Retail Gaze: A Dataset for Gaze Estimation in Retail Environments

Shashimal Senarath
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
shashimalsenarath.17@cse.mrt.ac.lk

Primesh Pathirana
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
primeshs.17@cse.mrt.ac.lk

Dulani Meedeniya
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
dulanim@cse.mrt.ac.lk

Sampath Jayarathna
Department of Computer Science, College of Science
Old Dominion University
Norfolk, USA
sampath@cs.odu.edu

2022 International Conference on Decision Aid Sciences and Applications (DASA) | 978-1-6654-9501-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/DASA54658.2022.9765224

***Abstract*—The concept of gaze object estimation predicts a bounding box that a person looks steadily. It is a applicable and contemporary technique in the retail industry. However, the existing datasets for gaze object prediction in retail is limited to controlled environments and do not consider retail product category area segmentation annotations. This paper proposes Retail Gaze, a dataset for gaze estimation in real-world retail environments. Retail Gaze is composed of 3,922 images of individuals looking at products in a retail environment, with 12 camera capture angles. Furthermore, we use state-of-the-art gaze estimation models to benchmark the Retail Gaze dataset and comprehensively analyze the results obtained.**

***Index Terms*—computer vision, gaze estimation, deep learning**

I. Introduction

One of the most widely utilized methods for studying human cognition and behaviour is eye gaze estimation and tracking [1]–[4]. In the literature, gaze estimation has been studied in multiple forms, like gaze point and direction estimation, gaze-following, and gaze object prediction. Gaze-Following introduced by Recasense et al. [5], follows human gaze to spot the location they are looking at. In the retail industry, it is vital to identify and analyze the products a customer is looking at to improve the shoppability of the store [6]. Following the gaze of a customer and predicting gaze object prediction can be used to determine the exact object being looked at. However, it is not accurate to capture a single product being looked at from a remote camera view in a real-world retail environment due to the high error margins in 2D gaze estimation methods.

A few datasets related to 2D gaze estimation in retail environments have been introduced with the introduction of gaze following and its applicability in retail environments. The GOO dataset, introduced by Tomas et al. [7], is a widely used dataset for 2D gaze estimation in retail environments. However, the controlled nature of the retail environment in the dataset, the limited number of camera capture angles, and single product item bounding box annotations can be seen as limitations of this dataset.

As the main contribution, we present Retail Gaze, a novel real-world retail environment 2D gaze estimation dataset. The dataset was captured in an uncontrolled retail environment with 12 camera capture angles. The dataset is annotated with area segmentation masks for products that belong to the same category on a retail shelf. Thus, the dataset can be used for gaze object prediction task that predicts an area of product items.

The next section explores the related datasets. Section III describes the Retail Gaze dataset and its tasks and provides a comparison with the GOO dataset. Section IV discusses the methodology used to benchmark the Retail Gaze dataset. Section V describes the experiments conducted and presents the obtained results for both the GOO and Retail Gaze datasets. Section VI concludes the paper.

II. Related Work

A limited number of datasets are available in the literature for gaze-following in retail environment, due to the application-specific nature of the problem [8]. Recasense et al. [5], have introduced the GazeFollow dataset, the first benchmark dataset for gaze-following. It is a large dataset which annotated with the area that the humans are looking at. This dataset has a large scene diversity in which people perform diverse activities in many everyday scenarios. The dataset is a multi-user gaze-following dataset with 122,143 images containing 130,339 people. However, this dataset is not captured in a retail environment and does not contain backhead images, and eye occluded images, which is a major limitation.

GOO [7] is widely used dataset for the task of remote gaze estimation in retail environments. This dataset was published by Tomas et al. [7] with benchmark results of other standard gaze estimation models [5], [9], [10]. The dataset consists of a real image dataset of 9,552 images of 100 subjects and a

Fig. 1. Datasets: (a) Gaze Follow, (b) GOO-Real, (c) Retail Gaze

synthetic dataset of 192,000 images. These images belong to 24 different products in shelves in a retail shop. Moreover, the dataset comprises with gaze point annotations, gazed object segmentation masks and bounding boxes. Compared with the GazeFollow dataset, the GOO dataset provides gaze estimation annotations for closely places many items, which is more suitable in a retail environment. However, GOO-Real and GOO-Synth datasets are captured in an experimental, controlled retail environment from two camera angles. Hence the applicability of this dataset to real retail environments is limited. The limitations of the GOO dataset are discussed in depth in Section 3, A. Sample images from GazeFollow and GOO-Real datasets are shown in Fig. 1 (a) and Fig.(b), respectively.

