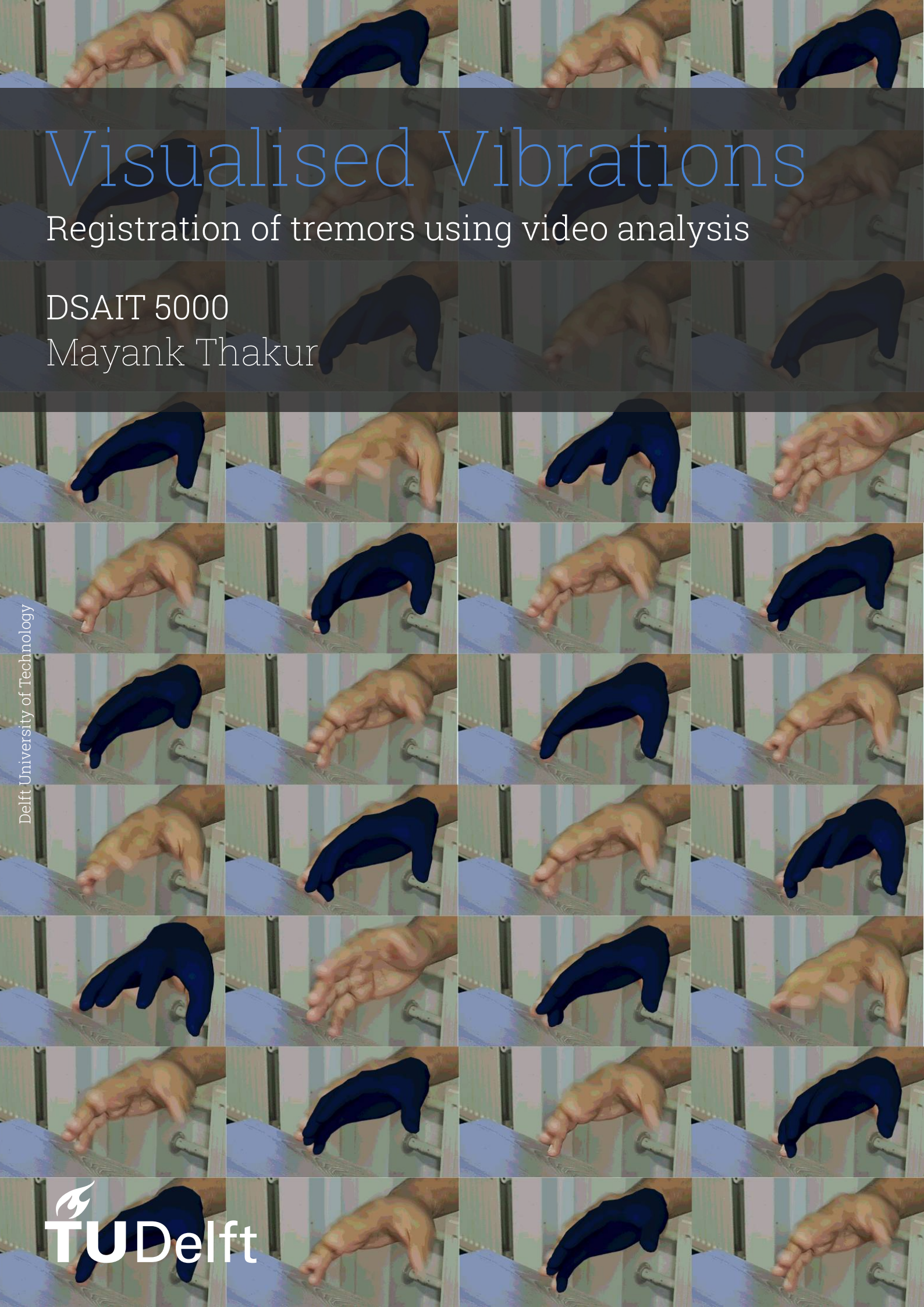


Visualised Vibrations

Registration of tremors using video analysis

DSAIT 5000

Mayank Thakur

Delft University of Technology

TU Delft

Visualised Vibrations

Registration of tremors using video analysis

by

Mayank Thakur

to obtain the degree of Master of Science in Data Science and AI Technology
at the Delft University of Technology,
to be defended publicly on November 26, 2026.

Supervisor:	Jan Van Gemert
Co-Supervisor:	Robert Van Den Berg
External Supervisor:	Michael Weinmann
Project Duration:	September 2025 - November 2026
Faculty:	Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft

Cover:	Designed by Mayank Thakur
Style:	TU Delft Report Style, with modifications by Daan Zwaneveld

TU Delft

Acknowledgments

I would like to thank my family and friends for helping me get through my thesis. None of this would be possible without the undying support of my parents. I also would not have been able to work on this thesis without the intellectual and emotional support of my friends.

I would also like to thank my supervisor, Jan Van Gemert, for stepping in so late and taking this project to completion.

Thank you!

Contents

Acknowledgments **i**

1 Introduction **1**

2 Preliminary Knowledge **3**
2.1 Tremor as a Temporal Signal 3
2.2 Representing a Hand in 3D 4
2.3 Neural Networks for Visual Estimation 4
2.4 Monocular Hand Pose Estimation 6
2.5 Temporal Consistency and Motion Priors 6
2.6 From Hand Motion to a One-Dimensional Signal 8
2.7 Evaluation of Frequency Estimates 9
2.8 Controlled and Real-World Evaluation 10

References **12**

3 Scientific Article **13**

A Declaration of AI Usage **29**

1

Introduction

Technology and medicine have long developed alongside one another, with technological advances enabling medical processes to become more efficient, quantitative, and, in many cases, more effective. Modern diagnostic and monitoring tools such as medical imaging, electroencephalography (EEG), and electrocardiography (ECG) allow physiological processes that cannot be assessed reliably through observation alone to be measured objectively. As both medicine and technology have developed, an increasing focus has therefore been placed on extracting useful information from these measurements to support clinical decision-making.

More recently, Artificial Intelligence (AI), and particularly Computer Vision (CV), has introduced new ways of extracting information from medical images and videos. Computer vision methods can identify patterns and structures in visual data that would otherwise require manual interpretation, and have consequently found applications across areas such as medical imaging and movement analysis. As these methods improve, increasingly detailed information can be obtained from ordinary visual recordings. This thesis investigates one such application: whether modern hand pose estimation and reconstruction methods can be used to quantify tremor from monocular videos.

Chronic Inflammatory Demyelinating Polyradiculoneuropathy (CIDP) is a rare immune-mediated neuropathy of the peripheral nervous system. In CIDP, the immune system damages components of the peripheral nerves, including their myelin sheath, interfering with the transmission of signals between the nervous system and the body. Tremor can occur as part of the clinical presentation of the disease and may affect the upper limbs, including the hands. One important characteristic of a tremor is its frequency: the number of oscillations occurring per second, expressed in Hertz (Hz). CIDP-associated upper-limb tremor has commonly been reported around 5-7 Hz when measured directly from hand or finger movement, although the frequency can vary between patients and between measurement locations [7, 1]. Quantifying this periodic movement can provide an objective description of the tremor rather than relying exclusively on visual assessment.

Current quantitative tremor assessment can be performed using an electrophysiological tremor registration. At Erasmus MC, this examination is performed by the Department of Clinical Neurophysiology and records the electrical activity of muscles using surface electrodes. Movement sensors may additionally be placed on the affected body parts to measure their movement, while the patient performs tasks in different postures and during directed movements [5]. Electrophysiological tremor analysis using surface electromyography and accelerometry is commonly used to characterize properties such as tremor frequency, rhythmicity, and the relationship between muscle activation and movement [2]. Although these measurements provide detailed information about the patient's movement, they require dedicated measurement equipment and an examination in a clinical setting. Video offers a substantially more accessible source of information.

Extracting tremor from video, however, introduces a different problem. Modern hand pose estimation methods can reconstruct the position and articulation of a hand from individual video frames, but these methods are generally developed to produce accurate or plausible hand poses rather than to measure

small periodic movements. Small frame-to-frame reconstruction errors can appear as artificial motion, while methods designed to smooth predictions over time may inadvertently suppress the tremor signal itself. Whether a visually plausible reconstruction also preserves the temporal information necessary for reliable tremor-frequency estimation is therefore not immediately clear.

Throughout this thesis, we investigate whether off-the-shelf monocular hand pose estimation and reconstruction methods can preserve and recover tremor-frequency information from video. We first study this problem under controlled conditions, where periodic hand movements with known reference frequencies can be recorded and evaluated directly. We then progressively move toward more complex activities and clinical recordings, while investigating which characteristics of the videos and reconstructed motion are associated with reliable or unreliable frequency estimates.

The remainder of this thesis is structured to make the work accessible to readers from different backgrounds. Chapter 2 introduces the clinical, computer vision, and signal-processing concepts required to understand the research. Chapter 3 then presents the main scientific contribution in CVPR article format.

2

Preliminary Knowledge

This chapter introduces the concepts required to understand the methods and experiments presented in the scientific article in Chapter 3. The work combines concepts from computer vision, 3D hand reconstruction, temporal motion modelling, and signal processing. Although each of these areas is broad, only the components directly relevant to tremor-frequency estimation from monocular video are discussed here.

The complete pipeline considered in this thesis can be viewed as a sequence of transformations. An RGB video is first converted into a temporally ordered representation of the hand using a hand tracking or 3D reconstruction model. The resulting hand trajectory is then converted into a motion signal, from which periodic components are estimated in the frequency domain. Finally, several measures are used to determine whether the resulting frequency estimate is likely to be reliable.

2.1. Tremor as a Temporal Signal

A tremor is an approximately rhythmic oscillatory movement of a body part. In this thesis, the main quantity of interest is its frequency, which describes how often the oscillation repeats per second and is expressed in Hertz (Hz). For example, a periodic motion repeating five times per second has a frequency of 5 Hz.

Tremor can occur as part of the clinical presentation of CIDP. In this work, the clinical mechanism causing the tremor is not modelled directly. Instead, the tremor is treated as a temporal movement pattern which may be visible in a sequence of hand poses.

Let the position of a point on the reconstructed hand at time t be

$$\mathbf{p}_t = \begin{bmatrix} x_t & y_t & z_t \end{bmatrix}^\top . \tag{2.1}$$

For a video containing T frames, the temporal trajectory of this point is given by

$$\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_T\} . \tag{2.2}$$

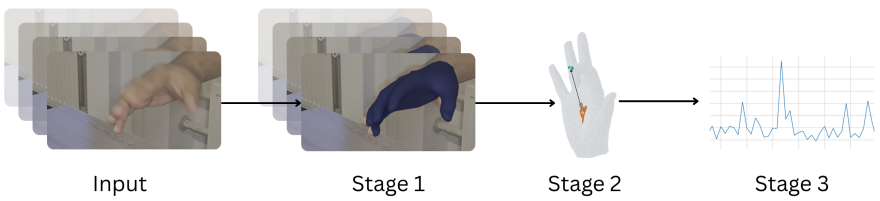


Figure 2.1: Overview of the proposed pipeline.

If the point undergoes periodic motion, this periodicity should be reflected in the trajectory, $\mathcal{P}$. The main objective of the computer vision stage of this thesis is to produce an anatomically plausible hand in every frame, while preserving small frame-to-frame changes that contain the tremor signal.

2.2. Representing a Hand in 3D

A central challenge in monocular hand reconstruction is that a 2D image does not uniquely determine a 3D hand configuration. Occlusion, or overlapping fingers in 3D poses may produce similar 2D projections. Modern hand reconstruction methods therefore often incorporate prior knowledge about the structure of human hands.

Two broad forms of output are particularly relevant to this thesis: hand landmarks and parametric hand meshes.

A landmark-based method represents the hand using a relatively small set of anatomically meaningful points. Such representations are computationally convenient, but they only describe a sparse subset of the hand surface. On the other hand, Mesh-based methods instead reconstruct a complete 3D surface. A hand mesh consists of vertices

$$\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}, \qquad \mathbf{v}_i \in \mathbb{R}^3, \tag{2.3}$$

connected by triangular faces. This representation makes it possible to analyse motion at many locations across the hand rather than only at a predefined set of joints.

