technology to aid clinician visual judgement, because video cameras are so ubiquitous in personal devices (no special equipment is required) and computing techniques have

the potential to precisely measure human movement from video. This could provide a new form of augmented or automatic remote healthcare assessment.

However, if similar methods are to be successfully applied to telemedicine, it is necessary to evaluate their accuracy when applied to videos obtained from standard laptop cameras (with a relatively low FPS) because this is one of the most common computer cameras used by patients for telemedicine consultations, including older adults [10] and carers, and especially for conditions that have a central focus on movement assessment [11]. So far, there has not been a published comparison of human movements measured via analysis of video collected through a laptop camera with a 'gold standard' measure obtained through wearable movement sensors.

In this study, for the first time, we determine the validity and reliability of DLC computer vision methods applied to 2D video collected via a standard laptop camera at 30 FPS, compared to a 3D gold-standard wearable sensors method collected at 250 FPS. To clarify, **the objective of this study** is not to classify participants into different positive and negative controls, but rather, to validate whether computer vision methods can extract accurate hand movement features compared with the gold standard Optotrak method using video data from a relatively low FPS (30FPS) laptop camera. We use a well-validated clinical test of human movement control, finger tapping, during which a person is asked to repetitively tap index finger and thumb together. This test is easy for participants to perform whilst seated and allows for a range of different frequencies and component measures to be evaluated. Additionally, finger tapping test is a common test being widely used in aiding the diagnosis of neuro logical diseases, and tackling small objects (like hand or finger) tracking from videos or images is challenging in computer vision domain. That's why it is a timely and necessary validity and reliability study of showing how accurately the computer vision method can measure finger tapping hand movement compared with the gold standard wearable sensor method. Through experiments, our method recognizes the need to loosen specific requirements, such as patient positioning and high-quality camera equipment, if objective measures of movement are to be successfully integrated into real-world telemedicine consultations.

2 Materials and Methods

2.1 Participants

A convenience sample of sixteen staff and students (9 female, mean age 34.5 years; range 24-52) at the University of Tasmania were recruited via an email invitation. Assessments took place at the University of Tasmania Sensorimotor Neuroscience and Ageing Laboratory. The study was approved by the institutional ethics review board at the University of Tasmania (Project ID: 21660) and participants provided written informed consent in accordance with the Declaration of Helsinki. Information on gender, age, dominant hand, and whether there was any history of neurological disorders, was collected from each participant.

2.2 Experiment Design

Participants sat facing a Dell Laptop (Model Precision 5540) placed on a table approximately 60cm in front of them. The laptop 2D camera captured video images at 30 FPS with a resolution of 1280 x 720 pixels. A high-speed 3D Optotrak camera system (Northern Digital Inc.) was fixed to the wall approximately 2m behind the participant; (Figures 1A and B). The Optotrak uses three co-linear detectors to record 3D (x, y, z) positional data of infrared Light Emitting Diodes (LEDs, active markers), at 250 FPS with an accuracy of 0.1mm; (Figure 1C [12]). A plain blue board placed 50cm behind the participant provided a uniform background for the laptop camera (Figure 1B). Standard ambient lighting was used.

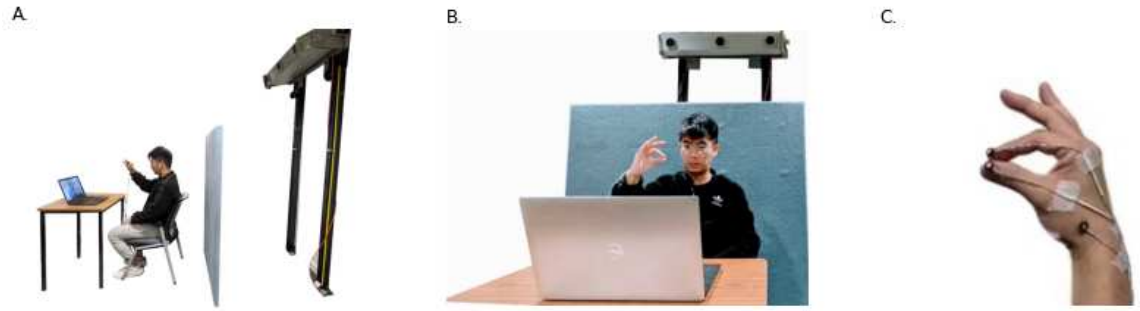

Figure 1: Experiment design. 1A: Lateral view with the participant sitting approximately 60cm from the laptop and approximately 2m from the Optotrak camera. 1B: Experimental set up from the front view, with a plain light blue board placed behind the participant and the hand visible to both the Optotrak camera and the laptop camera. 1C: The positions of the LED sensors with one secured on the index fingertip, one on the thumb-tip and one over the radial styloid process, each secured with adhesive tape. In this position, the sensors were visible to the Optotrak camera but not to the laptop camera, so as not to interfere with DLC video image analysis.

Three small (5mm diameter) lightweight (<1g) LED sensors were attached to the participant's right hand: one sensor on the lateral aspect of the index fingertip, one on the dorsal aspect of the thumb-tip and one over the radial styloid process; Figure 1C. The sensor positions were chosen so they were visible to the Optotrak camera while invisible to the laptop camera. In this way, both cameras could record the movement simultaneously, while the image of the sensors did not impact the deep learning algorithms to detect key points on the hand.

2.3 Protocol

The participant flexed their right elbow and held their right hand steady with the index finger and thumb opposed and their little finger facing towards the laptop camera. The researcher checked that the participant's hand position was captured by both the laptop camera (visible on the screen) and the Optotrak system (sensor positions detected) and then started recording from both systems. The participant was instructed to hold their hand still for a few seconds of recording and begin finger tapping when the researcher gave the 'start' command. The delayed start allowed time-synchronization between cameras during offline analysis.

The participant was instructed to tap their index finger against their thumb 'as big and fast as possible' (internally-paced). For the next four conditions, an electronic metronome (auditory tone) externally paced each period of finger tapping at 0.5Hz, 1Hz, 2Hz and 3Hz. After 20 seconds of finger tapping, the participant was instructed to stop. Each condition recorded thus comprised a preparation period of 3-5 seconds and a 'tapping' period of 20 seconds.

