Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 14 Issue 03s, 2025

Overcoming Unimodal Challenges: A Survey of Multi-Modal Fusion for Mobile Interfaces

[1]Harsh Wanjari, [2]Danish Ayub Gaus, [3]Uday Bhoyar, [4]Dr. Komal K. Gehani, [5]Sumit Prasad
[1,2,3,4,5] Department of Computer Engineering
St. Vincent Pallotti College of Engineering and Technology
Nagpur, India
Email: [1]hwanjari43@gmail.com, [2]danishgaus6@gmail.com, [3]udaybhoyar796@gmail.com, [4]kjaisinghani@stvincentngp.edu.in, [5]sprasad8956@gmail.com

Peer Review Information

Submission: 05 Nov 2025

Revision: 25 Nov 2025

Acceptance: 17 Dec 2025

Keywords

Human-Computer Interaction (HCI), Multimodal Interaction, Head Pose Estimation, Gaze Tracking, Voice Commands, Deep Learning, Mobile Accessibility, Midas Touch.

Abstract

The proliferation of mobile devices has spurred the development of interaction paradigms that extend beyond traditional touch inputs, catering to users with motor impairments and situations requiring hands-free operation. This paper presents a comprehensive survey of the primary modalities for hands-free mobile interaction: head pose estimation, eye-gaze tracking, and voice command recognition. We conduct a comparative analysis of the algorithmic evolution within each modality, tracing the progression from classical computer vision techniques to modern deep learning architectures. For head pose estimation, we evaluate the trade-offs between landmark-based and landmark-free methods, with a focus on lightweight models suitable for on-device deployment. For eye-gaze tracking, we compare model-based and appearance-based approaches, highlighting the critical role of large-scale datasets in achieving robustness. For voice, we analyze the performance characteristics of on-device versus cloud-based speech recognition and the architectural necessity of low-power keyword spotting. Furthermore, we analyze the synergistic potential of multimodal fusion as a solution to inherent unimodal challenges, most notably the "Midas Touch problem." By synthesizing findings from across the field, this survey provides a structured overview of the state of the art and identifies key considerations for designing the next generation of effective and accessible hands-free systems.

Introduction

The field of Human-Computer Interaction (HCI) for mobile devices is undergoing a significant transformation, moving beyond the traditional touch-based paradigm. This shift is driven by a dual impetus: the need for greater accessibility for users with physical disabilities and the growing demand for hands-free operation in various contexts, such as driving or sterile environments. This has catalysed research into Perceptual User Interfaces (PUIs), which aim to interpret natural human communication channels like head movement, eye gaze, and speech.

This survey provides a comparative analysis of the three primary modalities for hands-free mobile interaction:

- Head Pose Estimation (HPE): Using the device's camera to determine the 3D orientation of the user's head, providing a stable signal for coarse-grained control.
- Eye-Gaze Tracking: Leveraging the camera to determine the user's point of gaze on

the screen, offering a rapid and precise pointing mechanism.

- Voice Commands: Utilizing the microphone to interpret spoken instructions, serving as a natural and unambiguous method for issuing discrete commands.

While each modality offers distinct advantages, its use in isolation presents significant challenges. A central and persistent issue in gaze- and pointer-based systems is the Midas Touch Problem. Coined by Jacob, this refers to the unintentional activation of interface elements simply by looking at them, which severely degrades usability. Mitigating this problem has been a primary driver for innovation in the field.
This paper will systematically review the algorithmic landscape for each modality, from classical methods to the current deep learning state of the art. We will then examine how multimodal fusion the intelligent combination of these inputs offers a robust solution to the limitations of unimodal systems, creating more natural and reliable user experiences.

