The Efficiency Tradeoffs of Gaze Tracking

Heather Kong
University of Toronto Schools
heatheryykong@gmail.com

Alex Zhang
University of Toronto Schools
zhaal@utschools.ca

Daniel Ganjali
University of Toronto Schools
ganda@utschools.ca

Malcolm Frei
University of Toronto Schools
frema@utschools.ca

***Abstract*—Gaze estimation is a core component of assistive technology and accessibility systems. Existing gaze estimation methods are commonly divided into three major categories: 2D mapping-based, 3D model-based, and appearance-based approaches [1]. While 3D geometric systems can achieve high precision, they often require specialized calibration procedures, complex modeling, and dedicated hardware. Appearance-based deep learning methods are highly effective under natural conditions, but they typically require large annotated datasets and substantial computational resources [3]. In assistive settings, additional challenges such as unintended activation, increased cognitive load, and deployment constraints further complicate system design [2]. This paper presents a hybrid gaze tracking framework for real-time operation on RGB cameras. The proposed system combines geometric iris modeling, CLAHE-based contrast normalization, radial gradient validation, Kalman filtering, and blink-triggered facial landmark analysis. Experimental evaluation shows a mean performance of 75.36 FPS with a 68.38% reduction in positional variance, demonstrating a computationally efficient and accessible alternative for assistive gaze interaction.**

I. Introduction

Gaze direction is widely used in assistive technology and robotics as a proxy for user attention and intent [3]. For individuals with motor disabilities, eye tracking can function as a primary means of communication and interaction [2]. Liu et al. classify gaze estimation methods into three major categories: 2D mapping-based methods, 3D model-based methods, and appearance-based methods [1]. Each category presents trade-offs in hardware complexity, calibration requirements, computational cost, and robustness.

In assistive technology, Fischer-Janzen et al. identify several practical challenges, including the *Midas Touch* problem, calibration burden, and increased cognitive load during interaction [2]. Recent multimodal deep learning systems such as Dual Focus-3D combine eye appearance with head orientation to improve generalization under natural conditions [3]. However, such systems depend on large annotated datasets and computationally expensive training procedures.

A. Motivation

Existing gaze estimation methods present a clear trade-off between accuracy, complexity, and deployability. High-precision geometric systems can achieve excellent gaze estimation accuracy, but they often rely on specialized hardware, complex calibration procedures, and computationally expensive parameter estimation [1]. These requirements make them difficult to deploy in everyday environments, especially in assistive technology contexts where low-cost and accessible solutions are essential.

At the other end of the design spectrum, appearance-based deep learning systems offer strong performance under natural conditions and can generalize better across users and environments [3]. However, these approaches typically require large annotated datasets, high training cost, and substantial runtime resources, which limits their practicality for real-time deployment on commodity hardware.

There is therefore a need for systems that achieve reliable gaze estimation without the hardware burden of geometric systems or the computational cost of end-to-end deep learning regression models. This work proposes a hybrid framework that preserves the interpretability of geometric methods while incorporating machine learning selectively for optimization and event detection. By reserving heavier inference for blink-triggered events and maintaining a lightweight monitoring stage otherwise, the system balances efficiency, accessibility, and performance.

B. Related Works

Gaze tracking systems are typically divided into 2D mapping-based, 3D model-based, and appearance-based methods [1]. 2D mapping-based methods estimate the point of regard by learning a regression between eye features and calibration points. These systems can achieve high accuracy under constrained conditions, but they are sensitive to head motion and generally require explicit calibration [1].

3D model-based approaches reconstruct the optical axis using geometric modeling of the eyeball. These methods provide high precision, but the most effective implementations often require multiple cameras or accurate anatomical parameter estimation [1]. Appearance-based methods instead regress gaze direction directly from eye or face images. Recent hybrid solutions combine convolutional neural networks with head orientation measurements to improve performance in natural conditions [3]. Although such approaches improve generalization, they require large annotated datasets and computationally intensive training [3].

In assistive technology, gaze control introduces additional constraints such as the *Midas Touch* problem, calibration complexity, and cognitive load during interaction [2]. These constraints emphasize the need for systems that are not only accurate, but also efficient and deployable.

C. Problem Definition

Existing gaze tracking solutions present trade-offs between accuracy, robustness, and deployability. 2D mapping systems require calibration and controlled head motion [1], while 3D geometric systems often require specialized hardware and complex parameter estimation [1]. Appearance-based deep learning systems require large annotated datasets and high computational cost [3].

Assistive technology applications further require real-time operation, low hardware complexity, minimal calibration burden, and stable cursor control under natural head positions [2]. The problem addressed in this work is therefore the following:

> Design a real-time gaze tracking system that operates on a standard RGB laptop webcam, minimizes calibration requirements, preserves geometric interpretability, and maintains stable gaze estimation suitable for assistive interaction.

