Improving Retail Intelligence: A Unified Scale Invariant and Contrastive Learning Strategy for Enhanced Object Detection and Customer Behavior Analysis

Dhivya P
Assistant Professor, Department of Artificial Intelligence and Data Science,Karpagam College of Engineering, Coimbatore

***Abstract-** In today's information era, online stores seem to know the customer better than retail stores through advanced data analytics and tracking mechanisms. E-commerce have gained great insights into customer details such as their last purchase, customer preferences, shopping habits, and decision-making processes. To compete with the analytical potential of online stores, retail stores can indeed leverage scale-invariant object detection methods to better meet customer needs by enabling computers to interpret and understand visual information, just as humans do. The motivation of this paper is to detect and analyze customer behavior from video footage acquired by surveillance cameras by combining scale-invariant object detection algorithm and representation learning. With the help of scale invariant object detection method, we can exactly identify customers or objects irrespective of their distance from the surveillance cameras or variations in their appearance due to perspective changes. By exploiting contrastive feature learning, the system can extract selected features from customer interactionscaptured by the scale-invariant object detection process. This allowsfor a fine understanding of customer behavior beyond simple object detection. By accurately detecting and tracking customers at different scales provide detailed analysis of customer behavior suchas store entry and navigation pattern, in-store dwell time, purchase decision, influence of promotions and marketing. These details will help the retail owners to make data driven decisions based on historical data and patterns such as optimize store layouts, product placements, marketing strategies, estimating busy hours, optimize the staff allocation, etc., leading to enhanced better customer experiences.*

***Keywords-** Object detection, Scale-invariant object detection, Customer behavior,Contrastive feature learning, Data driven decision, Representation learning*

1. INTRODUCTION

Artificial intelligence (AI) offers an interesting branch of study called computer vision, which allows computers to examine photos or video data and extract information similar to human beginnings. In this paper we usecomputer vision techniques particularly scale invariant object detection along with Contrastive learning to optimize store operations, improve efficiency, and deliver exceptional shopping experiences that drive customer loyalty and business success.

A. Parameter to Understand Customer Behavior

Retailers need to understand customer behavior in order to improve the shopping experience, to arrange products in the best possible locations within the store, and eventually boost sales.In this paper we are going to track a few parameters to understand customer behavior and they are,

Store Entry and Navigation

Observing customer movement such as how customers enter the store, their navigation patterns within the store layout which include whether customers browse specific sections ornavigate based on signage and displays. Based on this observation,we can easily find out whether the customer is entering a shop byfocusing on the particular product or looking for all resources.

In-store Dwell Time

Tracking how long customers spend inside the store andin specific sections or areas. With the help of this parameter retailers may maximize shop layout and product placement by knowing which groups of consumers reach various regions of thestore and which areas draw the longest dwell times.

Purchase Decision

We need to analyze the factors that influence the customer purchase decision such as product features, pricing, promotions, and brand loyalty. Before buying a product,customers may interact with the product by touching, examining and comparing different products by reading labels and checking price and discount. Understanding which products attract the most attention and engagement can help retailers optimize productdisplays and promotions.

B. Scope of Object Detection in Retail Sector

In retail stores we implement object detection methods to understand customer behavior. By analyzing customer behavior the retail owner can improve business by attracting customers with various discounts and placing the product at the right position. There are various advantages in measuring the customer behaviorand it is listed as below

Inventory Management

Scale invariant object detection algorithm will automatically count and track inventory items on store shelves byanalyzing the images or video obtained from surveillance camera and alert the owner when inventory level falls below predefined threshold. It also identify deviations from the planogram such as misplaced items in shelves or display which help to optimize product placement strategies and enhance the overall store layout.

Enhancing security measures

The scale invariant object detection method works better then single object detection method by identifying the object even if it varies in size or appearance at different distances from the camera which help to prevent theft.

Checkout automation

The advancement in the development of multi object detection methods may help to automatically identify and track products as customers place them in their shopping card which leads to self-checkout or cashier-less stores in future which help to reduce waiting time and enhance the overall shopping experience.

