Zone-Based Joint Camera-Kiosk Analytics for Smart Retail Insights

1st Hassan Moin
CSEE. University of Essex
Colchester, United Kingdom
hm22803@essex.ac.uk

2nd Vahid Abolghasemi
CSEE. University of Essex
Colchester, United Kingdom
v.abolghasemi@essex.ac.uk

3rd Hossein Anisi
CSEE. University of Essex
Colchester, United Kingdom
m.anisi@essex.ac.uk

***Abstract*—This study presents a novel system that tightly integrates zone-based camera analytics with kiosk interaction logs to provide rich, multidimensional insights into customer behaviour in retail environments, particularly within smart retail settings in smart cities. Our approach enables retail administrators to define arbitrary, polygonal zones on a live floorplan overlay for each camera. In the proposed pipeline, video streams are processed at 30 fps using YOLOv8 for person detection and SORT for tracking, with entry and exit events logged per zone to compute footfall, dwell time, and bounce rates. Zones are classified as *kiosk zones* or *custom zones* based on metadata assigned during zone creation in the dashboard, where each polygon is manually linked to a kiosk identifier or defined as a general area of interest. By correlating zone presence with kiosk session data via timestamp overlap, we derive a robust kiosk engagement metric. We deployed the system in a mid-size retail store over four weeks, demonstrating improvements in the granularity and actionability of behavioural analytics and revealing optimal kiosk placements and display configurations in a smart retail context.**

***Index Terms*—Smart Retail, Edge Computing, IoT Analytics, Computer Vision, Behavioural Analytics, Data Fusion, Kiosk Zone, Engagement Rate, Camera Kiosk Integration**

I. Introduction

Today, data-driven decisions are essential for retailers aiming to improve customer engagement, optimise store layouts, and increase return on investment. In the context of smart cities, retail environments represent key components of urban infrastructure, where intelligent systems can enhance both economic activity and citizen experience. Many retailers use analytics platforms to understand shopper behaviour, but most applications only provide basic metrics such as footfall counts, point-of-sale data, or transaction summaries.

These metrics do not capture where customers spend time, which displays attract attention, or how they interact with kiosks. They also do not support rapid A/B testing or continuous optimisation, leading to inventory and staffing decisions often based on guesswork. As smart retail becomes part of smart city ecosystems, there is a growing need for detailed, real-time behavioural analytics across interconnected environments.

Recent advances in computer vision and real-time object detection enable continuous individual detection while preserving privacy, enabling deeper analysis of in-store behaviour. Algorithms such as YOLOv8 and its successors [1]–[3] can detect multiple people efficiently in crowded scenes, with inference times under 50 milliseconds at resolutions suitable for retail scenarios. Combined with lightweight tracking algorithms such as SORT (Simple Online and Realtime Tracking) [4] and multi-camera transformers [5], these methods enable unique tracking of customer paths without additional hardware. With GPU acceleration and efficient deployment, such models can process video streams at 25–30 frames per second across multiple cameras, enabling real-time analytics at scale. However, useful insights require structured behavioural interpretation beyond detection. For example, distinguishing between customers passing a kiosk and those actively engaging with it is critical for evaluating display effectiveness, staffing, and campaign Region of Interest (ROI).

Existing systems either (i) rely solely on vision-based analytics without incorporating interaction context, or (ii) use indirect proximity signals such as Bluetooth or WiFi to infer kiosk engagement. These approaches fail to capture the direct relationship between physical presence and actual user interaction. To the best of our knowledge, no prior work integrates spatial zone-based tracking with kiosk session logs in a unified, real-time framework, thereby limiting the ability to derive accurate, actionable behavioural insights in retail environments.

By integrating camera analytics with kiosk data, retailers can obtain a more complete view of shopper behaviour. We propose a zone-based analytics system that integrates vision-based detection and tracking with spatial zone management and real-time kiosk interaction logging. Each camera feed is processed using YOLOv8 for detection and SORT for tracking. At the same time, tracked individuals are mapped to polygonal zones defined via a dashboard and stored in a PostgreSQL/PostGIS database. Zone presence events are correlated with kiosk sessions using synchronised timestamps and a threshold-based temporal overlap rule, enabling precise computation of dwell time, bounce rate, and engagement. Unlike conventional proximity or heatmap methods [6], the proposed system supports dynamically editable zones, scalable spatial queries, and fusion of heterogeneous data sources. This provides a real-time foundation for behavioural analytics in retail environments and aligns with smart computing requirements in smart cities. The main contributions of this study are as follows:

- A zone-based spatial analytics framework with dynamically configurable polygonal regions for fine-grained behavioural analysis.
- A temporal-overlap method for accurate kiosk engagement detection by correlating zone presence with interaction logs.
- A unified integration of vision-based tracking and kiosk interaction data without requiring additional hardware.
- A real-world deployment and evaluation over four weeks in a retail environment with multiple cameras and kiosks.

