Intelligent Computer Vision System for Customer Behavior Analysis in Jarir Bookstores

Kaust Academy AI Specialization Final Report

Hassan Mohammed Nasr - Team Leader
hassan.m.nasr@outlook.com
4410606@upm.edu.sa

Abderrahmene Mehenni
mehenniabdou25@gmail.com

Khalid Alkaabi
kamk3bi@gmail.com

Mentor: Dr. Muhammad Mubashar
19230664@brookes.ac.uk

August 27, 2025

Abstract

This project presents a computer vision–based approach for analyzing customer behavior in retail stores, using Jarir Bookstores as a case study. We developed an object detection and tracking system that identifies and follows customers across CCTV footage. From these tracks, we generate region-of-interest heatmaps that highlight areas of high customer activity. Using this algorithm, we calculated the precise dwell time for each individual within specific zones. This allowed us to classify each customer as either "dwelled"—meaning they stayed long enough to show genuine interest—or "passed by." This crucial classification forms the basis of our behavioral analysis, enabling a deeper understanding of in-store engagement patterns.

Contents

Abstract **1**

1 Introduction **3**
1.1 Group Members and Their Contributions 3
1.2 Problem Motivation and Real-World Impact 3
1.3 Limitations of the Problem 3
1.4 Project Objectives and Goals 3
1.5 Problem Statement 4
1.6 Summary of the proposed system 4

2 System Overview **4**

3 Methodology **5**
3.1 Tools, Libraries, and Frameworks 5
3.2 Dataset or Test Input Creation 5
3.3 Steps of system operation 5
3.4 Criteria Used to Evaluate Performance 7

4 Results Discussion **8**

5 Reflections on Challenges **9**

6 Conclusion and Future Work **10**

1 Introduction

1.1 Group Members and Their Contributions

- **Hassan Mohammed Nasr - Team Leader** – Developed and implemented all core components of the final system, including Object Detection (YOLO), DeepSORT Tracking, PostTracker (improvement on deepSORT), the Counting System, Dwell Time Calculation, and Customers classification and analysis. Finalized the final project report and designed the poster.
- **Abderrahmene Mehenni** – Experimented with different YOLO models on different datasets, and tried to classify employees. Wrote the draft for the final report.
- **Khalid Alkaabi** – Prepared one PowerPoint for Jarir's meeting.

1.2 Problem Motivation and Real-World Impact

Retailers face a significant challenge in understanding customer behavior within physical stores. Manual observation is time-consuming, subjective, and often unreliable. As a result, layout and product placement decisions lack data-driven insights. Meanwhile, valuable data continuously captured by in-store CCTV cameras remains underutilized. For Jarir Bookstores, this gap represents a missed opportunity to optimize store layouts, enhance customer experience, and ultimately increase sales.

1.3 Limitations of the Problem

- Accuracy can be affected by extreme crowd density, leading to occlusions.
- Similar appearances and clothing (e.g., abaya for women, thobe for men) may lead to ID switches.
- Differentiating between customers and employees, especially female employees, remains a challenge.

1.4 Project Objectives and Goals

The objective of this project is to analyze customer interactions by detecting and tracking people within Jarir stores. Through advanced detection and analytical techniques, the system aims to provide insights into customer behavior and engagement patterns.

- **Objective 1:** To accurately detect and track individuals within a given video frame.
- **Objective 2:** To implement a robust system for counting visitors within predefined zones.
- **Objective 3:** To calculate the dwell time of each individual in specific areas to measure engagement.
- **Objective 4:** To generate visual analytics, such as heatmaps, to illustrate high-traffic areas and customer flow.

