GazeSwitch: Automatic Eye-Head Mode Switching for Optimised Hands-Free Pointing

BAOSHENG JAMES HOU, Lancaster University, United Kingdom
JOSHUA NEWN, Lancaster University, United Kingdom
LUDWIG SIDENMARK, University of Toronto, Canada
ANAM AHMAD KHAN, Lancaster University, United Kingdom
HANS GELLERSEN, Lancaster University, United Kingdom and Aarhus University, Denmark

This paper contributes GazeSwitch, an ML-based technique that optimises the real-time switching between eye and head modes for fast and precise hands-free pointing. GazeSwitch reduces false positives from natural head movements and efficiently detects head gestures for input, resulting in an effective hands-free and adaptive technique for interaction. We conducted two user studies to evaluate its performance and user experience. Comparative analyses with baseline switching techniques, Eye+Head Pinpointing (manual) and BimodalGaze (threshold-based) revealed several trade-offs. We found that GazeSwitch provides a natural and effortless experience but trades off control and stability compared to manual mode switching, and requires less head movement compared to BimodalGaze. This work demonstrates the effectiveness of machine learning approach to learn and adapt to patterns in head movement, allowing us to better leverage the synergistic relation between eye and head input modalities for interaction in mixed and extended reality.

CCS Concepts: • **Human-centered computing** → **Mixed / augmented reality**; **Virtual reality**; **Pointing**; **Gestural input**.

Additional Key Words and Phrases: Gaze interaction, Refinement, Eye Tracking, Eye-head Coordination, Computational Interaction, Machine Learning

ACM Reference Format:
Baosheng James Hou, Joshua Newn, Ludwig Sidenmark, Anam Ahmad Khan, and Hans Gellersen. 2024. GazeSwitch: Automatic Eye-Head Mode Switching for Optimised Hands-Free Pointing. *Proc. ACM Hum.-Comput. Interact.* 8, ETRA, Article 227 (May 2024), 20 pages. https://doi.org/10.1145/3655601

1 INTRODUCTION

The synergistic relationship between eye and head input modalities offers a promising approach for achieving hands-free pointing [22, 35, 37, 48]. The proposed BimodalGaze technique [37], for instance, allows for greater pointer control by automatically switching between 'Gaze Mode' for coarse positioning and 'Head Mode' for refinement. This seamless switch leverages eye-head coordination insights that allow the separation of natural from gestural head movement [34]. Natural head movement occurs when the head moves to support our visual system during a gaze shift so that we can see objects that are not right in front of us while keeping the eyes within a

Authors' addresses: Baosheng James Hou, Lancaster University, Lancaster, United Kingdom, b.hou2@lancaster.ac.uk; Joshua Newn, Lancaster University, Lancaster, United Kingdom, j.newn@lancaster.ac.uk; Ludwig Sidenmark, University of Toronto, Toronto, Canada, lsidenmark@dgp.toronto.edu; Anam Ahmad Khan, Lancaster University, Lancaster, United Kingdom, a.a.khan7@lancaster.ac.uk; Hans Gellersen, Lancaster University, Lancaster, United Kingdom and Aarhus University, Aarhus, Denmark, h.gellersen@lancaster.ac.uk.

ACM 2573-0142/2024/5-ART227
https://doi.org/10.1145/3655601

comfortable eye-in-head position. Gestural head movement occurs when the head is used in its own right, and is independent of gaze.

While prior research has demonstrated the potential utility of this insight into eye-head coordination for distinguishing both movements, achieving optimal mode switching performance based on these movements remains a challenge due to the similarities in relative movement patterns of the eyes and head that occur during natural and gestural head movements. When we shift our gaze to a target in ‘gaze mode’, the head will follow, at a slower pace, to maintain a comfortable eye-in-head position, although our eyes have already reached the target. Simultaneously, our eyes perform compensatory movements opposing to the head movement to stabilise vision (vestibulo-ocular reflex [5]). In ‘head mode’, the eyes remain fixated on the target to maintain visual acuity, but as the head moves, the eyes also move in the opposite direction to the head. Leveraging this insight, the BimodalGaze technique explored a threshold-based approach to classify these movements for automatic mode switching between gaze and head modes but found limitations in using fixed threshold values. When the threshold for transitioning from gaze mode to head mode is set too high, it will result in difficulty entering head mode. Conversely, if the threshold is set too low, it triggers head mode prematurely, resulting in more head movement required for pointing. Similarly, when determining the switch from head to gaze mode, a too-high threshold can result in difficulty entering gaze mode while in head mode, whereas a too-low threshold causes it to exit head mode too easily, leading to instability in the mode switching.

In this paper, we contribute GazeSwitch, an automatic ML-based approach for real-time switching between eye and head mode for optimised hands-free cursor control in mixed and extended reality. GazeSwitch leverages insights from HeadBoost [17], our previous work, which introduced a method for separating gaze-driven and gestural head movements using a machine learning approach. In evaluation, HeadBoost proved effective compared to BimodalGaze’s threshold-based approach [37]. It achieved better overall classification accuracy and detected the onset of head mode earlier, suggesting that an ML-based approach that learns the patterns of eye-head movements can not only optimally identify the onset of each mode but also prevent unintended mode switches, thereby contributing to a better user experience.

We evaluated GazeSwitch using an HTC Vive Pro Eye VR HMD with a 120 Hz integrated Tobii eye tracker through two user studies. The first study compared mode switching and target selection performance against Eye+Head Pinpointing [22], a manual technique where the user presses and releases a controller button to mode switch, and BimodalGaze [37], an automatic technique where users perform the head refinement movement when they intend, and the system automatically detects the mode-switch using a threshold-based algorithm. The second study compared the user experience of GazeSwitch against Eye+Head Pinpointing for hands-free cursor control on two tasks (i.e., tracing and colouring). Our findings indicate that GazeSwitch is effective for both discrete target selection and continuous interactions, affirming the overall validity of our proposed technique. Further, GazeSwitch allows for a smooth transition between modes, eliminating the need for users to perform manual clutching as required with Eye+Head Pinpointing, or execute exaggerated head movements as in the case of BimodalGaze.

In sum, we contribute: (1) An eye-head pointing technique where cursor control switches automatically between gaze mode for coarse positioning and head mode for precise positioning. (2) A mode switching approach that is based on ML classification of head movement to ensure that head mode is only activated when any head movements in support of a gaze shift have been completed as they would otherwise cause unintended input. (3) Evaluation of the performance and user experience of GazeSwitch against manual Eye-Head Pinpointing, showing our technique to be as performant while automating the mode switch, which reduces effort but trades of control.

2 RELATED WORK

Gaze has been widely explored as a hands-free alternative to manual input, as it functions as a fast and natural pointer for selection—people naturally look at objects before selecting them. However, using the eyes for input has limitations [28]. First, even during fixation, the eye is never completely still, which makes precise eye-based pointing challenging, especially for selecting small targets [52]. Second, although eye tracking has come a long way, its accuracy and precision are influenced by various factors, including calibration, lighting conditions, and the potential for drift over time. To address these inherent limitations, researchers have proposed a multitude of techniques, such as algorithms to smoothen eye tracking data [*e.g.* 12, 47], zooming techniques for accurate target selection [*e.g.* 1, 13, 42], and selection and disambiguation techniques that do not rely on calibration [*e.g.* 26, 32, 33, 45].

A promising approach involves harnessing the rapid pointing and hands-free capabilities of gaze for initial coarse positioning, and employing a complementary modality that affords more precise control for further positioning. A fundamental work that demonstrates this combination is MAGIC pointing [11, 51], where the cursor is "warped" to the gaze location and adjusted with manual mouse input, resulting in a substantial enhancement in pointing speed. In AR and VR, this principle has also been applied to controller movements [19]. Gaze-Shifting by Pfeuffer et al. [30] demonstrates the same principle with direct touch and pen input, where either input can be directly or indirectly mapped to the gaze area. The integration of gaze input with other modalities not only reduces physical movement and user fatigue but also enhances efficiency, fine control, and precision while capitalising on the natural speed and convenience of gaze pointing [2].

Besides hand-based input, head input has shown to be a promising input for disambiguation and refinement pointing for target selection, as the head affords hands-free fine control. In our previous work, we found that users have fine-grained over their head movement (~0.3 degrees) [17]. Eye-head combination capitalises on the strengths of both modalities, with the eyes providing fast and precise input while head movements enable finer adjustments. Moreover, eye-head techniques for pointing have been found to achieve faster speeds than head-only techniques [19, 21, 40].

Early works on desktop-based interaction combined head movement with gaze to refine gaze movements with leaning [48] or rotating head movements [27]. However, a key assumption for these works is that head movement is only used for interaction, not for controlling the viewport, as in VR. As the head position can easily be tracked in 3D interfaces, several techniques have been proposed that leverage head input—with many leveraging eye-head coordination insights for selection and manipulation. For example, using the head with estimation of gaze depth for target disambiguation [24], or for menu control [39].

