
From Gaze to Data: Creating New Marketing Insights

Hyeonsik Jo*
Kookmin University

Seoyoung Oh*
Kookmin University

Junhyeok Lee
Kookmin University

Abstract

This study proposes a gaze-journey analysis framework that addresses the inherent limitations of purchase-log-based approaches, which capture only outcomes but not the underlying consumer decision processes. The proposed model performs ROI-level gaze estimation and enhances robustness by selectively incorporating 6DRepNet head-pose cues only when the primary model, Gaze360, exhibits high predictive uncertainty. Frame-level gaze predictions are further integrated with multi-object tracking to reconstruct each shopper’s continuous gaze flow as a coherent temporal sequence. By quantitatively mapping gaze directions to shelf ROIs and restoring customer-specific gaze transitions over time, the framework generates gaze-based shopping-journey data that cannot be obtained from purchase logs alone. This design enables stable gaze estimation even under the challenging conditions of top-mounted CCTV—low facial resolution, occlusion, and frequent side/rear views. Moreover, the proposed framework provides practical utility for applications such as product placement optimization, attention analysis, and in-store path design. Our code is available at <https://anonymous.4open.science/status/gaze2data-72CC>.

1 Introduction

Consumer gaze behavior in retail environments is closely associated with the purchase decision-making process[1]. Prior studies[2] have experimentally demonstrated that gaze is a key predictor of product choice and that prolonged or repeated visual fixation on a specific item significantly increases the likelihood of its purchase. These findings indicate that gaze

data provide essential information that supports major decision-making tasks in retail marketing, including shelf layout optimization, advertising effectiveness evaluation, and purchase intention analysis.

Nevertheless, conventional marketing analysis has relied primarily on purchase outcome data—that is, information on products for which payment has been completed. Such purchase logs record only the final selected item and fail to capture behavioral signals such as which products customers visually attended to or which product alternatives they considered. The log merely states that a given product was purchased, while all gaze-based exploration, comparison, and switching processes disappear entirely. As a result, purchase-log-centered approaches possess structural limitations in understanding and analyzing consumer behavior.

If retailers could move beyond purchase outcomes and instead quantify the customer’s gaze-based shopping journey, it would be possible to conduct more multidimensional consumer behavior analyses that traditional methods cannot capture. Achieving this, however, requires that each customer’s gaze trajectory be tracked continuously over time, ensuring that sequences of entry, movement, shelf transitions, and exit within the scene remain coherent. Without this temporal continuity, gaze information becomes fragmented into isolated frames, making meaningful shopping journey analysis infeasible.

Motivated by this need, we propose a customer gaze journey tracking model that utilizes ROI (Region of Interest)-based Gaze360[3] gaze estimation as the primary signal, employs 6DRepNet[4] as an auxiliary model when the Gaze360 prediction exhibits high uncertainty (i.e., high quantile scores), and integrates multi-person tracking to maintain consistent identity across time. The proposed model maps frame-level gaze predictions onto shelf ROIs in a quantitative manner and links each customer’s gaze flow over time through tracking. This enables the construction of a new form of consumer gaze journey dataset suitable for practical marketing analysis—something that cannot be obtained from conventional purchase logs.

*Equal contribution.

2 Related Work

This section reviews prior studies relevant to CCTV-based customer gaze analysis in convenience-store environments.

2.1 Gaze Estimation

Early gaze-estimation research predominantly assumed high-resolution, frontal-face settings. Models trained on datasets such as MPIIGaze[5] and Eye-Diap[6] achieved strong accuracy by leveraging clearly visible eye regions, yet these assumptions do not hold in real-world CCTV footage. More recent approaches such as Gaze360 employ multi-frame architectures with six-degree-of-freedom rotation representations to handle side and rear viewpoints more robustly. However, prediction instability persists when the eyes are barely visible or the face is extremely small. In particular, abrupt gaze flips and sudden vector jumps frequently occur in retail surveillance videos, hindering direct deployment. To address this issue, our work introduces a stabilization mechanism that re-estimates unreliable frames using an auxiliary model rather than relying solely on per-frame outputs from Gaze360.