III. Retail Gaze

The Retail Gaze data is composed of images of a real-world retail environment where each image contains a human gazing upon an object or area on a shelf [11]. Each image captures the third-person view of the customer and shelves. Location of the gaze point, the Bounding box of the person's head, segmentation masks of product areas are provided. Retail gaze contains 3922 images of 2 participants, with each image consisting of a shelf packed with different products. The shelves are completely filled with different products and most of the time the same products are in the same area. Twelve different shelves are captured using only one angle, and in each angle, participants look at most of the product's areas.

We introduced the retail product category area segmentation masks as a novel method of annotating the object boundaries in a retail environment, which is more suitable for real-world scenarios. Moreover, the training, test and validation set consists 2745, 589 and 588 images, respectively, following the split ratio of 70%:15%:15%. Fig. 1 (c) shows a set of images from Retail Gaze dataset. In addition, this dataset is not specific to certain object types and supports many generalized object types. Further, the dataset can be increased by applying different synthetic data generations methods such as data augmentation techniques, change of illumination, intensity, noise, and GAN based methods [12].

For the creation of Retail Gaze, videos were taken using a developed device that uses Raspberry PI 3 development board and a 5MP camera module. All videos are captured under daylight conditions, and external light sources are not required due to the controlled light conditions inside the retail store, and it helps to represent the real-world retail store conditions. For the collection process, each participant walks through the shelf area and gazes at areas they are told to look at for a

few seconds. Each participant was instructed on which area he should look at on the shelf using a predetermined pattern. We extracted several frames from the video for each area. These patterns were used to annotate the ground-truth label. Since the data was collected during the COVID pandemic time, only single user images were captured due to the physical limitations of the collecting dataset at the time. Also, the dataset is captured with wearing face masks.

A. Comparison of Retail Gaze with GOO dataset

We compare the presented Retail Gaze dataset and the GOO dataset, which is the most suitable dataset available in the literature for remote gaze estimation in the retail industry. Most of the GOO and Retail Gaze datasets annotations, such as gaze point and head bounding box, are similar, but in the GOO dataset, product bounding boxes and segmentation masks are annotated for a single product. In the Retail Gaze dataset, segmentations are annotated by product area. The Goo dataset has 201,552 images of both real and synthetic images, but the Retail Gaze dataset is much more similar in size to the GOO-Real dataset. Retail Gaze and GOO datasets are focused on the retail environment. The GOO dataset is collected in a controlled environment, not in a real-world retail store. The Retail Gaze dataset is collected in a real retail store with natural environmental conditions such as lighting conditions, shelf structures, and product placements. The GOO-Real dataset has only two shelves, but the Retail Gaze dataset contains 12 different shelves. Retail gaze contains different product areas on each shelf, but GOO-Real only contains 24 different product categories, and it may be an issue driving models to overfit rather than generalizing the gaze estimation in the retail environment. But the GOO synthetic dataset contains more environmental variations than the Retail Gaze dataset, which may solve the overfit issue and help to improve the gaze estimation in other environments. The GOO dataset has good diversity because the participant count in the GOO dataset is greater than the Retail Gaze dataset.

B. Tasks of Retail Gaze dataset

The Retail Gaze dataset's comprehensive annotation allows it to be used to train systems for several challenging tasks, particularly in gaze estimation and object detection. The proposed Retail Gaze dataset supports the following tasks.

1) Gaze following: Racasens et al. [5] introduced the task of gaze following, that is used to predict the looking point of a human, by using the third-person view image of the person and his head location. This task has two stages: (1) gaze direction estimation using head and scene features, (2) gaze heat map generation using regression. The Retail Gaze dataset benchmarks the task of gaze following by stating the ground-truth gaze point on a scene in the dataset.

2) Gaze object prediction: Although gaze point prediction is challenging, studies have estimated a single gaze point using combining separate systems such as classification and detection, to locate the point a human is looking at. Thomas et al. [7] proposed an approach for gaze object prediction that classifies and predicts the boundaries of the object a human is looking at. This task is complex than gaze following because predicting an object person is looking at in a scene with many objects is more complicated than predicting with fewer objects in gaze following. By defining the gaze object as the ground truth product area that a person looks at, The Retail Gaze dataset can provide benchmarks on this task.

