Available on: http://gjetr.ep-journals.org/ E-ISSN: 3051-3782, P-ISSN: 3051-3774

Global Journal of Engineering and Technology Review

Volume: 02 Issue: 08 August 2026

Gaze-Based Interaction on Commodity Hardware: A Survey of Appearance-Based Estimation, Selection Techniques, And Multimodal Error Correction

ZIA RASHEED
University of Hertfordshire

Darda Rahman

University of Bedfordshire, UK

Mehmoona Akram
MSc International Business | Ulster University Birmingham | CIPD Level 5 (In Progress)

Published Date: August 31, 2026 **Article DOI : 10.65150/EP-gjetr/V2E8/2026-07**

ABSTRACT: Eye gaze is the fastest natural pointing signal available to interactive systems, but for three decades accurate gaze interaction has remained tied to dedicated infrared eyetracking hardware. The convergence of deep appearance-based gaze estimation, ubiquitous front-facing cameras, and on-device neural inference has opened a credible path to gaze interaction on unmodified consumer devices, and with it a distinct research agenda whose problems differ from those of laboratory eye tracking. This survey reviews that agenda across three layers. First, appearance-based gaze estimation: the CNN architectures, large-scale datasets, calibration strategies, and head-pose handling techniques that determine achievable accuracy on RGB cameras, and the persistent gap between benchmark and in-use performance. Second, gaze as an input modality: dwellbased selection, the Midas-touch problem, and the design space of confirmation mechanisms, smoothing filters, and target-aware interfaces that convert noisy fixation estimates into reliable selection. Third, the emerging layer of multimodal and implicit error correction, including physiological and behavioural signals as post-hoc feedback for detecting and repairing erroneous selections — an approach that relocates robustness from prevention to recovery. The survey synthesises evaluation practices across the three layers, identifies the accessibility deployment context as both the most demanding testbed and the strongest motivation for the field, and sets out open problems in personalisation, drift, privacy-preserving on-device processing, and the extension of gaze interaction to spatial computing platforms.

KEYWORDS: gaze estimation, eye tracking, human–computer interaction, appearancebased methods, dwell selection, Midas touch, multimodal interaction, survey

1. INTRODUCTION

The human oculomotor system orients the fovea to targets of interest with saccades reaching several hundred degrees per second, making gaze direction available to an interactive system hundreds of milliseconds before any manual pointing action can begin. This latency advantage has motivated gaze-based interaction research since the early 1990s, when Jacob's foundational work established both the promise of gaze pointing and its defining obstacle — the Midas-touch problem, whereby a channel that is simultaneously perceptual and communicative risks converting every act of looking into an unintended command [3]. Yet for most of the intervening period, gaze interaction remained confined to specialised contexts, because accurate gaze measurement required infrared-illuminated, purpose-built eye-tracking hardware whose cost and setup excluded mainstream deployment.

That constraint has weakened decisively. Deep appearance-based gaze estimation — learning a direct mapping from ordinary camera images to gaze direction — demonstrated on large in-the-wild datasets that useful accuracy is attainable from the RGB cameras already present in phones, tablets, and laptops [1], [2]. Simultaneously, on-device neural accelerators made real-time CNN inference feasible on consumer hardware without cloud dependence, and the accessibility community — for whom hands-free interaction is a necessity rather than a convenience — supplied both the most demanding requirements and the clearest use cases. The result is a coherent research area, distinct from laboratory eye tracking, whose central question is systems-level: how to build reliable gaze interaction from unreliable gaze estimates on uncontrolled hardware.

This survey reviews that area in three layers: the estimation layer (Section 2), which determines the quality of the raw gaze signal; the interaction layer (Section 3), which converts the signal into selection; and the correction layer (Section 4), an emerging body of work that uses multimodal and implicit feedback to detect and repair the errors the first two layers cannot eliminate. Section 5 reviews evaluation practice, Section 6 examines the accessibility deployment context, and Section 7 sets out open problems.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

2. THE ESTIMATION LAYER: APPEARANCE-BASED GAZE FROM RGB CAMERAS

Classical eye tracking measures gaze geometrically, using infrared illumination to create corneal reflections whose position relative to the pupil centre encodes gaze direction with sub-degree accuracy under controlled conditions [9]. Appearance-based estimation abandons the geometric apparatus and learns the image-to-gaze mapping directly, an approach transformed by deep learning and large-scale data collection. Zhang et al.'s MPIIGaze work established in-the-wild RGB estimation with a multimodal CNN combining eye images and head pose, and quantified the strong dependence of accuracy on head-pose variation [2], [5]. Krafka et al.'s GazeCapture demonstrated crowd-scale data collection on mobile devices — nearly 1,500 participants — and the accompanying iTracker network showed that population-trained models generalise usefully across users and devices [1]. Subsequent architectural work has explored pictorial intermediate representations, transformer backbones, and fewshot personalisation, with reported benchmark errors on standard datasets descending from roughly six degrees toward three degrees of visual angle over the past decade [6].

