Beyond Accuracy: Auditing Allocative Harms in Facial-Gesture Recognition for People with Motor Impairments

Siyu Zhang
School of Computer Science
University of Bristol
Bristol, United Kingdom
s.zhang@bristol.ac.uk

Yelu Gu
School of Engineering Mathematics and Technology
University of Bristol
Bristol, United Kingdom
yelu.gu@bristol.ac.uk

Kirsten Cater
School of Computer Science
University of Bristol
Bristol, United Kingdom
kirsten.cater@bristol.ac.uk

Oussama Metatla
School of Computer Science
University of Bristol
Bristol, United Kingdom
o.metatla@bristol.ac.uk

Abstract

Camera-based facial-gesture interfaces offer hands-free access for people with motor impairments (PwM), yet most recognition models are trained on able-bodied data and implicitly assume normative motor control and proprioception. We conducted a mixed-methods empirical study of 37 above-neck gestures performed by 11 PwM and 11 non-impaired participants. Results reveal systematic mismatches between user intention and model recognition in the PwM group, stemming from diverse patterns of body perception and control and leading to allocative harms. These mismatches concentrated in low-amplitude, asymmetric, and directional gestures. Building on these findings, we introduce FairGesture, a diagnostic auditing method for quantifying and interpreting such mismatches. FairGesture combines (1) the Perception Gap metric, (2) trajectory-based motion analysis, and (3) an analysis of user sensorimotor feedback, exploring the reasons behind these mismatches. The work reframes accuracy in gesture recognition as a problem of sensorimotor alignment, advancing user-centred evaluation and inclusive model design.

CCS Concepts

• **Human-centered computing** → **Empirical studies in accessibility**; *Human computer interaction (HCI)*; • **Computing methodologies** → *Artificial intelligence.*

Keywords

Facial gesture recognition, Accessibility, People with motor impairments, Algorithmic fairness, Allocative harms, Sensorimotor alignment, Perception gap, Diagnostic auditing

ACM Reference Format:
Siyu Zhang, Yelu Gu, Kirsten Cater, and Oussama Metatla. 2026. Beyond Accuracy: Auditing Allocative Harms in Facial-Gesture Recognition for People with Motor Impairments. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3772318.3791927

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791927

1 Introduction

Camera-based facial gesture recognition has emerged as a promising modality for hands-free interaction, with applications spanning mobile computing, smart environments [41, 70], and assistive technologies [27, 54, 94]. By leveraging facial movements such as eye blinks, brow raises, and head nods, these systems allow users to issue commands without relying on touch or fine motor hand control. This modality is particularly promising for individuals with upper limb impairment, providing a socially acceptable and contact-free means of interaction [48, 112]. As facial gesture interfaces become increasingly embedded in mainstream platforms, ranging from Augmented Reality (AR) and Virtual Reality (VR) headsets [47] to home automation systems [41, 70], the need to ensure their inclusivity and robustness across diverse motor abilities becomes critical.

However, most facial gesture recognition models are trained on data from able-bodied individuals [3, 61], thereby encoding assumptions of normative motor control and intact proprioceptive feedback. As a result, the behaviour of these models in PwM is poorly characterised. In such users, gestures may be expressed asymmetrically, low in amplitude, unstable in timing, or with inconsistent control, posing challenges for recognition models optimized for canonical motion patterns [29, 83]. Recent work in algorithmic fairness has shown that such biases in training distributions can result in allocative harm [31], whereby valid user inputs are systematically misclassified or excluded due to underrepresentation [50, 97]. In the context of facial gesture interaction, this may manifest as failures to recognise gestures that deviate from canonical movement patterns, despite being intentionally and effortfully performed by PwM [39]. While fairness research in facial analysis has examined disparities related to race [9, 33], gender [44, 90], and age [80], the role of *sensorimotor diversity* remains underexplored. Unlike hand gestures, which often fall within the user's visual field and benefit from real-time visual feedback, facial gestures are typically executed without visual confirmation, placing greater reliance on internal bodily awareness, such as muscle tension and proprioception [73].

These signals, however, may be degraded, delayed, or disrupted in users with neurological or muscular impairments [32], further compounding the mismatch between user intention and algorithmic recognition [83]. We argue that this performance discrepancy constitutes a critical form of allocative harm, where valid, effortful user inputs are systematically rejected due to models overfitted to able-bodied, normative movement patterns.

Rather than viewing these recognition failures as an isolated accuracy problem, we argue that they reflect a deeper misalignment between user intent and algorithmic interpretation, which we define as a sensorimotor alignment problem. To explore this perspective and systematically diagnose these misalignments, we introduce FairGesture, a diagnostic auditing method that assesses fairness in gesture recognition across users with diverse sensorimotor abilities. We adopt a mixed-methods approach involving 11 participants with motor impairments and 11 able-bodied controls, comparing recognition outcomes, self-assessed performance, motion trajectories, and qualitative feedback from interviews. Through this empirical foundation, our study reveals systematic disparities and intention-recognition mismatches concentrated in the PwM group, thereby surfacing the allocative harm inherent in current facial gesture recognition systems.

Our contributions are twofold:

(1) **Empirical Evidence of Intention–Recognition Misalignment in Motor-Diverse Users:** We conduct a mixed-methods comparative study (N=22; 37 gestures) showing that facial-gesture recognisers—trained on normative data—systematically misinterpret PwM's intended gestures. Mismatches are concentrated in low-amplitude, asymmetric, and directionally unstable gestures, and are supported by users' reports of proprioceptive difficulty, uncertain directionality, and diverse interpretations of gesture prompts. We thereby reframe recognition failures as a broader *sensorimotor alignment problem*, rather than isolated classification errors.
(2) **FairGesture: A diagnostic method with Actionable Auditing Assets.** We introduce FairGesture, a triangulation-based diagnostic method for analysing intention–recognition misalignment in facial-gesture interfaces. FairGesture integrates: (1) the *Perception Gap (PG) metric*, which quantifies user–model discrepancies as a proxy for ability-sensitive allocative risk; (2) *motion-dimension mismatch features* (amplitude, direction, temporal dynamics, activation region) derived from established kinematic foundations; and (3) *qualitative perception accounts from users* explaining certain executions feeling of gestures. We also release actionable auditing assets to support alignment-oriented evaluation in future inclusive-interface research. The resource is available at https://github.com/SiyuZhang-zsy/FairGesture

2 Related Work

2.1 Facial Gesture Recognition and Interfaces

Facial gesture recognition has gained prominence as a hands-free interaction to benefit individuals with motor impairments [17, 47, 55, 58, 68]. This approach provides a socially acceptable and contact-free means of interaction for individuals with upper limb impairment. Recent elicitation-based studies have been extended to facial gestures [47, 63, 96, 99, 112], underscoring the need for adaptable interfaces. However, gesture recognition systems specifically designed for PwM remain limited in both gesture diversity and recognition robustness. Most systems focus narrowly on eyelid gestures such as blinks [28] and achieve only moderate performance (e.g., 0.76 in user-dependent evaluations). Broader integration of personalised or atypical facial gestures remains underexplored, leading to recognition failures and reduced user satisfaction [83].

Traditional approaches often rely on facial landmarks and action units (AUs) to interpret expressions and gestures [62]. For instance, Baltrušaitis et al. introduced OpenFace, an open-source tool capable of real-time facial behaviour analysis, including AU detection and head pose estimation [3]. Its latest iteration, OpenFace 3.0, extends this pipeline with improved AU regression and streamlined real-time performance, making higher temporal stability and improved robustness to illumination and pose [45]. Similarly, Lugaresi et al. developed MediaPipe, a framework facilitating real-time face detection and tracking on mobile devices [61]. These landmark and AU based pipelines remain widely adopted in assistive technologies due to their low latency, computational efficiency, and on-device deployability. While these systems achieve high accuracy under controlled conditions, they are predominantly trained on datasets comprising able-bodied individuals [65, 87, 111], thereby encoding assumptions of normative motor control and intact proprioceptive feedback. As a result, they are highly sensitive to the atypical, asymmetric and low-amplitude movement patterns commonly observed in PwM.

In parallel, general-purpose Vision–Language Models (VLMs) have recently demonstrated impressive generalization in visual understanding tasks [14, 60]. However, despite their scalability, VLMs face severe computational and latency bottlenecks that limit their feasibility for real-time, privacy-sensitive assistive interfaces. Recent benchmarks and surveys show that multimodal inference typically demands high-end GPUs, extensive memory, and significant power consumption, even for single-image inputs [15, 49, 92, 107]. Consequently, lightweight, real-time AU-based systems such as OpenFace and MediaPipe remain the most deployable and widely adopted options in current accessibility contexts, making their potential biases toward motor diversity an urgent and understudied challenge.

2.2 Fairness in Computer Vision and Interaction Systems

The field of computer vision has witnessed growing concerns regarding algorithmic fairness, particularly in facial analysis systems. Buolamwini and Gebru revealed significant disparities in gender classification accuracy across different demographic groups[8] , while subsequent HCI work by Keyes and Hamidi demonstrated how facial analysis technologies can reinforce existing societal biases [2, 42, 52]. This line of work underscored the consequences of training models on unbalanced datasets and prompted calls for design practices that attend to the diverse experiences of marginalized

Figure 1: Remote and in-person participation settings. (a) A Participant with motor impairments completed the study remotely from home, seated in a wheelchair. (b) A participant completed the session in person at the university.

(a) (b)

communities [89]. Related work has also examined facial recognition through the lens of auditing and governance, highlighting ethical risks that arise when such systems are deployed without mechanisms to detect or correct systematic disparities [78].

In addition to demographic disparities, recent scholarship has highlighted how fairness challenges also manifest for users with disabilities. Allocative harms refer to instances where algorithmic systems disproportionately allocate or withhold opportunities, resources, or services from certain individuals or groups, often due to underrepresentation or historical biases in the training data, being salient in accessibility contexts [4, 31, 93, 97]. Wang et al. propose a taxonomy of algorithmic harms for disability, categorizing allocative, representational, and interactional forms of exclusion [104], while Venkatasubramanian et al. identify broader classes of negative outcomes arising from AI-driven systems for people with disabilities [102]. Although these frameworks provide essential conceptual grounding, they primarily focus on decision-making or content-moderation systems, leaving open how algorithmic harm manifests in sensorimotor interactions. PwM often face challenges when interacting with facial gesture systems due to inconsistent muscle control or impaired proprioception, which can affect their ability to perform stable, recognisable gestures. Ruocco et al. [83] conducted an evaluation of eyelid-based gesture systems and found that only 68% (13 out of 19) of participants with motor impairments agreed with the statement that "the software rarely misinterprets my input." This reflects a form of allocative harm, where valid user inputs are systematically rejected, not due to malicious bias, but due to the underrepresentation of atypical motion patterns in training data. Unlike the widely studied issues of race and gender bias, such sensorimotor-based disparities remain underexplored in the fairness literature. As facial gesture interfaces gain traction in assistive technologies, recognising and accommodating the diverse motor and perceptual profiles of users with disabilities is critical to ensuring equitable system performance.

2.3 Sensorimotor Perception and Gesture Mismatch

Gesture interaction depends not only on a user's ability to move, but also on their ability to perceive their movement. In motor-impaired populations, this internal feedback, particularly proprioception, is often disrupted [32, 73]. Conditions such as amyotrophic lateral sclerosis (ALS) or cerebral palsy can degrade sensory input from muscles and joints, making it difficult to monitor motion without visual cues [85]. Compounding this, users' *perceived motor competence* (PMC), which is their confidence in having executed a gesture, may be compromised [22, 26]. This perceptual uncertainty contributes to mismatches between user intent and model recognition, particularly for facial gestures that are difficult to self-monitor due to the lack of direct visual feedback. Prior research therefore suggests that gesture mismatch may reflect broader disruptions in sensorimotor perception and regulation rather than purely mechanical execution difficulties. A more detailed review of these constructs and their relevance to gesture interaction is provided in Appendix A.

Building on this body of work, recognition failures in gesture interaction may plausibly stem from perceptual and sensorimotor factors that are often overlooked during model training. However, little empirical research has systematically examined how such perceptual discrepancies manifest in camera-based facial gesture recognition systems or how they contribute to recognition disparities in people with motor impairments. To address this gap, we conducted an empirical study of above-neck gestures and introduce FairGesture, a diagnostic method designed to quantify mismatches between perceived execution and model recognition.

3 Methodology

3.1 Study Design

This study employed a mixed-methods approach to investigate fairness in facial gesture recognition across users with and without motor impairments. We combined quantitative metrics (recognition accuracy and motion trajectory analysis) with qualitative data from participant self-evaluations and post-task interviews to understand the causes and manifestations of intention-recognition mismatches.

3.2 Participant

We recruited a total of 22 participants, with 11 PwM forming Group A (Table 1), and another 11 participants without motor impairment comprising the control group, Group B (Table 2). All participants are Asian and provided their informed consent and were compensated for their time.

The participants in Group A (N = 11; 6 male, 5 female) were adults with motor impairments recruited through online social media platforms and non-profit disability organizations. All participants identified as Asian, with a mean age of 33.6 years (SD = 10). All participants reported motor impairments that affected their ability to use electronic devices. Specifically, one participant had spinal muscular atrophy (SMA II) with limited head mobility; three participants had spinal cord injuries, one of whom also experienced limited head mobility; three participants had cerebral palsy, resulting in hand tremors and difficulties with fine motor control; one participant had poliomyelitis with paraplegia; two participants had facial paralysis on the left side due to stroke and brainstem tumor, respectively; and one participant had arthritis combined with transverse myelitis. Five participants relied on wheelchairs for mobility, and two participants reported speech difficulties due to cerebral palsy, although these did not interfere with participation in the user study. Seven participants had prior experience using above-neck gestures for device control, and all participants were proficient in using flat-screen devices such as smartphones and computers. Table 1 summarizes participant demographics.

For Group B, we recruited 11 individuals (7 male, 4 female) without movement impairments from universities, the community, and online social media, aiming for a demographic profile similar to that of Group A. All participants are Asian. The demographic characteristics of Group B are detailed in Table 2.

3.3 Gesture Collection and Selection

To construct a comprehensive and diverse gesture set, we conducted a systematic review of over a decade of literature on above-neck gesture interfaces, including facial, head, gaze, cheek, and tongue-based gestures from ACM Library and IEEE Xplore. This review encompassed studies on standard, customised, and user-defined facial gestures, from which we initially extracted over 70 candidate gestures. To refine the set for experimental use, we applied several filtering criteria. First, we prioritised gestures with high frequency of use across prior systems to ensure ecological validity. Second, where gesture labels varied across sources (e.g., "smile" vs. "raise mouth corners"), we normalised terminology based on the underlying physical movement rather than affective intent. Third, we merged gestures with highly similar kinematic forms (e.g., "upper eyelid raise" and "wide open eyes") under unified labels. To ensure reliable detection and reduce participant burden, we excluded multi-bit or composite gestures, visually ambiguous gestures, and gestures that fell outside the user's visual field (e.g., sequential gaze paths, complex tongue motions). After filtering, we retained **37** one-bit or zero-bit gestures, which constitute the final gesture set used in our study. The full list of gestures retained after filtering, along with their categorisation and prior usage, is provided in Appendix B (Table 6). For each gesture, we prepared a short video clip used during the study to illustrate the intended motion; details of clip preparation and validation are provided in Appendix C.

3.4 Experimental Procedure

Participants were seated or reclined in front of a webcam (30 fps) in a controlled environment. Each of the 37 gestures was demonstrated using visual examples, after which participants were asked to replicate the gesture. During each gesture task, participants completed a self-evaluation on an 8-point Likert scale (1 = completely failed, 8 = perfect success), rating how successfully they believed they had performed the gesture. Ratings of 5 or higher were treated as self-perceived successful executions. These results were considered a self-perceived success. We used an even-numbered scale to avoid a neutral midpoint, which prior work has shown can function as a dumping ground for uncertainty and introduce additional measurement error [16, 51, 66, 101]. This design also aligns with prior facial-gesture elicitation studies that employ multi-point Likert-type scales to capture users' perceived execution quality [112]. Each gesture trial was followed by a short interview to gather open-ended feedback on participants' perception of the gesture. This process was repeated until all gestures had been completed. All sessions were video-recorded.

