People Counting and Face Recognition in Video-Based for Retail Analytic with YOLOv8-v10

Saranthorn Srimuang
Student Number 220979373
Anthony Constantinou
School of Electronic Engineering and
Computer Science
Queen Mary University of London

***Abstract*— Retail stores do not have sufficient insight data about customer demographic information to analyze customer behavior and improve service to target customer groups, creating efficiency in advertising, marketing strategy, and promotion and increasing sales.**

Video analytics methods are currently gaining popularity across various industries. For example, retail stores implement video-based recognition systems from CCTV cameras to detect faces to estimate age and gender for segmenting customer groups, and people counting systems to track customer counts at entrance and exit to analyze customers during peak periods and optimize customer flow with computer vision.

This paper's research is based on a deep convolutional neural network (DCNN) architecture for age and gender estimation and people counting systems (entrance and exit in a store). In addition, the performance accuracy was compared with the YOLOv8, YOLOv9, and YOLOv10 versions using datasets from the Adience database and the actual test video footage from a home care product retail store from the website. In terms of the experiment compared performance accuracy in YOLOv8, VOLOv9, and VOLOv10 versions, the best performance accuracy results were YOLOv10 at 75.0% for the people counting system and at 50.0% for gender recognition and age estimation. Additionally, this paper implements data visualizations for statistical analysis on computer and mobile application platforms.

Keywords—video analytics, face recognition, people counting, gender recognition and age estimation, YOLOv8, YOLOv9, VOLOv10

I. INTRODUCTION

Retailers operate in a dynamic market in a fast-paced environment where consumer trends and changing behaviours rapidly evolve consumer expectations and the competitive nature of the retail industry. Therefore, one of the crucial tasks of video analytics is age estimation, gender recognition, and tracking customer analysis in the retail industry, driven by computer vision in artificial intelligence (AI).

Computer vision can contribute to the advancement of retail analytics by providing valuable insights and enhancing the customer experience. Therefore, video-based age estimation and gender recognition are other vital parameters to analyze because each customer's gender and age group have different preferences and satisfactions. For example, the study shows that women were more concerned about the available menu choices, convenience, health, and restaurant brands in the restaurant business. In contrast, men were more concerned about the atmosphere and price. Therefore, knowledge of what gender is used to create targeted promotions or advertisements is essential [1].

For people, counting is used to monitor people entering and exiting the retail store based on data collected from a surveillance camera in real time. This purpose is to count the number of customers each day, which can track dwelling time and peak periods while customers are in a retail store [2].

YOLO (You Only Look Once) is a powerful object detection algorithm that operates in real-time, locating objects within images with impressive speed and accuracy. In this paper, we experiment with VOLOv8, VOLOv9, and VOLOv10 to compare their performance accuracy in age estimation, gender recognition, and people counting systems using the same dataset. This comparison provides valuable insights into the capabilities of these systems in real-world retail scenarios.

Problem Statement

In the retail industry, customer satisfaction and understanding customer behavior are critical factors to consider in enhancing store efficiency. Traditional retail stores use the records of cash registers or credit cards to analyze customers' buying behaviors [3]. Nevertheless, this information cannot be interpreted as the customer's interest in the front of the merchandise shelf, buy or not buy, and cannot count people shopping groups waiting in checkout lanes. Therefore, customer behavior recognition through CCTV cameras is more important [4].

The analysis of customer behavior from surveillance cameras is essential in marketing and beneficial store management. It allows for the segmentation of customers in a retail store during specific periods and for the analysis of

customer flow and peak-activity periods. This, in turn, enables retail owners to allocate staff more efficiently. Moreover, this technology plays a crucial role in providing retailers with more accurate and informed decision-making processes [5]. In this paper, recording Video footage from a home care product retail store with YOLOv10 is shown in Figure 1 and available link for implementation [6].

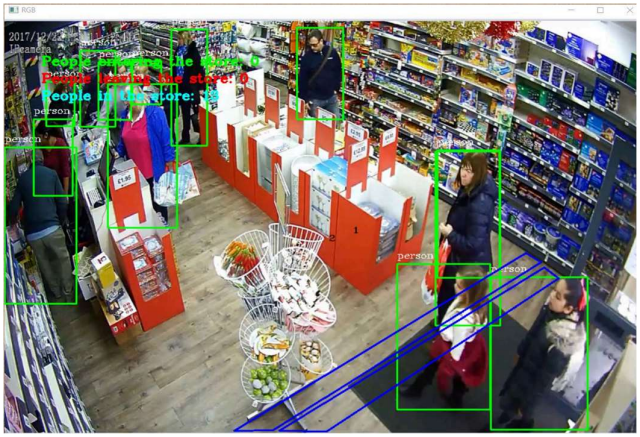


Fig.1 Frame from CCTV camera in retail store with YOLOv10 [6].

CCTV systems are usually installed in enterprises, such as entrances and exits, cash registers, and store corridors for security purposes. In addition, the existing surveillance cameras for data analytics are vital for using big data in analytic business to improve service levels.

In this paper, we discuss the difficulties from a computer vision perspective. Four problems are to be considered.

1) When a person enters the store, the camera cannot capture a high-quality image due to improper lighting, facial expressions, or heavy makeup. This can cause facial detection to be inaccurate because gender and age prediction are difficult. Additionally, the appearance of people of the same age can differ depending on genetics and lifestyle characteristics.