2.2 Head Pose Estimation

Head pose estimation provides relatively consistent directional cues even when facial details are limited, making it a valuable auxiliary signal for CCTV-based gaze analysis. Models such as 6DRepNet and Hopenet[7] can estimate head rotations from a single frame and thus generalize well across domains. Nonetheless, head pose does not encode fine-grained gaze information, and in daily scenarios human gaze often deviates from head orientation. Consequently, methods relying solely on head pose accumulate errors and fail to reach the precision required for applications such as ROI mapping. Recent work[8] has therefore incorporated head pose as a prior or stabilization constraint in gaze-estimation pipelines. Following this trend, we employ head-pose-based predictions as a fallback for abnormal frames, improving the temporal stability and continuity of the overall gaze sequence.

2.3 Object Tracking

To extend gaze analysis from single-person to multi-person retail environments, reliable object tracking is essential for maintaining consistent identities across frames and enabling behavior-level interpretation. Tracking-by-detection pipelines typically combine YOLO-based detectors[9] with Kalman filter[10] or IoU-based association[11] for online tracking. More advanced methods such as DeepSORT[12]

utilize deep appearance embeddings, while ByteTrack[13], OC-SORT[14], and related approaches leverage Transformer-based relational modeling or enhanced association strategies. Some studies aim to recover occlusions or model inter-object interactions. However, these trackers often rely on full-body appearance cues, which are limited in ceiling-mounted CCTV settings where subjects are partially visible and frequently occluded, making ID drift common. Considering these characteristics, we adopt a lightweight IoU-based tracker rather than embedding-heavy alternatives, achieving a practical balance between robustness and computational efficiency.

3 Proposed Framework

3.1 Overview

In this study, we design a gaze estimation pipeline that effectively handles the diverse eye-visibility conditions encountered in overhead CCTV environments by branching into three pathways according to the level of visibility. As shown in Figure 1, the model dynamically selects between different estimation branches for binocular, monocular, and non-visible cases. When both eyes are clearly visible, the system employs Gaze360 exclusively, as it provides the highest estimation accuracy; frames in which Gaze360 fails to return a valid output are discarded to ensure prediction quality. In cases where only one eye is visible, the reliability of direct gaze estimation decreases, and the system therefore switches to an indirect, head-pose-based approach using 6DRepNet. Frames for which 6DRepNet does not generate a valid output are similarly excluded. When neither eye nor even skin regions are detectable, we observe that gaze estimation becomes highly unreliable, and such frames are immediately discarded without attempting prediction. Building upon these visibility-aware branches, we implemented and evaluated a complete inference system, confirming that it achieves flexible and robust gaze estimation in complex CCTV environments by adapting to frame-level visibility conditions.

3.2 RoI Setting

The RoIs defined in this study represent semantically meaningful spatial units that reflect the actual structure of convenience-store shelving. Unlike prior gaze-estimation research that focuses on regressing a single gaze endpoint in image coordinates, our objective is not to determine “which pixel was observed,” but rather to precisely identify “which product category or which shelf region the customer attended to.” Accordingly, RoIs are defined not as pixel-level regions