The laptop and table were then moved 20cm further away from the participant and the same set of recordings were repeated. Thus, each participant completed ten recordings in total (see Table 1). After each recording, the participant had a 30-second period to rest. The order of conditions was fixed.

Six recruited participants (3 female, mean age 28.7 years; range 24-36) agreed to complete the full protocol twice. They had a five-minute break between the first and second recordings. In total, this resulted in 220 23-25 -second 2D videos collected via the 30 FPS laptop camera (around 165,000 frames in total) with paired 3D positional data collected via the 250 FPS 3D wearable sensor system.

Table 1: Finger tapping protocol

Condition number	Distance of participant's hand from laptop	Tapping frequency
1	60cm	as fast as possible
2	60cm	0.5Hz
3	60cm	1Hz
4	60cm	2Hz
5	60cm	3Hz
6	80cm	as fast as possible
7	80cm	0.5Hz
8	80cm	1Hz
9	80cm	2Hz
10	80cm	3Hz

2.4 Key Point Detection on Hand through 2D Video

We implemented 5 steps to detect key points (index fingertip and thumb-tip) on the hands through 2D videos, i.e., data pre-processing, data augmentation, model training, model evaluation and key point inference by using the DLC framework [13]. Figure 2 shows the whole process and details are given in the following sub-sections.

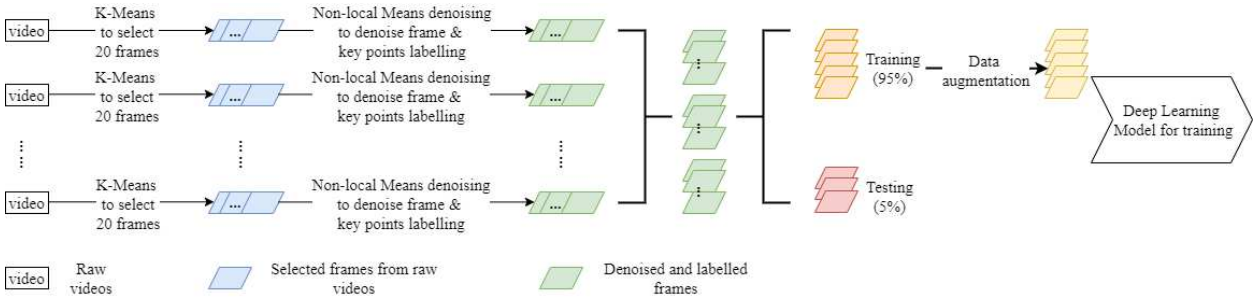

Figure 2: The process of key point detection on the hands using 2D video data.

2.4.1 Data Pre-Processing

In the data pre-processing step, initially 20 frames from each of the 220 finger tapping videos were selected by a K-means (K=10) clustering algorithm and the positions of the index fingertip and thumb-tip were manually labelled. These 20 selected frames were regarded as representations of the video, which contained 600 frames in total (30FPS $\times$ 20 seconds). To reduce the noise on individual frames, we applied a non-local means denoising algorithm [14]. Figure 3 shows the sample frames before and after denoising using this technique, demonstrating that the background behind the hand becomes cleaner after denoising.

2.4.2 Data Augmentation

Data augmentation settings in the DLC framework were implemented. Specifically, the probability of adding augmentation to a frame was set at 0.5, the scale crop ratio was 0.4, the rotation degree was +/-25 degrees, the fraction of applying rotation was 0.4, and the scale rate ranged between 0.5 and 1.25.

2.4.3 Model Training

In the model training step, 4,400 denoised and labelled frames were randomly split into training partition (95%) and testing (5%) datasets. There are many different neural network based key point detection methods, we selected 3 typical networks to be trained for finger tip detection, i.e., ResNet50 [15], EfficientNetB0 [16] and MobileNetV2 [17]. ResNet50, EfficientNetB0 and MobileNetV2 are commonly used as backbone networks for different computer vision tasks including key point detection [18, 19]. ResNet50 introduces the concept of skip-connection, which solves the gradient vanishing problem in neural networks. In this case, the network of ResNet can be very deep, allowing the network to learn deeper features in the network without compromising the vanishing gradient problem. EfficientNetB0 is well known for its scaling method that can uniformly scale different dimensions of depth/width/resolution using a

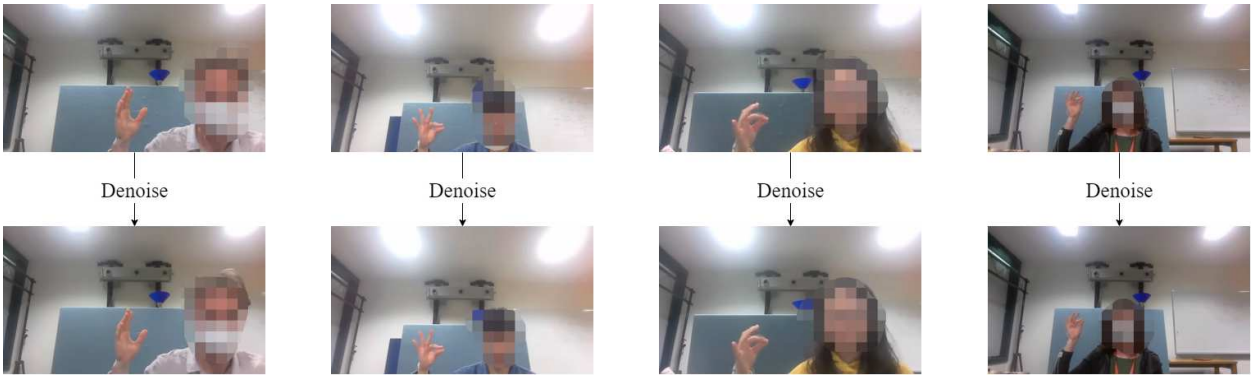

Figure 3: Using non-local means denoising algorithm to denoise the raw frames. This produced a more clear background. Faces have been covered with mosaics on the frames.

compound coefficient [16]. MobileNetV2 is a light weight network which can provide the real-time key point detection. Although it is light weight, it achieves good performance in key point detection tasks [20]. All these networks are state of the art methods in the computer vision field and they represent different types of novelty in terms of neural network theory. Additionally, these networks have been embedded into DLC software which brings convenience to neuroscientists and other non-computer scientists who may wish to use these methods to track movements. The output layers of each network architecture were 2 score maps (2D grid with values of 0-1 in each pixel) indicating the presence of the thumb-tip or index fingertip. The Adam optimizer [21] was applied in the learning process. To take advantage of transfer learning, training started from ImageNet pre-trained model and ended after 50,000 epochs. The loss function was calculated as the cross-entropy between ground truth score maps and predicted score maps.