2) The issue of object tracking inaccuracy is a significant challenge. When objects move out of the frame and re-enter rapidly, the resulting tracking errors can be severe. Moreover, poor camera angles or proximity can further hinder the detection of people entering and exiting.

3) The YOLO model's performance, while effective, comes at a cost. Its resource-intensive nature demands significant processing time and high machine resources, leading to delays in video processing.

4) Dataset limitations: The audience dataset may not cover all facial expressions or facial features because the prediction is inaccurate.

The following objectives are proposed:

1. Classify demographic customers by age group and gender recognition with computer vision.
2. Tracking and detecting customers who enter inside or outside a retail store.
3. A DCNN model that learns how to perform efficiently and accurately by comparing the ground truth.
4. Comparison of face recognition and people counting performance accuracy with YOLOv8, YOLOv9, and YOLOv10.
5. Report with dashboard: summary with visualization that can presented on computer desktops and mobile devices.

.

II. Related Work

Age and Gender Estimation

Several methods for face detection, including Capsule networks (Capnet), have been developed. Hinton et al. [7] introduced a new method for unsupervised learning called a capsule, which improved more effectively than plain CNNs. Recently, applying the pertained model integrated into the Caffenet CNN framework, along with the haar cascade,makes model outperform[8]. In addition, using the VGGNet Architecture of Deep Convolution Neural Network (D-CNN) is particularly noteworthy. This architecture can extract features from an image and provide output without relying on feature descriptors like Histogram Oriented Gradient (HOG) and Support vector machine (SVM). As a result, it has shown the highest accuracy compared to other Deep Convolution Neural Networks such as GoogleNet, Resnet, and DenseNet [9]. To illustrate, the VGG-16, a deep convolution network Model, is significant to be considered; it achieved an impressive 91.3% accuracy for gender prediction. [10].

People Counting System

Several researches have been performed on different object detection algorithms. Different techniques include machine learning, image processing, and deep learning, such as support vector machines (SVMs)[9] and convolutional neural networks(CNNs). The CNN consists of two phases: Object detection and tracking. The disadvantage of CNN is that it requires high computational power [11] whereas according to Loy al.[12] they are divided into three groups: counting by detecting, counting by clustering, and counting by regression. Recently, there has been a low-cost deep learning approach to estimate the number of people in retail stores in real-time and detect hot spots named RetailNet[2] the experiment outperforms straightforward CNN approaches.

III. METHODOLOGY

This section discusses the technical details of implementing face recognition (age group, gender classification, and people counting).

A. Programming Language

Python with Visual Studio Code (VS Code) is an open-source programming language widely used in data analysis, artificial intelligence, web development, and scientific computing. In this paper, we use the following libraries: Open CV, TKinter, YOLO, and Tracker to run programmers that detect tracking and processing video.

B. The proposed method

The proposed method, shown in Figure 2, consists of four main blocks: face detection, people identification, age group and gender classification, and people counting system. Each block is explained in the following sections.

1. *Face Detection:* It detects people's faces in video frames and crop faces. Object detection uses the YOLO model.

2. *People Identification:* This block presents a significant technical challenge, as people in videos are not static and can move in unpredictable ways. Therefore, each frame must accurately identify and label every person to prevent multiple detections of the same individuals. The process involves detecting three cropped image faces at subsequent time frames, with non-face images considered as noise that can affect face similarity checking. The eyes and bottom lips maintain position on every image, adding to the complexity of the task.

3. *Age group and gender Classification*: After detecting people in the video, the stage age group and gender classification were used using a transfer learning method. This technique is proper when the related task has limited data; it can improve the model's performance on the new task because it is less time-consuming than training a new model. This paper uses the Adience dataset, which consists of 26,580 images of human faces, to evaluate the performance. This dataset contains genders classified as male and female. For age range: 0-2, 4-6, 8-12, 15-20, 25-32, 38-43, 48-53, and 60-100[13]. However, these age groups have categorized too many age groups in the real world that have yet to be used. Therefore, the age group was reorganized into four groups: baby age between (0-2), child (4-13), adult (17-53), and elderly (60-100).

4. *People counting System:* people counting operates on a three-step approach: detection, tracking, and counting. In YOLO (You Only Look Once), an object detection iteration detects and assigns a bounding box in the frame. This is followed by tracking algorithms like SORT (simple online and real-time tracking) and deep SORT, which can track detected individuals across multiple frames. The people counting system extracts faces from the age group and gender classification. The output is presented as the total number of people, the total number of people by gender, the total number of people by age, and the total number of people grouped by gender and age.