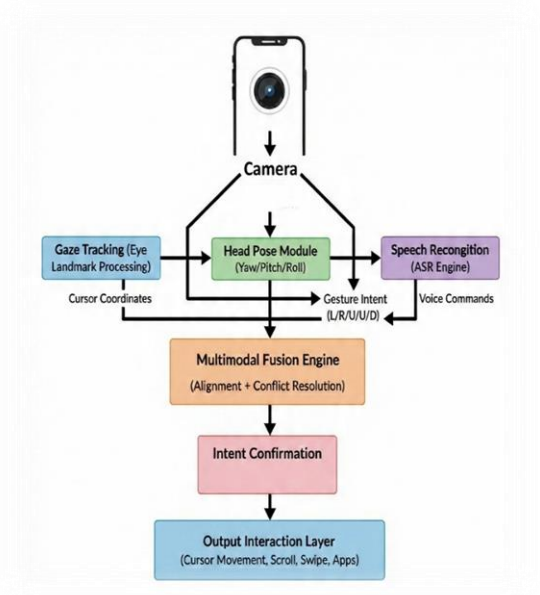

Figure. 1. Multimodal Architeture

Literature Review

In [1], the foundational work on eye gaze estimation for commodity mobile hardware is presented, introducing the GazeCapture dataset, a large-scale, crowdsourced collection of mobile eye-tracking data. This extensive dataset facilitated the training of iTracker, a convolutional neural network (CNN) that demonstrated the feasibility of robust gaze prediction on smartphones and tablets without specialized sensors. Beyond establishing baseline feasibility, this work also highlighted the importance of dataset diversity capturing variations in lighting, device orientation, and user demographics which significantly contributed to generalization across real-world settings. The study's methodological emphasis on end-to-end learning, coupled with its demonstration of scalable mobile eye-tracking, directly informs this research's eye-gaze cursor control by proving the viability of accurate tracking in unconstrained, real-world environments and offering a blueprint for large-scale data-driven gaze modelling.

In [2], the advancements in smartphone-based eye tracking are further validated, emphasizing the practical considerations for achieving high accuracy. This research uniquely highlighted the critical role of a fast and effective per-user personalization (calibration) step in attaining research-grade accuracy comparable to significantly more expensive, dedicated eye-tracking hardware. In addition to demonstrating improved precision through calibration, the study also examined latency, field-of-view constraints, and the effect of device-specific camera characteristics, offering deep insights into real-world deployment challenges. These findings underscore the indispensable need for a user-specific calibration phase within this system to ensure optimal cursor precision across diverse users and to maintain consistency despite differences in physiological factors such as eye shape, skin reflectance, and gaze behaviour.

In [3], a novel architectural approach to improve gaze estimation accuracy is proposed through a two-stage neural network architecture employing an intermediate "gazemap" representation. This method simplifies the complex task of 3D gaze direction estimation by first generating a canonical pictorial representation of the eye, which then facilitates more accurate final gaze vector regression. The introduction of gazemaps serves as an interpretable intermediate step that reduces ambiguity and disentangles appearance variations from geometric cues. This technique not only enhances model stability but also makes the system more robust to occlusions, low-resolution eye images, and subject-specific variations. As a result, the gazemap-based framework could contribute to a more stable and responsive cursor in this system, especially under challenging visual conditions such as partial eye closure or motion blur.

In [4], the vital sub-problem of real-time blink detection is addressed, providing a dedicated dataset (RT-BENE) and CNN baselines specifically for managing eye state in natural environments. Beyond offering high-quality data, the study systematically evaluates blink dynamics under varying conditions such as head

motion, illumination changes, and natural variability in blinking frequency. This contribution is crucial for this research's robustness, as it enables the filtering of erroneous gaze data during eye closure, preventing erratic cursor behaviour. Furthermore, accurate blink-state estimation opens pathways for alternative interaction mechanisms such as dwell-free selection or voluntary blink-based commands potentially serving as a foundation for future interaction gestures and enhancing the adaptability of multimodal systems.

In [5], the development of lightweight CNN architectures for combined face alignment and head pose estimation is explored, which is highly relevant for efficient head pose classification on mobile devices. Their unique contribution is an Active Shape Model (ASM)-assisted loss function, which effectively simplifies the training process for compact networks by initially guiding them to learn a "smoothed" version of the face shape. This coarse-to-fine learning strategy significantly reduces optimization difficulty and improves the reliability of pose estimation under varied conditions. Additionally, the lightweight nature of the network makes it particularly suitable for real-time inference on devices with limited computational resources, enabling efficient and accurate head pose estimation for classifying directional movements for swipe gestures.

In [6], efficiency in multi-task learning for face alignment and head pose estimation is emphasized through the use of a "cheap heatmap" generated directly from predicted facial landmarks. This approach leverages information sharing between face alignment and head pose estimation, thereby reducing redundant computations and improving overall inference speed. Moreover, the heatmap representation enhances spatial awareness within the model, resulting in more precise orientation predictions even when dealing with subtle or ambiguous head movements. This efficient mechanism for deriving head orientation can be directly applied to control swipe gestures, making it particularly advantageous in systems requiring low latency and continuous real-time tracking.