2.4.4 Model Evaluation

To evaluate the performance of the model, we calculated the End Point Error (EPE) between predicted point position and real point position, on the testing dataset for each of the three neural network architectures (Table 2). More neural networks training results were included in the supplementary material. Equation 1 shows the calculation of EPE.

$$\text{EPE} = \frac{1}{N \times J} \sum_{j=1}^{J} ||P_n^{(j)} - Y_n^{(j)}||_2 \tag{1}$$

where J is the number of types of hand key points (here $J = 2$ for index fingertip and thumb-tip), $P_n^{(j)}$ is the predicted position of the j^{th} hand key point on the n^{th} image and $Y_n^{(j)}$ is the true position of the j^{th} key point on the n^{th} image.

Table 2: End Point Error (EPE) for the three deep learning neural networks.

Deep learning neural network	EPE on training dataset	EPE on testing dataset
ResNet50	2.98 pixels	3.00 pixels
EfficientNetB0	1.52 pixels	1.53 pixels
MobileNetV2	2.21 pixels	2.30 pixels

To evaluate the complexity of the model, we showed the EPE vs number of parameters plot for different networks in Figure 4 to see the efficiency of different models. More neural networks' complexity evaluations were included in the supplementary material. Overall, different networks' fingertip tracking errors are all very small (around 1.5 to 3 pixels), while EfficientNetB0 achieves lowest EPE at a relatively small number of parameters.

2.4.5 Key Point Inference

In the key point inference stage, frames from the original videos were predicted by using the trained neural network architectures, and the output were a set of 2D (x, y) coordinates of index fingertip and thumb-tip in pixels.

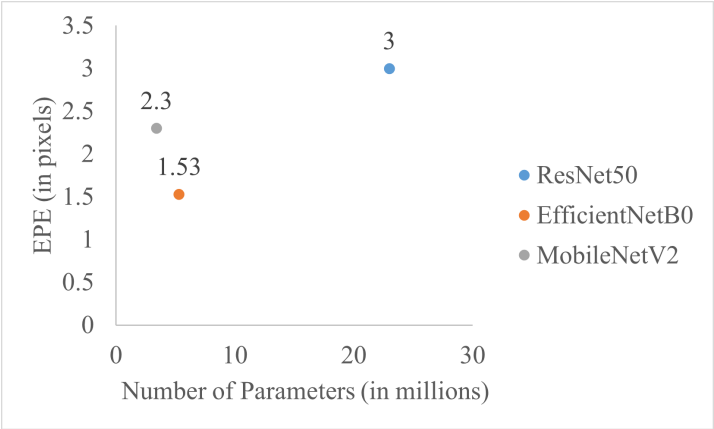

Figure 4: The performance (measured in EPE) vs number of parameters for different networks.

2.5 Extraction of Hand Movement Features

For the Optotrak system, Euclidean distance (D) (Equation 2) was calculated between the finger tip sensor and thumb tip sensor in 3D space and measured in millimeters. For the computer vision methods, displacement between the index fingertip and thumb-tip was measured in 2D space by the number of pixels (Equation 3). Figure 5 shows displacement vs time graph for both 2D space (based on 2D video) and 3D space (based on Optotrak).

$$D_{3D} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \tag{2}$$

$$D_{2D} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3}$$

where thumb-tip position is (x_1, y_1, z_1) and (x_1, y_1) for 3D and 2D space respectively, while index fingertip position is (x_2, y_2, z_3) and (x_2, y_2) for 3D and 2D space respectively.

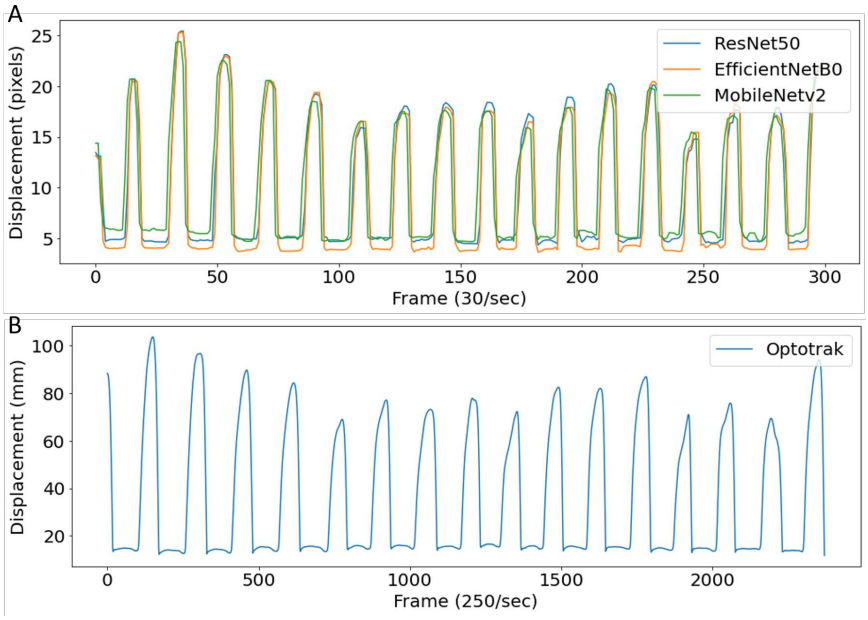

Figure 5: A, shows an example of the distance between the two finger sensors measured by computer vision methods (DLC) during the 1Hz condition. B shows the same finger-tapping motion calculated from the Optotrak method.