The MANO Hand Model

MANO, the *hand Model with Articulated and Non-rigid defOrmations*, is a parametric model designed to represent the pose and shape of human hands [10]. It provides a compact representation from which both a complete hand mesh and hand joints can be reconstructed. It is derived from detailed 3D scans and its parameters allow it to represent a large variety of hand positions and shapes. The resulting model provides a low-dimensional representation while maintaining an anatomically structured hand surface.

The main idea behind MANO is to separate two sources of variation. The first is *shape*, which describes differences between hands, such as proportions and morphology. The second is *pose*, which describes the articulation of the hand, including the rotation and bending of the fingers. A simplified representation of the MANO model can be written as

$$\mathbf{V} = M(\boldsymbol{\beta}, \boldsymbol{\theta}), \tag{2.4}$$

where $\boldsymbol{\beta}$ denotes the shape parameters and $\boldsymbol{\theta}$ denotes the pose parameters. The output $\mathbf{V}$ is the reconstructed 3D hand mesh.

The distinction between shape and pose is particularly useful for video. Ideally, the physical shape of a person's hand should remain approximately constant throughout a recording, whereas the pose changes from frame to frame. A MANO-based video reconstruction can therefore be represented as a sequence of changing pose parameters combined with a comparatively stable hand shape.

2.3. Neural Networks for Visual Estimation

Modern hand-pose estimation methods are generally based on neural networks. A neural network can be viewed as a parameterized function

$$\mathbf{y} = f_{\boldsymbol{\theta}}(\mathbf{x}), \tag{2.5}$$

where $\mathbf{x}$ denotes an input, $\mathbf{y}$ the predicted output, and $\boldsymbol{\theta}$ the learnable parameters of the network. These parameters compose what we refer to as the neural network. In an image-based hand-pose estimation

setting, $\mathbf{x}$ is typically an RGB image or a cropped image containing the hand, while $\mathbf{y}$ may contain 2D landmarks, 3D joint locations, mesh vertices, or the parameters of a parametric hand model.

The ideal parameters $\boldsymbol{\theta}$, are found through a training process. For each training sample, the network produces a prediction which is compared with a desired target using a *loss function*. The loss function converts the difference between the prediction and the target into a numerical value. A small loss indicates that the prediction agrees well with the desired output, while a large loss indicates a poor prediction.

Given a training sample $(\mathbf{x}, \mathbf{y}^*)$, where $\mathbf{y}^*$ represents the target output, the training objective can generally be written as

$$\mathcal{L} = \mathcal{L}\left(f_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{y}^*\right). \tag{2.6}$$

Training aims to find parameters that minimize this loss across the training data,

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i^*\right), \tag{2.7}$$

where N is the number of training examples.

This optimization is usually performed iteratively using gradient-based methods. The gradient of the loss describes how the network parameters should change in order to reduce the prediction error. These gradients are calculated through *backpropagation*, which propagates the contribution of the error backwards through the network. An optimizer then updates the parameters according to these gradients. A simplified update can be written as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}, \tag{2.8}$$

where η is the learning rate, which controls the size of each parameter update.

The choice of training objective strongly affects what information the resulting model learns to preserve. For example, a hand-pose estimation network may be trained to minimize the distance between predicted and ground-truth joints, the difference between predicted and ground-truth mesh vertices, or a re-projection error between the estimated 3D hand and its observed 2D image location. More complex systems often combine several objectives,

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 + \cdots + \lambda_m \mathcal{L}_m, \tag{2.9}$$

where the coefficients $\lambda_1, \ldots, \lambda_m$ determine the relative importance of the different objectives.

It is important to note that a model becomes good at the properties emphasized during its training. Hand-pose estimators are (generally speaking) trained mainly for spatial pose accuracy, and are therefore encouraged to reconstruct anatomically correct hands in individual images, but not necessarily encouraged to preserve small temporal details. Furthermore, The models used in this thesis are predominantly pre-trained models: their parameters have already been learned from large external datasets before being applied to the videos considered here.

We have to distinguish between the *training objective* of a model and the *downstream objective* of this thesis. The hand-pose models are generally trained to recover the correct spatial configuration of the hand, whereas the downstream goal here is to recover a temporally faithful motion signal from which tremor frequency can be estimated. These objectives overlap, but they are not identical.

Architectural complexity of the evaluated models. The description above intentionally presents neural networks in a simplified form, focusing only on the concepts required to understand how they are trained and used for prediction. The hand-pose estimation methods evaluated in this thesis (WiLoR, HAMBA, and Mediapipe) use more specialized architectures internally. For example, WiLoR uses a Transformer-based reconstruction network with attention mechanisms, while HAMBA uses graph-guided state-space modelling to capture relationships between different parts of the hand. MediaPipe uses a comparatively simpler landmark-estimation pipeline. These architectural differences are relevant because they influence how the hand is reconstructed, but their internal mathematical details are not required to understand the downstream tremor-frequency analysis performed in this thesis.

2.4. Monocular Hand Pose Estimation

Monocular hand pose estimation attempts to recover the configuration of a hand using images from a single camera. Unlike multi-camera or depth-based systems, no direct depth measurement is available. The model must therefore infer the 3D structure of the hand from visual information in a 2D image.

This task is difficult for several reasons. Hands contain many articulated joints and can assume a large variety of poses. Fingers frequently occlude one another, while objects can obscure parts of the hand. Motion blur, illumination, image resolution and camera viewpoint can further affect the available visual information. The ambiguity is particularly important for this thesis because the analysed clinical videos were not recorded specifically for modern hand-reconstruction methods.

The pose estimators evaluated later in this thesis use different representations and architectures. Their common purpose in the pipeline, however, is straightforward: each frame of an RGB video is converted into a structured representation of the hand which can subsequently be analysed over time. We use the following pose estimators:

MediaPipe is a lightweight hand-tracking method that estimates predefined landmarks at locations such as the wrist and finger joints [11]. These landmarks can be directly tracked over time and converted into motion signals, making MediaPipe a useful baseline for tremor-frequency estimation. That being said, this representation is sparse compared to mesh-based methods, which provide motion information across the full hand surface.

WiLoR is an end-to-end method for 3D hand reconstruction in unconstrained images [8]. It first detects the hand and then uses a Transformer-based neural network to predict MANO parameters, which are further refined for improved image alignment. WiLoR was designed for in-the-wild conditions with variation in pose, illumination, and occlusion, making it suitable for the less controlled videos considered in this thesis.

HAMBA is a monocular 3D hand-reconstruction model that estimates a structured hand representation from individual RGB images [3]. Unlike WiLoR's Transformer-based approach, HAMBA uses graph-guided state-space modelling to capture spatial relationships between different parts of the hand.

These methods differ in how they use temporal information. In MediaPipe, landmarks estimated in one frame are used to help track and localize the hand in subsequent frames, reducing the need to repeatedly run the full palm detector. However, the hand pose itself is still estimated from the current image rather than from an complete motion sequence [11]. WiLoR and HAMBA are more strictly frame-based, reconstructing each image independently without enforcing any temporal consistency [8, 3]. As a result, all three methods can still produce frame-to-frame inconsistencies.

2.5. Temporal Consistency and Motion Priors

A simple solution for temporally inconsistent motion would be to smooth the trajectory. For many motion-reconstruction applications this is desirable because abrupt frame-to-frame changes are often reconstruction errors. Tremor estimation complicates this assumption. The movement of interest is itself a comparatively small and rapid oscillation. A method which suppresses all rapid motion may therefore remove both reconstruction noise and genuine tremor.

The desired objective is consequently not maximum smoothness. Instead, the reconstruction should suppress motion caused by estimation instability while preserving motion originating from the physical hand.

Motion Priors

A motion prior encodes knowledge about the kinds of movement that are plausible. Rather than independently deciding whether every pose is valid, the prior models relationships across a sequence. Suppose an observed sequence of noisy poses is represented by

$$\hat{\mathbf{Q}} = \left\{ \hat{\mathbf{Q}}_1, \ldots, \hat{\mathbf{Q}}_T \right\}. \tag{2.10}$$

A motion prior can constrain the reconstructed sequence $\mathbf{Q}$ such that it remains compatible with the observations while also resembling plausible motion. In a general optimization setting this can be written as

$$\mathbf{Q}^* = \arg\min_{\mathbf{Q}} \left[\mathcal{L}_{\text{observation}}(\mathbf{Q}, \hat{\mathbf{Q}}) + \lambda \mathcal{L}_{\text{prior}}(\mathbf{Q}) \right]. \tag{2.11}$$

The observation term encourages the refined motion to remain close to the estimated hand, while the prior term encourages plausible temporal behaviour. The scalar λ controls the balance between these objectives. If the prior is weighted too weakly, reconstruction jitter may remain. If it is weighted too strongly, the reconstructed sequence may become smoother than the actual physical movement.

Motion priors are typically learned from collections of high quality recorded motion. The learned model captures recurring temporal patterns. In this thesis, the motion prior used is the Hand Motion Prior (HMP) proposed by Duran et al. [4].

Using a motion prior changes the inference problem. A frame-based estimator asks *What hand pose is most likely in this image?*, whereas a motion-prior method additionally asks *What sequence of hand poses is compatible with both these observations and plausible hand motion?*.

STRIDE and Test-Time Training

Conventional supervised learning separates training and inference: model parameters are learned on a training dataset and then kept fixed when unseen samples are processed. Test-time training changes this assumption by allowing a pre-trained model to adapt using information from the test sample itself. If the original model parameters are denoted by $\boldsymbol{\theta}_0$, this adaptation can be written as

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\text{TTT}}\left(\boldsymbol{\theta}; \mathbf{x}_{\text{test}}\right), \tag{2.12}$$

where the optimization uses constraints that can be evaluated directly from the test data. These constraints can include temporal continuity, physical plausibility, consistency with observed poses, or agreement with other geometric information. The central benefit is that the model can adapt to a video containing conditions that were uncommon in the original training distribution.

STRIDE applies this principle to temporally continuous 3D pose estimation [6]. It first obtains a sequence of noisy 3D pose estimates from an arbitrary off-the-shelf estimator and then adapts a pre-trained motion prior to the specific test video. The resulting sequence is intended to be smoother and more temporally coherent. Its model-agnostic structure can be summarized as

$$\text{RGB video} \rightarrow \text{initial pose estimates} \rightarrow \text{temporal refinement}. \tag{2.13}$$

The original STRIDE method was developed for human-body pose estimation under occlusion, but its framework is useful for studying temporal refinement for CIDP as well. In this thesis, this principle is applied using HMP as the hand-motion prior. We not only want to determine whether temporal refinement produces smoother trajectories, but whether it improves tremor-frequency estimation as well.

2.6. From Hand Motion to a One-Dimensional Signal

MANO produces a dense 3D hand mesh containing 778 vertices per frame. Rather than reducing the entire hand to a single centroid, as done in previous work [9], this thesis analyses localized anatomical regions of the mesh. This is useful because a tremor may originate from a specific part of the hand, such as the index finger, and averaging over the complete hand can suppress such local motion.

The selected local regions are used to construct lower-dimensional motion signals suitable for frequency analysis. Let $\mathcal{R}$ denote a set of vertices representing one anatomical region. The mean position of this region at frame t is

$$\bar{\mathbf{p}}_{\mathcal{R},t} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \mathbf{v}_{i,t}. \tag{2.14}$$

Averaging several nearby vertices reduces the dependence on a single potentially noisy vertex while preserving localized movement.

To further reduce global hand translation, the displacement between two regions can be used. For regions $\mathcal{R}_1$ and $\mathcal{R}_2$,

$$\mathbf{d}_t = \bar{\mathbf{p}}_{\mathcal{R}_1,t} - \bar{\mathbf{p}}_{\mathcal{R}_2,t}. \tag{2.15}$$

The corresponding scalar motion signal can then be derived from the magnitude of this displacement or its temporal variation. This representation can then be studied using the same general frequency-analysis pipeline.

The Discrete Fourier Transform

A temporal signal describes how a quantity changes over time, while a frequency-domain representation describes which oscillatory components are present in that signal.

For a discrete signal $x[n]$ with N samples, the Discrete Fourier Transform (DFT) is

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N}. \tag{2.16}$$

The magnitude of $X[k]$ indicates the strength of the corresponding frequency component. For a periodic signal such as

$$x(t) = A \sin(2\pi f_0 t), \tag{2.17}$$

the frequency spectrum contains a strong component around f_0. In this thesis, this principle is used to identify periodic motion in reconstructed hand trajectories, where tremor must be separated from slower voluntary movement and reconstruction noise.