In a study that compared variations of eyes for selection and other inputs for refinement, head correction of gaze is preferable even if manual input is available, as it requires less physical effort [22]. This eye+head variation, 'Eye+Head Pinpointing', is where the cursor is initially controlled with gaze and switches over to refinement mode when the user holds down a controller button to invoke head input. Head movements are then used to make precise adjustments to the cursor position, effectively "pinpointing" the target. When the user releases the button, the target returns to gaze pointing mode. In head-refinement mode, the CD-gain is adjusted to 0.5, allowing the technique to select small targets, as small as 0.5 degrees. While a manual switching technique affords users control over when to enter refinement mode, this switching process can be seamless, as shown with BimodalGaze [37].

The BimodalGaze technique seamlessly integrates eye and head movements, enabling automatic mode switching based on a threshold-based algorithm. The algorithm classifies and seamlessly transitions between gaze mode (gaze-driven head movements) and head mode (gestural head movements) using a set of thresholds. BimodalGaze enters 'head mode' when a head movement is detected (head velocity >15°/s) that started at least 150 ms after the previous gaze shift, and the angular difference between the trajectory of the eyes and head at least 20 degrees while 'gaze mode' when either a gaze shift is detected (gaze velocity >160°/s) or when the distance between gaze and cursor is more than 10 degrees. Hence, by classifying when the head supports gaze (natural) and when the head is used for interaction (gestural), the technique allows the seamless transition where the eyes are used for fast coarse pointing and head movements for refinement.

However, despite participants in their user study describing BimodalGaze's ability to automatic mode switch as smooth and effortless compared to Eye+Head Pinpointing, it displayed a greater frequency of initial selection errors. This impacted both the total selection time and the overall performance despite its shorter refinement time. BimodalGaze employed a high head velocity threshold as a heuristic to minimise consistent mode switching. This threshold, however, introduced challenges when only small movements were required, often leading to overshooting as users resorted to exaggerated head motions to trigger the algorithm to enter head mode. These issues primarily stemmed from the limitations inherent in a threshold-based approach, impacting mode switching performance.

In our work, we build on the insights from BimodalGaze for automatic eye-head mode switching, and the potential of head movement classification from our previous work, HeadBoost [17]. The HeadBoost classifier addresses the challenge of correctly classifying between two fundamental types of head movements: gaze-driven head movement (Head-Gaze) and gestural head movement (Head Gesture). The classifier, built using XGBoost [3], takes as input position and direction 3D vectors of both eye and head movements. It incorporates a comprehensive set of over 600 eye and head-related features sourced from eye and head movement classification literature, along with feature vectors from prior timestamps to capture and analyse user behaviour. These features encompass shape, noise, spectral, temporal, and correlation characteristics of the eye and head vectors, and in combination, facilitate the classification of head movements. This novel approach yielded exceptional results, boasting an offline classification accuracy with an F_1-Score of 0.89 for effectively discriminating between the two types of head movement.

In comparison with BimodalGaze, the HeadBoost classifier demonstrated better classification performance (F_1-Score: 0.89 vs 0.62), indicating a substantial improvement in overcoming the limitations of a threshold-based approach. Moreover, HeadBoost results showed that it predicted the onset of Head Gesture much earlier than BimodalGaze (119 ms earlier on average for all trials), an area that required improvement. In further analysis, the Headboost classifier accurately classified small head movements (<15°/s), compared to BimodalGaze. This performance enhancement can be attributed to the capacity of using a machine learning approach to learn and adapt to patterns in head movement, effectively overcoming classification challenges—a viable approach in light of the natural eye-head coordination behavioural complexities discussed in Introduction.

3 GAZESWITCH

To develop GazeSwitch, we first obtained a labelled dataset of eye and head movement data of participants as they performed cursor refinement tasks in a controlled study (detailed in Section 3.1). We closely followed the pipeline steps used to develop HeadBoost [17], including preprocessing and the initial steps for feature engineering (Section 3.2). With recursive feature selection, we obtained a classification rate above 120 Hz with a high classification performance of 0.91 F_1-Score (Section 3.3). We then apply the ability to classify head movement types with a simple logic to robustly define the mode switch between gaze pointing and head refinement (Section 3.4).

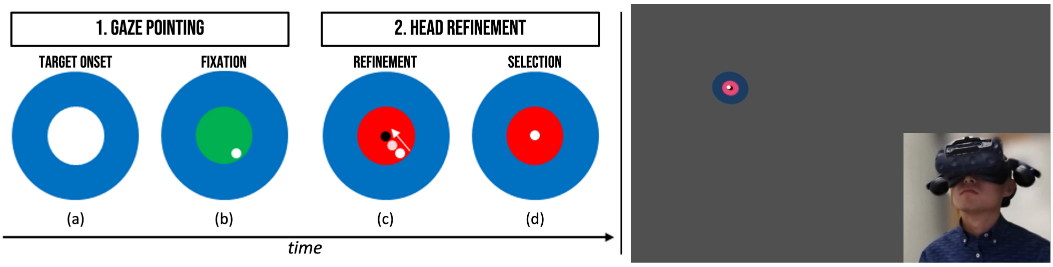


Fig. 1. Left: Task sequence for data collection. (a-b) Participants fixates on a target, receiving green feedback. 50% of the time, a new target appears, prompting a new gaze shift, the sequence repeats. (c-d) 50% of the time, the target centre turns red, and a black dot appears in the centre to prompt a refinement to place the cursor as close to the target as possible using head mode (thumbpad press) before selecting the target (thumbpad release). A new target appears, and the sequence repeats over. Right: Data collection setup.

3.1 Data Collection

We designed a target acquisition task and corresponding study procedure to collect eye and head movement data typical of gaze pointing and head refinement. We developed the apparatus using Unity 2020.3.32f1. Figure 1 illustrates the task sequence, uniquely designed to collect large variances of labelled eye and head movements for training. To collect gaze shifts of various directions and amplitudes, targets appeared at randomised positions of diverse patterns, some requiring only a gaze shift towards them and others demanding cursor refinement. Trials involving head mode selection occurred with a 50% probability, and the sequence of pointing modes was randomised. In cases requiring refinement, participants employed a technique akin to Eye+Head Pinpointing [22], toggling mode switching by pressing and releasing the thumbpad of a controller to place the cursor as close as possible to the target centre. The period with the controller button held down is labelled as 'Head Gesture', while the remaining samples are labelled as 'Head-Gaze'.

We followed the target design used in HeadBoost [17], featuring a diameter of 5.72 degrees, a transparent centre of 2.56 degrees (Figure 1a), and a black dot of 0.8 degrees in diameter (Figure 1c). The small size of the black dot was chosen to challenge gaze pointing, thereby encouraging participants to invoke head mode for refinement. The transparent centre provided feedback, transitioning from green when the user fixated on the target to red to indicate the need for closer placement to the centre. The target size was to ensure that the target was visible in the VR scene, facilitating participants in easily locating the subsequent target. For all trials, we collected the eye-in-world directional 3D vector, eye-in-head directional 3D vector, head position 3D vector, and head directional 3D vector.

We recruited 5 participants from our local university, aged 22-30 (M=26.8, SD=3.54, 1 female, 4 male). No prior VR or eye tracking experience was required, but participants needed to have normal or corrected-to-good vision. Upon arrival, participants were comfortably seated, briefed on the study procedure, and asked to sign a consent form before completing a demographic survey. They were then instructed to wear the HTC Vive Pro Eye VR HMD with integrated 120 Hz Tobii eye tracker, with assistance provided if needed, and underwent a five-point eye-tracking calibration. Following this, participants were asked to complete one sequence (30 trials) to familiarise themselves with the task before the data collection phase. Each participant completed 300 trials (10 sequences × 30 trials). Breaks were permitted between the sequences, and participants recalibrated each time they removed the HMD. Each session took approximately 40 minutes. The study procedure was approved by Lancaster University's research ethics committee.

3.2 Dataset Preprocessing and Feature Engineering

The data collection session resulted in 246898 timestamps from 1500 trials (300 trials per participant), with 65.3% of samples labelled as Head-Gaze, and 34.7% as Head Gesture. We preprocessed the raw data following best practices [4, 6, 8], filtering out samples with an inter-sample velocity exceeding $800°/s$. Following this, we applied cubic spline interpolation to standardise the sampling rate to 120 Hz (sampling frequency of the eye tracker). Lastly, we converted the 3D directional gaze and head vectors into 2D Fick angles using the Fick-gimbal method [15][1], mirroring the approach in our HeadBoost paper for consistency in feature generation. Furthermore, we adopted the hyperparameter choices used in HeadBoost, for both feature calculation and classifier training, determined through cross-validation.

