Single-User 2D Gaze Estimation in Retail Environment Using Deep Learning

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

Shashimal Senarath
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
shashimalsenarath.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 2nd International Conference on Advanced Research in Computing (ICARC) | 978-1-6654-0741-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICARC54489.2022.9754167

***Abstract*—Human gaze estimation is a widely used technique to observe human behavior. The rapid adaptation of deep learning techniques in gaze estimation has evolved human gaze estimation to many application domains. The retail industry is one domain with challenging unconstrained environmental conditions such as eye occlusion and personal calibration. This study presents a novel gaze estimation model for single-user 2D gaze estimation in a retail environment. Our novel architecture, inspired by the previous work in gaze following, models the scene and head feature and further utilizes a shifted grids technique to accurately predict a saliency map. Our results show that the model can effectively infer 2D gaze in a retail environment. We achieve state-of-the-art performance on Gaze On Objects (GOO) dataset. The obtained results have shown 25.2° angular error for gaze estimation. Furthermore, we provide a detailed analysis of the GOO dataset and comprehensively analyze the selected model feature extractor to support our results.**

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

I. Introduction

Human gaze estimation is one of the most frequently used techniques to observe human cognition and behavior. It is a widely studied field of area in application domains such as human-computer interaction, social behavior, medical health, business, and sports. [1]–[4]. Well-established image processing and computer vision-based related applications for object detection [5], face recognition [6] and human detection [7] have been addressed in the literature. However, eye gaze estimation research is still a trending area with the development of computer vision and deep learning techniques [8].

In the gaze estimation literature, multiple forms of gaze estimation such as gaze point estimation, gaze direction estimation, and gaze following have been studied broadly. However, the novel concept of gaze object prediction has not been extensively explored. Tomas et al. [9] have introduced this concept as the task of predicting the bounding box for a person's gazed-at object. The applications of gaze object prediction are mostly performed in unconstrained environments. The numerous variations in unconstrained environmental settings such as illumination, occlusion, head pose, subject count, subject distance variation make gaze object prediction a complex task. However, with the recent adaptation of deep learning techniques for gaze estimation, many promising approaches have been proposed to estimate gaze direction from images.

Recasens et al. [10] have made a significant breakthrough in this regard by introducing the concept of gaze following in the Convolutional Neural Network (CNN) based gaze estimation domain. Their work demonstrated the ability to recognize each person's attention target inside a single image using only image data. Chong et al. [11] have extended this approach to handling out-of-frame gaze targets using a multi-task learning approach. Another method motivated by the human gaze following behavior has been proposed by Lian et al. [12], which used multiple gaze direction fields of different scales to estimate the attention target robustly. However, most of these approaches have only considered front-head images [13]–[15], which is a significant limitation in a retail environment.

This paper presents a novel static CNN-based deep learning model to estimate the subject gaze in a retail environment using only 2D image data. Our goal is to robustly estimate human gaze using back-head images in the unconstrained retail environment. The proposed model is inspired by the previous work in [10], [14]. Our proposed model consists of three main parts namely saliency pathway, gaze pathway, and the shifted grids module. First, we use the gaze pathway module to generate a head feature map and an attention map from the extracted head image and its binary location map. Second, we use the saliency pathway to generate a scene feature map from the scene image, object channel, and head-binary location. Finally, the shifted grids module consisting of five shifted grids is used to produce the attention map robustly. For our task of 2D gaze estimation in a retail environment, we have used the Gaze On Objects (GOO) dataset [9], and

978-1-6654-0741-0/22/$31.00 ©2022 IEEE

GazeFollow dataset [10]. The proposed model outperforms all the benchmark baselines on the GOO dataset.

The paper is structured as follows. Section II explores the related literature. Section III describes the dataset used, design, and implementation details of the proposed approach. Section IV presents the obtained results together with a comparison of the existing studies and possible future research directions. Section V concludes the paper.

II. Related Work

Several studies have been addressed gaze estimation and gaze target prediction. In the gaze estimation literature, the concept of gaze following in 2D coordinates was first introduced by Recasens et al. [10], which is defined as identifying the object being looked at by the person given only the image. The authors have suggested a deep neural network-based gaze-following technique based on AlexNet and a new dataset, GazeFollow. Their dataset contains 122,143 images of 130,339 multi-users engaged in daily activities, with gaze location annotations inside the image. Moreover, they have proposed the novel shifted-grids approach to predict the gaze point by solving several overlapping classification problems. However, their work is not application-specific and mostly includes front-head images.