Power Spectral Density

The Power Spectral Density (PSD) describes how the power of a signal is distributed across different frequencies. A strong, narrow peak indicates that a substantial proportion of the signal is concentrated around one frequency, while a noisy or irregular signal generally distributes its power more broadly.

In this thesis, the PSD of the reconstructed motion signal is used to identify its dominant frequency. If $P(f)$ denotes the estimated PSD and $\mathcal{F}$ the analysed frequency range, the dominant frequency is defined as

$$\hat{f} = \arg\max_{f \in \mathcal{F}} P(f). \tag{2.18}$$

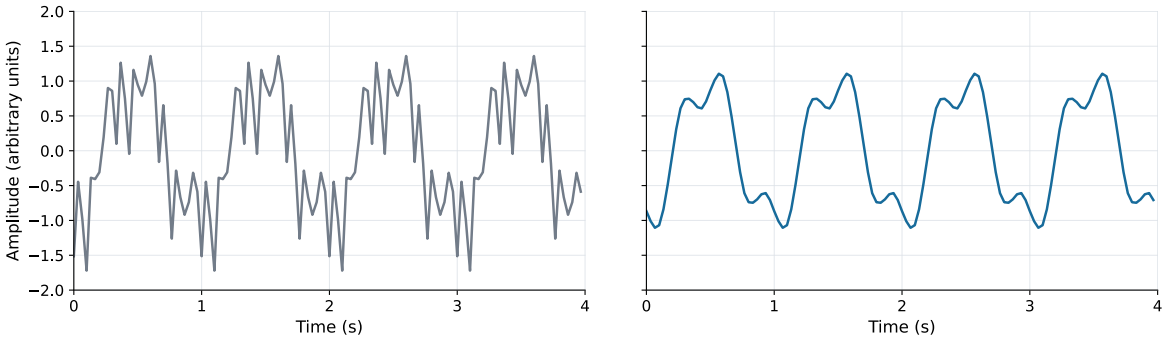


Figure 2.2: Impact of applying a 3rd order low-pass Butterworth filter on a periodic signal with high and low frequency components. The very high frequency components get removed

Importantly, a dominant frequency can still be returned when no clear periodic component is present, since some frequency will always contain the maximum power within the analysed range. The additional reliability measures introduced later are therefore used to determine how confidently this dominant frequency can be interpreted.

Signal Filtering

Video-derived hand trajectories contain more than just tremor. Voluntary movements, reconstruction noise, and camera motion can all appear in the same signal. Filtering is therefore used to reduce components that are not relevant to the frequency analysis and make the tremor motion easier to identify.

The analysis in this thesis uses a third-order Butterworth low-pass filter with a cut-off frequency of 4 Hz, and a high-pass filter with a cut-off frequency of 12 Hz, applied to signals sampled at 30 Hz. A low-pass filter preserves slower changes in the signal while progressively attenuating higher-frequency components. The Butterworth design is chosen for its smooth frequency response, avoiding strong fluctuations within the retained frequency range. The effects of using such a filter are visualized in Figure 2.2.

One consequence of this configuration is that, although frequencies up to 12 Hz are considered during peak selection, components above the 6 Hz filter cut-off are increasingly attenuated. The choice of filtering parameters can therefore directly influence which tremor frequencies remain visible in the resulting spectrum and must be considered when interpreting the estimated frequency.

2.7. Evaluation of Frequency Estimates

When a reference frequency f_{ref} is available, frequency-estimation accuracy is measured using the absolute error

$$E_{\text{abs}} = \left| \hat{f} - f_{\text{ref}} \right|, \tag{2.19}$$

and the Mean Absolute Error (MAE) across N recordings,

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{f}_i - f_{\text{ref},i} \right|. \tag{2.20}$$

MAE is expressed in Hertz and is therefore directly interpretable as the average difference between the estimated and reference frequencies.

Dummy baseline. A simple dummy estimator can be used as a reference point when evaluating a predictive method. Such an estimator ignores the input data and follows a fixed prediction strategy,

providing a baseline against which the performance of the actual estimation pipeline can be compared. When assessing the effectiveness of complex CV tasks, the dummy estimator can help ground the results, by comparing results against a very simple estimator.

Candidate Signal and Recording Measures

A dominant frequency can be extracted even from a noisy signal with no clear periodic component, since some frequency will always have the highest power within the analysed range. For recordings without frequency ground truth, additional characteristics of the reconstructed motion may therefore help indicate when an estimate is more or less trustworthy.

This thesis investigates six candidate measures, each motivated by a different possible failure mode. Importantly, these measures are not assumed in advance to be reliable predictors of frequency accuracy; their usefulness is evaluated experimentally in the scientific article.

- **Peak ratio** measures how strongly the dominant spectral peak stands out relative to the remaining power in the analysed frequency range. It is intended to capture spectral ambiguity.
- **Tremor amplitude** describes the magnitude of the reconstructed motion using its root-mean-square (RMS) value. It can indicate when an oscillatory component is weak or has been attenuated by processing.
- **Temporal noise** measures the variability of frame-to-frame motion and is intended to capture reconstruction jitter. A low value is not necessarily desirable, since excessive smoothing can also reduce genuine tremor motion.
- **Spatial coherence** measures how similarly neighbouring hand regions move over time. It is intended to distinguish structured hand motion from independently jittering mesh regions.
- **Palm orientation** describes the orientation of the reconstructed palm relative to the camera. Unlike the previous measures, it characterizes the recording geometry and is used to investigate sensitivity to viewpoint.
- **Frequency stability** measures how much the estimated dominant frequency changes across shorter temporal windows of the same recording. Large variation indicates that the frequency estimate is not temporally consistent.

2.8. Controlled and Real-World Evaluation

In this thesis, we gradually move from a controlled setting into the clinical setting. This structure is necessary because the final clinical videos do not contain synchronized tremor-frequency ground truth.

As it stands, an incorrect frequency estimate is difficult to diagnose. The error could originate from any component of the pipeline. The controlled dataset provides a setting in which these components can first be evaluated under simpler conditions. Participants simpler, tremor-like movements while recording conditions are changed in a controlled way. This allows us to diagnose potential issues in the pipeline.

In the controlled dataset we also vary characteristics such as participant and camera viewpoint. This allows the relationship between reconstruction conditions and estimation accuracy to be investigated while a reference frequency is still available.

Before moving to the clinical setting, we evaluate the pipeline on drawing tasks. Participants trace lines and spirals normally, and while emulating tremor-like motion. These recordings more closely resemble the activities in the clinical dataset. However, there is no frequency reference, and no ground truth. In this stage, the analysis begins to depend more strongly on the reliability metrics described earlier.

Finally, we apply the pipeline to the clinical setting, the EMC recordings. Here, we don't have any ground truth. The purpose of the reliability metrics is to investigate whether measurable characteristics of the reconstructed motion can distinguish recordings that provide a clear and stable tremor signal from recordings for which the resulting frequency estimate should be treated cautiously.

The progression can therefore be summarized as

$$\begin{array}{l} \text{controlled periodic motion} \\ \quad\downarrow \\ \text{periodic motion combined with voluntary activity} \\ \quad\downarrow \\ \text{unconstrained clinical recordings.} \end{array} \tag{2.21}$$

As the recording conditions become less controlled, direct supervision decreases and the importance of signal-reliability measures increases.

This experimental structure also separates two different questions. The first is whether a pose-estimation and signal-processing pipeline can recover a known periodic movement under favourable conditions. The second is whether the same pipeline provides enough information to determine when an estimate from an uncontrolled recording is reliable. The scientific article in the following chapter investigates both questions.

References

[1] "Chronic inflammatory demyelinating polyradiculoneuropathy-associated tremor: Phenotype and pathogenesis". In: *European Journal of Neurology* (2023). DOI: 10.1111/ene.15693.

[2] G. Deuschl et al. "The clinical and electrophysiological investigation of tremor". In: *Clinical Neurophysiology* 136 (2022), pp. 93–129. DOI: 10.1016/j.clinph.2022.01.004.

[3] Haoye Dong et al. "Hamba: Single-view 3D Hand Reconstruction with Graph-guided Bi-Scanning Mamba". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 2127–2160.

[4] Enes Duran et al. "HMP: Hand Motion Priors for Pose and Shape Estimation From Video". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Jan. 2024, pp. 6353–6363.

[5] Erasmus MC. *Tremorregistratie*. Patient information, Department of Clinical Neurophysiology. Accessed September 2026.

[6] Rohit Lal et al. "STRIDE: Single-Video Based Temporally Continuous Occlusion-Robust 3D Pose Estimation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Feb. 2025, pp. 794–803.

[7] "Postural tremor and chronic inflammatory demyelinating polyneuropathy". In: *Muscle & Nerve* 55 (2017), pp. 338–343. DOI: 10.1002/mus.25253.

[8] Rolandos Alexandros Potamias et al. "WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-wild". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 2025, pp. 12242–12254. DOI: 10.1109/CVPR52734.2025.01143.

[9] Noa Reijmers. *Trilling in beeld: Tremorregistratie middels videoanalyse*. Technical Medicine Internship Report. Delft, The Netherlands: Delft University of Technology, Nov. 2023.

[10] Javier Romero, Dimitrios Tzionas, and Michael J. Black. "Embodied Hands: Modeling and Capturing Hands and Bodies Together". In: *ACM Transactions on Graphics* 36.6 (Nov. 2017), 245:1–245:17. DOI: 10.1145/3130800.3130883.

[11] Fan Zhang et al. "MediaPipe Hands: On-device Real-time Hand Tracking". In: *arXiv preprint arXiv:2006.10214* (2020). arXiv: 2006.10214 [cs.CV].

3

Scientific Article

Visualized Vibrations: Registration of Tremors Using Video Analysis

Mayank Thakur
TU Delft

Abstract

Quantifying hand tremor from ordinary RGB videos offers a potentially accessible and contactless alternative to characterize patients with Chronic Inflammatory Demyelinating Polyradiculoneuropathy (CIDP). However, modern hand pose estimation methods are not optimized for the preservation of small, periodic movements. We investigate the reliability of monocular video-based tremor frequency estimation using a model-agnostic pipeline that extracts periodic motion from reconstructed hand trajectories. We compare four methods (MediaPipe, HAMBA, WiLoR, and STRIDE) on controlled recordings of emulated finger and wrist tremors, followed by more complex drawing activities and clinical recordings of CIDP patients. Under controlled conditions, WiLoR achieves the lowest IQR MAE of 0.12 Hz, while temporal refinement using STRIDE increases the error to 0.27 Hz, demonstrating that smoother hand reconstructions do not necessarily yield more accurate frequency estimates. Lastly, we investigate the reliability of tremor frequency estimates, and whether they can be assessed quantitatively without annotated data. We propose and apply metrics to controlled and clinical recordings, trying to characterize reconstruction reliability. The findings demonstrate the feasibility of our method under controlled conditions while highlighting the need for further research before extending to clinical settings.

1. Introduction

CIDP is a rare immune-mediated neuropathy of the peripheral nervous system, with an estimated pooled prevalence of approximately 2.8 cases per 100,000 people [1]. Tremor is a common manifestation of CIDP, with upper-limb postural and action tremor reported in approximately two-thirds of patients in a clinical cohort [12]. Tremor is generally assessed using clinical rating scales, while quantitative measurements of characteristics such as tremor frequency and amplitude can be obtained using instruments such as accelerometers and surface electromyography (EMG) [4].

However, tremor characteristics can vary over time and with daily activities, meaning that measurements obtained during a scheduled clinical examination may not fully represent a patient's symptoms in everyday settings [4, 8]. This motivates the investigation of more accessible methods that can complement clinical assessments and enable repeated measurements outside specialized clinical environments. Ordinary RGB videos offer a potentially contactless and widely accessible means of recording hand tremor, without requiring additional sensors to be attached to the patient [**?**].

Recent advances in hand pose estimation and 3D hand reconstruction have the potential to recover detailed hand motion from ordinary RGB videos. In this work, we focus primarily on tremor frequency and estimate the dominant periodic component of the reconstructed hand motion within a tremor-frequency band.

