Inclusive Hand-Based Region-of-Interest Selection Using On-Device Landmark Recognition

Fabian Glöggler[1,*], Markus Hahn[1,*] and Marianne von Schwerin[1]

[1]*Ulm University of Applied Sciences, Albert-Einstein-Allee 53–55, 89081 Ulm, Germany*

Abstract

Accessible camera-based assistance systems often rely on automatic scene analysis or predefined canonical hand gestures, which can limit accessibility for people with motor or sensory impairments. This paper presents *Region-of-Interest Hand Recognition* (*ROI HR*), an approach for explicit user-controlled region-of-interest (ROI) selection, in which the ROI is defined as a square window anchored above the centroid of a user-selected hand knuckle coordinate subset. The system is implemented as an edge-based artificial intelligence (AI) Android application using hand landmark detection and supports smartphones and smart glasses, including *Vuzix Shield*®, following an inclusive design approach. In contrast to canonical gesture-based interaction paradigms, *ROI HR* allows users to map stable and practically trackable hand configurations to task-specific ROIs, subject to common computer vision constraints such as occlusion, viewpoint variation, and motion blur. This supports interaction under non-standard hand postures, including clenched fists, limited finger extension, or missing fingers. All processing is performed fully on-device using the *MediaPipe Hand Landmarker*, ensuring low latency and visual privacy. The system was evaluated using a twofold methodology. A technical evaluation on a heterogeneous dataset of 1,000 hand images within the deployed application demonstrated a binary success rate of 98.9% for hand detection and landmark estimation, which does not measure ROI placement accuracy or task-level results. In addition, a formative user study with representative accessibility-oriented tasks showed high task completion rates, stable ROI positioning, and consistently positive subjective ratings on learnability, perceived control, interaction stability, and suitability for everyday use.

Keywords

Edge-based, artificial intelligence, Android application, hand recognition, smart glasses, inclusive design,

1. Introduction

This section outlines the problem context that motivates this work, defines its objective, and summarizes the methodological approach adopted in the following sections.

Definition of the problem Demographic change is leading to a steadily aging population in industrialized countries. In Germany, approximately 24% of the population is expected to be 65 years or older by 2026, with projections reaching around 30% by 2070 [1]. Aging is associated with an increased prevalence of sensory impairments and a growing demand for accessible support in everyday life. In 2021, more than 487,000 people in Germany were registered as visually impaired, in addition to more than 71,000 blind people [2]. Visual impairments often restrict independent access to information in public and commercial environments and contribute to persistent barriers to social participation and employment.

Camera-based assistive vision systems have emerged as a promising approach to address these challenges. However, many existing solutions, such as Microsoft *Seeing AI* and *Envision App*, rely on centralized cloud infrastructures for image analysis and interpretation [3, 4]. Such architectures can introduce latency, require stable network connectivity, and raise data protection concerns, thereby limiting their suitability for mobile, privacy-sensitive, and time-critical assistive use cases [5].

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops2026, March 23–26, 2026, Paphos, Cyprus
*Corresponding author.
Fabian.Gloeggler@thu.de (F. Glöggler); Markus.Hahn@thu.de (M. Hahn); Marianne.vonSchwerin@thu.de (M. von Schwerin)
https://www.thu.de/markus.hahn (M. Hahn); https://www.thu.de/marianne.vonschwerin (M. von Schwerin)

Edge-based AI enables decentralized processing on mobile platforms such as smartphones and smart glasses, offering reduced latency and improved privacy compared to cloud-based approaches [6]. Camera-based assistive vision systems typically comprise multiple processing stages, ranging from image acquisition and region selection to downstream tasks such as text recognition and semantic interpretation. Although recent advances in optical character recognition (OCR) and language models have substantially improved downstream analysis, practical experience using camera-based assistive applications such as *Seeing AI* indicates that reliable user interaction and explicit ROI selection remain critical bottlenecks in real-world assistive scenarios [7, 8, 3].

Two fundamental limitations can be observed in existing assistive systems. First, many camera-based reading aids rely on global scene descriptions or automatically selected text regions, providing limited transparency and user control over the analyzed image content. In everyday situations such as shopping or product inspection, users often require access to a specific information field rather than a full scene overview. Second, many systems presuppose intact fine motor control, for example, through canonical pointing gestures or precise touch input. Older adults and people with neurological conditions may experience tremor, spasticity, or contractures that hinder such patterns of interaction. Spasticity can result in stable but non-canonical hand postures, including clenched fists or *thumb-in-palm* configurations, which are poorly supported by gesture-based interfaces [9].

Although camera-based assistive vision systems are primarily motivated by visual impairments, their practical usability critically depends on the ability of users to interact with and control the system. In this work, the primary target population comprises users with motor impairments of the hand, including older adults and individuals with neurological conditions, for whom conventional interaction techniques such as precise touch input or canonical pointing gestures are often inaccessible. Visual assistance tasks such as reading or scene interpretation are considered downstream components that depend on, but are not addressed by, the proposed interaction approach.

Objective of the work Addressing interaction-related limitations in camera-based assistive vision systems, the objective of this work is to design and implement an interaction-focused front-end component for users with hand motor impairments. The proposed Android application *ROI HR* provides explicit, user-controlled, hand-based ROI selection using decentralized AI methods and is intended to operate as an upstream interaction module preceding downstream analysis stages such as OCR or task-specific reasoning.

By treating ROI selection as an explicit user-controlled interaction step, the proposed approach aims to improve transparency, controllability, and adaptability for users with various motor constraints. Exemplary downstream applications include targeted analysis of product packaging, such as identifying best-before dates within selected image regions, as demonstrated in previous work [10]. Rather than claim end-to-end accessibility outcomes, the proposed method is positioned as an interaction-focused front-end component that may facilitate independent and privacy-preserving access to visual information when integrated into modular assistive pipelines.

Approach of the work The remainder of this paper is structured as follows. Section 2 reviews related work on assistive vision systems. Section 3 describes the proposed Android application and the underlying methods. Section 4 presents and discusses the experimental results. Section 5 concludes the paper and describes directions for future work.