2. RELATED WORK

The traditional methods of tracking consumer behavior in retail environments is either based on manual observation or sensor-based systems have a number of shortcomings, including alack of efficiency in terms of time, inaccurate data, a restricted scope of insights, and an incapacity to pick up on minute details in the behaviors of the customers. To determine consumer behavior, a variety of machine learning techniques are employed, including time series analysis, clustering analysis, sequence analysis, classification and regression, etc. Jiahao Wen [1] paper proposed Case-Based Reasoning (CBR) combined with primitiverecognition to analyze video in retail stores. It decomposes customer behavior (CB) into smaller units called primitives and these primitives are used for pattern matching with predefined behavior patterns. Here the primitives may be single object motionor multiple object relationships. Primitives can be combined to describe a wide variety of distinct CBs and can be used again to characterize different CBs; hence, it can adapt flexibly to target changes in CBs in retail outlets, resulting in strong flexibility andacceptable identification accuracy. This approach utilizes trajectory segmentation and algorithms for recognizing primitives like "move," "stay," and "follow." The paper Jiahao Wen [1] acknowledges limitations like the inability to recognize complex relationships between multiple objects and the omission of "face to" primitive recognition. Tati Erlina [2] research uses the webcam and Raspberry Pi to detect the arrival of a person and The systemrecognizes a person as a possible thief if they cross the imaginaryborder specified by the software, in which case it transmits the image to the store owner via Telegram. Here the YOLOv4 modelis trained on Google Colab with Darknet. Human statuses such as presence in the image, pose, attire, and placement within the imageutilizing bounding boxes are detected through the use of dependent and independent variables. Tati Erlina [2] research only focused on security measures and it did not help to analyze the customer behavior. In their study, Song, H., et al. [3] employed a The Pyramid_Box incorporates a multiscale feature extractor alongside a context-aided module. This context-enhanced single- shot item locator enhances detection precision by managing scaledifferences. Zhang, X., et al [4] employ contrast convolutional neural networks to generate discriminative representations by pre- training them with a Pairwise loss function. One such contrastive feature learning framework is Convolutional Neural Networks for detecting objects. From the literature survey we find out the ML techniques, sensors are used to detect human behavior with less accuracy and time consuming to overcome the issue. We use deep learning techniques such as scale invariant multi object detection along with contrastive feature learning methods to improve accuracy and efficiency.

3. METHODOLOGY

In a retail setting, object recognition is difficult because of a number of variables, including the look, scale, and orientation of the objects as well as occlusions and complicated backgroundsbrought on by heavy traffic. Complex circumstances like occlusions and scale fluctuations provide challenges for established algorithms like SVM, Pyramid box, and CNN. Because of its ability to reliably identify and track items of interest across many sizes while accommodating changes in their size, shape, and appearance, the concept of "scale-invariant object detection" has attracted a lot of attention. To handle complex circumstances related to both occlusions and size fluctuations, it can be paired with contrastive learning. With the aid of a Pairwise loss function, contrastive feature learning is a self-supervised representation learning technique that seeks to identify meaningful data encoding by comparing positive and negative samples.

3 Primary component of Scale invariant detection with representation learning is given by

1. Scale invariant feature extractor
2. Representation learning module
3. Scale invariant object detection network

A. Scale invariant feature extractor

The primary objective of scale invariant adaptive featureextractor is to extract relevant features from images at multiple scales, which means capturing information at different levels of detail. It generates several feature maps with different grades andthen chooses the important attributes for every object detection. Regardless of their scale, objects of different sizes are representedusing feature maps. Fig1 shows the working principle of scale invariant adaptive feature extractor.

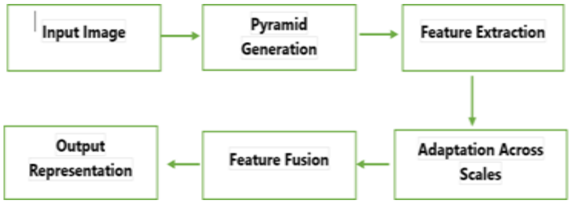

fig1 work flow of Scale invariant adaptive feature extractor

The process begins with an input image which may be a digital image or a frame from a video. Feature Pyramid Network (FPN) employs a pyramid representation of the input image. This involves creating multiple versions of the image, each at a different scale. This can be achieved through techniques like the Gaussian pyramid or Laplacian pyramid. Once the pyramid is generated, feature extraction algorithms are applied at each level of the pyramid. These algorithms are designed to identify and extract relevant information from the image. Features can includeedges, corners, textures, or any other distinctive characteristics. Adaptively, the most informative attributes are chosen for each item proposal from the activation maps at different sizes. ROI pooling is applied to each recommendation to generate a default- size attribute vector . Following this, the attribute vector s undergoprocessing through a sequence of densely connected layers, a softmax layer, and subsequent operations, resulting in a collectionof attention scores. The final feature representation for every proposal is produced by aggregating the feature maps that have been modified in accordance with the attention values.