However, accurate hand reconstruction does not necessarily imply accurate tremor estimation. A Tremor consists of relatively small and rapid movements, while hand pose estimation methods are generally optimized for spatial pose accuracy or visually plausible reconstruction. Therefore, standard training losses may be introducing artificial high-frequency motion, while temporal refinement methods (intended to reduce jitter) may be suppressing tremor signals. Furthermore, factors such as viewpoint, hand orientation, occlusion, camera motion, and the type of hand movement may influence the quality of the reconstructed temporal signal.

Previous work has mostly investigated Parkinsonian tremors. Recent Computer Vision (CV) research has substantially improved monocular hand pose estimation and reconstruction. Existing hand reconstruction methods are typically evaluated using spatial reconstruction accuracy and temporal consistency, rather than the accuracy with which they preserve small periodic motion for downstream frequency estimation. It remains unclear whether modern off-the-shelf hand pose estimation methods preserve the temporal information required for tremor-frequency analysis, particularly outside controlled recording conditions.

To investigate this, we develop a model-agnostic pipeline that converts hand-pose estimates across time into signals

suitable for frequency analysis. We evaluate the pipeline using both controlled and clinical recordings. We first study emulated periodic finger and wrist movements with known reference frequencies, allowing direct evaluation of frequency-estimation accuracy across different hand pose estimation and refinement methods. We then investigate whether characteristics of the reconstructed motion can explain variations in estimation accuracy. Finally, we extend the analysis to more complex controlled activities and in-the-wild clinical recordings for which reliable frequency ground truth is unavailable, to determine whether these measures can indicate when a frequency estimate should or should not be trusted.

1.1. Research Questions

The overarching research question addressed in this paper is:

How reliably can tremor frequency be estimated from monocular in-the-wild videos using computer vision–based hand tracking and reconstruction methods?

This question is divided into the following subquestions:

1. How accurately do different hand pose estimation and temporal refinement methods estimate the frequency of emulated periodic hand movements under controlled recording conditions?
2. Which characteristics of the reconstructed motion and recording conditions are associated with accurate and stable tremor-frequency estimation?
3. Can reliability measures developed under controlled conditions help assess the reliability of frequency estimates in recordings without ground truth, including clinical videos?

1.2. Contributions

The main contributions of this paper are:

1. A comparative evaluation of off-the-shelf hand tracking, 3D hand reconstruction, and temporal refinement methods for monocular video-based tremor-frequency estimation.
2. A model-agnostic framework for extracting and quantifying tremor-frequency information from temporally varying hand-pose estimates.
3. An investigation of objective reliability measures, examining their relationship with frequency-estimation accuracy under controlled conditions and their applicability to clinical recordings without ground-truth frequency measurements.

2. Related Works

2.1. Video-Based Tremor Analysis

CV has previously been used as a contactless alternative to sensor-based tremor measurement, with most work focusing on Parkinson's disease and essential tremors. Pintea et al. [9] investigated hand-tremor frequency estimation directly from RGB video using both Lagrangian and Eulerian approaches. The former estimates the frequency from a tracked hand trajectory, while the latter analyzes temporal changes in the image after compensating for larger hand movements. Williams et al. [13] similarly estimated tremor frequency from smartphone videos using optical flow and compared the resulting estimates with accelerometer measurements. These works demonstrated that clinically relevant tremor frequencies can be recovered from video without attaching additional sensors to the patient.

More recent methods have incorporated learned hand-tracking models into the analysis. Güney et al. [6] used MediaPipe to automatically track hand motion in high-frame-rate recordings of patients with Parkinson's disease and compared the extracted movement characteristics before and after medication. This removes some of the manual localization required by earlier approaches, but still relies on high-frame-rate recordings and a predefined examination procedure. More recently, Guarín [5] proposed a pipeline for estimating both tremor frequency and amplitude from standard videos. The method combines hand tracking with 3D body-pose estimation and image-based scaling, and evaluates modern hand-pose methods including WiLoR. Under a standardized postural tremor task, the automatically derived measurements showed strong agreement with manually obtained measurements.

These studies establish that video contains sufficient information for tremor quantification. However, the recording conditions are generally constrained through predefined patient poses, camera placement, frame rates, or examination tasks. Guarín also reports that existing hand-pose models can become unreliable when the hand is observed from postures that differ from those common in their training data, motivating the use of a custom hand-localization model. This distinction is important for the present work. Rather than determining whether tremor frequency can be recovered from video under favourable conditions, we investigate how reliably it can be recovered when the recording conditions, activities, hand orientations, and overall video quality are not strictly controlled. Furthermore, previous video-based tremor research has primarily focused on Parkinson's disease and essential tremor, while considerably less work has investigated video-based tremor analysis for CIDP.

2.2. Monocular Hand Pose Estimation

A second body of related work concerns the models used to obtain the hand trajectory itself. MediaPipe Hands [14] provides an efficient landmark-based representation and has consequently been adopted in several medical movement-analysis pipelines. While such landmarks provide a simple representation for frequency analysis, recent computer-

vision research has increasingly moved towards complete 3D reconstruction of the hand.

Many modern hand-reconstruction methods build on the MANO parametric hand model [11], which represents the hand using a compact set of pose and shape parameters. Recent approaches therefore recover not only individual landmarks, but an articulated 3D hand mesh from monocular RGB images. WiLoR [10] combines an in-the-wild hand detector with a transformer-based 3D reconstruction network. It was trained using a large collection of in-the-wild hand images, and predicts MANO-based hand reconstructions. Importantly for video analysis, WiLoR processes individual frames independently: its authors demonstrate smooth tracking from video without introducing an explicit temporal model.

Hamba [2] provides another recent approach to single-image 3D hand reconstruction. Instead of relying only on transformer attention, Hamba introduces graph-guided state-space modelling to explicitly represent the spatial relationships between hand joints and reports strong performance on established 3D hand benchmarks. Both WiLoR and Hamba are therefore representative of recent high-performing monocular hand reconstruction models, but neither is primarily designed to preserve small temporal oscillations between consecutive frames.

This creates an important distinction between spatial reconstruction accuracy and suitability for tremor analysis. A model can produce anatomically plausible hands on individual frames while small frame-to-frame errors introduce artificial high-frequency motion into the resulting trajectory. Conversely, small genuine movements may be lost even when individual reconstructed frames appear visually accurate. Existing 3D hand-pose benchmarks primarily evaluate spatial reconstruction errors rather than the fidelity of the temporal frequency content produced when the model is applied to video. The synthetic-to-real analysis of Zhao et al. [15] further demonstrates that hand-pose performance is affected by factors such as pose distributions, occlusions, the presence of the forearm, and visual appearance, illustrating why performance on established reconstruction datasets does not necessarily translate directly to a new recording domain.

2.3. Temporal Reconstruction and Motion Priors

Frame-wise pose estimation naturally leads to temporal inconsistencies when the predictions from consecutive frames are considered as a sequence. Duran et al. [3] address this problem with Hand Motion Priors (HMP). Instead of training directly on a limited video dataset, HMP learns a generative hand-motion prior from large-scale motion-capture data. During inference, noisy observations are optimized in the latent space of this prior, constraining the resulting sequence towards plausible hand motion. The optimization can be adjusted for the downstream task using hyperparameters. These include the learning rate and loss weights controlling the relative influence of initial pose estimates, 2D keypoint alignment, hand-shape regularization, temporal smoothness, and the learned motion prior. The optimization proceeds in two stages, each using different objectives and coefficients. The approach improves temporal stability and robustness to occlusion compared with conventional single-frame reconstruction methods.

A related approach is STRIDE [7], which introduces test-time training for temporally continuous 3D human-pose estimation. STRIDE first obtains noisy poses using an arbitrary off-the-shelf estimator and then adapts a pre-trained motion prior to the particular test video. This makes the framework model-agnostic and allows temporal information to be incorporated without retraining the original pose-estimation model. STRIDE is particularly relevant to this work because it represents a way of separating spatial pose estimation from temporal refinement.

However, temporal consistency presents a different problem when the movement of interest is itself a small and relatively high-frequency oscillation. A strong smoothness prior can remove estimation noise, but may simultaneously attenuate the physiological tremor signal. Temporal smoothness should therefore not automatically be interpreted as improved tremor reconstruction. This issue is also visible more generally in pose-estimation literature: the STRIDE evaluation notes that a low acceleration error can result from over-smoothing while the reconstructed poses themselves remain inaccurate. Therefore, we cannot assume that a smoother trajectory is necessarily a better one.

Existing work consequently addresses different parts of the problem considered in this paper. Video-based medical studies demonstrate that tremor frequency can be measured from video, while recent hand-pose models provide increasingly detailed 3D representations of the hand. Temporal reconstruction methods can further reduce noise and recover more coherent motion. However, these areas are generally evaluated using different objectives. Tremor-analysis methods commonly assume standardized recording conditions, hand-pose methods primarily optimize spatial reconstruction accuracy, and temporal methods explicitly encourage smooth and plausible motion. For tremor estimation from existing clinical videos, these objectives cannot be considered independently. The estimated signal depends not only on the frequency-analysis method, but also on whether the hand reconstruction remains temporally consistent under real-life conditions.



Figure 1. Overview of the proposed pipeline. From left to right, given an RGB video, an off-the-shelf hand tracking or reconstruction method estimates hand landmarks or meshes for each frame (Stage 1). For mesh-based estimates, localized vertex neighborhoods are selected around two anatomical locations, and the difference between their centroids defines a relative displacement trajectory (Stage 2). The trajectory is filtered and converted into a mean-centered magnitude signal, from which the power spectral density (PSD) is estimated (Stage 3). The dominant frequency within the tremor band is extracted with other measures.

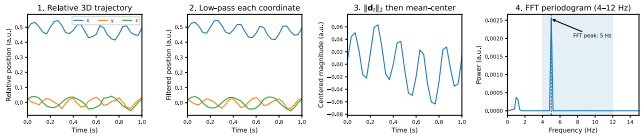


Figure 2. The four panels illustrate successive processing stages. The coordinates of a three-dimensional relative-displacement trajectory are low-pass filtered, converted to their Euclidean magnitude, and mean-centered. A one-sided FFT periodogram is then used to select the highest-power frequency bin within the 4-12 Hz analysis band. This procedure extracts the dominant periodic component of the reconstructed motion.

3. Method

3.1. Hand Motion Estimation

For MANO-based methods, each frame contains a reconstructed hand mesh and associated pose parameters. Since hand tremor may not affect the complete hand uniformly, reducing the complete mesh to a single centroid can suppress the tremor signal. The analysis therefore operates on localized regions of the hand.

For each selected anatomical region r, let $\mathcal{V}_r$ denote the corresponding set of mesh vertices or hand landmarks and let $\mathbf{x}_{i,t} \in \mathbb{R}^3$ be the location of point i in frame t. The mean position of the region is

$$\mathbf{c}_{r,t} = \frac{1}{|\mathcal{V}_r|} \sum_{i \in \mathcal{V}_r} \mathbf{x}_{i,t}. \tag{1}$$

For a selected pair of regions a and b, we form a relative 3D trajectory,

$$\mathbf{d}_t = \mathbf{c}_{a,t} - \mathbf{c}_{b,t}. \tag{2}$$

A local finger region can, for example, be measured relative to a palm region. Averaging within each region reduces the influence of errors at individual mesh points, while subtraction of the two region positions removes common global translation. Comparable anatomical regions are selected for each model output, whether represented by mesh vertices or landmarks.

This representation also makes the analysis largely model-agnostic: provided that a pose estimator produces hand landmarks or mesh points over time, the same downstream frequency-estimation procedure can be applied. Model predictions are ordered according to their original frame identifiers. Frames with missing or incomplete hand reconstructions are excluded from the analysis. The remaining observations are concatenated into a continuous sequence with spaces in between to maintain the temporal consistency of the data.