The rest of this paper is organised as follows. Section II reviews existing approaches in vision-based retail analytics and kiosk interaction studies. Section III describes the proposed system architecture. Section IV details the implementation aspects. Section V presents the experimental evaluation. Finally, Section VI concludes the paper and outlines future directions.

II. Related Work

Retail analytics has seen growing adoption of computer vision-based systems in recent years, particularly as retail environments become increasingly integrated into smart city infrastructure. Heatmaps, path tracking, and dwell-time estimation are among the most common applications. In [6], a deep learning-based person detector is introduced to generate heatmaps highlighting popular in-store regions. However, their work lacked support for dynamic or arbitrary zone definitions, which limits adaptability across different store layouts. Rubing Li et al. [7] proposed a method of estimating dwell time using background subtraction and trajectory clustering, applied to fixed displays. Although effective for static regions, this approach cannot account for dynamic promotional areas or integrate with digital interaction points such as kiosks, which are increasingly common in smart retail environments.

On the kiosk side, engagement has typically been inferred from indirect signals, such as Bluetooth or Wi-Fi proximity. James M et al. [8] used Bluetooth data to estimate customer attention near kiosks. However, this method required additional hardware and could not distinguish between active use and passive presence, limiting its effectiveness in scalable smart-city deployments.

More recent deep-learning frameworks, such as Valério et al. [9], adopt convolutional architectures for people counting and hotspot detection in retail spaces using monocular RGB cameras. Their pipeline focuses on region-level activity analysis but does not incorporate temporal reasoning or digital-interaction data. Similarly, Paolanti et al. [10] proposed a multi-camera RGB-D setup that integrates depth information to model shopper trajectories and detect shelf interactions. While effective for spatial behaviour analysis, this approach depends on calibrated 3D sensors and does not fuse vision data with kiosk or transaction logs, limiting its applicability in large-scale, distributed environments.

Other approaches, such as RFID tracking [11] and customer surveys [12], offer useful insights but are often costly, privacy-invasive, or unsuitable for real-time adaptation. Vision-based systems provide a passive and scalable alternative, particularly when combined with lightweight tracking algorithms such as SORT [4], making them more suitable for deployment in smart city contexts.

Recent breakthroughs in object detection, particularly with YOLOv8 [1], enable fast, accurate detection of individuals in crowded or complex environments. Combined with trajectory-tracking techniques and spatial reasoning tools, such as point-in-polygon testing, modern systems can provide robust behavioural analytics. However, few existing systems integrate these capabilities with digital kiosk interaction logs to provide a full behavioural loop from presence to interaction, which is critical for comprehensive smart retail analytics.

The release of YOLOv10 and related one-stage detectors has further pushed inference speed and accuracy beyond earlier versions, achieving state-of-the-art results suitable for crowded shopping environments [2], [3]. These detectors, combined with transformer-based multi-camera tracking approaches such as MCTR [5], allow consistent identity preservation across overlapping camera views, a key requirement for large-format stores where customer journeys span multiple aisles and entry points. Such advances address the long-standing challenge of fragmented trajectories and open the door to seamless store-wide analytics across distributed retail systems.

Beyond pure performance gains, several authors have highlighted the importance of scalable, ethical, and privacy-preserving deployment. Adanyin et al. [13] emphasise the need for transparent data governance and fairness in AI-powered retail systems, recommending practices such as edge-side anonymisation and explicit customer disclosure. These considerations complement technical innovation by ensuring compliance with emerging regulations, such as GDPR, and by fostering consumer trust—an increasingly critical factor as video analytics and interactive kiosks converge in smart city environments.

Broader surveys of video big-data analytics also reflect this trend toward holistic solutions. Thi-Thu-Trang Do et al. [14] review architectures for large-scale video processing and discuss how GPU-accelerated pipelines and cloud-edge hybrids can sustain real-time performance across dozens of simultaneous camera streams, directly supporting the type of high-throughput, distributed environment targeted by our framework.