Three problems dominate the gap between benchmark and deployed performance. Head motion: because the eye-image appearance of a gaze direction varies with head pose, free head movement is the largest single error source; the design space spans data augmentation across pose, explicit pose-conditioned models, and geometric normalisation that warps inputs to a canonical pose before inference. Personspecificity: population models leave a per-user residual — eye geometry, eyewear, appearance — that per-user calibration reduces substantially; the research frontier is minimising calibration burden through few-shot adaptation and implicit calibration from natural interaction events, since explicit point-fixation protocols are tedious for all users and impossible for some. Drift and environment: illumination change, device repositioning, and posture shifts degrade accuracy within a session, motivating continuous or opportunistic re-calibration rather than one-time setup. Across all three, the trend of the literature is away from treating estimation as a static per-frame problem and toward treating it as an adaptive, temporally extended one — a framing that connects the estimation layer directly to the interaction layer above it.

3. THE INTERACTION LAYER: FROM FIXATIONS TO SELECTION

A gaze estimate is not an input event; converting one into the other is the interaction layer's task, and its constraints have been stable since Jacob's analysis [3]. Dwell selection — triggering on sustained fixation within a target — remains the dominant technique for its simplicity and hands-free purity, and remains limited by the Midastouch dilemma: short dwell thresholds cause false activations during perceptual looking, long thresholds sacrifice the speed that motivates gaze input, and no threshold eliminates the conflict because looking is what the eyes are for. Fitts's-law analyses of gaze pointing confirm high potential throughput but show it eroded in practice by jitter, inaccuracy near small targets, and the dwell delay itself [13].

The design space of remedies is well mapped. Explicit confirmation adds a second action — blink, wink, facial gesture, key press in gaze-plus-key techniques such as EyePoint [15], or a second gaze at a confirm region — trading hands-free purity or speed for precision, and excluding users who cannot produce the confirming action. Signal engineering attacks the noise instead: velocity-adaptive smoothing filters that suppress fixation jitter without lagging saccades, fixation-detection algorithms that classify samples into fixations and saccades before selection logic runs, and confidence gating that discounts low-quality frames. Interface adaptation reshapes the target side: enlarged effective target areas, gaze-aware magnification, probabilistic target inference that snaps noisy fixations to the most plausible target, and prediction models that rerank targets by context and history, reducing the accuracy the estimator must deliver. The layers compose: deployed systems combine filtering, target inference, and a selection mechanism chosen for the user population — a composition whose overall behaviour is under-studied relative to its components.

4. THE CORRECTION LAYER: MULTIMODAL AND IMPLICIT ERROR HANDLING

A more recent strand accepts that upstream layers will err at some rate on commodity hardware and asks how errors can be detected and repaired after the fact, cheaply enough that they stop mattering. The enabling observation comes from physiological computing [10], [11]: users respond involuntarily to interaction errors, and those responses are measurable. Error-related brain potentials in EEG are the cleanest such signal but require instrumentation incompatible with everyday deployment; consumergrade channels — heart-rate variability from wrist wearables [12], electrodermal activity, pupil dilation from the same camera performing gaze estimation, and behavioural signatures such as post-selection gaze regressions and hesitation patterns — sacrifice signal cleanliness for deployability.

Architecturally, implicit correction reframes selection as a sequential decision problem with delayed, noisy reward [4]: a selection is emitted, an implicit response window is observed, a detected error signature triggers automatic repair — undo, or substitution of the next-ranked candidate target — and the reward stream trains personalisation over time. The approach's attraction is that it leaves correct selections fast (no added confirmation burden) while making incorrect ones self-repairing; its challenges are signal validity (arousal responses are individual, context-confounded, and slower than the interaction loop), trust calibration (an uncalibrated correction loop can degrade interaction below its gaze-only baseline, motivating staged-trust designs that disable the loop until per-user baselines are learned), and evaluation (separating the loop's contribution from the upstream layers requires ablation designs that few studies yet report). The strand is young, and its most complete instantiations remain system papers rather than controlled studies — but it addresses the precise residual that three decades of prevention-oriented techniques have not closed, which is why this survey treats it as a layer of the field rather than a curiosity.

5. EVALUATION PRACTICES

Evaluation conventions differ sharply across the three layers, complicating synthesis. The estimation layer reports angular error on benchmark datasets under standardised splits — rigorous but only loosely predictive of interactive usability, since interaction outcomes depend on error's temporal structure, not just its magnitude. The interaction layer reports throughput, selection time, and error rates in controlled pointing tasks, occasionally text-entry rates for gaze typing — comparable across studies but usually conducted with able-bodied participants on hardware better than deployment targets. The correction layer, newest and least standardised, reports repair latency, correction precision and recall, and longitudinal personalisation gains, rarely with common baselines. The survey's reading of this landscape is that the field's binding evidence gap is integrative:

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Corresponding Author: ZIA RASHEED

end-to-end evaluations of complete pipelines, on commodity devices, with the user populations for whom the systems are intended, over sessions long enough for drift and adaptation to matter. Component benchmarks are necessary; they are demonstrably not sufficient to predict system-level usability, and the accessibility context of the next section makes the gap concrete.