3.2. Frequency Estimation Pipeline

For each recording, we obtain a relative three-dimensional displacement trajectory $\mathbf{d}_t$ containing T retained observations. The frequency-analysis implementation assumes a sampling frequency of $f_s = 30$ Hz. A third-order Butterworth low-pass filter with a cutoff frequency of 4 Hz is and a high-pass filter with cutoff frequency of 12 Hz is applied independently to the x, y, and z coordinates. We then calculate the Euclidean magnitude of the filtered trajectory and subtract its temporal mean:

$$\widetilde{\mathbf{d}}_t = \mathcal{L}[\mathbf{d}]_t, \qquad y_t = \|\widetilde{\mathbf{d}}_t\|_2 - \frac{1}{T} \sum_{\tau=0}^{T-1} \|\widetilde{\mathbf{d}}_\tau\|_2, \tag{3}$$

where $\mathcal{L}$ represents the coordinate-wise filter. The resulting scalar signal y_t describes changes in the magnitude of relative hand displacement. The 4–12 Hz interval defines the subsequent frequency-search band rather than a band-pass filter. Frequencies above the 4 Hz cutoff therefore remain eligible for peak selection but are progressively attenuated by preprocessing, potentially reducing sensitivity to higher-frequency movements.

We estimate spectral power using a one-sided fast Fourier transform (FFT) periodogram, from Numpy.

For the isolated controlled movements, frequency-estimation error is $|\hat{f} - f_{\text{ref}}|$. Finger recordings have manually annotated movement cycles from which reference frequencies are obtained; when an independently measured reference is unavailable, the prescribed metronome frequency is used instead. Prescribed frequencies should not be interpreted as exact measurements of the performed motion.

For wrist based tremors, the lower point of the palm is used, and for drawing and EMC recordings, no sufficiently precise or synchronized frequency reference is available.

3.3. Metrics

Dominant frequency alone does not indicate whether a frequency estimate is reliable, as a spectral peak can be identified even in noisy or non-periodic signals. We therefore in-

troduce six additional metrics to characterize potential failure modes in the reconstructed hand motion.

Tremor Amplitude measures the root-mean-square (RMS) amplitude of the band-pass-filtered motion signal, characterizing the magnitude of the reconstructed oscillatory motion.

Peak Ratio measures the proportion of FFT spectral power concentrated at the dominant frequency within the 4–12 Hz band. It is calculated as $P_{\mathrm{FFT}}(\hat{f}) / \sum_{f_k \in [4,12]\,\mathrm{Hz}} P_{\mathrm{FFT}}(f_k)$. A more concentrated peak does not independently establish that its frequency is correct.

Temporal Noise measures the variance of frame-to-frame displacement magnitudes calculated from the unfiltered relative 3D trajectory between two anatomical regions. It characterizes motion variability and potential reconstruction jitter.

Spatial Coherence measures the mean valid Pearson correlation between frame-to-frame displacement magnitudes of spatially adjacent mesh vertices. It characterizes the consistency of reconstructed hand motion across neighboring regions. This metric is unavailable for MediaPipe in the current implementation.

Palm Orientation measures the mean angle of the palm relative to the camera and its variation over time, allowing us to investigate the influence of viewpoint on reconstruction quality.

Frequency Stability measures the standard deviation of dominant-frequency estimates across (50 %) overlapping temporal windows. The dominant frequency is estimated independently within each window using the same FFT procedure, and the population standard deviation of the resulting frequency estimates is reported. Lower values indicate greater temporal consistency.

Together, these metrics characterize different aspects of the reconstructed motion and recording conditions. Rather than assuming that each metric independently determines estimation quality, we investigate their relationships with frequency estimation accuracy. Their mathematical definitions and further interpretations are provided in Appendix C.

3.4. STRIDE

Most evaluated hand reconstruction methods do not explicitly enforce temporal consistency across the resulting sequence. To investigate whether temporal refinement improves frequency estimation, we additionally refine WiLoR predictions using a STRIDE-based workflow with the HMP (Hand Motion Priors) backend. This implementation initializes from frame-wise WiLoR hand reconstructions and uses an HMP learned hand-motion prior with per-video optimization to obtain temporally refined trajectories.

Temporal refinement is intended to reduce frame-to-frame reconstruction noise, but this may also attenuate genuine rapid hand oscillations. We therefore evaluate STRIDE as a refinement of WiLoR, rather than as an independent hand detector.

3.5. Implementation Details

All inference and analysis code was implemented in Python 3.10. Model inference was performed on the DelftBlue high-performance computing cluster using NVIDIA A100 GPUs, while the downstream signal analysis was implemented using NumPy and SciPy.

The same frequency-analysis implementation and parameter settings are applied to corresponding outputs from each evaluated model to ensure that differences in the reported frequency estimates originate from the reconstructed trajectories rather than from different downstream processing procedures. Code for model inference, signal analysis, and evaluation is planned for release at https://github.com/Person-Maink/localmodels-.

4. Experiments

4.1. Datasets

We use two complementary datasets. The Erasmus Medical Center (EMC) dataset contains clinical recordings of CIDP patients and represents the target real-world setting. Since these recordings do not contain synchronized tremor-frequency ground truth, we additionally construct a controlled dataset in which periodic hand movements are performed at prescribed frequencies.

4.1.1. EMC Dataset

The EMC dataset was collected at Erasmus Medical Center in collaboration with Amsterdam UMC and contains recordings of CIDP patients performing activities intended to expose upper-limb tremor. These include writing, drawing spirals and straight lines, pouring between cups, and several postural and full-body tasks.

For this study, only recordings containing hand activities suitable for monocular hand-pose estimation are considered. As the original recordings may contain identifying information and multiple activities within a single video, anonymized segments containing a single approximately homogeneous activity are extracted prior to analysis. Following screening, the final dataset comprises 11 clips from four participants. The recordings were acquired using a Sony camcorder, with a stored resolution of 1440 × 1080 pixels and an average frame rate of 25 frames per second.

4.1.2. Controlled Dataset

The controlled dataset was created to provide recordings for which the expected periodic motion is known and individual recording conditions can be varied independently.

Five volunteers without CIDP performed periodic movements at prescribed frequencies of 4 and 5 Hz while following a metronome. These frequencies were selected because participants could reproduce them more consistently than higher-frequency oscillations. Four participants were used for development and configuration selection, while the remaining participant was held out for the reported evaluation (24 recordings in the isolated-motion comparison).

Recordings were captured using a Samsung Galaxy S20 Ultra at 1920 × 1080 resolution and 30 frames per second. Participants performed the tasks over a plain tabletop, with the hand positioned near the centre of the frame and illuminated using a fill light. Recordings were acquired from multiple camera viewpoints. This allowed us to capture the same motion from the left side, the right side, and 3 other intermediary angles (i.e. 0, 45, 90, 135, 180 degrees), as well a birds-eye camera angle. All controlled recordings are 10 seconds in duration.

Two isolated motion types were recorded. In the finger condition, periodic motion originates primarily from the index finger, providing a localized and comparatively simple signal. In the wrist condition, periodic movement originates at the wrist and affects a larger part of the hand. The finger-motion recordings were manually annotated by marking movement cycles, from which reference frequencies were calculated. These cycle-derived frequencies were used as measured references in the evaluation.

Participants also completed three drawing tasks: a straight line, a large spiral, and a small spiral (Appendix A). Each task was performed once under a control condition and once while emulating a tremor. Unlike the previously described isolated motion, these recordings contain both voluntary task motion and an emulated tremor and therefore provide an intermediate setting between the simple controlled recordings and the clinical videos.

Of the 5 participants, the videos of 4 were used for the hyperparameter tuning stage, and 1 was used to report results.

4.2. STRIDE Hyperparameter Selection

Experimental question: *Which STRIDE configuration achieves the lowest frequency-estimation error across different movement conditions?*

STRIDE hyperparameter selection was treated as tuning the per-video refinement procedure, rather than training an additional model. The configurations were selected using four development participants, while the fifth participant was held out throughout the selection process.

A two-stage search was performed to limit the computational cost of per-video optimization. First, 48 configurations were evaluated on a small development subset. The search varied the number of HMP pose-optimization iterations (60, 120, or 180), pose learning rate (0.01, 0.03, or 0.06), temporal-window overlap (24, 48, or 72 frames), and the use and duration of shared hand-shape (β) optimization. All loss weights were held fixed.

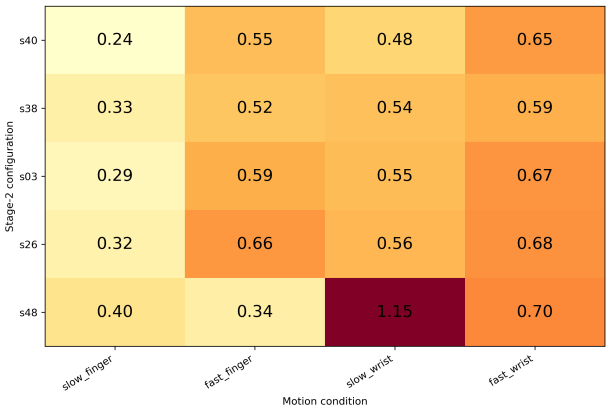


Figure 3. Mean absolute error (MAE) in dominant frequency estimation for the five selected STRIDE hyperparameter configurations, evaluated across slow- and fast-finger and wrist-motion conditions on the development set. Each cell reports MAE in Hz, with lighter colours indicating lower errors and darker colours indicating higher errors. Configuration s40 achieves the lowest condition-averaged MAE of approximately 0.48 Hz.

The five most promising configurations were subsequently evaluated on the complete development set of 96 clips, using participant-balanced frequency MAE as the selection criterion. Additional configuration details are provided in Appendix B.

Figure 3 demonstrates configuration `s40` achieving the lowest mean error across the four motion conditions, at 0.48 Hz. It performed particularly well for slow-finger and slow-wrist motion, achieving errors of 0.24 Hz and 0.48 Hz, respectively, while remaining competitive for faster motions. We selected `s40` based on its lowest aggregate MAE across all four movement conditions.

However, the results also reveal differences in performance across motion conditions. For example, configuration `s48` achieved a low error of 0.34 Hz for fast-finger motion, but its error increased substantially to 1.15 Hz for slow-wrist motion. This indicates that configurations performing well on one type of movement do not necessarily generalize to other movement conditions.

These differences suggest that the refinement hyperparameters influence the reconstruction of periodic hand motion differently depending on its frequency and anatomical origin. In particular, optimizing temporal consistency does not necessarily produce equally accurate frequency estimates across finger and wrist movements.

An important consideration is that the prescribed slow and fast frequencies differ by only 1 Hz (4 Hz and 5 Hz, respectively). A baseline randomly predicting either fre-

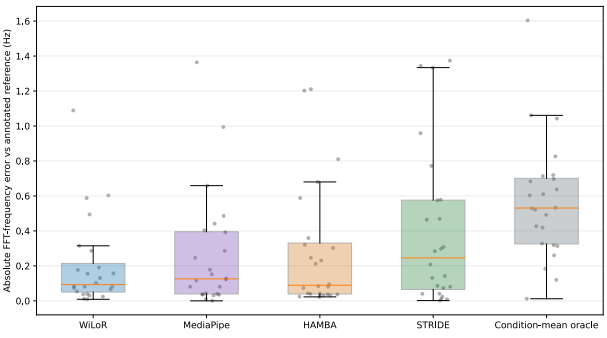


Figure 4. Comparison of absolute frequency-estimation error on the held-out participant for four hand pose methods and a condition-mean baseline. Grey dots indicate individual recordings ($n = 24$ per method), boxes show the interquartile range, orange lines mark the median, and whiskers extend to observations within 1.5 times the interquartile range. Values reported in the text are interquartile means (IQMs), calculated from the middle 50% of errors. WiLoR achieves the lowest IQM error (0.12 Hz), followed by HAMBA (0.15 Hz), MediaPipe (0.16 Hz), and STRIDE (0.27 Hz), while the baseline achieves 0.55 Hz. The results show that accurate frequency estimation is feasible under controlled conditions. Temporal refinement with STRIDE does not improve accuracy over the original WiLoR estimates.

quency with equal probability would therefore achieve an expected MAE of 0.5 Hz. While `s40` performs better than this baseline for slow-finger motion (0.24 Hz), its slow-wrist error (0.48 Hz) is only marginally lower, and its performance under fast-motion conditions does not demonstrate a similarly consistent improvement. Therefore, `s40` was selected because it achieved the lowest aggregate MAE (0.48 Hz) among the evaluated configurations on the development set. However, its performance varied across movement conditions, and the results do not establish that it consistently distinguishes between 4 and 5 Hz oscillations. The selection therefore represents the best overall configuration within the evaluated search space, rather than a universally optimal refinement strategy.