B. Representation learning module

The aim is to enhance the discrimination between similar and dissimilar entities by amplifying the resemblance among characteristics of similar items and diminishing the resemblance among characteristics of dissimilar items. Using the feature mapsobtained from the scale invariant extractor, this part applies a Pairwise loss function to train a representation specifically designed for object detection. Specifically, the module employs aSiamese network architecture to take two feature maps as arguments and produce the level of similarity among them. When comparing feature maps of the similar item with those of differentobjects, the Pairwise loss function will have a high similarity score.

feature maps. It does this by selecting and resizing feature maps that come from the scale invariant feature extractor. This producesa set of object proposals and confidence ratings. The modified RetinaNet architecture of the network encompasses a Feature Pyramid Network (FPN) with a classification and regression subnetwork, in addition to a backbone network tailored for specific tasks. The regression subnetwork predicts the offset and scale of each object proposal in relation to a specified set of anchor boxes, while the classification network forecasts the possibility that each object proposal contains an object of interest.

4. EXPERIMENTS AND RESULTS

A. Dataset

The model is trained using the Oxford Town Center dataset, sourced from surveillance cameras monitoring pedestrian activity in a bustling downtown area of Oxford. This dataset comprises a 5-minute video recorded at 25 frames per second (FPS), totalling 7500 annotated frames. Among these, 6500 frames are allocated for training, while the remaining 1000 frames serve as testing data for pedestrian detection. For evaluating the model's accuracy, the testing dataset utilized is the Mall dataset, acrowd counting dataset consisting of 2000 images captured within a mall environment.

The main 2 components of representation learning module is given by

Feature Embedding - The process of creating a condensed feature embedding involves passing the attribute vectors obtained from the scale invariant item detector through a sequence of fully connected layers. Each unit within these layers represents a dimension of the normalized feature embedding.

Pairwise loss function - The training process of the network involves the utilization of a Pairwise loss function. This function prefers correspondence among feature integration of identical objects while promoting dissimilarity between feature integration of unique objects. In particular, for every training instance, two samples are selected at random: an unfavourable sample, represented by a vector of attributes from an object that is distinct from the training instance, and an upward sample, represented by a feature vector from the same item. As a result, the description of the contrastive loss is the sum of the relationship among the margin parameter and the attribute vector of the proposed item, which includes the similarities between the positive and anchor examples.

C. Scale-invariant object detection network

The network is designed to find items in the selected

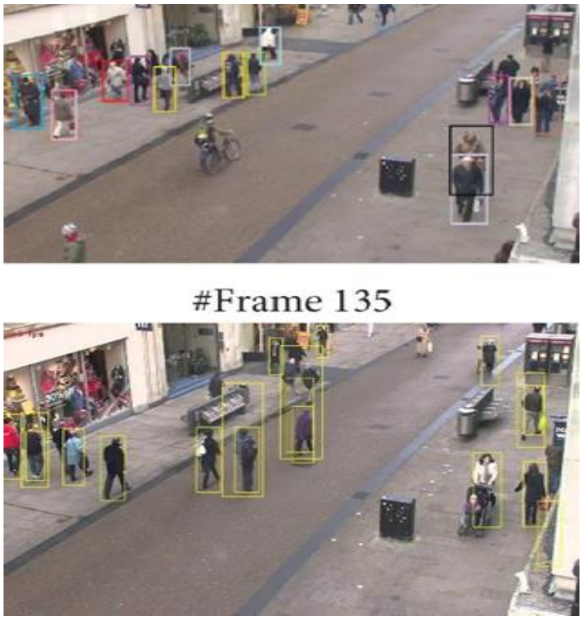

Fig2 : Oxford Town center Dataset

B. Preprocessing Data

Data preprocessing plays a crucial role in ensuring that the dataset is clean, well-structured, and suitable for training machine learning models, ultimately improving the model's performance and generalization ability. Preprocessing involves handling these issues by imputing missing values, detecting and removing outliers, and filtering noise. One common preprocessingstep involves Normalizing image pixel values to a standard scale,usually ranging from zero to one or -1 to 1. This normalization facilitates faster convergence of the model during training.

C. Feature Extraction

After preprocessing the dataset, the ContrastiveFeatureLearningModule() is employed to extract features for training. Contrastive learning commonly utilizes Siamese networks or analogous architectures. In a Siamese network, two identical subnetworks, termed twins, are employed, sharing identical architecture and weights. Each subnetwork processes an input instance (e.g., an image) to generate a fixed- size representation, known as an embedding.

D. Representation Learning

Representation Learning techniques using a Pairwise loss function are utilized to acquire similarity and dissimilarity features. It will show the customer's similar features up close andtheir different features farther away. This improves the discriminative power of the obtained features.

