

Meaningful and ergonomic hand poses for extended reality interaction

Received: 18 November 2025

Accepted: 15 July 2026

Published online: 08 August 2026

Cite this article as: Emporio M., Fattahi H., Caputo A. *et al.* Meaningful and ergonomic hand poses for extended reality interaction. *Virtual Reality* (2026). <https://doi.org/10.1007/s10055-026-01445-9>

Marco Emporio, Hassan Fattahi, Ariel Caputo & Andrea Giachetti

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

Meaningful and Ergonomic Hand Poses for Extended Reality Interaction

Marco Emporio¹, Hassan Fattahi¹, Ariel Caputo¹,
Andrea Giachetti^{1*}

^{1*}DIMI, Università di Verona, Strada Le Grazie 15, Verona, 37134, Italy.

*Corresponding author(s). E-mail(s): andrea.giachetti@univr.it;
Contributing authors: marco.emporio@univr.it;
hassan.fattahi@studenti.univr.it; ariel.caputo@univr.it;

Abstract

In this paper, we investigate ergonomic and technical factors limiting the effectiveness of gestural interfaces for Extended Reality based on hand pose recognition. For this purpose, we recorded and analyzed a novel dataset including hand poses known to be associated with semantics, performed by heterogeneous subjects wearing a popular Virtual Reality headset (Meta Quest 3) with finger-tracking capabilities. Unlike most related literature, which focuses on testing classifiers for recognizing static poses or dynamic gestures, we used the collected data to analyze these factors. We analyzed the variability of gesture execution and, through human visual labeling of renderings of the captured skeleton, verified the correspondence between the intended gesture labels and the recorded finger articulations. We investigated the effects of hand position, rotation, and skin color on the quality of the recorded poses. We evaluated the discriminability of the executions in feature spaces describing the hand poses, considering both the intended gesture labels and the visually assigned ones. We also collected information on semantic associations, familiarity with the gestures, and ergonomic issues through specific questionnaires.

The outcomes of our study show that many poses commonly used to control interfaces and strongly associated with semantic priors are not optimal, as they are not always well executed or robustly tracked. The gesture labels corresponding to participants' intentions do not always match those assigned during visual annotation, and poses are inadvertently modified during hand rotation. Our results provide valuable guidelines for designing effective gesture-controlled XR interfaces, including optimal dictionaries and system design choices.

Keywords: eXtended Reality, hand gestures, labeling, dictionary

1 Introduction

Hand pose estimation can be implemented using low-cost technologies and has been widely adopted across several domains. In ergonomics and workplace analysis, it enables the assessment of hand movements and the detection of potentially harmful patterns. In healthcare and rehabilitation, it supports quantitative evaluation of motor function and provides real-time feedback during therapy. In human-computer interaction, it enables intuitive, touchless command input, as exemplified by in-car gesture-control systems. Hand gesture-based control is also expected to become the standard interaction modality in extended reality (XR), where most modern headsets incorporate integrated finger-tracking systems.

However, the currently available solutions for controlling applications with gestures are limited (Yeh et al. (2022)), and the user experience is not always optimal (Graichen and Graichen (2023)). Even focusing on the simplest category of hand gestures used for XR control, "static" meaningful hand poses (Emporio et al. (2025)), it is challenging to find effective interaction solutions that exploit a rich gesture dictionary to issue different system commands.

In this work, we investigate the reasons for the development of gesture controls for XR based on rich hand pose dictionaries. Based on evidence from the literature and the XR development community, we identified potential causes of interaction problems and formulated a few specific research questions related to them. To find answers to these questions, we created a novel dataset comprising hand-skeleton data captured with a Meta Quest 3 headset while executing commonly used, meaningful hand poses.

It is essential to note that our analysis focuses on hand poses/shapes, independent of orientation, and does not consider dynamic gestures. This choice is intentional, as our goal is to understand the specific issues related to the execution, tracking, and recognition of the hand poses. Still, all the "gesture" recognition results are obviously limited to gestures where the semantics depend only on hand articulation (Emporio et al. (2025)). This work can be considered a first step in a systematic analysis of the ergonomic and technological issues related to gestural interaction that we plan to extend in future work to dynamic gestures.

The paper is structured as follows: Section 2 discusses the related literature on the ergonomic aspects of mid-air gestures, finger tracking, and the design of gesture vocabularies. Section 3 presents our goals and research questions, introduces the novel dataset (hand pose selection, acquisition protocol, and technical setup), and describes the methods used for its analysis (data preprocessing, annotation, and classification).

Section 4 summarizes the experimental outcomes, while Section 5 discusses the analysis and attempts to answer our research questions.

2 Related Work

The literature on mid-air gestural interfaces includes several papers analyzing the ergonomic aspects of the interaction. These studies investigate how specific hand postures, movements, and tracking conditions affect comfort, usability, and long-term feasibility. Rempel et al. (2014) exploited professional sign language interpreters to assess the discomfort caused by mid-air gestures designed for interaction. They

found clear associations between discomfort and hand postures that required a flexed wrist, discordant adjacent fingers, or extended fingers. [Chen et al. \(2016\)](#) analyzed the ergonomic issues of gestures involving hand articulation and motion. The main finding of their experiments is that hand dorsiflexion caused by the posture of opening the palm and facing forward is a relevant ergonomic issue. [Tang and Webb \(2018\)](#) evaluated a contactless gesture control system using the Rapid Upper Limb Assessment (RULA) human factor analysis tool ([McAtamney and Corlett \(1993\)](#)). The ergonomic analysis, however, did not consider hand poses. [Chen and Hu \(2025\)](#) evaluated the usability of a set of static and dynamic gestures, considering cognitive (discoverability, learnability) and electromyographic (EMG) factors, to identify an optimal set suitable for XR interaction among older adults.

Other works closely related to our research investigate the quality and limitations of finger tracking in headsets. These studies evaluate the accuracy, precision, and robustness of commodity tracking systems, often by comparing them with high-precision motion capture systems or by examining their performance across various interaction scenarios. [Schneider et al. \(2020\)](#) evaluated two commodity hand and finger tracking systems (HTC Vive and Leap Motion), finding non-negligible pointing position errors (Z-errors around 10mm in practice). [Soares et al. \(2021\)](#) performed accuracy and repeatability tests on HoloLens 2, finding position errors of about 20mm. [Bertolasini et al. \(2025\)](#) performed a fine-grained HoloLens 2 hand tracking evaluation for kinematic features assessment. They found that the tracker exhibits millimeter-level errors relative to Vicon's tracking system in a fingertip-tracking task and correlates well with the reference in pinching and grasping tasks.

[Borzelli et al. \(2025\)](#) quantitatively assessed the capability of tracking hand movements of the MetaQuest 2 tracker by comparing its kinematic measures with those provided by an OptiTrack motion capture system. They found that the Quest 2 offers reliable estimates of hand position and velocity, as well as fine-grained measures of grip aperture. Still, the precision of these measures might be affected by the subject's head movements.

[Nguyen et al. \(2023\)](#) surveyed the literature on finger-based interaction. A relevant challenge they found is the limited volume of interaction enabled by the devices, and the fact that the user's immersion is broken whenever their hands leave the hand-tracking coverage. The existing literature surveys on hand pose estimation are typically limited to the classification of the approaches used to annotate and classify data ([Ohkawa et al. \(2023\)](#); [Roumaissa and Mohamed Chaouki \(2025\)](#)), but do not address all the issues related to the quality of the data, recognizability of gestures, variability of gesture executions, and optimization of gesture vocabulary.

Regarding hand poses/gesture vocabularies, an interesting work has been presented by [Mukushev et al. \(2020\)](#), which uses an unsupervised approach to generate a hand-shape inventory. [Hanke \(2004\)](#) proposed a combinatorial approach to represent hand shapes that could ideally describe an extensive set of gestures, including those used in sign languages. [Xia et al. \(2022\)](#)'s work reports on an exhaustive literature review that identified 13 factors crucial to gesture vocabulary design. Some of them are out

of the scope of this paper. Still, it is essential to note that our analysis is consistent with their proposed approach, as it focuses on the specific physical and system factors relevant to the particular type of gestures and tracking system of interest. Other interesting works on gesture dictionaries focus more on developing tools for designers (Mo et al. (2021); Shen et al. (2022)) rather than analyzing the crucial factors for choosing vocabulary elements.

Other papers related to our research try to characterize gesture execution variability. Erazo et al. (2017) investigates this issue in the context of trajectory-based gestures. Al-Kalbani et al. (2017) presents an assessment of the variability in free-hand grasping of virtual objects in an exocentric mixed reality environment. A large dataset for analyzing hand pose variability has been proposed by Yuan et al. (2017); however, this dataset is limited to 2D data, and the space of variability in hand poses is not explored across different classes. Angrisani et al. (2024) present a study of the intra- and inter-individual variability in selection position and time in a point and pinch task performed with the HoloLens 2 headset.

While all the papers mentioned are related to our work and provide interesting insights into interaction design, none of them provide an analysis specifically focused on answering our research questions in the domain of hand-pose recognition for XR interaction. For this reason, we addressed this gap by capturing a novel dataset and performing a set of measurements and analyses specifically designed for our goals.

3 Materials and methods

3.1 Analysis of influencing factors

Although most research efforts related to gestural interaction are dedicated to the improvement of gesture classification algorithms, our hypothesis is that the difficulty in the development of rich and usable gesture-based interfaces for XR applications is not primarily due to this aspect, but may depend on several concurrent factors. Focusing on "static" gestures, where the goal is to recognize hand articulation only, and on the use of head-mounted displays for capturing gesture input, we can identify the following ones:

- The limited field of view of the sensor mounted on the headset does not always allow the sensor to capture the true pose of the hand due to occlusions and fingers extending beyond the tracking area (Parger et al. (2021)). The system's ability to recognize the gesture may critically depend on the gesture's pose, articulation, rotation, and execution location, and users are not necessarily aware of this limitation.
- Humans typically use hand poses (as well as dynamic gestures) to communicate; however, they are not consciously aware of precise execution rules associating meanings to specific hand joints' articulations (defined in terms of positions, angles, orientations), but, instead, learn the gestures by visual examples (Kendon (2004)). As a consequence, the way a gesture is executed can vary significantly across different subjects, even if they attempt to reproduce a target template with a known,

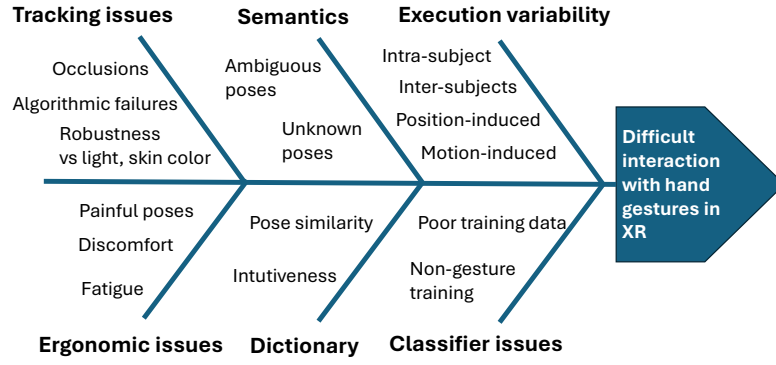


Fig. 1: Ishikawa diagram representing the hypothesized factors influencing the difficulty in the development of gesture-based interfaces for XR applications.

shared meaning. Some poses are not well executed by some subjects due to physical constraints.