We extracted shape-, noise-, spectral-, correlation-, and timing-based features (see Appendix B), computed over a window length of 512 ms. To address issues related to multicollinearity during classification, we refined the features using correlation distance and hierarchical clustering [25], resulting in a streamlined set of 80 representative features. We then included the features from the last 1024 ms for each labelled timestamp at every 6.25 Hz to capture the temporal context of users' behaviour. This resulted in a set of 600 features for each labelled time stamp. To overcome computational costs and the risk of overfitting [9], we applied Recursive Feature Addition (RFA). RFA involved incrementally adding features and assessing model performance on testing folds, retaining only features that improved performance. This process yielded a final set of 28 features for each labelled timestamp for classification. Fifteen of the final features are based on eye movement, 13 are based on head movement (Appendix B.2).

3.3 Model Classification and Evaluation

We use the final set of 28-dimensional features to train an XGBoost model (20 trees, max. depth 6) to classify between Head-Gaze and Head Gesture. As demonstrated in HeadBoost [17], XGBoost was superior in performance across the testing folds compared to other models. We then evaluated the classifier using leave-one-participant-out cross-validation, training the classifier five times, each time training the classifier on the data of four participants and evaluating it on the trials of the last participant. Model performance was evaluated using two metrics: the F_1-Score and the Area Under the Receiver Operating Characteristics Curve (AUC). F_1-Score combines precision and recall in a single metric, while AUC measures the classifier's ability to differentiate between classes. Both metrics range from 0 to 1, with 1 indicating perfect performance. The performance results of the user-independent model indicate that the built classifier can optimally classify head movements, achieving a high average F_1-Score of 0.91 (SD=0.01) and a high AUC score of 0.93 (SD=0.01), as well as high Precision and Recall scores, 0.92 (SD=0.01) and 0.90 (SD=0.02), respectively.

3.4 Mode Switching Logic

To switch into gaze mode, two conditions must be satisfied: (1) the trained ML classifier predicts gaze mode, and (2) the dispersion of the eye-in-head angles from the last 50 ms is greater than 3.6°. The second condition overwrites the ML prediction if the user is still fixating to maintain a steady head mode period. The 3.6° threshold chosen is twice the eye tracking precision of the HTC Vive Pro Eye during static head phases ($2 \times 1.8°$ mean intersample RMS) [41], and the 50 ms duration is a trade-off between window size and real-time classification responsiveness. The dispersion threshold serves to counter eye tracking imprecision and prevent unintended gaze mode activation due to minor jitters. Furthermore, requiring confirmation from both the ML model and the dispersion

[1] Functions for converting between gaze 3D vectors, Fick angles, and visual angles are authored by Per Bækgaard, available at https://github.com/baekgaard/fickpy

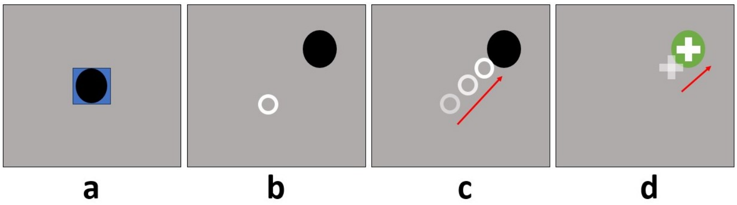


Fig. 2. Trial sequence. (a) The participant aligns their eyes and head to a centred neutral position following visual feedback by placing a black dot into a blue square. (b) Target onset, the cursor is visible as a white ring, indicating currently the user is currently in Gaze Mode. (c) In Gaze Mode, as the participant gaze shifts towards the target, the cursor follows eye gaze towards the target. (d) The participant may switch to Head Mode to refine the cursor position, if so, the cursor changes to a white cross to indicate Head Mode. The target turns green when acquired by the cursor. Selection is made with a button-up event of the thumbpad. The target is selectable in both Gaze and Head modes. The next trial begins after realigning eyes and head as in the first step. The cursor fading and the arrows are for illustrations only.

threshold mitigates accidental mode-switching resulting from single-frame false predictions of the ML model. This approach differs from BimodalGaze, which activates gaze mode by detecting larger gaze shifts with a velocity threshold. Through a pilot study, we found that thresholds worked well for the participants, giving confidence in our chosen parameters. The algorithm for GazeSwitch mode switching logic can be found in Appendix A.

4 STUDY 1: PERFORMANCE EVALUATION

We evaluated the performance of GazeSwitch against two existing eye-head mode switching techniques, Eye+Head Pinpointing (manual) and BimodalGaze (threshold-based). We used a 3 × 2 × 3 within-subject design with the three techniques, two target widths (0.8°, 1.5°) and three amplitudes (10°, 25°, 40°). We recruited 12 participants, aged 21 to 50, (M=29, SD=7.07, 6 female) through the university's mailing lists for this study. No prior VR or eye tracking experience was required, but participants needed to have normal or corrected-to-good vision. Eleven participants had either occasional or no VR experience, while one reported daily VR headset use. Six participants had no prior experience with eye tracking, whereas six reported occasional use. The study environment and tasks were developed in Unity version 2020.3.32f1. We collected the eye-in-world directional 3D vector, eye-in-head directional 3D vector, head position 3D vector, and head directional 3D vector using a HTC Vive Pro Eye VR HMD (90 Hz). The HMD has a field of view (FOV) of 100° in the horizontal plane, 110° in the vertical plane and a built-in eye tracker (120 Hz).

4.1 Task

We adopted a pinpointing task for this study, similar to the task used in BimodalGaze [37], which required participants to perform precise pointing for target selection using eye-head mode switching. Hence, targets can be selected in either eye or head mode, while confirmation is triggered using the controller. Given that the 0.8° target might be challenging to discern at larger amplitudes, we enhanced its visibility by introducing a white crosshair with a 3° transparent space at its centre and a thickness of 1° surrounding the target. Figure 2 illustrates the trial sequence.

At the onset of each trial, participants are guided visually to align their eyes and head, in which we enforce that the eyes and head position are within 5 and 2 degrees, respectively, from a centred neutral position, with the head velocity limited to less than $2°/s$. Once the alignment is completed, a black circular target appears, signalling the participant to look towards it. The cursor is visible throughout the trial and is initially attached to the filtered gaze point. We applied a 1€ filter with a minimum cutoff frequency of 1 Hz, slope beta value of 10, with the default 1 Hz cutoff frequency to smooth the cursor for visualisation, but the raw data streams were used as input to GazeSwitch.

Participants were required to place the cursor as close as possible to the target centre, with the option to switch to head mode to fine-tune the cursor position. In gaze mode, the cursor appears as a white ring (Figure 2b), while in head mode, it changes to a white cross (Figure 2d). Upon entering the target area, it turns green as hover feedback. The participant then completes the selection with a button-up event of the thumbpad of the controller. If the cursor is off-target at selection, or if no selection is made within 5 seconds, an error audio cue is played, and the trial is marked as failure. The target position will then be re-queued at the end of the block for a maximum of two additional attempts. If the target is selected within the 5-second window with the cursor inside the target area, the trial is marked as a success and will not be re-queued. The next trial begins after realigning the eyes and head back to the centre. The block concluded either upon selection of all targets or when the maximum attempts were exceeded (3 per target position).

4.2 Procedure

Upon arrival, participants were seated comfortably and provided with a briefing on the study. They were then given a consent form and a demographic questionnaire to be signed and filled out, respectively. They were then instructed to put on the HMD, with assistance provided if required, and to undergo the five-point eye tracking calibration. For each technique block, participants completed six sequences (2 Target Sizes × 3 Repetitions) of 24 trials (8 Directions × 3 Amplitudes) each, with the two target size levels randomly and evenly ordered. The techniques are counterbalanced with a Latin Square. Participants were then offered the opportunity to practice the current technique at the beginning of each block, involving one sequence of 24 trials with 0.8-degree targets.

At the end of each technique block, participants were asked to remove HMD, fill out a NASA TLX questionnaire [14] and provide verbal feedback about the technique they just used. Participants continued to the next block when ready.

In total, each participant performed 144 trials (3 Techniques × 2 Target sizes × 24 Trials). The study took 60 minutes to complete, after which we progressed to a second subsequent study that took a maximum of 30 minutes, which we report in Section 5. Participants were compensated with a £10 Amazon gift card for their time. The study procedures were approved by Lancaster University's research ethics committee.