2. State of the art

Camera-based assistance for low-vision and blind users has advanced substantially in recent years. Smartphone applications such as Microsoft *Seeing AI* and the *Envision App*, as well as wearable systems such as *OrCam*, enable text reading and visual recognition using device cameras [3, 4, 11]. However, these systems typically analyze the full camera view or automatically selected text regions. In visually complex scenes such as product packaging with multiple text fields, this can lead to ambiguous or non-transparent region selection from the user perspective. In addition, interaction is usually limited to predefined and canonical usage patterns, which offer little flexibility for individual motor abilities or alternative interaction strategies. Recent generative AI systems further improve image understanding and OCR, but solutions such as *ChatGPT* and *Gemini* usually rely on cloud-based processing. This introduces latency and raises privacy concerns when sensitive visual data must be transmitted externally [5]. Edge-based AI addresses these limitations by enabling real-time, on-device processing with reduced dependence on network connectivity [6].

Hand detection and hand gesture recognition are well-established topics in computer vision research, particularly in human-computer interaction and augmented reality (AR). Previous work has demonstrated robust real-time gesture recognition and large-scale hand image classification using learning-based methods on diverse datasets [12, 13]. Beyond availability and ease of integration, the selection of a pose estimation framework is influenced by empirical performance comparisons regarding detection accuracy and runtime behavior. A comparative evaluation of commonly used human pose estimation libraries reported that lightweight models such as *MoveNet* and *MediaPipe Pose* achieve competitive accuracy while maintaining higher robustness under challenging conditions, including self-occlusion and suboptimal camera viewpoints, while *OpenPose* showed reduced stability, particularly in video-based evaluation scenarios [14]. These findings illustrate the trade-off between generality and deployment efficiency, as multi-person frameworks often prioritize coverage over computational efficiency. In parallel, recent *YOLO*-based pose estimation approaches (You Only Look Once) such as *YOLOv8* pose variants, demonstrate strong accuracy and favorable runtime characteristics on standard pose benchmarks by incorporating architectural optimizations and improved loss formulations [15]. While these methods are well suited for full-body pose estimation in controlled environments, they primarily target large-scale detection scenarios and typically require higher computational resources. This limits their applicability for continuous on-device hand interaction on mobile hardware.

Efficient on-device frameworks such as Google's *MediaPipe Hand Landmarker* enable real-time hand landmark detection on mobile devices and are therefore widely adopted. The underlying approach combines a palm detector with a regression-based landmark model and achieves high accuracy at interactive frame rates on mobile hardware [16, 17]. By supporting explicit and continuous hand pose tracking rather than discrete gesture classification, such frameworks enable fine-grained and user-controlled interaction. This contrasts with many commercial assistive systems that rely on implicit region selection or a limited set of predefined gestures and, therefore, offer reduced adaptability to individual motor constraints. Landmark-driven interaction with explicit ROI control and fully on-device processing thus enables more independent access to visual information.

Beyond technical considerations, social participation remains a significant challenge for people with disabilities. In Germany, the employment participation rate of people with disabilities aged 15 to 64 years was approximately 58% in 2021, compared to about 82% for people without disabilities [18]. Labor market statistics also indicate persistent unemployment among severely disabled individuals, showing that barriers to participation extend beyond technical factors [19].

By combining explicit user-controlled ROI selection with decentralized on-device processing and adaptable hand-based interaction, this work is intended to support more transparent, privacy-aware, and flexible interaction in camera-based assistive vision systems.

3. Material and methods

This section describes the general context and workflow of the system in Section 3.1, the design and implementation of the Android application in Section 3.2, and the evaluation methodology in Section 3.3.

This work presents an edge-based Android application named *ROI HR*. The system enables explicit, user-controlled selection of an ROI based on a configurable subset of hand knuckle coordinates. The application is designed to run entirely on mobile devices and supports both Android smartphones and smart glasses, specifically the *Vuzix Shield*®. Its functional behavior is identical across smartphones and smart glasses, with differences limited to the display modality and interaction context.

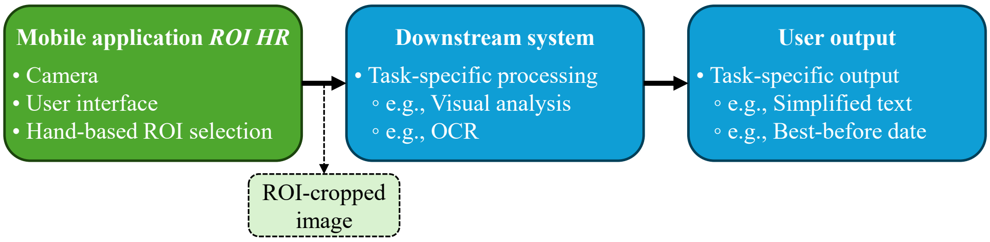


Figure 1: Conceptual overview of an edge-based assistive reading workflow with components highlighted in green indicating the scope of this paper.

Although camera-based assistive vision systems are often motivated by visual impairments, the practical usability of such systems critically depends on the user's ability to interact with and control the interface. Consequently, the primary target population of *ROI HR* comprises users with hand motor impairments, including older adults and individuals with neurological conditions, for whom conventional interaction techniques such as precise touch input or canonical pointing gestures are frequently inaccessible. Visual assistance tasks such as reading support or scene interpretation are considered downstream components that may benefit from explicit ROI selection but are not implemented or evaluated within the scope of the presented system.

3.1. System overview

The application is embedded in a modular, edge-based assistive reading workflow. Figure 1 illustrates this workflow and indicates the scope of the proposed contribution. As highlighted in Figure 1, the components shown in green correspond to the scope of this paper and focus exclusively on hand-based ROI selection. All remaining components represent potential downstream systems that can consume cropped ROI but are not implemented or evaluated in this work. The device's camera performs image acquisition, which provides a continuous image stream for further processing. Within this stream, the *ROI HR* application enables explicit hand-based interaction by detecting hand landmarks on-device and computing a user-controlled ROI based on configurable hand knuckle coordinates. The resulting ROI-cropped image serves as the interface to subsequent processing stages, which may include task-specific visual analysis such as OCR or information extraction, followed by presentation of results in an accessible output form.

3.2. Application concept

The application was developed using *Android Studio* and implemented in the *Java* programming language. The project follows a modular structure that separates camera handling, hand landmark detection, ROI computation, user interface rendering, and configuration management. All components are packaged within a standalone Android application archive and are executed locally on the device. *ROI HR* provides a live camera preview augmented with a real-time hand landmark overlay and a visual representation of the computed ROI. The overlay indicates the image region that will be cropped and forwarded for further processing. If a valid ROI is available, camera focus is prioritized in the central area of the ROI. Otherwise, the standard continuous autofocus is applied over the full image.