3.5 Recognition Models and Apparatus

We employed two facial-analysis pipelines with distinct design goals: MediaPipe Face Mesh[1] [61] and OpenFace 3.0[2] [45], enabling complementary analysis of practical recognition behaviour and underlying motor execution.

MediaPipe represents a deployed, lightweight recognition system commonly used in real-time mobile and assistive applications, providing real-time facial-landmark tracking and activation signals for several facial gestures (e.g., *MouthSmileLeft*, *MouthSmileRight*). Because MediaPipe has been widely adopted in mobile and assistive technologies [7, 18, 36, 47, 59, 71, 76], including accessibility-oriented interaction systems such as Google's Project Gameface,[3] we used its gesture outputs as our primary measure of recognition accuracy. OpenFace 3.0 provides detailed, interpretable measurements of facial muscle activation at the sensorimotor level, including intensity estimates for eight FACS Action Units (AU01, AU02, AU04, AU06, AU07, AU10, AU12, AU14), along with 68-point landmarks and head-pose parameters. We used OpenFace as a secondary recognition channel by directly analysing its AU intensity outputs to indicate gesture activation, without applying additional learning or threshold optimisation. All MediaPipe and OpenFace components were used with default pretrained weights and without fine-tuning.

Although both systems internally rely on facial landmarks, they expose different representational levels. MediaPipe outputs lightweight activation signals and dense geometric landmarks, whereas OpenFace converts landmark motion into semantically interpretable

[1] https://github.com/google-ai-edge/mediapipe
[2] https://github.com/CMU-MultiComp-Lab/OpenFace-3.0
[3] https://blog.google/innovation-and-ai/products/google-project-gameface/

Table 1: Demographic details about the participants with upper limb impairments (Group A) [1] Ir = Irreversible Diseases; Pr = Progressive diseases; Fa = Rapid fatigue; Dig = Difficulty gross motor; Dif = Difficulty fine motor; Tre = Tremor; [2] • means YES, - means No

Participant	Health condition	Since	Ir	Pr	Fa	Dig	Dif	Tre
P1(32, Female)	SMA II	6 months old	•	•	•	•	•	-
P2(31, Male)	C1 Spinal Cord Injury	10 years ago	•	-	•	•	•	•
P3(36, Female)	C5 Spinal Cord Injury	5 years ago	•	-	•	•	•	-
P4(18, Female)	Stroke(hemiplegia)	5 years ago	•	-	•	•	•	-
P5(32, Male)	Right Brainstem Tumor(hemiplegia)	3 years ago	-	-	•	-	•	-
P6(26, Male)	Cerebral palsy	born	•	-	•	•	•	-
P7(36, Female)	T1-T4 spinal cord injury	5 years ago	•	-	•	•	-	-
P8(31, Male)	Poliomyelitis and Paraplegia	28 years ago	-	•	•	•	-	-
P9(33, Male)	Cerebral Palsy	born	•	•	-	•	•	•
P10(28, Male)	Cerebral palsy	born	•	-	•	•	•	•
P11(24, Female)	Transverse Myelitis	10 years ago	•	-	•	-	•	-

Table 2: Demographic details about the participants without upper limb impairments (Group B)

Participant	Age	Sex	Health condition
Pb1	28	Male	no motor disorder
Pb2	27	Male	no motor disorder
Pb3	30	Male	no motor disorder
Pb4	57	Female	no motor disorder
Pb5	30	Male	no motor disorder
Pb6	21	Male	no motor disorder
Pb7	37	Male	no motor disorder
Pb8	57	Male	no motor disorder
Pb9	29	Female	no motor disorder
Pb10	35	Female	no motor disorder
Pb11	24	Female	no motor disorder

FACS AU intensities. Analysing both pipelines therefore enables a cross-model robustness check: if similar mismatch patterns arise across both systems, they are unlikely to be model-specific and instead reflect structural limitations in current facial gesture recognition pipelines.

3.6 Data Processing and Feature Extraction

To prepare the behavioural and recognition signals used in our analyses, as well as support downstream fairness evaluation, we collected four types of data: (1) recognition outputs from models, (2)participant self-evaluation scores, (3) motion trajectories extracted from models, and (4) participants' qualitative perception accounts

Recognition outputs. Recognition accuracy was computed per gesture participant using two independent pipelines: (a) MediaPipe gesture activations and (b) OpenFace AU-based gesture detection. MediaPipe-provided gestures (e.g., *Blink*, *MouthSmileLeft*) were taken directly from its activation outputs. For gestures not included in MediaPipe's default gesture set (e.g., *Raise Mouth Corners*), we constructed operational gesture definitions using the semantic activation signals exposed by MediaPipe (e.g., *Mouth Smile Left* + *Mouth Smile Right*). These definitions follow the underlying (1) published AU–gesture correspondences in FACS literature [25, 109];(2) the official demostration documented in MediaPipe's introduction [4] and developer documentation [37, 61]; and (3) prior accessibility and facial-gesture–interaction work that composes AU signals into gesture-level definitions. [39, 83, 112] and are detailed in Appendix D. OpenFace-based recognition was computed by thresholding the AU intensities (confidence scores) associated with each gesture, according to our predefined AU–gesture mapping. The full operational definitions for gesture computation are detailed in Appendix D.

Participant self-evaluation scores. After each gesture attempt, participants rated their perceived execution quality on an 8-point Likert scale (1 = "not successful at all," 8 = "fully successful"). The self-evaluation scores capture participants' internal perception of gesture completion independent of model output, providing an essential user-centric reference for interpreting intention–recognition mismatches.

Motion-trajectory analysis. To characterise how gestures were physically executed, we analysed temporal AU intensity traces (OpenFace) and landmark trajectories (MediaPipe) across the duration of each gesture. For each gesture instance, we extracted measures of amplitude (peak activation or displacement relative to baseline), duration (time from gesture onset to offset), and directional deviation (e.g., left–right asymmetry computed from bilateral AU or landmark differences). These motion features were used to identify sensorimotor patterns underlying recognition failures, such as low-amplitude execution, unstable timing, or directional drift.

Quantitative analysis. We compared group-level recognition accuracy and self-evaluation scores between participants with motor impairments and non-impaired controls. We tested between-group differences using two-sample (unpaired) t-tests for aggregated outcomes (e.g., overall accuracy and self-evaluation). For gesture-wise comparisons across the 37 gestures, we used Mann-Whitney U tests (Wilcoxon rank-sum). To account for multiple comparisons, we applied Benjamini-Hochberg false-discovery-rate (FDR) correction (q = .05), reporting raw p-values in the main text and FDR-adjusted values in the appendix tables.

[4] https://codepen.io/mediapipe-preview/pen/OJBVQJm

Qualitative analysis. Qualitative data were analysed using directed content analysis. Two independent coders reviewed transcribed interviews to identify and code instances related to proprioceptive difficulty, directional disorientation, and gesture interpretation. Inter-coder agreement was assessed at the code level, achieving Cohen's $\kappa = 0.69$, after which disagreements were resolved through discussion.

3.7 Ethical Considerations

The study was approved by the ethics committee of the University of Bristol. All participants were informed of their rights, including the right to withdraw. Data were anonymised and securely stored in an encrypted database in the University of Bristol.

4 FairGesture: A Diagnostic Auditing Method

FairGesture integrates three complementary evidence sources: Perception Gap (PG), landmark-based motion dimensions, and participants' perception accounts into a coherent diagnostic method for interpreting intention–recognition misalignment. PG highlights where mismatches occur, motion-dimension analysis characterises how gesture execution diverges from canonical templates, and participant accounts reveal why certain motions felt successful to users. Together, these components identify where and why gesture systems misinterpret users' intent, supporting researchers and developers in uncovering potential allocative harms that arise from reliance on normative motor patterns.

4.1 Perception Gap

To quantify misalignment between human intent and algorithmic recognition, we define the **Perception Gap (PG)** as the difference between self-perceived success and model recognition confidence:

$$\mathrm{PG}_i = \mathrm{SelfEval}_i - \mathrm{Accuracy}_i,$$

where $\mathrm{SelfEval}_i$ denotes the participant's perceived success. $\mathrm{Accuracy}_i \in [0, 1]$ denotes the recognition confidence output by the model. Because our study used an 8-point self-evaluation scale (1–8), we linearly map self-evaluation scores to $[0, 1]$:

$$\mathrm{PG}_i = \frac{\mathrm{SelfEval}_i - 1}{7} - \mathrm{Accuracy}_i.$$

Here, $\mathrm{SelfEval}_i \in [1, 8]$ is linearly scaled to $[0, 1]$ by subtracting 1 and dividing by 7.

A positive PG indicates that the user felt more successful than the model acknowledged, revealing potential failure interaction and ability-sensitive allocative risk. A negative PG indicates that the system recognised a gesture the user believed was unsuccessful, often linked to proprioceptive uncertainty .These cases do not constitute allocative harm and are therefore treated as secondary signals. PG is a continuous, descriptive measure, and the heatmap visualisation [46] allows us to observe salient distributional patterns, including **(1)** group-level divergence (e.g., consistently positive PGs in Group A but neutral in Group B), **(2)** cross-gesture and cross-model consistency (e.g., the same gestures showing positive PGs across both MediaPipe and OpenFace), and **(3)** within-group clustering (e.g., concentrated positive PGs for eye gestures in Group A). These distributional patterns align with later trajectory-level deviations and participants' perception accounts, strengthening their interpretive significance.

4.2 Motion-Trajectory Analysis for Diagnosing Sensorimotor Mismatch

To understand the motor sources of intention–recognition mismatches, FairGesture analyses gesture execution using landmark-based motion trajectories derived from MediaPipe. Our approach is grounded in established models of facial movement and motor-control variability [19, 20, 82, 100]. For each gesture instance, we use four fundamental motion dimensions as diagnostic lenses for interpreting how users' physical execution may diverge from the canonical templates implicitly assumed by recognition models:

(1) **Amplitude Dimension** — magnitude of landmark displacement relative to baseline (e.g., the amplitude of lip-corner elevation, brow raise). Reduced amplitude can reflect effortful yet low-intensity gestures that fall below model thresholds.
(2) **Direction Dimension** — Captures the orientation of the movement vector, including asymmetry, lateral drift, and left–right reversals. Directional deviation often leads recognisers trained on canonical, high-consistency motion to misinterpret the gesture.
(3) **Temporal-Dynamics Dimension** — Describes timing properties, such as onset speed, peak timing, duration, and intra-gesture stability. Slow onset, delayed peaks, or oscillatory motion can conflict with recognisers that assume temporally stable activation windows.
(4) **Activation-region Dimension** — Identifies which facial subregions are recruited when producing the gesture (e.g., cheek activation instead of brow activation). This allows us to detect compensatory or non-canonical activation pathways that recognisers may not be trained to interpret.

These features provide a physiologically grounded representation of how users execute gestures, enabling systematic comparison across groups and gestures. Trajectory analysis forms the motion-level evidence used later to interpret Perception Gap (PG) patterns.

4.3 Participant Perception Accounts as Interpretive Evidence

As part of FairGesture's triangulation pipeline, we incorporate participants' accounts of their perceived gesture execution to contextualise motion patterns and PG values. After each gesture trial, participants provided a brief open-ended reflection describing how the gesture felt, which muscles they believed they activated, and whether their execution aligned with their intention, following a consistent prompt. We applied a focused content-analysis procedure. Two coders independently annotated statements related to proprioceptive confidence, directional certainty, perceived amplitude, and compensatory muscle use; disagreements were resolved through discussion. These qualitative insights contextualise the quantitative mismatches and help explain why PG-identified gestures diverged along the four motion-dimension foundations (amplitude, direction, temporal dynamics, activation region), complementing trajectory-level observations with participants' own accounts of perceived execution.

4.4 Actionable Auditing Assets

To support adoption, we release several actionable auditing assets:

(1) **Perception Gap computation scripts** (Python)
(2) **Heatmap visualisation tool** for group- and gesture-level disparity inspection
(3) **Trajectory viewer** for AU and landmark time-series inspection
(4) **Interpretive checklist** to help developers diagnose mismatches

5 Results

This section presents our mixed-methods analysis of gesture recognition performance, user perception, and sensorimotor characteristics across participants with and without motor impairments. We report both quantitative recognition outcomes and qualitative explanations for the intention-recognition mismatch, structured across model performance, user self-evaluation, gesture trajectory deviations, and user-reported difficulties.

5.1 Recognition Accuracy by Group

We report recognition outcomes for both MediaPipe and OpenFace. MediaPipe outputs serve as our primary accuracy measure given their deployment relevance, while OpenFace provides an independent AU-based perspective. Reporting both allows us to examine whether mismatches are model-specific or consistent across architectures.

The recognition accuracy from MediaPipe varied significantly between the two groups of participants. As shown in the heatmap in Figure 2 and Figure 3 the model achieved an average accuracy of **56.2%** (± 26.8%) for gestures performed by Group A (motor-impaired participants), compared to **74.8%** (± 16.5%) for Group B (non-impaired). A two-sample t-test on participant-level means confirmed a significant group difference ($p < .0001$).

The accuracy also varied by gesture. Gestures involving small, localized movements, e.g., *LeftWink* and *RightWink*, had the lowest recognition rates in both groups, but especially in Group A (22.9%, 24.6%). In contrast, large or symmetric gestures such as *PuckerLips* and *RaiseEyebrows* exceeded 80% accuracy in Group A and 95% in Group B. To examine gesture-wise group differences, we conducted Mann–Whitney U tests across all 37 gestures comparing Group A and B. Recognition accuracy differed significantly ($p < 0.05$) on 26 gestures, most notably those involving lateral asymmetry or slow onset (e.g., *Move Head Backward*, *Move Jaw Forward*). Table 9 in the Appendix shows the accuracy comparison between the two groups for each gesture.

In addition, the results of *OpenFace 3.0* clearly show that *OpenFace 3.0* exhibits a pattern consistent with *MediaPipe*, demonstrating statistically significant fairness issues. For PwM, the mean recognition accuracy was **53.94%**, significantly lower than for able-bodied users (**77.03%**), with $p = 0.0049$. Detailed results are provided in Table 8 in the Appendix. This cross-model consistency suggests that the observed mismatches are not tied to a specific recognition architecture but reflect broader limitations in current facial-gesture pipelines.

To supplement the gesture-level visualisations, Table 3 summarises the group-level means and standard deviations for recognition accuracy from MediaPipe. For reference, the corresponding self-evaluation scores and perception gap values are also reported in the table and discussed in the following sections.

5.2 User Self-Evaluation and Perceived Competence

Participants in Group A reported a mean self-evaluation score of **7.00** (±1.85), while those in Group B averaged **7.54** (±0.98), indicating broadly comparable levels of reported gesture confidence. However, the greater variability observed in Group A suggests more heterogeneous experiences of perceived motor competence.

Within Group A, lower self-evaluation scores were concentrated in eye-related gestures (e.g., *Open Left Eye*, *Open Right Eye*) and certain head movements among participants with more severe motor impairments. By contrast, gestures involving mouth movements (e.g., *Raise Mouth Corners*, *Depress Mouth Corners*) were more frequently rated as successfully executed (self-evaluation ≥ 5), even when recognition accuracy was inconsistent.

In Group B, self-reported unsuccessful gestures were more evenly distributed across facial regions and exhibited closer correspondence with recognition accuracy patterns. These results indicate that perceived performance in Group A varies across gesture types and body regions. Table 10 in the Appendix provides gesture-level self-evaluation comparisons between the two groups.

5.3 Perception Gap and Misalignment Analysis

To quantify intention–recognition misalignment beyond raw accuracy, we compute a Perception Gap (PG) score for each gesture trial, defined as the difference between the user's normalised self-evaluation (scaled to $[0, 1]$) and the model's recognition outcome. We interpret PG magnitude in terms of its distributional patterns across gestures and user groups. A positive PG identifies cases where users believe they executed the gesture successfully but the system failed to recognise it; such cases correspond to the critical allocative harm scenario examined in this work. A negative PG value reflects the opposite pattern: the system recognises a gesture that users believe they performed unsuccessfully. While informative as a perceptual discrepancy, often tied to proprioceptive uncertainty, these negative PG cases do not constitute allocative harm and are therefore treated as secondary signals.