- Gesture poses may not be comfortable and can lead to pain and fatigue, thereby affecting the execution and usability of the interface (Rempel et al. (2014)).
- Some gestures can be difficult to recognize for automatic classifiers due to excessive similarity with other ones and high variability in the execution. Classifiers, on the other hand, need to account for variability in gesture execution, which means that data-driven approaches can improve effectiveness. Still, in this case, the training data should represent not only the different execution styles of the poses but also differentiate meaningful poses from non-gesture examples Emporio et al. (2025).

We summarized the variety of these potential factors with an Ishikawa diagram (Ishikawa (1989); Salerno et al. (2026), see Figure 1).

In this work, we address these issues by creating and analyzing a novel dataset, SHAPER (Significant HAND Poses for interaction in Extended Reality). The dataset includes hand poses captured on different subjects wearing the Meta Quest 3 headset in different, controlled conditions, which have been subsequently analyzed by performing an expert-based labeling of the acquired poses and performing a careful quantitative analysis of the influence of different factors on the variability and recognisability of the gestures, also recording information about familiarity and associated semantics. Practically, the goal of our work, e.g., the investigation of the factors challenging effective hand-gesture interaction in XR applications, can be refined to a series of research questions:

- **RQ1** Tracking reliability: Do head-mounted displays reliably capture meaningful static hand poses across different conditions?
- **RQ2** Execution quality: How consistently do people perform these poses, and how do factors like orientation, position, or demographics affect execution?

- **RQ3** Pose space: Can we clearly map and separate different hand poses in a feature space, or do some overlap and cause confusion?
- **RQ4** Human vs. machine recognition: How does human recognition of tracked poses compare to automatic classification?
- **RQ5** Practical vocabulary design: Which poses are the most ergonomic, robust, and widely understood—and how can we combine them into effective gesture dictionaries?

Following the Goal/Question/Metric (GQM) approach (Caldiera and Rombach (1994)), we can link these questions to specific metrics that can be evaluated on our specifically designed SHAPER dataset, both through objective measurements of gesture characteristics and through subjects' questionnaire outcomes. In detail, RQ1 relates to the accuracy of visual classification of captured poses. RQ2 can be linked to a quantitative analysis of pose descriptor variability and to the questionnaire's outcomes. RQ3 can also be studied by analyzing estimated pose descriptors, for example, by examining the separability of the dataset's classes in the feature space. RQ4 can be answered by comparing the labels provided by humans and computer-based annotations of the same data. All measurements performed on the dataset are useful for designing optimal vocabularies for specific applications (RQ5). In particular, combining the questionnaire's outcomes with classification and reliability scores, it is possible to provide gesture rankings and guidelines.

3.2 The SHAPER dataset

The design of a dataset suitable for our analysis required the choice of the hardware and software solution for the experimental setup, the definition of a good dictionary of hand poses considered meaningful across various contexts, a protocol to capture gesture executions with different levels of independent variables (hand positions and orientations), together with subject demographics and questionnaires to collect ergonomic issues, and a protocol to evaluate gesture execution and tracking accuracy through expert visual labeling of performed gestures.

3.2.1 Hardware/software acquisition setup

We performed gesture acquisition with a Meta Quest 3 headset, using a specifically designed Unity Technologies (2026) application that manages a pre-programmed series of recordings for each participant and displays overlays to guide subjects in following the protocol properly. The choice of the Meta Quest 3 headset is related to the fact that it is probably the most popular headset currently used for VR and MR applications, thanks to its video passthrough capabilities. Still, the results can be generalized to all the VR/MR headsets using finger tracking based on integrated RGBD sensors, employing similar technologies and viewpoints. The configuration used in our test is actually the MR one, with the headset in passthrough mode.

3.2.2 Dictionary

The gesture dictionary (shown in Figure 2) was developed by selecting poses commonly employed in interpersonal communication and gesture-based interfaces. We assigned a

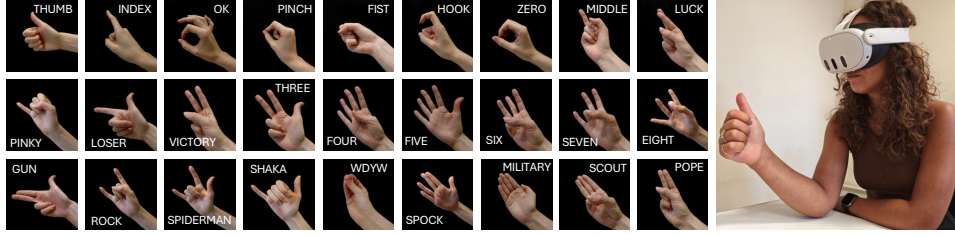


Fig. 2: The complete dictionary of hand poses potentially usable for gesture-based commands in extended reality considered in our study

reference name (superimposed to the corresponding image in Figure 2) to each gesture to enable us to perform a quick visual labeling and to refer to them in the following. However, these labels were not disclosed during the experiment, as our goal is also to investigate the association between the gesture and semantics using a properly designed questionnaire. In detail, we inserted

- poses already used in gestures employed in existing XR toolkits such as Meta SDK Meta Platforms, Inc. (2026) and MRTK Microsoft (2026), such as those we call INDEX, FIST, PINCH, or FIVE.
- other poses found in the gesture dictionaries proposed in research works (Zhang et al. (2018); Kapitanov et al. (2024); Nuzhdin et al. (2024)), such as those we call THUMB, OK, HOOK, ZERO, LUCK, LOSER, VICTORY, THREE, FOUR, FIVE, SIX, SHAKA, SPOCK, MILITARY, PINKY, POPE, WDW, ROCK, MIDDLE, GUN.
- We added the pose called EIGHT to have the complete list of 0-9 digits of the American Sign Language (Tennant and Brown (1998)), that could, for example, be used for menu control in XR.
- We added the poses we call SPIDERMAN, which is common and interesting because it differs from the thumb extension of ROCK, and might be challenging to recognize, and the pose we call SCOUT, which is also common and sometimes needs to be distinguished from the SIX pose.

Our research group agreed on the set, with all members considering gestures known and recognizable. We did not intend to represent all the poses humans can practically use for communication, nor did we include all the poses used in different sign languages. The primary goal is to have a sufficiently large set of gestures that are immediately recognized and associated with a meaning by general users (not trained in sign languages), while accounting for slight differences in hand articulation that may create issues for both human and automatic recognizers.

Given this set of 27 poses, we created corresponding visual templates (photos shown in Figure 2) to serve as examples during the recording phase, suggesting the execution of the pose without providing hints about its semantics.

Table 1: Subjective questionnaire completed for each gesture. Questions Q3 to Q6 are rated using a 5-point Likert scale ranging from 0 (strongly disagree) to 4 (strongly agree).

QN	Question	Response type
Q1	Are you familiar with Gesture n ?	Yes / No
Q2	If yes: What meaning would you attribute to it?	Open text (keywords)
Q3	How often do you perform this gesture?	Likert scale (0–4)
Q4	Do you feel discomfort while performing it?	Likert scale (0–4)
Q5	Do you feel fatigue while performing it?	Likert scale (0–4)
Q6	Do you feel that you are performing it correctly?	Likert scale (0–4)

3.2.3 Acquisition protocol

We performed the data acquisitions in a laboratory environment, with participants seated comfortably in front of a neutral white background. As we were not interested in assessing arm-related fatigue (which is well studied in the literature), but only issues related to hand articulation, we attempted to minimize upper-arm fatigue and avoid the "gorilla arms" effect. For this purpose, the participants were instructed to rest their forearms on a support. For each gesture of the dictionary, we started by showing its image to the participant (as in Figure 2) on a screen, without any associated label, asking them to answer a short questionnaire (shown in Table 1) designed to assess their familiarity with the gestures, their semantic associations, and the perceived ergonomic issues in their execution.

These questions have been chosen for several reasons. Gesture familiarity (Q1) was identified as a potential cause of preference in elicitation studies (Morris et al. (2010)). The recognition of a meaningful gesture, along with data on frequency of use (Q2/Q3), can be seen as a preliminary identification study proposed for dictionary design (Ali et al. (2019)). Discomfort (Q4) and fatigue (Q5) are critical for the design of hand gestures intended for frequent use. Rempel et al. (2014) demonstrate that physical symptoms such as pain and fatigue are not trivial in gesture-intensive tasks; in their study, 70% of professional signers reported fatigue, and 74% reported discomfort associated with specific hand postures. As we aimed at collecting data on tracking and user execution quality, we are finally interested in assessing the perceived correctness of execution (Q6), which is close to a metric already used in the literature, such as confidence in execution (Sellier et al. (2024)), but is more specific to our study's goals.

After this step, we asked subjects to perform the gestures while wearing the Quest 3 headset in pass-through mode, allowing them to see their hands and the surrounding environment. To support the analysis of specific influencing factors, we acquired the same gestures with varying hand orientations and positions. In detail, we instructed participants to perform the gesture in the *optimal* orientation, meaning with the palm facing the headset, no wrist torsion or hand dorsiflexion, and visible/recognizable fingers' articulation (Chen et al. (2016)). This strict orientation constraint was applied only during the initial 2 seconds of execution to ensure consistent tracking conditions for all subjects and gestures. Subsequently, we asked them to freely change the hand orientation for an additional 2 seconds, while maintaining approximately the same 3D

position and finger articulation. We classify the poses acquired during this interval as performed in the *rotated* orientation. The exact trajectory of the movement is left to the participant's discretion, allowing for user-driven variations.

We recorded the same gestures three times per subject, each executed in a different position. The first position is the *free* one (Figure 3a): the participant is asked to perform the gestures inside a ring-shaped overlay displayed in the Quest's hand tracker field of view, with no other constraints (both in the optimal and rotated orientation).