1.5 Problem Statement

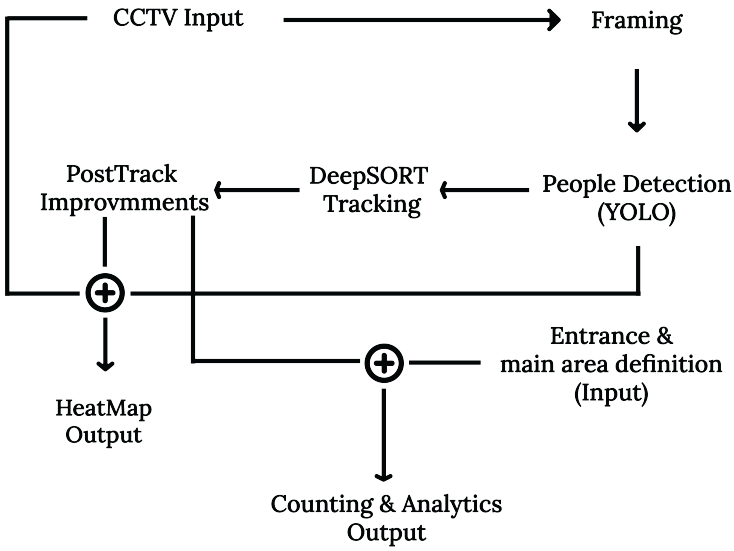
To design and implement an automated video analytics system that detects, tracks, and analyzes human movement in a Jarir Bookstore. The system will quantify visitor traffic, measure engagement through dwell time, differentiate between browsing customers and passersby, and visualize foot traffic patterns using heatmaps.

1.6 Summary of the proposed system

The proposed system is a multi-stage video processing pipeline. It begins by ingesting a video feed and using a YOLOv8 object detection model to identify people in each frame. The detected individuals are then passed to a DeepSORT tracker combined with a custom post-processing algorithm to assign and maintain unique IDs, even through temporary occlusions. The system then uses this tracking data to perform higher-level analysis, including counting individuals within user-defined polygonal zones, calculating their dwell time, and classifying their behavior. Finally, it generates comprehensive outputs, including an annotated video, CSV reports with dwell times and visitor analysis, and a cumulative heatmap image visualizing customer movement patterns over time.

2 System Overview

This section provides a high-level look at the components and logical flow of the analytics system.



The logical operation of the system is illustrated in the figure above. The process begins with raw video input from a CCTV source, which is broken down into individual frames. Each frame undergoes People Detection using the YOLO model. The detected individuals are then passed to the DeepSORT Tracking algorithm, which is enhanced with Post-Tracking Improvements to ensure ID consistency. This tracking data, combined with user-defined area inputs, feeds into two main output modules: the Heatmap Generation and the Counting & Analytics system, which produce the final data reports.

3 Methodology

3.1 Tools, Libraries, and Frameworks

- **Object Detection:** We utilized the YOLOv8 (large model version) implementation from the **Ultralytics** library for high-performance object detection.
- **Object Tracking:** The core tracking logic was built using **DeepSORT (deep-sort-realtime)**, while appearance features for re-identification were generated with the **OSNet** model via **Torchreid**.
- **Visualization:** OpenCV for video I/O, image manipulation, and drawing functions.
- **Data Handling:** NumPy for efficient numerical operations on arrays (e.g., bounding boxes, heatmap data).
- **Data Output:** Python's built-in csv module for exporting analysis results.
- **Defining Store Zones:** Image-map.net for extracting coordinate points for defining store zones.
- **Compressing Data:** Microsoft Clipchamp for combining video clips and reducing file size.
- **Framework:** The project's Python code was developed on the **Kaggle** platform

3.2 Dataset or Test Input Creation

For the development and validation of the analysis framework, the project relied on proprietary surveillance footage provided by Jarir Bookstore. The initial dataset was found unsuitable due to poor camera placement, which limited its analytical usefulness. As a result, Jarir supplied a new set of high-resolution CCTV recordings.

From this improved dataset, representative clips were selected. These clips were concatenated in chronological order to produce coherent video inputs for analysis. To balance efficiency with visual quality, the curated footage was downsampled and compressed. This prepared dataset served as the primary input for testing and refining the system.