The Optotrak and computer vision data were time-synchronized using the peak (i.e. index finger and the thumb maximally separated) of the second tap cycle and the subsequent 10 second period of data was included in the analysis.

The Mean Tapping Frequency (M-TF) was calculated as the average value of 1 divided by the time difference between each consecutive peak points (Equation 4). The Variation of TF (Var-TF) was calculated as the coefficient of variance, the ratio of standard deviation of 1 divided by the time difference between each consecutive peak points to the M-TF (Equation 5).

$$\text{M-TF} = \frac{1}{K_p - 1} \sum_{k=2}^{K_p} \frac{1}{t_{(k)} - t_{(k-1)}} \tag{4}$$

$$\text{Var-TF} = \sqrt{\frac{\sum_{k=2}^{K_p} (\frac{1}{t_{(k)} - t_{(k-1)}} - \text{M-TF})^2}{K_p - 1}} / \text{M-TF} \tag{5}$$

where K_p refers to the number of peaks and $t_{(k)}$ refers to the time point at k^{th} peak.

2.6 Statistical Analysis

M-TF and Var-TF outcomes were compared between the three DLC computer vision neural network architectures and the gold standard measure. Reliability of the computer vision methods for tracking hand motion at different distances from the laptop camera were calculated: Near-To-Laptop (60cm) versus Far-From-Laptop (80cm). Bland Altman [22] plots and paired Welch's t-tests measured the level of agreement, with +/-0.5Hz as a clinically acceptable error margin [23].

To evaluate the validity of the computer vision methods, we compared each of the three different artificial neural network architectures from the DLC platform [9] (i.e., ResNet50, EfficientNetB0 and MobileNetV2) based computer vision methods separately with the Optotrak measurements. We used Bland Altman [22] plots to measure the degree of error. To evaluate (whether the distance from the camera had significant impact on features extracted from different methods), we compared for each participant, the same finger tapping tests completed Near-To-Laptop (60cm) to their repeat tests completed Far-From-Laptop (80cm). We used Bland Altman [22] plots and paired Welch's t-tests to measure the degree of error. We considered $\pm$0.5Hz as a clinically acceptable error margin and in line with previous similar publications [23].

3 Results

3.1 Validation of Computer Vision Methods Compared to The Gold Standard System

When the participants finger tapped between 0.5Hz and 4Hz, the mean tapping frequencies obtained from the three-computer vision methods correlated highly with the Optotrak measures; see Figure 6 and 7, and Table 3. Almost all (95.8%; 538/552) of the computer vision measures were within +/-0.5Hz of the Optotrak measures in this frequency range; specifically 95.7%, 176/184 for ResNet50; 92.9%, 171/184 for EfficientNetB0 and 98.9%, 182/184 for MobileNetV2. However, as can be seen in Figure 6, when participants tapped at frequencies higher than 4Hz, there was a decline in the accuracy of the computer vision methods with significant differences between the computer vision and Optotrak methods (Table 3). The computer vision methods progressively under-estimated the tapping frequencies with fast movements, giving falsely low measures of frequency compared to the benchmark. On viewing the videos at higher tapping frequencies, it was noted that they had considerable motion blur on some frames. It was hard to manually label the correct positions of key points on these blurred frames, this blur led to inaccurate key point detection performance of the computer vision methods at higher speeds. The further validation assessments using a range of other neural networks, namely ResNet101, ResNet152, MobileNetV1, EfficientNetB3 and EfficientNetB6 are presented in the Supplementary Materials. In summary, all the neural networks generally showed accurate hand movement features extraction in low frequency finger tapping cases, but inaccurate extraction of hand movement features at tapping frequencies above 4HZ.

3.2 Reliability of Computer Vision Methods at Two Different Distances from Camera

There were no significant differences compared to the Optotrak system between tapping frequencies, or variation, measured Near-To-Laptop and Far-From-Laptop by the three computer vision methods ($p>0.05$); see Figure 8 and Table 4. It is important to note that there will be a natural variation between a participant's performance of the same condition (e.g. 1 Hz paced) in the Near-To-Laptop and Far-From-Laptop positions as humans very rarely reproduce movements 100% precisely at two different time points, even when paced. This is especially the case for internally paced 'Big/Fast' conditions, as exemplified by the variation in the Optotrak measures at higher frequencies too. The reliability