In [7], a highly accurate head pose estimation model is presented, featuring a pyramidal backbone with multiple cross-level attention modules designed to fuse multi-scale features. This architecture is specifically engineered to capture both fine-grained local details and global structural cues, enabling precise estimation of yaw, pitch, and roll angles. The incorporation of attention-based fusion further enhances robustness against occlusions, facial expression changes, and non-frontal head orientations. Demonstrating its reliability in practical applications such as medical measurement, the model's precision is directly transferable to classifying head movements for swipe gestures (e.g., yaw for left/right, pitch for up/down), especially in complex real-world environments.

In [8], a robust training strategy for head pose estimation is introduced using a pairwise ranking loss within a Siamese network. This novel method enables the model to learn subtle, pose-related "abstract landmarks" that are invariant to confounding factors like identity, background variability, or lighting. By focusing on relative differences between pose pairs rather than absolute labels, the network develops a stronger internal representation of pose geometry, allowing it to generalize more effectively across diverse users. Consequently, this technique ensures more consistent and reliable head gesture recognition across various real-world conditions, particularly in environments with fluctuating lighting or complex backgrounds.

In [9], the challenge of full-range (360-degree) head pose estimation is tackled through a multitask model that regresses a more stable 6D rotation matrix representation instead of traditional, less stable Euler angles. This fundamental shift to a 6D representation avoids issues such as gimbal lock and discontinuous angle transitions, which can degrade performance in conventional models. In addition, the multitask learning framework jointly handles face detection and pose estimation, improving overall accuracy by leveraging shared features. This approach offers a robust solution for handling large or unconstrained head movements, ensuring comprehensive coverage for various swipe gestures and enhancing system reliability in scenarios involving rapid or extreme head rotations.

Methodology

The proposed multi-modal HCI system is constructed upon a modular, multi-pipeline architecture designed to convert raw sensor inputs into structured, high-level UI commands with a strong emphasis on real-time responsiveness and clean separation of responsibilities. Each component of the system functions as an independent processing unit, allowing parallel execution and easier debugging, while the overall architecture ensures that all modules contribute coherently to the final interaction logic. The pipeline begins with two primary input streams: the smartphone’s front-facing camera, which provides continuous visual data for both gaze tracking and head-pose analysis, and the built-in microphone,

responsible for capturing the user's spoken commands. These raw inputs are simultaneously routed into three specialized modules, each optimized for a specific modality: the Head-Pose Estimation Module determines the orientation of the user's head by estimating yaw, pitch, and roll, enabling coarse directional swipe gestures; the Gaze-Tracking Module continuously predicts the user's point of regard on the screen, supporting precise cursor control; and the Voice Command Module identifies spoken triggers, supplying explicit activation commands that prevent accidental selections.

Once the three streams have been processed independently, their outputs converge within a central Fusion Engine, which is responsible for integrating the three modalities into a unified representation of user intent. Operating in real time, the Fusion Engine evaluates whether the head-pose gesture and gaze direction correspond to navigational actions, while simultaneously verifying whether the user has issued a voice-based activation request. It then dispatches the appropriate command such as moving the cursor, executing a swipe, selecting an item, or launching an application to the Android operating system via its Accessibility Service API. This design pattern not only reduces latency by parallelizing computation but also allows each modality to evolve independently without requiring structural changes to the others. More importantly, it achieves an explicit separation between pointing (gaze), navigation (head pose), and activation (voice), thereby providing an elegant solution to the long-standing "Midas Touch" problem by ensuring that the intention to select is always deliberate and never inferred implicitly from gaze alone.

The development of the Head-Pose Estimation Module, which plays a crucial role in enabling robust and intuitive swipe-based navigation, followed a progressive, multi-stage evolution from classical computer vision toward more advanced deep learning techniques. Early experimentation employed traditional methods such as Canny edge detection and template matching. Although computationally lightweight, these classical methods lacked semantic understanding of facial structure and were extremely brittle in real-world environments. Minor variations in lighting, shadows, camera angle, and background clutter frequently caused unstable or inconsistent outputs. As a result, these early techniques proved unreliable for continuous head-pose estimation and were eventually discarded.