A complementary binary mismatch analysis further reinforces this pattern of disproportionate misalignment. In Group A, only **75.8%** of trials aligned with the model output, and among all mismatches, **87.1%** were overestimations—cases where participants believed they had successfully executed the gesture but the recogniser failed to detect it. By contrast, Group B exhibited a substantially higher **95.7%** alignment rate, with mismatches more evenly distributed between over- and underestimation. These quantitative differences indicate that Group A experiences systematic intention–recognition disagreement rather than isolated mismatches. A Mann–Whitney U test comparing PG distributions between groups further confirmed significantly greater misalignment in Group A ($U = 92417.5$, $p = .0042$). Table 11 in the Appendix compares gesture-level alignment metrics between the two groups. Multiple

Table 3: Group-level comparison of recognition accuracy, self-evaluation scores, and perception gap. Means and standard deviations are reported, along with statistical test results and p-values. Match rate indicates the percentage of gestures where participant self-evaluation agreed with model recognition.

Metric	Group A (M ± SD)	Group B (M ± SD)	Test	p-value
Recognition Accuracy (%)	56.2 ± 26.8	74.8 ± 16.5	$t(20) = -11.87$	$< .0001$
Self-Evaluation Score (1–8)	7.12 ± 1.74	7.54 ± 0.99	$t(20) = -4.24$	$< .0001$
Perception Gap (scaled)	0.31 ± 0.25	0.19 ± 0.20	$U = 92417.5$	$= .0042$
Match Rate (%)(over/under)	75.8 (21.1 / 3.1)	95.7 (2.8 / 1.5)	–	–

comparisons were corrected using Benjamini–Hochberg FDR (q = .05). FDR-adjusted values are provided in Tables 9 and 10.

Figure 2 (right) and Figure 3 (right) visualise the distribution of PG across the gesture set. For non-impaired participants (Group B), PG values cluster tightly around zero, indicating strong alignment between perceived success and model recognition. Small pockets of negative PG suggest occasional underestimation (e.g., *Brow Raise*), but these instances are isolated rather than systematic. In contrast, Group A exhibits broad and structured regions of positive PG. These clusters appear most prominently in gestures involving directional isolation (*Close Left Eye*), fine motor articulation (*Eye Look Up Squint Eyes*), and complex multi-muscle coordination (*Raise/Depress Mouth Corners*, *Jaw Forward*). In these gestures, participants consistently rated their execution as successful, yet the recogniser repeatedly failed to detect the intended motion. Such patterns reflect systematic barriers to interaction rather than isolated classification errors.

5.4 Trajectory-Based Analysis of Mismatched Gestures

To further understand why intention–recognition mismatches arise, we analysed landmark-based motion trajectories for gestures with high Perception Gap (PG). We examined mismatches through four fundamental motion dimensions commonly used in modelling facial movement: amplitude, direction, temporal dynamics, and activation region (Table 4). These dimensions capture how a gesture unfolds physically and provide a neutral lens for diagnosing divergence from the canonical motion templates that recognition models implicitly assume. Across participants with motor impairments (Group A), we observed recurring forms of mismatch along these dimensions, whereas such deviations were rare in Group B. Figure 4 visualise representative exemplar trajectories of different dimensions, with extended plots in Figure 7 in Appendix G.

Mismatch along the Amplitude Dimension was commonly observed in many Group A trials. Gesture trajectories often exhibited the correct overall shape but with substantially reduced magnitude, resulting in landmark displacements that fell below the model's recognition thresholds. Despite this, participants frequently rated these gestures as successful, indicating that the movements felt complete from their perspective. This mismatch arose because diminished muscular activation led to insufficient AU signal strength for reliable recognition. For example, P4, a participant with hemiplegia, attempted "Raise Mouth Corners" exhibited limited movement amplitude of the right mouth corner (Figure 4 (a)) and was frequently misrecognised as "Raise Right Mouth Corner."

Directional Deviations reflected left–right disorientation or asymmetric motor control. In one case, P11's "Left Wink" gesture (Figure 4 (b)) was misclassified due to unintended activation in the opposite eye region.

Speed/Timing Mismatch arose from delayed initiation or unstable execution. P6's "Right Wink" gesture (Figure 4 (c)), for instance, had a prolonged onset and was mistakenly interpreted by the system as a "Close Right Eye".

Body Region Shifts occurred when compensatory movements redirected the activation locus away from the intended region. In P10's "Move Jaw Forward," (Figure 4 (d)) involuntary engagement of the cheeks and upper lips caused the gesture to be misclassified, revealing how compensatory behaviour can confound recognition. These trajectory-based analyses provide concrete evidence that sensorimotor diversity manifests in distinct, classifiable ways that challenge current recognition systems. They also underscore the need for models that can accommodate non-normative execution patterns rather than rejecting them as noise.

5.5 Qualitative Insights from Participants

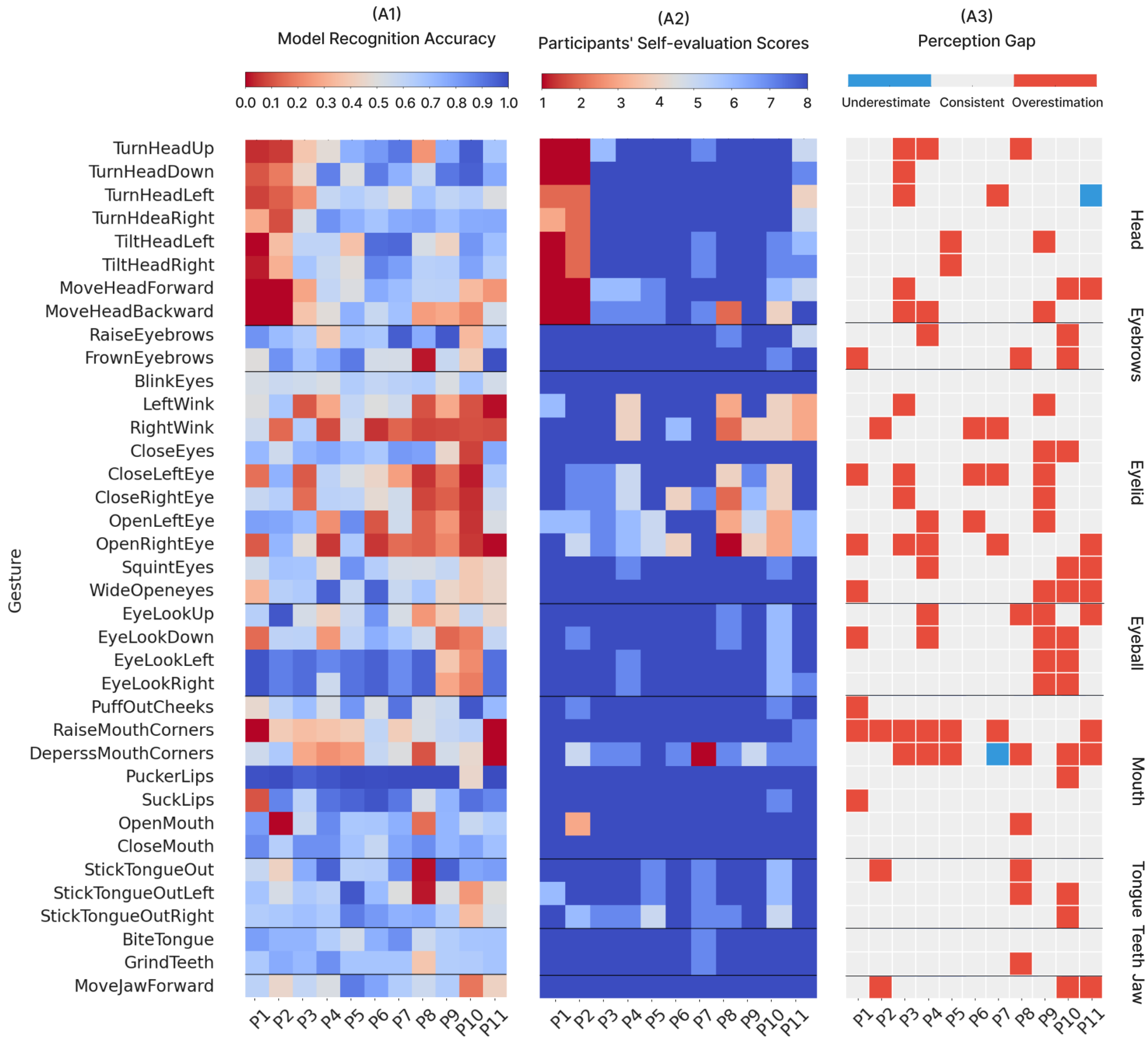
Participants' reflections shed light on the subjective motor experiences that shaped how they approached and evaluated facial gestures. Table 5 summarises representative links between participant feedback, recognition outcomes, and the underlying assumptions encoded in current gesture recognisers.

Impaired sensorimotor awareness. Many participants struggled to perceive their own facial movements, often seek external validation to confirm if a gesture had occurred (e.g., "Did I move my eyebrows?"). This lack of real-time feedback introduced doubt and inconsistency in their execution. This uncertainty often led them to rely on effort rather than outcome.

Spatial disorientation. Another phenomenon was spatial disorientation, particularly in distinguishing left from right. Such confusion impacted gestures like winks or lateral eye movements and sometimes led to directional misalignment. This perceptual ambiguity aligns with directional drift observed in trajectory analysis and explains why users confidently but incorrectly reported directional success.

Motor coordination difficulties. In addition, participants described difficulty controlling small, localised muscle groups, particularly during gestures requiring isolated movement without unintended co-activation(e.g., tongue gestures), which often resulted in unintended co-movements of adjacent muscles or regions. These

Figure 2: Recognition accuracy, self-evaluation, and perception gap for Group A (motor-impaired). Heatmaps show model accuracy (left), user self-evaluation (centre), and the resulting Perception Gap (right). Cooler colors represent higher values for accuracy and self-evaluation; in the PG map, warmer colors indicate positive gaps in which users perceived success but the recogniser failed to detect the gesture. Group A displays substantially lower and more variable recognition accuracy, greater spread in self-evaluation, and broad clusters of positive PG values, indicating systematic intention–recognition misalignment.



perceived coordination constraints correspond to the temporal instability and region-shift patterns identified in motion traces.

Diversity in gesture interpretation. Finally, we also observed divergent interpretations of what constitutes a gesture. Participants often enacted gestures based on their own embodied histories rather than on normative definitions used by recognition models. For example, participant P1, who has SMA II and is unable to lift the corners of her mouth due to muscle atrophy, described her smile as a "lip stretch," whereas most models, such as MediaPipe [61], associate "Mouth Smile" with "Raise Mouth Corner."

Together, these qualitative accounts help explain why many gestures that felt intentional, effortful, and complete to participants were nonetheless misrecognised by the system. Rather than reflecting user error or misunderstanding, these accounts indicate that participants often relied on embodied cues, such as perceived effort, muscle tension, or habitual movement patterns, that are not directly observable to recognition models. By articulating how gestures were perceived, executed, and interpreted by users themselves, these insights contextualise the Perception Gap patterns and trajectory deviations reported earlier, grounding them in participants' sensorimotor constraints and perceptual strategies. In

Figure 3: Recognition accuracy, self-evaluation, and perception gap for Group B (non-impaired). Heatmaps follow the same layout as Group A: accuracy (left), self-evaluation (centre), and Perception Gap (right). Group B shows consistently high accuracy and confidence, with PG values clustering tightly around zero and only small, isolated deviations. Unlike Group A, no systematic positive PG regions emerge, indicating strong alignment between perceived execution and model recognition.

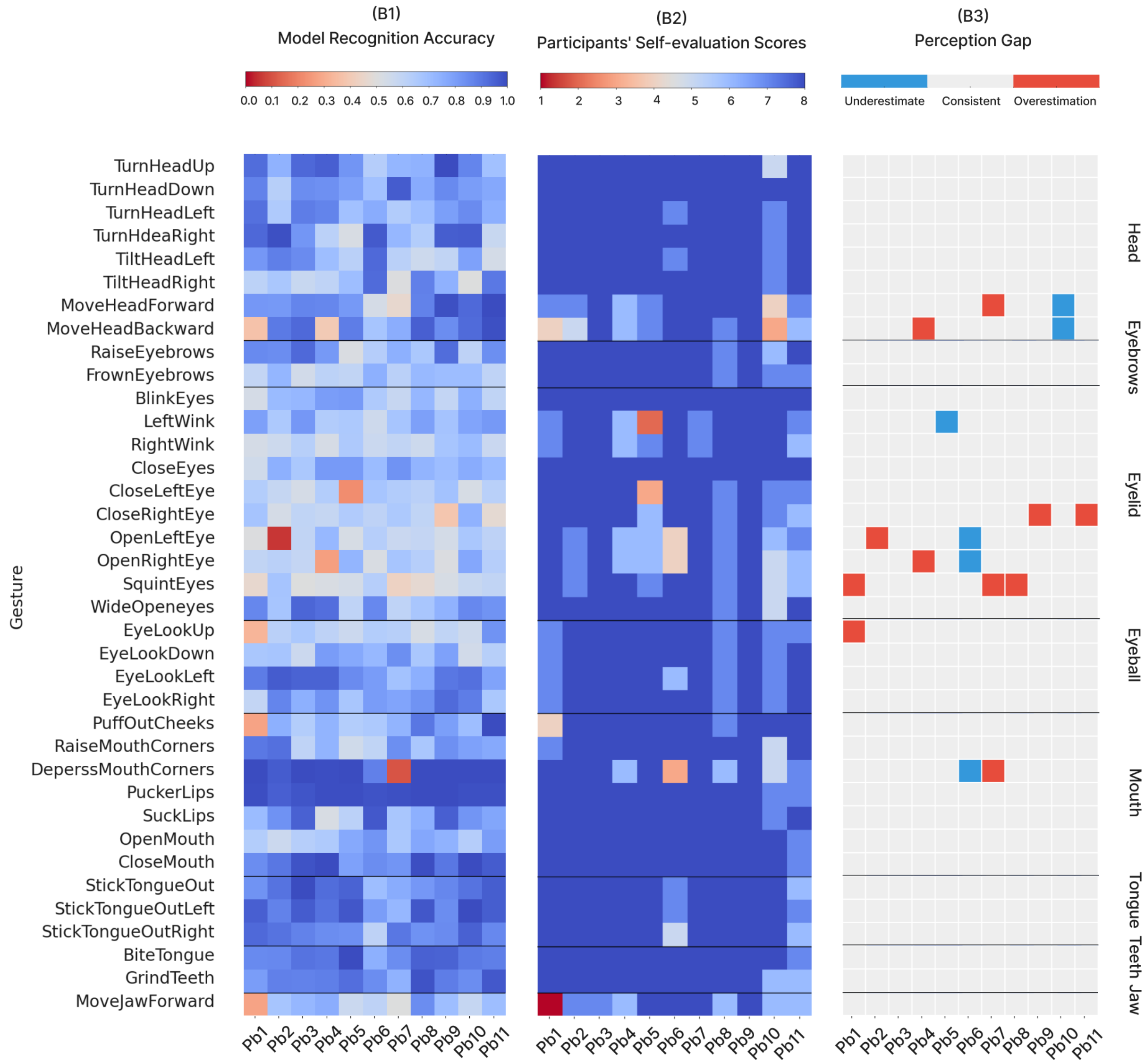


Table 4: Examples of Intention–Recognition Mismatch Cases. Each row highlights a gesture performed by a motor-impaired participant that was misrecognised despite a high self-evaluation score. AU = Action Unit. Related trajectories shown in Figure 4 (a)–(d).