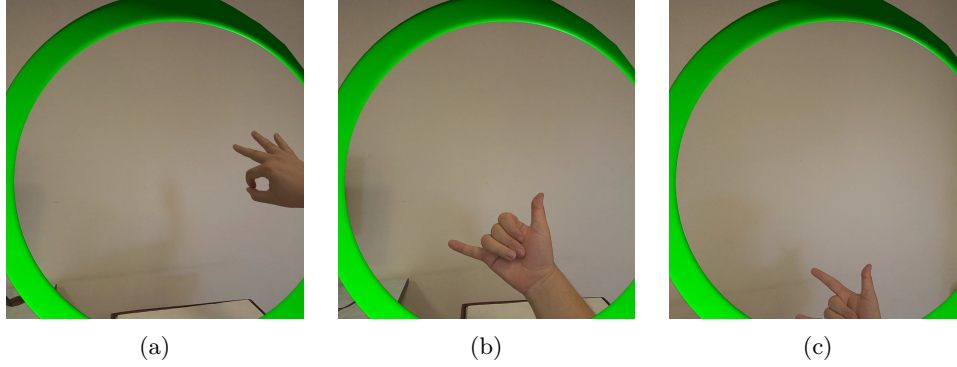


Fig. 3: Gestures acquired in the *free* position are executed with the hand in any position within the green circle (a). In the *central* position recordings, subjects are required to place the hand in specific positions, guided by a visual overlay that disappears when the position is reached (b). The same procedure is done to record gestures in the *bottom* position (c)

The second and the third positions are shown in Figure 3b and Figure 3c. In these cases, we used a visual overlay that disappears when the hand is in the correct position to guide the participant to perform the gesture correctly. For the *central* position, the overlay was put directly in front of the participant, aligned with the headset's central view (Figure 3b). For the *bottom* position, we put it near the bottom edge of the tracking area (Figure 3c). These additional acquisitions were intended to evaluate the tracker's performance across different regions of the 3D space.

After completing the execution of example-based gestures under all the previously described conditions, we also captured a "neutral" pose, verbally asking participants to maintain a relaxed hand, unrelated to a meaningful gesture. We performed this additional acquisition to capture the most frequent hand poses kept in non-meaningful gesticulation, and that should be classified as *non-gesture*. Meaningful gestures in interfaces should be easy to distinguish from non-gesture poses to avoid false-positive detections. We did not administer questionnaires to the participants regarding these poses.

For each subject, we captured pose streams lasting 2 seconds for all the previously described settings. No fixed pauses were enforced between gestures; however, the protocol allowed the subjects to require breaks at any time and for any duration. This

approach reduced the impact of fatigue or discomfort on gesture execution, preserving the quality and naturalness of the recorded data.

3.3 Data analysis methods

To process the captured data stream to obtain quantitative measurements for our analysis, we designed an annotation and processing pipeline that includes extracting representative samples of poses, visual annotation, estimation of local frames and normalization, estimation of feature descriptors, and statistical analysis/classification. Pose data have also been coupled with the questionnaires' outcomes and jointly analyzed. All the pose processing pipeline components, except the annotation tool, have been implemented in Python.

3.3.1 Pose data sampling

From the captured sequences, we selected a subset of frames for analysis. For the *optimal* orientation, we extract a representative frame per subject, while for the *rotated* orientations, captured with a moving hand, we sampled three frames equally spaced in time. We assigned to the extracted pose the reference names of the example image shown to the subjects before gesture execution, which we call the *intention* label of the gesture.

3.3.2 Visual annotation

The extracted, isolated, hand skeletons are subsequently labeled by human annotators based on a 3D visualization of the hand skeletons, without knowledge of the original ground-truth labels. The visual label allows us to assess the reliability of the finger tracker and identify execution issues across different pose classes and acquisition conditions. This process also enables characterization, for each class, of the variability of pose descriptors in well-tracked executions and of the performance of automatic recognition.

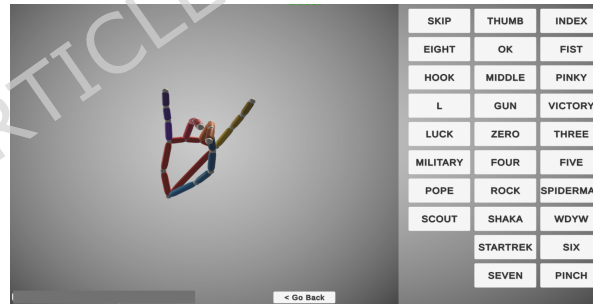


Fig. 4: Screenshot of the visual labeling application.

The procedure was performed using a simple user interface built in Unity. The tool visualizes the hand skeleton as a simple model composed of spheres and cylinders (see

Figure 4), allowing the user to modify the viewpoint interactively. It also allows the user to insert a corresponding label from a list. The tool allowed the authors of this paper, who agreed on the gesture semantics and the associated reference names, to label each pose with the recognized reference name. We refer to the obtained results as the *visual labels* of the sampled poses. In our protocol, three authors of the paper independently performed the labeling. The frames for which all of them assign the same reference name have the corresponding *visual labels* assigned. In contrast, any frame with at least one differing name or that at least one annotator did not recognize as an instance of a dictionary element was classified as NON GESTURE. Combining intention and visual labels, we also derived a set of *verified labels*, which are the same as the *intention* ones when they match the corresponding *visual labels*, but are always assigned as NON GESTURE when the two labels differ.

3.3.3 Pose data normalization

Our goal is to analyze and recognize poses independently of hand orientation and size, so we also transform them into a local frame and normalize their sizes to make the sum of bone lengths uniform. We create the local frame (Figure 5) by setting the x-axis directed from the wrist keypoint to the first middle finger keypoint, the y-axis directed as the normal of the triangle joining the first middle keypoint, the first index keypoint, and the wrist keypoint, and the z-axis given by the normalized cross product of the other two base unit vectors. We also translate the keypoints so that the local frame's origin is at the wrist.

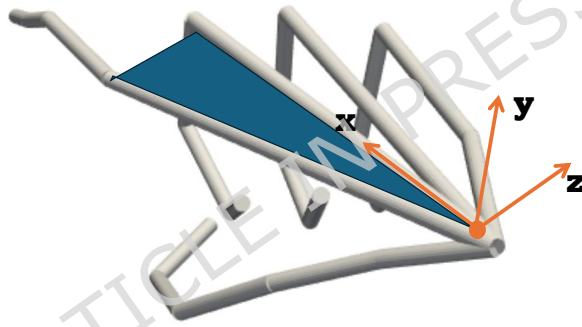


Fig. 5: Local hand frame used to normalize poses.

3.3.4 Pose descriptors

Given the goal of characterizing pose variability within and across classes, we need to introduce pose descriptors and metrics to compare them. A possible approach is to evaluate the distributions of the normalized coordinates, the differences between these coordinates and the average poses, or other derived quantities. However, in the domain of automatic hand pose classification, the most popular descriptor of the pose, not depending on the global hand orientation and position, is the Joint

Collection of Distances (JCD) (Zhang et al. (2017); Yang et al. (2019)). JCD is the vector including all the pairwise Euclidean distances between joints and provides a compact and orientation-invariant descriptor of the hand’s overall shape and topology.

3.3.5 Summary of measurements

Let us summarize all the variables collected in our experimental design. We have a set of measurements obtained during the execution of the gesture instances using the procedures previously described, and a set of subject-related variables obtained from a demographic questionnaire designed to collect basic information on nationality, age, gender, and skin color. These variables are shown in Table 2. To classify skin tones, we adopted the Monk Skin Tone Scale (Monk (2023)), a validated alternative to older classification systems such as the Fitzpatrick Scale applied in domains ranging from national surveys on inequality and health to improving fairness in computer vision and machine learning systems.

Furthermore, each pose is also associated with the responses to the subjective questionnaire reported in Table 1.

Table 2: Measured pose variables (top) and subject related variables in our experimental design.

Pose variable	Orientation	Position	Norm.joints	JCD	Intended lab.	visual lab.	verified lab.
level	opt/rot	f/c/b	24x3 array	276x1 array	string	string	string

Subject variable	Gender	Age	Country	Skin color
level	m/f	scalar	string(name)	string (Monk)

4 Experimental results

4.1 Selection of participants

A total of 20 subjects participated in the acquisition. All participants were right-handed and understood spoken English instructions. Before recording the gestures, they signed an informed consent module in accordance with the protocol approved by the local Ethics Committee and completed the demographic questionnaire.

According to the study design and the goals of the analysis of influencing factors, we selected participants with heterogeneous characteristics to increase the variability of hand poses and enable us to estimate the reliability of the hand-tracking system across different age, gender, and skin tone groups.

We summarize the main characteristics of the subjects in Table 3.

Table 3: Characteristics of the subjects recruited for our study.

Feature	Values range
Gender	13 Males / 7 females
Age	Min: 23, Mean: 33.1, Max: 65
Nationality	9 Iranian, 8 Italian, 1 Ethiopian, 1 Romanian, 1 Spanish
Skin color	1 C; 5 D; 11 E; 2 F; 1 H

4.2 Dataset analysis

The pose sampling procedure, applied on the data captured on the 20 subjects yielded 6,720 frames (20 subjects x 4 samples x 3 positions x 28 labels) used for our subsequent analysis, each one annotated with the 3D positions of the hand joints, a gesture/non-gesture label reflecting the subjects' intended actions (*intention* labels), and then additionally provided with the *visual* label following the previously described protocol.

4.2.1 Analysis of the poses

A first look at the characteristics of the collected hand skeletons reveals that the tracker cannot reconstruct the 3D hand model with consistent scaling. If we estimate the sum of the bone lengths by adding the distances between all the linked joints, we see that the obtained value is not uniform across the same subject. Figure 6 shows the distributions of the bone lengths for the 20 participants in our study. It is possible to see that, while the tracker captures differences in hand size of different subjects (there is a significant difference between the normalized sums of bone lengths across subjects), the tracking system is not providing a reliable estimation of the size, as the intra-subject variability can be relatively high. The distributions with 2/3 clusters of poses with different lengths are explained by the fact that subjects performed the poses in different standard positions. These subjects varied the distance between the HMD and the hand across conditions, and the estimated hand size depended on this distance.

With the normalized poses in a standard frame, we can visualize the variability across subjects in the various executions of the gesture poses with the same *intention* label (see Figure 7). There are significant variations in the poses of selected classes, which may affect gesture recognition, especially when dictionaries include classes with similar shapes (SPOCK/FIVE, THREE/GUN, SIX/SCOUT, etc.).