3.3 Steps of system operation

1. **Initialization:** The system loads the YOLOv8 model, initializes the DeepSORT, PostTracker, and HeatmapGenerator instances, and sets up the CountingSystem for each defined area.
2. **Frame Iteration:** The system reads the input video frame by frame in a loop.
3. **Detection:** Each frame is fed into our custom YoloDetector class, which serves as a wrapper for the YOLO model. It formats the detection output (bounding boxes, confidence scores, and class labels) to be directly compatible with the DeepSORT tracking algorithm.

4. **Tracking:** The detections are fed into the Tracker (DeepSORT), which assigns an initial track ID and extracts an appearance feature vector for each person.

5. **ID Refinement:** The `PostTracker` class serves as a crucial second layer of logic to resolve the ID switching inherent in real-time trackers such as DeepSORT. Its purpose is to maintain consistent identities even when individuals are occluded or leave and re-enter the camera's view.

   The re-identification process works as follows:

   1. Candidate Search When a new DeepSORT ID appears, the system compares it against all tracks currently in the "lost" state (i.e., tracks that recently disappeared but are still within a configurable timeout window).

   2. Combined Cost Function For each candidate match, a weighted cost is calculated based on two factors:

   - **Appearance Cost:** Computed as $1 - \text{cosine\_similarity}$ between the new track's appearance feature vector and the rolling historical average of the lost track. This prevents errors from sudden posture or lighting changes. A lower cost indicates a stronger visual match.
   - **Spatial Cost:** Computed as the normalized Euclidean distance between the new track's centroid and the last known position of the lost track. A lower cost means the new track reappeared close to where the old one vanished.

   These two metrics are combined as:

   $$\text{Combined Cost} = (1 - \lambda_{\text{motion}}) \cdot \text{Cost}_{\text{appearance}} + \lambda_{\text{motion}} \cdot \text{Cost}_{\text{spatial}}$$

   Here, λ_{motion} balances appearance vs. motion. For example, $\lambda_{\text{motion}} = 0.33$ assigns 67% weight to appearance similarity and 33% to spatial proximity.

   3. Decision and Re-association The system identifies the lost track with the lowest combined cost. If this minimum cost is below a defined threshold, the new DeepSORT ID is mapped to the persistent ID of the re-identified person. Otherwise, a new persistent ID is assigned.

   Outcome By separating temporary DeepSORT IDs from persistent IDs and applying this weighted cost function,

   `PostTracker` ensures robust identity consistency across occlusions, re-entries, and crowded scenarios.

6. **Analysis:** Once the tracking data is refined by the PostTracker to ensure stable IDs, it is fed into two specialized analysis modules: the **HeatmapGenerator** and the **CountingSystem**.

   Heatmap Generation: The `HeatmapGenerator` creates visual representations of foot traffic and engagement within the monitored space. It maintains two distinct heatmaps:

- **Decaying Heatmap:** A real-time view of activity. For each detected person, a Gaussian-like heat signature is applied to a grid centered at the base of the bounding box (approximating foot position). Heat values gradually decay, producing a dynamic visualization of current traffic flow and popular spots.
- **Cumulative Heatmap:** Aggregates all heat signatures over the full video duration with no decay. This historical record highlights long-term traffic patterns and underutilized areas, useful for layout optimization.

The final video blends the decaying heatmap with the original frame for intuitive visualization. A static cumulative heatmap image is also generated to summarize overall spatial usage.

Visitor Counting and Dwell Time Calculation: The `CountingSystem` accurately counts visitors and measures engagement using a two-zone approach:

- **Entrance Areas:** Buffer zones around the main area. A visitor must first appear in an entrance before being considered.
- **Main Area:** The primary zone of interest where counting and dwell time are calculated.

Counting logic ensures robustness against ID switches: a person is only counted if they move from an entrance to the main area and remain for longer than a threshold (e.g., 3 seconds). This prevents counting passersby and ensures only engaged visitors are included.

Customer Classification and Data Export: After processing, each visitor is classified based on total dwell time in the main area:

- **Dwelled:** Stayed longer than the threshold → considered engaged.
- **Passed By:** Entered but left before the threshold → not engaged.