Our work builds on these advances and introduces a tightly integrated pipeline that links vision-based detection, spatial zone tracking, and kiosk session logs within a unified smart-computing framework.

III. System Architecture

We consider a retail indoor environment equipped with multiple ceiling- or wall-mounted cameras and interactive

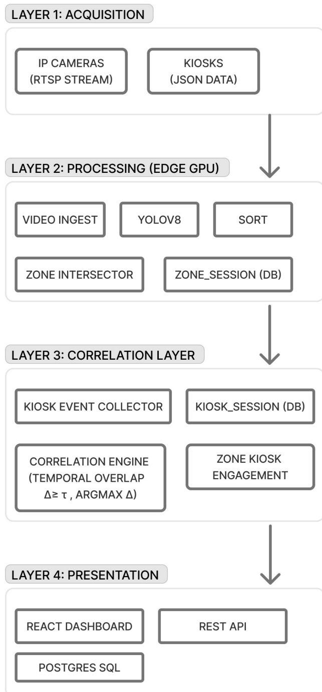

Fig. 1. Overall architecture of the proposed zone–kiosk analytics framework.

kiosks. Each camera observes a portion of the store, while kiosks are placed in predefined regions such as entrances or promotional areas. The objective is to detect and track customers, map their movements to spatial zones, and analyse behaviour by correlating presence with kiosk interaction logs.

The proposed system follows a modular, four-layer architecture integrating vision-based detection, spatial reasoning, and data correlation within a unified analytics pipeline. As illustrated in Fig. 1, the architecture comprises the following layers, which are described below:

A. Data Acquisition Layer

This layer captures behavioural signals from video and kiosk streams. It includes Video Input, where overhead and angled IP cameras stream H.265 video at 1280×720 resolution and 25 fps, balancing efficiency and detection accuracy (Fig. 2); and Kiosk Data, where touchscreen kiosks log session activity and transmit JSON-encoded interaction events, including kiosk ID, session ID, timestamps, and metadata.

All inputs are synchronised via an NTP (Network Time Protocol) based system clock to ensure temporal consistency.

B. Processing Layer

This layer performs detection, tracking, and spatial reasoning. It includes Object Detection, where YOLOv8 [1] detects persons per frame with a confidence threshold of 0.8; Tracking, where SORT [4] assigns consistent `track_ids` using Kalman filtering and Hungarian matching; and Zone Intersection, where bounding box centroids are evaluated using a point-in-polygon method accelerated via NumPy vectorisation and PostGIS indexing.

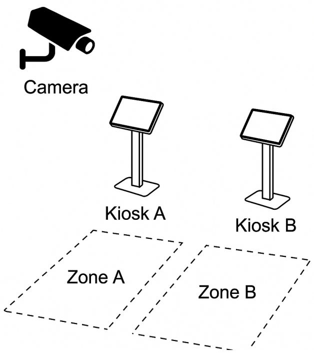

Fig. 2. Camera, kiosk, and zone setup.

Events are stored in the `zone_session` table with track ID, zone ID, and timestamps, with sessions terminating if tracking is lost for more than 30 frames.

C. Correlation Layer

This module links zone sessions with kiosk sessions to determine *engagement*.

A zone session is $Z = (z, t, t^{\text{in}}, t^{\text{out}})$, and a kiosk session is $K = (k, s, t^{\text{start}}, t^{\text{end}})$.

The temporal overlap is defined as:

$$\Delta(Z, K) = \max\Big(0,\ \min\big(t^{\text{out}}, t^{\text{end}}\big) - \max\big(t^{\text{in}}, t^{\text{start}}\big)\Big) \quad (1)$$

A session is classified as *engaged* if $\Delta(Z, K) \geq \tau$ for at least one kiosk session, where τ is a threshold (e.g., 3 s); otherwise, it is a *pass-by*.