Figure 2 shows the smartphone interface, including the main view with camera preview, a visualization of hand landmarks rendered in yellow, all 21 detected hand knuckle coordinates shown in orange, the user-selected coordinates highlighted in green, the centroid of the selected coordinates used for ROI positioning, and the resulting ROI box rendered in blue. The figure also includes the settings interface used for the configuration.

Figure 3 illustrates the same application workflow in the *Vuzix Shield*® AR smart glasses. The figure shows two separate screenshots corresponding to the rendered camera image and the overlay output. In actual use, the wearer perceives a single visually merged AR view in which the camera image and overlays are spatially aligned with the real world. This separation is used solely for documentation

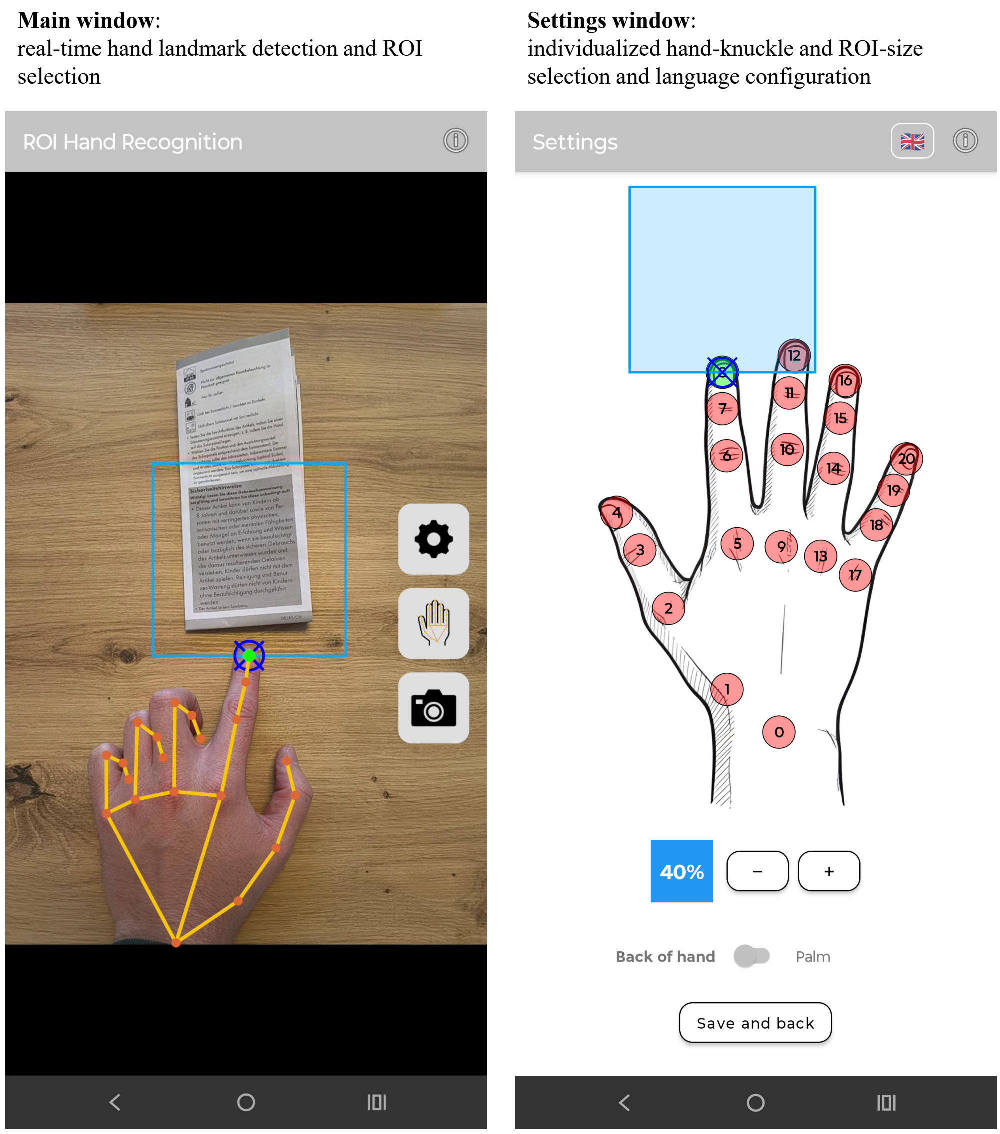


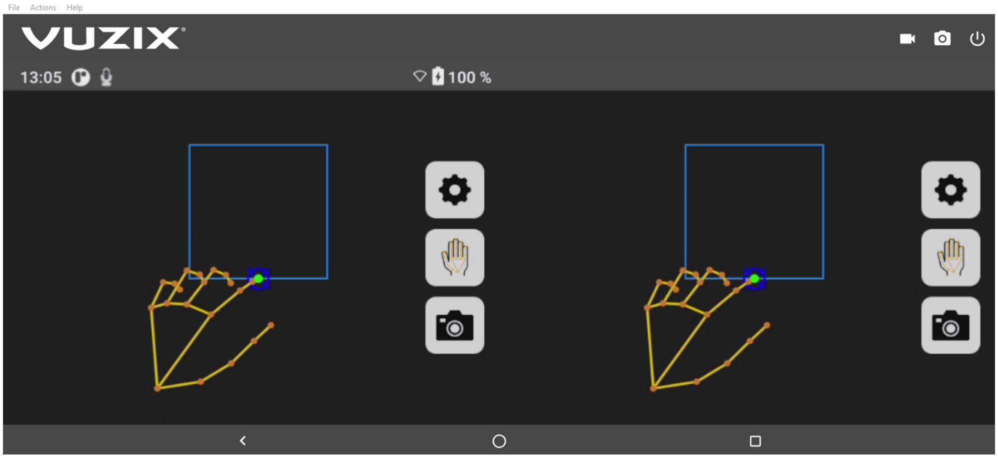
Figure 2: Frontend of the Android application *ROI HR* running on *motorola edge 60 pro* smartphone.

purposes. The configuration can be performed by either the user or, initially, by a supporting person. Once configured, the profile is stored locally and reused across sessions. This design supports hands-free interaction scenarios and adapts to individual impairment profiles.

Settings window The settings interface allows users to select which of the 21 detected hand knuckle coordinates are used for ROI computation. ROI size can be adjusted as a percentage of the image dimension in discrete ten percent steps. The visualization of the hand can be changed between the palm and dorsal views, and the language of the app can be changed between German and English. The landmark selection is independent of the left or right hand and of the palm or dorsal orientation. To ensure a stable and unambiguous mapping, the application processes only one hand at a time.