6. THE ACCESSIBILITY DEPLOYMENT CONTEXT

Assistive communication is simultaneously the field's strongest motivation and its hardest test. For users who cannot operate touch or speech input reliably, gaze is often the only high-bandwidth channel available, and commodity-hardware gaze interaction converts an expensive specialist provision into a mainstream tablet application — a cost transformation with direct access consequences. The same population, however, concentrates every difficulty the survey has catalogued: involuntary and repetitive movements that inflate false activations, inability to follow explicit calibration instructions (motivating stimulus-driven and implicit calibration), fatigue sensitivity that penalises long dwell thresholds, and carer-mediated configuration that demands robustness to non-expert setup. Systems targeting this context therefore integrate the full stack — per-user artifact modelling, compliance-free calibration, target prediction, and increasingly implicit correction — and their design choices anticipate needs of the broader population: techniques developed for users who cannot comply with calibration serve everyone who would rather not, and error-repair loops built for users who cannot issue explicit corrections benefit any hands-busy scenario. The accessibility literature also imposes the field's most serious treatment of privacy and consent, given continuous camera and physiological sensing of vulnerable users: on-device inference, nonpersistence of raw frames, and standards-compliant data handling appear here first and most rigorously [8].

7. OPEN PROBLEMS AND OUTLOOK

Five problems define the field's near-term agenda. Personalisation efficiency: closing the population-to-individual accuracy gap with seconds, not minutes, of implicit adaptation — few-shot and continual-learning formulations are promising but underevaluated in deployment. Drift as a first-class problem: session-scale robustness receives far less study than per-frame accuracy, despite dominating real usability; opportunistic re-calibration from natural interaction events needs systematic treatment. Trustworthy implicit correction: validating consumer-grade physiological and behavioural error signals against ground truth, characterising their failure modes, and designing staged-trust controllers whose worst case is the gaze-only baseline. Privacypreserving architecture: gaze and physiological streams are biometric and healthadjacent data; on-device processing, feature minimisation, and auditable nonpersistence should become reported, comparable properties of published systems rather than implementation footnotes. Spatial computing: head-mounted AR and VR platforms invert the hardware premise — gaze sensing is built in and head pose is free — making the interaction and correction layers surveyed here immediately portable to three-dimensional target spaces, where the Midas-touch problem returns in new geometry. The through-line of all five is the survey's central claim: on commodity hardware, gaze interaction is a systems problem, and progress now depends less on another degree of benchmark accuracy than on the integration, adaptation, and evaluation of complete pipelines in the contexts where they are needed most.

REFERENCES

1) K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2176–2184.
2) X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4511–4520.
3) R. J. K. Jacob, "What you look at is what you get: eye movement-based interaction techniques," in Proc. ACM CHI Conf. Human Factors in Computing Systems, 1990, pp. 11–18.
4) R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA: MIT Press, 2018.
5) X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 41, no. 1, pp. 162–175, 2019.
6) S. Park, A. Spurr, and O. Hilliges, "Deep pictorial gaze estimation," in Proc. European Conf. Computer Vision (ECCV), 2018, pp. 721–738.
7) Bulling and H. Gellersen, "Toward mobile eye-based human-computer interaction," IEEE Pervasive Computing, vol. 9, no. 4, pp. 8–12, 2010.
8) P. Majaranta and A. Bulling, "Eye tracking and eye-based human–computer interaction," in Advances in Physiological Computing, Springer, 2014, pp. 39–65.
9) D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 32, no. 3, pp. 478–500, 2010.
10) W. Picard, Affective Computing. Cambridge, MA: MIT Press, 1997.
11) H. Fairclough, "Fundamentals of physiological computing," Interacting with Computers, vol. 21, no. 1–2, pp. 133–145, 2009.
12) F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," Frontiers in Public Health, vol. 5, art. 258, 2017.
13) P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," Journal of Experimental Psychology, vol. 47, no. 6, pp. 381–391, 1954.
14) K. Ruhland et al., "A review of eye gaze in virtual agents, social robotics and HCI," Computer Graphics Forum, vol. 34, no. 6, pp. 299–326, 2015.
15) M. Kumar, A. Paepcke, and T. Winograd, "EyePoint: Practical pointing and selection using gaze and keyboard," in Proc. ACM CHI Conf. Human Factors in Computing Systems, 2007, pp. 421–430.
16) C. H. Morimoto and M. R. M. Mimica, "Eye gaze tracking techniques for interactive applications," Computer Vision and Image Understanding, vol. 98, no. 1, pp. 4–24, 2005.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Corresponding Author: ZIA RASHEED