Error Type	Gesture Track	Participant	Intention	Recognition
Amplitude	Figure 4 (a)	P4 (Hemiplegia)	Raise Mouth Corners	Raise Right Mouth Corner
Direction	Figure 4 (b)	P11 (Transverse Myelitis)	Left Wink	Right Wink
Speed	Figure 4 (c)	P6 (Cerebral palsy)	Right Wink	Close Right Eye
Region Shift	Figure 4 (d)	P10 (Cerebral palsy)	Move Jaw Forward	Pucker Mouth

Figure 4: Representative gesture trajectories illustrating mismatches. Each pair compares Group B (top row) and A (bottom raw) trajectories across four error axes: (a) amplitude, (b) direction, (c) speed, and (d) body region activation.

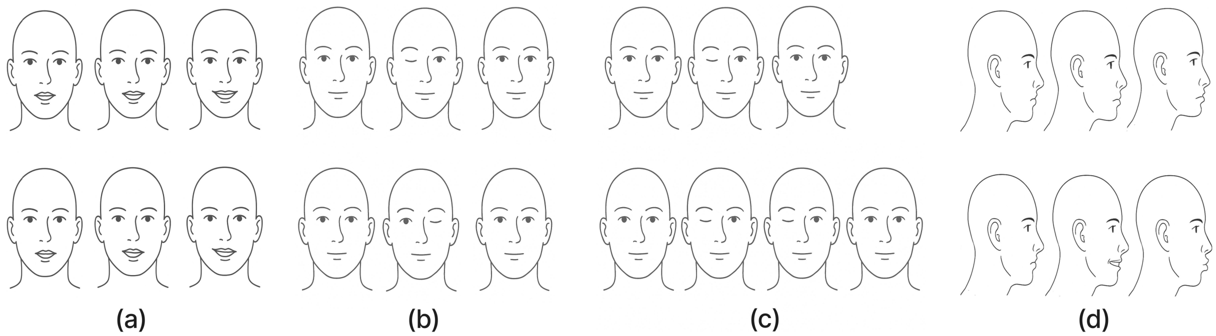


Table 5: Linking participant feedback with recognition errors and algorithmic implications. Quotes highlight how user intent diverges from system assumptions.

User Feedback	Recognition Outcome	Implication for Models
"Did I move my both sides?"	Not recognised (low AU)	Thresholds too strict for low-amplitude input
"I am confused of left and right"	Misclassified direction	Lacks tolerance for directional ambiguity
"I can't do it so precisely."	Gesture not accepted	Overfitted to rigid, idealized motion patterns
"This is my smile looks like."	Misrecognised or ignored	Treats atypical expressions as invalid

this way, participant accounts provide critical interpretive evidence for understanding intention–recognition mismatches beyond what recognition metrics alone can reveal. Extended qualitative examples and trajectory-level evidence are provided in Appendix H. Building on these empirically grounded findings, we further abstract recurring patterns into a set of higher-level algorithmic failure modes in Appendix I.

6 Discussion

6.1 Reframing Recognition into Sensorimotor Alignment

Our findings suggest that conventional accuracy metrics are insufficient to characterise performance for facial-gesture interfaces used by people with motor impairments. Perception Gap (PG) analysis and trajectory inspection reveal systematic misalignment between user intention and model interpretation. Rather than treating these failures as isolated misclassifications, we argue that they reflect a sensorimotor alignment problem: recognition pipelines are implicitly tuned to canonical movement patterns and intact proprioception, while many users rely on atypical, asymmetric, or compensatory motor strategies. This perspective helps reconcile two seemingly contradictory observations. On the one hand, participants with motor impairments often reported high confidence in their gesture execution; on the other, models consistently rejected or misinterpreted those gestures. PG makes this misalignment measurable, and motion-dimension analysis shows that it is not random noise but linked to characteristic differences in amplitude, direction, temporal stability, and activation region. In this sense, "accuracy" is not simply a property of the model but of the relationship between a model's assumptions and a user's motor repertoire.

6.2 From Perception Gap to Ability-Sensitive Allocative Harm

Recognition errors in facial-gesture interfaces do not merely reduce technical performance, they carry ability-sensitive allocative consequences. Participant P10, who has strabismus, described his experience with face-based authentication on his mobile phone: when he intends and genuinely believes he has performed "look left," but the system repeatedly fails to recognise it (mismatch → failed recognition), he is unable to access his financial app despite multiple attempts (failed recognition → reduced access). This substantially disrupts his daily life (reduced access → allocative harm). In this interaction, he expends far more effort than other users yet receives no effective feedback or successful outcome.

Existing gesture recognisers implicitly encode normative assumptions about amplitude, symmetry, temporal stability, and muscle recruitment—assumptions learned from datasets that lack motor-diverse facial gestures. As a result, P10's strabismus-related gaze patterns fall outside the model's representational space. These assumptions systematically penalise genuine but non-canonical motor expressions, such as P10's non-standard rightward gaze. These recurring failures directly translate into reduced access: the user cannot trigger commands, navigate interfaces, make selections, or confirm actions, even when exerting maximal effort and believing the gesture was successfully executed. The system effectively withholds opportunities, functionality, or access from a specific ability group, thereby constituting allocative harm. Unlike previously documented demographic allocative harms, these harms are ability-sensitive: they arise from users' diverse motor and proprioceptive profiles and the model's failure to adapt to them.

6.3 Implications for Inclusive Model, Dataset and Interface Design

By reframing recognition errors through the lens of ability-sensitive allocative harm, our analysis highlights that improving accessibility requires aligning recognition behaviour with the ways users can and do express intent. In the following, we organise these implications across model, data, and interface design considerations.

6.3.1 ***Data and Model Implications.*** To translate Perception Gap (PG) analysis into concrete improvements, we propose several dataset-centric strategies aimed at enhancing the inclusivity of recognition systems. These strategies use PG distributions not merely as a diagnostic tool, but as a guide for principled intervention in data composition and model training.

Targeted personalisation via Micro-Datasets. At the level of individual deployment, PG analysis enables efficient, user-specific adaptation. Instead of full retraining or large-scale data collection, the system can identify gestures with consistently high PG for a particular user and request a minimal micro-dataset (5–10 samples) only for those gestures. These samples can then support lightweight adaptation techniques, such as threshold calibration or parameter-efficient fine-tuning. In this way, PG serves as a prioritisation signal that localises adaptation to a user's unique motor profile, reducing burden while improving alignment for that individual.

Strategic Sampling Informed by Perception Gaps. At the population level, aggregated PG distributions reveal gesture classes that are systematically misrecognised across users with motor impairments. Consistently positive PG clusters indicate categories where valid user intent is repeatedly rejected, signalling the need for expanded sampling, reweighting, or targeted augmentation. Rather than indiscriminate data expansion, developers can prioritise under-represented execution styles—such as low-amplitude movements, directionally unstable trajectories, or compensatory activation patterns. This strategy aligns with fairness-aware data curation approaches [79], but grounds prioritisation in empirically observed misalignment patterns.

Structuring Datasets by Motor-Complexity Tier. Recognizing that gesture execution varies naturally in complexity, we propose organizing datasets into explicit difficulty tiers. A three-tier structure: ranging from canonical, high-amplitude gestures, through moderately variable ones, to highly variable or compensatory patterns, enables curriculum-style training. Models can first learn stable prototypes from clear examples before gradually generalizing to more diverse and challenging executions. This structuring datasets approach, proven in related machine-learning studies, builds robustness and reduces brittleness when faced with out-of-distribution expressions [5, 95], enables models to progressively generalise from stable exemplars to more diverse and atypical executions. This approach explicitly acknowledges motor variability as a structured dimension of learning, rather than treating it as noise or out-of-distribution artefacts

6.3.2 ***Interface Design Implications.*** Beyond model- and data-level improvements, the mismatches revealed by FairGesture point to several concrete interface-design opportunities for improving accessibility.

Mirror Feedback. First, facial gestures differ from hand gestures in that most motion occurs outside the user's visual field. Many participants, especially those with reduced proprioceptive sensitivity, repeatedly reported uncertainty about whether they had executed a gesture correctly. Interfaces can therefore provide perceptual feedback or uncertainty visualisation to make system interpretation visible. Examples include a small, non-intrusive mirror widget showing the user's live face, overlaid with landmarks or simplified motion traces, and confidence indicators that reveal how close the gesture is to being recognised. Such feedback can help users adjust their motion in real time and reduce intention–execution discrepancies.

Intent-centric Personalisation. Second, interfaces can support intent-centric personalisation. Rather than requiring users to conform to canonical gestures, systems can treat stable, non-canonical expressions as valid personalised gestures. FairGesture diagnostics can guide this process by identifying gestures that are difficult for a given user and suggesting alternatives that better match their sensorimotor profile. During onboarding, the system could:(1) recommend gestures with consistently low PG for that user, (2) adjust decision thresholds to align with the user's amplitude range, or (3) re-map gestures that exhibit consistent directional drift (e.g., treating a user's midline "wink" as their personalised left-wink gesture).

Taken together, these implications suggest that improving gesture recognition accessibility is not a matter of isolated model fixes, but of aligning sensing, interpretation, and interaction with the ways users can and do express intent. FairGesture offers one step toward this alignment by making allocative harms visible and actionable, but the broader challenge lies in designing recognition systems that treat motor diversity as a first-class design constraint rather than an edge case.

7 Limitations and Future Work

This study focused on a limited population, excluding older adults and individuals with cognitive impairments. Our gesture set was constrained to primitive, visually discernible one-bit gestures, omitting more complex multi-bit and intraoral actions due to ambiguity and limited proprioceptive feedback. We also note that our qualitative annotations achieved moderate inter-rater agreement (Cohen's $\kappa = 0.69$), highlighting opportunities for refining annotation guidelines or incorporating complementary validation signals in future work. Future work will generalise FairGesture across models, especially vision-language models (VLMs), which have the potential to directly understand user gestures and provide more specific textual feedback. The triangulation-based diagnostic method could also be extended to other user groups to examine additional fairness challenges.

8 Conclusion

Through a mixed-methods analysis, we revealed systematic mismatches between user intention and model recognition in facial-gesture interaction, driven by sensorimotor variability, perceptual ambiguity, and rigid algorithmic assumptions. Our findings show that these errors are not isolated, but reflect allocative harms embedded in model design and training data. By combining self-evaluation, recognition accuracy, and trajectory analysis, FairGesture enables interpretable fairness diagnostics and identifies gesture types most vulnerable to exclusion. This work reframes gesture recognition as a sensorimotor alignment problem and highlights the need for evaluation frameworks that better accommodate motor diversity.

Acknowledgments

We thank all participants for their time and thoughtful contributions to this study. We are grateful to our colleagues for insightful discussions and feedback that helped shape this work.

References

[1] Wajeeha Akram, Laura Tiberii, and Margrit Betke. 2006. A Customizable Camera-Based Human Computer Interaction System Allowing People with Disabilities Autonomous Hands-Free Navigation of Multiple Computing Tasks. In *Universal Access in Ambient Intelligence Environments*. https://api.semanticscholar.org/CorpusID:7903987

[2] Fabio Bacchini and Ludovica Lorusso. 2019. Race, again: how face recognition technology reinforces racial discrimination. *Journal of information, communication and ethics in society* 17, 3 (2019), 321–335.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[4] Fabian Beigang. 2022. On the advantages of distinguishing between predictive and allocative fairness in algorithmic decision-making. *Minds and Machines* 32, 4 (2022), 655–682.

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[6] Hongliang Bi and Jiajia Liu. 2022. Csear: Metalearning for head gesture recognition using earphones in internet of healthcare things. *IEEE Internet of Things Journal* 9, 22 (2022), 23176–23187.

[7] Michael Bruner. 2024. Exploring Embodied Interactive Techniques: with Touchdesigner and Mediapipe. In *ACM SIGGRAPH 2024 Labs* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 11, 2 pages. doi:10.1145/3641236.3665164

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, 77–91.

[9] Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. 2020. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE transactions on biometrics, behavior, and identity science* 3, 1 (2020), 101–111.

[10] J Charlier, Cathy Buquet, Frederic Dubus, Jean Pierre Hugeux, and Bernard Degroc. 1997. VISIOBOARD: a new gaze command system for handicapped subjects. *Medical and Biological Engineering and Computing* 35, 416 (1997), 433–436.

[11] Sotirios Chatzidimitriadis, Saber Mirzaee Bafti, and Konstantinos Sirlantzis. 2023. Non-intrusive head movement control for powered wheelchairs: A vision-based approach. *IEEE Access* 11 (2023), 65663–65674.

[12] Lei Chen, Yingtao Zhang, Xuanming Zhang, and Yun Zhang. 2024. Immersive Teleoperation Platform Based on Eye Tracker. In *2024 6th International Symposium on Robotics & Intelligent Manufacturing Technology (ISRIMT)*. IEEE, 282–286.

[13] Victor Chen, Xuhai Xu, Richard Li, Yuanchun Shi, Shwetak Patel, and Yuntao Wang. 2021. Understanding the Design Space of Mouth Microgestures. In *Designing Interactive Systems Conference 2021*. 1068–1081.

[14] Xinyu Chu et al. 2023. MobileVLM: A Fast, Strong and Open Vision-Language Model. *arXiv preprint arXiv:2312.16886* (2023). https://arxiv.org/abs/2312.16886

[15] Xinyu Chu et al. 2024. MobileVLM V2: Faster and Stronger Baseline for Vision-Language Models. *arXiv preprint arXiv:2402.03766* (2024). https://arxiv.org/abs/2402.03766

[16] Seung Youn Chyung, Katherine Roberts, Ieva Swanson, and Andrea Hankinson. 2017. Evidence-Based Survey Design: The Use of a Midpoint on the Likert Scale. *Performance Improvement* 56, 10 (2017), 15–23. doi:10.1002/pfi.21727

[17] Muratcan Cicek, Ankit Dave, Wenxin Feng, Michael Xuelin Huang, Julia Katherine Haines, and Jeffry Nichols. 2020. Designing and evaluating head-based pointing on smartphones for people with motor impairments. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–12.

[18] Davide Ciraolo, Antonio Celesti, Maria Fazio, Mirjam Bonanno, Massimo Villari, and Rocco Salvatore Calabrò. 2023. Emotional artificial intelligence enabled facial expression recognition for tele-rehabilitation: a preliminary study. In *2023 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 1–6.

[19] Jeffrey F Cohn and Karen L Schmidt. 2004. The timing of facial motion: Facial expression and movement analysis. *Philosophical Transactions of the Royal Society B* 359, 1441 (2004), 1241–1249.

[20] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. 2001. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 6 (2001), 681–685.

[21] Cristiana D'Anna, Michele Mucci, and Rodolfo Vastola. 2021. Perceived motor competence and self-efficacy in children: Competitive sports vs sedentary lifestyle. (2021).

[22] An De Meester, Lisa M Barnett, Ali Brian, Steven J Bowe, Judith Jiménez-Díaz, Femke Van Duyse, J Megan Irwin, David F Stodden, Eva D'Hondt, Matthieu Lenoir, et al. 2020. The relationship between actual and perceived motor competence in children, adolescents and young adults: A systematic review and meta-analysis. *Sports Medicine* 50, 11 (2020), 2001–2049.

[23] Christine Dierk, Scott A Carter, Patrick Chiu, Tony Dunnigan, and Don Kimber. 2019. Use Your Head! Exploring Interaction Modalities for Hat Technologies.. In *Conference on Designing Interactive Systems*. 1033–1045.

[24] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski. 1999. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21, 10 (1999), 974–989. doi:10.1109/34.799905

[25] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[26] Isaac Estevan and Lisa M Barnett. 2018. Considerations related to the definition, measurement and analysis of perceived motor competence. *Sports Medicine* 48, 12 (2018), 2685–2694.

[27] Cheikh Latyr Fall, Francis Quevillon, Martine Blouin, Simon Latour, Alexandre Campeau-Lecours, Clément Gosselin, and Benoit Gosselin. 2018. A multimodal adaptive wireless control interface for people with upper-body disabilities. *IEEE transactions on biomedical circuits and systems* 12, 3 (2018), 564–575.

[28] Mingming Fan, Zhen Li, and Franklin Mingzhe Li. 2020. Eyelid gestures on mobile devices for people with motor impairments. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–8.