3) Head Detection: Retail Gaze has annotated the bounding box of each participant's head for usage in various applications. Several studies have used head location or image of the head as an input for the gaze estimation task [5], [9], [10]. Generally, face detection identifies the location of the head or the image of the head. However, it cannot handle total or half head occlusion. Thus, training a custom model to recognize the head's bounding box in complex head orientations will be a good solution. The retail gaze dataset is suitable for this task since it contains various head orientations and their associated head bounding boxes. However, we focus only on gaze estimation tasks in our work.

IV. Design and Methodology

A. Baseline Selection

Since we focus on gaze estimation task on the proposed Retail Gaze dataset, we used standard baseline models in gaze estimation. The existing performance benchmark on other datasets can be used to compare and verify the correctness of the proposed baseline implementations. Considering the similarity of the dataset, we used the same baseline methods used in GOO-Real dataset for better comparison of our work [7]. The state-of-the-art benchmark models for the GOO-Real dataset are used for the implementation of gaze estimation, prior to the experiment with the Retail Gaze dataset.

Baseline model architectures should only get the scene image, head location, and object locations as input, and the gaze heatmap should be the output from those models. The highest confidence value of the gaze heatmap indicates the gaze point, and gaze estimation evaluation is done using the gaze heatmap and extracted gaze point.

B. Baseline Methods

The models presented by Racasens et al. [5], Chong et al. [9] and Lian et al. [10] are used as the baseline methods to evaluate retail gaze dataset based on above criteria. These baseline architectures have common hand-made sub modules to solve gaze estimation tasks. These modules solve three subproblems in gaze estimation.1) scene module - extract feature from the scene image 2) head module - extract feature from the head image and head location 3) decoder module - generate a gaze point confidence heatmap from the scene and head feature maps. Recasens et al. [5] have used a shifted-grids approach to predict the gaze point by solving several overlapping classification problems. Chong et al. [9] have extended this approach to handling out-of-frame gaze targets using a multi-task learning approach. Instead of a single gaze direction field, Lian et al. [10] have introduced the concept

of multiple gaze direction fields to generate the confidence heatmap.

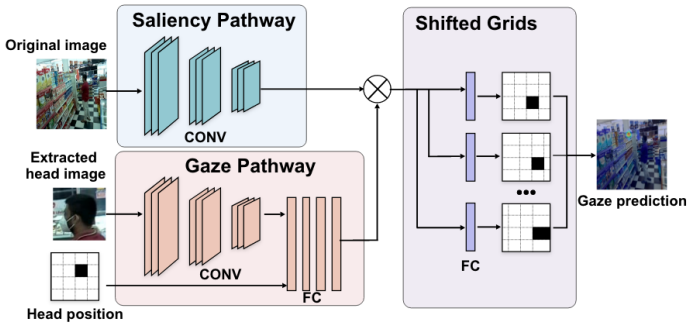

Fig. 2. Dual pathways and shifted grid architecture by Recasens et al. [5]

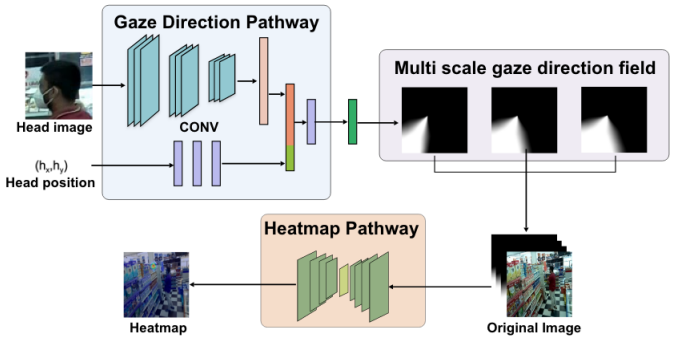

Fig. 3. Multi scale gaze direction field and heatmap pathway architecture by Lian et al. [10]

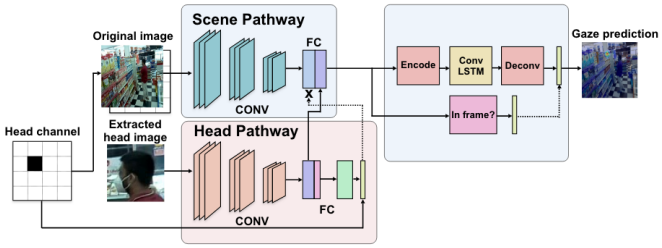

Fig. 4. Dual pathways and attention mechanism based architecture by Chong et al. [9]

