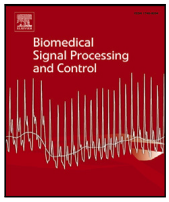
Hand tracking for clinical applications: Validation of the Google MediaPipe Hand (GMH) and the depth-enhanced GMH-D frameworks

Biomedical Signal Processing and Control 96 (2024) 106508

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Hand tracking for clinical applications: Validation of the Google MediaPipe Hand (GMH) and the depth-enhanced GMH-D frameworks

Gianluca Amprimo [a,b,*], Giulia Masi [a], Giuseppe Pettiti [b], Gabriella Olmo [a], Lorenzo Priano [c,d], Claudia Ferraris [b,**]

[a] Politecnico di Torino - Control and Computer Engineering Department, Corso Duca degli Abruzzi, 24, Turin, 10129, Italy
[b] National Research Council - Institute of Electronics, Information Engineering and Telecommunications, Corso Duca degli Abruzzi, 24, Turin, 10029, Italy
[c] Università di Torino - Neurosciences Department "Rita Levi Montalcini", Via Verdi, 8, Turin, 10124, Italy
[d] Istituto Auxologico Italiano – Department of Neurology and Neurorehabilitation S. Giuseppe Hospital, Via Cadorna 90, Oggebbio, 28824, Italy

ARTICLE INFO

Keywords:
Hand tracking
Deep learning
Google mediaPipe
Azure kinect
Finger tapping
Hand dexterity assessment
RGB
RGB-depth

ABSTRACT

Accurate 3D tracking of hand and fingers movements poses significant challenges in computer vision. The potential applications span across multiple domains, including human–computer interaction, virtual reality, industry, and medicine. While gesture recognition has achieved remarkable accuracy, quantifying fine movements remains a hurdle, particularly in clinical applications where the assessment of hand dysfunctions and rehabilitation training outcomes necessitate precise measurements. Several novel and lightweight frameworks based on Deep Learning have emerged to address this issue; however, their performance in accurately and reliably measuring finger movements requires validation against well-established gold standard systems. In this paper, the aim is to validate the hand-tracking framework implemented by Google MediaPipe Hand (GMH) and an innovative enhanced version, GMH-D, that exploits the depth estimation of an RGB-Depth camera to achieve more accurate tracking of 3D movements. Three dynamic exercises commonly administered by clinicians to assess hand dysfunctions, namely hand opening–closing, single finger tapping and multiple finger tapping are considered. Results demonstrate high temporal and spectral consistency of both frameworks with the gold standard. However, the enhanced GMH-D framework exhibits superior accuracy in spatial measurements compared to the baseline GMH, for both slow and fast movements. Overall, our study contributes to the advancement of hand tracking technology, and the establishment of a validation procedure as a good-practice to prove efficacy of deep-learning-based hand-tracking. Moreover, it proves that GMH-D is a reliable framework for assessing 3D hand movements in clinical applications.

1. Introduction

Monitoring, recognising, and interpreting the natural movement of the body, without the aid of devices and instrumentation that can alter its characteristics, are among the most currently addressed research topics in Computer Vision (CV) [1,2] for a multitude of scientific and consumer applications [3,4]. In fact, Human Pose Estimation (HPE) and Human Action Recognition (HAR) are finding their way into the fields of human–computer interaction, virtual reality, robotics, sports, video surveillance, industry, biomechanics, and medicine [1,5–12]. Despite advances in the recognition and estimation of static or quasi-static poses and gestures [13], accurate tracking and measurement of motion characteristics is still an open challenge, mainly when focusing on small body parts such as the hand and fingers [14–16].

The human hand has a complex and fully articulated anatomical structure suitable for performing coarse and fine-grained movements. The development of real-time, robust, non-invasive, cost-effective, and accurate algorithms for tracking human hand and finger movements is complex [17], and constraints are often established according to specific needs [18,19]. After several attempts to develop contact-based solutions for constrained hand tracking using wearable devices and supporting aids (such as instrumented gloves) [20–23], the first bare-hand solutions were proposed [24–28]. In most cases, the latter were limited only to offline processing or low frame rate, thus preventing

* Corresponding author at: Politecnico di Torino - Control and Computer Engineering Department, Corso Duca degli Abruzzi, 24, Turin, 10129, Italy.
** Corresponding author.
E-mail address: gianluca.amprimo@polito.it (G. Amprimo).
URLs: https://www.polito.it/personale?p=gianluca.amprimo (G. Amprimo), https://www.researchgate.net/profile/Giulia-Masi-2 (G. Masi), https://www.ieiit.cnr.it/people/Pettiti-Giuseppe (G. Pettiti), https://www.sysbio.polito.it/analytics-technologies-health/ (G. Olmo), https://neuroen.campusnet.unito.it/do/docenti.pl/Alias?lorenzo.priano#tab-profilo (L. Priano), https://www.ieiit.cnr.it/people/Ferraris-Claudia (C. Ferraris).

https://doi.org/10.1016/j.bspc.2024.106508
Received 21 August 2023; Received in revised form 24 April 2024; Accepted 24 May 2024
Available online 4 June 2024

their practical use where real-time motion capture and analysis are required, such as for clinical applications.

Bare-hand tracking through vision systems has attracted much interest as it may overcome the main limitations of contact-based solutions: interfering effects on natural movements, discomfort, and bulkiness. CV techniques have been extensively investigated to capture bare-hand movements from videos, including skin colour segmentation and mean shift algorithms and its variants [29,30]. The availability of early RGB-D cameras quickly led to a more comprehensive 3D analysis by exploiting the potential of distance estimation through depth maps [31–33]. The first complete 3D hand skeletal models appeared, which have been successfully applied, for instance, in various clinical studies [34–37], albeit with specific performance-related constraints [38,39]. The current explosion of DL in CV tasks, hand-tracking included, has contributed to the further development of these methods. Indeed, recent frameworks for *in-the-wild* (i.e., unconstrained) hand tracking from RGB or RGB-D cameras leveraging DL showcase great potential, especially in solving self-occlusions.

From a clinical perspective, the study of hand motion is particularly relevant. Indeed, hands play a crucial role in daily life tasks as they are pivotal for interacting autonomously with the environment, other humans, and machines. The fine motor control of fingers derives from complex neuronal networks, which leverage excitation and inhibition pathways to generate hand dexterity [40]. It is not surprising then that several neurodegenerative pathologies such as Parkinson's disease (PD), Ataxia, Amyotrophic Lateral Sclerosis, and acute neurological events such as stroke have evident manifestations in this body district [41–44]. For instance, in PD, testing hand functionality is crucial for assessing symptoms such as bradykinesia and tremor. Clinicians usually conduct this evaluation during outpatient visits by performing a qualitative scoring of motor tasks such as SFT, OC, or pronation-supination of the hand. These exercises, whose standard scoring is part of scales such as the Movement Disorders Society's revision of Unified Parkinson's Disease Rating Scale (MDS-UPDRS), provide an easy tool for observing and assessing the symptoms of PD. However, this subjective clinical evaluation may be affected by intra- and inter-rater variability [45]. This limitation has drawn the attention of researchers toward a more quantitative perspective, using new technologies such as video-based hand tracking and inertial sensors for hand movement assessment. In addition, rehabilitation of hand functionality is central for patient recovery and independence, especially after acute events such as stroke. New rehabilitation paradigms, e.g., exergaming [46,47], require new technologies for patient-computer interaction through hand movements, such as robotics [48] or *smart* gloves that embed electromyographic and inertial sensors [47]. In this scenario, video-based methods, enhanced by DL, may provide an alternative non-invasive and easy-to-use approach to implement these rehabilitation strategies.

Nevertheless, the clinical acceptability of these innovative 3D hand tracking frameworks requires objective evidence of high accuracy and reliability [49]. This validation should leverage the comparison of tracking performance against gold reference systems for human movement analysis, such as motion capture systems, rather than other manual instrumentation [50]. This rigorous procedure, however, is still rare in DL-based hand tracking for clinical applications.

This paper presents a validation procedure against a gold standard system of two candidate frameworks for the aforementioned clinical applications, namely GMH [51] and GMH-D [52]. The latter is an enhanced version of GMH that runs on top of a RGB-D camera, which provides simultaneous and calibrated colour and depth video streams. This validation considers three standard tasks taken from clinical examination and rehabilitation of hand motor functions in subjects with PD to compare the two frameworks. The main innovative contributions of this work are the following:

- to validate the accuracy and reliability of basic GMH and enhanced GMH-D frameworks against measurements obtained by a motion capture system, in terms of 3D trajectories and estimated spatial, temporal, and spectral features;
- to compare the suitability of GMH and GMH-D to track hand movements during the three selected dynamic exercises, also verifying their adaptability to different scenarios;
- to establish good-practice guidelines for the validation of DL-based 3D hand tracking frameworks, since most of the current solutions lack a proper validation as a measurement system (see Section 2.2).

This work unfolds as follows. Section 2 provides an overview of recent DL frameworks for hand tracking, focusing on solutions already applied in clinical scenarios. Section 3 describes the validation protocol and experimental setup. Section 4 presents the methodological approach to analyse and validate the frameworks. Section 5 reports and discusses the main results of the experimental tests. Finally, Section 6 illustrates some final remarks.

2. Background

2.1. DL for in-the-wild hand tracking

DL approaches for hand tracking from *in-the-wild* video sequences can be organised in a taxonomy, according to their input modality: RGB, depth map, or mixed RGB-D [53]. Researchers investigated depth approaches to allow 3D reconstruction of the hand following the increase in market availability of depth sensors. In the most common architecture, Convolutional Neural Networks (CNN) process depth data to extract hand tracking information [54–56], possibly enforcing kinematics-based rules to improve the estimation [57]. Even if accurate, depth methods have downsides such as large energy consumption, poor form factor, poor near-distance coverage, and limited outdoor usage due to light interaction with ToF technology [58].

In multimodal methods (RGB-D), either the RGB stream identifies the hand in two dimensions (2D) and then the associated depth stream allows to uplift joints [16], or the two modalities are fused to perform a single-shot estimation [59,60]. The mixed RGB-D modality is also frequently used to augment the training of DL models performing inference on RGB-only data [53].

Concerning this last modality, 2D hand landmarks extraction is often part of many state-of-the-art HPE estimation methods such as OpenPose [61] and AlphaPose [62]. Few approaches perform 2D tracking through dedicated architectures [63,64] since it limits any analysis to hand movements happening for the most in a planar projection. Studying more complex gestures using these methods requires multiple-camera setups to perform geometrical triangulation [65], thus *uplifting* coordinates from 2D to 3D. This approach, however, increases the complexity and the cost of the final acquisition system.

Several recent works have focused on the challenging task of directly estimating 3D coordinates from depth clues in monocular RGB videos. Since the first work from Zimmerman et al. [66], many architectures have been investigated [67–69]. However, these works report little information about the efficiency [70,71], or claim real-time performance (>30 fps) without providing code to reproduce the results [72]. Moreover, to achieve top-tier accuracy on benchmark datasets, these methods often exploit high-performance GPUs, which makes these solutions infeasible for applications outside research laboratories.

2.2. DL hand tracking in clinical applications

Despite the numerous recent works investigating *in-the-wild* 3D hand tracking from a single RGB camera, only a few clinical applications implement these novel DL models or similar custom solutions [73–78]. Indeed, most of the clinical investigations still utilise simpler but well-established tools such as OpenPose [42,79–86], GMH [87–90], and DeepLabCut [91–96].

Works exploiting OpenPose and DeepLabCut derive 2D hand joints so the subsequent motor assessment is limited to parameters easily

retrievable from a planar view (e.g., temporal/frequency parameters of motion or angles between joints). Custom DL models often infer 3D hand coordinates by employing backbone networks pre-trained on popular *in-the-wild* hand tracking datasets, possibly fine-tuning on data collected for the specific clinical study. These methods typically require GPU acceleration, often with low throughput (<30 fps) during hand joint estimation. The popular choice for the medical field of simple, out-of-the-shelf methods with respect to state-of-the-art, and *in-the-wild* DL approaches, seems to be caused by the complexity of the best-performing networks, which can be hard to replicate and apply in practice. Moreover, clinical applications often do not require actual *in-the-wild* tracking since acquisition settings are typically standardised and within a controlled scenario, favouring the usage of simpler but effective solutions.

Furthermore, it must pointed out that most of these frameworks are solely validated as a component inside a broader automatic pipeline for impairment prediction (e.g., automatic PD rating). This soft validation verifies the consistency of the final predictions from the entire assessment pipeline against qualitative clinical examinations [73–78], but it does not provide a quantitative measure of the accuracy and reliability of the hand tracking framework as an actual measurement tool. The final predictions may be biased by the data collected specifically for the study, and the fitting of the prediction model at the end of the whole pipeline. However, without validation is impossible to establish whether poor prediction performance depends on these factors or on low-quality hand tracking at the source and whether these systems are identifying a reliable evidence of impairment in the subjects' parameters. Understanding these aspects is essential to increase acceptability of these solutions by clinical personnel, who are often hesitant toward approaches for which a high degree of interpretability and trust cannot be provided [49].

Only one study [83] validated OpenPose performance as a measurement tool in the SFT task by analysing joint angles during task execution. Another study [93] validated 2D poses given by DeepLabCut with respect to movement frequency using an Optotrack motion capture system. A similar approach, focusing on resting tremor in PD, was done by [87] for GMH. However, no 3D tracking-based work provides performance validation compared to video-based motion capture, i.e., the gold standard for human motion analysis. This type of validation is especially significant for 3D approaches, since it may compare spatial parameters in terms of relative distances between hand joints rather then only in terms of angles between them.