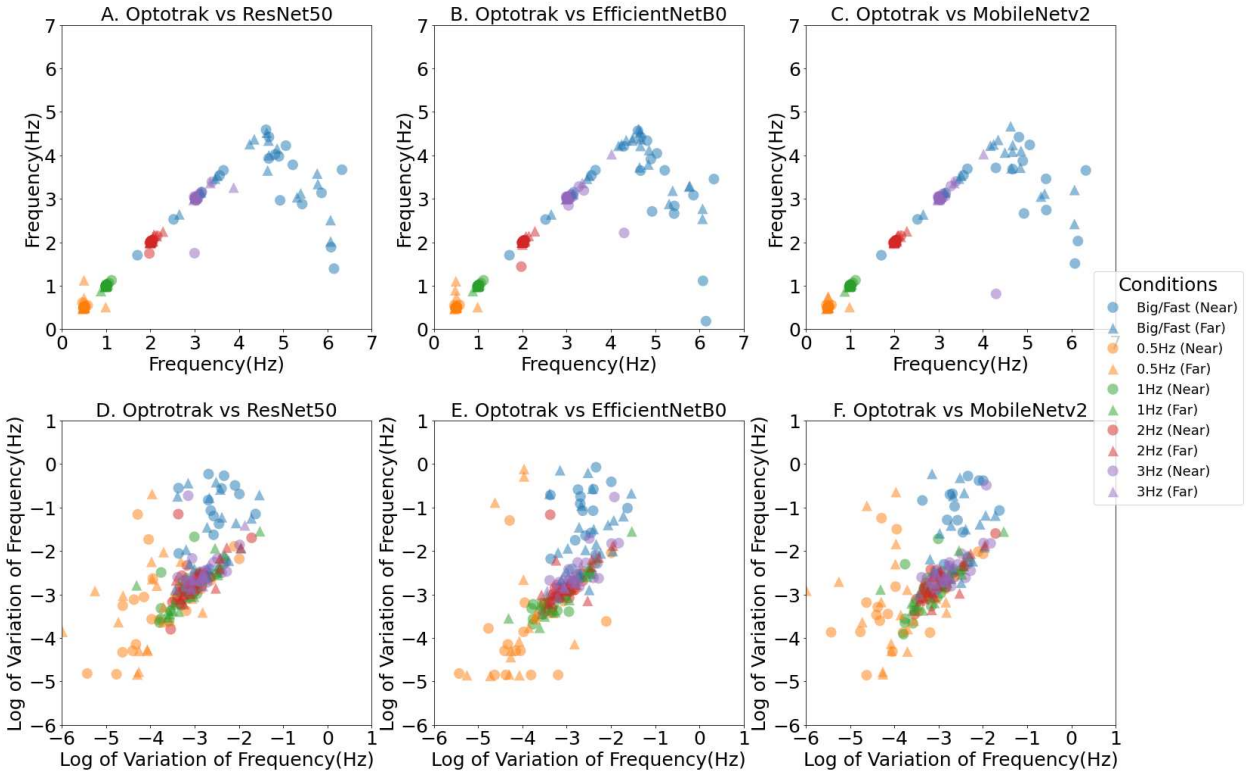

Figure 6: Tapping Frequency x-y scatter plot between different computer vision methods (y axis) and the Optotrak method (x axis). A, B and C show the scatter plots of mean tapping frequency. D, E and F show the scatter plots of the logarithm of variation in the tapping frequency. The colored marks represent the different finger tapping conditions with blue denoting the 'Big/Fast' self-paced conditions, and yellow, green, red and purple the externally paced conditions at frequencies of 0.5Hz, 1Hz, 2Hz and 3Hz respectively. Circles are for conditions performed Near-To-Laptop (60cm) and triangles are for conditions performed Far-From-Laptop (80cm).

Table 3: Accuracy of each computer vision method compared with the Optotrak measures

Methods	M-TF t-value (p-value)	Var-TF t-value (p-value)
'As fast as possible' condition with **frequency < 4Hz**		
ResNet50	0.52 (0.48)	1.48 (0.24)
EfficientNetB0	0.72 (0.47)	2.60 (0.01)
MobileNetV2	0.39 (0.54)	1.89 (0.19)
'As fast as possible' condition with **frequency > 4Hz**		
ResNet50	107.61 (0)	59.40 (0)
EfficientNetB0	7.44 (0)	6.57 (0)
MobileNetV2	109.67 (0)	41.56 (0)
0.5Hz condition		
ResNet50	0.39 (0.53)	6.41 (0.02)
EfficientNetB0	0.84 (0.40)	1.81 (0.08)
MobileNetV2	0.02 (0.88)	7.63 (0.01)
1Hz condition		
ResNet50	0.02 (0.90)	3.40 (0.70)
EfficientNetB0	0.02 (0.99)	0.58 (0.56)
MobileNetV2	0.05 (0.82)	3.63 (0.06)
2Hz condition		
ResNet50	0.02 (0.89)	6.10 (0.02)
EfficientNetB0	0.93 (0.36)	1.43 (0.16)
MobileNetV2	0.01 (0.92)	4.94 (0.03)
3Hz condition		
ResNet50	4.26 (0.04)	3.92 (0.05)
EfficientNetB0	1.61 (0.11)	1.79 (0.08)
MobileNetV2	2.16 (0.15)	3.63 (0.06)

Table 4: Difference between Near-To-Laptop and Far-From-Laptop computer measures compared with the Optotrak

Methods	M-TF t-value (p-value)	Var-TF t-value (p-value)
'As fast as possible' condition		
Optotrak	0 (1)	0.43 (0.52)
ResNet50	0.55 (0.46)	0.36 (0.55)
EfficientNetB0	1.86 (0.07)	0.45 (0.66)
MobileNetV2	3.63 (0.07)	0.94 (0.34)
0.5Hz condition		
Optotrak	0.48 (0.5)	2.64 (0.12)
ResNet50	1.07 (0.31)	0.64 (0.43)
EfficientNetB0	1.50 (0.15)	1.33 (0.20)
MobileNetV2	1.07 (0.31)	0.13 (0.72)
1Hz condition.		
Optotrak	2.87 (0.1)	1.71 (0.2)
ResNet50	2.16 (0.15)	0.29 (0.6)
EfficientNetB0	1.39 (0.17)	1.30 (0.20)
MobileNetV2	1.76 (0.19)	0.45 (0.51)
2Hz condition.		
Optotrak	0.84 (0.37)	0.52 (0.48)
ResNet50	1.66 (0.21)	0.03 (0.85)
EfficientNetB0	1.39 (0.18)	0.26 (0.80)
MobileNetV2	0.62 (0.44)	0.32 (0.57)
3Hz condition.		
Optotrak	0 (0.93)	1.86 (0.18)
ResNet50	0.98 (0.34)	1.33 (0.26)
EfficientNetB0	1.30 (0.20)	1.63 (0.12)
MobileNetV2	1.28 (0.27)	2.18 (0.16)