All metrics—including visitor counts, dwell times, and classifications—are exported to CSV files for further reporting or integration into business intelligence dashboards.

7. **Visualization:** The system draws the bounding boxes, track IDs, counting zone polygons, and a heatmap overlay onto the current frame.
8. **Output:** The annotated frame is written to an output video file. This process repeats for all frames.
9. **Finalization:** After processing the entire video, the system finalizes the dwell time calculations and saves the analysis and the raw dwell times to separate CSV files. A final cumulative heatmap image is also saved.

3.4 Criteria Used to Evaluate Performance

1. **Visitor Count:** The total number of unique individuals counted within each defined area.

2. **Dwell Time:** The time in seconds each tracked individual spends in an area, used as a proxy for engagement.

3. **Dwelled vs. Passed-By Ratio:** The percentage of visitors who stay in an area versus those who quickly pass through. This metric helps classify customer intent.

4. **Heatmap Visualization:** The clarity and accuracy of the generated heatmap in representing high-traffic zones.

4 Results Discussion

The developed video analytics system successfully fulfilled all primary objectives outlined in the problem statement, demonstrating a robust capability to detect, track, and analyze human movement within the Jarir Bookstore environment. When applied to CCTV footage from the high-traffic laptops section, the system provided a multi-faceted analysis of customer behavior in real-time. Visually, the system's output transformed the raw video feed into an intelligent analytical tool. As shown in Figure 1, each customer was robustly identified and encapsulated within a precise bounding box, and more importantly, assigned a unique tracking ID. This ID was persistent across frames, enabling the longitudinal analysis of each individual's complete journey through the monitored area, a foundational step for all subsequent behavioral metrics.

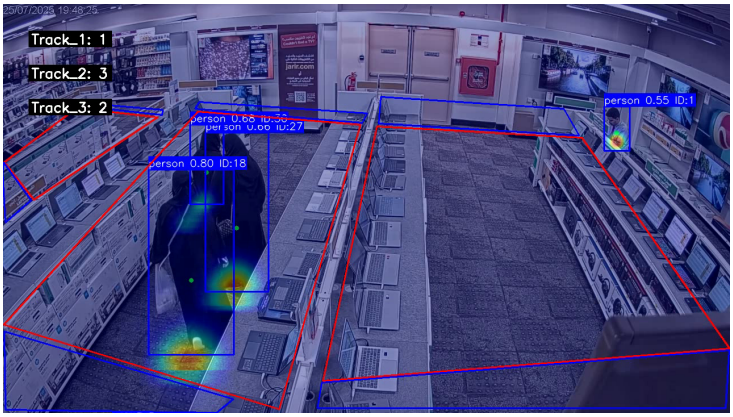


Figure 1: Heatmap overlay generated by the system.

To directly address the objective of visualizing foot traffic patterns, this tracking data was aggregated to generate a dynamic heatmap. This powerful tool turns thousands of data points into a clear picture of shopper behavior, showing where customers gather, the busiest spots, and their common paths. Rather than simply displaying presence, the heatmap was generated by accumulating heat at the foot level of each person in every frame, creating a nuanced and accurate depiction of where customers physically stood and lingered over time. Furthermore, to fulfill the requirement of quantifying visitor traffic, the monitored area was logically segmented into three distinct regions of interest. The system maintained a dynamic, real-time count of individuals within each zone, offering an immediate and quantifiable measure of customer distribution and density.

Beyond this crucial visual analysis, the system was engineered to extract quantitative metrics, thereby satisfying the objective of measuring customer engagement and differentiating behavior types. A key achievement was the successful classification of visitors, which addressed the challenge of distinguishing between customers staying in the area showing interest and those merely passing through. As summarized in Table 1, the system categorized each individual as either "dwelled" or "passed by" based on a predefined time-duration threshold within specific zones. For those classified as "dwelled," the system precisely calculated their total engagement time, as detailed in Table 2. This dwell time metric serves as a direct and powerful proxy for customer interest, highlighting which products or displays are most effective at capturing shopper attention.