[29] Yuting Fan et al. 2022. Eyelid Gestures for People with Motor Impairments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–24.

[30] Wenxin Feng, Jiangnan Zou, Andrew Kurauchi, Carlos H Morimoto, and Margrit Betke. 2021. Hgaze typing: Head-gesture assisted gaze typing. In *ACM Symposium on eye tracking research and applications*. 1–11.

[31] Victor Galaz, Miguel A Centeno, Peter W Callahan, Amar Causevic, Thayer Patterson, Irina Brass, Seth Baum, Darryl Farber, Joern Fischer, David Garcia, et al. 2021. Artificial intelligence, systemic risks, and sustainability. *Technology in society* 67 (2021), 101741.

[32] Simon C Gandevia. 1996. Proprioceptive loss and motor control. *Brain* 119, 4 (1996), 1285–1307.

[33] Clare Garvie and Jonathan Frankle. 2016. Facial-recognition software might have a racial bias problem. *The Atlantic* 7, 04 (2016), 2017.

[34] Tan Gemicioglu, R Michael Winters, Yu-Te Wang, Thomas M Gable, and Ivan J Tashev. 2023. TongueTap: Multimodal Tongue Gesture Recognition with Head-Worn Devices. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 564–573.

[35] Muzzamil Ghaffar, Shakil R Sheikh, Noman Naseer, and Fraz Ahmed. 2021. Assistive smart home environment using head gestures and EEG eye blink control schemes. In *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*. IEEE, 1–6.

[36] Lorenzo Giampietro. 2025. Serious Games for patient rehabilitation using Mediapipe and Unity. In *Master thesis*. Universitat Politècnica de Catalunya.

[37] Google. 2023. MediaPipe Face Landmarker. https://github.com/google-ai-edge/mediapipe-samples/tree/main/examples/face_landmarker. Accessed: 2025-01-18.

[38] Philip Graybill and Mehdi Kiani. 2018. Eyelid drive system: An assistive technology employing inductive sensing of eyelid movement. *IEEE transactions on biomedical circuits and systems* 13, 1 (2018), 203–213.

[39] Anhong Guo et al. 2019. Toward fairness in AI for people with disabilities: A research roadmap. *ACM SIGACCESS Accessibility and Computing* 125 (2019), 31–35.

[40] Peter Hafner and Michal Čerňanskỳ. 2024. Enhancing Piano Accessibility Using Augmented Reality for Education in Piano Classes. In *2024 International Conference on Emerging eLearning Technologies and Applications (ICETA)*. IEEE, 162–167.

[41] Md Farhan Hamid and Md Ashraful Alam. 2017. Virtual wardrobe for physically impaired using Microsoft Kinect sensor. In *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*.

[42] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–13.

[43] Md Mehedi Hassain, Md Fakwer Uddin Mazumder, Md Reyad Arefin, and Mohammed Abdul Kader. 2023. Design and implementation of smart head-motion controlled wheelchair. In *2023 IEEE Engineering Informatics*. IEEE, 1–8.

[44] Agneta Herlitz and Johanna Lovén. 2013. Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition* 21, 9-10 (2013), 1306–1336.

[45] Jiewen Hu, Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. 2025. OpenFace 3.0: A Lightweight Multitask System for Comprehensive Facial Behavior Analysis. *arXiv preprint arXiv:2506.02891* (2025).

[46] Md Touhidul Islam, Imran Kabir, Md Alimoor Reza, and Syed Masum Billah. 2025. IKIWISI: An Interactive Visual Pattern Generator for Evaluating the Reliability of Vision-Language Models Without Ground Truth. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA.

[47] Puneet Jain and Chris Lloyd Salter. 2025. " Can I use my mouth to navigate in VR?": Integrating Mouth Gestures in XR with (and for) Disabled Artists through Criptastic Hacking. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[48] Hairong Jiang, Bradley S Duerstock, and Juan P Wachs. 2014. A machine vision-based gestural interface for people with upper extremity physical impairments. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44, 5 (2014), 630–641.

[49] Yifan Jin, Yi Li, Qiaolin Chen, et al. 2024. Efficient Multimodal Large Language Models: A Survey. *arXiv preprint arXiv:2405.10739* (2024). https://arxiv.org/abs/2405.10739

[50] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*. PMLR, 2439–2448.

[51] Miloš Kankaraš and Stefania Capecchi. 2025. Neither Agree nor Disagree: Use and Misuse of the Neutral Response Category in Likert-Type Scales. *METRON* 83, 1 (2025), 111–140. doi:10.1007/s40300-024-00276-5

[52] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–22.

[53] Jinhyuk Kim, Jaekwang Cha, Hojun Lee, and Shiho Kim. 2017. Hand-free natural user interface for VR HMD with IR based facial gesture tracking sensor. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*. 1–2.

[54] Tomasz Kocejko, Jacek Ruminski, Piotr Przystup, Artur Polinski, and Jerzy Wtorek. 2017. The role of EMG module in hybrid interface of prosthetic arm. In *2017 10th International Conference on Human System Interactions (HSI)*. IEEE, 36–40.

[55] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A Lee, and Mark Billinghurst. 2018. Pinpointing: Precise head-and eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.

[56] Amit Laddi, Vijay Bhardwaj, Nishant Kapoor, Dinesh Pankaj, and Amod Kumar. 2015. Unobtrusive head gesture based directional control system for patient mobility cart. In *2015 International Conference on Signal Processing, Computing and Control (ISPCC)*. IEEE, 236–240.

[57] Arto Laukkanen, Donna Niemistö, Kaisa Aunola, Lisa M Barnett, and Arja Sääkslahti. 2023. Child perceived motor competence as a moderator between physical activity parenting and child objectively measured physical activity. *Psychology of Sport and Exercise* 67 (2023), 102444.

[58] Zhen Li, Mingming Fan, Ying Han, and Khai N Truong. 2020. iWink: Exploring eyelid gestures on mobile devices. In *Proceedings of the 1st International Workshop on Human-Centric Multimedia Analysis*. 83–89.

[59] Jiajie Lin, Hanhong Tan, Jinglin Deng, Xiaoyan Li, and Yangzhou Feng. 2024. Design of Mobile Phone Holder Based on Face Recognition. In *2024 5th International Conference on Intelligent Design (ICID)*. 218–221. doi:10.1109/ICID64166.2024.11024823

[60] Haotian Liu, Chunyuan Li, et al. 2023. Improved Baselines with Visual Instruction Tuning. *arXiv preprint arXiv:2310.03744* (2023). https://arxiv.org/abs/2310.03744

[61] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[62] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. 2017. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing* 10, 3 (2017), 325–347.

[63] Katsutoshi Masai, Kai Kunze, Daisuke Sakamoto, Yuta Sugiura, and Maki Sugimoto. 2020. Face commands-user-defined facial gestures for smart glasses. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 374–386.

[64] Denys JC Matthies, Jan-Niklas Antons, Frank Heidmann, Reto Wettach, and Robert Schleicher. 2012. NeuroPad: use cases for a mobile physiological interface. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*. 795–796.

[65] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. 2013. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing* 4, 2 (2013), 151–160.

[66] J. T. Nadler, R. Weston, and E. C. Voyles. 2015. Stuck in the Middle: The Use and Interpretation of Mid-Points in Items on Questionnaires. *The Journal of General Psychology* 142, 2 (2015), 71–89. doi:10.1080/00221309.2014.994590

[67] Donna Niemistö, Lisa M Barnett, Arto Laukkanen, Asko Tolvanen, and Arja Sääkslahti. 2023. Perceived motor competence in early childhood predicts perceived and actual motor competence in middle childhood. *Scandinavian Journal of Medicine & Science in Sports* 33, 10 (2023), 2025–2038.

[68] Tomi Nukarinen, Jari Kangas, Oleg Špakov, Poika Isokoski, Deepak Akkil, Jussi Rantala, and Roope Raisamo. 2016. Evaluation of HeadTurn: An interaction technique using the gaze and head turns. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. 1–8.

[69] Sharan Pai and Anish Bhardwaj. 2019. Eye gesture based communication for people with motor disabilities in developing nations. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[70] Milan Pandey, Kushal Chaudhari, Rajnish Kumar, Anoop Shinde, Divyanshu Totla, and ND Mali. 2018. Assistance for paralyzed patient using eye motion detection. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. 1–5.

[71] Raji Pandurangan, S Gowtham, and J Gugan. 2025. Neural Network based Vision-Driven Navigation Aid for Individuals with Physical Disabilities. In *2025 International Conference on Electronics and Renewable Systems (ICEARS)*. IEEE, 1436–1441.

[72] Theja Ram Pingali, Sarthak Dubey, Anurag Shivaprasad, Arpit Varshney, Satish Ravishankar, Govinda Ram Pingali, Niyanth Krishna Polisetty, Nikhil Manjunath, and Padmaja KV. 2014. Eye-gesture controlled intelligent wheelchair using Electro-Oculography. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2065–2068.

[73] Uwe Proske and Simon C Gandevia. 2012. Proprioception: the sense of self. *Physiological reviews* 92, 4 (2012), 1651–1697.

[74] Wahyu Rahmaniar, Alfian Ma'Arif, and Ting-Lan Lin. 2022. Touchless head-control (thc): Head gesture recognition for cursor and orientation control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022), 1817–1828.

[75] Reza Rahnama-ye Moqaddam and Hamed Vahdat-Nejad. 2015. Designing a pervasive eye movement-based system for ALS and paralyzed patients. In *2015 5th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE, 218–221.

[76] Akula Kumar Raja, Chidakash Sugandhi, Gorantla Nymish, Nama Sai Havish, and Manazhy Rashmi. 2023. Face Gesture Based Virtual Mouse Using Mediapipe. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*. IEEE, 1–6.

[77] Vijay Rajanna and Tracy Hammond. 2018. A gaze gesture-based paradigm for situational impairments, accessibility, and rich interactions. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. 1–3.

[78] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Remi Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 145–151.

[79] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)* 54, 9 (2021), 1–40.

[80] Matthew G Rhodes and Jeffrey S Anastasi. 2012. The own-age bias in face recognition: a meta-analytic and theoretical review. *Psychological bulletin* 138, 1 (2012), 146.

[81] Khandaker Jannatul Ritu, Khalil Ahammad, Md Mohibullah, Mahmuda Khatun, Md Zasim Uddin, Md Kamal Uddin, Yoshinori Kobayashi, and Mahmudul Hasan. 2023. SelfBOT: An Automated Wheel-Chair Control Using Facial Gestures Only. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 1–6.

[82] Lawrence D Rosenblum, Mark A Schmuckler, and Jennifer A Johnson. 2002. The Kinematics of Motor Control. *Journal of Experimental Psychology: Human Perception and Performance* 28, 3 (2002), 489–497.

[83] Michael Ruocco, Jack Duggan, Chetan Jaiswal, Brian O'Neill, and Karen Majeski. 2024. Leveraging AI Face-Tracking and Gesture Recognition for Hands-Free Computing: Bridging the Gap for Users with Physical Disabilities. In *2024 IEEE Global Humanitarian Technology Conference (GHTC)*. IEEE, 232–239.

[84] Katsuya Sakai, Yuichiro Hosoi, Yusuke Harada, Yumi Ikeda, and Junpei Tanabe. 2024. Overestimation associated with walking and balance function in individuals diagnosed with a stroke. *Physiotherapy Theory and Practice* 40, 7 (2024), 1404–1411.

[85] Katsuya Sakai and Yumi Ikeda. 2019. Clinical assessment of motor imagery and physical function in mild stroke patients. *Journal of Physical Therapy Science* 31, 12 (2019), 992–996.

[86] Ryota Sakurai, Yoshinori Fujiwara, Hiroyuki Suzuki, Susumu Ogawa, Takahiro Higuchi, and Kuniyasu Imanaka. 2021. Changes in self-estimated step-over ability among older adults: A 3-year follow-up study. *The Journals of Gerontology: Series B* 76, 10 (2021), 2003–2012.

[87] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. 2008. Bosphorus database for 3D face analysis. In *Biometrics and Identity Management: First European Workshop, BIOID 2008, Roskilde, Denmark, May 7-9, 2008. Revised Selected Papers 1*. Springer, 47–56.

[88] Daniel L Schacter, Daniel T Gilbert, and Daniel M Wegner. 2009. *Psychology.* Macmillan.

[89] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. 2020. HCI guidelines for gender equity and inclusivity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[90] Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. 2020. Diagnosing gender bias in image recognition systems. *Socius* 6 (2020), 2378023120967171.

[91] Barbara Seebacher, Markus Reindl, and Turhan Kahraman. 2023. Factors and strategies affecting motor imagery ability in people with multiple sclerosis: a systematic review. *Physiotherapy* 118 (2023), 64–78.

[92] Ahmed Sharshar et al. 2025. Vision-Language Models for Edge Networks: A Comprehensive Survey. *arXiv preprint arXiv:2502.07855* (2025). https://arxiv.org/abs/2502.07855

[93] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.

[94] Muhammad Faiz Ahmad Sobri, Zakaria Hussain, Saiful Zaimy Yahaya, Rozan Boudville, and Noraiza Aqilah Abdul Aziz. 2022. Facial Expression Electric Wheelchair Control Instruction Using Image Processing. In *2022 IEEE 12th International Conference on Control System, Computing and Engineering (ICCSCE)*. IEEE, 168–173.

[95] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision* 130, 6 (2022), 1526–1565.

[96] Qin Sun, Yunqi Hu, Mingming Fan, Jingting Li, and Su-Jing Wang. 2024. "Can It Be Customized According to My Motor Abilities?": Toward Designing User-Defined Head Gestures for People with Dystonia. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–11.

[97] Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* 2, 8 (2019), 73.

[98] Muhammed Oğuz Tas and Hasan Serhan Yavuz. 2022. A human-computer interaction system based on eye, eyebrow and head movements. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi* 28, 5 (2022), 632–642.

[99] Jingze Tian, Yingna Wang, Keye Yu, Liyi Xu, Junan Xie, Franklin Mingzhe Li, Yafeng Niu, and Mingming Fan. 2024. Designing Upper-Body Gesture Interaction with and for People with Spinal Muscular Atrophy in VR. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[100] Michel Valstar and Maja Pantic. 2006. Spontaneous vs. posed facial behaviour: Automatic analysis of brow actions. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 162–167.

[101] Pedro Velez and Stephen D. Ashworth. 2007. The Impact of Item Readability on the Endorsement of the Middle Response Option. *Survey Research Methods* 1, 2 (2007), 69–79.

[102] Krishna Venkatasubramanian, Haven Hardie, and Tina-Marie Ranalli. 2025. Toward a Taxonomy of Negative Outcomes from the Use of AI-Driven Systems for People with Disabilities. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*. 1–18. doi:10.1145/3663547.3746359

[103] Ker-Jiun Wang, Quanbo Liu, Yifan Zhao, Caroline Yan Zheng, Soumya Vhasure, Quanfeng Liu, Prakash Thakur, Mingui Sun, and Zhi-Hong Mao. 2018. Intelligent wearable virtual reality (VR) gaming controller for people with motor disabilities. In *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 161–164.

[104] L. Wang, V. Kameswaran, and H. Kacorri. 2025. Toward a Taxonomy of Algorithmic Harms for Disability: A Systematic Review. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. 2649–2665. doi:10.1609/aies.v8i3.36745

[105] Davy PY Wong and Pai H Chou. 2023. TongueMendous: IR-Based Tongue-Gesture Interface with Tiny Machine Learning. In *Proceedings of the 8th international Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*. 1–8.

[106] Liwei Wu, Ben Lafreniere, Tovi Grossman, Thomas White, and Stephanie Santosa. 2024. Body Language for VUIs: Exploring Gestures to Enhance Interactions with Voice User Interfaces. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 133–150.

[107] Jiajun Xu, Zhiyuan Li, Wei Chen, Qun Wang, Xin Gao, Qi Cai, and Ziyuan Ling. 2024. On-Device Language Models: A Comprehensive Review. *arXiv preprint arXiv:2409.00088* (2024). https://arxiv.org/abs/2409.00088

[108] Yukang Yan, Chun Yu, Xin Yi, and Yuanchun Shi. 2018. Headgesture: Hands-free input approach leveraging head movements for hmd devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–23.