2.3. A focus on GMH and GMH-D

GMH is a DL approach based on RGB input for hand tracking, included in MediaPipe [51], the solution for light-weight and portable Machine Learning (ML) pipelines by Google LLC. The GMH framework is composed of two sub-modules: a Palm Detection (PDM) module and a Hand Landmarks Detection (HLD) model. First, the PDM identifies the region of interest corresponding to the hand, then the HLD detects the 21 key points corresponding to hand joints within it. Fig. 1 summarises the coordinate systems provided by GMH. As it can be observed, the framework provides both *Image Coordinates* in pixels, coupled with a dimensionless parameter $z_{j,im}$ that estimates the relative depth of the joint j with respect to the wrist reference, and a set of 3D *World Coordinates*, expressed in metres and centred in the bounding box of the palm detected by PDM.

This framework balances accuracy and time efficiency. Indeed, it supports a frame rate in excess of 50 fps on a Google Pixel 6 phone using CPU only, or even faster (>80 fps) exploiting GPU acceleration, as reported by the pipeline official web page [97]. This aspect is indeed crucial to developing easy-to-use and widely employable assessment systems. Regarding its application to the medical field, it has been used to identify [3] and measure [87] resting tremor in Parkinson's disease, validating its accuracy against that of an accelerometer. Moreover,

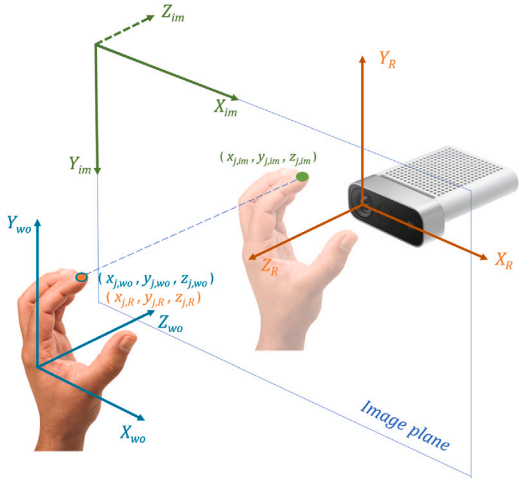


Fig. 1. Set of coordinates tracked by GMH and GMH-D: for GMH, in green *Image Coordinates* (pixels), centred in the upper left corner of the image; in blue, *World Coordinates* (metres) centred in the middle of the detected palm. In orange, the *Real-World Coordinates* (metres) estimated by the GMH-D framework, centred in the RGB-D recording camera. For *Image Coordinate* of GMH, axis Z_{im} expresses an adimensional depth parameter, relative to the wrist and scaled as the other two axes.

in [90], GMH was used to measure finger excursion from static images and was validated using standard manual goniometry. However, the lack of dynamic and continuous tracking and an acquisition protocol with several restrictions on hand positioning and environmental conditions limit the significance of these validation procedures to a very narrow scenario.

Another study performed GMH validation for evaluating finger tapping and hand opening–closing clinical tasks [52]. The study also proposed an enhanced version of the framework exploiting an RGB-D camera (i.e., MAK), namely GMH-D. According to [52], GMH-D showed time performance comparable to GMH but enhanced 3D tracking accuracy by leveraging both the depth estimation performed by the DL model and the depth-map provided by the RGB-D camera. Indeed, the depth estimation for each joint ($\hat{d}_j$, Eq. (1)) is derived from the depth value of the wrist as measured by the on-board depth sensor (d_{wrist}) and the estimation done by the neural network ($z_{j,im}$, refer to Fig. 1).

$$\hat{d}_j = d_{wrist} + z_{j,im} d_{wrist} \tag{1}$$

The wrist is the origin of the reference system of GMH and the most stable joint since it is tracked within a body surface much larger than the fingers. These characteristics reduce the likelihood of errors when retrieving its depth value from the depth map provided by the depth sensor, which could depend on virtual marker misplacement by GMH or boundary interference due to motion [52]. The authors validated the improvement provided by GMH-D over GMH by comparing the measurement of maximum and minimum peaks in the distance between relevant hand joints during SFT and OC movements [52]. They achieved this by using a ruler placed close to the hand during the execution of the movements to retrieve real-world distances from a offline video analysis. This preliminary validation, although achieving promising results, provided a limited estimation of the quality of the tracking over the complete tasks, for which the support of a motion capture system is required.

For these reasons, the goal of this work is to further expand the previously obtained results, since GMH-D and GMH seem to be a promising marker-less and non-invasive solutions for clinical assessment of hand and finger motion due to their stability, easiness of deployment, and low-computational power.

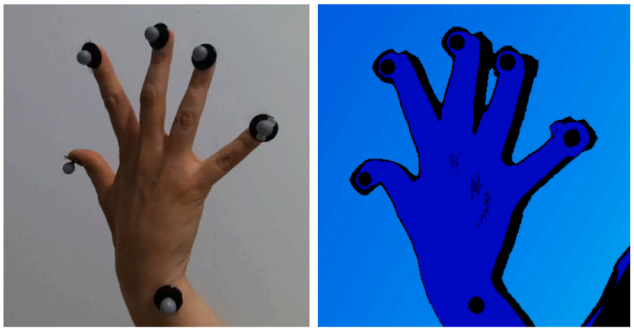


Fig. 2. On the left, hand appearance when applying minimal marker configuration for motion capture (only the tips of fingers and the wrist reference); on the right, the same hand as seen by the depth sensor of the RGB-D camera. As it can be observed, passive-reflective markers produce holes in the estimated depth map.

3. Setup and validation protocol

3.1. Challenges of motion capture validation

Using motion capture for the validation of video-based HPE methods is a well-established practice [98,99] since marker configuration does not excessively alter the appearance of the complete body shape nor impairs the motor performance of the subject. However, the same validation is more cumbersome in the case of hand tracking, especially using RGB-D cameras. Indeed, the density of markers required to track all the degrees of freedom of the hand is much higher than for validating HPE methods. Therefore, complete hand coverage with markers inevitably alters its appearance, likely causing a reduction of accuracy in the DL model (Fig. 2, left image). Moreover, an excessive number of markers could reduce mobility, limiting the possible tasks to validate.

In addition, passive reflective markers cause holes in the depth map provided by RGB-D cameras, as shown in Fig. 2 (image on the right). Consequently, marker placement and recording camera viewing angle should be considered carefully, to avoid estimation errors due to such phenomenon.

Finally, time synchronisation between the motion capture system and the device running the DL framework is mandatory to avoid jitters in the two recording streams that could make realignment between trajectories unfeasible. When dealing with RGB-D cameras that work in the same infrared spectrum of motion capture, synchronisation is also crucial to avoid interference that may lead to wrong depth estimation by the RGB-D camera. All these items were taken into account in defining the setup and validation protocol for this study.

3.2. Validation setup

Data acquisition sessions were organised at the Engineering for Health and Well-Being (EHW) Laboratory of the National Research Council (Institute of Electronics, Information Engineering, and Telecommunications) in Turin, where a gold-standard motion capture system is available. The system is an OPT solution with six Prime13 cameras (1280 × 1024 px resolution). OPT cameras operate at 120 fps, covering a working volume of approximately 6 × 4 × 3 m^3. The system was calibrated before each acquisition session, obtaining a residual value of 0.6 mm, which is an average offset distance between the converging rays when reconstructing a marker; hence, it is related to the OPT reconstruction precision. The final estimated measurement error was less than 2.8 mm. Reflective markers of size 25 mm were exploited. All tests were performed in the central zone of the recorded volume, thus ensuring maximum tracking accuracy. MAK was positioned in this area, stably fixed on a tripod (1 m high), to capture videos of the participants' performance. MAK was connected to a laptop (Alienware

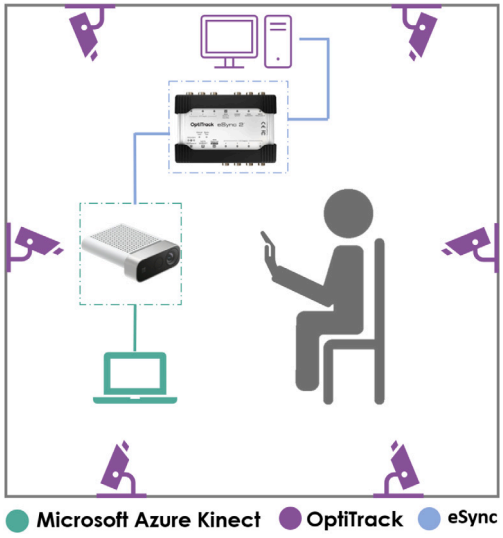


Fig. 3. Setup for the data acquisition sessions: an OptiTrack system (OPT) with 6 cameras is employed together with the Microsoft Azure Kinect DK (MAK). The Optitrack eSync 2 device is used for synchronisation. The subjects are told to seat in front of MAK in the centre of the working volume of the motion capture system and perform the assessment tasks.

m15 R2 I7-9750H, 16 GB RAM, NVIDIA GeForce RTX-2070 MaxQ with 8 GB of GDDR6). Fig. 3 illustrates the complete experimental setup.

The two systems were synchronised using a sync generator (OptiTrack eSync2). The eSync2 was configured to operate as a *master* sync generator by driving MAK with a 30 Hz sync signal and synchronising the OPT cameras at 120 fps using the internal 4x frequency multiplier.

3.3. Selected tasks and participants

The validation of GMH and GMH-D considered three reference movements: OC of the hand, SFT, and MFT. SFT consists of repeatedly tapping the thumb and index fingers. MFT consists of repeatedly and sequentially tapping the index, middle, ring, and little finger against the thumb.

These tasks are dynamically challenging and commonly used to measure fine hand dexterity and motor dysfunctions in the elderly and in subjects with chronic conditions (i.e., stroke) [100–103]. In particular, the SFT task is frequently addressed in works that apply DL hand tracking for PD diagnosis and staging through an automatic assessment pipeline [73,76,81,83,88,89,91,92]. In addition to the selected exercises, an initial SOH phase was also recorded to extract participants' hand size.

As this study only aimed to validate the frameworks, healthy adult volunteers were involved. Specifically, ten subjects (4 females, 6 males), age 31.10 ± 7.80 years old were recruited. Average hand length (from middle finger tip to wrist, as retrieved from SOH task) for the male and female groups was 18.92 ± 0.83 cm and 16.5 ± 0.78 cm, respectively. None of the participants had physical hand/wrist/arm problems that could prevent them from performing the planned tasks. The experimental study was organised according to the Declaration of Helsinki (1964) and the latest amendments, supervised by a clinician, and approved by the local Ethical Committee of Istituto Auxologico Italiano. Each participant signed an informed consent after receiving details on the study purposes and instrumentation.

3.4. Validation protocol

Table 1 summarises the protocol and the number of trials acquired for each task. Each recorded trial lasted 15 s. Eventually, a dataset

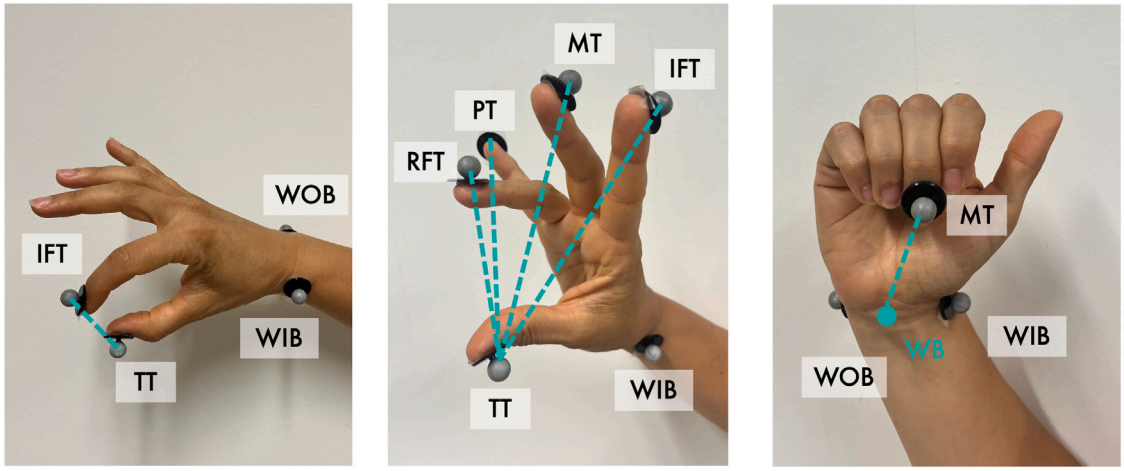


Fig. 4. Marker configuration for the three tasks: on the left, marker set for SFT involves only wrist, index and thumb tips; in the centre, marker set for MFT involves all finger tips and the wrist; on the right, marker set for OC and SOH involves the wrist reference and the middle finger tip.

composed of 200 videos was collected considering all the three tasks, plus 10 initial SOH trials. The validation procedure considered three significant influencing factors:

- **Distance from camera**: the displacement from the camera could affect how the underlying DL model identifies depth clues in the videos.
- **Velocity of motion**: in low frame rate recording devices (<60 fps), motion blur alters the appearance of fingers [52]. This alteration can produce inaccurate virtual marker positioning and distorted hand reconstruction.
- **Camera viewing angle**: especially in the case of SFT, different camera perspectives can modify the number of self-occluded joints during motion.

Regarding distance from the camera, two ranges were considered: NEAR distance, between 60 cm and 80 cm from MAK, and FAR distance, between 80 cm and 100 cm from MAK. This factor was studied in all tasks. About velocity of motion, three speeds were investigated by coordinating the task execution with the rhythm of a metronome. For SFT and OC, low speed at 75 beats per minute (bpm), normal speed at 115 bpm, and high speed at 140 bpm were considered. Subjects were also asked to achieve different ranges of motions, compatible with the requested speed: slow speed-wide excursion, normal speed-free excursion, and high speed-small excursion. This request ensured variability in the movements performed for the same task. For MFT, only the normal speed (115 bpm) was considered since correct movement execution at high speed is complex even for healthy subjects. Finally, the camera viewing angle was studied for SFT only, since both a lateral and a frontal perspective may capture the movement with different degrees of self-occlusion. In contrast, for OC and MFT only frontal viewing angle is feasible, since in the lateral positioning the tracking of the palm by GMH is challenging and likely to result in poor accuracy.