4.3. Controlled Frequency Estimation

Experimental question: *How accurately can different hand-pose estimation models recover the frequency of tremors under controlled conditions?*

We first evaluate how the accuracy reconstructed hand motion preserves the dominant frequency of tremor under controlled conditions. The evaluation uses 24 recordings from a held-out participant, covering finger- and wrist-emulated tremors at frequencies of 4 and 5 Hz across multiple camera viewpoints. We compare MediaPipe, HAMBA, WiLoR, and STRIDE, measuring the absolute difference between the estimated dominant frequency and the reference frequency. We additionally include a condition-mean baseline.

As shown in Figure 4, WiLoR achieves the lowest interquartile mean (IQM) absolute error of 0.12 Hz, followed by HAMBA (0.15 Hz) and MediaPipe (0.16 Hz). STRIDE produces a higher IQM error of 0.27 Hz, despite having refined WiLoR trajectories to improve temporal consistency. Its error distribution is also larger, suggesting that refinement does not consistently improve frequency estimation.

This difference highlights an issue with temporal refinement. STRIDE encourages temporally plausible and consistent hand motion (through HMP). However, genuine tremor also consists of rapid temporal variations, and the refinement process may suppress these movements with the noise. A smoother trajectory can therefore exhibit low temporal noise while retaining an incorrect dominant frequency. The controlled results are consistent with this trade-off, although they do not indicate whether excessive smoothing is the cause of every estimation error.

The condition-mean baseline (which always predicts the average reference frequency observed among the four development participants for each movement condition) achieves an IQM error of approximately 0.55 Hz, exceeding the errors of all four evaluated methods. This suggests that the estimated frequencies reflect differences in how individual participants performed the movements. WiLoR's IQM error of approximately 0.12 Hz suggests that accurate frequency estimation is feasible under controlled recording conditions. However, the evaluation consists of only 24 recordings from a single held-out participant, while the reference frequencies are derived from annotations rather than synchronized instrument-based measurements. Furthermore, the IQM does not characterize the largest estimation failures. Therefore, these results show the feasibility of the pipeline under the tested conditions, but do not establish accuracy across different participants, a broader range of tremor frequencies, or clinical recordings. Temporal refinement also does not necessarily improve downstream frequency accuracy.

4.4. Reliability Measures under Controlled Motion

Experimental question: *Which of the proposed motion and signal-quality metrics are associated with frequency-estimation error when reference measurements are available?*

As shown in Figure 5, FFT peak ratio exhibits the strongest and most consistent relationship with frequency-estimation error. Negative correlations are observed for all four methods, with WiLoR ($\rho = -0.67$) and STRIDE ($\rho = -0.62$) showing the strongest associations, followed by MediaPipe and HAMBA (both $\rho = -0.40$). This suggests that frequency estimates obtained from signals with a more prominent spectral peak tend to have lower errors. However, a prominent peak does not independently estab-

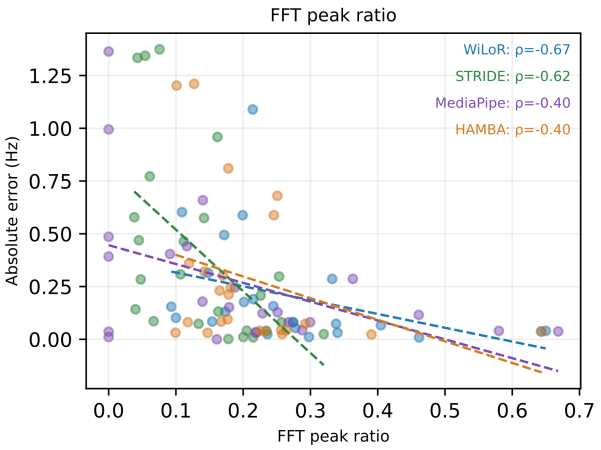


Figure 5. Relationship between peak ratio and absolute frequency-estimation error for the held-out participant. Each point represents a recording, with colours distinguishing the reconstruction methods. Dashed lines indicate model-specific linear trends, and the reported Spearman correlation coefficients quantify the monotonic relationships. All four methods exhibit negative correlations, strongest for WiLoR (ρ = -0.67) and STRIDE (ρ = -0.62). A more concentrated spectral peak is therefore associated with lower estimation error.

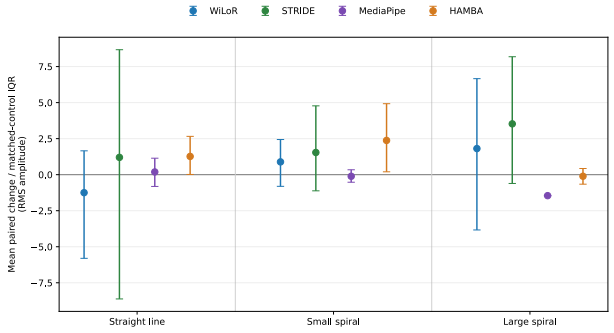


Figure 6. Effect of emulated tremor on reconstructed RMS amplitude during drawing tasks. The horizontal axis distinguishes straight-line, small-spiral, and large-spiral drawing tasks. Coloured markers show the mean paired change in RMS amplitude between emulated-tremor and matched control recordings for each reconstruction method, normalized by the matched-control IQR. Vertical error bars indicate standard deviation. Positive values indicate increased reconstructed amplitude under emulated tremor. The effect varies across tasks and methods: WiLoR shows reduced amplitude for straight-line drawing but increased amplitude for both spiral tasks. This implies that introduced tremor-like motion is not consistently reflected in stronger reconstructed oscillations.

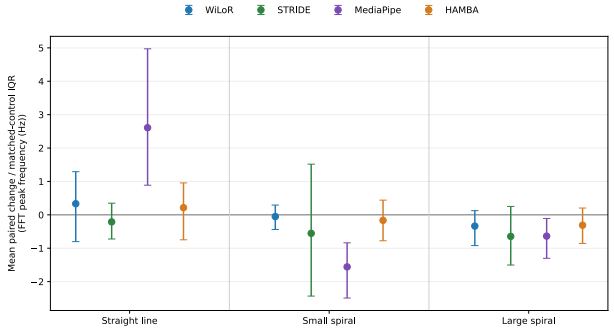


Figure 7. Effect of emulated tremor on estimated dominant frequency during drawing tasks. The horizontal axis distinguishes straight-line, small-spiral, and large-spiral drawing tasks. Coloured markers show the mean paired change in RMS amplitude between emulated-tremor and matched control recordings for each reconstruction method, normalized by the matched-control IQR. Vertical error bars indicate standard deviation. All four methods exhibit decreases in mean dominant frequency for the spiral tasks, whereas straight-line drawing produces model-dependent changes. These differences demonstrate that introducing tremor-like motion alters the estimated dominant frequency, but cannot establish estimation accuracy without frequency ground truth.

lish that the estimated frequency is correct.

The remaining metrics (Figure 15)show weaker and less consistent relationships across reconstruction methods. For WiLoR, frequency-stability standard deviation is positively correlated with error ($\rho = 0.39$), suggesting that less stable frequency estimates tend to be less accurate. Palm-orientation variance shows a negative correlation ($\rho = -0.39$). STRIDE exhibits a different pattern, with mean palm orientation ($\rho = 0.35$) and RMS amplitude ($\rho = 0.28$) showing positive correlations, while frequency stability has almost no relationship with error ($\rho = -0.05$). These differences indicate that the characteristics associated with frequency-estimation error depend partly on the reconstruction method.

We therefore consider peak ratio a potentially useful indicator of frequency-estimation reliability, while retaining the other metrics to characterize reconstruction behaviour and recording conditions in subsequent experiments. Although the correlation of peak ratio with error does not establish that it can independently predict estimation accuracy.

4.5. Drawing Tasks Without Frequency Ground Truth

Experimental question: *How does introducing tremor-like motion during drawing affect the reconstructed motion and its frequency characteristics when ground truth is unavailable?*

We next move from isolated periodic movements to tasks that more closely resemble the activities performed in the clinical recordings. Participants draw a straight line, a large spiral, and a small spiral under both a control condition and a condition in which tremor-like motion is intention-

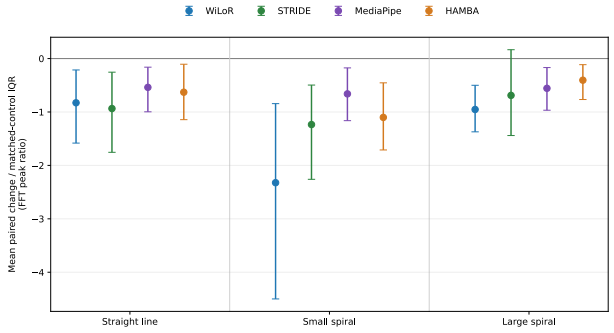


Figure 8. Effect of emulated tremor on peak ratio during drawing tasks. The horizontal axis distinguishes straight-line, small-spiral, and large-spiral drawing tasks. Coloured markers show the mean paired change in RMS amplitude between emulated-tremor and matched control recordings for each reconstruction method, normalized by the matched-control IQR. Vertical error bars indicate standard deviation. All four methods exhibit negative mean changes across the three drawing tasks, suggesting that introducing tremor-like movement distributes spectral power across a broader range of frequencies rather than producing a more clearly defined dominant peak.

ally introduced. Since the emulated tremor does not have a sufficiently precise reference frequency, direct frequency-error evaluation is not possible. We therefore compare each tremor recording with its matched control recording from the same task and viewpoint, using the held-out participant.

Figure 6 shows that the introduction of tremor-like motion does not consistently increase the RMS amplitude of the reconstructed hand-displacement signal. For WiLoR, the mean paired amplitude decreases by 0.00120 for the line task, but increases by 0.00125 and 0.00121 for the large and small spirals, respectively. STRIDE shows an increase in RMS amplitude across all three tasks, whereas MediaPipe shows decreases for both spiral tasks. These differences indicate that the response of the reconstructed motion to the introduced tremor depends on both the activity and the reconstruction method.

The dominant-frequency estimates show similarly varied behaviour, as illustrated in Figure 7. For both spiral tasks, all four methods show a decrease in mean dominant frequency between the control and tremor conditions. On the other hand, the straight-line task produces differing results based on the model. Since the actual frequency of the emulated tremor is unknown, these changes cannot be interpreted as improvements or deteriorations in estimation accuracy.

The peak ratio (Figure 8) has a more consistent pattern. The mean peak ratio always decreases when tremor-like motion is introduced, suggesting that the additional movement results in a less concentrated frequency spectrum, rather than introducing a single clearly dominant periodic

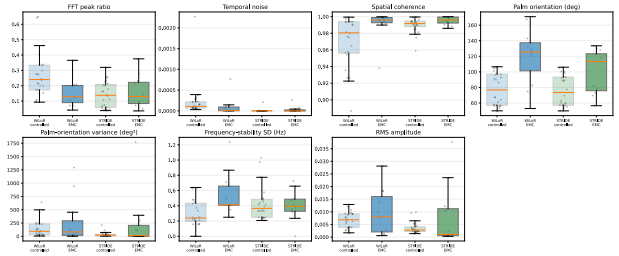


Figure 9. Comparison of reconstructed motion characteristics between controlled and clinical EMC recordings for WiLoR and STRIDE. Each panel shows the distribution of a different signal or reconstruction metric, with individual points representing recordings. The boxplots illustrate differences in metric distributions between the two datasets, allowing investigation of changes in spectral concentration, temporal stability, hand orientation, and reconstructed motion amplitude.

component. It should be noted that this decrease can occur even when RMS amplitude increases, meaning that stronger reconstructed oscillations might not always correspond to a clearer spectral peak.

Overall, introducing tremor-like motion during drawing produces changes in the estimated frequency which are model-dependent, while consistently reducing spectral peak concentration. However, changes in dominant frequency do not necessarily indicate changes in accuracy, as the observed peak may not be indicating the tremor frequency. Similarly, a lower peak ratio indicates that spectral energy is more broadly distributed, rather than an inherently less accurate frequency estimate. Although peak ratio was associated with estimation error in the controlled experiments, this relationship cannot be assumed to generalize. Currently, these metrics characterize changes in reconstructed motion but cannot establish whether tremor-frequency estimation becomes more or less accurate.