Following the work of Recasens et al. [10] multiple studies [11], [12], [14], have addressed the problems of handling out-of-frame gaze targets and detecting attention targets in the video. Chong et al. [11] have extended the GazeFollow dataset to include out-of-frame gaze target annotations. Chong et al. [14] have addressed the problem of dynamic attention in videos by introducing the VideoAttentionTarget dataset. They have proposed a Spatio-temporal architecture to infer time-varying attention targets. An interesting approach for gaze target prediction is described that is inspired by human behavior in gaze following in [12]. Their work has used multiple gaze direction fields of different scales to estimate the gaze direction of the person robustly. Our work is complementary to these studies. However, we focus more on back-head images in an application-specific environment.

Application-specific scenarios of gaze target prediction have been studied in [13], [15]. Sugano et al. [13] have presented AggreGaze, a method for predicting Spatio-temporal attention of people on public displays in an unconstrained environment. Their work highlights the importance of appearance-based methods with deep learning for multi-person gaze target prediction without personal calibration and special equipment. However, their work only considers front-head images which contrasts with our application. Furthermore, Bermejo et al. [15] have proposed a system for tracking the gaze of a retail shopper using 3D gaze estimation technology. Even though their work achieves better results in a retail environment, it requires personal calibration, which is a major limitation.

In another study, Tomas et al. [9] have presented the task of gaze object prediction that predicts the bounding box of a person's gazed object. They have further presented the GOO dataset consisting of a large-synthetic image dataset and a small real image dataset of people gazing at objects in a retail environment. This work that closely resembles our work has applied recent state-of-the-art gaze target estimation models to predict the gaze targets in a retail environment.

III. Design and Methodology

A. Dataset

We used GOO dataset for the proposed gaze estimation approach, and this section provides a comprehensive analysis of the GOO dataset. Fig. 1 shows statistics of the dataset. The majority of existing gaze estimation datasets include the pixel being looked at, not the boundaries of a specific item of interest. Tomas et al. [9] have introduced the task of gaze object prediction along with the GOO dataset for the retail environment to address this issue. GOO dataset contains images of shelves packed with 24 different product items, and each image includes a customer (subject) looking upon a product item. All objects in the image are annotated with their respective bounding boxes, classes, points, segmentation masks, and gaze points, and head locations are provided as existing datasets. There are two parts to this dataset: GOO-Real and GOO-Synth.

The GOO-Real dataset contains 6229 images of 100 people (32 female and 68 male), with each image consisting of shelves packed with 24 distinct product categories.

Furthermore, the dataset consists of 2450 images of the train set, 2146 images of the test set, and 1633 images of the validation set with 40%, 34%, and 26% as split ratios. Head position and gaze point distributions of the dataset are shown in Fig. 1. The head positions have accumulated to one place in both camera angles as shown in Fig. 1(a) and Fig. 1(b). Furthermore, It can be assumed as an optimal place for participants to look at all the objects in the scene.

Gaze points in camera angle 0 as shown in Fig. 1(c) are evenly distributed among all product items, but gaze points in camera angle one as shown in Fig. 1(d) have a significant bias for the right shelf. All gaze points in the dataset are on the product items, and it is an issue because, in retail environments, customers may look at other areas rather than product items. Gaze areas, except for product items and out-of-the-frame gaze, will improve the overall accuracy and robustness of the gaze estimation system. The distribution of the distance between head and gaze point has shown in Fig. 1(e) and Fig. 1(f) as a histogram (distance between head and gaze point calculated using image coordinate system). Small distances and large distances value count is small, and mid-range distances are significantly high in both camera angles (normally distributed). Camera angle 0 as shown in Fig. 1(g), the participants look at their right-hand and left-hand sides evenly, but in camera angle one as shown in Fig. 1(h), most of the participants look at their right-hand side. Thus, several gaze points in the camera angle one bias to the right shelf.