A minimal marker configuration for motion assessment was selected for each task to reduce its impact on tracking. Physical markers were placed on the back of the hand to avoid depth map holes and in close correspondence to the positions where virtual joints of GMH should lie to limit the systematic positioning error with respect to OPT. For SFT (Fig. 4, left image), Wrist Outer Bone (WOB), Wrist Inner Bone (WIB), IFT, and TT were selected. Wrist markers provide a reference for the hand structure, while IFT and TT joints are those actively involved and whose relative distance is leveraged in the literature for kinematic assessment of the task. For MFT (Fig. 4, middle image), MT, RFT, and Pinkie Tip (PT) were also marked to evaluate the relative distance between all fingers tips and TT along the whole task. Finally, for OC and SOH (Fig. 4, right image) just WOB, WIB, and MFT markers were applied.

Table 1
To validate the two frameworks, a total of 200 videos lasting each around 15 s were recorded. Ten participants performed 3 different hand dexterity tasks: hand opening and closing (OC), single finger tapping (SFT) and multi-finger tapping (MFT). Requested speed (slow, normal, and fast) of execution, distance from the camera (near, far), and viewing angle (frontal, lateral) were modified in order to explore the performances of GMH and GMH-D as they vary. The number of videos recorded in each set-up is reported in the table as speed, distance, and viewing angle vary.

	Distance	Viewing angle	Speed			TOT
			Slow 75 bpm	Normal 115 bpm	Fast 140 bpm	
OC	Near 60–80 cm	Frontal	20	10	20	50
	Far 80–100 cm	Frontal	/	10	/	10
SFT	Near 60–80 cm	Lateral	20	10	20	50
		Frontal	20	20	20	60
	Far 80–100 cm	Lateral	/	10	/	10
MFT	Near 60–80 cm	Frontal	/	10	/	10
	Far 80–100 cm	Frontal	/	10	/	10
TOT			60	80	60	200

4. Comparison methods and metrics

The comparison between GMH, GMH-D, and OPT focused on the relative spatial distance between fingers whose motion mainly describes the task to assess. Considering only inter-fingers distances does not require complex calibrations among the tracking frameworks since relative distances are invariant to translation and rotation of the reciprocal reference systems. Fig. 4 visualises all the investigated spatial distances as dotted lines. For OC, MT-WB distance was evaluated, with WB as the middle point between WIB and WOB markers of OPT. For SFT, the IFT-TT distance was considered. For MFT, IFT-TT, MT-TT, RFT-TT, and PT-TT distances were studied, plus a virtual overall trajectory defined as the sum of these sub-trajectories (TT-ALL).

4.1. Whole-trajectory comparison

Before comparison, raw inter-fingers distances measured by GMH, GMH-D, and OPT were realigned. The vertical offset, due to misalignment between real and virtual markers, and the horizontal offset, due to residual time-shift between MAK and OPT, were removed. To

vertically realign the OPT trajectory with those of GMH-D and GMH, the mean distance between each point was evaluated and subtracted to OPT trajectory (physical markers lie on top of virtual markers being attached to the finger-tips). The residual temporal shift, instead, was automatically removed using a cross-correlation method [104].

After this procedure, the selected distances as computed by OPT, GMH, and GMH-D were compared in terms of RMSE and its relative version, the PRMSE, for a fairer comparison among trials with different finger excursions. This metric is defined in Eq. (2) as:

$$PRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_{OPT} - \hat{Y}}{Y_{OPT}}\right)^2} \times 100\% \quad (2)$$

where Y_{OPT} is the measurement obtained from OPT, whereas $\hat{Y}$ is the measurement estimated by either GMH or GMH-D.

In addition, Pearson's correlation coefficient ρ was used to measure the coherence between the curves measured by GMH and GMH-D and the reference given by OPT.

4.2. Single-segments comparison

In the second step of the validation, a finer comparison involved the estimation of the Range of Motion (*ROM*) in cm and the time duration (*DUR*) in seconds of single repetitions (movements) of the task, (e.g., single taps of the TT and IFT joints in SFT). To this aim, a trivial segmentation algorithm was implemented in Matlab 2020b to identify each movement segment as the portion of trajectory between two consecutive local minima, containing a single local maximum inside. Since all the investigated movements have a periodic nature, this segmentation procedure was possible for all tasks. In the case of MFT, segmentation was applied only to the TT-ALL distance, which should summarise the motion of all involved fingers.

For each task, all the collected videos were considered, to achieve a dataset containing 1430, 1944, and 482 single segments of movement respectively for OC, SFT, and MFT. Segment-level parameters (*ROM* and *DUR*) in each task were compared between OPT and GMH/GMH-D using Bland–Altman plots. In such plots, the difference in the estimations by two measurement systems is compared to the mean value between them (*bias*). To indicate a good level of agreement, around 95% of the points should fall inside the Limits of Agreement (LoA), defined as ±1.96SD, and this range should be sufficiently small given the desired application. Moreover, CCC [105] and the ICC [106] were estimated to measure the level of agreement. In particular, CCC is an alternative metric to ICC, for assessing inter-rater variability between measurement systems [12,107] and it is defined in Eq. (3) as

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (3)$$

where ρ is Pearson's correlation coefficient between random variables X and Y (either parameters from OPT and GMH or from OPT and GMH-D), μ and σ are respectively the mean value and the standard deviation of the distributions of X and Y. Following [12,107,108], a value over 0.8 denotes high agreement between the two systems. For ICC, a threshold of 0.75 detects high agreement between different raters [106].

Finally, to provide an overview of the tasks in the spectral domain, the dominant frequency of the voluntary movement spectral band (F_{DOM}) was identified together with its associated spectral power (POW_{DOM}). A comparison of F_{DOM} and POW_{DOM}, as estimated by GMH, GMH-D, and OPT in each trial, was carried out using Bland–Altman plots, ICC, and CCC.

5. Results and discussion

The results of the validation procedure are organised according to motion complexity. The OC task was considered the least complex because all fingers move together. SFT follows, as it involves the coordinated movement of two specific fingers. Finally, MFT, which involves dynamic and coordinated movement of all fingers.

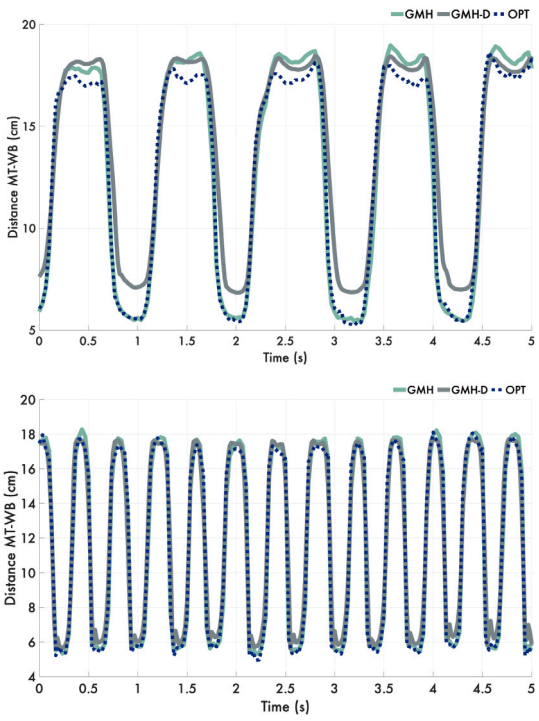


Fig. 5. MT-WB distance during slow (top) and fast (bottom) OC trials as measured by GMH and GMH-D with respect to the gold standard OPT (dotted line). The three curves have been vertically and horizontally realigned for a direct comparison.

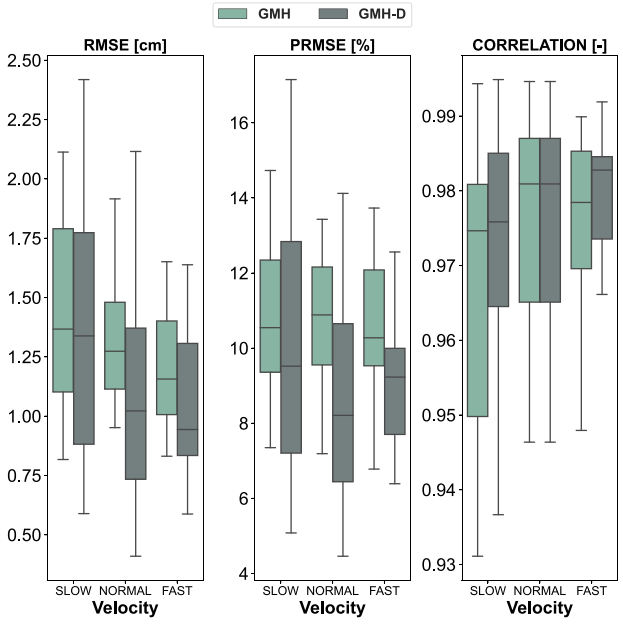


Fig. 6. RMSE (left), PRMSE (middle), and Pearson's ρ (right) box plots, in trials at different velocity: Slow (75 bmp), Normal (115 bpm), Fast (140 bpm).

5.1. OC task validation

Fig. 5 reports an example of the estimation of the MT-WB distance performed by GMH (green) and GMH-D (grey) with respect to OPT (dotted line), considering a slow (top) and a fast (bottom) execution. As it can be observed, both methods reconstruct with good precision the inter-fingers distance, with perfect temporal alignment.

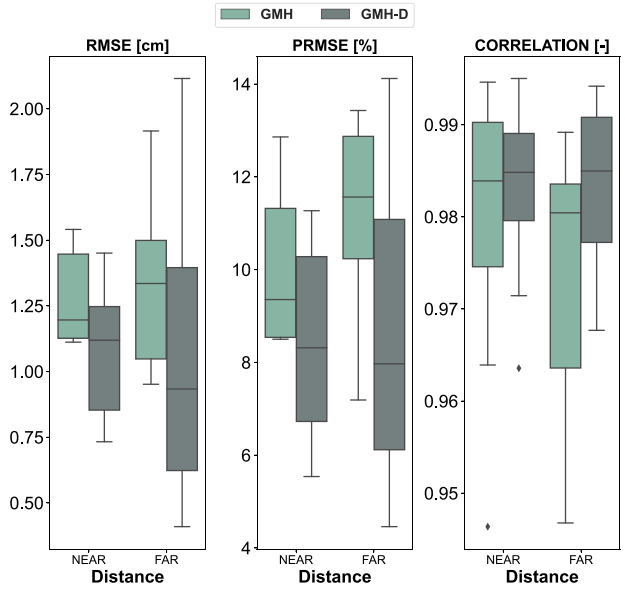


Fig. 7. RMSE (left), PRMSE (middle), and Pearson's ρ (right) box plots, in trials at different distances from camera: NEAR distance (60–80 cm) and FAR distance (80–100 cm).

Results on RMSE, PRMSE, and Person's ρ are organised separately according to velocity and distance factors, in Figs. 6 and 7 respectively. Both in terms of RMSE and PRMSE, GMH-D achieves a smaller error compared to GMH as the speed of motion increases, but the two methods are overall comparable and both with a high correlation with the OPT reference ($\rho > 0.93$ in all tests). Both methods appear to slightly worsen, with box plots for RMSE and PRMSE exhibiting larger interquartile ranges when passing from the NEAR to the FAR condition. However, correlations remains high, suggesting high temporal agreement even when spatial agreement reduces. Overall, the median error lies below 10% in the PRMSE for GMH-D and between 9.5%–11% for GMH, independently of motion velocity and distance from camera. These results overall denotes very good accuracy and reliability of both methods compared to OPT.

From the segment-level analysis, the Bland–Altman plots for the extracted parameters are reported in Fig. 8. As for ROM, it can be appreciated that 95.67% and 94.76% of estimations falls in the LoA, with a slightly narrower error range for GMH-D. The same holds true for *DUR*, with 96.02% for GMH-D and 92.66% for GMH of points inside the LoA and a mean difference of 0.0 s for both. For F_{DOM}, a perfect agreement with OPT is reached by both methods, as confirmed by ICC and CCC values. As for POW_{DOM}, GMH exhibits a narrower error range than GMH-D, but both methods have some large outliers outside the LoA. The investigation of these points revealed a delayed closing phase for GMH-D and GMH with respect to OPT, which may have produced a shift in the distribution of the spectral power among frequencies. These events may be connected to a temporary hardware desynchronisation between MAK and OPT. However, these events cannot be imputed to the two frameworks themselves and verified only in few trials thus such outliers could be reasonably neglected.