Hand landmark detection with *MediaPipe Hand Landmarker* Hand tracking is implemented using the *MediaPipe Hand Landmarker* framework. The deployed model bundle integrates palm detection and hand landmark estimation and is loaded from the task file *hand_landmarker.task*. The *HandLandmarker* model is executed in *LIVE_STREAM* mode using the CPU delegate. In this mode, palm detection is only re-triggered when tracking confidence drops below a defined threshold, improving runtime efficiency on mobile hardware.

Main window: real-time hand landmark detection and ROI selection



Settings window: individualized hand-knuckle and ROI-size selection and language configuration

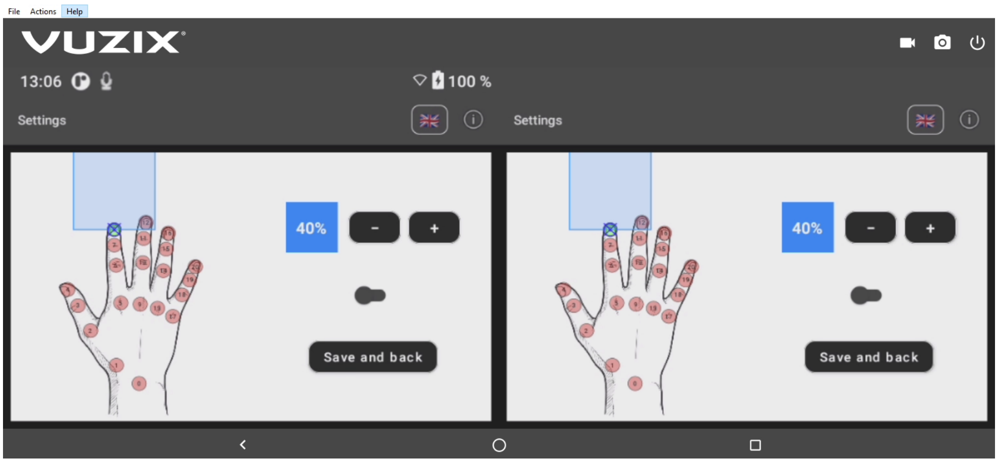


Figure 3: Frontend of the Android application *ROI HR* running on *Vuzix Shield*® smart glasses.

The *MediaPipe Hand Landmarker* is a composite model bundle consisting of a palm detection network and a hand landmark estimation network. The palm detector is based on a convolutional neural network with a single-shot detection architecture, while the hand landmark model employs a convolutional regression architecture to estimate precise landmark locations. For each detected hand, the model predicts 21 two-dimensional landmark coordinates following the standard *MediaPipe* indexing scheme, with landmark 0 representing the wrist, landmarks 1 to 4 the thumb, 5 to 8 the index finger, 9 to 12 the middle finger, 13 to 16 the ring finger, and 17 to 20 the little finger.

As defined by the architecture of the *MediaPipe Hand Landmarker*, palm detection operates on inputs of *192* × *192* pixels, while hand landmark estimation uses *224* × *224* pixel inputs. These resolutions are fixed by the model design. Input images are automatically resized by the framework, and explicit input resolution is not specified in the application code. The model bundle uses *float16* quantization. The minimum hand detection confidence, the minimum tracking confidence, and the minimum hand presence confidence are not explicitly set and therefore remain at the default value of *0.5* as defined by the *MediaPipe Hand Landmarker* API. The maximum number of simultaneously detected hands is limited to *1* by default, regardless of whether the detected hand is left or right. [16]

ROI computation from selected hand knuckle coordinates A selected subset of the hand knuckle coordinate indices of the set $\{0, \dots, 20\}$ is used for the ROI calculation. For each frame with valid hand detection, the centroid of the selected landmark coordinates is computed. A square ROI is then defined, with its side length determined by the configured size percentage relative to the shorter image dimension. In the current implementation, the ROI is positioned above the centroid, with the centroid corresponding to the bottom center of the square. This placement supports pointing above the hand, while keeping the hand itself outside the cropped region. If no valid hand landmarks are detected in a given frame, no ROI can be derived, and the application falls back to processing the full area of the image. This ensures continuous operation even in the absence of a detectable hand and prevents undefined or unstable ROI behavior.

On-device analysis and exported outputs All hand detection, landmark processing, and ROI computation are performed fully on-device. The application was deployed and tested on two representative Android-based devices that cover handheld and head-worn usage scenarios, namely the *motorola edge 60 pro* smartphone and the *Vuzix Shield®* smart glasses. The smartphone provides a high-resolution rear camera and sufficient on-device processing capabilities to support real-time execution of the *MediaPipe Hand Landmarker* in live stream mode [20]. The *Vuzix Shield®* features an integrated forward-facing camera, a transparent waveguide display, and an Android-based processing unit optimized for continuous vision workloads and hands-free interaction with persistent visual overlays [21].

In addition to live camera operation, the application provides an upload and analyze mode for single images. For each analyzed image, the application stores the original input, an annotated version with rendered hand landmarks, and the cropped ROI. The cropped output can be processed later by external assistive systems or analysis pipelines that are not part of this work.

3.3. Evaluation procedure

The proposed system was evaluated using a twofold methodology. A technical on-device evaluation assessed the robustness of hand detection and landmark availability in static images under heterogeneous visual and morphological conditions, as described in Section 3.3.1. In addition, a formative user study examined the learnability, perceived controllability, comfort, fatigue, and accessibility of ROI selection based on landmarks under varying hand conditions, as detailed in Section 3.3.2.

The evaluation deliberately focuses on hand availability as a prerequisite for ROI computation and on the usability of the interaction concept. Task-level performance metrics such as OCR accuracy, end-to-end task completion time, or video-based temporal stability are beyond the scope of this work and are discussed as directions for future research.

3.3.1. Technical evaluation

The technical evaluation examines the robustness of the detection of hand signals on the device and landmark estimation in diverse image conditions and hand appearances. For this purpose, a test dataset comprising 1,000 static hand images was constructed. At the time of this study, no publicly available dataset explicitly covered clinically annotated spastic or deformed hands suitable for on-device evaluation. Therefore, images from multiple public hand datasets were combined with a small number of recordings captured by the authors.