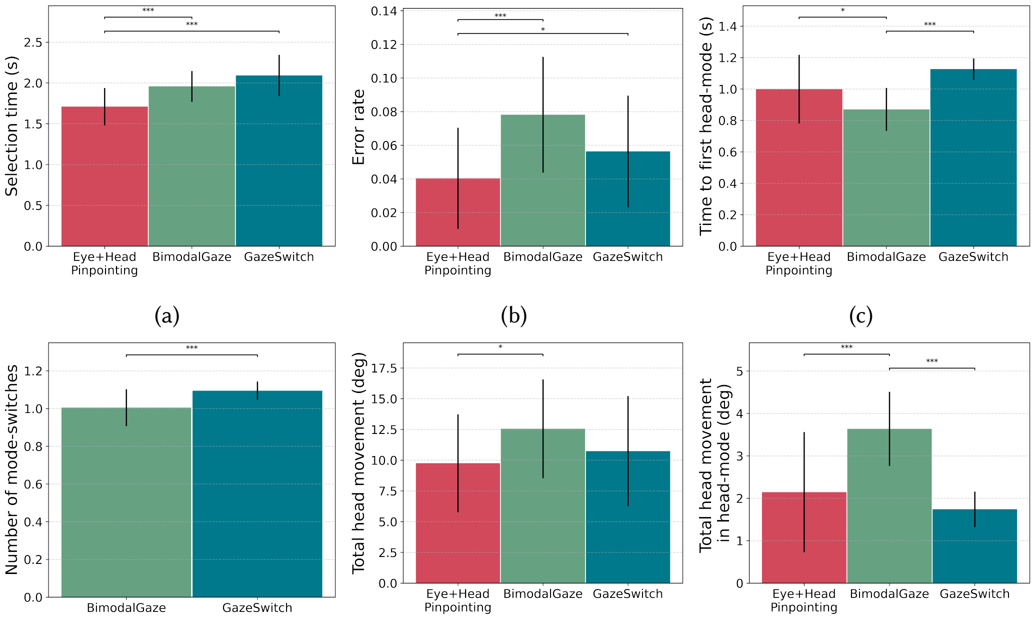
4.3 Results

We performed a three-way repeated-measures ANOVA with interaction technique, target size, and target amplitude as independent variables, using a significance level of $\alpha = 0.05$. In cases where the data was ordinal or conventional transformations did not address normality, we applied the Aligned Rank Transform (ART) technique [49] and confirmed that the aligned responses approximately summed to zero. When the assumption of sphericity was violated, as indicated by Mauchly's test, we employed Greenhouse-Geiser correction. Post hoc tests were carried out using pairwise t-tests with Bonferroni corrections or the ART procedure for multifactor contrast tests [10]. We analysed usability Likert-scale data using Friedman tests with Bonferroni-corrected Wilcoxon tests for post hoc analysis. Table 1 shows the mean and standard deviation for each performance evaluation metric.

4.3.1 Selection time. Selection time, measured from the onset of a trial to a successful selection, serves as an indicator of the overall technique speed. We found a significant main effect for Technique ($F_{1.98,21.74} = 11.27, p < 0.001$), Target Size ($F_{1,11} = 27.2, p < 0.001$), and Target Amplitude ($F_{1.37,15.12} = 56.1, p < 0.001$). Post hoc examination demonstrated that Eye+Head Pinpointing

Table 1. Performance metric for the techniques, with mean and standard deviation (in parenthesis).

	Eye+Head Pinpointing	**BimodalGaze**	**GazeSwitch**
Selection time (s)	1.74 (0.25)	2.00 (0.25)	2.10 (0.27)
Error rate	0.04 (0.03)	0.08 (0.03)	0.06 (0.03)
Time to first head-mode (s)	1.00 (0.22)	0.87 (0.14)	1.13 (0.07)
Number of mode switches	-	1.00 (1.00)	1.09 (0.05)
Total head movement (deg)	9.75 (3.98)	12.56 (4.02)	10.73 (4.48)
Total head movement in head-mode (deg)	2.15 (1.42)	3.64 (0.87)	1.74 (0.42)
Subjective performance rating (NASA-TLX)	2.00 (0.58)	3.00 (1.35)	2.50 (1.19)



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 3. The performance metrics of all mode switching techniques evaluated from target selection task. * p<0.05, ** p<0.01, *** p<0.001

exhibited a significantly shorter selection time compared to both BimodalGaze and GazeSwitch ($p < 0.001$) (see Figure 3a). Selection times were significantly longer at the 40° amplitude compared to all other amplitudes ($p < 0.001$) and at the smaller (0.8°) target size ($p < 0.001$). No significant interactions were observed.

4.3.2 Error rate. We define an error as missing the target due to trial timing out or having an inaccurate cursor position at the time of selection, measured as error rate (percentage of unsuccessful initial attempts). We found significant main effects for Technique ($F_{2,22} = 12.88, p < 0.001$) and Target Amplitude ($F_{2,22} = 6.88, p < 0.01$). Post hoc analysis showed that the error rate of Eye+Head Pinpointing is significantly lower than BimodalGaze ($p < 0.001$) and GazeSwitch ($p < 0.05$) (see Figure 3b). Further, the error rate of all techniques was significantly lower at 10° amplitude than at 40° amplitude ($p < 0.05$). No significant interactions were observed.

4.3.3 Time to first Head Mode. This metric measures the time from the onset of the trial to when the participant first entered the head mode, reflecting how quickly each technique facilitates an intended switch to head mode. We observed that some participants never entered head mode under certain conditions, causing the data to deviate from normal distribution even after standard transformations were applied. Thus, for this metric, we only considered Technique and Target Amplitude as factors for the ANOVA analysis. We found significant main effects for Technique ($F_{2,22} = 32.03, p < 0.001$) and Target Amplitude ($F_{2,23} = 28.14, p < 0.001$). Moreover, we observed a significant interaction effect ($F_{4,44} = 2.80, p < 0.05$) for time to first head mode. Post hoc analysis showed that BimodalGaze has a significantly shorter time to enter the first head mode than both Eye+Head Pinpointing ($p < 0.05$) and GazeSwitch ($p < 0.001$). Further, BimodalGaze showed a significantly earlier transition to head mode compared to GazeSwitch at 25° and 40° amplitudes ($p < 0.001$) (see Figure 3c). Lastly, we found that GazeSwitch transitioned to head mode significantly earlier at 10° compared to 25° and 40° ($p < 0.001$). Conversely, we observed that BimodalGaze and Eye+Head Pinpointing switched to head mode significantly earlier at 10° compared to only 40°.

4.3.4 Number of mode switches. This metric quantifies how many times head mode is entered, providing insights into the overall stability of the mode switching for each technique. A count of 0 or 1 signifies complete stability in the technique, under the assumption that participants do not intentionally execute more than one mode switch. Our analysis revealed significant main effects for Technique ($F_{1,11} = 19.70, p < 0.001$), Target Size ($F_{1,11} = 18.17, p < 0.01$), and Target Amplitude ($F_{1,11} = 15.8, p < 0.001$). Furthermore, we observed significant two-way interaction effects ($p < 0.05$). Post hoc examination revealed that for every identical amplitude and target size, BimodalGaze exhibited a significantly lower number of mode switches compared to GazeSwitch ($p < 0.01$) (see Figure 3d).

4.3.5 Total head movement. This metric is derived from the sum of the inter-sample Euclidean distance of head movement throughout the entire trial. It provides an assessment of the overall head movement performed by the participant and is useful for determining if the differences in head movement during head mode are meaningful when considering the demands of the entire task. Our analysis revealed significant main effects for Technique ($F_{2,22} = 4.91, p < 0.05$), Target Size ($F_{1,11} = 9.13, p < 0.05$), and Target Amplitude ($F_{2,22} = 183.03, p < 0.001$), along with interaction effects between Technique × Target Size ($F_{2,22} = 3.58, p < 0.05$) and Target Size × Target Amplitude ($F_{2,22} = 0.027, p < 0.05$). For the larger target (1.5°), BimodalGaze required significantly greater overall head movement compared to only Eye+Head Pinpointing ($p < 0.05$) (refer to Figure 3e). However, for smaller targets (0.8°), we observed that BimodalGaze exhibited significantly more head movement than both Eye+Head Pinpointing and GazeSwitch ($p < 0.05$).

We further calculated total head movement in head mode only, which measures the overall effort of the selection technique, as more head movement during refinement may suggest more action from the users. We found a significant main effect for Technique ($F_{2,22} = 12.93p < 0.001$), Target Amplitude ($F_{2,22} = 26.62, p < 0.001$), and the interaction between Target Size and Target Amplitude ($F_{2,22} = 26.62, p < 0.05$). Subsequent post hoc examination revealed that BimodalGaze necessitated significantly more head movement during refinement in comparison to the other techniques ($p < 0.001$) (see Figure 3f). However, we observed no significant difference in head movement between Eye+Head Pinpointing and GazeSwitch. As expected, the analysis indicated that targets with a 40° amplitude required significantly more head movement during refinement compared to all the smaller amplitudes ($p < 0.05$).