Another evident fact is that the variability is much greater in the skeletons captured during orientation changes. In particular, the poses captured with the hand not in the optimal orientation and with moving hands exhibit a larger variability in the bending of the fingers around the knuckles (Figure 8b, Figure 8d). This fact makes it more difficult for an automatic method to correctly label poses, especially those with open fingers, such as FIVE (Figure 8b), which may be easily mistaken for NON GESTURE (Figure 8d). Similarly, in rotated orientations, the poses executed as THREE are difficult to distinguish from those executed as GUN. The mean distance from the average pose of the fingertips in the FIVE gestures is 2.3 cm for the optimal orientation

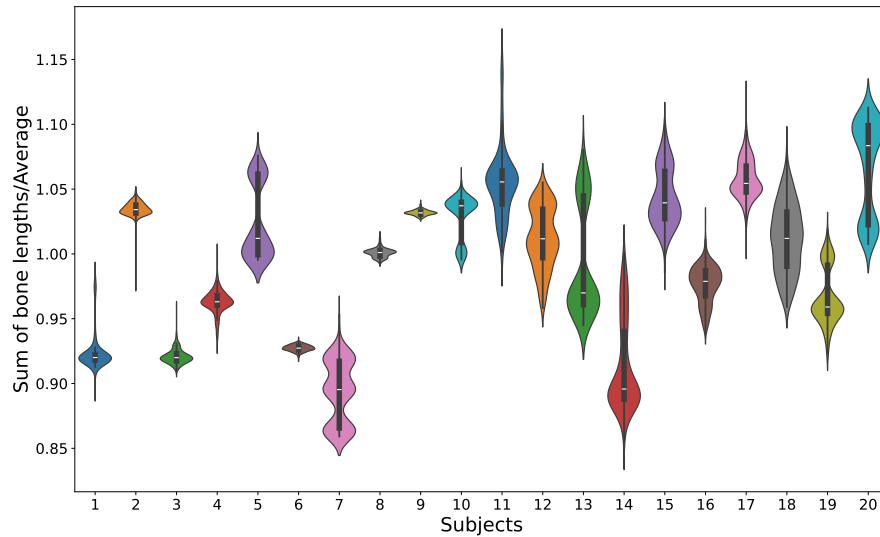


Fig. 6: The violin plot visualizes the distributions of the bone lengths for each participant, estimated on the selected frames.

and 3.5 in the rotated one. In contrast, for the THREE pose, the distance increases from 1.7 cm in the optimal orientation to 3.0 cm in the rotated one.

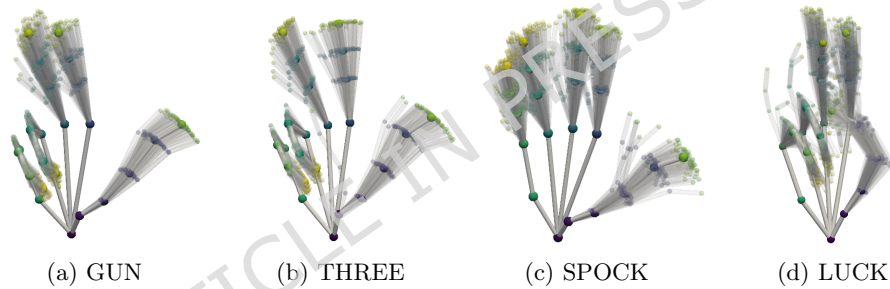


Fig. 7: Distributions of the normalized keypoints around the average pose of four gestures. Large variations appear, especially in finger bending around the knuckles. Due to the variability, similar shapes (a) and (b) can be difficult to label. Some shapes like SPOCK (c) or LUCK (d) are not well tracked and hard to recognize looking at the rendered skeletons.

Figure 9 shows the variability around the average JCD vector descriptors of all the poses recorded as belonging to the same classes. It is possible to see that there are significant variations, as expected from the visualizations of Figure 7 and Figure 8. Poses with open hands obviously have larger variability. The NON GESTURE class is not remarkably varied, as the subjects mostly kept a neutral pose with relaxed hands.

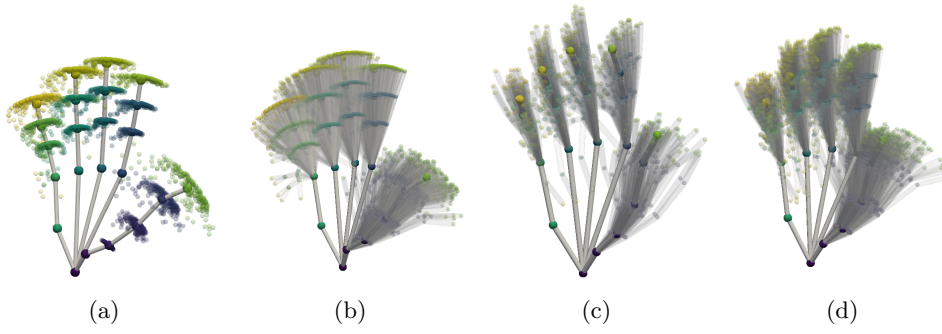


Fig. 8: Pose distributions for the FIVE gesture for optimal orientation (a) and generic orientations (b); and for the relaxed-hand NON GESTURE executions for optimal orientation (c) and generic orientations (d). The rotated poses tend to present a larger bending around the knuckles, especially for the FIVE pose, which becomes in some cases quite similar to the NON GESTURE .

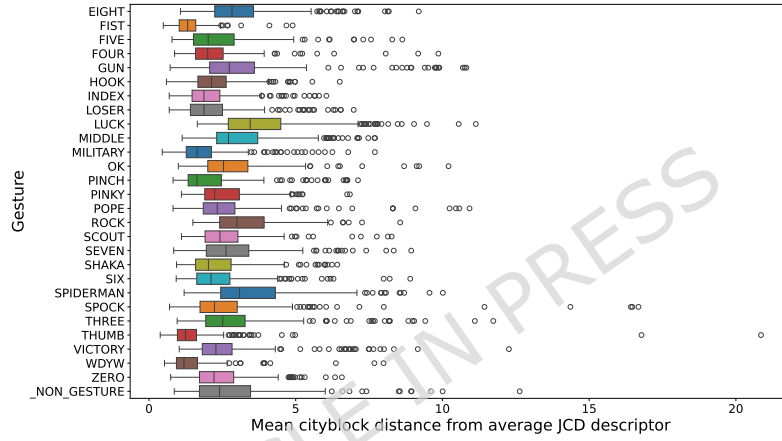


Fig. 9: The variability around the average JCD vector descriptors of all the poses recorded as belonging to the same classes.

We reported in [Figure 10](#) the analysis of the intra- and inter-class similarity of the poses. [Figure 10a](#) shows the heatmap similarity, and poses that are supposed to be challenging to distinguish present inter-class distances close to the intra-class ones. Performing a hierarchical clustering ([Figure 10b](#)) based on this matrix, we can detect groups of poses that may be difficult to separate and should be merged or not included in the same dictionary (GUN/THREE, SCOUT/SIX, PINCH/HOOK/INDEX, FIVE/SPOCK).

Independent of the descriptor employed, we can think of assigning a semantic to each meaningful pose based on margins on the descriptor components. However, this is not necessarily straightforward, as shown in [Figure 11a](#), where we mapped the JCD

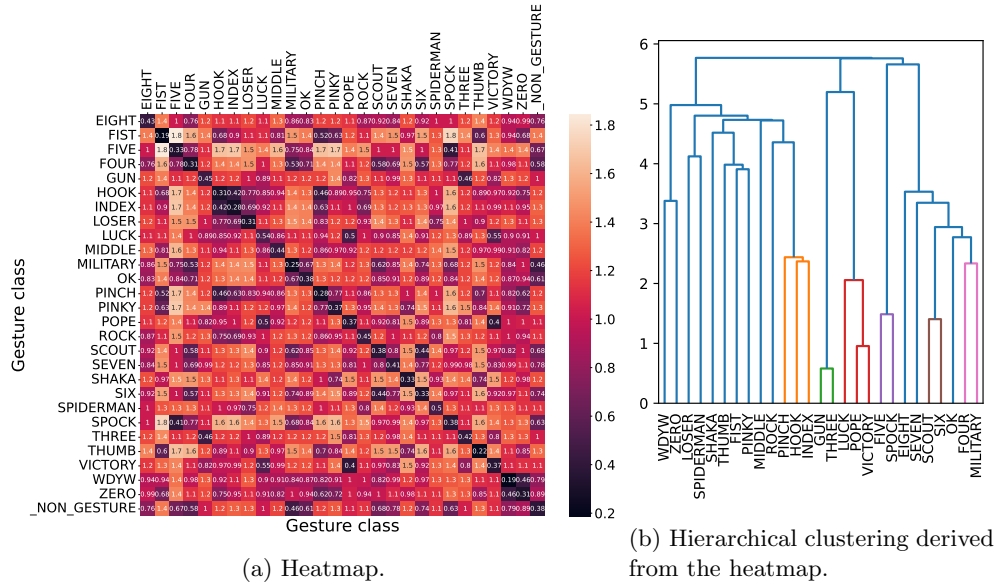


Fig. 10: Analysis of the average intra- and inter-class cityblock distance of the JCD descriptors.

descriptors of all the poses in a 2D space using t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton (2008)). It is possible to see that, while clusters corresponding to the gesture labels are clearly visible, there are large superimpositions and misplaced items. If we evaluate the separability of class pairs using a linear support vector machine (Figure 11b), most of the classes are not separable, which means that it would be tough to obtain an effective automatic classification based on the captured poses. The results improve substantially when considering the same visualizations performed with the poses captured in the *optimal* orientation Figure 12, with better defined clusters, but still, surprisingly, few classes are separable from all the others (PINKY, SPIDERMAN).

We need, however, to remember that the poses have so far been grouped using *intention* labels.

We need to understand if they are correct and if we can trust the output of the tracker and the subjects' ability to articulate the fingers as expected.

4.2.2 Visual labels vs intention labels

Comparing the two sets of labels, we can see that there is a non-negligible number of instances where *visual* labels do not correspond to the *intention* labels. Considering the *visual* labels as a prediction of a hypothetical ground truth represented by the *intention* labels, we can estimate per-class and average classification scores that are sometimes not very good. Table 4 shows the scores for the gestures recorded in the *optimal* orientation only, while the scores obtained on the entire set of recorded poses

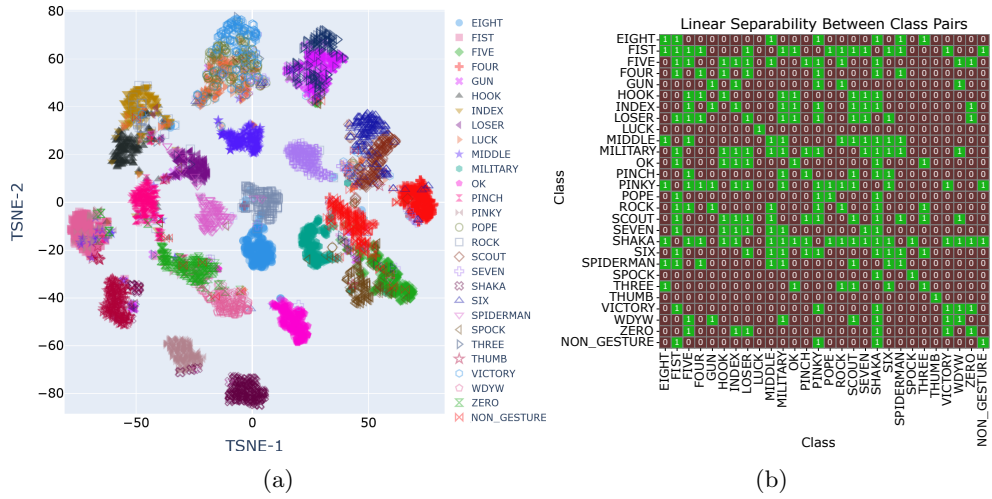


Fig. 11: (a) Visualization with 2-dimensional t-SNE projection of the JCD features describing all the poses. Clusters are not completely separated, and there are many outliers. (b) matrix showing the linearly separable class pairs in the JCD space (corresponding to green cells). Most of the pairs are not separable (brown cells).

are shown in Table 5. Even in the *optimal* orientation, errors occur (gestures are not recognized as intended), indicating issues with the tracking system or gesture execution. Labels in the tables are sorted by F-score. The highest F-scores correspond to gestures that are not the most commonly used in interactive systems. Scores are mostly high for the optimal orientation, indicating that humans recognize the gestures from the tracked skeletons with few errors or false positives; however, there are notable exceptions, such as LUCK, POPE, ZERO, SPOCK, INDEX, and VICTORY.