The dataset was designed to achieve broad variability in hand orientation (palmar and dorsal), viewpoint, image resolution, background complexity, and gesture morphology. Particular emphasis was placed on compact and flexed hand configurations that resemble reduced finger extension and *thumb-in-palm*-like postures, which represent relevant stress cases for the intended target group.

The final dataset comprises 1,000 images and is provided in the supplementary material in folder *1_Dataset*, with contributions from the following sources:

- 17 images recorded by the authors
- 316 images from *11K Hands* [22]
- 300 images from *FreiHAND* [23]
- 367 images from the *Hand gestures dataset* [24]

Dataset *11K Hands* *11K Hands* provides high-resolution hand images captured under controlled conditions with systematic variation in the hand side and palmar or dorsal orientation. In the constructed test set, this dataset primarily contributes viewpoint-consistent samples that span open-hand to closed-hand configurations, allowing controlled evaluation of landmark detection stability across similar capture geometries. [22]

Dataset *FreiHAND* *FreiHAND* contributes substantial variability in appearance, background, and imaging conditions that is closer to in-the-wild scenarios. Although the provided 3D annotations are not used, the dataset supports the evaluation of hand detection robustness under less constrained visual conditions. [23]

Dataset *Hand gestures dataset* The *Hand gestures dataset* contains low-resolution images grouped into gesture categories such as closed fist, bent fingers, and open palm. These compact and occlusion-prone finger configurations approximate reduced finger extension and *thumb-in-palm*-like postures and increase morphological diversity in the test set. [24]

On-device evaluation procedure All 1,000 images were analyzed directly within the Android application *ROI HR* on a *motorola edge 60 pro* smartphone using an upload-and-analyze mode. This ensures that the evaluation reflects the complete on-device deployment pipeline, including model inference and landmark post-processing.

The technical evaluation focuses exclusively on successful hand detection and landmark availability. ROI placement accuracy and task-level outcomes were not explicitly measured, as ROI computation in the proposed system is deterministic once valid landmarks are available. Consequently, the availability of landmarks is considered a necessary condition for the selection of landmark-based ROI. The evaluation output is limited to the rendered hand landmark overlay.

3.3.2. User study

In addition to the technical evaluation, a formative user study was conducted to assess the usability and accessibility of ROI selection based on landmarks. The study focused on learnability, perceived controllability, interaction stability, physical fatigue, and perceived suitability for everyday use.

Ten participants with heterogeneous backgrounds in age, smartphone experience, and hand mobility participated in the study. To approximate a range of motor constraints relevant to the target group, several participants performed tasks with simulated hand impairments, including clenched fists, limited finger extension, and missing fingers. Both left- and right-handed interaction was represented. The study followed a within-subject design and was conducted exclusively for research purposes. No personal data were collected, and all responses were anonymously evaluated. The user questionnaires were provided in both German and English and are included in the supplementary material (folder *3_EvaluationUserStudy*).

Participants interacted with the *ROI HR* application on a smartphone and completed a sequence of three representative interaction tasks. The study explicitly focused on the interaction process rather than the performance of the downstream task and did not include automatic text recognition or image analysis.

- **Task 1: Configuration of hand landmarks.** Participants selected a personalized subset of hand knuckle coordinates in the application settings to control ROI, aiming to identify a configuration that felt comfortable and controllable.
- **Task 2: Selection of a newspaper column.** Using the configured landmarks, the participants selected a column of text on a printed newspaper.
- **Task 3: Selection of a best-before date.** Participants used the application to select the best-before date on a product package.

Task-level interaction was assessed using binary success indicators. For Task 1, the successful completion and the need for assistance were recorded. For Tasks 2 and 3, participants indicated whether the ROI covered the intended target and whether the positioning was sufficiently stable or precise. These measures are reported as the percentage of affirmative “yes” responses among participants.

Subjective user experience was assessed using a post-study questionnaire consisting of six statements rated on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree). The statements addressed learnability, perceived control, understandability of landmark configuration, interaction stability, physical fatigue, and suitability for everyday use. For these measures, the mean and standard deviation were calculated across participants. In addition, two open-ended questions collected qualitative feedback on aspects that participants particularly liked about the interaction and elements they found unclear or tiring.

4. Results and discussion

This section presents and discusses the results of the proposed landmark-based ROI selection system with respect to its intended use in accessibility-oriented mobile interaction. Qualitative task-driven examples of selection based on landmarks are presented in Section 4.1 to illustrate how different hand configurations enable user-specific interaction. Quantitative results on the robustness of on-device hand detection and landmark estimation are reported in Section 4.2 and discussed with respect to practical deployment on mobile devices. Empirical findings from a formative user study are presented and discussed in Section 4.3.

4.1. Task-driven ROI selection examples

Figure 4 illustrates the task-driven selection of ROI based on different subsets of hand landmarks. Each example demonstrates how the proposed approach enables task-specific ROI placement without relying on predefined or canonical hand gestures.

In the newspaper reading scenario shown in Example (a), the ROI is positioned to cover a user-selected text column. This enables selective magnification or further automated processing, which is particularly beneficial for users with visual impairments or reduced reading ability. The extracted ROI can be forwarded to downstream systems, such as text-to-speech engines or language simplification pipelines, supporting the comprehension of dense or complex documents.

Example (b) focuses on accessing product information, such as ingredient lists or usage instructions, printed in small fonts and dense layouts. For many people with disabilities, reading this information is difficult due to limited legibility and visual clutter. By defining ROI through empirically selected hand knuckle coordinates, relevant text regions can be isolated independently of hand shape or finger availability. The extracted content can then be processed by additional systems, for example, to summarize ingredients, highlight allergens, or provide explanations in simplified language.