GOO-Synth contains 192000 training image data, and GOO-Synth was built using Unreal Engine. It contains synthetic images similar to GOO-Real retail environment scenes. Images are taken from 5 different camera angles (randomly selected

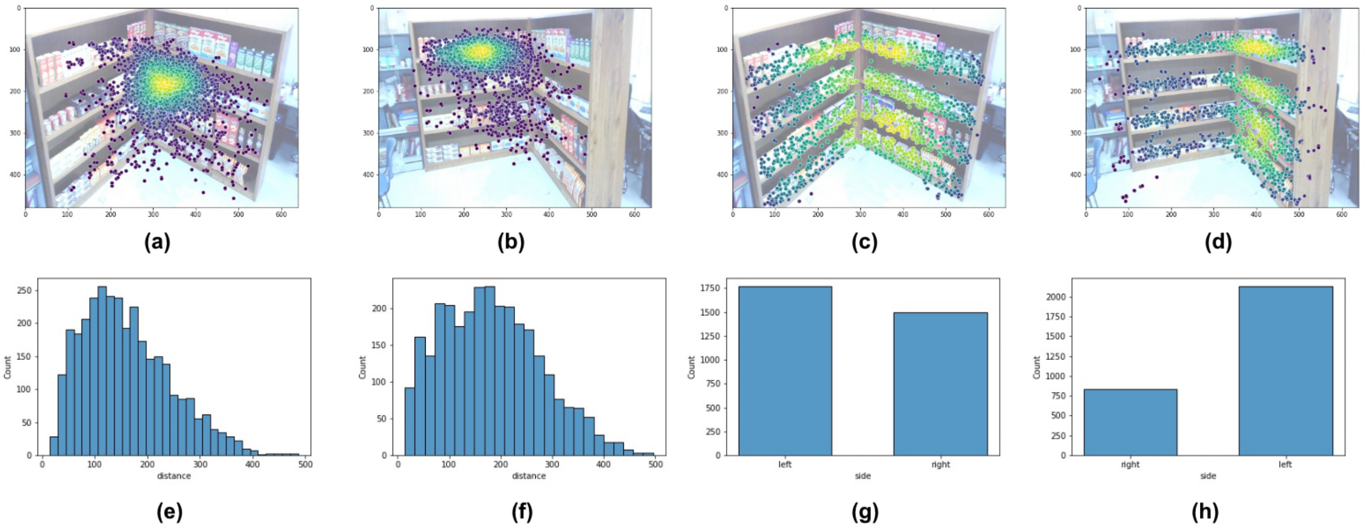

Fig. 1. Dataset Details: Distribution of (a), (b) Head positions. (c), (d) Gaze points. (e), (f) Distances between head and gaze point. (g), (h) Shelves' sides.

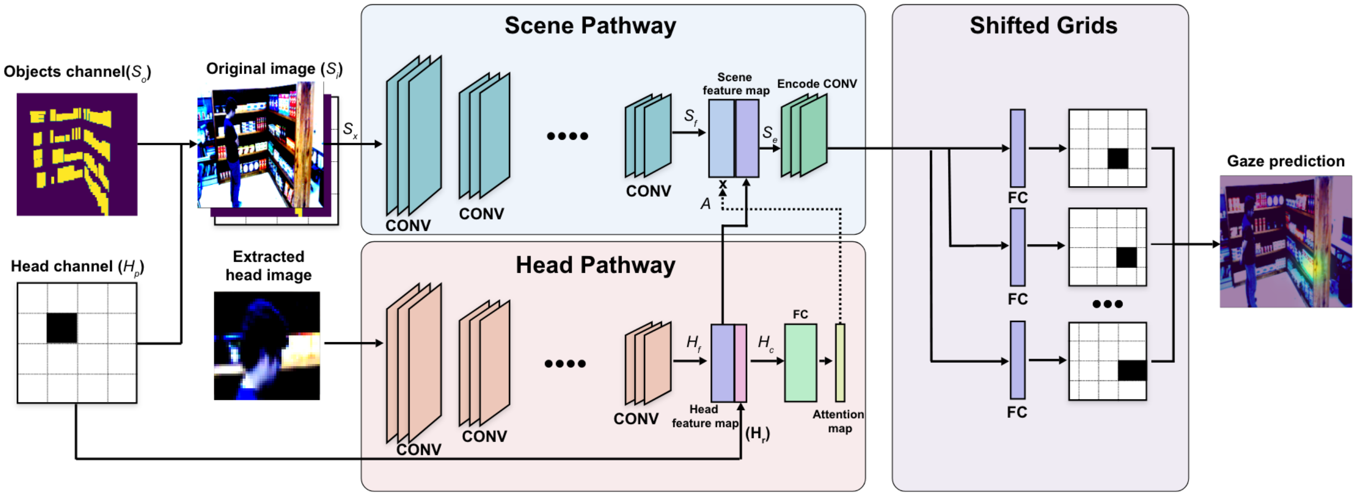

Fig. 2. Gaze Estimation Model Architecture with scene pathway, head pathway, and a shifted grids module.