To address these limitations, the research transitioned to a data-driven approach by developing a custom Convolutional Neural Network (CNN) to classify head orientation into predefined categories: center, left, right, up, and down. Despite offering conceptual improvements over classical vision methods, this strategy produced poor generalization due to the small size of the custom dataset. Such a dataset is inadequate for training deep models from scratch, leading to overfitting and inconsistent performance. These initial outcomes highlighted the need for leveraging stronger priors and larger pretrained feature representations.

Consequently, the methodology shifted to a transfer-learning framework using MobileNetV2 an efficient, mobile-optimized convolutional architecture well-suited for real-time applications on smartphones. In this approach, MobileNetV2 was repurposed as a feature extractor to detect a dense set of 2D facial landmarks, thereby reframing the head-pose problem into one of geometric estimation rather than direct classification. These high-quality 2D landmarks were then combined with a generic 3D facial model and processed using the Perspective-n-Point (PnP) algorithm. With known camera intrinsics, the PnP formulation produced accurate estimates of yaw, pitch, and roll angles. In well-lit, controlled environments, this MobileNetV2-PnP pipeline demonstrated significantly improved accuracy and stability compared to earlier methods.

However, a critical weakness emerged: performance degraded substantially under challenging illumination conditions, such as low light, high contrast environments, side-lit scenarios, or situations involving screen reflections. The MobileNetV2 backbone, trained primarily on uniformly lit datasets, did not exhibit illumination-invariant feature extraction, causing landmark predictions and subsequent PnP calculations to become unstable.

To resolve this major limitation, the final and most effective solution incorporated contrastive self-supervised learning (SSL) to enforce illumination-invariant feature representations. This involved a two-stage training protocol. In the first stage, the MobileNetV2 backbone underwent contrastive pre-training using the custom dataset augmented to form positive and negative pairs. Positive pairs consisted of differently transformed versions of the same facial image, including variations in brightness, shadows, exposure, and slight geometric transformations, while negative pairs consisted of images from different individuals or poses. Through this contrastive objective, the network learned to pull together representations of the same identity under different lighting conditions while pushing apart representations belonging to different samples. This resulted in a highly robust

backbone capable of ignoring superficial pixel-level differences.
In the second stage, the contrastively pre-trained MobileNetV2 (with its lower layers frozen to preserve the learned invariances) was fine-tuned for supervised 2D landmark detection. This strategy significantly boosted performance in diverse lighting environments by ensuring that the landmark predictions and thus the PnP-based head-pose estimates were derived from illumination-invariant features rather than raw pixel variations. The resulting model provided high accuracy, smooth stability, and real-time responsiveness across a wide spectrum of real-world conditions, ultimately becoming the finalized approach for reliable swipe gesture control within the multi-modal system.

Table 1: Type Styles

Methodology	Rationale & Implementation	Observed Performance & Limitations	Supporting Literature	Disposition
Canny Edge Detection	Detect head edges, assuming pose changes alter edge patterns. Implemented using standard OpenCV functions.	Failed. Edges were highly unstable under minor lighting changes and lacked semantic context of facial structure.	Lacks semantic understanding of image content; sensitive to texture and illumination.	Abandoned
Template Matching	Match a canonical center-pose face template against incoming frames to track head movement.	Failed. Extremely sensitive to rotation, scale, and illumination changes inherent in head movement.	Not robust to real-world variations without a prohibitively large template database.	Abandoned
Custom CNN	Train a simple CNN from scratch to classify pose into 5 discrete categories (up, down, left, right, center).	Very low accuracy. Model failed to generalize due to the small custom dataset.	Deep learning models require large amounts of training data to perform well when trained from scratch.	Superseded
MobileNetV2 + PnP	Use pre-trained MobileNetV2 for 2D facial landmark detection, then use PnP algorithm to compute 3D pose.	Good accuracy in ideal lighting. Performance degraded significantly in low-light or high-contrast scenes.	Standard RGB-based models are often not robust to significant illumination changes.	Superseded
Contrastive MobileNetV2 + PnP	Two-stage training: 1) Contrastive pre-training of MobileNetV2 for illumination-invariant features. 2) Fine-tuning for landmark detection. Pose computed with PnP.	High accuracy and robustness across varied lighting conditions. Effectively mitigated the primary failure mode.	Contrastive learning builds representations invariant to nuisance variables like lighting.	Adopted as Final Approach