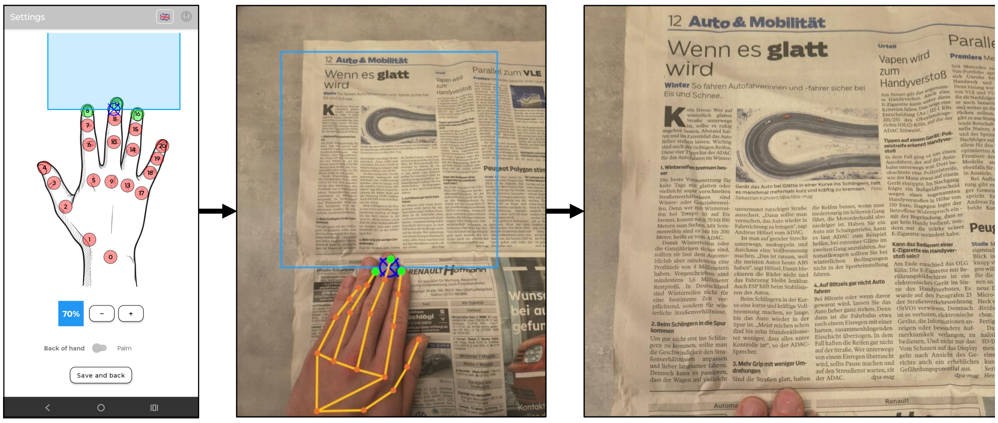
The extraction of the best-before date (BBD) is illustrated in Example (c). Correctly identifying BBDs is critical, as failure to do so may pose health risks. Since BBDs are often printed in small fonts and appear in varying locations on packaging, task-specific ROI selection provides a direct and reliable means to isolate the relevant region for further automated recognition and accessible presentation.

Beyond the illustrated examples, the proposed approach supports a wide range of everyday scenarios, including form completion, instruction reading, and recipe assistance. In this sense, the system functions as a flexible front end for assistive visual processing rather than as a task-specific solution.

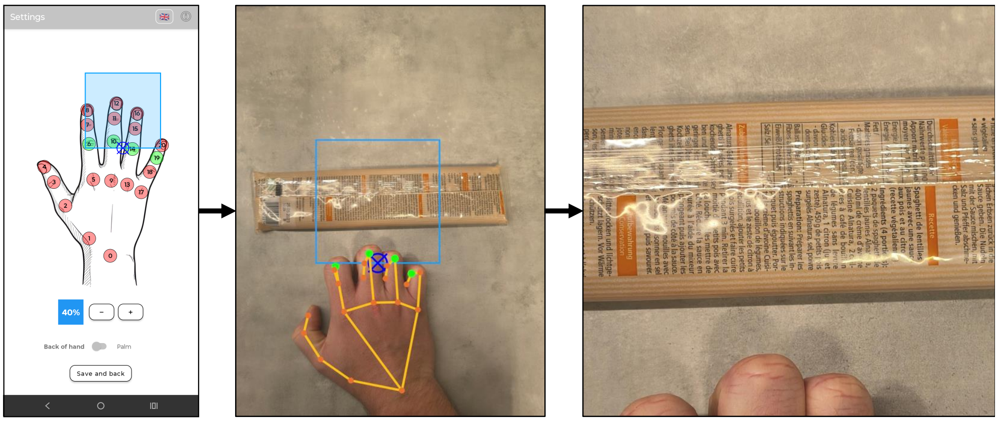
4.2. Hand detection performance on the 1,000-image test set

Across the 1,000-image test dataset, *ROI HR* successfully detected a hand and estimated landmarks in 989 images, corresponding to a detection success rate of **98.9%**. The remaining 11 failure cases are shown in Figure 5. Missed detections were mainly caused by extreme viewpoints, strong occlusions, or motion blur, which reduced the visibility of hand structures required for reliable landmark estimation.

Example (a): Newspaper article reading using hand-knuckle coordinates 8, 12, and 16



Example (b): Product information selection using hand-knuckle coordinates 6, 10, 14, and 19



Example (c): Best-before date reading using hand-knuckle coordinates 6, and 10

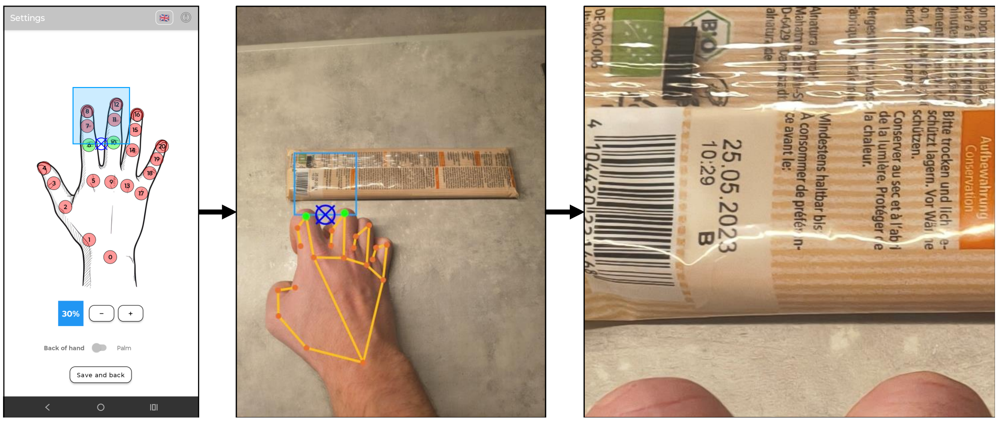


Figure 4: Task-specific ROI selection examples based on hand landmark subsets: newspaper reading (a), product information access (b), and best-before date extraction (c).

In practical use, such failures can be mitigated through capture guidance and by configuring landmark subsets that remain visible under constrained or atypical hand postures. Importantly, incomplete or partially unstable landmark estimation does not necessarily render the system unusable. Since the mapping from landmarks to ROI is defined empirically for each user, the interaction remains functional even in the presence of anatomical differences or amputations. For example, users who do not have one or more fingers can define a closed fist as a stable reference configuration for ROI control.

Figure 6 shows representative successful detections of the test set in the form of a 5×5 matrix subset. Each image overlays detected hand landmarks and the corresponding knuckle coordinates, providing visual confirmation of the stable landmark estimation. The examples cover a wide range

Figure 5: Missed hand detections: 11 failure cases out of 1,000 test images.

of hand appearances, orientations, and background conditions. Consistent placement of landmarks across the matrix demonstrates robust detection when sufficient visual information is available, which is essential for subsequent landmark-based ROI computation and task-driven interaction.

The principle of interaction changes from smartphone-based use to smart glasses, as illustrated in Figures 2 and 3. Although smartphones require explicit device handling, smart glasses enable hands-free operation. In this setting, head motion directly affects the camera stream, making stable landmark selection and careful ROI size configuration particularly important.

A central strength of the proposed approach is its high degree of **customizability**. Arbitrary hand configurations can be assigned to task-specific ROIs through user-defined landmark subsets. This enables interaction even when standard pointing gestures are not possible, for example, due to missing fingers, limited extension, or clenched fists. As a result, users can perform task-specific interactions independently using a smartphone or smart glasses without relying on external assistance.