The results for ICC and CCC are reported in Table 2, using a 95% confidence level (low and high confidence ranges are reported). For the two metrics, *p-values* are all below $p < 0.001$. The ICC values highlight a high level of agreement (>0.90) for temporal and spectral properties, either using GMH or GMH-D. Slightly worse results, albeit still in excess of 0.8, are achieved for ROM, with GMH-D slightly outperforming GMH (0.89 vs. 0.81). The CCC confirms these results, pointing out a larger discrepancy between ROM when exploiting GMH-D versus GMH, hence favouring the first method. This is above the threshold

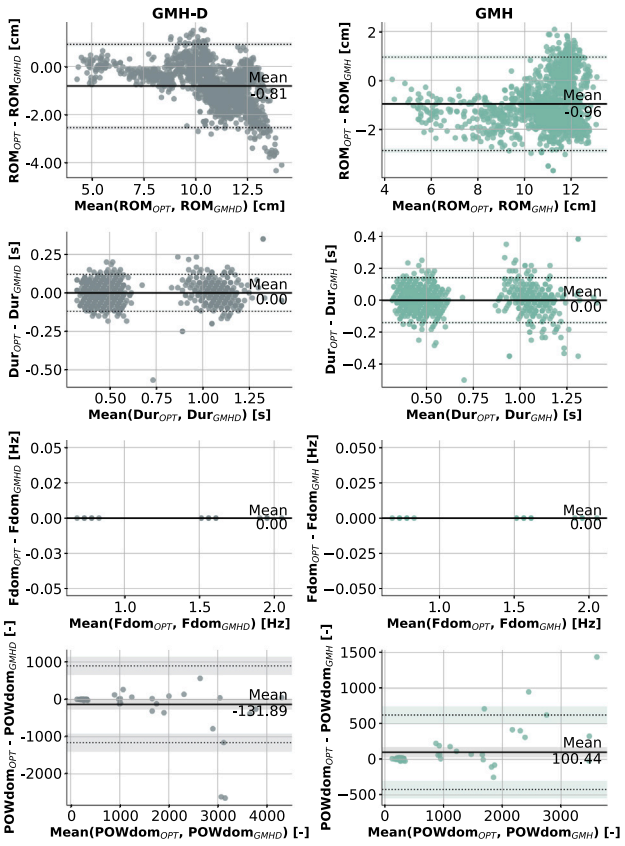


Fig. 8. Bland–Altman plots for ROM (top), *DUR* (middle), F_{DOM} and POW_{DOM} (bottom) estimated from single repetitions of the OC task. Colour coding for GMH and GMH-D is the same as in Fig. 7.

of 0.8, denoting nevertheless a good-to-excellent agreement. It must be considered that the discrepancy of ROM is related to the 10% PRMSE average discrepancy between the trajectories, but could also be affected by the trivial segmentation algorithm. Indeed, how minimum and maximum points are identified depends on the morphology of the MT-WB distance measured by the GMH and GMH-D and, consequently, could have an impact on the evaluated parameters.

5.2. SFT task validation

Fig. 9 reports an example of the IFT-TT distances by GMH and GMH-D with respect to OPT (dotted line), considering a slow (top) and a fast (bottom) execution. In contrast with OC, it is evident that GMH suffers from a squeezing effect in the estimation of ROM, which is attenuated for GMH-D. This alteration is magnified by the increase in the execution speed, with an evident error in the spatial tracking of the IFT and TT joints when they are in close contact. This effect was already identified in [52], where an error in the z_{WO} component (Fig. 1) of IFT and TT was observed during the task execution.

For SFT, the analysis of the velocity factor takes into consideration also the recording viewing angle (either lateral or frontal). The distribution of the RMSE, the PRMSE, and the Person's ρ values for the collected trials are shown in Fig. 10. From the box plots, the following considerations can be derived. A decrease in the error and a slightly better correlation is achieved when moving from the frontal to the lateral viewing angle for GMH-D, since this view possibly improves the evaluation of the depth by the ToF sensor of MAK. In addition, for GMH-D, the velocity of execution seems to have a marginal effect, especially considering PRMSE and the correlation. The slight change in RMSE is likely connected to the difference in the achieved ROM (wide

Table 2
Intraclass Correlation (ICC) and Lin's Concordance Correlation Coefficient (CCC) values for segment-level and frequency parameters in OC task, both for GMH and GMH-D methods with respect to the gold-standard OPT.

	GMH						GMH-D					
	ICC			*CCC*			*ICC*			*CCC*		
	Low Conf.	*Value*	*High Conf.*	*Low Conf.*	*Value*	*High Conf.*	*Low Conf.*	*Value*	*High Conf.*	*Low Conf.*	*Value*	*High Conf.*
ROM	0.79	**0.81**	0.82	0.68	**0.71**	0.72	0.88	**0.89**	0.90	0.80	**0.82**	0.83
DUR	0.96	**0.96**	0.97	0.96	**0.96**	0.97	0.97	**0.97**	0.98	0.97	**0.97**	0.98
F_{DOM}	1.00	**1.00**	1.00	1.00	**1.00**	1.00	1.00	**1.00**	1.00	1.00	**1.00**	1.00
POW_{DOM}	0.94	**0.96**	0.98	0.93	**0.96**	0.97	0.84	**0.90**	0.94	0.83	**0.90**	0.93

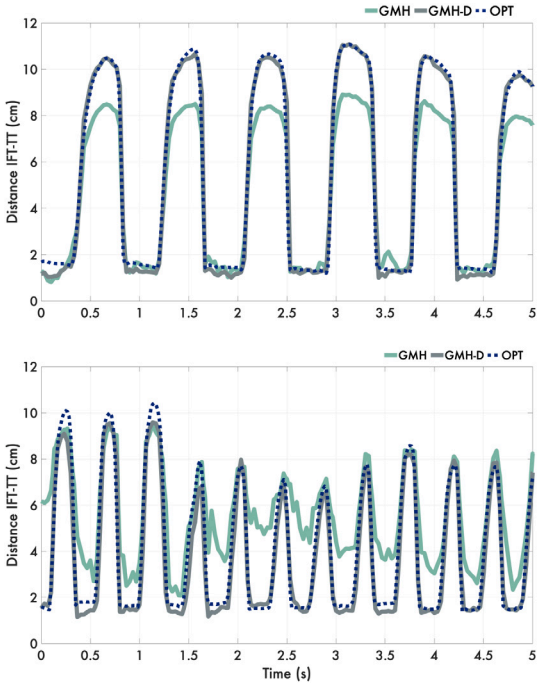


Fig. 9. IFT-TT distance during slow (top) and fast (bottom) SFT trials, as measured by GMH and GMH-D with respect to the gold standard OPT (dotted line). The three curves have been vertically and horizontally realigned to allow a direct comparison.

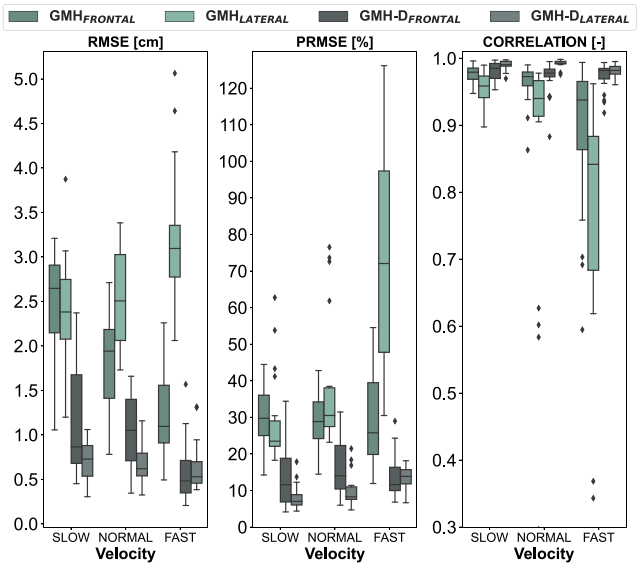


Fig. 10. Box plots of RMSE (top), PRMSE (middle), and Pearson's ρ (bottom) in trials at different velocity: low speed (75 bmp), normal speed (115 bpm), fast speed (140 bpm). Results are reported for the two studied recording viewing angle, either lateral ($GMH_{LATERAL}$, $GMH\text{-}D_{LATERAL}$) or frontal ($GMH_{FRONTAL}$, $GMH\text{-}D_{LATERAL}$).

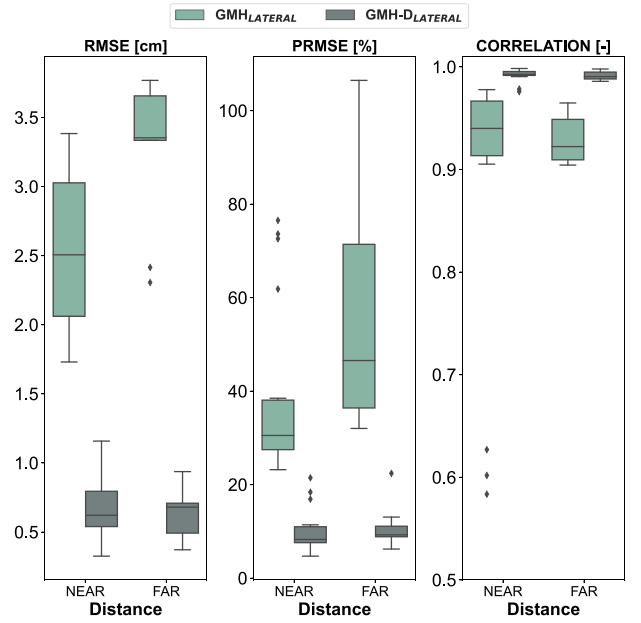


Fig. 11. RMSE (top), PRMSE (middle) and Pearson's ρ (bottom) box plots in SFT trials at different distance from recording camera: NEAR distance (60–80 cm) and FAR distance (80–100 cm). Only lateral viewing angle is considered for GMH ($GMH_{LATERAL}$) and GMH-D ($GMH\text{-}D_{LATERAL}$).

vs. small excursions) in the three type of trials, which does not affect instead PRMSE. Overall, the median value for this metric falls in the range 8%–15% for both perspectives.

Much greater error is measured for GMH both in the frontal and the lateral viewing angle (even four times more than that of GMH-D in the fast trials), with a median value much larger than 20% in all scenarios and a very large interquartile. While RMSE may be misleading, due to the natural reduction of *ROM* as velocity increases, PRMSE provides a more meaningful description. As it can be observed, the error amplifies in the lateral viewing angle and steadily increases while increasing the speed, whereas these effect is less evident in the frontal viewing angles. Correlation values confirm this reduction in accuracy both in the frontal and lateral viewing angles for GMH, with some large outliers suggesting low reliability at higher speeds. Overall, GMH-D provides the best tracking quality for this task, independently on the camera viewing angle.

For the sake of brevity, the remaining part of the analysis focuses on the differences between GMH and GMH-D for the lateral viewing angle alone -i.e., the one for which the smallest error and the largest correlations were achieved, considering all possible combinations of frameworks (GMH, GMH-D) and camera viewing angle (frontal, lateral).

First, the effect of distance from the camera is evaluated, by considering RMSE, PRMSE, and correlation. This result is reported in Fig. 11. GMH appears influenced by the distance factor, with an evident

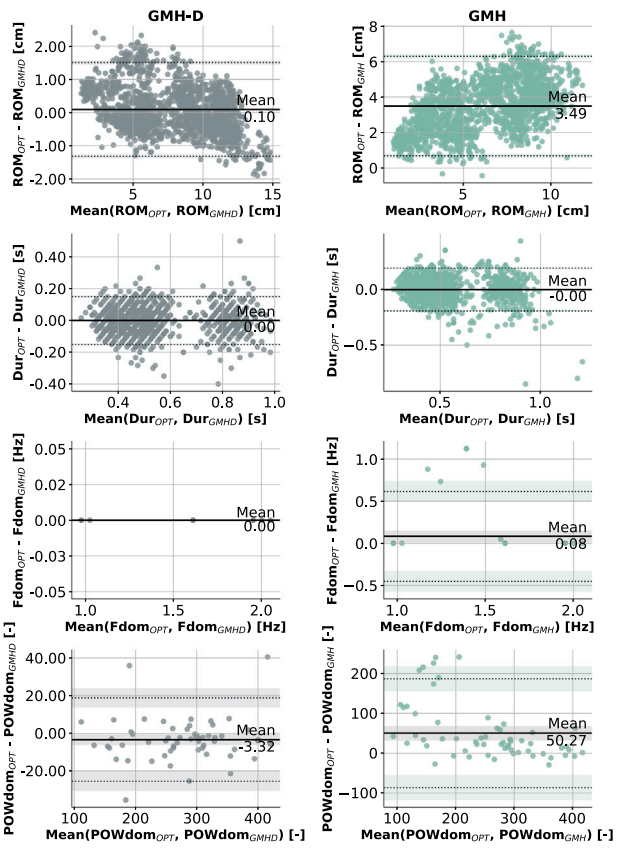


Fig. 12. Bland–Altman plots for *ROM* (top), *DUR* (top-middle), F_{DOM} (middle-bottom), POW_{DOM} (bottom) estimated from single repetitions of the SFT task. Colour coding for GMH and GMH-D is the same as in Fig. 7.

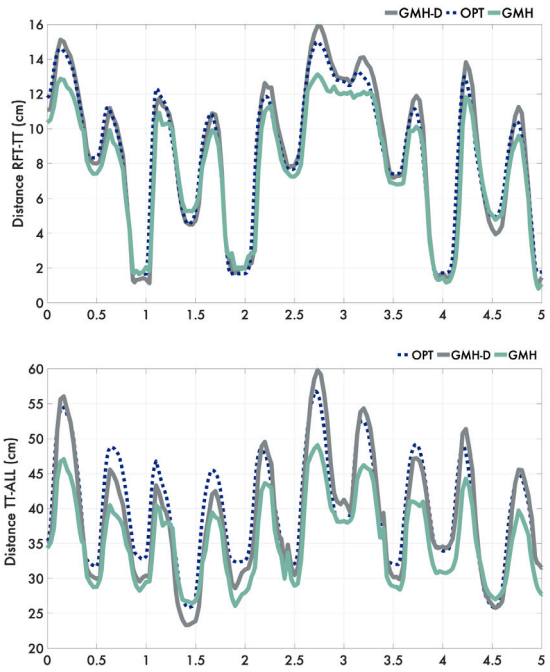


Fig. 13. RFT-TT distance (top) and TT-ALL (bottom) distance during a MFT task, as measured by GMH and GMH-D with respect to the gold standard OPT (dotted line). The three curves have been vertically and horizontally realigned to allow a direct comparison.

increase of the error at the FAR distance (larger PRMSE and RMSE and lower correlation with large outlier), whereas GMH-D shows a much smaller and stable error and an almost perfect correlation -i.e, median value close to 99% and very small interquartile range.