Rather than fully replacing geometric reasoning with deep learning, this work uses a hybrid approach that combines geometric modeling with machine learning-assisted optimization.

II. Methodology

This section presents the design process of the proposed hybrid gaze tracking system. We first describe the data used for development and evaluation, then detail the system architecture and processing pipeline, followed by the evaluation methodology and additional analysis procedures.

A. Data

The system is designed for, developed with, and evaluated using monocular RGB webcam input under natural indoor conditions. Eye regions are extracted from full-face frames using heuristic eye detection techniques. Full facial landmark inference is only performed after a blink event is detected, using previously buffered frames in which the eyes remain open.

To align with assistive deployment constraints discussed in [2], data collection is performed using commodity hardware without infrared illumination or multi-camera setups. For validation and benchmarking, eye detection and crop accuracy are compared against landmark-based eye approximations, which serve as proxy ground truth references [1].

The dataset used in development consists of the following:

- Continuous webcam frames at a resolution of 1280×720;
- Annotated eye landmark approximations for benchmarking;
- Temporal frame sequences for motion modeling and blink-triggered inference.

Only limited calibration is required, consistent with the objective of reducing calibration burden in assistive systems [2].

B. Proposed Solution

The proposed system combines geometric modeling with learning-based parameter optimization. This design is motivated by the trade-offs identified in the gaze estimation taxonomy of Liu et al. [1]: 2D mapping methods require calibration and are sensitive to head position, 3D model-based methods require complex geometric reconstruction, and appearance-based deep learning systems require extensive training data and computational cost [3].

Rather than reproducing one classical paradigm directly, this work integrates geometric reasoning with selectively used machine learning optimization.

Eye regions are extracted from RGB frames and preprocessed using:

- Contrast Limited Adaptive Histogram Equalization (CLAHE),
- Adaptive thresholding,
- Dark percentile filtering.

CLAHE enhances local contrast and improves robustness under varied lighting conditions, addressing lighting variability issues commonly encountered in geometric systems [1].

Dark blob detection is then performed to identify candidate pupil regions. Each candidate is filtered using area constraints, ellipticity constraints, and radial gradient validation. Radial gradient validation ensures that intensity transitions follow the expected dark-to-light structure of the pupil–iris–sclera boundary, reducing false positives caused by shadows and skin texture.

Iris modeling is performed using Random Sample Consensus (RANSAC). The pupil boundary is approximated using the circular equation

$$(x - x_c)^2 + (y - y_c)^2 = r^2. \tag{1}$$

To mitigate outliers from eyelashes, eyelids, and reflections, RANSAC is used for robust circle fitting. This preserves geometric interpretability while improving robustness under partial occlusion, consistent with the geometric modeling principles discussed in [1].

Cursor stability is improved using a Kalman filter with a constant-velocity state model,

$$x_k = [p_x,\ p_y,\ v_x,\ v_y]^T. \tag{2}$$

The prediction and update equations smooth jitter across frames. Temporal stabilization directly addresses usability concerns in assistive systems, where jitter can increase cognitive load and unintended activation risk [2].

Instead of full gaze regression using CNN architectures as in appearance-based approaches [3], machine learning is used for hyperparameter tuning, threshold optimization, and blink event recognition. This selective integration preserves performance while improving robustness.

C. Pipeline

The proposed system follows a two-stage event-driven processing pipeline designed to balance computational efficiency with reliable gaze estimation.

Incoming frames from a monocular RGB webcam are first processed through a lightweight monitoring stage. In this stage, eye regions are extracted from the full frame using heuristic eye detection techniques. These cropped eye regions are analyzed using a small convolutional neural network that performs continuous blink detection.

The system maintains a temporal frame buffer storing recent frames in which the eyes are detected as open. When a blink exceeding a predefined duration threshold is detected, the system transitions to the second stage. The buffered open-eye frames are then processed using full-frame facial landmark estimation. Stable facial landmarks, including nose and eye landmarks, are used to refine the eye center estimate. A gaze vector is computed from the relative displacement between the pupil region and the estimated eye center, and a combined gaze estimate is then formed from both eyes.

By restricting full facial landmark inference to blink-triggered events, the system reduces the computational cost associated with continuous gaze estimation while still enabling reliable gaze direction inference for assistive interaction scenarios.

D. Evaluation Methodology

The system is evaluated using quantitative metrics aligned with assistive usability requirements [2]. The primary performance metrics are:

- Frames per second (FPS),
- Capture rate (%),
- Area reduction (%),
- Temporal stability, measured by variance reduction,
- Proxy localization error in pixels.

Evaluation is performed on continuous frame sequences under multiple lighting conditions and mild head motion. Results are reported as averages across five sequences. This evaluation structure aligns with the robustness criteria summarized in [1] and the usability considerations discussed in [2].