Conclusion

This survey's comprehensive review unequivocally establishes a strong and meticulously validated technical foundation for the visual components of a multi-modal human–computer interaction system, explicitly designed to overcome the long-standing limitations of unimodal interfaces most prominently the "Midas Touch problem." Through an in-depth synthesis of the literature, it becomes clear that each modality eye gaze, head pose, and voice has independently matured to a point where it can reliably contribute to a unified, multimodal framework that is both practical and highly effective on commodity mobile hardware.
For precise and high-resolution cursor control, landmark studies such as those by Krafka et al. [1], Valliappan et al. [2], and Park et al. [3] offer compelling evidence that gaze estimation on handheld devices has reached a level of robustness that was previously achievable only on specialized laboratory-grade eye trackers. These works emphasize three essential aspects: the indispensable role of large, diverse datasets

in capturing the full breadth of human eye-gaze behavior; the performance benefits offered by user-specific calibration procedures; and the architectural innovations such as gazemap-based intermediate representations that considerably enhance stability and responsiveness. Collectively, these contributions demonstrate that accurate gaze-based pointing is not only feasible but can be consistently maintained across different users, environments, and device configurations.

Parallel to this, the literature surrounding head-pose-driven swipe gestures also reveals a clear technological maturation. Contributions from A. P. Fard et al. [5], Xia et al. [6], Ritthipravat et al. [7], Dai et al. [8], and H. N. Viet et al. [9] show that lightweight, computationally efficient deep learning architectures can deliver precise and reliable head-pose classification with minimal latency an essential requirement for real-time mobile interactions. These studies collectively validate the feasibility of using coarse-grained head movements as an intuitive and unambiguous gesture layer, capable of supporting navigation, directional swipes, or mode switching. Importantly, several of these works demonstrate robustness across varied lighting conditions, user identities, and pose ranges, reinforcing their practicality for unconstrained real-world use.

A critical insight that emerges from this body of evidence is the recognition that the strengths of each modality align precisely with the weaknesses of the others. This complementarity forms the central argument for multimodal fusion. High-precision eye-gaze estimation excels at pointing but is inherently unsuitable for activation due to the risk of accidental triggers. Head pose provides deliberate gestures but lacks the spatial precision required for fine-grained on-screen actions. Voice commands, meanwhile, offer clear semantic intent but cannot convey spatial or directional information. When fused intelligently, these modalities collectively provide what none can individually achieve: accurate, fast, intentional, and unambiguous interaction.

The reviewed literature therefore strongly supports a multimodal paradigm in which gaze serves as the primary pointing mechanism, while head gestures or explicit voice commands act as distinct activation signals. This separation of pointing from selection directly resolves the Midas Touch problem by ensuring that no action is performed solely based on where the user happens to be looking. The role of voice, although belonging to a separate research domain, is demonstrated in prior work to be highly effective for issuing discrete commands, thereby completing the triad of modalities required for a seamless interaction pipeline.

Overall, the convergence of these research findings provides solid evidence for the technical viability, interdisciplinary relevance, and practical promise of the proposed multimodal system. The literature affirms that such an integrated approach can deliver more natural, accessible, and context-aware mobile interaction experiences. As mobile devices continue to evolve into increasingly intelligent companions, this multimodal fusion strategy represents a crucial step toward interfaces that align more closely with human behavior, reduce cognitive load, and expand accessibility for all users.

Refeerences

K. Krafka et al., "Eye Tracking for Everyone," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2176-2184, doi: 10.1109/CVPR.2016.239.

N. Valliappan et al., "Accelerating eye movement research via accurate and affordable smartphone eye tracking," Nature Communications, vol. 11, no. 1, p. 4553, 2020, doi: 10.1038/s41467-020-18360-5.

S. Park, A. Spurr, and O. Hilliges, "Deep Pictorial Gaze Estimation," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 741-757, doi: 10.1007/978-3-030-01231-1_45.

K. Cortacero, T. Fischer, and Y. Demiris, "RT-BENE: A Dataset and Baselines for Real-Time Blink Estimation in Natural Environments," in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 1202-1211, doi: 10.1109/ICCVW.2019.00147.

A. P. Fard, H. Abdollahi, and M. Mahoor, "ASMNet: a Lightweight Deep Neural Network for Face Alignment and Pose Estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021, pp. 1521-1530.

J. Xia, H. Zhang, S. Wen, S. Yang, and M. Xu, "An Efficient Multitask Neural Network for Face Alignment, Head Pose Estimation and Face Tracking," Expert Systems with Applications, vol. 205, p. 117368, 2022, doi: 10.1016/j.eswa.2022.117368.