Algorithm 1 Zone-Kiosk Engagement Correlation

Require: Zone session $Z = (z, t, t^{\text{in}}, t^{\text{out}})$; set of kiosk sessions $\mathcal{K}$ in kiosk-zones; threshold τ
engaged $\leftarrow$ FALSE; $\Delta_{\max} \leftarrow 0$; $K^\star \leftarrow$ NIL
for all $K = (k, s, t^{\text{start}}, t^{\text{end}}) \in \mathcal{K}$ **do**
  $\Delta \leftarrow \max\big(0,\ \min(t^{\text{out}}, t^{\text{end}}) - \max(t^{\text{in}}, t^{\text{start}})\big)$
  if $\Delta \geq \tau$ **then**
    engaged $\leftarrow$ TRUE
    if $\Delta > \Delta_{\max}$ **then**
      $\Delta_{\max} \leftarrow \Delta$; $K^\star \leftarrow K$
    end if
  end if
end for
if engaged **then**
  label Z as `engaged` and optionally associate $K^\star$
else
  label Z as `pass-by.`
end if
Persist result in `zone_kiosk_engagement` (zone_id, track_id, engaged, $\Delta_{\max}$, $K^\star$)

The procedure in Algorithm 1 evaluates the intersection length Δ between the time interval of a zone session and each kiosk session in the relevant kiosk-zone(s). If any overlap meets or exceeds τ, the zone session is marked *engaged*.

When multiple kiosk sessions overlap, the algorithm optionally assigns the session with maximum overlap $K^\star$ (argmax of Δ) to avoid double counting.

D. Presentation Layer

The dashboard provides both real-time analytics and historical insights through an interactive interface. It includes a Zone Editor, a polygon-based tool for defining spatial regions with associated metadata; an Analytics Dashboard, which displays key metrics such as footfall, dwell time, bounce rate, and engagement rate; and a Session View, offering filterable session logs with peak time and behavioural insights.

E. Technical Stack Overview

The framework is implemented as a modular cloud-based system. The backend is developed in Python, integrating YOLOv8 and SORT within a Django architecture, with PostgreSQL/PostGIS handling spatial-temporal data processing. The frontend uses React.js and Node.js to provide real-time visualisation through APIs and WebSockets. Deployment is containerised using Docker on AWS, with EC2 hosting services behind Nginx, S3 for storage, and CloudWatch for monitoring.

IV. Implementation Details

The system is implemented as a modular pipeline supporting real-time processing, extensibility, and high availability. Key components include zone management, video inference, session tracking, kiosk correlation, and analytics computation.

A. Zone Definition

Zones are defined through a web-based React.js dashboard using the React Bits library. Retail administrators create polygonal regions on live or static camera feeds, supporting multiple zones per view. Each zone includes: `zone_id`, `zone_type`, linked `camera_id` and GeoJSON polygon. Zones are stored in PostgreSQL with PostGIS support and indexed using GiST for efficient spatial queries.

B. Video Inference and Tracking

RTSP (Real-Time Streaming Protocol) streams are ingested via OpenCV, resized to 1280×720, and processed at 30 fps. YOLOv8 performs person detection with a confidence threshold of 0.8 and NMS (Non-Maximum Suppression) of 0.4, balancing recall and precision. Inference runs on an AWS-hosted NVIDIA Tesla T4 GPU with CUDA acceleration, enabling real-time processing across multiple streams. SORT is used for tracking, associating detections via Kalman filtering and Hungarian matching to maintain consistent `track_ids`.

C. Zone Session Logging

Each tracked person's centroid is evaluated against zone polygons using a ray-casting algorithm. On zone entry:

- A record is created in `zone_session`
- `enter_time` is logged using an NTP-synchronised clock

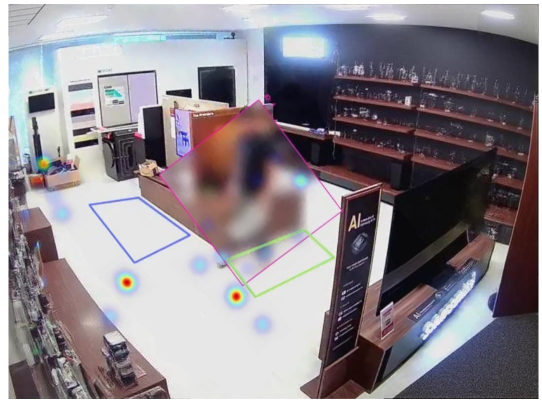

Fig. 3. Retail pilot store: camera views and kiosk-zone placement.

- On exit, `exit_time` and `duration_seconds` are computed

Sessions include `session_id`, `track_id`, `zone_id`, `enter_time`, and `exit_time`. A single user may trigger multiple overlapping zones.