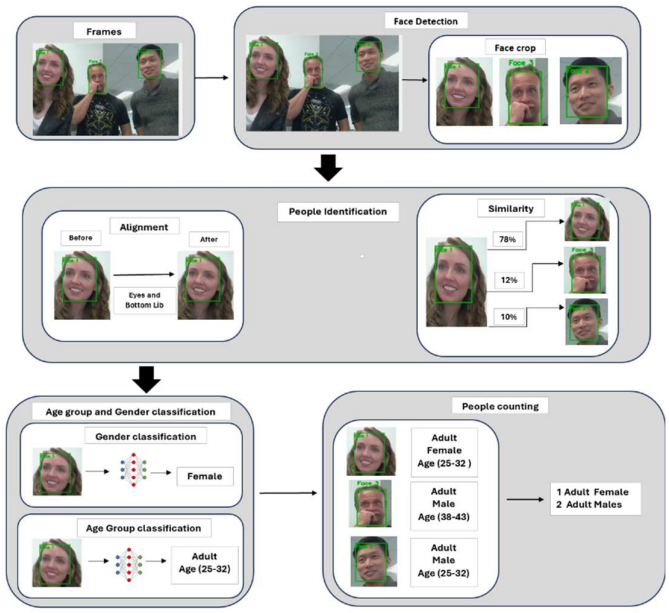


Fig.2 Flow chart for Face recognition and People Counting

C. Object Detection

YOLO (You Only Look Once) is a widespread real-time object detection and image segmentation module. YOLO is a neural network that predicts an image's bounding boxes and class probabilities. Several versions of YOLO have improved the accuracy and speed of previous versions. This paper compares performance accuracy with YOLOv8, YOLOv9, and YOLOv10 versions.

1. YOLOv8 Model

YOLOv8 is a series of object detection models, image classification, and instance segmentation tasks designed to be more efficient and improved over previous versions, as shown in Figure 3.

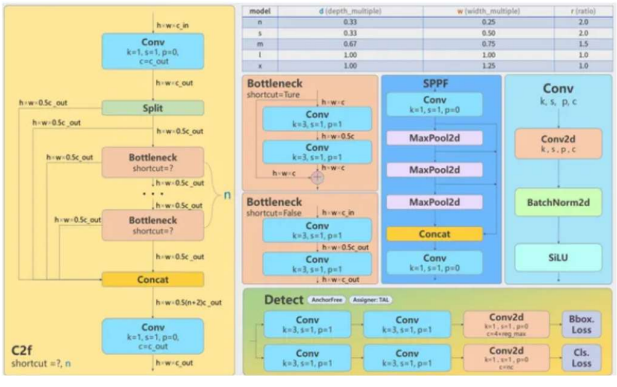


Fig.3 Architecture of YOLOv8

Key Feature and Architecture of YOLOv8

There are three significant blocks in the algorithm as follows:

1. *Backbone*: YOLOv8 is designed to extract features from the input image. It helps to reduce computation with high accuracy. The backbone consists of convolutional layers, down-sampling layers, and residual blocks.

2. *Neck*: The neck of YOLOv8 aggregates features from different stages of the backbone and prepares predictions using PANet (Path Aggregation Network), which improves information flow and enhances the feature pyramid. It fuses different scales and integrates contextual information to improve detection accuracy. In addition, it reduces the spatial resolution and dimensionality and increases speed.

3. *Head:* generates the network's output, such as bounding boxes and a confidence score for object detection. It generates bounding boxes and assigns a confidence score to each bounding box.

2. *YOLOv9 Model*

YOLOv9 is designed to improve speed and accuracy in real-time object detection tasks. It consists of two key innovations: Programmable Gradient Information (PGI) and the Generalized Efficient Layer Aggregation Network (GELAN). These features help solve issues related to information bottlenecks and gradient loss, as shown in Figure 4.

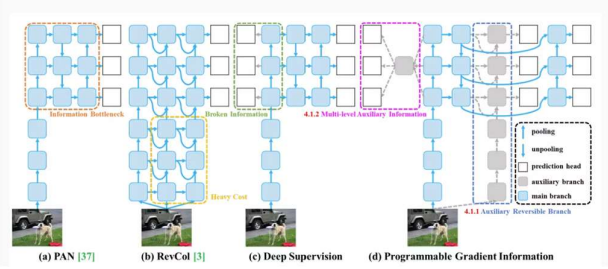


Fig.4 Architecture of YOLOv9

Key Feature and Architecture of YOLOv9

1. *Programmable Gradient Information (PGI)*:
PGI is a novel concept introduced in YOLOv9 to tackle the information bottleneck problem and preserve of essential data across deep network layers. This allows for the generation of reliable gradients, accurate model updates and improved detection performance.

2. Generalized Efficient Layer Aggregation Network (GELAN):
GELAN represents a strategic architectural advancement, enabling YOLOv9 to achieve superior parameter utilization and computational efficiency. It combines the strengths of CSPNet, which enhances gradient path planning, with GELAN, as shown in Figure5.

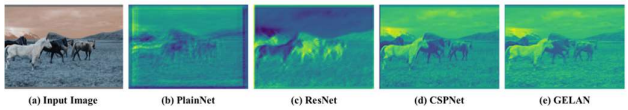


Fig.5 Visualization result from different network architectures: (e) GELAN

3. YOLOv10 Model

YOLO10 introduces a new approach to real-time object detection by eliminating non-maximum suppression (NMS) and optimizing various model components, as shown in Figure 6.