4.6. Clinical Video Evaluation

Experimental question: *How do reconstructed motion characteristics differ between controlled and clinical recordings, and can the proposed metrics identify unusual reconstruction behaviour?*

Figure 9 compares the distributions of the proposed metrics between the controlled and clinical recordings for WiLoR and STRIDE. Several differences are observed between the two datasets. For WiLoR, the clinical recordings generally exhibit lower FFT peak ratios and higher frequency-stability standard deviations, suggesting less clearly defined and less temporally stable dominant frequencies. STRIDE shows smaller changes in these two measures. Both methods exhibit larger palm-camera orientation angles in the clinical recordings, indicating differences in the hand orientations encountered across the

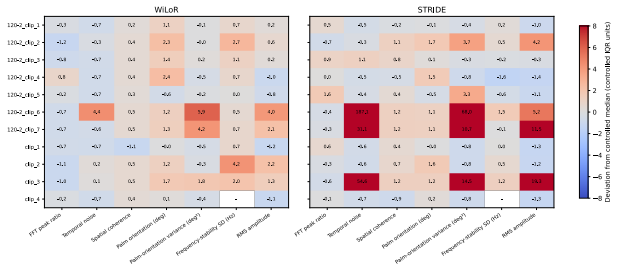


Figure 10. Per-clip metrics for the EMC recordings using WiLoR and STRIDE. Each cell represents the deviation of a clinical recording's metric from the corresponding controlled median, normalized by the controlled interquartile range (IQR). These values describe differences in reconstructed motion characteristics. Several clinical clips exhibit substantial deviations, particularly under STRIDE, demonstrating that reconstruction behaviour varies considerably between recordings.

datasets. The clinical recordings also show greater variability in several metrics, particularly RMS amplitude and palm-orientation variance.

Several recording characteristics may contribute to these differences. Unlike the isolated movements in the controlled dataset, clinical recordings contain more complex activities, including drawing and pouring, with voluntary motion occurring alongside potential tremor. The larger palm-camera angles also indicate differences in viewing geometry between the datasets. Such conditions may complicate hand reconstruction and introduce additional motion components into the estimated signal. However, the individual effects of viewpoint, occlusion, illumination, camera movement, and activity type were not independently isolated. Their contribution to the observed metric differences therefore remains a possible explanation rather than an established causal relationship.

Figure 10 shows the deviation of individual clinical recordings from the controlled reference distributions, normalized by the controlled interquartile range. Substantial differences are observed between recordings, particularly for STRIDE. Clips such as `120-2_clip_6`, `120-2_clip_7`, and `clip_3` exhibit large deviations in temporal noise, palm-orientation variance, and RMS amplitude. In contrast, FFT peak ratio remains relatively close to the controlled reference values across most recordings. These results indicate that clinical recordings can exhibit unusual reconstruction characteristics without corresponding large changes in spectral peak concentration.

Clinical recordings show greater variation in reconstructed motion than controlled recordings. While the proposed metrics identify unusual motion characteristics, the lack of ground truth prevents their validation as indicators of frequency-estimation accuracy.

5. Conclusion

In this work, we investigated the reliability of tremor-frequency estimation from monocular RGB videos using off-the-shelf hand pose estimation and reconstruction methods. We developed a model-agnostic frequency-analysis pipeline and evaluated it across controlled periodic movements, drawing tasks, and clinical recordings of CIDP patients.

Under controlled conditions, WiLoR achieved the lowest frequency-estimation error on the held-out participant, demonstrating that modern hand reconstruction methods can recover frequencies close to the reference measurements. However, STRIDE's temporal refinement did not improve frequency accuracy, demonstrating that smoother reconstructed motion does not necessarily preserve the underlying periodic signal. Among the proposed reliability measures, FFT peak ratio showed the strongest and most consistent negative correlation with frequency-estimation error, making it a promising indicator of estimate reliability.

In the drawing experiments, introducing tremor-like motion produced inconsistent changes in amplitude and dominant frequency, while reducing spectral peak concentration across all evaluated methods. The clinical recordings then revealed large differences in metrics compared to the controlled dataset. Although the proposed metrics can identify unusual reconstructions, their relationship with frequency-estimation accuracy could not be verified in the absence of clinical ground truth.

Overall, our findings suggest that monocular video-based tremor-frequency estimation is feasible under controlled conditions, but its reliability in less constrained clinical recordings remains uncertain. The results highlight the importance of evaluating hand reconstruction methods based on their preservation of relevant temporal information, rather than spatial accuracy or temporal smoothness alone.

6. Limitation

The primary limitation of this study is the availability of suitable evaluation data. The clinical dataset is limited in size and lacks ground-truth measurements, preventing direct evaluation of frequency-estimation accuracy. The controlled dataset contains five healthy participants emulating periodic movements at 4 and 5 Hz, with the final evaluation performed on a single held-out participant. These movements do not necessarily reproduce the frequency variability, amplitude, and irregularity of pathological tremors. Furthermore, the participants rely on prescribed metronome frequencies that they may not reproduce precisely.

A second limitation arises from the frequency-analysis methodology itself. The proposed pipeline measures the relative displacement between selected anatomical regions,

making it suitable for localized movements but potentially suppressing tremors involving approximately rigid motion of the entire hand, as they would require reference points that MANO cannot model (i.e. base of the wrist). Furthermore, the pipeline cannot necessarily distinguish genuine tremor from periodic reconstruction artifacts, particularly in recordings without a clear tremor signal.

Finally, the evaluated hand reconstruction methods introduce additional uncertainties. Their accuracy may depend on camera viewpoints, occlusions, and differences between their training distributions. That being said, we have yet to independently isolate these factors. Similarly, the proposed reliability metrics attempted to characterize the reconstructed motion but do not independently establish measurement accuracy. Most of the metrics also didn't succeed in effectively characterizing the motion. While STRIDE demonstrates that temporal refinement can reduce frequency-estimation accuracy, this conclusion is limited to the evaluated HMP-based refinement procedure and hyperparameter configurations, rather than temporal refinement methods in general.

Clinical and Ethical Considerations Beyond technical accuracy, applying the proposed pipeline in clinical practice raises some concerns. Patient recordings, which may contain identifiable information, will require appropriate storage and access, in accordance with the GDPR. More importantly, incorrect estimates could mislead clinical assessments if interpreted without sufficient caution. The proposed pipeline should only be considered a potential tool for supporting clinical assessment. Clinicians could use the estimated frequencies and accompanying reliability measures alongside direct observations and conventional assessments, while treating uncertain estimates cautiously. Clinical deployment would require further validation against synchronized reference measurements, as well as input from clinicians and patients to establish appropriate use and acceptable levels of uncertainty.

7. Future Work

Future research should prioritize the development of larger, clinically validated datasets containing synchronized video and instrument-based tremor measurements. Such datasets would enable direct evaluation of frequency-estimation accuracy. The new dataset can also include recordings without tremor, alongside more diverse recording conditions and higher frame rates, which can help establish the conditions under which the proposed pipeline produces reliable estimates.

Further methodological improvements could address the limitations of the current hand-motion representation, by investigating alternative motion priors and tremor-aware temporal refinement methods, incorporating frequency-preserving objectives to distinguish reconstruction noise from genuine periodic movement.

Finally, the proposed reliability metrics provide a foundation for developing objective criteria for usable video recordings. Their predictive value could be evaluated using independently annotated clinical data, potentially allowing unreliable frequency estimates to be identified without ground truth during inference. A systematic analysis of camera viewpoints, hand orientations, occlusions, and differences between model training and clinical data distributions could further inform model adaptation and practical recording guidelines. Together, these developments could improve the reliability and accessibility of video-based tremor quantification outside controlled clinical environments.

References

[1] Merel C. Broers, Carina Bunschoten, Daan Nieboer, Hester F. Lingsma, and Bart C. Jacobs. Incidence and prevalence of chronic inflammatory demyelinating polyradiculoneuropathy: A systematic review and meta-analysis. *Neuroepidemiology*, 52(3-4):161–172, 2019. 1

[2] Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *Advances in Neural Information Processing Systems*, 37:2127–2160, 2024. 3

[3] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J. Black. HMP: Hand motion priors for pose and shape estimation from video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6353–6363, 2024. 3

[4] Rodger J. Elble and William Ondo. Tremor rating scales and laboratory tools for assessing tremor. *Journal of the Neurological Sciences*, 435:120202, 2022. 1

[5] Diego L. Guarín. Video-based quantification of hand postural tremor without external references. integrating postural tremor quantification into visionMD. *npj Parkinson's Disease*, 11:351, 2025. 2

[6] Gökhan Güney, Talisa S. Jansen, Sebastian Dill, Jörg B. Schulz, Manuel Dafotakis, Christoph Hoog Antink, and Anne K. Braczynski. Video-based hand movement analysis of parkinson patients before and after medication using high-frame-rate videos and MediaPipe. *Sensors*, 22(20): 7992, 2022. 2

[7] Rohit Lal, Saketh Bachu, Yash Garg, Arindam Dutta, Calvin-Khang Ta, Hannah Dela Cruz, Dripta S. Raychaudhuri, M. Salman Asif, and Amit K. Roy-Chowdhury. STRIDE: Single-video based temporally continuous occlusion-robust 3d pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 794–803, 2025. 3

[8] Zoltan Mari and Dietrich Haubenberger. Remote measurement and home monitoring of tremor. *Journal of the Neurological Sciences*, 435:120201, 2022. 1

[9] Silvia L. Pintea, Jian Zheng, Xilin Li, Paulina J. M. Bank, Jacobus J. van Hilten, and Jan C. van Gemert. Hand-tremor frequency estimation in videos. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018. 2

[10] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12242–12254, 2025. 3

[11] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017. 3

[12] Matthew Silsby, Alessandro F. Fois, Con Yiannikas, Karl Ng, Matthew C. Kiernan, Victor S. C. Fung, and Steve Vucic. Chronic inflammatory demyelinating polyradiculoneuropathy-associated tremor: Phenotype and pathogenesis. *European Journal of Neurology*, 30(4):1059–1068, 2023. 1

[13] Stefan Williams, Hui Fang, Samuel D. Relton, David C. Wong, Taimour Alam, and Jane E. Alty. Accuracy of smartphone video for contactless measurement of hand tremor frequency. *Movement Disorders Clinical Practice*, 8(1):69–75, 2021. 2

[14] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. MediaPipe Hands: On-device real-time hand tracking. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality (CV4ARVR)*, 2020. 2

[15] Zhuoran Zhao, Linlin Yang, Pengzhan Sun, Pan Hui, and Angela Yao. Analyzing the synthetic-to-real domain gap in 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12255–12265, 2025. 3

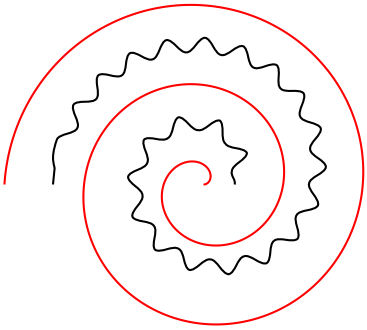


Figure 11. Tremor spiral figure. Volunteers were asked to stay within the red lines, and emulate a tremor following the black lines. A small (Diameter = 8cm) and large (Diameter = 10cm) version were used.

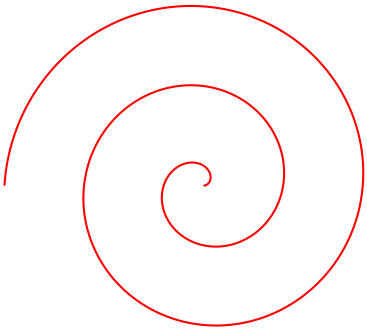


Figure 12. Spiral figure. Volunteers were asked to stay within the red lines, and draw a spiral. A small (Diameter = 8cm) and large (Diameter = 10cm) version were used.



Figure 13. Straight line figure. Volunteers were asked to stay within the red lines, and draw a straight line, which spanned the width of an A4 sheet (approximately 12 cm).

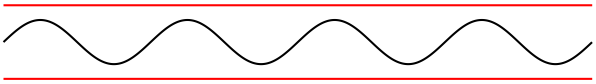


Figure 14. Wavy line figure. Volunteers were asked to stay within the red lines, and draw a wavy line, following the black line, which spanned the width of an A4 sheet (approximately 12 cm).