D. Kiosk Integration

Kiosks transmit interaction data via API (Application Programming Interface). Each session includes: `kiosk_id`, `session_id`, `start_time` and `end_time`

Sessions are stored in the `kiosk_session` table and joined with `zone_session` for kiosk zones. Overlapping timestamps indicate engagement; otherwise, the event is classified as a pass-by.

E. Analytics Calculation

Metrics are computed via SQL views refreshed in near real time. *Footfall* is derived from distinct zone-entry events. *Dwell time* is the mean duration between entry and exit per visitor. *Bounce rate* represents sessions shorter than 5 s. *Engagement rate* is defined as the ratio of engaged to total sessions.

V. Results and Discussion

A. Pilot Setup

The system was deployed in a $200\,\mathrm{m}^2$ apparel retail store with two IP cameras and two touchscreen kiosks as in Fig. 3. Each kiosk has an associated *kiosk zone*. Camera 1 covers the entrance and Kiosk A, while Camera 2 covers a promotional table and Kiosk B. All devices are synchronised via an NTP-based clock to ensure timestamp alignment.

B. Metric Definitions

We evaluate (i) zone counting accuracy using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), (ii) dwell estimation error using MAE and median absolute error (median AE), (iii) kiosk engagement rate (Sec. III-C), and (iv) goal conversion during engaged sessions.

For intervals $i{=}1,\ldots,N$:

TABLE I
KIOSK METRICS OVER FOUR WEEKS (95% CI FOR ENGAGEMENT).

Kiosk	Entries	Avg dwell (s)	Eng. (95% CI)	Eng. (n)
A	1200	45	12.0 [10.2, 13.8]	144
B	900	40	18.0 [15.5, 20.5]	162

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|c_i^{\text{sys}} - c_i^{\text{gt}}\right| \tag{2}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(c_i^{\text{sys}} - c_i^{\text{gt}}\right)^2} \tag{3}$$

For dwell estimation:

$$\text{MAE}_{\text{dwell}} = \frac{1}{M}\sum_{j=1}^{M}\left|d_j^{\text{sys}} - d_j^{\text{gt}}\right| \tag{4}$$

Engagement and conversion are reported with Wald 95% confidence intervals.

C. Kiosk-Level Summary (4 Weeks)

Table I summarises aggregated zone analytics for the two kiosks over four weeks, including entries (zone sessions), mean dwell, engagement with 95% CI, and engaged counts.

To further evaluate the difference between kiosks, we performed a two-proportion z-test comparing engagement rates ($p_A = 0.12$, $p_B = 0.18$). The result ($z = 3.86$, $p < 0.001$) indicates that kiosk B achieves significantly higher engagement than kiosk A. This difference is consistent with placement effects, as kiosk B is located near the promotional table while kiosk A is positioned near the entrance. Following repositioning, kiosk A showed an approximate 5% relative increase in engagement, demonstrating the practical actionability of the proposed analytics framework for optimising retail layouts.

D. Counting Accuracy vs Manual Tallies

Eight 10-minute intervals were sampled (four per kiosk). Table II reports results, with **MAE** $= 2.13$ and **RMSE** $= 2.21$. These results show high agreement with manual counts, corresponding to $< 5\%$ error in typical footfall volumes. The low RMSE indicates stable performance across varying conditions, supporting reliable zone-level analytics suitable for smart retail environments without additional sensing infrastructure.

E. Dwell Estimation Error

Dwell duration accuracy was evaluated across 12 manually timed sessions (six per kiosk). The system achieved $\textbf{MAE}_{\textbf{dwell}} = 2.42\,\text{s}$ and **Median AE** $= 2.0\,\text{s}$, with all errors below 5 s.

This demonstrates near-human precision in estimating behavioural duration. Minor underestimation in longer sessions arises from centroid-based boundary effects when users linger near zone edges. Such accuracy is sufficient for practical

TABLE II
COUNTING ACCURACY (MANUAL VS. SYSTEM) ON INTERVALS.