4.3.6 Subjective feedback. We observed a statistically significant difference in the NASA-TLX workload for performance, with participants rating BimodalGaze significantly lower than Eye+Head

Pinpointing ($p < 0.05$). No other significant differences were found. We further analysed participants' verbal feedback and found that all three techniques were generally well-received. However, it was evident that each technique had its own set of limitations.

For Eye+Head Pinpointing, participants reported feeling "more in control" (P2) due to the ability to manually mode switch, resulting in the selection task being perceived as "convenient" (P4) and "fast" (P1, P3, P4). However, some participants (e.g., P3, P12) found it challenging when timing button presses and remembering to release the button, leading to more head movement required if the button was pressed too early.

For BimodalGaze, several participants found target selection to require "effort" (P2, P5, P9, P10), mainly because they perceived it as "less like an automatic switch" (P9), requiring more head movement and being "inconsistent" (P12, P10, P11) for mode switching. P7 noted, "*It is more inconsistent, sometimes the cross appears when I didn't need it, other times it didn't appear when I wanted. I have to learn the head movement to turn on the cross.*". However, some participants acknowledged that when mode switching was accurate, BimodalGaze could make the task feel "easier" (P4, P11, P12) and "smooth" (P3, P5).

For our proposed GazeSwitch technique, participants recognised that automatic mode switching facilitated "fast" (P1, P6) and "easy" (P3, P6) selection. Participants further commented that GazeSwitch was "responsive" (P8) and "precise" (P9), making the overall experience seamless. P1 noted, "*I felt it's the best... It's like the computer is helping you rather than complicating things. It's the most assisted, least rushed, and is consistent.*". However, some participants also noted that the accuracy of mode switching heavily relied on eye tracking quality, as GazeSwitch could become more "unstable" (P3, P12, P11) and "shaky towards the corners" (P3, P4) or during "accidental [unintentional] head movements" (P8, P13).

5 STUDY 2: USER EXPERIENCE EVALUATION

We developed two applications in Unity 2020.3.32f1 to compare the user experience of automatic and manual mode switching: (1) tracing the outline of an object with precise marker placement and (2) colouring objects in the scene. In the tracing task, participants have the flexibility to switch between gaze and head mode, affording them to utilise both long sweeping lines and short successive selections. The colouring task was designed to highlight the affordance of gaze mode to quickly move across the sides of the screen, while using head mode to select small targets precisely.

In this second study, we exclusively compared GazeSwitch and Eye+Head Pinpointing techniques, as BimodalGaze operates on the same automatic mode switching principle but received lower perceived performance in our prior evaluation (see Section 4.3.6). This study followed immediately after the performance evaluation study (Section 4); hence, the same participants and procedures when taking breaks were used. At the start of this study, we briefed participants on both tasks and the operation of both techniques. We then asked the participants to wear the HMD and perform a five-point eye-tracking calibration. Participants first performed the tracing application using both techniques, but the order of the techniques was counterbalanced. After each task, participants were invited to comment on their overall experience of using each technique. Participants were also free to report their preferred technique for performing the task.

5.1 Tasks

5.1.1 Tracing. This task is inspired by Gaze-Shifting [30] and demonstrates that users can precisely control the cursor to follow the contours of a car using mode switching. Hence, participants are tasked to trace the outline of the car by placing markers on it. Figure 4-Left shows the application scene, where a car is positioned at the centre, spanning 80° horizontally and 51° vertically. This is achieved by utilising the gaze mode to cover longer distances and the head mode to trace around

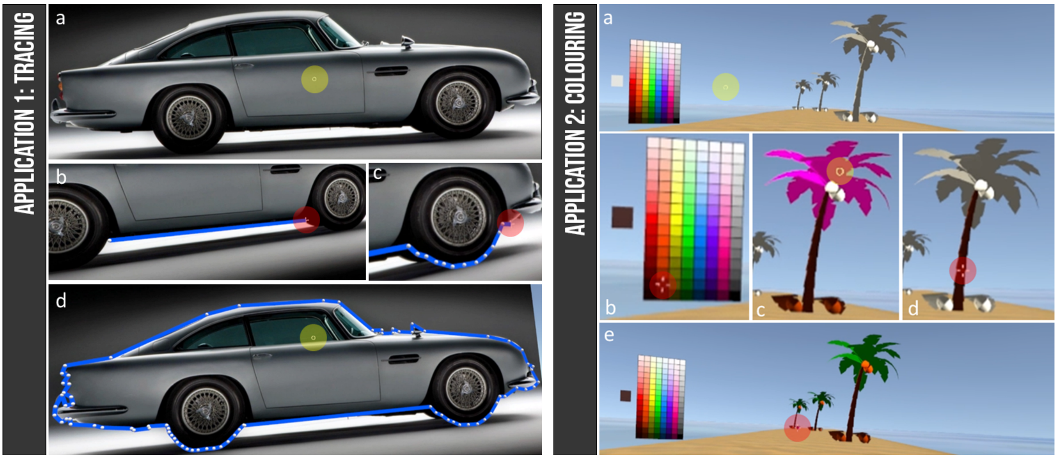


Fig. 4. Applications. The cursor highlighting serves as a visual distinction for the mode and is for illustration purposes only. Yellow indicates gaze mode, while red indicates head mode. Both examples shown used the GazeSwitch to complete the respective task. **Left**: The user can leverage the eyes' saccadic movement to cover a large distance before switching to head mode to place a marker (b). The user can activate and stay in head mode to place multiple markers in close proximity to trace out details, e.g., the curvature of the wheel (c). **Right**: The user selects the desired colour, typically in head mode (b). The user can leverage gaze mode to saccade quickly to a target (c). If it is a big target, e.g., a wide leaf, they can select without refinement. If the target is small, for e.g., the thin tree trunk, the user can activate head mode to refine the cursor position (d).

smaller features. The tracing line is produced by extending it from a previous marker to the current cursor position. An outline marker is placed by releasing the thumbpad on the controller.

5.1.2 Colouring. This task evaluates participants' experience of using the techniques, where gaze mode can be used to select and interact with larger targets and only use head refinement when needed to interact with small targets. As shown in Figure 4-Right, the application scene displays three palm trees positioned at increasing distances from the user. Participants are assigned the task of applying colour to various parts of the palm trees, with different sections available for colouring. The process involves selecting a colour from a palette situated 25° to the left of the beach scene and then choosing the specific part of the palm tree to be coloured. Hover feedback is provided as the cursor lands on colourable parts of the tree. The tree closest to the user will appear larger in visual angle compared to the farthest tree.

5.2 Results

Following the general inductive approach [44], two researchers independently coded interview transcripts focusing on participant experiences with the techniques. Initially, the first coder proposed six themes, which the second coder refined by removing two. A final consistency check, where both coders independently re-applied the themes, yielded 90% agreement and disagreements resolved through discussion. Both GazeSwitch and Eye+Head Pinpointing were generally well-received, with four out of the twelve participants preferring GazeSwitch for tracing and eight for colouring. Thematic analysis revealed key findings centred around the effectiveness and consistency of mode switching:

5.2.1 Effort. Seven participants stated that GazeSwitch requires less effort than Eye+Head Pinpointing mainly because it offers automatic mode switching. Participants further commented that

GazeSwitch makes the experience more "fluent" (P5) than Eye+Head Pinpointing as it allows them to focus on the task without the need to press a button to mode switch: "*GazeSwitch saves clicking... I can zone out on how to do it and just focus on what you are doing.*" (P1).

5.2.2 Speed. Six participants mentioned that Eye+Head Pinpointing was faster than GazeSwitch, primarily because they found pressing a button to switch into head mode an easy action: "*Eye+Head Pinpointing is quicker as it is straightforward and intuitive.*" (P3). Participants further noted that GazeSwitch could sometimes be time-consuming, particularly when the cursor got stuck in head mode due to their unfamiliarity with the technique. In contrast, five participants reported that GazeSwitch enabled them to complete tasks quicker than Eye+Head Pinpointing. This was attributed to the "accurate" (P1, P4) automatic mode switching offered by GazeSwitch: "*Automatic is quicker, more efficient, especially when you get the hang of it.*" (P6).

5.2.3 Control over Mode Switching. Nine participants noted that Eye+Head Pinpointing offers greater control over mode switching compared to GazeSwitch. This enhanced control makes the mode switching more "stable" (P1, P3) and allows the participants to explore the visual scene with their eyes and head freely: "*Eye+Head Pinpointing allows more manual control as I can move the head around without thinking about switching to head mode. I like the extra power... I can manually and precisely enter head mode when I want.*" (P8).

5.2.4 Stability of Mode Switching. Eight participants agreed that GazeSwitch is less stable than Eye+Head Pinpointing, resulting in a cursor that is "shaky" (P4, P10) and "jittery" (P3, P11). Participants noticed the instability is worse at the edges of the field of view (FOV), possibly due to eye tracking loss or when the "*eyes move away at the last minute before selection, presumably already moving on to the next outline point, causing the cursor to jump, which led to mistakes.*" (P2).