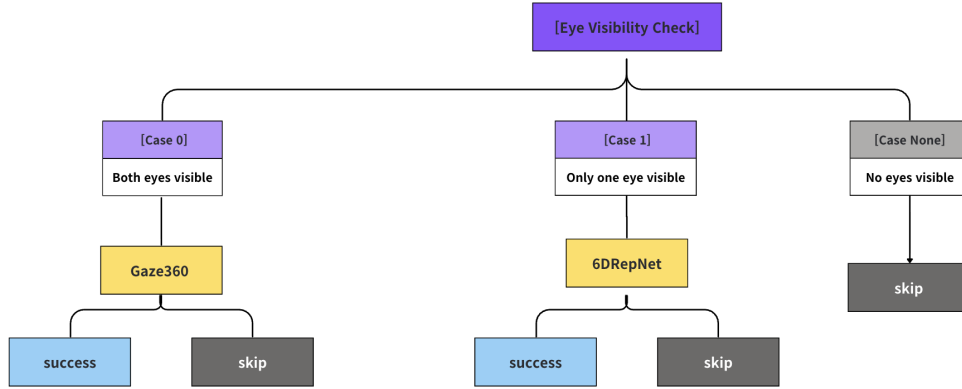


Figure 1: Overview of the proposed visibility-aware gaze estimation framework.

but as semantically grounded areas corresponding to product categories commonly used in marketing and retail analytics.

The experimental environment was constructed in a convenience store located on a university campus in Seoul. The store was first divided into six large Zones (A–F), and each Zone was further subdivided into fine-grained RoIs based on product type and shelving structure. Figure 2 illustrates the layout of these Zones and RoIs from a top-down perspective, serving as the reference coordinate system for all subsequent data processing and evaluation.

A key aspect of our RoI design is that regions are not restricted to simple rectangles; instead, the system supports point-based (poly) definitions, enabling RoIs of diverse shapes. Convenience-store shelves often form trapezoidal, tilted, or perspective-distorted shapes that cannot be accurately represented with drag-based rectangular annotations. In our framework, users specify four points sequentially to construct arbitrary quadrilateral RoIs—including trapezoids and skewed parallelograms—thereby allowing the actual geometric structure of the shelves to be mapped into digital coordinates with high fidelity.

3.3 Gaze + Head Estimation Integration

The gaze estimation module of this study is designed as a visibility-aware adaptive pipeline that addresses the constraints of overhead CCTV environments—low facial resolution, unstable eye visibility, and frequent lateral or rear appearances. For every frame, the system determines the eye visibility state (both, single, none) from the detected face and selectively applies either Gaze360 or 6DRepNet accordingly.

Following face detection via RetinaFace[15], the eye-

state classifier categorizes each frame into both, single, or none based on yaw–pitch conditions and additional quality cues. When both eyes are visible, Gaze360 is used as the primary gaze estimator. Gaze360 normalizes each facial crop to 224×224 , extracts frame-level features using a ResNet18 backbone, and processes the seven-frame sequence through a BiLSTM[16] to predict the yaw–pitch of the central frame. While this approach provides high accuracy under frontal or near-frontal conditions, the overhead CCTV perspective often yields blurry eye regions or unstable temporal sequences. To suppress unreliable outputs, the system delays inference when fewer than seven frames are available, applies a prediction interval to avoid excessive updates, and detects abnormal rapid changes (e.g., $\text{yaw} > 35$, $\text{pitch} > 25$) as outliers to be discarded.

In the single-eye state, the input quality for Gaze360 degrades significantly; thus, 6DRepNet is used as the primary head-pose estimator. 6DRepNet encodes a single facial image using a RepVGG[17] backbone, reconstructs a physically valid 3×3 rotation matrix via the Ortho6D representation, and converts it to Euler angles (pitch–yaw–roll). The recovered yaw–pitch values are then transformed into a 3D gaze-direction vector according to Equation (1), and this vector serves as an approximation of the user’s gaze in the single-eye condition.

$$\mathbf{g} = \begin{pmatrix} \cos(p) \sin(y) \\ \sin(p) \\ -\cos(p) \cos(y) \end{pmatrix} \quad (1)$$

When no eyes—and insufficient facial features overall—are visible, gaze estimation is not performed, and the frame is excluded from ROI mapping.



Figure 2: Visualization of three RoI examples defined through our point-based polygon annotation system.