V. Experiments and Evaluations

The GOO-Real dataset has used several baselines [5], [9], [10] for benchmarking. We benchmarked the GOO-Real dataset on these baselines and compared the results to identify the correctness of our baseline implementations. Then we used these baselines to benchmark the Retail Gaze dataset and benchmarks of these models shown in Fig. 2, Fig. 3 and Fig. 4. These models were selected for the better comparison of our study with the existing work.

These models are based on the gaze follow behaviour of humans in real world. Initially, we look at the head or eyes to identify the area that the person is looking at. Then, we identify the salient objects in their view point to predict the direction and object they are looking at. Recasense et al. [5] have introduced a multi-model predictions to predict the fixation point with the Gaze Follow dataset. This contained two different CNN pathways as shown in Fig. 2. The saliency pathway for scene image and gaze pathway for closed head image that was cropped from scene image. Features from these two pathways fed into the shifted girds consist of fully connected layer with an attention mechanism to predict the gaze.

Following the model of Lian et al. [10], Fig. 3 consists of a gaze direction pathway, multi-scale gaze direction field and heatmap pathway. Further, the model presented by Chong et al. [9], as shown in Fig. 4, the binary representation of the head location is used, where white and black pixels indicate the bounding box of the head and the other area of the image, respectively. The head pathway computes the feature map from the head image in the scene. Scene pathway computes the scene feature map by taking input as the sequence of the scene image, head position channel, and object channel.

A. Implementation Details

All baseline models are implemented using PyTorch and PyTorch lighting frameworks and trained and tested on Colab Pro. To recreate results as accurately as possible, all relevant pretraining and initialization methods and hyper-parameters such as learning rate, epochs, and batch sizes were obtained from the respective publications for each model.

B. Evaluation Criteria

The baseline models were evaluated on the Retail Gaze dataset using the standard metrics for evaluating gaze following; Area Under the ROC Curve (AUC), L2 Distance, and angular error. The AUC in gaze following proposed by Judd et al. is defined as the area under the ROC curve where the saliency maps are thresholded by categorizing pixels as fixated and unfixated. L2 distance is the mean euclidean distance between the ground-truth gaze point annotation and the predicted gaze point in 2D coordinates. The angular difference between the ground truth gaze vector and the predicted gaze vector in 2D coordinates is considered as the angular error. First, the baseline models were trained on the GOO-Real dataset, with 0-shot, 1-shot, and 5-shot training. After validating the models, they were benchmarked on the Retail Gaze dataset by subjecting them to the same set of learning without pre-training.

C. Results and Analysis

Table I shows the comparison of our results with the existing state of the art models presented by Recasense et al. [5], Lian et al. [10], and Chong et al. [9]. The existing algorithms [7], were trained on both GOO-Real and Retail Gaze datasets to obtain the results.

TABLE I
RESULTS ON GOO-REAL TEST SET.

Model	Existing studies			Proposed work		
	AUC	*Dist.*	*Ang.*	*AUC*	*Dist.*	*Ang.*
[5]	0.850	0.220	44.4	0.848	0.231	45.6
[10]	0.840	0.321	43.5	0.831	0.319	42.8
[9]	0.796	0.252	51.4	0.810	0.253	51.7

TABLE II
RESULTS ON RETAIL GAZE TEST SET.

Model		AUC	Dist.	Ang.
Recasense et al. [5]	0-shot	0.573	0.426	90.2
	1-shot	0.754	0.291	46.5
	5-shot	0.799	0.255	36.4
Chong et al. [9]	0-shot	0.677	0.449	66.8
	1-shot	0.696	0.268	39.8
	5-shot	0.735	0.241	35.2
Lian et al. [10]	0-shot	0.410	0.518	79.8
	1-shot	0.487	0.536	71.8
	5-shot	0.522	0.418	62.4

Table II shows the results obtained using baseline models with the proposed Retail Gaze dataset. The standard 0-shot, 1-shot, and 5-shot training results state the performance improvements of the baseline models after training with the Retail Gaze dataset. The 0-shot evaluation shows the model weight initialization in each baseline model. Considering lowest L2 Distance and angular error, the model by Chong et al. [9] provided better results on Retail Gaze for 1-shot training results, with fewer training iterations. Further, Chong et al. [9] and Recasense et al. [5] provided the best converging models for the dataset, considering 5-shot training.