E. Additional Analysis

To determine optimal configurations, systematic parameter analyses were performed. The following parameters were varied:

- CLAHE clip limit,
- Dark percentile threshold,
- RANSAC tolerance,
- Kalman process noise covariance,
- Kalman measurement noise covariance.

Each configuration was evaluated using capture rate and stability metrics. The optimal configuration was selected by maximizing the score

$$\text{Score} = \alpha \cdot \text{Capture Rate} - \beta \cdot \text{Variance}, \tag{3}$$

where α and β are empirically chosen weights.

This methodology intentionally avoids heavy end-to-end deep learning architectures [3], avoids multi-camera geometric reconstruction [1], and minimizes calibration requirements emphasized in assistive technology research [2]. Instead, it integrates robust geometric modeling with lightweight machine learning optimization to achieve real-time gaze tracking suitable for assistive deployment.

III. Results

This section presents the quantitative evaluation of the proposed hybrid geometric–machine learning gaze tracking framework.

A. Real-Time Performance

Real-time capability is crucial for assistive deployment scenarios. The system was evaluated on an Apple Silicon M4 CPU with integrated GPU at a resolution of 1280×720. The proposed pipeline achieved:

- Mean FPS: 75.3554,
- Standard deviation: 2.4455.

These results confirm that the hybrid geometric and machine learning approach maintains real-time performance without requiring dedicated GPU acceleration, in contrast to more computationally intensive appearance-based systems [3].

B. Temporal Stability

Temporal stability was evaluated by comparing raw pupil center variance with Kalman-filtered variance. Stability improvement is defined as

$$\text{Stability Improvement} = \frac{\sigma^2_{\text{raw}} - \sigma^2_{\text{filtered}}}{\sigma^2_{\text{raw}}}. \tag{4}$$

TABLE I
TEMPORAL STABILITY RESULTS.

Metric	Value
Raw variance	6.502913
Filtered variance	2.056533
Variance reduction (%)	68.38%

Kalman filtering reduced jitter substantially, while the added smoothing introduced negligible latency.

C. Comparative Positioning

Compared to classical 2D mapping methods, the proposed approach maintains effectiveness under mild variations in head and eye positioning without requiring explicit and complex calibration procedures [1]. Compared to deep learning and appearance-based systems that integrate multimodal features [3], the proposed method achieves real-time performance on commodity hardware with reduced computational complexity, making it more practical for assistive technology environments.

Although the system does not reconstruct full 3D gaze vectors, it achieves stable gaze tracking sufficient for assistive cursor control tasks, which is an important design goal for practical human-centered systems [2].

D. Summary of Findings

The results demonstrate the following:

- The system operates in real time on commodity hardware;
- Geometric modeling improves detection reliability;
- Temporal filtering reduces jitter without compromising responsiveness;
- Machine learning-assisted optimization improves parameter selection without requiring full end-to-end training.

These findings support the design objective of balancing deployability, effectiveness, and computational efficiency.

IV. Conclusion

This paper presented a hybrid gaze tracking framework designed for efficient assistive interaction using a standard RGB webcam. The proposed system combines heuristic eye detection, geometric iris modeling, radial gradient validation, Kalman filtering, and blink-triggered facial landmark analysis. By reserving heavier computation for blink-triggered events and maintaining a lightweight monitoring stage otherwise, the system achieves strong real-time performance while preserving interpretability and low deployment cost.

Experimental evaluation showed a mean speed of 75.36 FPS and a 68.38% reduction in positional variance after temporal filtering. These results demonstrate that a hybrid geometric approach, combined with lightweight machine learning optimization, can provide stable gaze-based interaction without relying on high-cost hardware or large end-to-end regression models.

Future work should focus on improving robustness under more extreme lighting conditions, stronger head pose variation, and more diverse user-specific eye characteristics. Extending the system with more refined calibration, stronger blink classification, and broader user testing would also improve its readiness for real assistive deployment.

References

[1] J. Liu, J. Chi, H. Yang, and X. Yin, "In the eye of the beholder: A survey of gaze tracking techniques," *Pattern Recognition*, vol. 132, p. 108944, 2022.

[2] A. Fischer-Janzen, T. M. Wendt, and K. Van Laerhoven, "A scoping review of gaze and eye tracking-based control methods for assistive robotic arms," *Frontiers in Robotics and AI*, vol. 11, p. 1326670, 2024.

[3] A. Bendimered, R. Iguernaissi, M. M. Nawaf, R. Cherif, S. Dubuisson, and D. Merad, "Dual Focus-3D: A Hybrid Deep Learning Approach for Robust 3D Gaze Estimation," *Sensors*, vol. 25, no. 13, p. 4086, 2025.