6 DISCUSSION

In this paper, we extended the insights from prior research to overcome limitations in existing eye-head mode switching techniques. Our contribution, GazeSwitch, leverages machine learning to optimise real-time switching between eye and head modes, enabling fast and precise hands-free pointing. Our findings demonstrate that adopting an ML-based classification approach reduces the occurrence of false positives resulting from natural head movements while efficiently detecting head gestures for input. The results from our two user studies not only validate the effectiveness of GazeSwitch in discrete target selection but also highlight its capability for continuous interaction, as demonstrated in our tracing task. This capability is significant for hands-free gaze and head interaction as it is traditionally only available for manual clutch-based techniques (*e.g.* Eye+Head Pinpointing) or other gaze-combined manual techniques (*e.g.* Gaze-Shifting).

GazeSwitch facilitates a smooth transition between pointing and refinement modes without requiring manual actions like Eye+Head Pinpointing or exaggerated head movements due to threshold limitations, as in BimodalGaze. The fast and adaptive mode switching facilitated by our classifier does not impose specific behaviours on users but instead allows them to act more freely. This has an impact on other parts of GazeSwitch. In both Eye+Head Pinpointing and BimodalGaze, feedback is of utmost importance in showing the current mode. As in the original implementations, Eye+Head Pinpointing forced users to go into head mode for selection, as gaze mode does not display any feedback. In BimodalGaze the cursor switches colour to signify a mode switch, which is necessary to ensure that users perform an exaggerated enough movement. In our work, we also implemented mode switch feedback by changing the circle into a crosshair. However, as our findings show that users could easily and seamlessly switch between modes, it minimises the need for explicit broadcasting of modes, potentially making the technique feel more fluid and synergistic.

However, the results of our studies also revealed trade-offs between GazeSwitch and the baseline switching techniques. Compared to Eye+Head Pinpointing, we found that GazeSwitch exhibited a higher error rate and longer selection time, but no significant differences were found in terms of overall head movement or the head movement required to enter head mode. There was also no significant difference in the onset of head mode or other performance metrics. These findings suggest that GazeSwitch allowed users to naturally utilise their heads, as participants commented on its effortless operation compared to manually activating head mode. However, the manual mode switch in Eye+Head Pinpointing offered greater control and stability, resulting in quicker and more accurate selections.

In comparison to BimodalGaze, GazeSwitch was perceived as less stable, possibly due to switching to gaze mode right before selection. Further analysis showed that participants attempted head refinement in 82.22% (SD=23.77) of failed selections were eventually made in gaze mode, and 97% (SD=5.49) of these could have succeeded if participants had selected in head mode. In these failed trials, participants maintained a final stable head mode for 0.81 seconds (SD=0.28). However, gaze velocity rises around 0.14 seconds before selection, unlike successful trials, where gaze velocity only increases after selection. This distinct pattern (shown in Figure 5 in Appendix C) suggests participants might have looked away before selection, triggering gaze mode an unnoticeable 0.14 seconds (SD=0.12) before selection, thus undoing head-mode refinement. This aligns with research showing fixation probability peaks before interaction [18, 36], potentially leading to "Late-Trigger errors" [20].

Moreover, participants entered head mode later with GazeSwitch compared to BimodalGaze, but also required less head movement. We also found no differences in the selection time or error rate between the two techniques, highlighting the difference between threshold-based and ML-based techniques. When using BimodalGaze, participants commented that they needed exaggerated head movements to activate head mode, which resulted in increased effort, and the early activation of head mode did not translate into shorter selection times. While the threshold-based approach demonstrated stability, it also contributed to a decrease in perceived performance, as participants may require time to familiarise themselves with the necessary head movement for activating head mode in BimodalGaze.

Our work and study findings highlight the effectiveness of the machine learning classification approach for classifying head movements into head-gaze and head gestures for hands-free and adaptive interaction, which we initially proposed as part of our HeadBoost paper [17]. In contrast with this prior work, where we evaluated the HeadBoost classifier in an offline context, this paper demonstrates its feasibility for real-time classification and eye-head pointing. This breakthrough opens up exciting opportunities for enabling various expressive and robust head movements for interaction, including head-based gestures, inferring user intentions based on head movements, and further exploration of other application areas.

6.1 Limitations and Future Work

When GazeSwitch performed smoothly, participants enjoyed its efficiency and seamless interaction. However, when it failed to perform optimally, participants noticed unexpected switched modes that interrupted task completion. Participants' feedback indicated mode switching instability as the main limitation of GazeSwitch, particularly noticeable around the edges of the field of view (FOV). Some found it helpful to adjust their head positioning slightly to centralise the target before attempting refinement. Further, participants favoured GazeSwitch over Eye+Head Pinpointing in the colouring application, which had a narrower scene compared to the tracing application. These observations suggest that GazeSwitch's performance may be affected by large visual angles. Given that GazeSwitch heavily relies on eye tracking for mode prediction, a decrease in eye tracking precision at extreme visual angles could contribute to instability and premature gaze shifts ("Late-Trigger errors"). In contrast, Eye+Head Pinpointing allows users to switch to head mode even if eye tracking fails, providing a fallback option to continue the task. Exploring alternatives such as defaulting to head mode when eye tracking fails, as proposed in error-aware gaze-based interaction techniques [38], or algorithms capable of identifying intended targets [*e.g.* 18], could mitigate these challenges.

We recognised two key limitations concerning machine learning. Firstly, our user studies demonstrated that GazeSwitch can effectively operate in diverse tasks and contexts, suggesting a sufficiently diverse dataset. However, while we collected 1500 trials in the training data, this was derived from only five participants. Future investigations could explore more representative data collection methods and alternative ML models to potentially enhance head-based classification performance and improve overall user experience. Secondly, like any ML-based classifier, the performance of GazeSwitch heavily relies on the quality and diversity of the collected data. Although we gathered data from a selection task with various target sequences, expanding data collection to encompass a broader range of eye tracking quality levels and different tasks and environments may result in a more robust classification system.

In this work, we evaluated our proposed technique within a virtual reality (VR), utilising a robust VR HMD equipped with accurate motion tracking using base stations. GazeSwitch, like Eye+Head Pinpointing and BimodalGaze techniques, is intended to function across various environments. As long as gaze and head tracking capabilities are available, any of these techniques, including GazeSwitch, can be applied in any environment. Hence, this concept could theoretically be extended to desktop-based environments by employing a remote eye tracker and a standard webcam for eye and head tracking, offering potential direction for further exploration in future research.

7 CONCLUSION

In this paper, we contribute GazeSwitch, an ML-based technique designed to enhance real-time mode switching for fast and accurate hands-free pointing. This approach allows users to leverage fast gaze pointing for covering long distances and efficiently switch to refine head pointing in various contexts, enabling the selection of discrete small targets and facilitating continuous interaction. Through our evaluation of GazeSwitch with two baseline switching techniques (Eye+Head Pinpointing and BimodalGaze), we observed that our proposed technique demands less effort for mode switching and enables users to interact seamlessly without the need for exaggerated head movements to trigger mode switching. However, our findings also highlight the performance limitations of GazeSwitch, which accounts for occasional instability in mode switching. In conclusion, GazeSwitch demonstrates the substantial potential for future developments in expressive head-based interactions and other application areas, broadening the possibilities for hands-free interaction.

ACKNOWLEDGMENTS

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant No. 101021229 GEMINI: Gaze and Eye Movement in Interaction).

REFERENCES

[1] Michael Ashmore, Andrew T. Duchowski, and Garth Shoemaker. 2005. Efficient eye pointing with a fisheye lens. In *Proceedings of Graphics Interface 2005* (Victoria, British Columbia) *(GI '05)*. Canadian Human-Computer Communications Society, Waterloo, CAN, 203–210. https://doi.org/10.5555/1089508.1089542

[2] Ishan Chatterjee, Robert Xiao, and Chris Harrison. 2015. Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA) *(ICMI '15)*. ACM, New York, NY, USA, 131–138. https://doi.org/10.1145/2818346.2820752

[3] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[4] Antoine Coutrot, Janet H. Hsiao, and Antoni B. Chan. 2018. Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods* 50, 1 (01 Feb 2018), 362–379. https://doi.org/10.3758/s13428-017-0876-8

[5] John D. Crawford and Tutis Vilis. 1991. Axes of eye rotation and Listing's law during rotations of the head. *Journal of Neurophysiology* 65, 3 (1991), 407–423. https://doi.org/10.1152/jn.1991.65.3.407 PMID: 2051188.