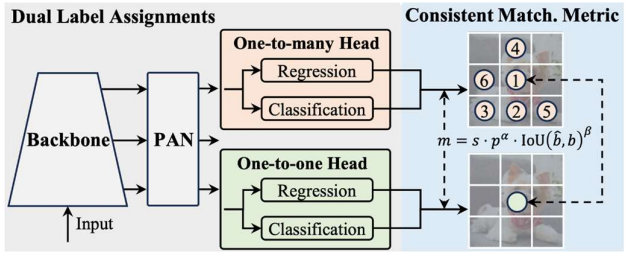


Fig.6 Architecture of YOLOv10

Key Feature and Architecture of YOLOv10

The architecture of YOLOv10 has the strengths of previous YOLO models. The model architecture consists of the following:

1. *Backbone*: Responsible for feature extraction, the backbone in YOLOv10 uses an enhanced version of CSPNet (Cross Stage Partial Network) to improve gradient flow and reduce computational redundancy.

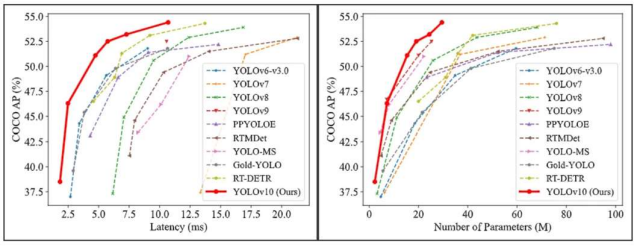
2. *Neck*: The neck is designed to aggregate features from different scales and passes them to the head. It includes PAN (Path Aggregation Network) layers for effective multiscale feature fusion.

3. *One-to-Many Head*: Generates multiple predictions per object during training to improve learning accuracy.

4. *One-to-One Head*: This method generates a single best prediction per object during inference, eliminating the need for NMS, reducing latency, and improving efficiency.

Comparison with YOLOv10- with other models

YOLO has several versions, from YOLOv1 to VOLOv10, to improve accuracy, speed, and functionality. Figure 7 shows YOLOv10's performance proficiency with other versions.



YOLOv10 Performance from the official YOLOv10 repository.

Fig.7 Comparisons with others in terms of latency-accuracy (left) and size-accuracy (right) trade-offs

YOLOv10 architectural improvements reduce parameter and latency across various model scales. Speed and accuracy VOLOv10 are rapid and high, with the small size processing each image in just one millisecond (1000fps) with high speed on the CPU.

D. Flow diagram of face detection and people counting

The following flowchart demonstrates the workflow of a video processing system. The process involves pertained models, including a face detection network, age and gender classification, and people counting using the YOLO algorithm. It is shown in Figure 8.

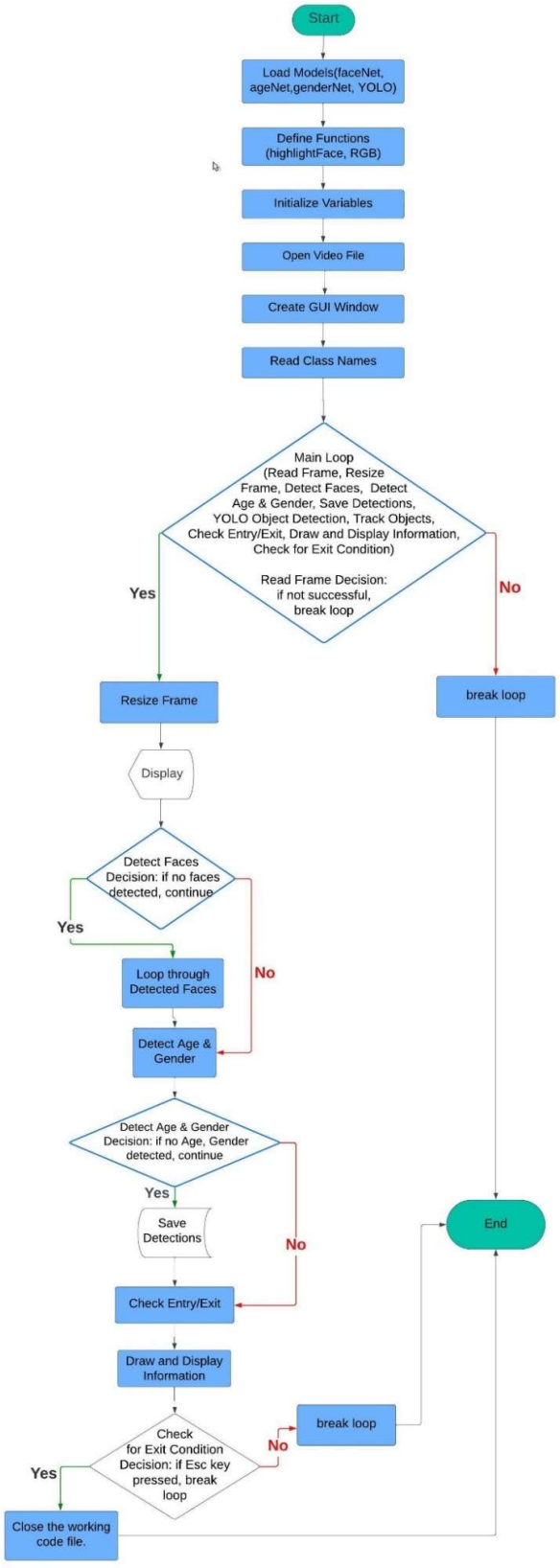


Fig 8 Overall flow diagram of face detection and people counting

Installation and Operation

1) *Import Libraries*: Open CV, TKinter, Yolo, and Tracker

2) Function Highlight in Face function in image processing or machine learning involves highlighting or tracking faces in images or videos involving face detection and object tracking techniques.