Interval	Zone	Manual	System	Abs. err.
I_A1	Kiosk A	52	49	3
I_A2	Kiosk A	40	42	2
I_A3	Kiosk A	65	62	3
I_A4	Kiosk A	55	54	1
I_B1	Kiosk B	45	47	2
I_B2	Kiosk B	58	60	2
I_B3	Kiosk B	70	68	2
I_B4	Kiosk B	62	64	2
MAE = 2.13 RMSE = 2.21				

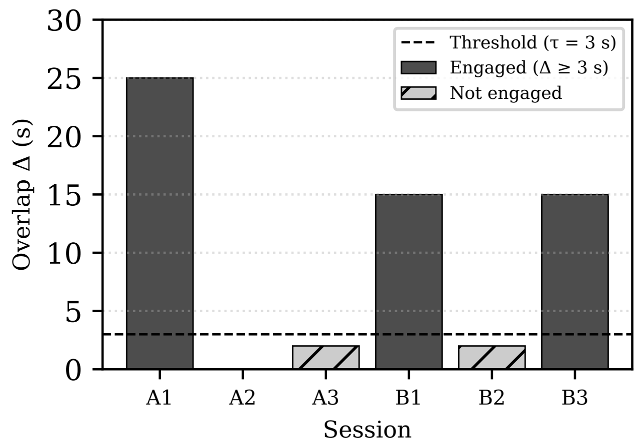

Fig. 4. Overlap durations (Δ) between zone and kiosk sessions. Values above $\tau = 3\,\text{s}$ indicate engagement.

analytics in smart retail systems, where relative dwell comparisons across zones and layouts are more critical than subsecond precision.

F. Engagement Correlation Analysis

Fig. 4 shows overlap durations (τ=3 s) between zone and kiosk sessions. Bars above the threshold indicate engagement, clearly separating pass-bys from interactions. Engaged sessions typically range between 10–25 s, while non-engaged cases show negligible overlap. Across 60 manually verified pairs, the method achieved precision $= 0.93$ and recall $= 0.88$, validating the temporal-overlap rule as a reliable mechanism for real-time behavioural analytics.

G. Conversion / Goal Completion

We define a goal as an *information request* (e.g., product info tap/QR scan) during an engaged kiosk session. Kiosk A: 36 completions out of 144 engaged (**25.0%**, 95% CI [17.9, 32.1]). Kiosk B: 49/162 (**30.2%**, 95% CI [23.2, 37.3]). A two-proportion z-test gives $z{=}-1.02$ (p=0.31), i.e., no significant difference in conversion given engagement. End-to-end zone$\rightarrow$goal rates are 3.0% (A) and 5.4% (B).

H. Comparison with Prior Work

Tables III and IV present quantitative and qualitative comparisons between the proposed framework and representative studies in vision-based retail analytics. Since the literature

TABLE III
QUANTITATIVE COMPARISON WITH RELATED WORK.

Method	Count MAE	Dwell MAE (s)	Int./Eng. Acc.
Ours	**2.13**	**2.42**	**0.93**
RetailNet [9]	≈ 1.2	–	–
Paolanti et al. [10]	0.5% err.	–	0.926

TABLE IV
CAPABILITY COMPARISON WITH RELATED WORK.

Capability	**Ours**	Zhang [6]	Rubing Li [7]
Arbitrary polygonal zones	✓	×	×
Dwell in dynamic promo areas	✓	×	Limited
Kiosk integration via logs	✓	×	×
Extra hardware (beacons/RFID)	×	×	×
Temporal-overlap engagement (τ)	✓	×	×
Real-time store-scale analytics	✓	Mixed	Mixed

reports diverse metrics, we focus on directly comparable indicators, namely *counting error* (MAE), *dwell-time estimation error* (MAE in seconds), and *interaction or engagement accuracy*. Where prior studies reported accuracy rather than absolute error, values were converted to equivalent error percentages for consistency. Valério et al. [9] represents a deep-learning-based people-counting approach deployed in a real retail store, achieving a mean absolute error of approximately 1.2 persons per interval but without addressing dwell time or digital engagement, focusing instead on coarse heatmap generation. Paolanti et al. [10] employed an RGB-D camera network for multi-camera tracking and shelf-interaction recognition, achieving around 99.5% counting accuracy (0.5% error) and 92.6% interaction accuracy; however, it required depth sensors, controlled calibration, and did not estimate dwell durations or integrate kiosk interaction data. Our method extends prior approaches by integrating vision-based tracking with kiosk interaction logs within a unified zone-based framework. Using only RGB cameras, it achieves counting accuracy (MAE 2.13) and dwell-time MAE of 2.42 s, while delivering 93% engagement accuracy through timestamp-based correlation. This enables reliable measurement of true customer engagement, reducing false positives common in vision-only or proximity-based systems. The approach combines polygonal zones in PostgreSQL/PostGIS with a temporal-overlap rule (τ=3 s) and synchronised kiosk data to provide actionable behavioural insights without additional hardware. Compared to prior work [6]–[8], it supports arbitrary zones, direct integration with kiosk logs, and threshold-based engagement detection.