[6] Brendan David-John, Candace Peacock, Ting Zhang, T. Scott Murdison, Hrvoje Benko, and Tanya R. Jonker. 2021. Towards Gaze-Based Prediction of the Intent to Interact in Virtual Reality. In *ACM Symposium on Eye Tracking Research and Applications* (Virtual Event, Germany) *(ETRA '21 Short Papers)*. ACM, New York, NY, USA, Article 2, 7 pages. https://doi.org/10.1145/3448018.3458008

[7] Michael Dietz, Daniel Schork, Ionut Damian, Anika Steinert, Marten Haesner, and Elisabeth André. 2017. Automatic Detection of Visual Search for the Elderly using Eye and Head Tracking Data. *KI - Künstliche Intelligenz* 31, 4 (01 Nov 2017), 339–348. https://doi.org/10.1007/s13218-017-0502-z

[8] Stefan Dowiasch, Svenja Marx, Wolfgang Einhäuser, and Frank Bremmer. 2015. Effects of aging on eye movements in the real world. *Frontiers in Human Neuroscience* 9 (2015), 46. https://doi.org/10.3389/fnhum.2015.00046

[9] Robert P. W. Duin. 2002. The combining classifier: to train or not to train?. In *2002 International Conference on Pattern Recognition*, Vol. 2. IEEE, 765–770 vol.2. https://doi.org/10.1109/ICPR.2002.1048415

[10] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. ACM, New York, NY, USA, 754–768. https://doi.org/10.1145/3472749.3474784

[11] Ribel Fares, Shaomin Fang, and Oleg Komogortsev. 2013. Can We Beat the Mouse with MAGIC?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. ACM, New York, NY, USA, 1387–1390. https://doi.org/10.1145/2470654.2466183

[12] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. 2017. Toward Everyday Gaze Input: Accuracy and Precision of Eye Tracking and Implications for Design. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 1118–1130. https://doi.org/10.1145/3025453.3025599

[13] Dan Witzner Hansen, Henrik H. T. Skovsgaard, John Paulin Hansen, and Emilie Møllenbach. 2008. Noise Tolerant Selection by Gaze-Controlled Pan and Zoom in 3D. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (Savannah, Georgia) *(ETRA '08)*. ACM, New York, NY, USA, 205–212. https://doi.org/10.1145/1344471.1344521

[14] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*. Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[15] Thomas Haslwanter. 1995. Mathematics of three-dimensional eye rotations. *Vision Research* 35, 12 (1995), 1727–1739. https://doi.org/10.1016/0042-6989(94)00257-M

[16] Kenneth Holmqvist, Marcus Nyström, and Fiona Mulvey. 2012. Eye Tracker Data Quality: What It is and How to Measure It. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Santa Barbara, California) *(ETRA '12)*. ACM, New York, NY, USA, 45–52. https://doi.org/10.1145/2168556.2168563

[17] Baosheng James Hou, Joshua Newn, Ludwig Sidenmark, Anam Ahmad Khan, Per Bækgaard, and Hans Gellersen. 2023. Classifying Head Movements to Separate Head-Gaze and Head Gestures as Distinct Modes of Input. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. ACM, New York, NY, USA, Article 253, 14 pages. https://doi.org/10.1145/3544548.3581201

[18] Michael Xuelin Huang, Tiffany C.K. Kwok, Grace Ngai, Stephen C.F. Chan, and Hong Va Leong. 2016. Building a Personalized, Auto-Calibrating Eye Tracker from User Interactions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5169–5179. https://doi.org/10.1145/2858036.2858404

[19] Shahram Jalaliniya, Diako Mardanbegi, and Thomas Pederson. 2015. MAGIC pointing for eyewear computers. In *Proceedings of the 2015 ACM International Symposium on Wearable Computers* (Osaka, Japan) *(ISWC '15)*. ACM, New York, NY, USA, 155–158. https://doi.org/10.1145/2802083.2802094

[20] Manu Kumar, Jeff Klingner, Rohan Puranik, Terry Winograd, and Andreas Paepcke. 2008. Improving the accuracy of gaze input for interaction. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (Savannah, Georgia) *(ETRA '08)*. ACM, New York, NY, USA, 65–68. https://doi.org/10.1145/1344471.1344488

[21] Andrew Kurauchi, Wenxin Feng, Carlos Morimoto, and Margrit Betke. 2015. HMAGIC: Head Movement and Gaze Input Cascaded Pointing. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Corfu, Greece) *(PETRA '15)*. ACM, New York, NY, USA, Article 47, 4 pages. https://doi.org/10.

1145/2769493.2769550

[22] 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* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173655

[23] Linnéa Larsson, Marcus Nyström, Richard Andersson, and Martin Stridh. 2015. Detection of fixations and smooth pursuit movements in high-speed eye-tracking data. *Biomedical Signal Processing and Control* 18 (2015), 145–152. https://doi.org/10.1016/j.bspc.2014.12.008

[24] Diako Mardanbegi, Tobias Langlotz, and Hans Gellersen. 2019. Resolving Target Ambiguity in 3D Gaze Interaction through VOR Depth Estimation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300842

[25] Fionn Murtagh and Pedro Contreras. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2, 1 (2012), 86–97. https://doi.org/10.1002/widm.53

[26] Joshua Newn, Eduardo Velloso, Marcus Carter, and Frank Vetere. 2016. Multimodal Segmentation on a Large Interactive Tabletop: Extending Interaction on Horizontal Surfaces with Gaze. In *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces* (Niagara Falls, Ontario, Canada) *(ISS '16)*. ACM, New York, NY, USA, 251–260. https://doi.org/10.1145/2992154.2992179

[27] 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* (Gothenburg, Sweden) *(NordiCHI '16)*. ACM, New York, NY, USA, Article 43, 8 pages. https://doi.org/10.1145/2971485.2971490

[28] Marcus Nyström, Richard Andersson, Kenneth Holmqvist, and Joost Van De Weijer. 2013. The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods* 45 (2013), 272–288. https://doi.org/10.3758/s13428-012-0247-4

[29] Pontus Olsson. 2007. Real-time and Offline Filters for Eye Tracking. (2007), 42.

[30] Ken Pfeuffer, Jason Alexander, Ming Ki Chong, Yanxia Zhang, and Hans Gellersen. 2015. Gaze-Shifting: Direct-Indirect Input with Pen and Touch Modulated by Gaze. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. ACM, New York, NY, USA, 373–383. https://doi.org/10.1145/2807442.2807460

[31] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying Fixations and Saccades in Eye-Tracking Protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Palm Beach Gardens, Florida, USA) *(ETRA '00)*. ACM, New York, NY, USA, 71–78. https://doi.org/10.1145/355017.355028

[32] Ludwig Sidenmark, Christopher Clarke, Joshua Newn, Mathias N. Lystbæk, Ken Pfeuffer, and Hans Gellersen. 2023. Vergence Matching: Inferring Attention to Objects in 3D Environments for Gaze-Assisted Selection. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. ACM, New York, NY, USA, Article 257, 15 pages. https://doi.org/10.1145/3544548.3580685

[33] Ludwig Sidenmark, Christopher Clarke, Xuesong Zhang, Jenny Phu, and Hans Gellersen. 2020. Outline Pursuits: Gaze-Assisted Selection of Occluded Objects in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376438

[34] Ludwig Sidenmark and Hans Gellersen. 2019. Eye, Head and Torso Coordination During Gaze Shifts in Virtual Reality. *ACM Trans. Comput.-Hum. Interact.* 27, 1, Article 4 (Dec 2019), 40 pages. https://doi.org/10.1145/3361218

[35] Ludwig Sidenmark and Hans Gellersen. 2019. Eye&Head: Synergetic Eye and Head Movement for Gaze Pointing and Selection. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. ACM, New York, NY, USA, 1161–1174. https://doi.org/10.1145/3332165.3347921

[36] Ludwig Sidenmark and Anders Lundström. 2019. Gaze behaviour on interacted objects during hand interaction in virtual reality for eye tracking calibration. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (Denver, Colorado) *(ETRA '19)*. ACM, New York, NY, USA, Article 6, 9 pages. https://doi.org/10.1145/3314111.3319815

[37] Ludwig Sidenmark, Diako Mardanbegi, Argenis Ramirez Gomez, Christopher Clarke, and Hans Gellersen. 2020. BimodalGaze: Seamlessly Refined Pointing with Gaze and Filtered Gestural Head Movement. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Full Papers)*. ACM, New York, NY, USA, Article 8, 9 pages. https://doi.org/10.1145/3379155.3391312

[38] Ludwig Sidenmark, Mark Parent, Chi-Hao Wu, Joannes Chan, Michael Glueck, Daniel Wigdor, Tovi Grossman, and Marcello Giordano. 2022. Weighted Pointer: Error-aware Gaze-based Interaction through Fallback Modalities. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3585–3595. https://doi.org/10.1109/TVCG.2022.3203096

[39] Ludwig Sidenmark, Dominic Potts, Bill Bapisch, and Hans Gellersen. 2021. Radi-Eye: Hands-Free Radial Interfaces for 3D Interaction Using Gaze-Activated Head-Crossing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, Article 740, 11 pages. https://doi.org/10.1145/3411764.3445697

[40] Ludwig Sidenmark, Franziska Prummer, Joshua Newn, and Hans Gellersen. 2023. Comparing Gaze, Head and Controller Selection of Dynamically Revealed Targets in Head-Mounted Displays. *IEEE Transactions on Visualization and Computer Graphics* 29, 11 (2023), 4740–4750. https://doi.org/10.1109/TVCG.2023.3320235

[41] Alexandra Sipatchin, Siegfried Wahl, and Katharina Rifai. 2020. Accuracy and precision of the HTC VIVE PRO eye tracking in head-restrained and head-free conditions. *Investigative Ophthalmology & Visual Science* 61, 7 (2020), 5071–5071.