If we include non-optimal orientations (right part of Table 4), we instead observe a significant decrease in visual recognition of the gestures and a different ranking of the classification scores. This increased number of errors may depend on both the higher probability of occlusions and the interplay between global motion and the subjects' ability to maintain a fixed hand shape. In particular, we observed substantial decreases in recognition rates for poses such as FIVE or THREE. We already noted that this can be because subjects tend to bend the fingers when changing globally the hand orientation (see Figure 7 and Figure 8), making the open hand pose more similar to a relaxed hand or the THREE gesture closer to the GUN pose.

We can examine the confusion matrices to analyze the mislabeled data in more detail. Figure 13a shows the misclassifications for gestures executed in the optimal orientation. There are visible issues with the tracking system. The fact that the executions intended as LUCK are interpreted as POPE or VICTORY from the rendered skeletons means that the tracker does not work well with crossed fingers. Other issues are clearly due to the interplay between variable execution and pose similarity (HOOK appearing as INDEX, SPOCK appearing as FIVE, PINCH appearing as ZERO). Figure 13b shows the label assignments for all the orientations. In this case, many executions were

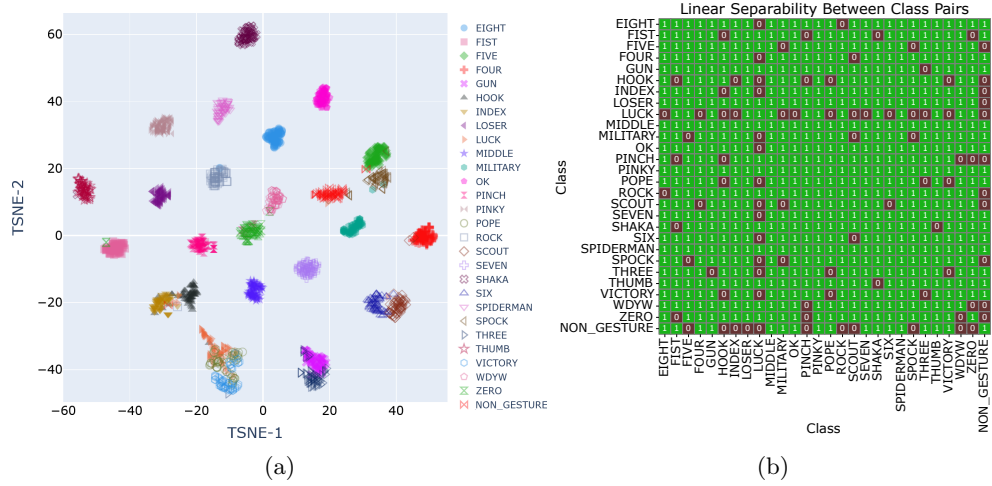


Fig. 12: (a) Visualization with 2-dimensional t-SNE projection of the JCD features describing the poses for the *optimal* orientation only. Clusters are better defined. (b) Matrix showing the linearly separable class pairs in the JCD space (corresponding to green cells). Most of the pairs are linearly separable, but there are still ambiguous cases.

not recognized with certainty and were assigned to the NON GESTURE class in the *visual* labeling.

The classification results are not affected by variations in hand position or skin tone. Figure 14 shows the average F-scores obtained for different levels of these variables, which appear negligible. This fact shows that the tracking system of the Quest 3 is robust against them. We also evaluated differences by grouping subjects according to characteristics such as nationality (Iranians vs. Italians), gender (M vs. F), and age (< 35 vs. ≥ 35), without finding any statistically significant differences in average scores.

4.2.3 Automatic classification of poses

To complete our analysis of the dataset, we performed simple tests of automatic labeling using the JCD pose descriptors and a linear decision boundary classifier (Linear Discriminant Analysis, LDA). The goal was not, in fact, to evaluate or compare the performance of complex learning methods, but to assess the influence of data quality and label assignment on the automatic classification of poses. We used a Leave-One-Subject-Out procedure, comparing each subject's poses with classifiers trained on the data from all other subjects, and collecting all predictions into a single vector.

A first, interesting result is that, using as training and testing labels the original *intention* labels (the same ones compared to the visual labeling in Table 4), the classification scores are higher (F-score=0.89) than those obtained in the visual recognition (F-score=0.81, see Table 6).

	Prec.	Rec.	F-Sc.
MIDDLE	1	1	1
OK	1	1	1
PINKY	1	1	1
SEVEN	1	1	1
SPIDERMAN	1	1	1
EIGHT	1	0,98	0,99
FOUR	0,98	1	0,99
LOSER	1	0,98	0,99
SHAKA	1	0,98	0,99
THUMB	0,98	1	0,99
WDYW	0,98	1	0,99
MILITARY	1	0,97	0,98
ROCK	0,98	0,98	0,98
FIST	0,97	0,97	0,97
SIX	0,95	0,98	0,97
THREE	0,97	0,97	0,97
FIVE	0,95	0,97	0,96
GUN	1	0,92	0,96
SCOUT	1	0,9	0,95
HOOK	1	0,88	0,94
PINCH	0,96	0,92	0,94
SPOCK	1	0,87	0,93
VICTORY	0,87	1	0,93
INDEX	0,84	0,98	0,91
ZERO	0,91	0,83	0,87
POPE	0,68	0,78	0,73
NON_GESTURE	0,48	0,98	0,64
LUCK	1	0,03	0,06
Average	0,94	0,93	0,92

Table 4: Classification scores in visual labeling of skeletons captured in optimal orientation. Gestures are sorted by F-Scores, taking both missed and false detections. Top rated gestures in the left table are the most robust against issues related to bad execution.

	Prec.	Rec.	F-Sc.
OK	0,99	0,92	0,95
SHAKA	0,99	0,89	0,94
WDYW	0,97	0,90	0,94
FOUR	0,92	0,93	0,93
MILITARY	0,95	0,88	0,92
SEVEN	1,00	0,84	0,91
THUMB	0,86	0,94	0,90
EIGHT	1,00	0,81	0,89
PINCH	0,88	0,91	0,89
LOSER	0,87	0,88	0,88
MIDDLE	0,97	0,80	0,88
FIST	0,78	0,98	0,87
PINKY	0,99	0,78	0,87
SIX	0,92	0,82	0,87
SPIDERMAN	0,99	0,78	0,87
VICTORY	0,88	0,84	0,86
FIVE	0,85	0,83	0,84
HOOK	0,98	0,73	0,84
ZERO	0,86	0,81	0,83
ROCK	0,90	0,75	0,82
SCOUT	0,99	0,70	0,82
THREE	0,93	0,74	0,82
GUN	0,87	0,75	0,81
INDEX	0,70	0,91	0,79
POPE	0,63	0,77	0,69
SPOCK	1,00	0,52	0,68
NON_GESTURE	0,23	0,96	0,38
LUCK	1,00	0,01	0,02
Average	0,89	0,80	0,81

Table 5: Classification scores in visual labeling of skeletons captured generic orientations. Gestures are sorted by F-Scores, taking both missed and false detections. Top-rated gestures are the most robust against orientation changes during execution.

Training label	Test labels	Avg. precision	Avg. recall	Avg. F1-Score
No training-Visual classification	Intention	0,89	0,80	0,81
Intention	Intention	0,89	0,89	0,89
Verified	Intention	0,89	0,80	0,82
Intention	Verified	0,86	0,80	0,76
Verified	Verified	0,88	0,88	0,87

Table 6: Classification scores obtained with supervised labeling in a leave-subject one method (see text) using the original execution labels or the verified labels for training and as ground truth for testing.

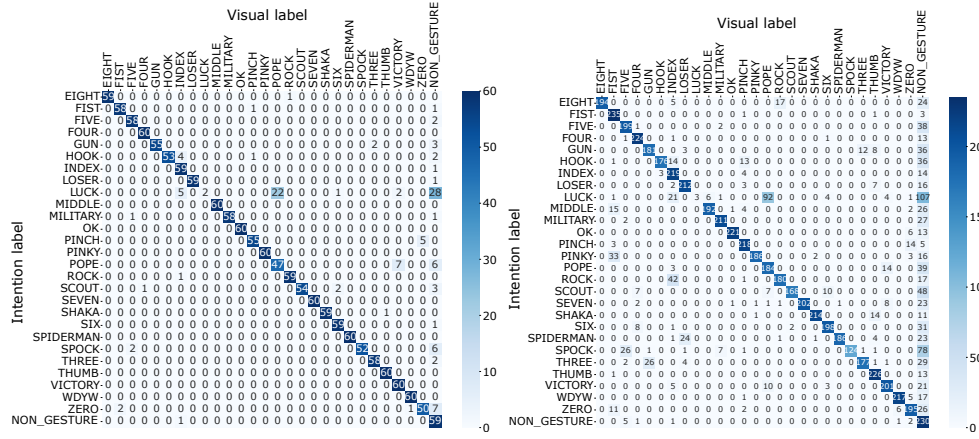


Fig. 13: Confusion matrices in visual labeling of captured skeletons.



Fig. 14: The variability of the F-scores of the visual classification.

The improvement is because, while the tracking and execution variations make the skeletons captured not easy to be visually recognized as corresponding to the *intention* label, biases in the executions and trackings are similar across different subjects, so that an example-based algorithm can predict the *intention* labels successfully. This behavior is particularly evident when looking at per-class scores and can be seen in Figure 15. Here, the blue bars represent the F-score of the human visual classification (blue), and the red ones are the JCD/LDA classification scores using the *intention* labels for training and test. The difference is significant for poses like LUCK and SPOCK, which are not properly tracked but can be successfully labeled using the example-based method.