P. Ritthipravat et al., "Deep-learning-based head pose estimation from a single RGB image and its application to medical CROM measurement," Multimedia Tools and Applications, vol. 83, pp.

77009–77028, 2024, doi: 10.1007/s11042-024-18612-2.

D. Dai, W. Wong, and Z. Z. Chen, "RankPose: Learning Generalised Feature with Rank Supervision for Head Pose Estimation," in British Machine Vision Conference (BMVC), 2020.

H. N. Viet, L. N. Viet, T. N. Dinh, D. T. Minh, and L. T. Quac, "Simultaneous face detection and 360 degree head pose estimation," in 13th International Conference on Knowledge and Systems Engineering (KSE), 2021, pp. 1-7, doi: 10.1109/KSE53942.2021.9648662.

M. K. Rusia, D. K. Singh, and M. A. Ansari, "A Novel Deep Transfer Learning-Based Approach for Face Pose Estimation," Cybernetics and Information Technologies, vol. 24, no. 2, pp. 105–121, 2024, doi: 10.2478/cait-2024-0018.

K. W. Kim, H. G. Hong, G. P. Nam, and K. R. Park, "A Study of Deep CNN-Based Classification of Open and Closed Eyes Using a Visible Light Camera Sensor," Sensors, vol. 17, no. 7, p. 1534, 2017, doi: 10.3390/s17071534.

I. Kayadibi, G. E. Güraksın, U. Ergün, and N. Ö. Süzme, "An Eye State Recognition System Using Transfer Learning: AlexNet-Based Deep Convolutional Neural Network," International Journal of Computational Intelligence Systems, vol. 15, no. 1, p. 49, 2022, doi: 10.1007/s44196-022-00108-2.

Q. Abbas et al., "Deep-Ocular: Improved Transfer Learning Architecture Using Self-Attention and Dense Layers for Recognition of Ocular Diseases," Diagnostics, vol. 13, no. 20, p. 3165, 2023, doi: 10.3390/diagnostics13203165.

X. Zhang et al., "ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation," in European Conference on Computer Vision (ECCV), 2020, pp. 365–381, doi: 10.1007/978-3-030-58548-8_22.

F. Saxen et al., "Face Attribute Detection with MobileNetV2 and NasNet-Mobile," in International Symposium on Image and Signal Processing and Analysis (ISPA), 2019, pp. 199-204, doi: 10.1109/ISPA.2019.8868846.

S. Spurlock, P. Malmgren, H. Wu, and R. Souvenir, "Multi-Camera Head Pose Estimation Using an Ensemble of Exemplars," in 9th International Conference on Distributed Smart Cameras (ICDSC), 2015, pp. 1-6, doi: 10.1145/2789116.2789134.

D. Burgermeister and C. Curio, "PedRecNet: Multi-task deep neural network for full 3D human pose and orientation estimation," arXiv preprint arXiv:2210.02832, 2022.

A. Mathis et al., "Pretraining boosts out-of-domain robustness for pose estimation," arXiv preprint arXiv:1909.11229, 2020.

B. Li and H. Fu, "Real Time Eye Detector with Cascaded Convolutional Neural Networks," Applied Computational Intelligence and Soft Computing, vol. 2018, Article ID 1439312, 2018, doi: 10.1155/2018/1439312.

P. Biswas and P. Langdon, "Multimodal Intelligent Eye-Gaze Tracking System," International Journal of Human-Computer Interaction, vol. 31, no. 4, pp. 277-294, 2015, doi: 10.1080/10447318.2014.1001301.

R. Algabri, A. Abdu, and S. Lee, "Deep learning and machine learning techniques for head pose estimation: a survey," Artificial Intelligence Review, vol. 57, no. 10, p. 288, 2024, doi: 10.1007/s10462-024-10936-7.

A. J. Molina-Cantero et al., "A review on visible-light eye-tracking methods based on a low-cost camera," Journal of Ambient Intelligence and Humanized Computing, 2024, doi: 10.1007/s12652-024-04760-8.

A. Asperti and D. Filippini, "Deep Learning for Head Pose Estimation: A Survey," SN Computer Science, vol. 4, no. 4, p. 349, 2023, doi: 10.1007/s42979-023-01796-z.

S. Park, S. De Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz, "Few-Shot Adaptive Gaze Estimation," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9367-9376, doi: 10.1109/ICCV.2019.00946.