Table 1: Visitor Classification for Track_3

Area	Category	Count	Percentage
Track_3	Dwelled	3	37.5
Track_3	Passed By	5	62.5
Track_3	Total	8	100.0

Table 2: Dwell Times for Visitors in Track_3

Track ID	Dwell Time (seconds)
4	41.3
14	21.6
41	4.73

Ultimately, the successful integration of these visual (heatmaps, real-time counts) and statistical (dwell time, classification) outputs provides a comprehensive solution for understanding in-store behavior. The extraction of these metrics demonstrates the system's capacity to move beyond simple people-counting to capture nuanced behavioral patterns. This equips Jarir Bookstore with actionable intelligence, laying a solid, data-driven foundation for strategic decisions aimed at optimizing store layouts, improving product placement, and fundamentally enhancing the overall customer experience.

5 Reflections on Challenges

Throughout the development of this project, several technical and logistical challenges were encountered. This section reflects on these hurdles, the strategies employed to overcome them, and the unexpected insights gained during the process.

One of the most significant technical challenges was the prevalence of ID switching in the initial tracking implementation. Due to factors such as customer occlusion and similar clothing (e.g., abayas and thobes), the tracking algorithm frequently lost and incorrectly reassigned unique identifiers to individuals. To mitigate this, a PostTracker module was developed. This secondary tracking layer implemented a more robust re-identification logic, significantly improving tracking consistency and reducing erroneous ID switches.

A related challenge was balancing the strictness of the counting algorithm. To prevent a single person from being counted multiple times due to ID switches. Fine-tuning the parameters of both the tracker and the counting logic was an iterative process that required careful adjustment to achieve a reliable balance between preventing false positives and ensuring accurate counts.

Another challenge was obtaining suitable camera angles for analysis. The initial dataset provided limited analytical value due to poor positioning. Once better angles were secured, the next issue was handling the large file sizes of the high-resolution CCTV recordings.

Finally, cultural aspects of the Saudi context introduced additional complexity. For example, it is common for multiple women wearing abayas to walk closely together. In such cases, the detection model occasionally grouped them into a single bounding box, misidentifying several individuals as one. This remains a limitation and highlights the need for more advanced detection or segmentation techniques in future work.

6 Conclusion and Future Work

This project successfully developed and implemented a comprehensive video analytics system tailored for the retail environment of Jarir Bookstore. By integrating a YOLOv8 detection model with a DeepSORT tracker enhanced by a custom PostTracker module, the system effectively addresses the critical challenge of ID switching caused by occlusions and visually similar attire. The key finding is that a multi-layered approach, combining real-time tracking with a sophisticated re-identification logic based on both appearance and spatial heuristics, is essential for maintaining robust tracking in a real-world setting. The system successfully transforms raw CCTV footage into actionable business intelligence by quantifying visitor counts, measuring engagement through dwell time, classifying customers as "dwelled" or "passed by," and generating intuitive heatmap visualizations of foot traffic. By creating a custom, fine-tuned solution, this system provides a powerful and adaptable analytics tool that can deliver more granular insights than many generic, off-the-shelf commercial products.

While the current system demonstrates strong performance in detecting and analyzing customer behavior, several directions remain open for improvement. First, a dedicated neural network can be trained to distinguish store employees from regular customers, ensuring that staff presence does not bias the analytics. This would yield more accurate insights into genuine shopper engagement. Second, a major enhancement would be to detect customer-product interactions. By defining micro-zones around key product displays, the system could be trained to identify and count instances where a customer physically engages with an item, providing direct, quantitative feedback on product visibility and appeal. Finally, to enable a store-wide deployment, the system must be architected for multi-camera scalability. This involves building a robust pipeline capable of processing multiple, simultaneous video feeds and, more importantly, implementing cross-camera re-identification to seamlessly track a single customer's journey as they move between different camera views, providing a holistic view of in-store navigation patterns.