The problem with the *intention* labels associated with poorly executed or poorly tracked data (and thus with skeletons that are not recognized as belonging to the class) is that they cannot be considered a true "ground truth". Ideally, these data should not be used for classifier training, even if it is effective, and even though using

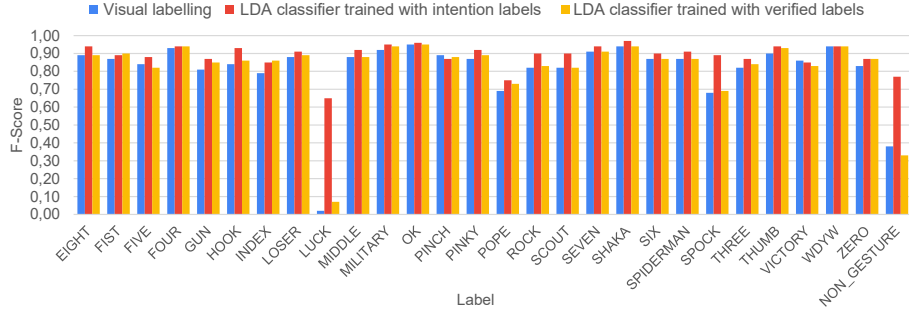


Fig. 15: Per-class scores of visual and automatic classification using the *intention* label as "ground truth". Training with the intention labels helps predict the subject's intentions (red bars) and is more effective than visual recognition of skeleton renderings (blue bars). An automatic classifier trained with the verified labels performs like visual recognition (yellow bars).

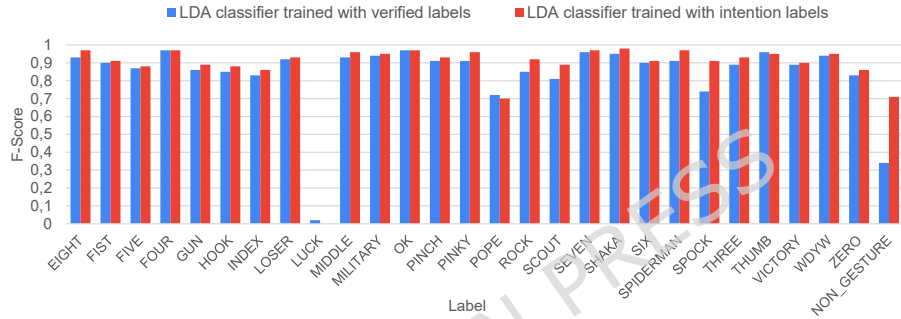


Fig. 16: Per-class scores of automatic classification using the *verified* label as "ground truth". Verified training data (red bars) results in better prediction performances with respect to the *intention* labels (blue bars).

the subject's intention to label gesture examples when collecting this kind of dataset is common.

For this reason, we introduced the *verified* labels, which are identical to the *intention* labels only when they match the corresponding *visual* labels and are assigned as NON GESTURE in all other cases. This procedure creates imbalanced classes with many examples of the NON GESTURE class that no longer represent only relaxed neutral poses, but also ambiguous poses that fill gaps between the gesture classes in the pose feature space. If we use these "clean" *verified* labels to train the LDA classifiers, the predicted labels no longer match the *intention* ones better than the visually recognized ones (see yellow bars in Figure 15 and Table 6). It is worth noting that if our goal is to achieve good classification of well-executed/tracked poses, we should train

our method on verified data, as this choice provides much better results (Table 6). Figure 16 compares the F-scores of the predictions based on LDA classifiers trained and tested with *intention* and *verified* labels.

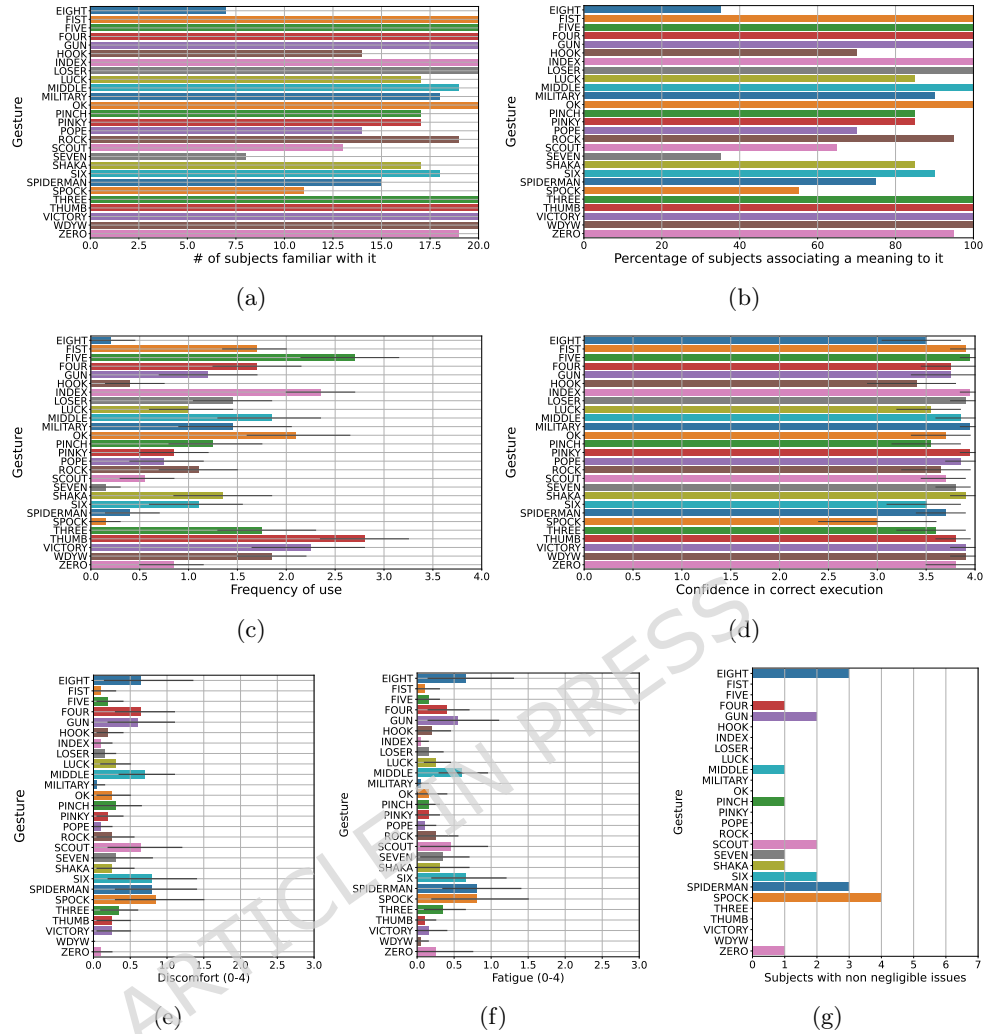


Fig. 17: Outcomes of the questionnaire. Most of the poses are associated with meanings by almost all the subjects (a), even if the frequency of use is often limited (b). Confidence in the correct execution of the gesture is relatively high across all poses. The average values of discomfort and fatigue reported are low (d,e), but a few subjects reported critical values (> 2) (f).



Fig. 18: Word clouds with the collected terms associated with the subjects and hand shape templates.

4.3 Questionnaire outcomes

Figure 17a displays the number of subjects considering each gesture familiar. Figure 17b shows the percentage of subjects knowing some meanings for the exemplar gestures, and Figure 17c displays the frequency of use of the different poses. Most of the proposed gestures are known and associated with one or more meanings by most participants. Figure 18 shows the word clouds with the terms related to the provided templates. There are, however, notable exceptions. SEVEN, EIGHT, and SPOCK are associated with meaning by less than 60% of the subjects and are also practically not used. HOOK, POPE, and SCOUT are also not associated with meaning by more than one-fourth of the subjects. Most of the gestures are not frequently used, and only FIVE, INDEX, OK, THUMB, and VICTORY have an average value larger than 2.

It is worth noting that the lack of semantic association or the infrequent use does not influence the confidence in the correct execution of the gestures [Figure 17c](#) as well as the results about the recognizability of the poses reported in [subsubsection 4.2.2](#).

The subjects were mostly confident to perform the gesture correctly according to the example-based semantics ([Figure 17c](#)), with only HOOK and SPOCK reporting an average score lower than 3.5 (but higher than 3). Average values of discomfort and fatigue were low for all the labels ([Figure 17d](#), [Figure 17e](#)), but a few outliers assigned critical scores (> 2) to some of them ([Figure 17f](#))

5 Discussion

Even though we designed it to address a set of specific research questions, our study also yields several findings that can be extremely useful to interaction designers in creating effective gesture-based interfaces. In this section, we discuss how our measurements address our research questions, summarize key takeaways, and clarify the analysis's limitations and potential future research directions.

5.1 Answers to the research questions

5.1.1 Tracker performance (RQ1)

Regarding the performance of the Quest 3 hand-tracking system, our tests demonstrate that it is effective and robust across various skin tones and acquisition positions. It can also produce accurate skeletons in cases of non-optimal orientation and partial occlusions. However, it is worth noting that the tracking system sometimes fails on certain poses. For example, crossed fingers (LUCK) present significant errors, and the angles between adjacent fingers are likely not always accurate in unusual poses, such as SPOCK. These tracking issues can lead to gesture recognition errors and can occur when the gesture is performed in a nearly optimal orientation, avoiding significant occlusions. The tracker constantly reconstructs the correct number of fingers. Still, gestures that differ in whether adjacent fingers are attached or form a small angle (e.g., POPE vs. VICTORY) can be challenging to distinguish. This fact suggests that, using a skeleton-based recognizer, we should avoid dictionaries with gestures differing only in the angles between open fingers. In this case, an image-based classifier would likely provide more accurate predictions.

5.1.2 Execution issues (RQ2)

The questions on discomfort and fatigue revealed that the ergonomic issues for executing the proposed poses are limited. However, there are non-negligible issues for a few subjects and poses (see [Figure 17f](#)). In some cases, these issues pertain to particular characteristics of poses already associated with discomfort in the literature, such as discordant adjacent fingers or extended fingers ([Rempel et al. \(2014\)](#)). We find these characteristics in EIGHT, MIDDLE, GUN, SIX, SPIDERMAN, and SPOCK, for example. Other poses, however, feature extended fingers (FIVE) or discordant adjacent fingers (LOSER, PINKY), but were not associated with critical scores by the subjects. Apart from SPOCK, which is often mistaken for FIVE, the other poses

rated as critical are not associated with increased difficulty in recognition. However, it may be a good guideline to avoid including gesture dictionaries poses where recognizing discordant adjacent fingers or differences in finger extension is essential for correct labeling. Some problems, however, could be resolved by modifying the template/instructions to provide the "semantics" of the gesture classes. For example, the issues reported during the execution of SIX, SEVEN, or EIGHT stem from the difficulty of keeping the fingers from contacting the thumb when it is extended. However, they are well recognized, even if they are often poorly executed, as the key features for recognition are not the finger opening or the evident visibility of discordant fingers, but the touching of the fingertips. If the gesture templates for them had presented relaxed fingers, the issues would have disappeared.