Moving to the segment-level analysis of all trials, the Bland–Altman plots for *ROM*, *DUR*, F_{DOM}, and POW_{DOM} are shown in Fig. 12. In 93.63% of *ROM* measurements, the error between GMH-D and OPT lies in a smaller LoA ([−1.32, 1.51] cm), whereas 95.85% of GMH errors fall in a much larger range ([0.69, 6.30] cm, mean value: 3.49 cm). For *DUR*, the two frameworks appear more comparable, with 95.07% and 95.25% of measurement errors respectively for GMH-D and GMH inside similar LoA. Therefore, performance are closer, but still GMH-D provides an higher accuracy. In the estimation of F_{DOM}, no error is performed by GMH-D, whereas for GMH 91.53% of the errors are between [−0.45, 0.62] Hz (mean: 0.08 Hz). Finally, for POW_{DOM}, 94.2% of the errors for GMH-D fall in a narrower range ([−25.49, 18.85]) than GMH where 89.83% is in a wider range, proving again the higher accuracy of GMH-D.

The results for ICC and CCC, with their confidence intervals, are reported in Table 3, using a 95% confidence level. For both metrics, *p-values* are all below $p < 0.001$, so they are not reported. For GMH-D, the ICC and CCC values suggest an excellent level of agreement (>0.90) for all the four investigated parameters. In contrast, for GMH, it is confirmed by both metrics how the method wrongly estimates *ROM*, producing measurements largely affected by errors. Nevertheless, a good level of agreement is achieved for the other temporal and spectral parameters.

5.3. MFT task validation

Fig. 13 reports an example of the RFT-TT distance (top) and the TT-ALL summarising distance (bottom), performed by GMH and GMH-D with respect to the OPT measurement (dotted line). As it can be observed, a reconstruction adherent to OPT trajectory is challenging for both frameworks. Furthermore, it is worth noticing how the TT-ALL distance reflects, as hypothesised, the cumulative motion of all fingers, thus providing a way to observe all the single tapping movements from one single trajectory.

Results on RMSE, PRMSE, and Person's ρ are organised according to the distance factor (Fig. 14). The values are reported for the IFT-TT (INDEX), MT-TT (MIDDLE), RFT-TT (RING), and PT-TT (PINKIE) distances that compose the cumulative TT-ALL trajectory.

When considering the NEAR distance, GMH-D performs overall better than GMH, with a median PRMSE value below 15% and a median RMSE ≤ 1 cm for all fingers. Both GMH and GMH-D achieve a correlation value with OPT > 0.9. In the FAR condition, both GMH and GMH-D show an increase in the median error and its interquartile range, and a small decrease in the correlation, suggesting an influence of the distance factor in the quality of the temporal reconstruction of the MFT trajectories.

Moving to the segment-level analysis on the TT-ALL distance, the Bland–Altman plots for *ROM* and *DUR*, F_{DOM} and POW_{DOM} are shown in Fig. 15. As a first remark, it must be noted the mean error and the LoA are inevitably larger than for the other tasks due to the propagation in TT-ALL of the error in the tracking of each finger involved in this fictitious trajectory. Therefore, an error range around four times larger than for SFT was expected and observed. Proceeding with the analysis, in 94.62% of the evaluations of *ROM*, the error between GMH-D and *OPT* lies in the range [−4.50, 6.50] cm (mean: 0.89 cm), whereas 96.48% of errors for GMH are in the range [−3.40, 10.76] cm (mean: 3.68 cm). For *DUR*, 95.45% of measurements have an error in the range [−0.12, 0.12] s for GMH-D and 95.65% in the range [−0.13, 0.13] s for GMH, in line with what observed for SFT. Regarding

Table 3
ICC and CCC values for segment-level and frequency parameters in SFT task, both for GMH and GMH-D methods with respect to the gold-standard OPT, considering a 95% confidence interval.

	GMH						GMH-D					
	ICC			*CCC*			*ICC*			*CCC*		
	Low Conf.	*Value*	*High Conf.*	*Low Conf.*	*Value*	*High Conf.*	*Low Conf.*	*Value*	*High Conf.*	*Low Conf.*	*Value*	*High Conf.*
ROM	−0.07	**0.49**	0.81	0.47	**0.49**	0.51	0.97	**0.98**	0.98	0.97	**0.98**	0.98
DUR	0.84	**0.85**	0.86	0.84	**0.85**	0.86	0.89	**0.90**	0.91	0.89	**0.90**	0.91
F_{DOM}	0.68	**0.80**	0.87	0.66	**0.86**	0.87	1.00	**1.00**	1.00	1.00	**1.00**	1.00
POW_{DOM}	0.59	**0.74**	0.84	0.51	**0.65**	0.75	0.98	**0.99**	0.99	0.98	**0.99**	0.99

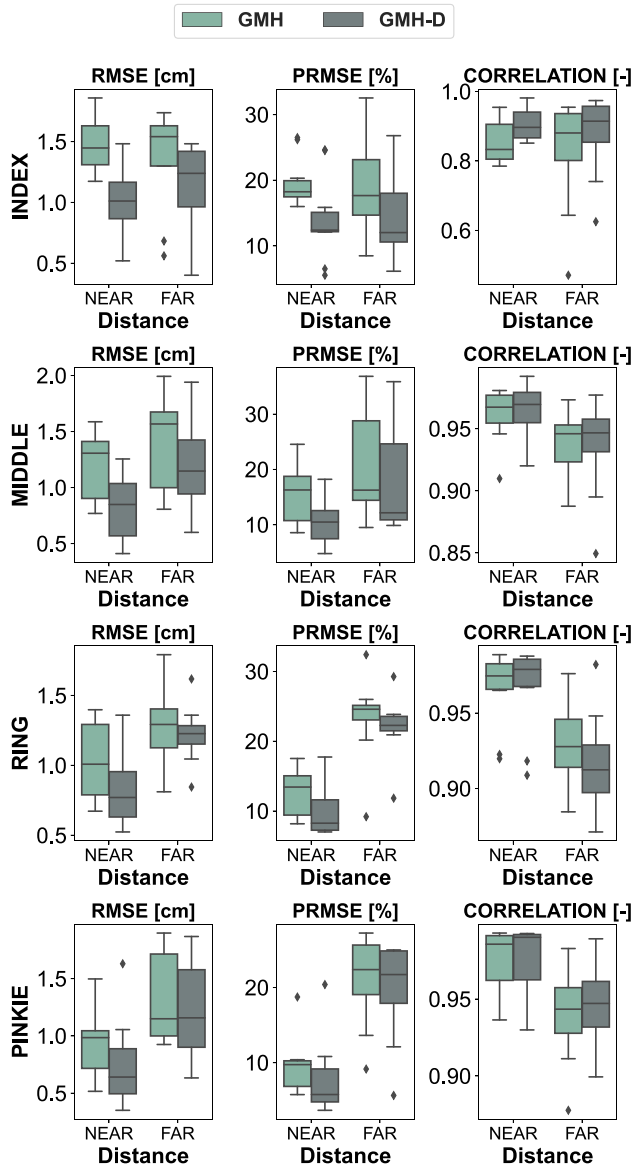


Fig. 14. RMSE (top), PRMSE (middle), and Pearson's ρ (bottom) box plots, in MFT trials at different distances from camera: NEAR distance (60–80 cm) and FAR distance (80–100 cm).

the estimation of F_{DOM}, no error is performed by GMH-D with respect to OPT, whereas 88.89% of GMH estimations are producing an error between [−0.03, 0.02] Hz (mean: 0.00 Hz), almost negligible. Finally, for POW_{DOM} 93.75% of estimations have an error in the range [−2.22, 121.62] (mean: 58.75) for GMH and in the range [−26.32, 76.37] (mean: 25.02) for GMH-D, supporting again the overall higher accuracy of GMH-D also in terms of spectral parameters.

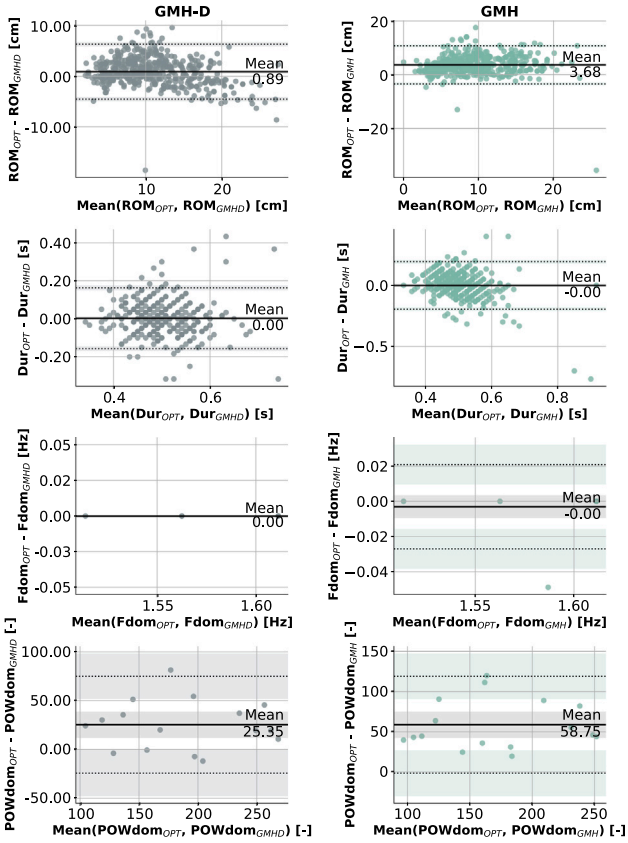


Fig. 15. Bland–Altman plots for *ROM* (top), *DUR* (top-middle), F_{DOM} (middle-bottom), POW_{DOM} (bottom) estimated from single segments of the TT-ALL distance in MFT task. Colour coding for GMH and GMH-D is the same as in Fig. 7.

The results for ICC and CCC, with their confidence intervals, are reported in Table 4, using a 95% confidence level. For both metrics, *p-values* are all below $p < 0.001$, so they are not reported. It must be noted that, by enforcing just one possible execution speed, the internal variability in terms of segments duration is reduced. In addition to the limited number of segments for the task, this produces low values of ICC and CCC for *DUR* parameters, either for GMH and GMH-D. However, from Bland–Altman plots, we can observe that actually almost all points are falling in a narrow error range, comparable with that achieved in the previous tasks. Therefore, the validity of the two metrics for *DUR* parameter is really limited since the inter-variability (between GMH/GMH-D and OPT) and the intra-variability of the single segments in terms of duration in the dataset are unbalanced- i.e., even mistakes with small magnitude produce a variability larger than the internal variability of *DUR*, biasing the results. Moreover, it must be considered that the virtual TT-ALL distance, due to its definition as sum of other trajectories, propagates their error, which can alter significantly its morphology (e.g., Fig. 13). This alterations may affect the trivial segmentation procedure established and consequently the estimation of

Table 4
ICC and CCC values for segment-level and frequency parameters in MFT task, both for GMH and GMH-D methods with respect to the gold-standard OPT.

	GMH						GMH-D					
	ICC			*CCC*			*ICC*			*CCC*		
	Low Conf.	*Value*	*High Conf.*	*Low Conf.*	*Value*	*High Conf.*	*Low Conf.*	*Value*	*High Conf.*	*Low Conf.*	*Value*	*High Conf.*
ROM	0.67	**0.71**	0.76	0.52	**0.70**	0.76	0.84	**0.86**	0.89	0.83	**0.85**	0.86
DUR	0.36	**0.44**	0.51	0.38	**0.44**	0.46	0.39	**0.46**	0.53	0.42	**0.46**	0.48
F_{DOM}	0.88	**0.90**	0.93	0.84	**0.89**	0.93	1.00	**1.00**	1.00	1.00	**1.00**	1.00
POW_{DOM}	0.57	**0.83**	0.85	0.73	**0.75**	0.78	0.69	**0.88**	0.96	0.81	**0.82**	0.84

Table 5
Comparison of computational performance of GMH and GMH-D in terms of mean frame rate (FPS) during data processing, compared to similar solutions previously employed for hand tracking in clinical applications. [–] stands for information not provided.

Framework	Hardware(CPU/GPU/RAM)	Input size	FPS	Source
GMH	Intel i7-9750H/-/16 GB RAM	1280 × 720 px	30 fps	This work
GMH-D	Intel i7-9750H/-/16 GB RAM	1280 × 720 px	30 fps	This work
Openpose	-/Nvidia GTX 1080 Ti/11 GB RAM	1280 × 720 px	22 fps	[113]
A2J/ST-A2J	-/Nvidia GTX 1080 Ti/-	176 × 176 px	105 fps	[110,112]
HandGraphCNN	-/Nvidia GTX 1080/-	256 × 256 px	>50 fps	[111]

the *DUR* parameter for the single segments. The values of ICC suggest a good level of agreement (>0.80) for all the remaining investigated parameters, with almost a perfect agreement for spectral properties either using GMH or GMH-D. Again, GMH-D outperforms GMH in terms of accuracy of the estimation of *ROM* (for ICC: 0.86 vs. 0.71; for CCC: 0.85 vs. 0.70).

5.4. Computational performance

Table 5 reports the computational performance of GMH and GMH-D during the processing of the collected videos in terms of mean frame rate, expressed as frame per seconds (fps). For a more comprehensive comparison, performance of other hand tracking frameworks, previously employed in similar clinical applications, were included, namely OpenPose [109], A2J/ST-A2J [110] and HandGraphCNN [111]. All these methods, for instance, were used to assess SFT in PD [79,96,112], even though without reporting computational performance. Therefore, these information was retrieved either from the original studies presenting them or from reports of bench-marking performance on *in-the-wild* datasets for hand tracking [113]. Therefore, values were achieved on different input data with respect to GMH and GMH-D, and this must be taken into account in comparing the reciprocal differences.

Overall, GMH and GMH-D are the only methods that offer real-time (above >30 fps) without GPU acceleration, whereas all the other methods require at least a Nvidia GTX 1080 graphic card. Moreover, both models can process high quality input frames (1280 × 720px) without reducing their speed.