[109] Peng Yang, Qingshan Liu, and Dimitris N Metaxas. 2010. Exploring facial expressions with compositional features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2638–2644.

[110] Junsong Yuan, Zicheng Liu, and Ying Wu. 2009. Discriminative subvolume search for efficient action detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2442–2449.

[111] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. 2014. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing* 32, 10 (2014), 692–706.

[112] Xuan Zhao, Mingming Fan, and Teng Han. 2022. "I Don't Want People to Look At Me Differently" Designing User-Defined Above-the-Neck Gestures for People with Upper Body Motor Impairments. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[113] Peizhao Zhu, Yuzheng Zhu, Wenyuan Li, Yanbo He, Yongpan Zou, Kaishun Wu, and Victor CM Leung. 2025. Char: Composite head-body activities recognition with a single earable device. *IEEE Transactions on Mobile Computing* (2025).

A Sensorimotor Background

Gesture-based interfaces rely not only on the user's ability to perform a recognisable motion, but also on their internal capacity to perceive that motion accurately. In populations with motor impairments, this perceptual ability is often disrupted. Understanding how individuals experience and evaluate their own motor actions, particularly in facial gestures, requires attention to the cognitive and physiological basis of motor perception.

A.1 Proprioception and Motor Feedback

Proprioception refers to the sense of the relative position and movement of one's body parts, informed by sensory receptors in muscles, joints, and tendons [73]. It enables individuals to monitor and adjust movement without relying solely on visual cues. In healthy users, proprioceptive signals allow for precise coordination and correction of facial gestures, such as blinking or brow movement [73]. However, in motor-impaired individuals, especially those with neurological disorders such as cerebral palsy, ALS, or muscular dystrophy, proprioceptive feedback can be degraded or delayed [32]. This can result in inaccurate internal representations of movement, leading to over- or under-estimation of gesture execution [84–86, 91].

A.2 Perceived Motor Competence in Motor-Impaired Populations

Perceived motor competence (PMC) is the individual's subjective awareness of their ability to perform a movement [22, 26, 88]. In rehabilitation science and developmental psychology, PMC is recognised as a critical factor influencing motivation, performance, and self-regulation in motor tasks [21, 57, 67]. For individuals with motor impairments, PMC may be compromised due to inconsistent movement outcomes, reduced sensory feedback, or lifelong experiences of motor difficulty.

In our study, this phenomenon emerged repeatedly in participant interviews. For example, P10 commented,"I tried moving my brows, but I'm not sure if both moved or just one," while P1 noted, "I can't tell if my smile counts, this is what my smile looks like." These statements reflect not only mechanical execution difficulties but also perceptual ambiguity: the inability to verify one's own motor state in real time. Facial gestures are particularly vulnerable to such ambiguity because they often occur outside the user's visual field and offer limited tactile or kinesthetic feedback.

A.3 Implications for Gesture Recognition

Gesture recognition models assume that users can reproduce spatially consistent gestures that match learned patterns [110]. However, when proprioception and PMC are compromised, users may overestimate success even when motion trajectories deviate from canonical templates. These mismatches reflect **allocative harms**, which are not just failed recognition, but systematic exclusion due to unmodeled sensorimotor diversity. Our diagnostic method addresses this by aligning model interpretation with user experience through multi-level data: accuracy, perception, and gesture kinematics.

B Facial Gesture Set

This appendix provides a structured summary of the final facial gesture set used in our study, including gesture categories, associated interaction mappings, and prior evidence of user preference reported in the literature. Table 6 organises these gestures by facial region and action mode, and situates them within prior above-neck interaction research.

C Video Clips for Gesture Set

We created video clips (fig.5) for each gesture. All the video clips have been reviewed by two researchers to confirm they are clearly expressed and unambiguous. Then, these clips were tested by the MediaPipe Face Tracking API[61]. All the clips can be recognised as the corresponding gesture and achieve a confidence score over 0.6 to ensure accurate recognition by the facial gestures and head pose gesture recognition API from MediaPipe. To facilitate participant understanding and reduce cognitive load associated with directional interpretation, all video clips were horizontally mirrored, allowing participants to perceive the gestures as if performed by themselves.

D Gesture Operationalisation Rules

This appendix documents the operational rules used to compute gesture activation for the two recognition pipelines evaluated in this study: MediaPipe-based gesture activations and OpenFace AU-based recognition. The goal is to make the gesture definitions transparent and reproducible, rather than to introduce additional modelling or learning.

MediaPipe exposes semantically defined activation signals for coarse facial actions (e.g., *EyeBlinkLeft*, *MouthSmileLeft*). For gestures included in MediaPipe's native gesture set, we directly used the provided activation outputs. For additional gestures not explicitly defined by MediaPipe, we derived operational definitions by combining these activation signals using fixed, rule-based computations. These MediaPipe-based computation rules are summarised in Table 7. No thresholds or parameters were adapted during the study.

OpenFace provides per-frame AU intensity estimates. In our study, OpenFace outputs were used directly without additional rule-based computation, except for eyebrow-raise gestures. Specifically, eyebrow raising was operationalised as $\max(\mathrm{AU1}, \mathrm{AU2})$ to capture the stronger activation between inner and outer brow movements. No other aggregation or thresholding was applied to OpenFace AU signals.

E Results of OpenFace 3

At the time of submission, OpenFace 3.0 provides intensity estimates for eight facial action units (AUs): AU1 (Inner Brow Raiser), AU2 (Outer Brow Raiser), AU4 (Brow Lowerer), AU6 (Cheek Raiser), AU9 (Nose Wrinkler), AU12 (Lip Corner Puller), AU25 (Lips Part), and AU26 (Jaw Drop). Based on prior literature and our formative investigation (Table 6), we found that AU6 (Cheek Raiser) and AU9 (Nose Wrinkler) were rarely selected by users as intentional control gestures. We therefore focused our analysis on five AUs that better aligned with users' preferred gesture semantics, covering mouth opening, mouth corner raising, eyebrow frowning, eyebrow raising, and jaw movement. The operational definitions used to compute these gestures from OpenFace outputs are detailed in Appendix D. Detailed quantitative results are reported in Table 8.

F Per-Gesture Group Comparisons

This section presents detailed comparisons between Group A (motor-impaired) and Group B (non-impaired) on a per-gesture basis. We report gesture-level recognition accuracy, self-evaluation scores, and perception–recognition alignment statistics. Table 9 summarizes the recognition performance across 37 gestures, highlighting significant disparities in accuracy between groups. Table 10 provides corresponding self-evaluation score comparisons, revealing subjective confidence trends. Table 11 reports gesture-level alignment metrics, including match rates and the prevalence of over- and underestimation errors, offering a more granular view of how intention–recognition mismatches manifest across gestures.

G Trajectory-Based Gesture Variation

We generated illustrations based on the landmark trajectories to demonstrate the diverse gesture trajectories in Group A, thereby providing readers with an intuitive understanding of the variability of gestures among PwM. In these illustrations, the gesture motion trajectories directly mirror the landmark trajectories to objectively depict the gestures. To protect the participants' privacy, a neutral outline was used for the head and facial features. Figures 6 (a')–(e') illustrate how landmark trajectories were rendered step by step to visualise gesture execution while preserving participant anonymity.

We visualised the trajectories of 12 representative gestures that exhibited an intention-recognition mismatch in 7. These gestures were categorised based on key movement characteristics, including amplitude (distance) differences, directional differences, speed differences, and differences in the activated body region. For comparison, we selected the trajectories with the highest recognition accuracy for the corresponding gestures in Group B. In each figure, the top row shows the trajectories from the control group, while the bottom row shows the trajectories of participants with motor impairments.

Table 6: Summary of Above-Neck Gestures Identified in Prior Literature

Taxonomy	Gestures	Action Mode	Related Interaction	Interface	User Preference
Head Gesture	Turn/Pitch Head Up	one	Tap, Flick, Scroll/Volume/Pan/Cursor Up, Drag, Previous container, Zoom in, Go forward	[6, 11, 27, 41, 113][56, 74, 76, 81]	[23, 43, 96, 108, 112]
	Turn/Pitch Head Down	one	Drag, Next Container, Volume down, Scroll Down, Select, zoom out, cursor down, go backward	[6, 11, 27, 41, 56, 74, 76, 81, 113]	[43, 96, 108, 112]
	Turn Head Left	one	Swipe/Pan/Cursor/Go to Left, Drag, Rotate, Open, Previous/Next button/target/app/screen, Pause/Hold	[1, 53, 64, 98]	[23, 96, 106, 108, 112]
	Turn Head Right	one	Swipe/Pan/Cursor/Go to Right, Drag, Open, Next button/target/app, Pause/Hold,	[1, 53, 64, 98]	[23, 96, 106, 108, 112]
	Roll/Tilt Head Left	one	Glid to Left, Swipe Left, Rotate, Home, zoom in, left mouse click, activate, go left	[6, 30, 35, 74, 113]	[43, 96, 108, 112]
	Roll/Tilt Head Right	one	Glid to Right, Swipe Right, Rotate, Home, zoom out, right mouse click, go right	[6, 30, 35, 74, 113]	[43, 96, 108, 112]
	Protrude Head	one	Zoom out, Next button, Open background, drag, hold, select,activate	[35]	[96, 108, 112]
	Retract Head	one	Zoom in, Previous Button, drag, select,activate	[35]	[96, 108, 112]
Eyebrow Gesture	Raising Eyebrows	one	Double Click	[76]	
	Frown Eyebrows	one	Go right	[94]	
Eyelid Gesture	Blink Eyes	one	Open the App/System/Container, Next button, Lock, Tap, Screenshot, Change, Lock/Unlock, Select	[28, 35, 41, 58, 69, 70]	[28, 96, 112]
	Left Wink	one	Volume down, Left Click, Go left, Activate	[10, 58, 64, 98]	
	Right Wink	one	Selection, Pause, Mouse Right Click, Point Right	[28, 38, 58, 69, 76, 81]	[28]
	Close/Shut Eyes	zero	Long press, Phone lock, initialize, stop, select	[28, 56, 58, 75]	[28, 96, 112]
	Close/Shut Left Eye	zero	-	[28, 58]	[28]
	Close/Shut Right Eye	zero	-	[28, 58]	[28]
	Squint / Squeez eyes	zero	Zoom out, Activate/Deactivate Scrolling	[28, 58, 81]	[28, 96, 112]
	Wide Open eyes	Single	Go Forward	[24, 103]	
Eye Ball Gesture	Eye Look Up	One	Scroll Up, Drag, Previous container, Go North/Forward, Rotate Up	[12, 54, 69, 70, 72, 77]	[96, 112]
	Eye Look Down	one	Scroll Down, Drag, Next container, Go South/Backward, Rotate Down	[12, 54, 69, 70, 72, 77]	[96, 112]
	Eye Look Left	One	Back, Swipe Left, Drag, Move to the previous/next/left screen/target/app, Go west/left, Rotate left	[12, 54, 69, 70, 72, 75, 77] [96, 112]	
	Eye Look Right	One	Forward, Swipe/Pan/Go/Rotate to Right, Drag, Move to previous/next/right button/app	[12, 69, 70, 72, 75, 77]	[23, 96, 112]
Cheek Gesture	Puff Out Cheeks	one	Long Press		[96]
Mouth Gesture	Flick, Go forward	zero	Selection,Go Backward	[94]	[13, 96, 112]
	Depress Mouth Corners	zero	Drag, Go right	[94]	[96, 112]
	Pucker/Frown Lips	zero	Volume Down		[13, 96, 112]
	Suck Lips	zero	Volume Down		[96, 112]
	Open Mouth	zero	Selection, Mouse Right Click,Shooting	[40, 76, 81]	[13, 96, 112]
Tooth Gesture	GrindTeeth	zero	Drag	[34]	[13, 96]
Tongue Gesture	Stick Tongue Out	zero	Zoom out, Drag, Volume up	-	[105] [13, 96]
	Stick Tongue Out Left	zero	-	[13, 105]	
	Stick Tongue Out Right	zero	-	[13, 105]	
Jaw Gesture	Move Jaw Forward	zero	Selection	-	[13]

(a) Amplitude differences. (a1) and (a1') are the 'turn heads up' gestures of Pb9 and P8. (a1) shows a large-amplitude upward head tilt, recognised by the algorithm (Accuracy = 99.68%) and self rated highly (Self-score = 8). In contrast, (a1') shows limited movement due to motor variability; it was not recognised (Accuracy = 23.53%), though **participant P8** self-rated it as completed (Self-score = 8).

(a2) and (a2') show the gesture 'Raise eyebrows' of Pb9 and P10. (a2) was recognised (accuracy = 92. 55%) and self-rated (self-score = 8). In (a2'), limited movement in the right eyebrow which due to muscle strength variation led to misrecognition as "Raise Right Eyebrow" (Accuracy = 62.43%); The accuracy of "Raise Eyebrows" is only 33.86%, **participant P4** rated it as completed (Self-score = 8).

(a3) and (a3') depict the "Raise Mouth Corners" gesture from Pb2 and P4. (a3) was recognised (Accuracy = 91.57%) and rated highly (Self-score = 8). In (a3'), reduced amplitude due to proprioceptive

Figure 5: The video clips of Head Gestures.

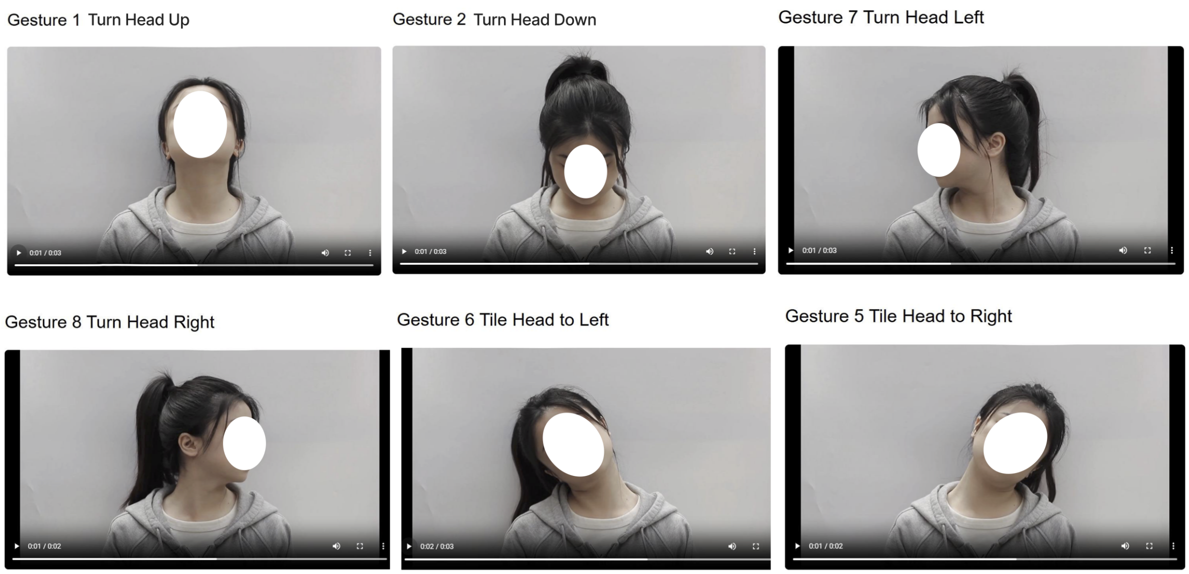


Table 7: AU-based Computation Rules for Gestures

Gesture	Computation Rule
FrownEyebrows	(browDownLeft + browDownRight) / 2
BlinkEyes	(eyeBlinkLeft + eyeBlinkRight) / 2
LeftWink	if eyeBlinkLeft > 0.5 > eyeBlinkRight, = eyeBlinkLeft, else eyeBlinkLeft - eyeBlinkRight
RightWink	if eyeBlinkRight > 0.5 > eyeBlinkLeft, = eyeBlinkRight, else eyeBlinkRight - eyeBlinkLeft
CloseEyes	max(eyeBlinkLeft, eyeBlinkRight)
CloseLeftEye	if eyeBlinkLeft > 0.5 > eyeBlinkRight, = eyeBlinkLeft, else eyeBlinkLeft - eyeBlinkRight
CloseRightEye	if eyeBlinkRight > 0.5 > eyeBlinkLeft, = eyeBlinkRight, else eyeBlinkRight - eyeBlinkLeft
OpenLeftEye	if eyeBlinkRight > 0.5 > eyeBlinkLeft, = eyeBlinkRight, else eyeBlinkRight - eyeBlinkLeft
OpenRightEye	if eyeBlinkLeft > 0.5 > eyeBlinkRight, = eyeBlinkLeft, else eyeBlinkLeft - eyeBlinkRight
SquintEyes	(eyeSquintLeft + eyeSquintRight) / 2
WideOpenEyes	(eyeSquintLeft_end + eyeSquintRight_end - eyeSquintLeft_start - eyeSquintRight_start) / ((eyeSquintLeft_end + eyeSquintRight_end)/2)
EyeLookUp	(eyeLookUpLeft + eyeLookUpRight) / 2
EyeLookDown	(eyeLookDownLeft + eyeLookDownRight) / 2
EyeLookLeft	min(eyeLookOutLeft, eyeLookInRight)
EyeLookRight	min(eyeLookOutRight, eyeLookInLeft)
PuffOutCheeks	1 - (cheekSquintLeft + cheekSquintRight) / 2
RaiseMouthCorners	(mouthSmileLeft + mouthDimpleLeft + mouthSmileRight + mouthDimpleRight) / 2
DepressMouthCorners	1 - RaiseMouthCorners
MoveJawForward	(jawForward_end - jawForward_start) / jawForward_end

differences led to misrecognition (Accuracy = 37.38%), although **participant P4** self-rated it as completed (Self-score = 8).