3) *YOLO model:* configure and set regions of interest that define the area in the image (area: enter and exit).

```
# Load YOLO10 model for object detection
model = YOLO(r'C:\Users\asus\Desktop\age and gender detection\age and gender d
```

```
#Define two areas as lists of tuples representing coordinates
area1 = [(366, 715), (830, 406), (854, 422), (428, 714)]
area2 = [(438, 715), (859, 431), (882, 450), (513, 716)]
```

4) RGB function: identifying and highlighting using color information (RGB values) for tracking objects of a specific color or segmentation of an image..

5) Create a GUI window by TKinter.

6) Read coco.txt for object tracking.

```
my_file = open("coco.txt", "r")
data = my_file.read()
class_list = data.split("\n")
```

7) Read class names: detect age and gender, track objects, detect faces, etc.

8) Main Loop for Video Processing: Read Frame, Resize Frame, Detect Age, and Gender Yolo Object Detection. Each tracked object checks whether the object is in the correct position. Is it in a given zone (area 1 `or area 2) or not? Entry or exit: If an object (person) is in area 2, it is considered to be "entering" the area. If the object (person) was in Area 2 before and is currently in Area 1, it will be considered to have "entered" the area successfully. If the object (person) was previously in Area 1 and is currently in Area 2, it will be considered to have "exited" the area. In addition, if unsuccessful, Track Object will be a break loop.

9) Draw polygons for areas 1 and 2: counting people entering, exiting, and currently in the store.

10) Save to a JSON file and close Windows.(Figure 9)


```
{
    "id": 1,
    "gender": "Male",
    "age": "(60-100)",
    "position": {
        "x1": 725,
        "y1": 314,
        "x2": 780,
        "y2": 382
    }
```


Fig 9 Example a Json file

E. Dataset

The dataset used in this paper is the Adience dataset, available at the Kaggle website. It comprises face photos in real-world imaging conditions like noise, lighting, pose, and appearance, collected from Flickr albums and distributed under the Creative Commons (CC) license. It has 26,580 photos of 2,284 subjects in eight age ranges and is about 1 GB in size. This dataset was chosen for this work because it has many images of faces and uses different accessories, such as earrings, hats, and glasses, with emotions (happy, neutral, unhappy) and different gestures and resolutions. Figure 10 shows a sample image in the Adience dataset.

Fig 10 Sample of each age group and gender from the Adience dataset.

The actual test video footage is from the iProx CCTV HD Dome 3MP and 4MP high-resolution CCTV cameras sold on hdcctvcameras.net. This video was recorded via a Hikvision 8-channel POE NVR. To view more high-resolution CCTV videos and photos. They are recorded in the home care product retail store as shown in figure 11.

Fig 11 Video footage from HD CCTV Camera video HDCCTVCameras.net

IV. Result and Discussion

Several factors can be estimating age, gender, and people counting system. In this paper, it is necessary to compute the accuracy metrics to measure the overall correctness of the model by calculating the proportion of correctly classified instances (both true positives and true negatives). The formula is as follows:

$$\text{Accuracy Rate} = \frac{TP + TN}{TP + TN + FP + FN}$$

Our evaluation of performance accuracy in YOLOv8, VOLOv9, and VOLOv10 is illustrated. YOLOv10 achieved the best results, with 75.0% accuracy for the people counting system in Table I, 50.0% accuracy for gender in Table II, and 50.0% accuracy for age group in Table III. Finally, Table IV and Figure 12 show a summary comparison of the accuracy of YOLO versions.

TABLE I: Comparison People Counting accuracy with YOLO

Models	People Counting	Number of People	Number of People Correct Prediction	Accuracy
YOLOv8	entering	7	5	71.4%
	Exiting	5	3	60.0%
	Total	**12**	**8**	**66.7%**
YOLOv9	entering	7	6	85.7%
	Exiting	5	2	40.0%
	Total	**12**	**8**	**66.7%**
YOLOv10	entering	7	6	85.7%
	Exiting	5	3	60.0%
	Total	**12**	**9**	**75.0%**

TABLE II: Comparison Gender Accuracy with YOLO

Models	Gender	Number of People	Number of People Correct Prediction	Accuracy
YOLOv8	Female	12	6	50.0%
	Male	6	1	16.0%
	Total	**18**	**7**	**38.9%**
YOLOv9	Female	12	6	50.0%
	Male	6	1	16.0%
	Total	**18**	**7**	**38.9%**
YOLOv10	Female	12	7	58.3%
	Male	6	2	33.3%
	Total	**18**	**9**	**50.0%**

TABLE III: Comparison Age Group Accuracy with YOLO

Models	Age Group	Number of People	Number of People Correct Predictions	Accuracy
YOLOV8	(0-2)	0	0	0.0%
	(4-6)	2	0	0.0%
	(8-12)	3	0	0.0%
	(15-20)	2	1	50.0%
	(25-32)	2	2	100.0%
	(38-43)	3	0	0.0%
	(48-53)	3	2	66.7%
	(60-100)	3	0	0.0%
YOLOV9	(0-2)	0	0	0.0%
	(4-6)	2	0	0.0%
	(8-12)	3	0	0.0%
	(15-20)	2	1	50.0%
	(25-32)	2	2	100.0%
	(38-43)	3	1	33.3%
	(48-53)	3	2	66.7%
	(60-100)	3	0	0.0%
YOLOV10	(0-2)	0	0	0.0%
	(4-6)	2	1	50.0%
	(8-12)	3	0	0.0%
	(15-20)	2	1	50.0%
	(25-32)	2	2	100.0%
	(38-43)	3	3	100.0%
	(48-53)	3	2	66.7%
	(60-100)	3	0	0.0%
	Total	**18**	**9**	**50.0%**