5.5. GMH vs. GMH-D: remarks

In general, the results suggest that both the RGB framework (GMH) and the RGB-D enhanced framework (GMH-D) hold potentiality in their application to track fine hand movements, as the one required by clinical assessment tasks. Overall, GMH-D seems to provide a more accurate and reliable motion reconstruction in the three tasks, with a mean PRMSE always smaller than 15% and an almost perfect correlation ($\rho > 0.97$) between the tracked inter-finger distances in all the tasks. Moreover, GMH-D appears robust to different execution speed and also to the distance from the recording camera, at least in the range [60 cm–100 cm]. These results clearly depend also on the quality of the depth sensor of the RGB-D camera. In this study, they were achieved using a MAK, which has an high accuracy for depth tracking up to 3.5 m [114]. When adapting GMH-D to other RGB-D devices, the quality of their depth sensors may clearly affect the result and should be evaluated beforehand. Still, this method can struggle with self-occlusions of fingers, as observed in SFT, where the lateral viewing angle improved the quality of tracking over the frontal one. Therefore, when applying this kind of approach, optimal positioning of the hand should be taken into consideration. Overall, GMH-D allows also to evaluate, with a good level of agreement, all the four investigated parameters, namely *ROM*, *DUR*, F_{DOM}, and POW_{DOM}. Clearly, even if powerful, this framework requires an additional depth sensor, which could still represent a limitation for certain applications.

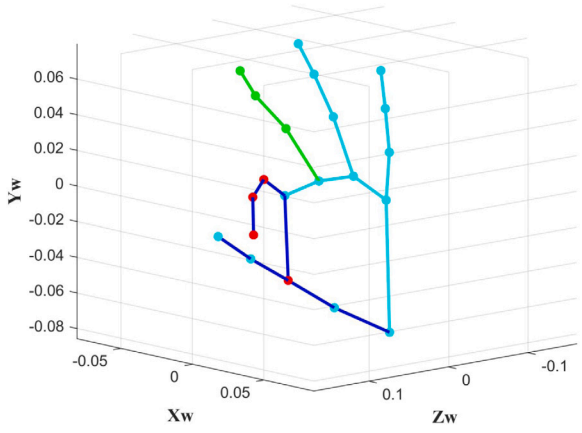


Fig. 16. GMH fails in reconstructing the contact phase between the two finger tips in SFT, in the lateral perspective. This depth error is solved by GMH-D using depth from ToF sensor of MAK.

On the other hand, GMH requires only an RGB camera and appeared promising in the tracking of the OC task, with accuracy comparable to GMH-D. However, the analysis of SFT and MFT tasks highlighted how GMH fails in precisely reconstructing 3D motion when finer movements are considered, producing very large RMSE and PRMSE values. From the analysis of trials with high error, it is evident how GMH correctly marks the hand in 2D, but due to self-occlusions of joints, fails in properly reconstructing the 3D shape of the hand. For instance, in the lateral SFT task when fingers are touching, GMH often does not identify the two finger tips as aligned over the *z*-axis, but misplaces one in front of the other, producing a wrong value in the minima of the IFT-TT distance. An example of this failure is shown in Fig. 16.

Table 6
Summary of recommendations (distance range, viewing angle, and speed of movements) for the optimal capture of the three validated tasks (OC, SFT, MFT) with GMH and GMH-D. The last row reports the best tracking framework for each task according to the main results of the validation study.

	OC	SFT	MFT
Distance range	60–80 cm	60–100 cm	60–80 cm
Viewing angle	Frontal	Lateral	Frontal
Speed of motion	Slow/Normal/Fast	Slow/Normal/Fast	Slow/Normal/Fast
Best framework	**GMH/GMH-D**	**GMH-D**	**GMH-D**

Overall, trajectories reconstructed for SFT and MFT tasks appear to be influenced by a squeezing effect stemming from the inaccurate depth estimation by the DL model of GMH. This phenomenon is less pronounced in OC tasks, where the consistent movement of all the fingers together may simplify the complexity of the tracking. Additionally, the model might lack sufficient training in reconstructing specific and finer hand movements such as SFT and MFT compared to the more common opening–closing gesture. This failure in estimating depth could also account for GMH's dependence on camera distance, leading to degraded performance in the FAR positioning, and also for the worsening of accuracy due to increased motion velocity, which alters hand appearance due to motion blur. These issues are mitigated or eliminated by GMH-D, thanks to the use of the ToF depth sensor. It is worth noting that GMH, as explicitly stated by developers in [97], was not primarily designed for this type of applications and was mainly thought for working with close hand recordings, such as egocentric videos from a smartphone.

Nevertheless, even if GMH seems to be trustworthy in spatial analysis of motion only for OC task, it could still be applied to estimate temporal and spectral properties also in the other two tasks. Indeed, the framework shows good to excellent agreement in terms of ICC and CCC for *DUR*, F_{DOM}, and POW_{DOM}.

It must also be considered that some factors can have slightly enlarged the error for both frameworks: the residual offset between physical and virtual markers; the alteration of the appearance of the hand due to the physical markers; for segment-level analysis, the trivial segmentation algorithm, which might have introduced an additional error in the evaluation of ICC and CCC between parameters.

Unfortunately, due to the lack of comparable validation procedures for 3D RGB or RGB-D hand tracking methods based on DL in the literature, it is not possible to carry out a proper comparison with other studies. However, the good results obtained in estimating frequency parameters for GMH are in line with the results from [87], which compared frequency of tremor from PD measured with GMH and an accelerometer (mean absolute error 0.229±0.174 Hz). Moreover, the small error obtained for estimating *ROM* is coherent with [115], which found a very good agreement between 2D tracking of GMH and measures taken from a touchscreen device (mean RMSE 0.28±0.064 normalised pixel).

Overall, the results obtained in this validation study are promising and demonstrate the potential, the strengths, and the weaknesses of the two DL-based frameworks, especially for a perspective use in clinical applications requiring high accuracy and reliability. Therefore, Table 6 summarises the previous observations and provides concise guidelines for researchers interested in using the two frameworks for the three tasks or other similar hand and finger movements, taking into consideration the investigated influencing factors.

6. Conclusions

This paper presented the validation against a gold standard system for motion capture of two DL-based hand tracking frameworks, namely Google MediaPipe Hand (GMH) and its enhanced version GMH-D. This validation was focused especially on proving accuracy and reliability of these frameworks for their perspective usage in clinical applications, such as automatic assessment of three tasks commonly administered to patients with Parkinson's disease, namely hand opening and closing, single finger tapping, and multiple finger tapping. This work aimed also at remarking the importance of carrying out a rigorous validation of DL-based tracking frameworks as measurement systems before their application in clinical scenarios. This is especially relevant considering that most *off-the-shelf* DL solutions for hand tracking are not specifically designed for deployment in clinical applications and could not adapt well to this scenario.

Three possible influencing factors were investigated, namely distance from recording camera, recording camera viewing angle, and velocity of tracked motion. Results suggest that for a more accurate reconstruction of 3D motion, GMH-D provides good to excellent level of agreement to the gold standard, by exploiting depth information coming from an RGB-D camera (in this study, Microsoft Azure Kinect). GMH, by leveraging only an RGB input, proved to be less accurate in spatial domain. Still, it may be employed for evaluating temporal and spectral properties of motion with a good level of trust.

As a limitation, this validation study did not take into account different light conditions that could alter tracking for both frameworks. However, since also the gold standard of motion capture has strict requirements on light conditions to function properly, how to carry out this evaluation is still an open challenge to investigate as a future direction of development. Moreover, motor tasks involving hand rotation, such as the hand pronation-supination, were not yet investigated, but could represent an additional challenge for both methods. Also this application is left open for future investigations.

To conclude, thanks to the rapid growth of DL-based solutions for accurate hand tracking from RGB and RGB-D videos, a lot of new possibilities will arise in the next future, especially for clinical applications. In this scenario, providing rigorous validation will become crucial to prove the efficacy and the reliability of such frameworks, which is the direction to which this work is providing its main contribution.

Code availability statement

The code to run both GMH and GMH-D on videos acquired using Microsoft Azure Kinect is available in GitHub:
https://github.com/gianluca-amprimo/GMH-D.git.

CRediT authorship contribution statement

Gianluca Amprimo: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Giulia Masi:** Formal analysis, Investigation, Software, Writing – review & editing. **Giuseppe Pettiti:** Funding acquisition, Methodology, Resources, Supervision, Writing – review & editing. **Gabriella Olmo:** Conceptualization, Supervision, Writing – review & editing. **Lorenzo Priano:** Supervision, Writing – review & editing. **Claudia Ferraris:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Upal Mahbub, Tauhidur Rahman, Md Atiqur Rahman Ahad, Contactless human monitoring: Challenges and future direction, in: Contactless Human Activity Analysis, Springer International Publishing, Cham, 2021, pp. 335–364.

[2] Santosh Kumar Yadav, Kamlesh Tiwari, Hari Mohan Pandey, Shaik Ali Akbar, A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions, Knowl.-Based Syst. 223 (106970) (2021) 106970.

[3] Xinyi Wang, Saurabh Garg, Son N Tran, Quan Bai, Jane Alty, Hand tremor detection in videos with cluttered background using neural network based approaches, Health Inf. Sci. Syst. 9 (1) (2021) 30.

[4] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, Ioannis A Kakadiaris, 3D human pose estimation: A review of the literature and analysis of covariates, Comput. Vis. Image Underst. 152 (2016) 1–20.

[5] Preksha Pareek, Ankit Thakkar, A survey on video-based human action recognition: recent updates, datasets, challenges, and applications, Artif. Intell. Rev. 54 (3) (2021) 2259–2322.

[6] Sharath Chandra Akkaladevi, Christoph Heindl, Action recognition for human robot interaction in industrial applications, in: 2015 IEEE International Conference on Computer Graphics, Vision and Information Security, CGVIS, IEEE, 2015.

[7] Gisela Miranda Difini, Marcio Garcia Martins, Jorge Luis Victória Barbosa, Human pose estimation for training assistance: A systematic literature review, in: Proceedings of the Brazilian Symposium on Multimedia and the Web, ACM, New York, NY, USA, 2021.

[8] Shradha Dubey, Manish Dixit, A comprehensive survey on human pose estimation approaches, Multimed. Syst. 29 (1) (2023) 167–195.

[9] Mahmoud Al-Faris, John Chiverton, David Ndzi, Ahmed Isam Ahmed, A review on computer vision-based methods for human action recognition, J. Imaging 6 (6) (2020) 46.

[10] Mais Yasen, Shaidah Jusoh, A systematic review on hand gesture recognition techniques, challenges and applications, PeerJ Comput. Sci. 5 (2019) e218.

[11] Noraini Mohamed, Mumtaz Begum Mustafa, Nazean Jomhari, A review of the hand gesture recognition system: Current progress and future directions, IEEE Access 9 (2021) 157422–157436.

[12] Serena Cerfoglio, Paolo Capodaglio, Paolo Rossi, Federica Verme, Gabriele Boldini, Viktoria Cvetkova, Graziano Ruggeri, Manuela Galli, Veronica Cimolin, Tele-rehabilitation interventions for motor symptoms in COVID-19 patients: A narrative review, Bioengineering (Basel) 10 (6) (2023).

[13] Pramod Kumar Pisharady, Martin Saerbeck, Recent methods and databases in vision-based hand gesture recognition: A review, Comput. Vis. Image Underst. 141 (2015) 152–165.

[14] Ammar Ahmad, Cyrille Migniot, Albert Dipanda, Hand pose estimation and tracking in real and virtual interaction:A review, Image Vis. Comput. 89 (2019) 35–49.

[15] Gavin Buckingham, Hand tracking for immersive virtual reality: Opportunities and challenges, Front. Virtual Real. 2 (2021).

[16] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, Christian Theobalt, Real-time hand tracking under occlusion from an egocentric RGB-D sensor, in: 2017 IEEE International Conference on Computer Vision, ICCV, IEEE, 2017.

[17] Hiske van Duinen, Simon C. Gandevia, Constraints for control of the human hand: Control of the hand, J. Physiol. 589 (Pt 23) (2011) 5583–5593.

[18] Lisa Reissner, Gabriella Fischer, Renate List, Pietro Giovanoli, Maurizio Calcagni, Assessment of hand function during activities of daily living using motion tracking cameras: A systematic review, Proc. Inst. Mech. Eng. H 233 (8) (2019) 764–783.

[19] Ali Erol, George Bebis, Mircea Nicolescu, Richard D Boyle, Xander Twombly, Vision-based hand pose estimation: A review, Comput. Vis. Image Underst. 108 (1–2) (2007) 52–73.

[20] Tommaso Lisini Baldi, Mostafa Mohammadi, Stefano Scheggi, Domenico Prattichizzo, Using inertial and magnetic sensors for hand tracking and rendering in wearable haptics, in: 2015 IEEE World Haptics Conference, WHC, IEEE, 2015.

[21] Claudio Pacchierotti, Gionata Salvietti, Irfan Hussain, Leonardo Meli, Domenico Prattichizzo, The hring: A wearable haptic device to avoid occlusions in hand tracking, in: 2016 IEEE Haptics Symposium, HAPTICS, IEEE, 2016.

[22] Antonio H J Moreira, Sandro Queiros, José Fonseca, Pedro L Rodrigues, Nuno F Rodrigues, Joao L Vilaca, Real-time hand tracking for rehabilitation and character animation, in: 2014 IEEE 3nd International Conference on Serious Games and Applications for Health, SeGAH, IEEE, 2014.

[23] Julien Stamatakis, Jérome Ambroise, Julien Crémers, Hoda Sharei, Valérie Delvaux, Benoit Macq, Gaëtan Garraux, Finger tapping clinimetric score prediction in Parkinson's disease using low-cost accelerometers, Comput. Intell. Neurosci. 2013 (2013) 717853.