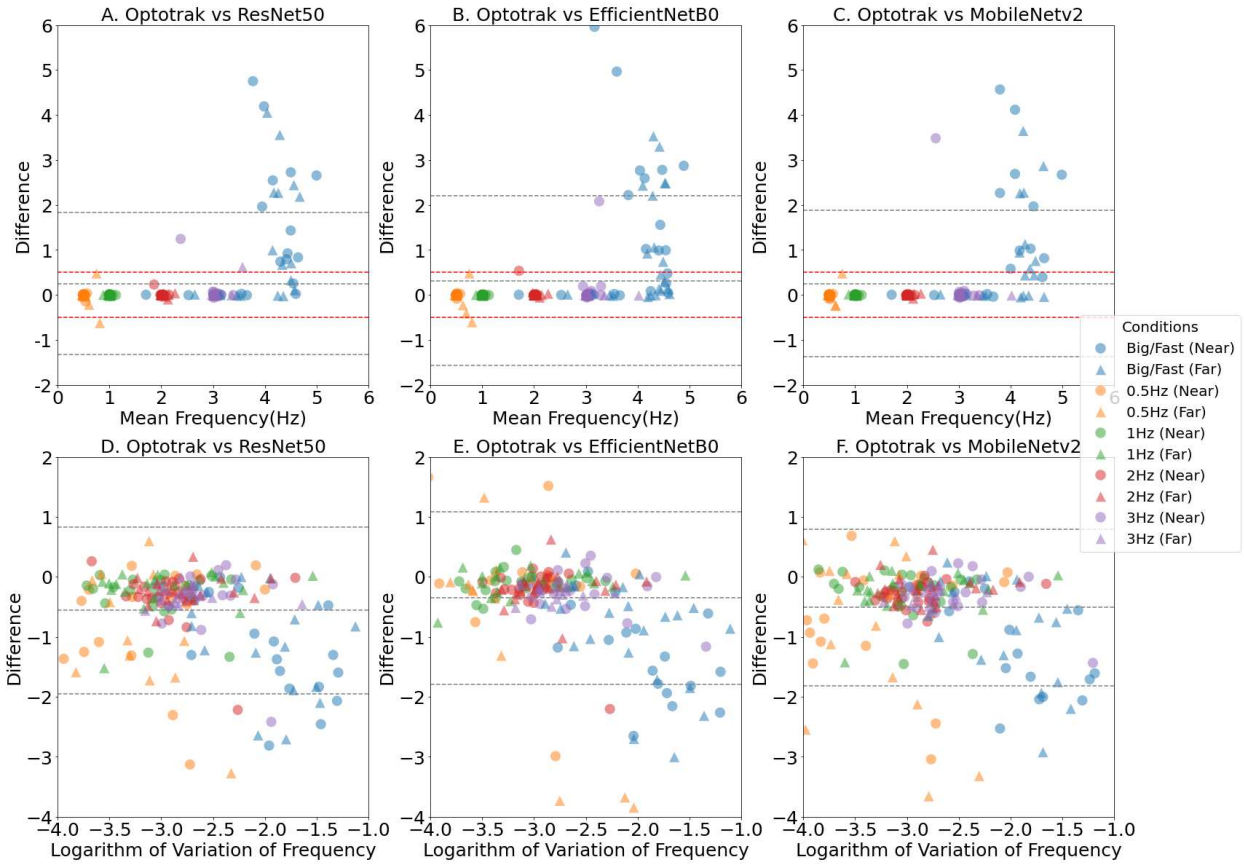

Figure 7: Validity of computer measures of Tapping Frequency, demonstrated by the Bland Altman Plots, with the representation of the limits of agreement (red dashed lines), and from -1.96 standard deviation to +1.96 standard deviation (lower and upper grey dashed lines). A, B and C show the mean tapping frequency comparison between the Optotrak system and the three computer vision methods. D, E and F show a measure of tapping rhythm - the logarithm of variation in the tapping frequency. The colored marks represent the different finger tapping conditions with blue denoting the 'Big/Fast' self-paced conditions, and yellow, green, red and purple the externally paced conditions at frequencies of 0.5Hz, 1Hz, 2Hz and 3Hz respectively. Circles are for conditions performed Near-To-Laptop (60cm) and triangles are for conditions performed Far-From-Laptop (80cm).

assessments using a range of other neural networks, namely ResNet101, ResNet152, MobileNetV1, EfficientNetB3 and EfficientNetB6 are presented in the Supplementary Materials. In summary, all the neural networks generally showed the distance between participant and camera (in a range of 60 to 80cm) does not affect the feature extraction.

4 Discussion

Our results demonstrate that when tapping frequencies were between 0.5Hz and 4Hz, the accuracy of the computer vision methods employing 2D video data collected at 30 FPS were comparable to the 'gold standard' wearable sensor method. These computer vision methods were also reliable when the hand was at different distances from the laptop camera. This is the first study to apply DLC methods to videos from a standard laptop camera to measure human hand movements. The three DLC deep learning models, ResNet50, EfficientNetB0 and MobileNetV2, showed similar validity and test-retest reliability.

The implication of this study is that existing hardware currently used for video consultations may be sufficient to objectively measure movement in order to augment clinician judgement. The accuracy of the method reduced above 4Hz due to inaccurate fingertip tracking on some blurred frames related to using the low FPS laptop camera. However, this may have little relevance for clinical use, as few patients are likely to tap at such high frequencies; for example a study [24] that quantified Parkinson's finger tapping frequency found that the mean tapping frequency was around 2Hz.