TABLE IV: Summary Comparison accuracy YOLO versions

Accuracy	People `Counting	Gender	Age
Yolov8s	66.7%	38.9%	27.8%
Yolov9s	66.7%	38.9%	33.3%
Yolov10s	**75.0%**	**50.0%**	**50.0%**

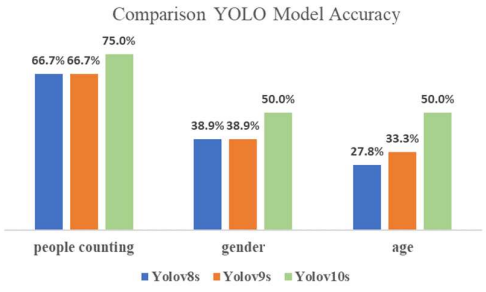


Fig 12 Comparison YOLO Model Accuracy

Our model's gender accuracy of 50% with VOLOv10, while lower than the previous state-of-the-art of 88.2% in Dual et al.'s report [14] and 89.7% in another study's result [15] in the same Adience dataset, still holds significant implications for real-world applications. To find out the problem of low accuracy due to video footage recorded in retail stores, which cannot capture a high-quality image due to improper lighting and inaccuracy because of poor camera angles or proximity, which can hinder the detection of people entering and exiting.

Figure 13: The results show that males, comprising 33% and 67% of females, correspond to correct predictions for female individuals.

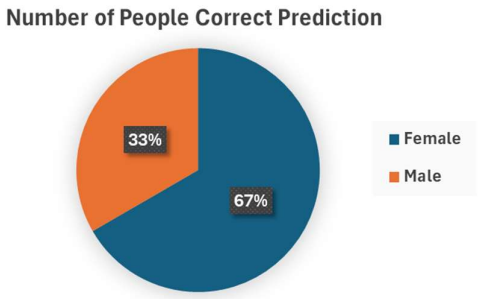


Fig 13 Number People Correct Prediction Accuracy (YOLOv10)

Figure 14: The pie chart illustrates the accuracy of age group prediction using the YOLOv10 model. The chart is divided into four age groups: Adults, Children, Babies, and Elderly. The largest segment is the adult age group at 65%. This indicates that the YOLOv10 model is the most accurate in prediction in the adult age segment. At the same time, the model did not correctly predict any individuals in the Baby and Elderly segment at 0%. This means there is a significant limitation in the model prediction in identifying and classifying these age groups.

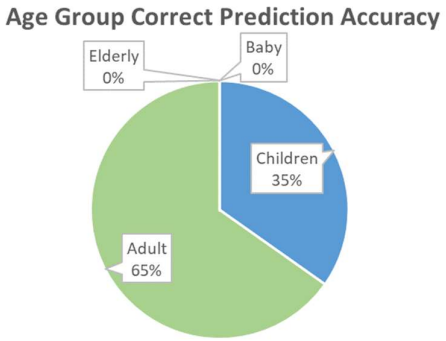


Fig 14 Age Group Correct Prediction Accuracy (YOLOv10)

Data Analysis & Visualization

- Data visualization helps retail owners analyze data to support decision-making to boost sales and improve service.

- The created dashboard will display information in the computer and mobile applications at various levels, as shown in Figure 15.

    1. To specify the period of interest (daily, weekly, monthly)
    2. To display groups of objects of interest such as gender, age range, and people counting (total entries and total exits),
    3. Analyzing the busiest hours, busiest days, and average stay.
    4. Demonstrate customer analysis summary.

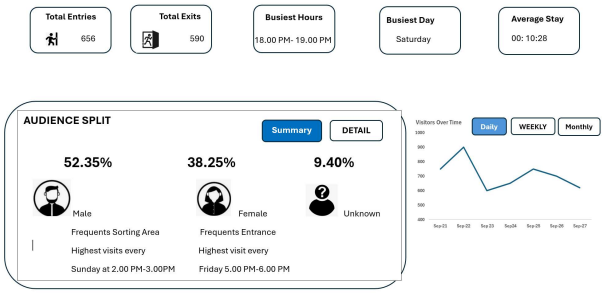


Fig 15 Data Visualization: Dashboard

V. CONCLUSION

This paper's objective is facing recognition and a people counting system that captures images and video footage with a fixed CCTV camera in retail stores. Our paper experiment's accuracy result is lower than that of the previous state-of-the-art experiment because of unqualified video footage. However, to illustrate, comparing performance accuracy in YOLOv8, VOLOv9, and VOLOv10 versions, the VOLOv10 performs the best accuracy results in both face recognition and people counting systems, whereas YOLOv8 and VOLOv9 performance accuracy results look similar. More effective accuracy can be achieved by reducing the model size or processing time and using diverse datasets that cover a wide range of facial features to train models. Additionally, high-quality images and videos closely resembling real-world scenarios can significantly enhance the model's accuracy.