[24] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, Daniel Freedman, Pushmeet Kohli, Eyal Krupka, Andrew Fitzgibbon, Shahram Izadi, Accurate, robust, and flexible real-time hand tracking, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, 2015.

[25] Iason Oikonomidis, Nikolaos Kyriazis, Antonis Argyros, Efficient model-based 3D tracking of hand articulations using Kinect, in: Procedings of the British Machine Vision Conference 2011, British Machine Vision Association, 2011.

[26] Robert Y. Wang, Jovan Popović, Real-time hand-tracking with a color glove, in: ACM SIGGRAPH 2009 Papers, ACM, New York, NY, USA, 2009.

[27] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, Jian Sun, Realtime and robust hand tracking from depth, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2014.

[28] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, Christian Theobalt, Fast and robust hand tracking using detection-guided optimization, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2015.

[29] Zhigeng Pan, Yang Li, Mingmin Zhang, Chao Sun, Kangde Guo, Xing Tang, Steven Zhiying Zhou, A real-time multi-cue hand tracking algorithm based on computer vision, in: 2010 IEEE Virtual Reality Conference, VR, IEEE, 2010.

[30] Tjokorda Agung Budi Wirayuda, Habbi Ananto Adhi, Didik Hari Kuswanto, Retno Novi Dayawati, Real-time hand-tracking on video image based on palm geometry, in: 2013 International Conference of Information and Communication Technology, ICoICT, IEEE, 2013, pp. 241–246.

[31] Hui-Shyong Yeo, Byung-Gook Lee, Hyotaek Lim, Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware, Multimedia Tools Appl. 74 (8) (2015) 2687–2715.

[32] Chia-Ping Chen, Yu-Ting Chen, Ping-Han Lee, Yu-Pao Tsai, Shawmin Lei, Real-time hand tracking on depth images, in: 2011 Visual Communications and Image Processing, VCIP, IEEE, 2011.

[33] Wenhuan Cui, Wenmin Wang, Hong Liu, Robust hand tracking with refined CAMShift based on combination of depth and image features, in: 2012 IEEE International Conference on Robotics and Biomimetics, ROBIO, IEEE, 2012.

[34] Abdul H Butt, E Rovini, C Dolciotti, G De Petris, P Bongioanni, MC Carboncini, F Cavallo, Objective and automatic classification of Parkinson disease with leap motion controller, Biomed. Eng. Online 17 (1) (2018) 1–21.

[35] Maryam Khademi, Hossein Mousavi Hondori, Alison McKenzie, Lucy Dodakian, Cristina Videira Lopes, Steven C Cramer, Free-hand interaction with leap motion controller for stroke rehabilitation, in: CHI '14 Extended Abstracts on Human Factors in Computing Systems, ACM, New York, NY, USA, 2014.

[36] Francesco Luke Siena, Bill Byrom, Paul Watts, Philip Breedon, Utilising the intel RealSense camera for measuring health outcomes in clinical research, J. Med. Syst. 42 (3) (2018) 53.

[37] Paulina J M Bank, Johan Marinus, Carel G M Meskers, Jurriaan H de Groot, Jacobus J van Hilten, Optical hand tracking: A novel technique for the assessment of bradykinesia in Parkinson's disease, Mov. Disord. Clin. Pract. 4 (6) (2017) 875–883.

[38] C Ferraris, D Pianu, A Chimienti, G Pettiti, V Cimolin, N Cau, R Nerino, Evaluation of finger tapping test accuracy using the LeapMotion and the intel RealSense sensors, in: Proceedings of the 37th International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2015), Milan, Italy, 2015, pp. 25–29.

[39] Jože Guna, Grega Jakus, Matevž Pogačnik, Sašo Tomažič, Jaka Sodnik, An analysis of the precision and reliability of the leap motion sensor and its suitability for static and dynamic tracking, Sensors (Basel) 14 (2) (2014) 3702–3720.

[40] Marc H. Schieber, Motor cortex – hand movements and plasticity, in: Marc D. Binder, Nobutaka Hirokawa, Uwe Windhorst (Eds.), Encyclopedia of Neuroscience, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN: 978-3-540-29678-2, 2009, pp. 2431–2433, http://dx.doi.org/10.1007/978-3-540-29678-2_3586.

[41] Francesco Menegoni, E Milano, C Trotti, Manuela Galli, M Bigoni, S Baudo, Alessandro Mauro, Quantitative evaluation of functional limitation of upper limb movements in subjects affected by ataxia, Eur. J. Neurol. 16 (2) (2009) 232–239.

[42] Yan Pang, Jake Christenson, Feng Jiang, Tim Lei, Remy Rhoades, Drew Kern, John A Thompson, Chao Liu, Automatic detection and quantification of hand movements toward development of an objective assessment of tremor and bradykinesia in Parkinson's disease, J. Neurosci. Methods 333 (108576) (2020) 108576.

[43] Rajesh Benny, Kishore Shetty, The split hand sign, Ann. Indian Acad. Neurol. 15 (3) (2012) 175–176.

[44] S.M. Hunter, P. Crome, Hand function and stroke, Rev. Clin. Gerontol. 12 (1) (2002) 68–81.

[45] Bart Post, Maruschka P Merkus, Rob MA de Bie, Rob J de Haan, Johannes D Speelman, Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? Mov. Disord.: Off. J. Mov. Disord. Soc. 20 (12) (2005) 1577–1584, http://dx.doi.org/10.1002/mds.20640.

[46] Matthew Demoe, Alvaro Uribe-Quevedo, André L Salgado, Hidenori Mimura, Kamen Kanev, Patrick CK Hung, Exploring data glove and robotics hand exergaming: lessons learned, in: 2020 IEEE 8th International Conference on Serious Games and Applications for Health, SeGAH, IEEE, 2020, pp. 1–8.

[47] Luanne Cardoso Mendes, Angela Abreu Rosa de Sá, Isabela Alves Marques, Yann Morère, Adriano de Oliveira Andrade, RehaBEElitation: the architecture and organization of a serious game to evaluate motor signs in Parkinson's disease, PeerJ Comput. Sci. 9 (2023) e1267.

[48] Giada Devittori, Daria Dinacci, Davide Romiti, Antonella Califfi, Claudio Petrillo, Paolo Rossi, Raffaele Ranzani, Roger Gassert, Olivier Lambercy, Unsupervised robot-assisted rehabilitation after stroke: feasibility, effect on therapy dose, and user experience, J. NeuroEng. Rehabil. 21 (1) (2024) 1–11.

[49] Carlos Antonio Godoy Junior, Francesco Miele, Laura Mäkitie, Eleonora Fiorenzato, Maija Koivu, Lytske Jantien Bakker, Carin Uyl-de Groot, William Ken Redekop, Welmoed Kirsten van Deen, Attitudes toward the adoption of remote patient monitoring and artificial intelligence in Parkinson's disease management: Perspectives of patients and neurologists, in: The Patient-Patient-Centered Outcomes Research, Springer, 2024, pp. 1–11, http://dx.doi.org/10.1007/s40271-023-00669-0.

[50] Ameur Latreche, Ridha Kelaiaia, Ahmed Chemori, Adlen Kerboua, Reliability and validity analysis of MediaPipe-based measurement system for some human rehabilitation motions, Measurement (London) 214 (112826) (2023) 112826.

[51] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, Matthias Grundmann, MediaPipe hands: On-device real-time hand tracking, 2020, arXiv:2006.10214.

[52] Gianluca Amprimo, Claudia Ferraris, Giulia Masi, Giuseppe Pettiti, Lorenzo Priano, GMH-D: Combining google MediaPipe and RGB-depth cameras for hand motor skills remote assessment, in: 2022 IEEE International Conference on Digital Health, ICDH, IEEE, 2022.

[53] Theocharis Chatzis, Andreas Stergioulas, Dimitrios Konstantinidis, Kosmas Dimitropoulos, Petros Daras, A comprehensive study on deep learning-based 3D hand pose estimation methods, Appl. Sci. (Basel) 10 (19) (2020) 6850.

[54] Jonathan Tompson, Murphy Stein, Yann Lecun, Ken Perlin, Real-time continuous pose recovery of human hands using convolutional networks, ACM Trans. Graph. 33 (5) (2014) 1–10.

[55] Naima Otberdout, Lahoucine Ballihi, Driss Aboutajdine, Hand pose estimation based on deep learning depth map for hand gesture recognition, in: 2017 Intelligent Systems and Computer Vision, ISCV, IEEE, 2017.

[56] Meysam Madadi, Sergio Escalera, Xavier Baró, Jordi Gonzàlez, End-to-end global to local convolutional neural network learning for hand pose recovery in depth data, IET Comput. Vis. 16 (1) (2022) 50–66.

[57] Xingyi Zhou, Qingfu Wan, Wei Zhang, Xiangyang Xue, Yichen Wei, Model-based deep hand pose estimation, 2016, arXiv:1606.06854.

[58] Rui Li, Zhenyu Liu, Jianrong Tan, A survey on 3D hand pose estimation: Cameras, methods, and datasets, Pattern Recognit. 93 (2019) 251–272.

[59] Evangelos Kazakos, Christophoros Nikou, Ioannis A Kakadiaris, On the fusion of RGB and depth information for hand pose estimation, in: 2018 25th IEEE International Conference on Image Processing, ICIP, IEEE, 2018.

[60] Jordi Sanchez-Riera, Kathiravan Srinivasan, Kai-Lung Hua, Wen-Huang Cheng, M Anwar Hossain, Mohammed F Alhamid, Robust RGB-D hand tracking using deep learning priors, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2018) 2289–2301.

[61] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh, OpenPose: Realtime multi-person 2D pose estimation using part affinity fields, IEEE Trans. Pattern Anal. Mach. Intell. 43 (1) (2021) 172–186.

[62] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, Cewu Lu, Rmpe: Regional multi-person pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2334–2343.

[63] Nicholas Santavas, Ioannis Kansizoglou, Loukas Bampis, Evangelos Karakasis, Antonios Gasteratos, Attention! A lightweight 2D hand pose estimation approach, IEEE Sens. J. 21 (10) (2021) 11488–11496.

[64] Filippos Gouidis, Paschalis Panteleris, Iason Oikonomidis, Antonis Argyros, Accurate hand keypoint localization on mobile devices, in: 2019 16th International Conference on Machine Vision Applications, MVA, IEEE, 2019.

[65] Guan Ming Lim, Prayook Jatesiktat, Christopher Wee Keong Kuah, Wei Tech Ang, Camera-based hand tracking using a mirror-based multi-view setup, Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 2020 (2020) 5789–5793.

[66] Christian Zimmermann, Thomas Brox, Learning to estimate 3D hand pose from single RGB images, in: 2017 IEEE International Conference on Computer Vision, ICCV, IEEE, 2017.

[67] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, Christian Theobalt, GANerated hands for real-time 3D hand tracking from monocular RGB, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018.

[68] Yiming He, Wei Hu, 3D hand pose estimation via regularized graph representation learning, in: Artificial Intelligence, Springer International Publishing, Cham, 2021, pp. 540–552.

[69] Shaoxiang Guo, Eric Rigall, Yakun Ju, Junyu Dong, 3D hand pose estimation from monocular RGB with feature interaction module, IEEE Trans. Circuits Syst. Video Technol. 32 (8) (2022) 5293–5306.

[70] Adrian Spurr, Jie Song, Seonwook Park, Otmar Hilliges, Cross-modal deep variational hand pose estimation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018.

[71] Sanjeev Sharma, Shaoli Huang, Dacheng Tao, An end-to-end framework for unconstrained monocular 3D hand pose estimation, 2019, arXiv:1911.12501.

[72] Yujun Cai, Liuhao Ge, Jianfei Cai, Nadia Magnenat Thalmann, Junsong Yuan, 3D hand pose estimation using synthetic data and weakly labeled RGB images, IEEE Trans. Pattern Anal. Mach. Intell. 43 (11) (2021) 3739–3753.

[73] Ning Yang, De-Feng Liu, Tao Liu, Tianyuan Han, Pingyue Zhang, Xuenan Xu, Siyu Lou, Huan-Guang Liu, An-Chao Yang, Cheng Dong, Mang I Vai, Sio Hang Pun, Jian-Guo Zhang, Automatic detection pipeline for accessing the motor severity of Parkinson's disease in finger tapping and postural stability, IEEE Access 10 (2022) 66961–66973.

[74] Stefan Williams, Samuel D Relton, Hui Fang, Jane Alty, Rami Qahwaji, Christopher D Graham, David C Wong, Supervised classification of bradykinesia in Parkinson's disease from smartphone videos, Artif. Intell. Med. 110 (101966) (2020) 101966.

[75] David C Wong, Samuel D Relton, Hui Fang, Rami Qhawaji, Christopher D Graham, Jane Alty, 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), IEEE, 2019.

[76] Dmitry Viatkin, Begonya Garcia-Zapirain, Amaia Méndez Zorrilla, Deep learning techniques applied to predict and measure finger movement in patients with multiple sclerosis, Appl. Sci. (Basel) 11 (7) (2021) 3137.

[77] David Ahmedt-Aristizabal, Clinton Fookes, Simon Denman, Kien Nguyen, Tharindu Fernando, Sridha Sridharan, Sasha Dionisio, A hierarchical multi-modal system for motion analysis in patients with epilepsy, Epilepsy Behav. 87 (2018) 46–58.

[78] Bo Lin, Wei Luo, Zhiling Luo, Bo Wang, Shuiguang Deng, Jianwei Yin, Mengchu Zhou, Bradykinesia recognition in Parkinson's disease via single RGB video, ACM Trans. Knowl. Discov. Data 14 (2) (2020) 1–19.