(b) Directional differences. (b1) and (b1') show "Left Wink" from Pb3 and P11. The image of Figure (b1) was recognised (Accuracy = 81.76%) and self-rated (Self-score = 8). In (b1'), directional deviation due to left and right disorientation caused recognition failure (Accuracy = 0.09%), though **participant P11** reported success (Self-score = 8).

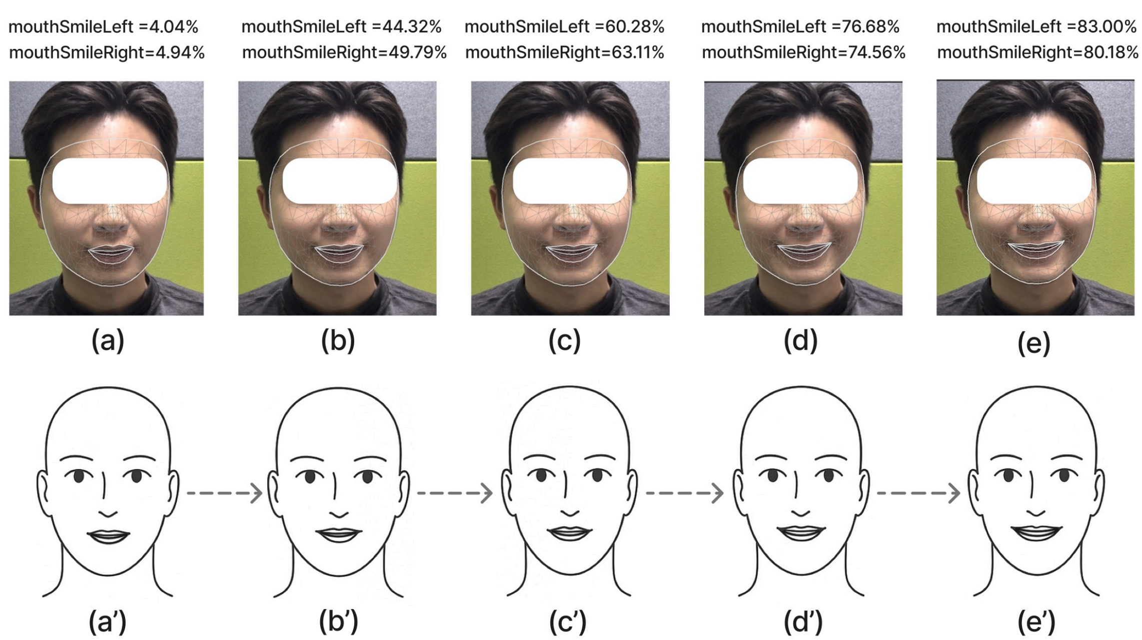
(b2) and (b2') show "Eye Look Left" from Pb2 and P10. The image of (b2) was recognised (Accuracy = 96.34%) and self-rated (Self-score = 8). In (b2'), directional drift due to motor control diversity caused misrecognition (Accuracy = 21.50%), though **participant P10** self-rated it as completed (Self-score = 8).

(b3) and (b3') again show "Raise Mouth Corners" from Pb2 and P1. (b3) was recognised (Accuracy = 91.57%) and self-rated (Self-score = 8). In (b3'), direction was limited due to strength asymmetry,

Table 8: OpenFace 3.0 action unit activation values for facial gestures across participants.

Participants	RaiseEyebrows (max(AU1,AU2))	FrownEyebrows (AU4)	RaiseMouthCorners(AU12)	OpenMouth(AU25)
P1	0.813158	0.947782	0	0.997394
p2	0.881189	0.995293	0.0104065	0.996935
p3	0.115445	0.0373973	0.0609479	0.997265
p4	0.571817	0.220495	0.0153467	0.996642
p5	0.993033	0	0.261731	0.998906
p6	0.499899	0	0.739779	0.994848
p7	0.699864	0.00160084	0.604374	0.997898
p8	0.991935	0	0.803924	0.978151
p9	0.740064	0.00486127	0.0214302	0.941494
p10	0.295701	0.153301	0.702918	0.938162
p11	0.0647594	0.837812	0	0.810929
Pb1	0.99189	0.52239	0.830408	0.998887
Pb2	0.975746	0.0849971	0.873705	0.997516
Pb3	0.997973	0.998792	0.579678	0.997421
Pb4	0.823124	0.993204	0.994875	0.995307
Pb5	0.895094	0.996867	0.757573	0.995689
Pb6	0.588846	0.962703	0.467995	0.997415
Pb7	0.997752	0.010618	0.0207168	0.995915
Pb8	0.00275466	0.00364981	0.159954	0.988634
Pb9	0.905312	0.998595	0.867902	0.994219
Pb10	0.621479	0.776821	0.715707	0.981305
Pb11	0.660214	0.885362	0.989801	0.997816

Figure 6: The process of deriving a gesture trajectory for a 'raise mouth corners' gesture from frame-level activation signals.



resulting in misrecognition as "Mouth Stretch" (Accuracy = 41.59%); the accuracy of Raise Mouth Corners was 0.21%, **participant P1** self-rated it as completed (Self-score = 8).

(c) Speed differences. (c1) and (c1') are "Turn Head Right" from Pb2 and P3. (c1) was recognised (Accuracy = 99.01%) and self-rated (Self-score = 8). In (c1'), a delayed motion due to impaired motor control caused misrecognition as "Blink Eyes" (Accuracy = 61.72%),

Table 9: Per-gesture comparison of recognition accuracy between Group A and Group B with Benjamini Hochberg (BH) FDR correction.

Gesture	A (%)	B (%)	U	p	BH-adjusted p	BH threshold
TurnHeadUp	55.0 ± 33.0	82.1 ± 12.2	33.0	0.076	0.2560	0.0405
TurnHeadDown	63.2 ± 29.3	80.2 ± 9.1	45.0	0.325	0.4716	0.0345
TurnHeadLeft	47.0 ± 22.0	77.4 ± 9.7	2.0	0.000999	0.00924	0.0054
TurnHeadRight	63.3 ± 23.5	79.2 ± 18.1	40.0	0.189	0.3253	0.0291
TiltHeadLeft	57.0 ± 27.9	71.9 ± 14.0	40.0	0.189	0.3253	0.0291
TiltHeadRight	58.3 ± 24.2	69.2 ± 15.3	47.0	0.393	0.5193	0.0378
MoveHeadForward	42.4 ± 26.9	81.5 ± 17.1	11.0	0.001	0.00435	0.0115
MoveHeadBackward	37.9 ± 25.4	78.4 ± 21.8	13.0	0.002	0.00705	0.0142
RaiseEyebrows	69.1 ± 19.8	76.2 ± 13.8	51.0	0.555	0.6417	0.0432
FrownEyebrows	61.3 ± 27.0	65.1 ± 6.9	51.0	0.555	0.6417	0.0432
BlinkEyes	58.1 ± 5.6	67.9 ± 9.3	23.0	0.015	0.0370	0.0203
LeftWink	33.8 ± 23.5	71.6 ± 8.8	6.0	0.000999	0.00924	0.0054
RightWink	22.9 ± 23.1	59.9 ± 5.9	12.0	0.002	0.00705	0.0142
CloseEyes	64.1 ± 21.7	73.0 ± 8.0	51.0	0.555	0.6417	0.0432
CloseLeftEye	33.9 ± 26.1	58.1 ± 12.9	27.0	0.030	0.0707	0.0270
CloseRightEye	40.0 ± 24.4	59.0 ± 10.0	30.0	0.049	0.1030	0.0311
OpenLeftEye	45.1 ± 30.7	54.6 ± 17.5	53.0	0.646	0.7030	0.0459
OpenRightEye	23.9 ± 25.7	59.4 ± 13.6	18.0	0.006	0.0171	0.0176
SquintEyes	57.7 ± 12.4	53.8 ± 7.6	73.0	0.431	0.5499	0.0392
WideOpeneyes	60.1 ± 20.4	77.8 ± 12.3	28.0	0.036	0.0814	0.0284
EyeLookUp	56.2 ± 20.0	59.6 ± 11.9	43.0	0.264	0.4070	0.0324
EyeLookDown	46.5 ± 23.2	69.8 ± 9.7	19.0	0.007	0.0185	0.0189
EyeLookLeft	80.1 ± 25.4	85.9 ± 9.1	66.0	0.743	0.7744	0.0480
EyeLookRight	76.2 ± 28.5	78.2 ± 10.3	80.0	0.212	0.3410	0.0311
PuffOutCheeks	66.8 ± 17.1	71.2 ± 18.3	45.0	0.325	0.4716	0.0345
RaiseMouthCorners	38.9 ± 21.4	75.8 ± 12.4	5.0	0.000999	0.00924	0.0054
DeperessMouthCorners	38.0 ± 21.5	90.2 ± 26.7	9.0	0.000999	0.00924	0.0054
PuckerLips	94.0 ± 16.3	98.8 ± 1.3	62.0	0.948	0.9480	0.0500
SuckLips	76.1 ± 26.0	78.0 ± 14.6	66.0	0.743	0.7744	0.0480
OpenMouth	58.3 ± 26.6	70.0 ± 8.6	50.0	0.511	0.6302	0.0405
CloseMouth	75.0 ± 9.4	91.3 ± 7.7	11.0	0.001	0.00435	0.0115
StickTongueOut	67.1 ± 27.0	87.4 ± 9.1	27.0	0.030	0.0707	0.0270
StickTongueOutLeft	53.9 ± 24.9	90.7 ± 8.7	9.0	0.000999	0.00924	0.0054
StickTongueOutRight	67.1 ± 14.5	85.4 ± 9.1	15.0	0.003	0.00925	0.0162
BiteTongue	69.0 ± 9.1	88.9 ± 6.6	2.0	0.000999	0.00924	0.0054
GrindTeeth	66.8 ± 11.2	87.5 ± 6.9	3.0	0.000999	0.00924	0.0054
MoveJawForward	57.2 ± 18.9	63.1 ± 15.2	46.0	0.358	0.4906	0.0365

the accuracy of “Turn Head Right” was 53.28%, though **participant P3** rated it as completed (Self-score = 8).

(c2) and (c2’) show “Left Wink” from Pb3 and P6 (c2) was recognised (Accuracy = 81.76%) and rated highly (Self-score = 8). In (c2’), motion speed and direction were limited, leading to misrecognition of “Close Left Eye” (Accuracy = 45.04%); The accuracy of “Left Wink” is only 5.04% **participant P6** rated it as completed (Self-score = 6).

(c3) and (c3’) show “Eye Look Up” from Pb11 and P8. (c3) was recognised (Accuracy = 82.93%) and rated highly (Self-score = 8). (c3’) shows directional deviation due to motor and perceptual differences, unrecognised (Accuracy = 24.54%), though **participant P8** self-rated it as completed (Self-score = 6).

(d) Differences in active regions. (d1) and (d1’) show “Turn Head Up" from Pb9 and P8. (d1) was recognised (Accuracy = 99.69%) and self-rated (Self-score = 8). In (d1’), deviations in muscle activation due to proprioceptive differences led to recognition failure (Accuracy = 23.53%), though **participant P8** self-rated it as successful (Self-score = 8). (d2) and (d2’) show “Left Wink” from Pb3 and P9 . (d2) was recognised (Accuracy = 81.76%) and rated highly (Self-score = 8). (d2’) shows shifted activation due to individual variability in motor control, leading to misrecognition as “Blink Eyes” (Accuracy = 52.38%). The accuracy of “Left Wink” is 29.22%, though **participant P9** rated it as completed (Self-score = 8). (d3) and (d3’) depict “Tongue Out Left” from Pb3 and P10 (d3) was recognised (Accuracy = 99.93%) and rated highly (Self-score = 8). In (d3’), muscle

Table 10: Per-gesture comparison of self-evaluation scores between Group A and Group B, with Benjamini–Hochberg FDR-corrected p-values.

Gesture	Group A (M ± SD)	Group B (M ± SD)	p-value	p_{FDR}
Turn Head Up	7.36 ± 1.21	7.55 ± 0.69	0.33	0.51
Turn Head Down	7.36 ± 1.03	7.64 ± 0.67	0.43	0.51
Turn Head Left	7.27 ± 1.42	7.64 ± 0.67	0.23	0.65
Turn Head Right	7.18 ± 1.47	7.55 ± 0.69	0.39	0.52
Tilt Head Left	7.09 ± 1.51	7.45 ± 0.69	0.36	0.56
Tilt Head Right	7.00 ± 1.67	7.45 ± 0.69	0.30	0.56
Move Head Forward	6.91 ± 1.58	7.27 ± 0.90	0.42	0.52
Move Head Backward	5.45 ± 2.19	7.36 ± 0.81	0.01	0.25
Raise Eyebrows	7.55 ± 1.21	7.82 ± 0.41	0.41	0.52
Frown Eyebrows	6.91 ± 1.58	7.55 ± 0.69	0.16	0.54
Blink Eyes	7.27 ± 1.42	7.73 ± 0.47	0.23	0.65
Left Wink	6.00 ± 2.45	7.27 ± 0.90	0.11	0.51
Right Wink	6.36 ± 2.15	7.45 ± 0.69	0.09	0.55
Close Eyes	7.45 ± 0.93	7.64 ± 0.67	0.53	0.58
Close Left Eye	7.27 ± 1.42	7.55 ± 0.69	0.46	0.53
Close Right Eye	7.18 ± 1.47	7.55 ± 0.69	0.39	0.52
Open Left Eye	6.91 ± 1.58	7.55 ± 0.69	0.30	0.56
Open Right Eye	5.18 ± 2.71	7.45 ± 0.69	0.01	0.25
Squint Eyes	7.00 ± 1.67	7.27 ± 0.90	0.56	0.58
Wide Open Eyes	7.45 ± 0.93	7.82 ± 0.41	0.36	0.56
Eye Look Up	6.00 ± 2.18	7.45 ± 0.69	0.04	0.49
Eye Look Down	7.27 ± 1.42	7.64 ± 0.67	0.33	0.51
Eye Look Left	6.73 ± 1.79	7.64 ± 0.67	0.09	0.56
Eye Look Right	6.82 ± 1.72	7.64 ± 0.67	0.13	0.51
Puff Out Cheeks	7.36 ± 1.21	7.82 ± 0.41	0.36	0.56
Raise Mouth Corners	7.45 ± 1.21	7.64 ± 0.67	0.58	0.58
Depress Mouth Corners	6.64 ± 1.50	7.36 ± 0.81	0.24	0.59
Pucker Lips	7.27 ± 1.42	7.82 ± 0.41	0.23	0.65
Suck Lips	7.09 ± 1.45	7.82 ± 0.41	0.09	0.56
Open Mouth	7.64 ± 0.67	7.82 ± 0.41	0.53	0.58
Close Mouth	7.64 ± 0.67	7.82 ± 0.41	0.53	0.58
Stick Tongue Out	7.00 ± 1.41	7.45 ± 0.69	0.33	0.51
Stick Tongue Out Left	6.91 ± 1.58	7.45 ± 0.69	0.30	0.56
Stick Tongue Out Right	6.36 ± 2.15	7.45 ± 0.69	0.13	0.51
Bite Tongue	7.00 ± 1.67	7.45 ± 0.69	0.36	0.56
Grind Teeth	7.00 ± 1.67	7.45 ± 0.69	0.36	0.56
Move Jaw Forward	6.00 ± 2.26	7.45 ± 0.69	0.06	0.56

control variance caused a spatial shift in activation. The accuracy of "Tongue Out Left" is 16.85%. During the middle of the motion, there were misrecognitions as "PuckerLips" (Accuracy = 69.93%) and "Raise Right Mouth Corner" (Accuracy = 50.71%); **participant P10** self-rated the gesture as completed (Self-score = 6).