The evaluation outputs for hand detection are provided in the supplementary material (folder *2_EvaluationHandDetection*).

4.3. User study results

The formative user study provides empirical insight into the usability and perceived accessibility of ROI selection based on landmarks.

All ten participants successfully completed Task 1 (landmark configuration), resulting in a task completion rate of 100%. Assistance was required in 60% of cases, primarily during the initial exploration of the landmark configuration options. This indicates that while the configuration process is learnable, additional onboarding or visual guidance could further improve first-time usability. For Task 2 (newspaper column selection), the ROI covered the intended target in 100% of cases, and stable positioning was reported in 90% of the cases. For Task 3 (best-before date selection), both ROI coverage and positioning precision reached 100% affirmative responses. These results indicate that once configured, landmark-based ROI selection enables reliable targeting of both large regions and small visual details, even under simulated motor impairments.

Figure 6: Successful hand detections: representative 5×5 subset from the 1,000-image test dataset.

Likert-scale ratings reveal consistently positive subjective evaluations. Learnability achieved a mean score of 4.30 with a standard deviation σ of 0.48, indicating that participants generally found the interaction easy to learn with high agreement between users. Perceived control over ROI positioning reached a mean of 4.10 ($\sigma = 0.74$), reflecting a positive assessment with moderate inter-individual variability. The understandability of the landmark configuration received a mean rating of 4.40 ($\sigma = 0.52$), suggesting that the underlying concept of interaction was well understood after a short familiarization period. The stability of the interaction was rated with a mean of 4.10 ($\sigma = 0.74$), indicating a generally stable behavior while also reflecting differences in hand posture, simulated impairment type, and selected landmark subsets. Physical fatigue received a low mean score of 1.60 ($\sigma = 0.70$), suggesting that the interaction was perceived as largely non-fatiguing. The suitability for daily use achieved a mean rating of 4.40 ($\sigma = 0.70$), indicating a strong general acceptance of the interaction concept.

In general, the reported mean values indicate consistently positive subjective ratings in all aspects evaluated. The associated standard deviations are generally low to moderate, showing substantial agreement among the participants despite heterogeneous hand conditions and simulated impairments. Slightly higher variability for perceived control and interaction stability reflects individual differences in hand mobility and preferred landmark configurations, rather than systematic usability issues. These findings underscore the importance of allowing users to customize the subset of landmarks used for ROI control.

Detailed results of the user study are provided in the supplementary material (folder *3_EvaluationUser-Study*).

5. Conclusion

This paper introduced *ROI HR*, an edge-based AI system for explicit user-controlled ROI selection on mobile and wearable devices. The proposed approach defines a square ROI deterministic anchored above the centroid of a user-selected subset of hand knuckle coordinates and is implemented as an Android application that supports smartphones and smart glasses, including *Vuzix Shield*®. In contrast to camera-based assistance systems that rely on automatic scene analysis or predefined canonical hand gestures, *ROI HR* follows an inclusive and individualize interaction paradigm. Users can map stable and practically trackable hand configurations to task-specific ROIs, enabling interaction under non-standard hand postures such as clenched fists, limited finger extension, or missing fingers.

All processing is performed fully on-device using the *MediaPipe Hand Landmarker*, enabling low-latency interaction while preserving visual privacy by avoiding cloud-based image transmission. This design choice is particularly relevant for assistive applications that operate on sensitive visual content. A technical evaluation conducted directly within the deployed application on a heterogeneous dataset of 1,000 hand images demonstrated a binary success rate of 98.9% for hand detection and landmark estimation. Although this metric does not assess ROI placement accuracy or task-level performance, it indicates robust landmark availability under realistic conditions, which is a prerequisite for reliable landmark-based ROI interaction. In addition to the technical evaluation, a formative user study with representative accessibility-oriented tasks provided empirical insights into usability and perceived accessibility. Participants achieved high task completion rates and reported stable ROI positioning across tasks that involved large regions and small visual details. The subjective ratings also indicated positive assessments of learnability, perceived control, stability of the interaction, and suitability for everyday use. Together, these findings suggest that ROI selection based on landmarks constitutes a viable and accessible interaction technique for camera-based assistive systems on mobile devices.

To support reproducibility and further research, the dataset, detailed evaluation results, and an archived version of the *ROI HR* source code are provided as supplementary material in the ACM Digital Library. The supplementary material is organized into the following folders: *1_Dataset*, *2_EvaluationHandDetection*, *3_EvaluationUserStudy*, and *4_Software_ROI_HR*. In addition, *ROI HR* software is available as open-source under the *MIT* license on GitHub (https://github.com/fgloeggler/ROI-HR) to support ongoing development and community contributions.

Despite promising results, several limitations remain. The user study was exploratory in nature and involved a limited number of participants with simulated impairments. Furthermore, the evaluation did not quantify ROI placement accuracy in pixel space, temporal stability in continuous video streams, or task-level performance gains when combined with downstream components such as OCR or task-specific information extraction, for example, identifying BBDs within selected image regions, as demonstrated in previous work [10].

Future work will therefore focus on clinical validation with participants with diagnosed motor impairments, longitudinal and video-based stability analyses, runtime evaluations across different device classes, and end-to-end assessments of assistive tasks such as reading support and information extraction. Beyond technical extensions, the integration of additional sensing modalities and hardware platforms represents a further step toward reducing interaction barriers and increasing accessibility in everyday camera-based assistance.

Acknowledgments

This work was conducted at Ulm University of Applied Sciences within the interdisciplinary *MIKE* project (Ger. ***M**ehr **I**nklusion durch smarte **K**I auf **E**ndgeräten*, en. *more inclusion through smart AI on end devices*), funded by the *Federal Ministry of Research, Technology and Space* (*BMFTR*) as part of the *DATIpilot* initiative. The project is conducted by the *EdgeAI* research group and focuses on the development of user-friendly edge AI systems to promote inclusion, particularly for people with disabilities in everyday life and work.

Declaration on Generative AI

During the preparation of this work, the authors made limited use of *ChatGPT 5.2* to support the drafting process, primarily to overcome writer's block and to help with the formulation of definitions and expressions. All content generated with tool support was critically reviewed, evaluated, and refined by the authors to ensure scientific accuracy, originality, and integrity. In addition, *LanguageTool* was used for grammar and spelling checks, and *DeepL Translator* was used for translation purposes. The authors assume full responsibility for the final content of the manuscript, including its accuracy, originality, and compliance with scientific standards.