3.4 RoI Mapping

Once the final gaze vector is obtained, the system proceeds to the RoI mapping stage, which determines the semantic region corresponding to the user’s gaze. Conventional endpoint-based approaches require estimating a distance scale k that represents the length of the gaze vector. However, in overhead fixed-view CCTV settings where depth information is unavailable, such scale estimation becomes highly unstable. The apparent size of a person in the frame does not reliably encode the actual camera–shelf geometry, meaning that even small errors in k can significantly distort the inferred gaze endpoint.

To address this limitation, the proposed method avoids explicit distance estimation and instead maps gaze to RoIs using the shortest-distance principle between the gaze ray and each RoI center. Specifically, the face center P_{face} and the normalized gaze direction vector $\hat{g}$ define a ray, and for each RoI center P_{roi} , the projection length and shortest distance are computed as in Equation (2):

$$t = (P_{\text{roi}} - P_{\text{face}}) \cdot \hat{g}, \quad d = \|P_{\text{roi}} - (P_{\text{face}} + t\hat{g})\|. \quad (2)$$

The RoI with the minimum distance d is selected as the final mapped region, enabling stable direction-based mapping without relying on depth estimation. Furthermore, to ensure temporal robustness, the system performs RoI mapping only when gaze estimation is sustained for at least 20 consecutive frames, thereby preventing spurious assignments caused by transient noise or momentary head movements.

4 Experiments

4.1 Experimental Settings

This experiment was designed to evaluate how reliably the proposed gaze–RoI mapping framework operates in a real convenience-store environment. Data collection was conducted inside a convenience store located within a university building in Seoul, where the spa-

tial layout—such as shelf positions, height and depth, aisle width, and product-category arrangement—was preserved to reflect actual commercial conditions. A key objective of the experiment was to explicitly include challenges that are rarely addressed in prior gaze estimation research, particularly those characteristic of overhead CCTV viewpoints: low facial resolution, ambiguous eye visibility, lack of frontal facial views, frequent exposure of the occipital region, and partial occlusions caused by other customers or store fixtures.

The recording setup used ceiling-mounted fixed CCTV cameras similar to those deployed in commercial convenience stores, with a mounting height of approximately 2.3–2.6 meters and a comparable field of view. Unlike the high-resolution frontal cameras commonly used in academic datasets, this configuration introduces inherent noise factors encountered in real-world retail environments. Under such an overhead viewpoint, the facial region occupies only a small pixel area, fine features such as eyes, nose, and mouth are sparsely represented, and side- or back-facing views naturally occur as customers browse shelves. By intentionally embracing these characteristics, the study aims to validate the practical feasibility of CCTV-based gaze analysis.

RoIs serve as the spatial reference framework for evaluation. As described in Section 3.2, the entire store is partitioned into six major zones (Zones A–F), each of which is further subdivided into semantic RoIs (A1, A2, ..., F2, etc.) corresponding to actual product categories and shelf units. Since the goal of gaze analysis is not precise coordinate regression but semantic determination—i.e., which shelf the customer was looking at—all evaluations focus on the accuracy of the final RoI assignments.

The environment also includes both single-person and multi-person scenes, with up to three individuals appearing simultaneously. This setup enables assessment not only of raw gaze estimation accuracy but also of the robustness of tracking (ID consistency), geometry-based corrections, and selective correction mechanisms under realistic operational conditions.

In summary, the experimental environment was designed to explicitly model the constraints unique to real convenience stores—constraints not adequately represented in conventional frontal-camera datasets—and to provide a dataset suitable for evaluating the proposed framework under near-real-world conditions. This setup offers essential evidence for determining whether the proposed approach can be effectively deployed in practical commercial environments.

4.2 Scenario Design and Data Collection

The dataset constructed for this study consists of approximately 40 video clips (each 10–20 seconds in length), with the same three actors participating across all scenarios. All recordings were conducted in an actual convenience store environment, where the camera was installed at a height and angle similar to commercial overhead CCTV systems. This setup preserves the unique constraints of real-world retail surveillance, including low facial resolution, frequent lateral and rear views, and repeated partial occlusions.