VI. CONCLUSION

This paper presented Retail Gaze, a dataset for gaze estimation in real-world retail environments, consisting of 3,922 images captured from 12 camera angles. As the currently available retail gaze estimation datasets are captured under controlled environment conditions, we captured the presented Retail Gaze dataset inside a supermarket in an uncontrolled environment from diverse camera capture angles to improve the real-world applicability of the dataset. Moreover, we applied retail product category area segmentation annotations, as a novel method of annotating the object boundaries in a retail environment, which is more suitable for real-world scenarios. Finally, we used state-of-the-art baseline gaze estimation models to benchmark the proposed Retail Gaze dataset through comprehensive experiments, and we analyzed the obtained results quantitatively.

This dataset can be extended to improve the real-world applicability, as it is essential to expand the subject diversity in the dataset. In retail environments, often multiple user scenarios can be seen. Hence, there is an opportunity to expand the dataset to multi-user scenarios to facilitate multi-user gaze estimation in retail environments.

ACKNOWLEDGMENT

We appreciated the support received from the Mae Fah Luang University, Thailand for the conference registration.

REFERENCES

[1] M. K. Eckstein, B. Guerra-Carrillo, A. T. M. Singley, and S. A. Bunge, "Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?" *Developmental cognitive neuroscience*, vol. 25, pp. 69–91, 2017, doi: 10.1016/j.dcn.2016.11.001.

[2] S. De Silva, S. Dayarathna, G. Ariyarathne, D. Meedeniya, S. Jayarathna, A. M. Michalek, and G. Jayawardena, "A rule-based system for ADHD identification using eye movement data," in *Proc. of the Moratuwa Engineering Research Conference (MERCon)*, Moratuwa, Sri Lanka, 2019, pp. 538–543, doi: 10.1109/MERCon.2019.8818865.

[3] J. Kerr-Gaffney, A. Harrison, and K. Tchanturia, "Eye-tracking research in eating disorders: A systematic review," *International Journal of Eating Disorders*, vol. 52, no. 1, pp. 3–27, 2019, doi: 10.1002/eat.22998.

[4] M. Meißner and J. Oll, "The promise of eye-tracking methodology in organizational research: A taxonomy, review, and future avenues," *Organizational Research Methods*, vol. 22, no. 2, pp. 590–617, 2019.

[5] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?" in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[6] P. Pathirana, S. Senarath, D. Meedeniya, and S. Jayarathna, "Single-user 2D gaze estimation in retail environment using deep learning," in *Proc. of the 2nd International Conference on Advanced Research in Computing (ICARC)*, Sri Lanka, 2022.

[7] H. Tomas, M. Reyes, R. Dionido, M. Ty, J. Mirando, J. Casimiro, R. Atienza, and R. Guinto, "GOO: A dataset for gaze object prediction in retail environments," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, USA, 2021, pp. 3119–3127, doi: 10.1109/CVPRW53098.2021.00349.

[8] P. Pathirana, S. Senarath, D. Meedeniya, and S. Jayarathna, "Multi-user eye gaze estimation: A survey on deep learning-based approaches," *Expert Systems with Applications*, 2022, (to appear).

[9] E. Chong, Y. Wang, N. Ruiz, and J. M. Rehg, "Detecting attended visual targets in video," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, 2020, pp. 5395–5405, doi: 10.1109/CVPR42600.2020.00544.

[10] D. Lian, Z. Yu, and S. Gao, "Believe It or Not, We Know What You Are Looking At!" in *Proc. of the 14th Asian Conference on Computer Vision (ACCV)*, Perth, Australia, 2019, pp. 35–50.

[11] P. Pathirana and S. Senarath. Retail Gaze: Gaze Estimation in Retail Environment. Accessed Feb. 16, 2022. [Online]. Available: https://www.kaggle.com/dulanim/retailgaze

[12] H. Padmasiri, J. Shashirangana, D. Meedeniya, O. Rana, and C. Perera, "Automated license plate recognition for resource-constrained environments," *Sensors*, vol. 22, no. 4, 2022, doi: 10.3390/s22041434.