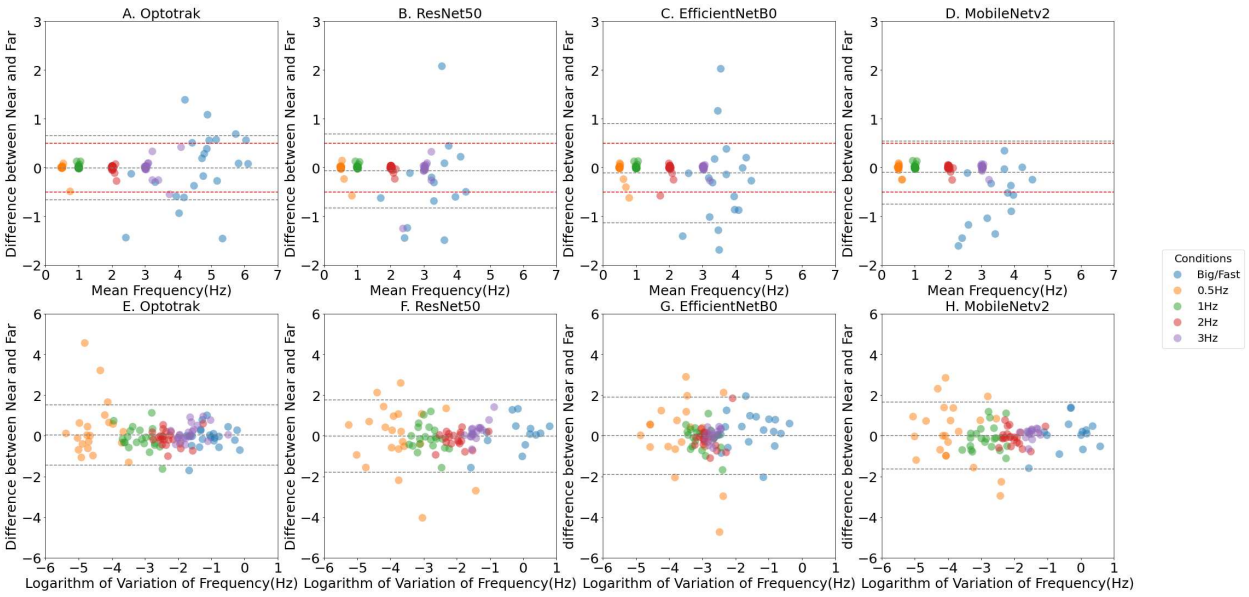

Figure 8: Reliability of computer measures of Tapping Frequency at two distances from camera, demonstrated by the Bland Altman Plots, with the representation of the limits of agreement (red dashed lines), and from -1.96 standard deviation to +1.96 standard deviation (lower and upper grey dashed lines). A, B, C and D show the mean tapping frequency scatter plots between Near-To-Laptop (60cm camera to hand) and Far-From-Laptop (80cm camera to hand) conditions for the Optotrak system and the three computer vision methods respectively. E, F, G and H show a measure of tapping rhythm - the logarithm of variation in the tapping frequency. The colored marks represent the different finger tapping conditions with blue denoting the 'Big/Fast' self-paced conditions, and yellow, green, red and purple are the externally paced conditions at frequencies of 0.5Hz, 1Hz, 2Hz and 3Hz respectively. The middle grey dashed line represents the mean difference between Near-To-Laptop and Far-From-Laptop conditions. The upper and lower grey dashed lines represent the upper and lower borders at 95% confidence level. The upper and lower red dashed lines represent the +/-0.5Hz agreement levels.

Our study extends existing understanding of methods to quantify the finger tapping examination. A variety of studies have shown that devices can be used to record finger tapping and extract clinically useful information. For example, Djuric et al. [25] proposed a method to assess finger tapping task using 3D gyroscopes; Summa et al. [26] used magneto-inertial devices to record hand motor tasks (including finger tapping tests) to assess motor symptoms. There are several reports of using video tracking to measure finger tapping [27, 28, 29, 23] including one with laptop cameras, but none have validated their method against precise wearable sensors. A particular strength of the work presented here is the use of a gold standard kinematic measure as the benchmark to test laptop camera validity. Optotrak can accurately measure movement at different distances from the infrared cameras with no constraints to ambient lighting or a cluttered background [30]. Optotrak markers collected position data in x, y and z directions at a high frequency with an accuracy of 0.1 mm. The fast and accurate data ensured the reliability of 'ground truth data. It is technically challenging to compare computer vision with established technology to record clinical examination, since 'wearables' will add relevant markers to the video, potentially improving the performance of computer vision tracking. However, we avoided this problem by novel positioning of the Optotrak markers and camera on the opposite side to the laptop camera, making the markers invisible on the video.

This is a timely study as telemedicine use dramatically increases around the world, and clinicians and researchers need accurate methods to measure hand movements. Limitations of our study include the relatively small number of participants and the homogeneity of our sample i.e. younger adults accustomed to using technology who did not have any cognitive deficit or motor impairment. Future steps would include assessing movement tracking in a wider range of participants with positive and negative controls, validating in other types of movement, and undertaking a classification-based research study using different computer vision methods.

5 Conclusion

Remote video consultation forms an expanding part of healthcare systems globally. Widespread availability of devices that allow a remote video assessment are alleviating the burden of frequent travel to clinic appointments for people living with frailty and impaired mobility [31, 32]and reducing the inequity of access to healthcare systems for people who live in rural or remote locations. There is potential for computer vision techniques to provide precise objective measures of movement to augment clinician judgement during video calls. Webcams are standard hardware for video consultation, but the accuracy of computer vision using that low-cost equipment for computer tracking of clinical examination has never been tested before now [33, 34]. Our study provides evidence that deep learning technologies have advanced to the stage where it is now feasible to integrate computer vision into remote healthcare systems using standard computer equipment. This could improve the clinical consultation, not only remotely but also when face-to-face if a camera was used to video record the examination, as it would allow clinicians to view an overlay of live extracted movement features during their clinical evaluation.

6 Acknowledgements

The authors would like to thank all the participants for taking part in the study.

7 Summary

- This is the first study to compare tracking of human hand movements using deep learning methods applied to 2D laptop videos to a 3D wearable sensor method.
- The deep learning video methods were able to accurately measure finger tapping frequency in the 0 to 4Hz range.

References

[1] Jon Currie, Ben Ramsden, Cheryl McArthur, and Paul Maruff. Validation of a clinical antisaccadic eye movement test in the assessment of dementia. *Archives of neurology*, 48(6):644–648, 1991.

[2] R Benecke, JC Rothwell, JPR Dick, BL Day, and CD Marsden. Performance of simultaneous movements in patients with parkinson's disease. *Brain*, 109(4):739–757, 1986.

[3] Renjie Li, Xinyi Wang, Katherine Lawler, Saurabh Garg, Quan Bai, and Jane Alty. Applications of artificial intelligence to aid detection of dementia: a scoping review on current capabilities and future directions. *Journal of Biomedical Informatics*, page 104030, 2022.

[4] Keisuke Shima, Toshio Tsuji, Akihiko Kandori, Masaru Yokoe, and Saburo Sakoda. Measurement and evaluation of finger tapping movements using log-linearized gaussian mixture networks. *Sensors*, 9(3):2187–2201, 2009.

[5] Tobias Wissel, Tim Pfeiffer, Robert Frysch, Robert T Knight, Edward F Chang, Hermann Hinrichs, Jochem W Rieger, and Georg Rose. Hidden markov model and support vector machine based decoding of finger movements using electrocorticography. *Journal of neural engineering*, 10(5):056020, 2013.