VI. FUTURE WORK

There is a broader scope of retail analysis to be monitored in customer behaviors; steps will be taken to add this to the dashboards: Analysis of dwelling time, assessment of foot traffic, generation of customer movement heatmap [2], and evaluation of customer response to in-store promotions [16]. We plan to develop a system to recognize different customers' behaviors on the front of the shelf: no interest, viewing, turning to the shelf, touching, picking and returning to the shelf, picking and putting into the basket [17]. Additionally, the most important thing to be considered is improving the security and privacy of insightful data for customers. Therefore, we have to implement strong passwords and authenticity.

REFERENCES

[1] Jefferson James Keh et al., (2020). "Video -based gender Profiling on Challenging Camara Viewpoint for Restaurant Data Analytics." International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management(HNICEM).IEEE.

[2] Val´ erio Nogueira et al., (2019). "RetailNet: A deep learning approach for people counting and hot spots detection in retail stores." Conference on Graphics Patterns and Images, IEEE.

[3] Yilin Song et al.,(2017). "Online Cost Efficient Customer Recognition System For Retail Analytics." Winter Application of Computer Vision Workshops, IEEE.

[4] Jingwen Liu et al.,(2015). "Customer Behavior in Retail Store from Surveillance Camara." IEEE International Symposium on Multimedia, IEEE.

[5] Ahmed Hossam et al.,(2024). "Revolutionizing Retail Analytics: Inventory and Customer Insight with AI." Computer Vision and Pattern Recognition.

[6] https://drive.google.com/file/d/1QmVaHyjcOCoIFm5kZ6iPdPCogo-Bqbgw/view?usp=sharing

[7] Sepidehsadat Hosseini et al.,(2019). "GF-CapsNet: Using Gabor Jet and Capsule Networks for Facial Age, Gender, and Expression Recognition." IEEE International Conference on Automatic Face & Gesture Recognition, IEEE.

[8] B.Abirami et al.,(2020). "Gender and age prediction from real time facial images using CNN" Materials Today: Proceedings, Vol. 33, Part 7, pp. 4708-4712. https://doi.org/10.1016/j.matpr.2020.08.350

[9] Jayaprada S et al.,(2022). "Age Detection based on Facial Features Using Support Vector Machine." International Conference on Knowledge Engineering and Communication System, IEEE.

[10] ARSHIN RIZWANA.S,. et al. (2022). "Age & Gender Recognition Using Deep Learning" Third International Conference on Intelligent Computing, Instrumentation and Control Technologies,Vol.72, pp563-571. https://doi.org/10.1016/j.patcog.2017.06.028

[11] Prafful Javare et al., (2020)." Using Object detection and data analysis for development customer insights in retail setting." International Conference on Advances in Science & Technology.

[12] C. C. Loy et al.,(2013), "Crowd counting and profiling: Methodology and evaluation" Springer, pp. 347–382

[13] Karthick R.,(2018) "Deep learning for Age Group Classification System", nt. J.Adv.Sig.Img.Sci, vol. 4, no. 2, pp. 16–22. doi: 10.29284/ijasis.4.2.2018.16-22.

[14] M Duan et al.,(2018). "A hybrid deep learning CNN-ELM for age and gender classification." Neurocomputing, vol.275, pp.448-461. https://doi.org/10.1016/j.neucom.2017.08.062

[15] Karthick R., (2018). "Deep learning for age group classification system," International Journal of Advances in Signal and Image Science, vol.4, no.2,pp.16-22. https://doi.org/10.29284/ijasis.4.2.2018.16-22

[16] Shashimal Senarath et al., (2022)."Customer Gaze Estimation in Retail Using Deep Learning." IEEEAccess, IEEE.

[17] Vishvesh Soni., (2021) "Deep Learning and Computer Vision-Based Retail Analytics for Customer Interaction and Response Monitoring." Eigenpub Review of Science and Technology, vol.5, no 1. https://studies.eigenpub.com/index.php/erst/article/view/43

MSc Project - Reflective Essay

Project Title:	**People Counting and Face Recognition in Video-Based for Retail Analytic with YOLOv8-v10**
Student Name:	**Saranthorn Srimuang**
Student Number:	**220979373**
Supervisor Name:	**Anthony Constantinou**
Programme of Study:	**School of Electronic Engineering and Computer Science**

1 Analysis of the strengths and weakness of the project.

1.1 Strengths

- Using pre-trained models, we are already trained on large datasets and can be fine-tuned for specific tasks, reducing the amount of data and computation time and improving accuracy.
- The VOLOv10 version stands out with its superior performance in both people counting and face recognition, surpassing the accuracy of YOLOv8 and VOLOv9. This should install confidence in the system's performance.