H Qualitative Factors Underlying Intention–Recognition Mismatches

We conducted a qualitative analysis of participant feedback from Group A, triangulated with recognition outcomes and gesture trajectory data. Our goal was to understand how mismatches arise between user intention and system recognition. We observed four recurring categories of mismatch factors: impaired sensorimotor awareness, spatial disorientation, motor coordination difficulties, and diverse interpretations of gesture semantics.

H.1 Impaired Sensorimotor Awareness

Many participants struggled to perceive their own movements, especially those involving fine motor control in the face and eyes. Due to impaired proprioception and sensory feedback, they often required external confirmation. For instance, P9 asked, Are my eyeballs moving?" and P10 questioned, Am I moving both sides? I don't know either." P1, during a left wink, asked, "Did I also close my right eye?" These questions reflect not only mechanical challenges but also perceptual ambiguity. Gesture execution was often slowed down deliberately, as noted by P10: "I have to do it very deliberately."

Table 11: Per-gesture perception–recognition alignment metrics for both groups. MR = Match Rate, OR = Overestimation Rate, UR = Underestimation Rate. A = Group A (motor-impaired), B = Group B (non-impaired). Metrics reflect the proportion of gesture trials in which self-perception and model prediction align or diverge.

Gesture	MR (%) (A)	OR (%) (A)	UR (%) (A)	MR (%) (B)	OR (%) (B)	UR (%) (B)
TurnHeadUp	72.7	27.3	0.0	100.0	0.0	0.0
TurnHeadDown	90.9	9.1	0.0	100.0	0.0	0.0
TurnHeadLeft	72.7	18.2	9.1	100.0	0.0	0.0
TurnHeadRight	100.0	0.0	0.0	100.0	0.0	0.0
TiltHeadLeft	81.8	18.2	0.0	100.0	0.0	0.0
TiltHeadRight	90.9	9.1	0.0	100.0	0.0	0.0
MoveHeadForward	72.7	27.3	0.0	81.8	9.1	9.1
MoveHeadBackward	72.7	27.3	0.0	81.8	9.1	9.1
RaiseEyebrows	81.8	18.2	0.0	100.0	0.0	0.0
FrownEyebrows	72.7	27.3	0.0	100.0	0.0	0.0
BlinkEyes	100.0	0.0	0.0	100.0	0.0	0.0
LeftWink	81.8	18.2	0.0	90.9	0.0	9.1
RightWink	72.7	27.3	0.0	100.0	0.0	0.0
CloseEyes	81.8	18.2	0.0	100.0	0.0	0.0
CloseLeftEye	54.5	45.5	0.0	100.0	0.0	0.0
CloseRightEye	81.8	18.2	0.0	81.8	18.2	0.0
OpenLeftEye	72.7	27.3	0.0	81.8	9.1	9.1
OpenRightEye	54.5	45.5	0.0	81.8	9.1	9.1
SquintEyes	72.7	27.3	0.0	72.7	27.3	0.0
WideOpeneyes	63.6	36.4	0.0	100.0	0.0	0.0
EyeLookUp	63.6	36.4	0.0	90.9	9.1	0.0
EyeLookDown	63.6	36.4	0.0	100.0	0.0	0.0
EyeLookLeft	81.8	18.2	0.0	100.0	0.0	0.0
EyeLookRight	81.8	18.2	0.0	100.0	0.0	0.0
PuffOutCheeks	90.9	9.1	0.0	100.0	0.0	0.0
RaiseMouthCorners	36.4	63.6	0.0	100.0	0.0	0.0
DepressMouthCorners	36.4	54.5	9.1	81.8	9.1	9.1
PuckerLips	90.9	9.1	0.0	100.0	0.0	0.0
SuckLips	90.9	9.1	0.0	100.0	0.0	0.0
OpenMouth	90.9	9.1	0.0	100.0	0.0	0.0
CloseMouth	100.0	0.0	0.0	100.0	0.0	0.0
StickTongueOut	81.8	18.2	0.0	100.0	0.0	0.0
StickTongueOutLeft	81.8	18.2	0.0	100.0	0.0	0.0
StickTongueOutRight	90.9	9.1	0.0	100.0	0.0	0.0
BiteTongue	100.0	0.0	0.0	100.0	0.0	0.0
GrindTeeth	90.9	9.1	0.0	100.0	0.0	0.0
MoveJawForward	72.7	27.3	0.0	100.0	0.0	0.0

P4, who reported facial numbness on one side, was unaware that her eyebrow elevation was asymmetrical in gesture (a2'). In some cases, participants unintentionally triggered multiple gestures (e.g., eye look up resulting in lateral movements; see Figure 7(b3')).

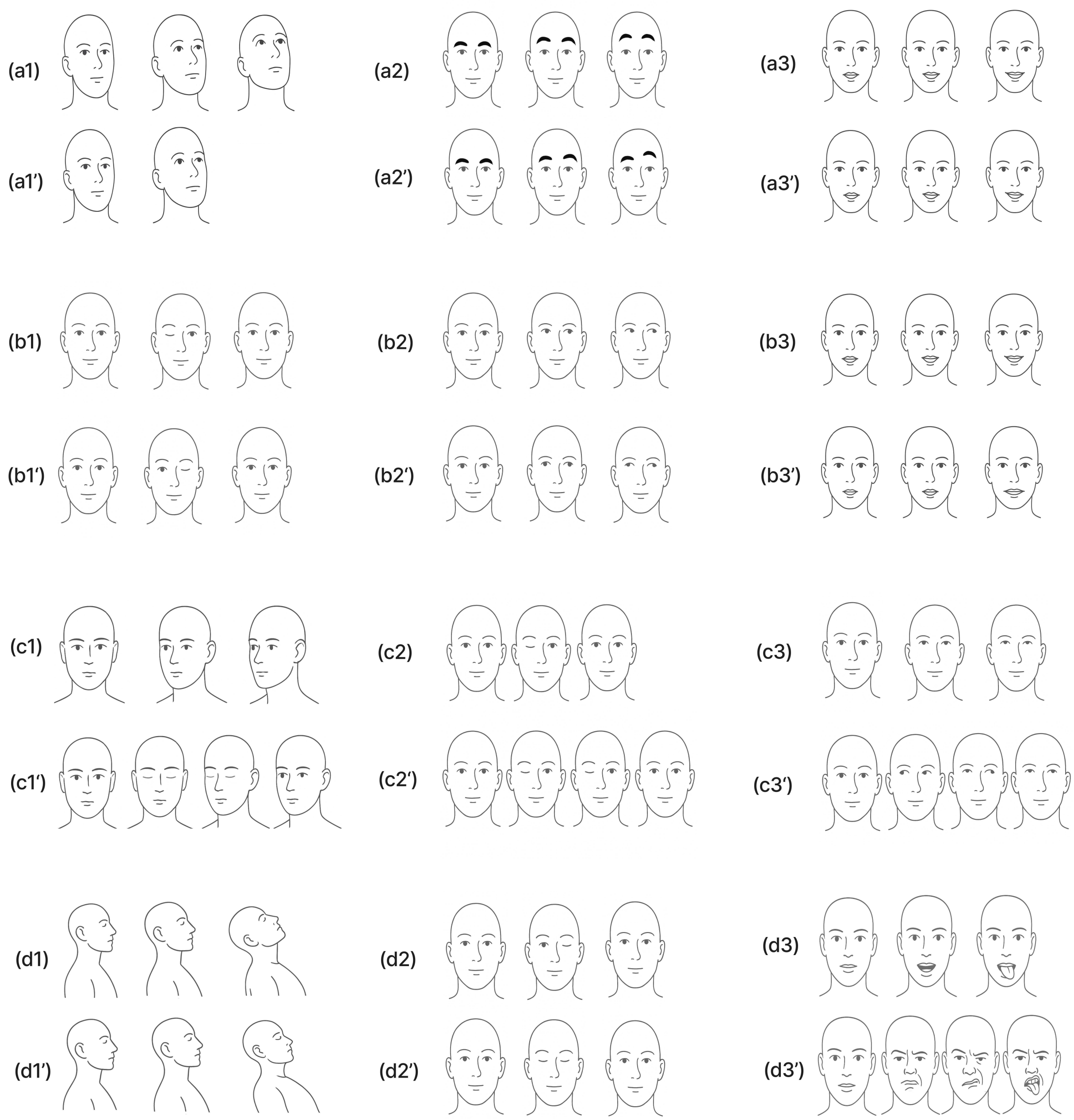
H.2 Spatial Disorientation and Directional Confusion

Several participants exhibited confusion in distinguishing left from right, a recurring issue for P1, P2, P6, P8, P9, and P10. P10 noted, I can't really tell left from right," and P6 said after a long pause, I haven't figured out left and right yet." P8, after a blink, asked, Was that my left eye?" These difficulties also extended to recalling recent movements. For instance, P8 asked immediately after a head turn, Which direction did I just go?" Such spatial disorientation affected both gesture comprehension and execution. P2 and P10 noted difficulty coordinating left-right gestures, particularly tongue and eye gestures. These problems often resulted in unintended co-movements or misdirection.

H.3 Motor Coordination Difficulties and Compensatory Movements

Motor-impaired participants often exhibited involuntary compensatory strategies without being aware of them. For instance, during the "Turn Head Up" gesture, P2, P4, and P11 arched their backs

Figure 7: Gesture Trajectory Variations Across Dimensions between Group A and Group B across four aspects: (a) amplitude, (b) direction, (c) speed, and (d) active regions



or leaned their upper body instead of moving the head alone. P4 used shoulder movement to compensate for head retraction but was unaware of this substitution. In contrast, Pb1 (non-impaired) made similar compensatory movements but retained proprioceptive awareness. These findings were supported by both self-evaluation and recognition data. P1 noted asymmetry in executing right-sided gestures. P2's performance on right-sided gestures was consistently lower, aligning with both self-reports and system outputs.

Table 12: Representative mismatched gestures from Group A (motor-impaired participants). Each case shows a gesture whose motion led to recognition failure or misclassification due to sensorimotor variation. Entries are cross-referenced with Figure 7.

Type	Gesture	Participant	Accuracy	Self-Eval	Mismatch Description
Amplitude	Turn Head Up	P8	23.53%	8	Limited neck movement; under-threshold tilt (Fig. 7(a1'))
Amplitude	Raise Eyebrows	P10	33.86%	8	Right brow weak; misrecognised as "Right Eyebrow" (Fig. 7(a2'))
Amplitude	Raise Mouth Corners	P4	37.38%	8	Low amplitude; failed recognition (Fig. 7(a3'))
Direction	Left Wink	P11	0.09%	8	Left-right confusion; failed recognition (Fig. 7(b1'))
Direction	Eye Look Left	P10	21.50%	8	Directional drift due to motor variability (Fig. 7(b2'))
Direction	Raise Mouth Corners	P1	0.21%	8	Direction misread as "Mouth Stretch" (Fig. 7(b3'))
Speed	Turn Head Right	P3	53.28%	8	Delayed motion; misread as "Blink Eyes" (Fig. 7(c1'))
Speed	Left Wink	P6	5.04%	6	Slow onset; misrecognised as "Close Left Eye" (Fig. 7(c2'))
Speed	Eye Look Up	P8	24.54%	6	Slow and unstable upward gaze (Fig. 7(c3'))
Region Shift	Turn Head Up	P8	23.53%	8	Compensated with back arch; altered region (Fig. 7(d1'))
Region Shift	Left Wink	P9	29.22%	8	Co-activation of upper/lower eye muscles (Fig. 7(d2'))
Region Shift	Move Jaw Forward	P10	16.85%	6	Tongue motion misread as multiple gestures (Fig. 7(d3'))

H.4 Diverse Interpretations of Gesture Semantics

Differences in movement ability and expressive history led to non-standard interpretations of gesture prompts. For instance, gestures like "Raise Mouth Corners," often defined in datasets as "smile," were interpreted differently. P1 stated, "I don't have the strength to lift my mouth corners into an arc... but this stretch should count as a gesture. This is what my smile looks like." These expressions, while intentional and effortful, were often unrecognised due to deviations from normative training data. This highlights a mismatch not of intent, but of semantic and expressive representation.

Summary: These findings illustrate that intention-recognition mismatches are not random failures but reflect deeper disconnects between embodied experience and algorithmic assumptions. Together, they characterise the qualitative conditions under which gesture recognition systems break down for motor-impaired users. See Figure 7 and Table 12 for illustrative trajectories and examples.

I Algorithmic Implications for Inclusive Gesture Recognition

The qualitative observations in Appendix H suggest that current gesture recognition models operate under normative assumptions that are frequently violated by PwM. These mismatches reflect deeper algorithmic blind spots in both sensor interpretation and decision logic. We summarize below the key failure modes that characterise how current gesture recognition systems break down for users with motor impairments:

I.1 Limited feedback modelling.

Many gestures in our study (such as subtle eye movements or brow lifts) occur outside the user's visual field or proprioceptive awareness. As reported by P9 and P10, users often could not confirm whether a gesture had been completed without external validation. However, most recognition models assume that users receive and act on implicit motor feedback. This highlights a mismatch between users' limited feedback awareness and models that rely on implicit motor feedback.

I.2 Overfitting to idealized motion patterns.

Gestures performed by participants with limited mobility often exhibited reduced amplitude or asymmetric motion. These were frequently rejected by the model despite users rating them as successful (e.g., P1's "Raise Mouth Corners"). This points to overfitting to normative data distributions. This reveals the absence of mechanisms for accounting for reduced amplitude or asymmetric effort in current models.

I.3 Directional ambiguity and mirror confusion.

Several users (e.g., P6, P8) struggled to differentiate left from right or exhibited mirrored co-activation. Standard models treat gestures like "Left Wink" and "Right Wink" as mutually exclusive, assuming crisp lateral differentiation. This exposes a lack of explicit representation for lateral uncertainty in standard gesture recognition models.

I.4 Lack of graceful degradation.

Current models often return binary classifications, with no representation of near-miss gestures. This penalizes users whose expressions fall just outside trained distributions. This reflects the absence of intermediate or near-miss representations in current classification pipelines.

I.5 Normative semantic assumptions.

Users interpreted gesture prompts through the lens of personal experience and embodied constraints (e.g., "This is what my smile looks like"). Models that rely solely on externally imposed labels (e.g., "smile" = AU12) risk invalidating meaningful expressions. This exposes a mismatch between user-aligned semantic intent and label-driven gesture definitions.

Together, these observations characterise a systematic misalignment between embodied motor intent and the assumptions embedded in current gesture recognition systems.