5.1.3 Gesture semantics (RQ3)

In our experiments, we defined the semantics of the gestures with a visual example. This procedure corresponds to how we learn to communicate with gestures ([Kendon \(2004\)](#)), and worked well in our experiments. In fact, the gestures previously unknown to the subjects are successfully executed and recognized without any additional issues. Considering the labels SEVEN and EIGHT, which are those that are declared as not associated with meaning by most of the subjects, in fact, we have that, for the first label, the average F-score of the visual classification is 0.91 for the subjects not associating meaning and 0.84 for those associating a meaning. For EIGHT, the average F-score is 0.89 for the subjects not associating meaning and 0.83 for those associating a meaning.

Given the descriptors (normalized joint coordinates or JCD) of the well-executed gestures of each class, e.g., those identified by the *verified* labels, we could derive a rule-based semantic association for novel gesture instances, defined by ranges of acceptable values for the parameters. In practice, however, this is not straightforward, as the variability in pose execution is vast and not consciously controlled. Our analysis demonstrated that, when using dictionaries with similar gestures, ambiguities and errors can occur, especially when subjects are allowed to move or rotate poses, as they inadvertently modify them without realizing it. However, we plan to exploit the captured data to derive a rule-based definition of selected gestures based on the range of variability of the pose descriptors in the example (visually verified) data, at least for simplified vocabularies.

5.1.4 Human vs automatic labeling: biases and the role of the NON GESTURE class (RQ4)

The visual classification of the captured skeletons is analyzed to understand the trackers' limitations and execution inaccuracies. However, in a practical application, we must deal with an automatic classification typically based on a supervised classifier trained on labeled data. In existing benchmarks, gesture labels are usually those here called the *intention* labels. We have seen that, using a simple multi-label classifier trained on JCD features and intention labels, the machine can recognize gestures that humans would not consider poorly executed or tracked (e.g., LUCK). This happens

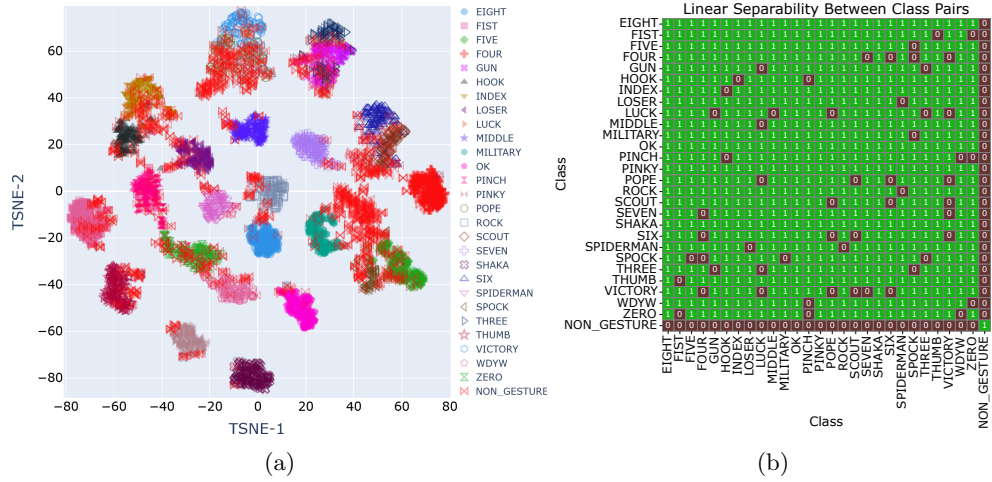


Fig. 19: (a) Visualization with 2-dimensional t-SNE projection of the JCD features describing all the poses grouped by *verified* labels. Clusters of gesture labels are clearer than in the case of intention labels, but NON GESTURE elements are here mixed with them. (b) Gesture classes are mostly linearly separable (green cells), but the labeling of ambiguous poses as non gestures makes the classes not linearly separable from NON GESTURE in the JCD space.

because the related poses, even if they do not correspond to the visual examples, exhibit similar biases in the training and test data. This effect can be exploited in practical recognizers, but it would create problems if poses with these execution biases are similar to those of other classes or can be executed involuntarily by subjects. A more correct way to train automatic classifiers is to create a *verified* set of labels, assigning all poorly executed or poorly tracked examples to the NON GESTURE class. However, the relabeling can create, as in our case, poorly represented classes in the training set (e.g., LUCK) and generate many NON GESTURE examples similar to gesture ones, making them difficult to discriminate (Table 6, right).

This behavior also clarifies the role of the NON GESTURE class for the training gesture recognizers. If we use a supervised learning method trained on gesture-class examples, without setting margins on the parameters describing the pose, all non-meaningful poses would be labeled as belonging to the closest class in the latent representation. This choice would lead to many false-positive detections in practical use. This issue can be avoided by adding the NON GESTURE class to the training set, but, to be effective, this method would require regular sampling of all feasible poses that don't have meaning in our dictionary. We created our original NON GESTURE class using relaxed poses, but this covers only a small subset of the non-meaningful hand shapes (the most frequent ones), as shown in Figure 11a. This issue led to significant false-positive detections, affecting the precision values reported in Table 6.

On the other hand, labeling all the poses not recognized unanimously by human annotators as NON GESTURE in the verification step would make it quite challenging to create effective classifiers, due to the high similarity of gesture and non-gesture examples. This fact is evident in [Figure 19a](#), where we repeated the 2D t-SNE visualization of the JCD descriptors for the poses, now grouped by *verified* labels. Clusters are well defined, but the NON GESTURE elements are interleaved with the elements of the other classes, and all the classes are not linearly separable from the NON GESTURE one in the JCD space.

An optimal recognizer for a gesture dictionary should find optimal margins in the pose descriptor spaces that separate the classes from each other and from the NON GESTURE region. This region is challenging to model due to its poor representation in the training data. The visually verified labeled examples obtained with our protocol can be used to estimate class margins with a data-driven approach, as well as to create a pose distribution model ([Cootes et al. \(1995\)](#)) that can generate valid synthetic examples of NON GESTURE.

5.1.5 Optimal dictionaries (RQ5)

Based on the outcomes of our analysis, guidelines can be derived to create gesture dictionaries that facilitate robust and ergonomic interaction. We should avoid using classes that present execution/tracking issues corresponding to low scores in [Table 4](#) (left or right, depending on whether we plan to use the pose in optimal or variable orientation). Note that we have many poses with high values for the optimal orientation. Still, some notable examples of poses widely used in practice, such as the PINCH, INDEX, and VICTORY poses, present significant issues. These issues can be reduced by merging classes (e.g., VICTORY/POPE/LUCK), but this would change how we assign gesture semantics. For the rotated poses, we encounter more relevant problems due to occlusions, poor tracking, and changes in hand shape. Gestures like the open hand (FIVE) become critical, as they can be mistaken for relaxed NON GESTURE poses.

A good dictionary could be created by choosing the gestures with the highest F-Score (or precision if false positives are more critical), removing the labels creating ergonomic issues (e.g., corresponding to non-zero values in [Figure 17f](#)). In this case, the best five gestures would be OK, PINKY, LOSER, THUMB, and WDW for executions in optimal orientation; OK, WDW, LOSER, MILITARY, and THUMB, for generic ones. If we ignore the self-reported ergonomic scores, gestures like EIGHT and SHAKA are also particularly interesting, being linearly separable from the other gesture classes even in rotated executions, but also FOUR and SEVEN are good.

Our analysis shows that, given specific design requirements (e.g., the number of labels required, orientation types, and the criticality of ergonomic factors), it is possible to optimize the dictionary by introducing cost functions based on classification accuracy and class separability, while also considering label merging. In our dataset, we collected semantic associations assigned to the hand poses (independent of orientation) by the subjects. An accepted association of a pose with a specific command to issue might be an obvious choice to improve intuitiveness or learnability ([Xia et al. \(2022\)](#)), but, apart from this particular case, we don't consider the existing semantic

associations collected as a factor to be considered in a dictionary optimization. The association, in fact, does not influence the execution, and the representativity of the set of subjects performing the gestures is limited.

5.2 Key takeaways

Some of the outcomes of our measurements can directly provide guidelines for hand gesture-based interaction design:

- If a supervised gesture recognizer is based on skeleton tracking, it is important to consider the issue of label noise (Frénay and Verleysen (2013)) that may affect *intention* labels. In fact, they may be biased by tracking failures or inaccurate executions of some subjects (see Figure 13). Biases, however, may increase recognition rates in particular cases.
- Among the many criteria that should be considered in gesture-vocabulary design (Xia et al. (2022)), our results suggest that XR gesture dictionaries should prioritize poses whose salient articulatory features are minimally affected by execution errors (see Table 4, Table 5) and remain stable under normal user variability and the sensing limitations of head-mounted hand tracking. This criterion favors poses whose defining features are robustly captured even when non-essential fingers vary, such as OK, WDW, SEVEN, and EIGHT, while discouraging poses whose recognition depends on subtle inter-finger angles, crossed fingers, or small differences from nearby poses. SEVEN and EIGHT could be better defined with relaxed, non-salient fingers rather than requiring unnecessary finger extension, to reduce potential discomfort Rempel et al. (2014).
- If hands can be rotated while gestures are being recognized, use poses robust against orientation change (e.g., SHAKA), and avoid poses that are weak against them (GUN, THREE, FIVE)
- Many poses popularly used in gestural interaction, like INDEX, PINCH, GUN, or ZERO, present serious discriminability issues, and could be replaced by more robust poses. The idea of combining gestures in dictionaries, considering their discernibility from other elements, proposed in dictionary design tools (Mo et al. (2021)), should also be applied to commonly used poses.
- Do not include too similar poses (e.g., GUN/THREE, FIVE/SPOCK, or PINCH/HOOK/INDEX. Consider using a more relaxed semantic definition by merging related classes (POPE, VICTORY, and LUCK could, for example, be merged into a class defined as "index and middle finger open").
- When creating a training set, consider the importance of non-gesture examples to define correct margins in the descriptor spaces. They should ideally cover the feasible poses, or at least separate the gesture-label clusters. This has been done in recent works, such as (Nuzhdin et al. (2024)). If this is not possible, as the gesture recognition task may be considered an open-set problem, the classifier should implement specific strategies to mitigate the risk of failures (e.g., Scheirer et al. (2012)).
- If there are no particular requirements, e.g., discoverability, avoid relying on existing knowledge of gestures and related meaning to define the vocabulary. It is preferable

to rely on actual execution ability, robustness to tracking issues, and separability from other dictionary elements. Humans often overestimate their ability to perform recognizable poses, and knowledge of gestures does not significantly improve recognition rates. Elicitation studies may be misleading, suggesting gestures that are not optimal for execution and recognition.