1.2 Weaknesses

- Because objects move out of the frame and re-enter rapidly, we have to adjust the confidence threshold and parameter setting to trial to get better accuracy.
- In comparison, our gender accuracy result of 50.0% with YOLOv10 is lower than the previous state-of-the-art of 88.2% in Duan et al.'s report [1]and 89.7% in other study results [2] because video footage is unqualified, and it is very hard to find the best solution that can achieve the higher accuracy than the previous report.
- Before undertaking this project, I was unable to find actual video footage from CCTV cameras in retail stores due to data protection and the General Data Protection Regulation (GDPR), so I had to find video footage in public. Due to dataset limitations, which include a short video clip and only one clip, we cannot analyze people counting, face recognition, and gender on weekdays and weekends. In addition, we do not have timestamp data for customer flow analysis and dwelling time.
- The AIENCE dataset used to train the model may not cover all facial features or expressions, reducing the prediction accuracy. This is a factor causing our performance accuracy result to be inaccurate.
- Finding the critical factors that can improve video footage accuracy is tough.
- For coding, incorrect or missing command-line parameter settings may cause the program to fail to run, and an improper confidence threshold setting may cause false filtering of detected objects.
- Compatibility of libraries and Python versions: Sharing multiple libraries can cause problems if the library or Python versions do not match, so updating the library or Python version may cause some functions not to work.

2. Presentation of possibilities for further work

Further work can be done to improve our project of retail analysis from a wider perspective, as follows:

- We plan to develop a system that can accurately recognize and categorize different customer behaviors on the front of the shelf. This system will identify actions such as no interest, viewing, turning to the shelf, touching, picking and returning, and picking and putting into the basket [3].
- Store dwelling time analysis might include foot traffic analysis, heat maps, and infrared sensors to track customer movements. In addition, assessing customer reactions to in-store promotions by analyzing facial expressions, body language, and gaze estimation is necessary for evaluating effective promotion advertising and tailoring future marketing strategies to understand customer interest and engagement levels [4].
- Improved face detection can increase the accuracy of gender and age prediction by using actual video records in real situations to find out the problem in business or finding videos and images suitable for counting people entering and exiting to improve the face detection model that can cover multiple facial features.
- We will adjust object tracking using high-performance object trackers suitable for video motion characteristics and optimizing the tracker parameters.
- Enhance model performance by processing it with a high-performance machine, reducing the size of the model, or using a lighter model to reduce processing time.
- Increase dataset coverage, such as using a diverse and comprehensive dataset of facial features to train models using high-quality and real-world images and videos.
- Our focus on user-friendliness and authenticity in the design of our web applications will make the audience feel comfortable and secure when using our improved dashboards.

3. Critical analysis of the different between theory and practical work produced

The practical work in video analysis for face recognition differs from the theory because of constraints, unpredictability, and limited resources. As follow,

Data Quality: In theory, high-quality datasets and video footage that can extract nose, libs, and eyes for face recognition and shape and texture for people counting are required. However, in practical work, video footage is unqualified due to poor lighting, camera angle, makeup, and distance, so the leading model cannot predict precisely.

Real-Time Performance: Theoretical models focus on the highest accuracy and speed, but there is a necessary trade-off between accuracy and speed in practical work. This balance requires reducing video resolution and degrades performance.

Computation: In theory, complex models with multiple layers, such as deep neural networks, are implemented, whereas in practical work, constraints with limited hardware are used. Therefore, models performance predicts inaccuracy.

4. Awareness of Legal, Social Ethical Issues and Sustainability.

Retail stores implement video-based recognition systems from CCTV cameras in stores to detect faces for age and gender, and people counting system will be considered in the General Data Protection Regulation (GDPR) legislation. These laws mandate strict data protection with privacy issues in personal data from video analysis that must be collected and processed transparently for legitimate purposes. People from CCTV cameras' privacy and security are a crucial challenge for individual consent to capture video data in public. Hence, the ethical issue of consent for video analysis is to be considered. Individuals filmed and analyzed may not be aware they are being monitored and may need help to understand what technology can infer about them. Therefore, video analysis technologies must ensure transparency and the purpose of the data usage with public trust. In addition, it can impact public perception and behavior, especially when monitored by age and gender, and it might lead to dissatisfaction or changes in how individuals express themselves in public and societal behavior.

Furthermore, deep learning models may predict bias if a dataset used to train is not representative of the diverse range of human appearance in different ages and genders, potentially leading to biased performance and inaccurate counts or misclassifications of age and gender. In addition, the critical issue is data security, especially data collection from video footage, which should be anonymized to protect personal data, such as blurring faces and changing voice recordings, because anonymizations can alleviate privacy and data privacy. Furthermore, for data storage, strong encryption of the video files and access control policies should be granted with authentication mechanisms. Besides that, data reduction techniques should use only parts of the video frame that are essential for age and gender estimation to minimize unrelated areas and individuals' data.

References

[1] C. C. Loy et al.,(2013), "Crowd counting and profiling: Methodology and evaluation" Springer, pp. 347–382

[2] Karthick R.,(2018) "Deep learning for Age Group Classification System", nt. J.Adv.Sig.Img.Sci, vol. 4, no. 2, pp. 16–22. http:// doi: 10.29284/ijasis.4.2.2018.16-22.

[3] Vishvesh Soni., (2021) "Deep Learning and Computer Vision-Based Retail Analytics for Customer Interaction and Response Monitoring." Eigenpub Review of Science and Technology, vol.5, no 1. https://studies.eigenpub.com/index.php/erst/article/view/43

[4] Shashimal Senarath et al., (2022)."Customer Gaze Estimation in Retail Using Deep Learning." IEEE Access, IEEE.