E. Scale invariant Object Detection model

Now integrate the YOLO model with Contrastive Feature Learning Module by defining weight. Here we proposed YOLOv5s because it is ideal for running inference on the CPU. Followed by Feature Fusion, Behavior Analysis, Insights Generation and Evaluation. The overall arguments used in YOLOV5 is given by

-- data: Load Oxford Town Center Dataset by dividing it into training dataset, test dataset.

-- weight: As we are running inference on CPU, YOLOV5s modelis used and the value is given by yolov5s.pt

-- img: is used to control image size, here we resize them to 640 pixels which is most commonly used.

-- epochs: The parameter "epochs" specifies how many complete passes the model makes forward and backward through the entiredataset during the training phase

--batch-size: It indicates how many samples the model processes in a single forward and reverse pass. It directly impacts the training speed and memory requirements. When training the model, you can specify the batch size as an argument to determine how manysamples are processed together in each iteration.

-- name: We can provide a custom directory name where all the results will be saved.

F. Evaluation Metrics

The precision, recall, and F1-score metrics serve as key indicators for evaluating performance. They offer valuable insights into the model's ability to detect objects in images and accurately localize them. These metrics are essential for evaluating accuracy, robustness, and generalization, among otherfeatures of object detection. Achieving a mean average precision (mAP) of 0.81 throughout evaluation, the algorithm showed remarkable effectiveness in detecting clients across various scalesand sizes. Moreover, the system accurately identified and categorized various customer behavior patterns, such as store entry, navigation, purchase decision, and dwell time. It achieved an overall precision of 0.78, recall of 0.84, and an F1-score of 0.82.

5. CONCLUSION AND FUTURE WORK

By combining the scale invariant object detection algorithms such as YOLOv5 and contrastive feature learningusing Pairwise loss function open the door for accurate object localization and detailed feature extraction. By analyzing consumer behavior such as browsing, choosing, and buying, the retail owner can enhance company tactics including inventory management and positioning the appropriate product at the properlocation. The efficiency and scalability of this strategy can be improved in the future by implementing real-time analysis and deployment on devices with limited resources, among other things. Furthermore, extending the assessment to more extensive and varied datasets would confirm its efficacy in a range of retail contexts. For a more thorough understanding of customer behavior and to improve recommendation accuracy and operational efficiency, further contextual data like weather, store layout, and promotional events can be incorporated

REFERENCE

[1] Jiahao Wen, ToruAbe and TakuoSuganuma,” ACustomer Behavior Recognition Method for Flexibly Adapting to Target Changes in Retail Stores” MDPI, Sensors 2022, 22, 6740.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Tati Erlina , Muhammad Fikri, “Yolo Algorithm-Based VisitorDetection System for Small Retail Stores Using Single Board Computer” Journal of Applied Engineering and Technological Science Vol 4(2) 2023: 908-920.

[3] Song, H., et al, “Pyramid box: A context-assisted single shot object detector”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778),2018.

[4] Zhang, X., et al, “CPL: Contrastive pre-training of convolutional neural networks for object detection”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14944-14953), 2020.

[5] A. H. Ahmed, K. Kpalma, and A. O. Guedi, “Human Detection Using HOG-SVM, Mixture of Gaussian and Background Contours Subtraction,” 2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2017.

[6] K. C. Lai, Y. P. Chang, K. H. Cheong, and S. W. Khor, “Detection and classification of object movement - an application for video surveillance system,” 2010 2nd International Conference on Computer Engineering and Technology, 2010.

[7] B. Tian, L. Li, Y. Qu, and L. Yan, “Video Object Detection for Tractability with Deep Learning Method,” 2017 Fifth International Conference on Advanced Cloud and Big Data (CBD), 2017.

[8] Aradhya, HV Ravish. "Elegant and efficient algorithms for realtime object detection, counting and classification for video surveillance applications from single fixed camera." 2016International Conference on Circuits, Controls, Communications and Computing (I4C). IEEE, 2016.

[9] Ravish, H. V., Aradhya Mohana, and Kiran Anil Chikodi. "Real time objects detection and positioning in multiple regions using single fixed camera view for video surveillance applications." IEEE International Conference on Electrical, Electronics, Signals, Communication & Optimization, 2015.

[10] Singh, Shekhar, and S. C. Gupta. "Human object detection byHoG, HoB, HoC and BO features." 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC).IEEE, 2016.

[11] Ding, Sheng, and Kun Zhao. "Research on daily objects detection based on deep neural network." IOP Conference Series:Materials Science and Engineering. Vol. 322. No. 6. IOP Publishing, 2018.