5.3 Limitations

While our analysis provides helpful guidelines, we would like to highlight its limitations. We recorded the data using a specific tracker, so we cannot immediately apply all the results to other systems. However, the issues due to viewpoint and occlusions are the same for all the head-mounted trackers. The Quest 3 tracker performed well, even though some tracking biases (e.g., failure to represent crossed fingers or the SPOCK pose) may be related to specific priors of the software used. We performed additional tests with various devices (e.g., HoloLens 2 and Leap Motion) that yielded similar results.

Our analysis is limited to static poses and orientation-invariant ([Emporio et al. \(2025\)](#)) gesture dictionaries. While this may seem like a limitation, it was a deliberate choice to specifically analyze how different hand articulation capabilities and tracking accuracy influence pose recognition alone, regardless of gesture dynamics. We plan to extend the analysis in future work to dynamic gestures, leveraging the results from this study.

The number of subjects performing the gestures, though diverse, remains small for generalization. While this is a limitation, the data acquired were sufficient to identify several issues, derive applicable guidelines, and emphasize the importance of adopting specific strategies for designing effective gesture recognizers. We plan, however, to enlarge the dataset with future acquisitions. We based our ergonomic evaluation solely on self-reported discomfort and fatigue after a short task. Further experiments using physiological measurements and longer tasks could improve the significance of the results. The dictionary of poses used is not exhaustive of all the potentially meaningful hand shapes. While it is not possible to eliminate this effect, it is worth noting that the choice of dictionary influences pose classification, as the presence of classes with similar hand shapes can make assigning visual labels more difficult.

6 Conclusions

Thanks to the development of effective vision-based hand trackers, XR applications are now often provided with gesture-based interaction options. However, the gesture dictionaries used in these applications are typically limited, and the user experience is not always good due to missed or false detections. We believe that our study, based on an extensive collection of hand poses executed by heterogeneous subjects in different conditions, helps to understand several factors that limit the success of gestural interfaces for XR, that do not depend on the machine learning approach used for recognition, but are strongly affected by tracking devices' limitations, human factors, and the design of the dictionary. A carefully designed interactive XR application, along

with its associated gesture dictionaries, informed by the study's outcomes presented here, could facilitate the widespread adoption of this type of interface.

Compliance with Ethical Standards. The study with human subjects was conducted in accordance with recognized ethical standards and was approved by the "Comitato di Approvazione della Ricerca sulla Persona" (CARP) of the University of Verona. All participants provided written informed consent prior to taking part in the experiment.

Acknowledgements. This study was carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from NextGeneration EU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE000000004). CUP MICS D43C22003120001 - Cascade funding project CollaborICE.

References

- Al-Kalbani M, Frutos-Pascual M, Williams I (2017) Freehand grasping in mixed reality: analysing variation during transition phase of interaction. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. Association for Computing Machinery, New York, NY, USA, ICMI '17, p 110–114, <https://doi.org/10.1145/3136755.3136776>, URL <https://doi.org/10.1145/3136755.3136776>
- Ali AX, Morris MR, Wobbrock JO (2019) Crowdlicit: A system for conducting distributed end-user elicitation and identification studies. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp 1–12
- Angrisani L, De Benedetto E, Duraccio L, et al (2024) Exploring variability in human-machine interaction through gesture recognition based on hand-tracking within xr. In: 2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine), pp 816–820, <https://doi.org/10.1109/MetroXRaine62247.2024.10796138>
- Bertolasi J, Garcia-Hernandez NV, Memeo M, et al (2025) Evaluation of hololens 2 for hand tracking and kinematic features assessment. In: Virtual Worlds, MDPI, p 31
- Borzelli D, Boarini V, Casile A (2025) A quantitative assessment of the hand kinematic features estimated by the Oculus Quest 2. Scientific Reports 15(1):8842
- Caldiera VRBG, Rombach HD (1994) The goal question metric approach. Encyclopedia of software engineering pp 528–532
- Chen LC, Chu PY, Cheng YM (2016) Exploring the ergonomic issues of user-defined mid-air gestures for interactive product exhibition. In: Streitz N, Markopoulos P (eds) Distributed, Ambient and Pervasive Interactions. Springer International Publishing, Cham, pp 180–190
- Chen Y, Hu Y (2025) Ergonomics evaluation of gestures interaction for the elderly in extended reality. International Journal of Industrial Ergonomics 107:103749.

- <https://doi.org/https://doi.org/10.1016/j.ergon.2025.103749>, URL <https://www.sciencedirect.com/science/article/pii/S0169814125000551>
- Cootes TF, Taylor CJ, Cooper DH, et al (1995) Active shape models-their training and application. *Computer vision and image understanding* 61(1):38–59
- Emporio M, Ghasemaghaei A, Laviola Jr JJ, et al (2025) Continuous hand gesture recognition: Benchmarks and methods. *Computer Vision and Image Understanding* p 104435
- Erazo O, Rekik Y, Grisoni L, et al (2017) Understanding Gesture Articulations Variability. In: Bernhaupt R, Dalvi G, Joshi A, et al (eds) *Lecture Notes in Computer Science, Human-Computer Interaction - INTERACT 2017*, vol LNCS-10514. Springer International Publishing, Bombay, India, pp 293–314, https://doi.org/10.1007/978-3-319-67684-5_18, URL <https://inria.hal.science/hal-01678464>, part 4: Information on Demand, on the Move, and Gesture Interaction
- Frénay B, Verleysen M (2013) Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems* 25(5):845–869
- Graichen L, Graichen M (2023) Do users tolerate errors? effects of observed failures on the subjective evaluation of a gesture-based virtual reality application. In: da Silva HP, Cipresso P (eds) *Computer-Human Interaction Research and Applications*. Springer Nature Switzerland, Cham, pp 232–243
- Hanke T (2004) Hamnosys—representing sign language data in language resources and language processing contexts. In: *sign-lang@ LREC 2004*, European Language Resources Association (ELRA), pp 1–6
- Ishikawa K (1989) *Introduction to quality control* [internet]
- Kapitanov A, Kvanchiani K, Nagaev A, et al (2024) Hagrid – hand gesture recognition image dataset. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp 4572–4581
- Kendon A (2004) *Gesture: Visible action as utterance*. Cambridge University Press
- Maaten Lvd, Hinton G (2008) Visualizing data using t-sne. *Journal of machine learning research* 9(Nov):2579–2605
- McAtamney L, Corlett EN (1993) Rula: a survey method for the investigation of work-related upper limb disorders. *Applied ergonomics* 24(2):91–99
- Meta Platforms, Inc. (2026) Meta XR SDK. <https://developers.meta.com/horizon/documentation/unity/>, accessed: 2026-05-16
- Microsoft (2026) Mixed Reality Toolkit 3. <https://github.com/MixedRealityToolkit/MixedRealityToolkit-Unity>, accessed: 2026-05-16

- Mo GB, Dudley JJ, Kristensson PO (2021) Gesture knitter: A hand gesture design tool for head-mounted mixed reality applications. In: Proceedings of the 2021 CHI conference on human factors in computing systems, pp 1–13
- Monk E (2023) The monk skin tone scale URL <https://doi.org/10.31235/osf.io/pdf4c>
- Morris MR, Wobbrock JO, Wilson AD (2010) Understanding users' preferences for surface gestures. In: Proceedings of graphics interface 2010. p 261–268
- Mukushev M, Imashev A, Kimmelman V, et al (2020) Automatic classification of hand-shapes in russian sign language. In: Proceedings of the LREC 2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives, European Language Resources Association (ELRA)
- Nguyen R, Gouin-Vallerand C, Amiri M (2023) Hand interaction designs in mixed and augmented reality head mounted display: a scoping review and classification. *Frontiers in Virtual Reality* 4:1171230
- Nuzhdin A, Nagaev A, Sautin A, et al (2024) Hagridv2: 1m images for static and dynamic hand gesture recognition. arXiv preprint arXiv:241201508
- Ohkawa T, Furuta R, Sato Y (2023) Efficient annotation and learning for 3d hand pose estimation: A survey. *International Journal of Computer Vision* 131(12):3193–3206
- Parger M, Tang C, Xu Y, et al (2021) Unoc: Understanding occlusion for embodied presence in virtual reality. *IEEE Transactions on Visualization and Computer Graphics* 28(12):4240–4251
- Rempel D, Camilleri MJ, Lee DL (2014) The design of hand gestures for human–computer interaction: Lessons from sign language interpreters. *International journal of human-computer studies* 72(10-11):728–735
- Roumaissa B, Mohamed Chaouki B (2025) Deep learning based on hand pose estimation methods: a systematic literature review. *Multimedia Tools and Applications* pp 1–38
- Salerno F, Ulrich L, Maculotti G, et al (2026) A metrological approach for augmented reality tooltip tracking assessment. *Computers in Industry* 175:104430
- Scheirer WJ, de Rezende Rocha A, Sapkota A, et al (2012) Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence* 35(7):1757–1772
- Schneider D, Otte A, Kublin AS, et al (2020) Accuracy of commodity finger tracking systems for virtual reality head-mounted displays. In: 2020 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW), IEEE, pp 804–805

- Sellier Q, Sluÿters A, Vanderdonckt J, et al (2024) Evaluating gesture user interfaces: Quantitative measures, qualitative scales, and method. *International Journal of Human-Computer Studies* 185:103242
- Shen J, Dudley J, Mo G, et al (2022) Gesture spotter: A rapid prototyping tool for key gesture spotting in virtual and augmented reality applications. *IEEE transactions on visualization and computer graphics* 28(11):3618–3628
- Soares I, B. Sousa R, Petry M, et al (2021) Accuracy and repeatability tests on hololens 2 and htc vive. *Multimodal Technologies and Interaction* 5(8):47
- Tang G, Webb P (2018) The design and evaluation of an ergonomic contactless gesture control system for industrial robots. *Journal of Robotics* 2018(1):9791286
- Tennant RA, Brown MG (1998) *The American sign language handshape dictionary*. Gallaudet University Press
- Unity Technologies (2026) Unity editor. <https://unity.com/>, version 6.4, accessed 5/16/2026
- Xia H, Glueck M, Annett M, et al (2022) Iteratively designing gesture vocabularies: A survey and analysis of best practices in the hci literature. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29(4):1–54
- Yang F, Wu Y, Sakti S, et al (2019) Make skeleton-based action recognition model smaller, faster and better. In: *Proceedings of the 1st ACM International Conference on Multimedia in Asia*, pp 1–6
- Yeh SC, Wu EHK, Lee YR, et al (2022) User experience of virtual-reality interactive interfaces: a comparison between hand gesture recognition and joystick control for xrspace manova. *Applied Sciences* 12(23):12230
- Yuan S, Ye Q, Stenger B, et al (2017) Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 4866–4874
- Zhang S, Liu X, Xiao J (2017) On geometric features for skeleton-based action recognition using multilayer lstm networks. In: *2017 IEEE winter conference on applications of computer vision (WACV)*, IEEE, pp 148–157
- Zhang Y, Cao C, Cheng J, et al (2018) Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia* 20(5):1038–1050