[42] Henrik Skovsgaard, Julio C. Mateo, John M. Flach, and John Paulin Hansen. 2010. Small-Target Selection with Gaze Alone. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (Austin, Texas) *(ETRA '10)*. ACM, New York, NY, USA, 145–148. https://doi.org/10.1145/1743666.1743702

[43] Robert M. Steinman. 1965. Effect of Target Size, Luminance, and Color on Monocular Fixation. *J. Opt. Soc. Am.* 55, 9 (Sep 1965), 1158–1164. https://doi.org/10.1364/JOSA.55.001158

[44] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2 (2006), 237–246. https://doi.org/10.1177/1098214005283748

[45] Eduardo Velloso, Marcus Carter, Joshua Newn, Augusto Esteves, Christopher Clarke, and Hans Gellersen. 2017. Motion Correlation: Selecting Objects by Matching Their Movement. *ACM Trans. Comput.-Hum. Interact.* 24, 3, Article 22 (apr 2017), 35 pages. https://doi.org/10.1145/3064937

[46] Mélodie Vidal, Andreas Bulling, and Hans Gellersen. 2012. Detection of Smooth Pursuits Using Eye Movement Shape Features. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Santa Barbara, California) *(ETRA '12)*. ACM, New York, NY, USA, 177–180. https://doi.org/10.1145/2168556.2168586

[47] Oleg Špakov. 2012. Comparison of Eye Movement Filters Used in HCI. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Santa Barbara, California) *(ETRA '12)*. ACM, New York, NY, USA, 281–284. https://doi.org/10.1145/2168556.2168616

[48] Oleg Špakov, Poika Isokoski, and Päivi Majaranta. 2014. Look and Lean: Accurate Head-Assisted Eye Pointing. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Safety Harbor, Florida) *(ETRA '14)*. ACM, New York, NY, USA, 35–42. https://doi.org/10.1145/2578153.2578157

[49] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. ACM, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[50] Raimondas Zemblys, Diederick C. Niehorster, Oleg Komogortsev, and Kenneth Holmqvist. 2018. Using machine learning to detect events in eye-tracking data. *Behavior Research Methods* 50, 1 (01 Feb 2018), 160–181. https://doi.org/10.3758/s13428-017-0860-3

[51] Shumin Zhai, Carlos Morimoto, and Steven Ihde. 1999. Manual and Gaze Input Cascaded (MAGIC) Pointing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. ACM, New York, NY, USA, 246–253. https://doi.org/10.1145/302979.303053

[52] Xinyong Zhang, Xiangshi Ren, and Hongbin Zha. 2008. Improving eye cursor's stability for eye pointing tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. ACM, New York, NY, USA, 525–534. https://doi.org/10.1145/1357054.1357139

A GAZESWITCH MODE SWITCHING LOGIC

Algorithm 1: GazeSwitch logic for switching between Gaze mode and Head mode

```
$Mode \leftarrow Gaze\ Mode$;
if $ML\_label$ is $Head\ Mode$ then
    $Mode \leftarrow Head\ Mode$;
else if $ML\_label$ is $Gaze\ Mode$ then
    if $Previous\_Mode$ is $Head\ Mode$ AND $Dispersion \leq 3.6$ then
        $Mode \leftarrow Head\ Mode$;
    else
        $Mode \leftarrow Gaze\ Mode$;
```

B GAZESWITCH FEATURES

B.1 Full List of Features

Table 2. Full feature list, as proposed by HeadBoost [17]

Category	Features
Shape-based	Slope, Range, Mean Velocity, Peak Velocity, Mean Acceleration, Peak Acceleration, Integral, Energy, Wavelength [46], Spatial features in the positional signal (P_D, P_{CD}, P_{PD}, P_R) defined by Larsson et al. [23]
Noise-based	Dispersion [31], Standard Deviation, RMS, BCEA [16, 43, 50], RMS-diff, BCEA-diff, Mean-diff, Median-diff[29, 50], Rayleightest [23, 50]
Spectral	Rolloff, Centroid, Entropy [7], Flatness
Correlation-based	Correlation
Timing-based	Time since last saccade (200 °/s)

B.2 Final RFA selected 28 features

The final selected features in the GazeSwitch ML model is listed below, 15 of the final features are based on eye movement, 13 are based on head movement, and spanned all 1024ms of the sampled past feature vectors, suggesting a combination of eye-head dynamics contributed to the model performance.

- at current timestamp: time since last saccade, eye-in-head energy, eye-in-world energy, eye-in-world wavelength in the polar direction, head energy, head slope, head wavelengths in the combined (Az and Pol) direction and the polar direction, head integral in the polar direction, RMS difference between first and second half of the window for eye-in-world and head angles, median difference in eye-in-world between the first and second half of the window, the PCA PCD measure for the eye-in-world angle.
- at -0.16s, RMS difference between first and second half of the window for head angles.
- at -0.32s, eye-in-world energy, eye-in-world wavelength in the polar direction, RMS difference between first and second half of the window for head angles.
- at -0.48s, eye-in-world energy.
- at -0.64s, eye-in-world wavelength in the azimuth direction, head positional wavelength in the Z axis, head spectral centroid.
- at -0.8s, eye-in-world wavelength in the polar direction, head spectral centroid in the polar direction, the PCA dispersion of the head.
- at -1s, RMS difference between the first and second half of the window for the eye-in-world angle.
- at -1.024s, head wavelength in the polar direction, time since last saccade, eye-in-world integral in the polar direction.

C “LATE-TRIGGER ERROR” VISUALISATION

Figure 5 visualises the “Later-Trigger error” observed during interaction using the GazeSwitch technique, error is characterised by a last-minute saccade away from the target just before selection, undoing head refinement, causing selection error. Failed selections in gaze mode displayed a notable increase in gaze velocity approximately 140 ms before selection. In contrast, successful trials showed an increase in gaze velocity only after the selection, indicating a distinct temporal pattern associated with selection success. In trials where selection occurred in gaze mode but failed, participants maintained a final stable head mode for 0.81 seconds (SD=0.28), only breaking into gaze mode 0.14 seconds (SD=0.12) before selection. During this head refinement period, 97.47% (SD=5.49) of attempts were able to align the cursor on the target, with a minimal average cursor-target offset of $0.22°/s$ (SD=0.08). However, at the last moment before selection, the refinement was undone by breaking into gaze mode, most likely due to the increased gaze velocity indicative of a ‘saccade away’ from the target, causing selection error. These results suggest the “Late-Trigger error” may be a top contributor to errors and perceived instability when using GazeSwitch.

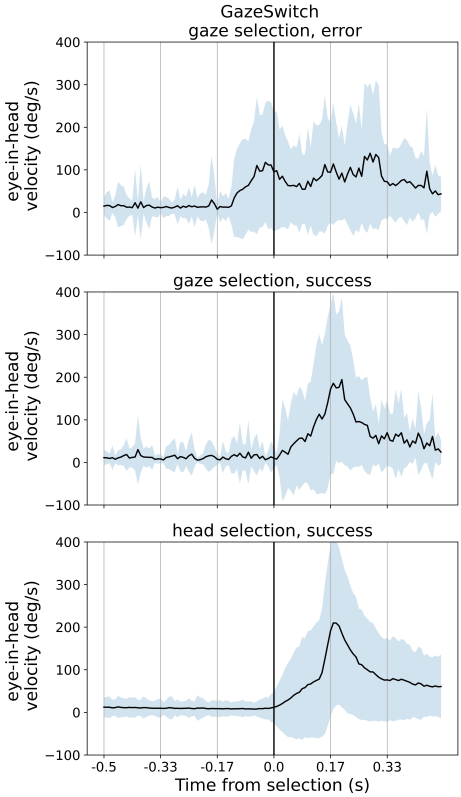


Fig. 5. Eye-in-head velocity from 500 ms before, to 500 ms after selection by selection mode and outcome. Mean over all trials is shown as the solid black line, standard deviation in blue shade.

Received November 2023; revised January 2024; accepted March 2024