The single-person scenarios include roughly 25 video clips. In each clip, the actor was instructed to gaze at products located within predetermined RoIs, with the frequency and pattern of RoI transitions systematically varied across scenarios. Some videos contain no RoI transitions, wherein the actor continuously fixates on a single region to assess the stability of RoI classification in static gaze conditions. Other clips require the actor to sequentially inspect different RoIs, thereby simulating dynamic shifts of visual attention. In select scenarios, the actor moves between Zones while searching within RoIs, with transitions structured to reflect natural changes between product categories. This design aims to evaluate how reliably the model captures semantically meaningful shifts in consumer attention rather than merely physical movement.

The multi-person scenarios contain approximately 15 video clips and simulate situations in which two actors appear simultaneously, after which one exits the scene and another enters—an in-out replacement pattern. This mimics real convenience-store aisles where one customer leaves and another approaches the same shelf, posing a common challenge for person tracking in overhead CCTV. Each actor is assigned a role such as “in,” “out,” or “stay,” depending on their entrance and exit timing, providing controlled conditions for analyzing ID-maintenance failures. Scenes were designed to include crossing paths in narrow aisles and instances where actors stand in front of the same RoI, naturally producing occlusions or partial visibility of the face. RoI transition frequencies were independently assigned to each actor so that, for example, one actor might remain in the same RoI while another transitions two or

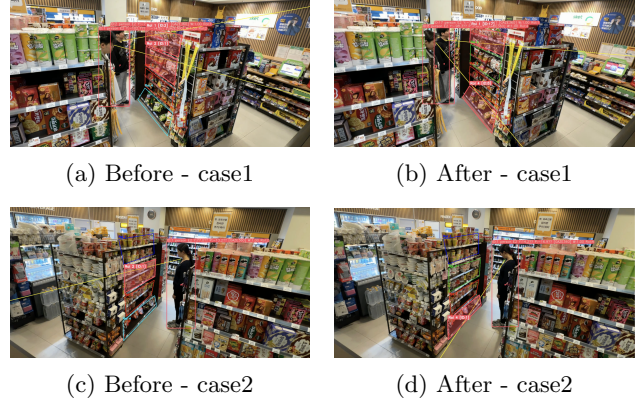


Figure 3: Examples of RoI mapping results produced by our point-based polygon annotation system, illustrating stable detection and assignment in both single- and multi-person scenarios.

three times, creating diverse gaze-shift patterns within a single scene. These scenarios provide essential evidence for evaluating whether the proposed selective correction mechanism remains stable under complex gaze dynamics and irregular occlusion conditions.

The overall dataset was constructed to avoid bias toward specific viewpoints or RoIs. Left- and right-leaning camera perspectives were balanced, and within each Zone, RoIs corresponding to different shelf layers were evenly represented. This allows independent analysis of how physical distance, height, and camera angle influence gaze estimation accuracy. RoI transition frequencies were also evenly distributed across the dataset, ranging from static fixation (0 transitions) to multi-step transitions (up to 3), enabling a broad evaluation of model robustness.

The labeling procedure followed predefined, scenario-specific ground-truth sequences of gaze-to-RoI transitions, and each final RoI journey was recorded as a comma-separated sequence based on the points at which RoI changes occurred.

4.3 Experimental Results

We evaluated the proposed selective-correction framework using 25 RoI-transition samples from the single-person dataset. The evaluation metric is the *Set-Matching Accuracy*, which considers a prediction correct if the predicted RoI set P_t contains the ground-truth RoI GT_t . Formally:

$$Acc_{Set-Matching} = \frac{1}{N} \sum_{t=1}^N \mathbb{I}(GT_t \cap P_t \neq \emptyset).$$

Under this evaluation, 21 out of the 25 samples were

correctly matched, yielding a Set-Matching Accuracy of 0.84. Performance remained consistently high across different RoI-transition complexities: 0.83 for sequences with 0, 1, and 2 transitions, and 0.86 for the most complex 3-transition cases. Participant-wise evaluation also demonstrated strong generalization, with accuracies of 0.88 for Hyunsik, 0.80 for Junhyeok, and 0.84 for Seoyoung. These results indicate that the selective-correction mechanism effectively stabilizes noisy Gaze360 outputs, while geometry-based cues and temporal smoothing contribute to maintaining temporal consistency and spatial robustness even under challenging movement and occlusion conditions. As shown in Figure 3, the point-based polygon RoI definitions used in this study provide a consistent spatial structure that supports stable region identification and assignment in both single-person and multi-person scenarios.