VI. CONCLUSION

This work presents a zone-based integration of camera analytics and kiosk interaction logs to enhance behavioural insights in retail environments. The proposed system combines real-time person detection and tracking with flexible polygonal zone definitions. It correlates these with kiosk sessions through a temporal-overlap rule, enabling the transition from coarse footfall monitoring to actionable metrics such as dwell time, bounce rate, engagement, and goal conversion. A four-week deployment with two cameras and two kiosks demonstrates the effectiveness and practical utility of the framework, showing how integrating computer vision with interactive retail systems can transform passive monitoring into a proactive decision-support tool for optimising store layouts and customer engagement. More broadly, the approach contributes to smart retail within smart city ecosystems by enabling data-driven commerce, adaptive retail spaces, and improved customer experiences, while addressing key limitations of existing retail analytics methods.

Future work will focus on extending the framework with predictive modelling to forecast shopper behaviour and optimise store layouts proactively. The system can be enhanced by integrating point-of-sale and inventory data to enable full end-to-end conversion analysis. Additionally, incorporating multi-camera re-identification will improve tracking across larger retail environments. Further research will also explore anomaly detection for applications such as loss prevention and queue management. Finally, large-scale deployments across diverse store formats will be conducted to validate scalability and generalisability.

REFERENCES

[1] N. M. Muriyah, J. H. Sim, and A. Yulianto, “Evaluating yolov5 and yolov8: advancements in human detection,” *Journal of Information Systems and Informatics*, vol. 6, no. 4, pp. 2999–3015, 2024.

[2] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, “Yolov10: real-time end-to-end object detection,” in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS ’24. Red Hook, NY, USA: Curran Associates Inc., 2024.

[3] C.-Y. Wang and H.-y. Liao, “Yolov1 to yolov10: The fastest and most accurate real-time object detection systems,” *APSIPA Transactions on Signal and Information Processing*, vol. 13, 01 2024.

[4] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” 09 2016, pp. 3464–3468.

[5] A. Niculescu-Mizil, D. Patel, and I. Melvin, “Mctr: Multi camera tracking transformer,” 08 2024.

[6] M. Şimşek and M. K. Tekbaş, “Heatmap creation with yolo-deep sort system customized for in-store customer behavior analysis,” *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering*, vol. 66, no. 1, pp. 118–131, 2024.

[7] R. Li, A. Ghose, K. Xu, and B. Li, “Predicting consumer in-store purchase using real-time retail video analytics,” *Available at SSRN 4513385*, 2023.

[8] J. M. Rehg, M. Loughlin, and K. Waters, “Vision for a smart kiosk,” in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1997, pp. 690–696.

[9] V. Nogueira, H. Oliveira, J. Silva, T. Vieira, and K. Oliveira, “Retailnet: A deep learning approach for people counting and hot spots detection in retail stores,” 07 2019.

[10] M. Paolanti, R. Pietrini, A. Mancini, and E. Frontoni, “Deep understanding of shopper behaviours and interactions using rgb-d vision,” *Machine Vision and Applications*, vol. 31, p. 66, 09 2020.

[11] N. Brégaint, “Leveraging rfid in retail: From stock management to customer experience enhancement,” 2025.

[12] R. Esmeli, M. Bader-El-Den, H. Abdullahi, and D. Henderson, “Implicit feedback awareness for session based recommendation in e-commerce,” *SN Computer Science*, vol. 4, no. 3, p. 320, 2023.

[13] A. Adanyin, “Ethical ai in retail: Consumer privacy and fairness,” *European Journal of Computer Science and Information Technology*, vol. 12, pp. 21–35, 07 2024.

[14] T.-T.-T. Do, T. Huynh Quyet, K. Kim, and V.-Q. Nguyen, “A survey on video big data analytics: Architecture, technologies, and open research challenges,” *Applied Sciences*, vol. 15, p. 8089, 07 2025.