[6] Haroon Khan, Farzan M Noori, Anis Yazidi, Md Zia Uddin, MN Khan, and Peyman Mirtaheri. Classification of individual finger movements from right hand using fnirs signals. *Sensors*, 21(23):7943, 2021.

[7] Trisha Greenhalgh, Gerald Choon Huat Koh, and Josip Car. Covid-19: a remote assessment in primary care. *Bmj*, 368, 2020.

[8] Espen Saxhaug Kristoffersen, Else Charlotte Sandset, Bendik Slagsvold Winsvold, Kashif Waqar Faiz, and Anette Margrethe Storstein. Experiences of telemedicine in neurological out-patient clinics during the covid-19 pandemic. *Annals of clinical and translational neurology*, 8(2):440–447, 2021.

[9] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018.

[10] Wei Duan-Porter, Courtney H Van Houtven, Elizabeth P Mahanna, Jennifer G Chapman, Karen M Stechuchak, Cynthia J Coffman, and Susan Nicole Hastings. Internet use and technology-related attitudes of veterans and informal caregivers of veterans. *Telemedicine and e-Health*, 24(7):471–480, 2018.

[11] Gregor Durner, Joachim Durner, Henrike Dunsche, Etzel Walle, Robert Kurzreuther, and René Handschu. 24/7 live stream telemedicine home treatment service for parkinson's disease patients. *Movement disorders clinical practice*, 4(3):368–373, 2017.

[12] RA States and E Pappas. Precision and repeatability of the optotrak 3020 motion measurement system. *Journal of medical engineering & technology*, 30(1):11–16, 2006.

[13] Tanmay Nath, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie Weygandt Mathis. Using deeplabcut for 3d markerless pose estimation across species and behaviors. *Nature protocols*, 14(7):2152–2176, 2019.

[14] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Non-local means denoising. *Image Processing On Line*, 1:208–212, 2011.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[18] Shaoen Wu, Junhong Xu, Shangyue Zhu, and Hanqing Guo. A deep residual convolutional neural network for facial keypoint detection with missing labels. *Signal Processing*, 144:384–391, 2018.

[19] Feng Hong, Changhua Lu, Chun Liu, Ruru Liu, Weiwei Jiang, Wei Ju, and Tao Wang. Pgnet: Pipeline guidance for human key-point detection. *Entropy*, 22(3):369, 2020.

[20] Savina Colaco and Dong Seog Han. Facial landmarks detection with mobilenet blocks. *Proceedings of the Korea Telecommunications Society Conference*, pages 1198–1200, 2020.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] Davide Giavarina. Understanding bland altman analysis. *Biochemia medica*, 25(2):141–151, 2015.

[23] Stefan Williams, Hui Fang, Samuel D Relton, David C Wong, Taimour Alam, and Jane E Alty. Accuracy of smartphone video for contactless measurement of hand tremor frequency. *Movement Disorders Clinical Practice*, 8(1):69–75, 2021.

[24] Rocco Agostino, Antonio Currà, Morena Giovannelli, Nicola Modugno, Mario Manfredi, and Alfredo Berardelli. Impairment of individual finger movements in parkinson's disease. *Movement disorders*, 18(5):560–565, 2003.

[25] Milica Djurić-Jovičić, Nenad S Jovičić, Agnes Roby-Brami, Mirjana B Popović, Vladimir S Kostić, and Antonije R Djordjević. Quantification of finger-tapping angle based on wearable sensors. *Sensors*, 17(2):203, 2017.

[26] Susanna Summa, Jacopo Tosi, Fabrizio Taffoni, Lazzaro Di Biase, Massimo Marano, A Cascio Rizzo, Mario Tombini, Giovanni Di Pino, and Domenico Formica. Assessing bradykinesia in parkinson's disease using gyroscope signals. In *2017 international conference on rehabilitation robotics (ICORR)*, pages 1556–1561. IEEE, 2017.

[27] Taha Khan, Dag Nyholm, Jerker Westin, and Mark Dougherty. A computer vision framework for finger-tapping evaluation in parkinson's disease. *Artificial intelligence in medicine*, 60(1):27–40, 2014.

[28] Kjersten Criss and James McNames. Video assessment of finger tapping for parkinson's disease and other movement disorders. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 7123–7126. IEEE, 2011.

[29] David C Wong, Samuel D Relton, Hui Fang, Rami Qhawaji, Christopher D Graham, Jane Alty, and Stefan Williams. Supervised classification of bradykinesia for parkinson's disease diagnosis from smartphone videos. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 32–37. IEEE, 2019.

[30] Jill Schmidt, Devin R Berg, Heidi-Lynn Ploeg, and Leone Ploeg. Precision, repeatability and accuracy of optotrak® optical motion tracking systems. *International Journal of Experimental and Computational Biomechanics*, 1(1):114–127, 2009.

[31] Carlo M Contreras, Gregory A Metzger, Joal D Beane, Priya H Dedhia, Aslam Ejaz, and Timothy M Pawlik. Telemedicine: patient-provider clinical engagement during the covid-19 pandemic and beyond. *Journal of Gastrointestinal Surgery*, 24(7):1692–1697, 2020.

[32] Bhaskar Roy, Richard J Nowak, Ricardo Roda, Babar Khokhar, Huned S Patwa, Thomas Lloyd, and Seward B Rutkove. Teleneurology during the covid-19 pandemic: a step forward in modernizing medical care. *Journal of the neurological sciences*, 414:116930, 2020.

[33] Stefan Williams, Zhibin Zhao, Awais Hafeez, David C Wong, Samuel D Relton, Hui Fang, and Jane E Alty. The discerning eye of computer vision: Can it measure parkinson's finger tap bradykinesia? *Journal of the Neurological Sciences*, 416:117003, 2020.

[34] Stefan Williams, Samuel D Relton, Hui Fang, Jane Alty, Rami Qahwaji, Christopher D Graham, and David C Wong. Supervised classification of bradykinesia in parkinson's disease from smartphone videos. *Artificial Intelligence in Medicine*, 110:101966, 2020.