References

[1] Statistisches Bundesamt, 15. koordinierte Bevölkerungsvorausberechnung in Deutschland: Variante 1: Moderate Entwicklung der Geburtenhäufigkeit und Lebenserwartung bei niedrigem Wanderungssaldo (G2L2W1), Online, 2025. URL: https://service.destatis.de/bevoelkerungspyramide, accessed on November 12, 2025.

[2] Deutscher Blinden- und Sehbehindertenverband e. V., Zahlen und Fakten zu Blindheit und Sehbehinderung, Online, 2025. URL: https://www.dbsv.org/zahlen-fakten.html, accessed on January 14, 2025.

[3] Microsoft, Seeing AI: An app for visually impaired people that narrates the world around you, Online, 2025. URL: https://www.microsoft.com/en-us/garage/wall-of-fame/seeing-ai/, accessed on November 2, 2025.

[4] Envision Technologies, Envision App, Online, 2025. URL: https://www.letsenvision.com/app, accessed on April 17, 2025.

[5] Rane, Nitin and Choudhary, Saurabh and Rane, Jayesh, Gemini Versus ChatGPT: Applications, Performance, Architecture, Capabilities, and Implementation, SSRN Electronic Journal (2024). doi:http://dx.doi.org/10.2139/ssrn.4723687.

[6] Raghubir Singh and Sukhpal Singh Gill, Edge AI: A survey, Internet of Things and Cyber-Physical Systems 3 (2023) 71–92. doi:https://doi.org/10.1016/j.iotcps.2023.02.004.

[7] Geewook Kim and Teakgyu Hong and Moonbin Yim and Jeongyeon Nam and Jinyoung Park and Jinyeong Yim and Wonseok Hwang and Sangdoo Yun and Dongyoon Han and Seunghyun Park, OCR-free Document Understanding Transformer, 2022. doi:https://doi.org/10.48550/arXiv.2111.15664.

[8] Haoran Wei and Yaofeng Sun and Yukun Li, DeepSeek-OCR: Contexts Optical Compression, 2025. doi:https://doi.org/10.48550/arXiv.2510.18234.

[9] Yulia Rivelis and Nowera Zafar and Karen Morice, Spasticity, in: StatPearls, StatPearls Publishing, Treasure Island (FL), 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK507869/, updated on August 8, 2023.

[10] F. Glöggler, M. Hahn, N. von Hohenzollern, E. Khademi, L. Litzkow, M. von Schwerin, Domain-adapted text recognition for best-before dates: Integrating synthetic and real data into an edge-based ai system for reading best-before dates to support people with visual impairments, in: Embedded Software Engineering Kongress 2025, ELEKTRONIKPRAXIS, Vogel Communications Group GmbH & Co. KG and MicroConsult Academy GmbH, Würzburg and Munich, Germany, 2025, pp. 421–433.

[11] OrCam, OrCam - die revolutionäre Vorlesekamera, Online, 2025. URL: https://orcam.de/de/willkommen, accessed on April 17, 2025.

[12] Kapitanov, Alexander and Kvanchiani, Karina and Nagaev, Alexander and Kraynov, Roman and Makhliarchuk, Andrei, HaGRID – HAnd Gesture Recognition Image Dataset, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4572–4581. doi:https://doi.org/10.48550/arXiv.2206.08219.

[13] Sahoo, Jaya Prakash and Prakash, Allam Jaya and Pławiak, Paweł and Samantray, Saunak, Real

Time Hand Gesture Recognition Using Fine Tuned Convolutional Neural Network, Sensors 22 (2022). doi:https://doi.org/10.3390/s22030706.
[14] Chung, Jen-Li and Ong, Lee-Yeng and Leow, Meng-Chew, Comparative Analysis of Skeleton-Based Human Pose Estimation, Future Internet 14 (2022) 380. doi:https://doi.org/10.3390/fi14120380.
[15] Xu, Xunqian and Wu, Tao and Du, Zhongbao and Rong, Hui and Wang, Siwen and Li, Shue and Chen, Dakai, Enhanced Human Pose Estimation Using YOLOv8 with Integrated SimDLKA Attention Mechanism and DCIOU Loss Function, PLOS ONE 20 (2025) 1–16. doi:https://doi.org/10.1371/journal.pone.0318578.
[16] Google, Hand landmarks detection guide, Online, 2025. URL: https://ai.google.dev/edge/mediapipe/solutions/vision/hand_landmarker, accessed on December 17, 2025.
[17] Zhang, Fan and Bazarevsky, Valentin and Vakunov, Andrey and Tkachenka, Andrei and Sung, George and Chang, Chuo-Ling and Grundmann, Matthias, MediaPipe Hands: On-Device Real-Time Hand Tracking, 2020. doi:https://doi.org/10.48550/arXiv.2006.10214. arXiv:2006.10214.
[18] Statistisches Bundesamt (Destatis), Mikrozensus - Lebenslagen der behinderten Menschen: Endergebnisse 2021, 2024. URL: https://www.destatis.de/DE/Themen/Querschnitt/Grundgesetz/behinderte_menschen.html, accessed on December 17, 2025.
[19] Klaus, Anton and Deyerler, Yvonne, Arbeitsmarktsituation schwerbehinderter Menschen 2024, 2025. URL: https://statistik.arbeitsagentur.de/DE/Statischer-Content/Statistiken/Themen-im-Fokus/Menschen-mit-Behinderungen/generische-Publikation/Arbeitsmarktsituation-schwerbehinderter-Menschen.pdf, accessed on December 17, 2025.
[20] Motorola Mobility LLC, motorola edge 60 pro, Online, 2025. URL: https://www.motorola.com/at/de/p/phones/motorola-edge/edge-60-pro/pmipmht41mf, accessed on December 17, 2025.
[21] Vuzix Corporation, Vuzix Shield® Smart Glasses, Online, 2025. URL: https://www.vuzix.com/products/vuzix-shield-smart-glasses, accessed on April 17, 2025.
[22] M. Afifi, 11K hands: Gender recognition and biometric identification using a large dataset of hand images, Multimedia Tools and Applications (2019). doi:https://doi.org/10.48550/arXiv.1711.04322.
[23] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, T. Brox, FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019. doi:https://doi.org/10.48550/arXiv.1909.04349.
[24] R. Giridhar, Hand gestures dataset, Online, 2022. doi:https://doi.org/10.34740/kaggle/ds/2206851, kaggle dataset.