# RoI Transitions	Samples	Set-Matching
0 transitions	6	0.83
1 transition	6	0.83
2 transitions	6	0.83
3 transitions	7	0.86
Overall	25	0.84

Table 1: Accuracy by Number of RoI Transitions

5 Conclusion

This study proposes a selective-correction-based gaze estimation framework designed to perform RoI-level customer gaze and behavior analysis using only overhead CCTV footage. The framework employs Gaze360 as the primary estimator, while selectively incorporating 6DRepNet-based head-pose information in frames with low prediction confidence, thereby effectively mitigating structural challenges inherent to overhead CCTV environments such as low facial resolution, non-frontal viewpoints, and partial occlusions. In addition, a direction-distance-based RoI mapping approach is adopted to enable shelf- or product-category-level gaze interpretation required in real retail environments. To evaluate the proposed method, we constructed a dataset based on an actual convenience-store layout, covering diverse behavioral and viewing conditions. Experimental results show that the framework achieves stable RoI-level tracking performance across various gaze-transition patterns, demonstrating that it can recover gaze-journey information—such as exploration and switching behaviors—that cannot be captured using purchase-log-based analytics alone.

References

- [1] Clement, J. (2007). Visual Influence on In-Store Buying Decisions: An Eye-Track Experiment on the Visual Influence of Packaging Design. *Journal of Marketing Management*, 23, 917–928.
- [2] Gidlöf, K., Anikin, A., Lingonblad, M., Wallin, A. (2017). Looking Is Buying: How Visual Attention and Choice Are Affected by Consumer Preferences and Properties of the Supermarket Shelf. *Appetite*, 116, 29–38.
- [3] Kellnhofer, P., Recasens, A., Stent, S., Mëch, R., Matusik, W. (2019). Gaze360: Physically Unconstrained Gaze Estimation in the Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [4] Hempel, T., Abdelrahman, A. A., Al-Hamadi, A. (2022). 6D Rotation Representation for Unconstrained Head Pose Estimation. In *IEEE International Conference on Image Processing (ICIP)*.
- [5] Zhang, X., Sugano, Y., Fritz, M., Bulling, A. (2017). MPIIGaze: Real-world Dataset and Deep Appearance-based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1), 162–175.
- [6] Funes Mora, K. A., Monay, F., Odobez, J.-M. (2014). EyeDiap: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-D Cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*.
- [7] Doosti, B., et al. (2020). Hope-Net: A Graph-based Model for Hand-Object Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [8] Ghosh, S., Hayat, M., Dhall, A., Knibbe, J. (2021). MTGLS: Multi-Task Gaze Estimation with Limited Supervision. *arXiv preprint arXiv:2110.12100*.
- [9] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779–788).
- [10] Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems.
- [11] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S. (2019). Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 658–666).

- [12] Wojke, N., Bewley, A., Paulus, D. (2017). Simple Online and Realtime Tracking with a Deep Association Metric. In *IEEE International Conference on Image Processing (ICIP)* (pp. 3645–3649).
- [13] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., ... Wang, X. (2022). ByteTrack: Multi-object Tracking by Associating Every Detection Box. In *European Conference on Computer Vision*.
- [14] Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K. (2023). Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9686–9696).
- [15] Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S. (2020). RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [16] Huang, Z., Xu, W., Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint arXiv:1508.01991*.
- [17] Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J. (2021). RepVGG: Making VGG-Style ConvNets Great Again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.