A. Emulation Figures

B. STRIDE Hyperparameter Tuning

We selected STRIDE refinement hyperparameters with a two-stage development-only search. Stage 1 screened 48 configurations on a stratified 18-clip subset. Five candidates (s03, s26, s38, s40, and s48) then underwent Stage 2 eval-

Table 1. Selected STRIDE refinement hyperparameters for configuration `s40`.

Parameter	`s40` value
Temporal chunk overlap	48 frames
Pose-refinement iterations	60
Pose learning rate	0.01
Shape (β) optimization	Enabled
Shape-refinement iterations	60
Shape learning rate	0.03
Pose rotation-loss weight	0.35
Pose position-loss weight	0.05
Pose root-loss weight	0.12
Pose latent-prior weight	0.05
Shape joint-loss weight	0.60
Shape-prior weight	0.40

Table 2. FFT-periodogram frequency MAE of configuration `s40` by motion condition.

Condition	MAE (Hz)
Slow finger	0.237
Fast finger	0.547
Slow wrist	0.477
Fast wrist	0.652

uation on all 96 controlled clips. The locked 24-clip split was not used for configuration selection.

For every clip, the predicted frequency was obtained from the peak of the FFT periodogram, and error was computed as the absolute difference from the annotated reference frequency. Ties were resolved by worst-condition MAE, worst-viewpoint MAE, frequency-stability SD, and configuration identifier. Configuration `s40` achieved the lowest participant-balanced FFT-frequency MAE, 0.478 Hz (95% CI half-width across participant means: 0.203 Hz). Its worst motion-condition MAE was 0.652 Hz and its worst viewpoint MAE was 0.737 Hz. The next-best candidate, `s38`, achieved a participant-balanced MAE of 0.495 Hz. Thus, `s40` was used for all downstream STRIDE experiments. This ranking is a deterministic development-set selection procedure rather than a formal significance comparison between configurations.

B.0.1. Selected configuration: `s40`

The 48-frame overlap determines how much adjacent temporal refinement chunks share. Pose refinement optimizes the per-frame hand pose, whereas β-optimization refines the hand-shape parameters shared across the sequence.

For context, Table 2 reports the FFT-periodogram MAE of `s40` by motion condition.

Its highest average error occurred for the behind-camera (BE) viewpoint, at 0.737 Hz.

C. Metrics in Detail

Dominant frequency alone does not indicate whether a frequency estimate is reliable. A dominant spectral component is always returned when a search is restricted to a frequency interval, even when the underlying reconstruction is noisy or does not contain a clear periodic signal. We therefore calculate additional measures intended to characterize different potential failure modes of the reconstructed hand motion.

Let x_t denote the mean-centered scalar motion signal obtained using the procedure described in Section 3.3, $P(f)$ its estimated spectral power, and $\mathcal{F}$ the set of discrete frequency bins within the 4–12 Hz tremor band. The estimated dominant frequency is denoted by $\hat{f}$.

Alternative Hand Representations The metrics are computed from two different hand representations. WiLoR, HAMBA, and STRIDE produce MANO-based hand meshes. For these methods, the relative displacement trajectory $\mathbf{d}_t$ is calculated using the centroids of selected local vertex neighborhoods. MediaPipe instead produces 21 anatomical hand landmarks. For MediaPipe, the trajectory is constructed using anatomically corresponding landmarks, without averaging over mesh neighborhoods.

The frequency-estimation pipeline applies a third-order Butterworth low-pass and high-pass filter with a cutoff frequency of 4 Hz and 12 Hz to each coordinate of the relative 3D hand trajectory. The filtered trajectory is then converted into a scalar magnitude signal for spectral analysis. The 4–12 Hz interval defines the frequency-search band. This distinction is important when interpreting estimates near the upper end of the search band. Some reliability metrics, as specified below, are instead calculated directly from unfiltered reconstructed motion.

Tremor Amplitude Tremor amplitude characterizes the magnitude of oscillatory motion present in the reconstructed hand. Given the extracted motion signal x_t, we apply an additional band-pass filter within the 4–12 Hz tremor band to obtain $\tilde{x}_t$. The root-mean-square (RMS) amplitude is then calculated as

$$A_{\mathrm{RMS}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \tilde{x}_t^2}. \tag{4}$$

A higher RMS amplitude indicates stronger oscillatory motion within the selected frequency band, while a lower value indicates weaker oscillations. However, this measure does not independently indicate reconstruction quality, as the measured amplitude may contain both genuine tremor

and reconstruction noise. Furthermore, temporal smoothing may reduce RMS amplitude by suppressing genuine tremor motion. Since reconstructed coordinates are not necessarily calibrated to physical units, RMS amplitude should not be interpreted as an absolute physiological measurement. We therefore use this metric to compare the preservation of oscillatory motion across reconstruction methods, rather than treating higher or lower amplitude as inherently desirable.

Peak Ratio The peak ratio describes how strongly the dominant frequency stands out from the remaining spectral content within the tremor band. It is calculated as the spectral power at the dominant frequency relative to the total power within the band:

$$R_{\text{peak}} = \frac{P(\hat{f})}{\sum_{f \in \mathcal{F}} P(f)}. \tag{5}$$

A high peak ratio indicates that the estimated spectrum is dominated by a relatively distinct periodic component. A low value indicates that spectral energy is distributed across several frequencies and that the selected dominant frequency may therefore be less informative. However, a concentrated spectral peak may also result from periodic reconstruction artifacts and does not independently establish the accuracy of the estimated frequency.

Temporal Noise Temporal noise characterizes frame-to-frame variability in the reconstructed hand motion. Unlike the frequency-estimation procedure, this metric is calculated directly from the unfiltered relative 3D displacement trajectory $\mathbf{d}_t$ between the selected anatomical regions.

The frame-to-frame displacement magnitude is defined as

$$v_t = \|\mathbf{d}_{t+1} - \mathbf{d}_t\|_2, \qquad t = 1, \ldots, T-1. \tag{6}$$

Temporal noise is then calculated as the variance of this displacement magnitude over time:

$$N_{\text{temp}} = \text{Var}_{t=1,\ldots,T-1}(v_t). \tag{7}$$

This measure characterizes the variability of frame-to-frame movement and is intended primarily to capture reconstruction jitter. However, genuine hand movements also contribute to this variability. Low temporal noise is therefore not automatically desirable for tremor analysis, since excessive temporal smoothing can remove genuine oscillatory motion. Temporal noise is consequently analyzed in relation to frequency accuracy rather than treated as an independent quality score.

Spatial Coherence Spatial coherence characterizes how consistently neighboring vertices of the reconstructed hand move over time. Unlike temporal noise, which operates on the relative trajectory between two anatomical regions, spatial coherence considers the motion of individual mesh vertices.

For a selected vertex i, with reconstructed 3D position $\mathbf{x}_{i,t}$ at frame t, the frame-to-frame displacement magnitude is defined as

$$m_{i,t} = \|\mathbf{x}_{i,t+1} - \mathbf{x}_{i,t}\|_2, \qquad t = 1, \ldots, T-1. \tag{8}$$

For two spatially adjacent vertices i and j, their motion similarity is measured using the Pearson correlation coefficient:

$$\rho_{ij} = \text{corr}\,(m_i, m_j), \tag{9}$$

where m_i and m_j denote the corresponding temporal sequences of displacement magnitudes.

Let $\mathcal{E}$ denote the set of neighboring vertex pairs for which a valid correlation can be calculated. The overall spatial coherence is defined as the mean valid pairwise correlation:

$$C_{\text{spatial}} = \frac{1}{|\mathcal{E}|} \sum_{(i,j) \in \mathcal{E}} \rho_{ij}. \tag{10}$$

Low spatial coherence may indicate independently varying or noisy vertex trajectories, while high coherence indicates that neighboring vertices exhibit similar motion. However, highly coherent motion may arise from either genuine coordinated hand movement or excessive temporal regularization. Spatial coherence therefore has no assumed optimal value and cannot independently establish reconstruction accuracy.

This metric is calculated only for mesh-based reconstruction methods. It is unavailable for MediaPipe in the current implementation, as the corresponding mesh-neighborhood relationships are not provided.

Palm Orientation Palm orientation describes the orientation of the reconstructed palm relative to the camera. Let $\mathbf{n}_t$ denote the palm normal at frame t and $\mathbf{c}$ the camera viewing direction. The palm-camera angle is calculated as

$$\theta_t = \cos^{-1}\left(\frac{\mathbf{n}_t^\top \mathbf{c}}{\|\mathbf{n}_t\|_2 \|\mathbf{c}\|_2}\right). \tag{11}$$

We characterize each recording using both the mean palm orientation,

$$\bar{\theta} = \frac{1}{T} \sum_{t=1}^{T} \theta_t, \tag{12}$$

and its variance over time,

$$V_\theta = \frac{1}{T}\sum_{t=1}^{T}(\theta_t - \bar{\theta})^2. \tag{13}$$

The mean angle describes the typical viewing geometry, while its variance describes how much the relative orientation changes during the recording. These changes may arise from hand movement, camera movement, or both.

Palm orientation is an explanatory recording characteristic rather than a reliability score. It is included to investigate whether particular viewpoints make the hand pose, and consequently the tremor signal, more difficult to reconstruct. This measure is particularly relevant when comparing the controlled recordings, in which viewpoint is deliberately varied, with the less constrained EMC recordings.

Frequency Stability Frequency stability measures whether the estimated dominant frequency remains consistent across shorter temporal windows from the same recording. The signal is divided into overlapping two-second windows, with the dominant frequency estimated independently for each window. Let $\hat{f}_k$ denote the dominant-frequency estimate obtained from window k, for a total of K windows.

The mean window-level frequency is calculated as

$$\bar{f} = \frac{1}{K}\sum_{k=1}^{K}\hat{f}_k, \tag{14}$$

and frequency instability is measured using the standard deviation

$$S_f = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{f}_k - \bar{f})^2}. \tag{15}$$

Because the analyzed clips contain approximately homogeneous activities, a reliable periodic signal is expected to produce similar frequency estimates across windows. Lower values of S_f therefore indicate greater temporal consistency in the estimated dominant frequency. However, a consistently incorrect frequency or a stable periodic reconstruction artifact may also produce a low standard deviation. Frequency stability consequently measures the consistency of frequency estimates rather than their accuracy.

D. Metric results

Figure 15 Shows more detailed metrics with regards to motion reconstruction.

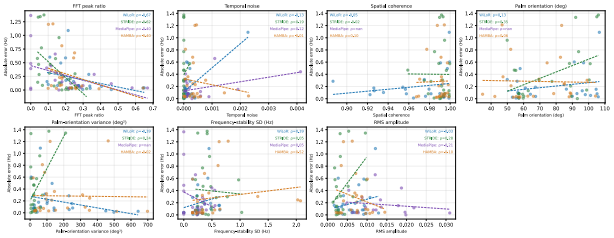


Figure 15. Relationship between reconstructed motion characteristics and absolute frequency-estimation error for the held-out participant. Each point represents a recording, with colours distinguishing the reconstruction methods. Dashed lines indicate linear trends, while the reported Spearman correlation coefficients quantify the monotonic relationships between each metric and frequency error. Beyond the consistently negative relationship observed for FFT peak ratio, the remaining metrics exhibit weaker or method-dependent associations. These findings suggest that reconstruction characteristics cannot be interpreted as universal indicators of frequency-estimation accuracy.

A

Declaration of AI Usage

Generative AI tools, primarily ChatGPT and Codex by OpenAI, were used in a limited supporting capacity during this thesis project. Their use did not replace the research, analysis, or interpretation carried out by the author.

For programming-related tasks, generative AI was occasionally used to assist with boilerplate code, particularly for generating plotting and visualization scripts and for writing or modifying shell scripts used to run experiments. All generated code was reviewed, adapted where necessary, and tested by the author before use. Generative AI was not used to develop the core methodological contribution of this thesis or to interpret the experimental results.

For the written thesis, generative AI was used mostly as a structural aid, and occasionally for developing an academic tone in the writing. The arguments, interpretation and discussions were produced by the author.

The author retains full responsibility for the accuracy, originality, and integrity of all code, analyses, results, interpretations, references, and written content presented in this thesis. Any suggestions provided by generative AI were independently assessed and verified before being used. No sensitive, confidential, or personal research data were provided to generative AI tools.