[79] Gareth Morinan, Yuriy Dushin, Grzegorz Sarapata, Samuel Rupprechter, Yuwei Peng, Christine Girges, Maricel Salazar, Catherine Milabo, Krista Sibley, Thomas Foltynie, Ioana Cociasu, Lucia Ricciardi, Fahd Baig, Francesca Morgante, Louise-Ann Leyland, Rimona S Weil, Ro'ee Gilron, Jonathan O'Keeffe, Computer vision quantification of whole-body Parkinsonian bradykinesia using a large multi-site population, NPJ Park. Dis. 9 (1) (2023) 10.

[80] Xing Liang, Epaminondas Kapetanios, Bencie Woll, Anastassia Angelopoulou, Real time hand movement trajectory tracking for enhancing dementia screening in ageing deaf signers of British Sign Language, IFIP Adv. Inf. Commun. Technol. 3 (2019) 377–394.

[81] Hao Li, Xiangxin Shao, Chencheng Zhang, Xiaohua Qian, Automated assessment of Parkinsonian finger-tapping tests through a vision-based fine-grained classification model, Neurocomputing 441 (2021) 260–271.

[82] Mandy Lu, Qingyu Zhao, Kathleen L Poston, Edith V Sullivan, Adolf Pfefferbaum, Marian Shahid, Maya Katz, Leila Montaser Kouhsari, Kevin Schulman, Arnold Milstein, Juan Carlos Niebles, Victor W Henderson, Li Fei-Fei, Kilian M Pohl, Ehsan Adeli, Quantifying Parkinson's disease motor severity under uncertainty using MDS-UPDRS videos, Med. Image Anal. 73 (102179) (2021) 102179.

[83] Letizia Gionfrida, Wan M R Rusli, Anil A Bharath, Angela E Kedgley, Validation of two-dimensional video-based inference of finger kinematics with pose estimation, PLoS One 17 (11) (2022) e0276799.

[84] Taeho Lee, Eun-Tae Jeon, Jin-Man Jung, Minsik Lee, Deep-learning-based stroke screening using Skeleton data from neurological examination videos, J. Pers. Med. 12 (10) (2022) 1691.

[85] Kye Won Park, Eun-Jae Lee, Jun Seong Lee, Jinhoon Jeong, Nari Choi, Sungyang Jo, Mina Jung, Ja Yeon Do, Dong-Wha Kang, June-Goo Lee, Sun Ju Chung, Machine learning-based automatic rating for cardinal symptoms of Parkinson disease, Neurology 96 (13) (2021) e1761–e1769.

[86] Trung-Hieu Hoang, Mona Zehni, Huaijin Xu, George Heintz, Christopher Zallek, Minh N Do, Towards a comprehensive solution for a vision-based digitized neurological examination, IEEE J. Biomed. Health Inform. 26 (8) (2022) 4020–4031.

[87] Gökhan Güney, Talisa S Jansen, Sebastian Dill, Jörg B Schulz, Manuel Dafotakis, Christoph Hoog Antink, Anne K Braczynski, Video-based hand movement analysis of Parkinson patients before and after medication using high-frame-rate videos and MediaPipe, Sensors (Basel) 22 (20) (2022) 7992.

[88] Zhu Li, Kang Lu, Miao Cai, Xiaoli Liu, Yanwen Wang, Jiayu Yang, An automatic evaluation method for Parkinson's dyskinesia using finger tapping video for small samples, J. Med. Biol. Eng. 42 (3) (2022) 351–363.

[89] Gianluca Amprimo, Irene Rechichi, Claudia Ferraris, Gabriella Olmo, Objective assessment of the finger tapping task in Parkinson's disease and control subjects using azure Kinect and machine learning, in: 2023 IEEE 36th International Symposium on Computer-Based Medical Systems, CBMS, 2023, pp. 640–645, http://dx.doi.org/10.1109/CBMS58004.2023.00293.

[90] Fanbin Gu, Jingyuan Fan, Zhaoyang Wang, Xiaolin Liu, Jiantao Yang, Qingtang Zhu, Automatic range of motion measurement via smartphone images for telemedicine examination of the hand, Sci. Prog. 106 (1) (2023) 368504231152740.

[91] Stefan Williams, Zhibin Zhao, Awais Hafeez, David C Wong, Samuel D Relton, Hui Fang, Jane E Alty, The discerning eye of computer vision: Can it measure Parkinson's finger tap bradykinesia? J. Neurol. Sci. 416 (117003) (2020) 117003.

[92] Adonay S Nunes, Nataliia Kozhemiako, Christopher D Stephen, Jeremy D Schmahmann, Sheraz Khan, Anoopum S Gupta, Automatic classification and severity estimation of ataxia from finger tapping videos, Front. Neurol. 12 (2021) 795258.

[93] Renjie Li, Rebecca J St George, Xinyi Wang, Katherine Lawler, Edward Hill, Saurabh Garg, Stefan Williams, Samuel Relton, David Hogg, Quan Bai, Jane Alty, Moving towards intelligent telemedicine: Computer vision measurement of human movement, Comput. Biol. Med. 147 (105776) (2022) 105776.
[94] Xinrui Huang, Xi Chen, Xiaoteng Shang, Shiwen Zhang, Jiyan Jin, Shuyang Li, Feifei Zhou, Ming Yi, Image-recognition-based system for precise hand function evaluation, Displays (102409) (2023) 102409.
[95] Jung Hwan Shin, Jed Noel Ong, Ryul Kim, Sang-Min Park, Jihyun Choi, Han-Joon Kim, Beomseok Jeon, Objective measurement of limb bradykinesia using a marker-less tracking algorithm with 2D-video in PD patients, Park. Relat. Disord. 81 (2020) 129–135.
[96] Gaëtan Vignoud, Clément Desjardins, Quentin Salardaine, Marie Mongin, Béatrice Garcin, Laurent Venance, Bertrand Degos, Video-based automated analysis of MDS-UPDRS III parameters in Parkinson disease, 2022, bioRxiv.
[97] MediaPipe, Hand landmarks detection guide, 2019, https://developers.google.com/mediapipe/solutions/vision/hand_landmarker. (Accessed 21 July 2023).
[98] Veronica Cimolin, Luca Vismara, Claudia Ferraris, Gianluca Amprimo, Giuseppe Pettiti, Roberto Lopez, Manuela Galli, Riccardo Cremascoli, Serena Sinagra, Alessandro Mauro, Lorenzo Priano, Computation of gait parameters in post stroke and Parkinson's disease: A comparative study using RGB-d sensors and optoelectronic systems, Sensors (ISSN: 1424-8220) 22 (3) (2022) http://dx.doi.org/10.3390/s22030824, URL https://www.mdpi.com/1424-8220/22/3/824.
[99] Chang Soon Tony Hii, Kok Beng Gan, Nasharuddin Zainal, Norlinah Mohamed Ibrahim, Shahrul Azmin, Siti Hajar Mat Desa, Bart van de Warrenburg, Huay Woon You, Automated gait analysis based on a marker-free pose estimation model, Sensors (ISSN: 1424-8220) 23 (14) (2023) http://dx.doi.org/10.3390/s23146489, URL https://www.mdpi.com/1424-8220/23/14/6489.
[100] Gilles Naeije, Antonin Rovai, Massimo Pandolfo, Xavier De Tiège, Hand dexterity and pyramidal dysfunction in Friedreich ataxia, a finger tapping study, Mov. Disord. Clin. Pract. 8 (1) (2021) 85–91.
[101] S Summa, J Tosi, F Taffoni, L Di Biase, M Marano, A Cascio Rizzo, M Tombini, G Di Pino, D Formica, Assessing bradykinesia in Parkinson's disease using gyroscope signals, IEEE Int. Conf. Rehabil. Robot. 2017 (2017) 1556–1561.
[102] Desiree Joy Lanzino, Megan N Conner, Kelli A Goodman, Kathryn H Kremer, Maegan T Petkus, John H Hollman, Values for timed limb coordination tests in a sample of healthy older adults, Age Ageing 41 (6) (2012) 803–807.
[103] R Colombo, A Raglio, M Panigazzi, A Mazzone, G Bazzini, C Imarisio, D Molteni, C Caltagirone, M Imbriani, The SonicHand protocol for rehabilitation of hand motor function: A validation and feasibility study, IEEE Trans. Neural Syst. Rehabil. Eng. 27 (4) (2019) 664–672.
[104] Matlab, Finddelay tutorial, 2020, https://www.mathworks.com/help/signal/ref/finddelay.html. (Accessed 22 July 2023).
[105] L.I. Lin, A concordance correlation coefficient to evaluate reproducibility, Biometric 45 (1989) 255–268.
[106] Terry K. Koo, Mae Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, J. Chiropr. Med. 15 (2) (2016) 155–163.
[107] R. Kotas, M. Janc, M. Kamiński, P. Marciniak, E. Zamysłowska-Szmytke, W. Tylman, Evaluation of agreement between static posturography methods employing tensometers and inertial sensors, IEEE Access 7 (2019) 164120–164126, http://dx.doi.org/10.1109/ACCESS.2019.2952496.
[108] Douglas G. Altman, Practical Statistics for Medical Research, CRC Press, 1990.
[109] Z. Cao, G. Hidalgo, T. Simon, S. Wei, Y. Sheikh, OpenPose: Realtime multi-person 2D pose estimation using part affinity fields, IEEE Trans. Pattern Anal. Mach. Intell. (ISSN: 1939-3539) 43 (01) (2021) 172–186, http://dx.doi.org/10.1109/TPAMI.2019.2929257.
[110] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Zhou Tianyi, Junsong Yuan, A2J: Anchor-to-joint regression network for 3D articulated pose estimation from a single depth image, in: Proceedings of the IEEE Conference on International Conference on Computer Vision, ICCV, 2019.
[111] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, Junsong Yuan, 3D hand shape and pose estimation from a single rgb image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10833–10842.
[112] Zhilin Guo, Weiqi Zeng, Taidong Yu, Yan Xu, Yang Xiao, Xuebing Cao, Zhiguo Cao, Vision-based finger tapping test in patients with Parkinson's disease via spatial-temporal 3D hand pose estimation, IEEE J. Biomed. Health Inf. 26 (8) (2022) 3848–3859, http://dx.doi.org/10.1109/JBHI.2022.3162386.
[113] Openpose, OpenPose 1.1.0 benchmark, 2023, https://docs.google.com/spreadsheets/d/1-DynFGvoScvfWDA1P4jDInCkbD4lg0IKOYbXgEq0sK0/edit. (Accessed 30 October 2023).
[114] Michal Tölgyessy, Martin Dekan, L'uboš Chovanec, Peter Hubinský, Evaluation of the azure Kinect and its comparison to Kinect V1 and Kinect V2, Sensors (ISSN: 1424-8220) 21 (2) (2021) http://dx.doi.org/10.3390/s21020413, URL https://www.mdpi.com/1424-8220/21/2/413.
[115] Vaidehi P. Wagh, Matthew W. Scott, Sarah N. Kraeutner, Quantifying similarities between MediaPipe and a known standard for tracking 2D hand trajectories, 2023, http://dx.doi.org/10.1101/2023.11.21.568085, bioRxiv.

Gianluca Amprimo received his Master's Degree in Computer Engineering from the Politecnico di Torino in 2020. He is currently a PhD student at the Control and Computer Engineering Department of Politecnico di Torino and a Research Fellow at the National Research Council (CNR-IEIIT) of Italy. His main research interests include human pose estimation from video, innovative technologies for telemonitoring and rehabilitation, and AI for medical application.

Giulia Masi received her Master's Degree in Biomedical Engineering at Politecnico di Torino in 2021, specialising in biosignal processing. She is currently a PhD student in Control and Computer Engineering at Politecnico di Torino. In particular, she is interested in the study of sleep and motion in the neurodegenerative diseases, as well as stress, using objective and quantitative measurements, such neurophysiological signals and motion trajectories.

Giuseppe Pettiti graduated in Electronic Engineering at Polytechnic of Turin in 1986. Since 1987, he is a researcher at CNR. Along the years his activity has been mainly carried out in image and signal processing scope. He participated to many projects in different contexts such as TV coding, industrial robotics, 3D reconstruction, cultural heritage conservation and restoration. During the last years he dealt with non-invasive technologies for movement analysis, sleep monitoring and rehabilitation of motor disabilities.

Gabriella Olmo (IEEE Senior Member) received her M.E. and Ph.D. degrees in Electronic Engineering from Politecnico di Torino, Italy, in 1986 and 1992, respectively. In 2016, she received her Master's degree in Medicine and Surgery from Università di Torino, Italy. She is currently a full professor in the Department of Control and Computer Engineering, Politecnico di Torino, Italy. Her main research interests are in the fields of wearable sensors, signal processing, and machine learning techniques for medical applications. She is coauthor of more than 250 publications in international journals and proceedings in international conferences.

Lorenzo Priano received his degree in Medicine and specialisation in Neurology from the University of Turin (Italy) in 1993 and in 1997, respectively. Currently, he is a researcher and associate professor at the Department of Neuroscience, University of Turin (Italy) and a neurologist at the Division of Neurology and Neurorehabilitation, San Giuseppe Hospital, I.R.C.S.S. Istituto Auxologico Italiano, in charge of the Laboratory of Clinical and Experimental Neurophysiology and Sleep Medicine Service for neurological disorders. His main research activities are in clinical neurology, sleep medicine, clinical and experimental neurophysiology for diagnostic and neurorehabilitation purposes,clinical and experimental neuropharmacology. He is the author and co-author of numerous scientific papers published in national and international journals, books, and conference proceedings.

Claudia Ferraris received her degree in Computer Science from the University of Turin (Italy) in 1997, then joined CNR working on image/video coding, compression techniques, and motion estimation algorithms. After a long work experience in the industrial field, she joined again CNR-IEIIT, to work on non-invasive technologies for motion analysis, remote monitoring, and rehabilitation in elderly and pathological conditions. Since 2020, she has been a permanent researcher at the same organisation. She received her Ph.D. in Neuroscience from the University of Turin (Italy) in 2021. She is the author and co-author of numerous research papers published in national and international journals, books, and conference proceedings.