from 50 different virtual camera angles), and one of 20 synthetic customer models interacts with the scene in each image. These human models had a wide range of skin tones, genders, physique types, and outfits. The product items are the same as GOO-Real product items, and each scene contains one of 38400 background environments.

B. Model Architecture

Our gaze estimation model architecture consists of three main components: scene pathway, head pathway, and a shifted grids classification module. The model architecture is depicted in Fig. 2.

Head Pathway: The head pathway computes the head feature map from the head image of the person in the scene. The convolution part of this head pathway is a pre-trained VGG-16 [16]. The head position channel (H_p), which is a binary representation of the head location with white pixels indicating the head bounding box and black pixels indicating the other area of the image. The head position channel is reduced to 14×14 using three max-pooling operations. This head feature map (H_f) is concatenated with the reduced head position channel (H_r) as shown in (1).

$$H_c = H_f \oplus H_r \tag{1}$$

A fully connected layer, which models the attention mechanism is then used to compute the attention map (A) using these concatenated features (H_c), and this method is influenced by Chong et al. [14].

Scene pathway: Scene pathway computes the scene feature map (S_f) by taking input (S_x), as the concatenation of the scene image (S_i), head position channel, and object channel (S_o) as shown in (2).

$$S_x = S_i \oplus S_o \oplus H_p \tag{2}$$

Few existing models [10], [14] have provided head position as a spatial reference, allowing the model to learn quicker, and we followed that method in our model. Apart from that, we found that providing gaze object bounding boxes help the model to learn faster and get more accurate gaze fixation from the scene. The gaze object channel is a binary image of product items boundaries, with white pixels representing object boundary boxes and black pixels representing the other area of the image. The convolution part of the scene pathway is also a pre-trained VGG-16 [16] with an additional convolution layer. Based on the attributes of the head, we applied an attention mechanism similar to [14] to pay greater attention to scene features that are more likely to be looked to. The computed scene feature map (S_f) was then multiplied with the attention map (A) computed by head pathway as shown in (3).

$$S_c = S_f \otimes A \tag{3}$$

The head feature map (H_f) is concatenated with the weighted scene feature map (S_c) as shown in (4),

$$S_e = S_c \oplus H_f \tag{4}$$

and this concatenated feature is encoded using convolution layers.

Shifted Grids: We use multi-model predictions to predict the fixation point that is introduced in [10]. They have formulated this prediction task as a classification task rather than a regression task, which naturally supports multi-model outputs. In this method, fixation location is quantized into a $N \times N$ grid, and the network classifies the input into one of the N^2 classes. When N is small, the prediction will suffer from poor precision, and selecting a significant N learning problem becomes more challenging. We have used the N value proposed by Recasense et al. [10] in our model. Their proposed shifted grids which predict overlapping outputs from the model, improved the confidence of the classification. Finally, we calculate the average of the shifted outputs to get the final prediction.

C. Implementation Details

We implemented our models in Pytorch and Pytorch Lighting frameworks. Scene image, cropped face image, head channel, and object channel are used as inputs to our model. Head channel and object channel are created using the head bounding box and the object bounding boxes of each frame. The scene image and the cropped face image are resized to 224×224 and normalized into the corresponding backbone of the model. The attention layer generates 7×7 spatial soft-attention weights. The output of the last convolution layer feeds into four fully connected layers, and each fully connected layer is the size of 699, 400, 200, 169, respectively. The fully connected layer's output goes through a Sigmoid activation and returns five shifted grids of size 5×5 each. We utilize backpropagation to train our model, and we employ a negative-log-likelihood loss for each shifted grid, averaging their losses. We use data augmentation such as random crops and color profiles. To prevent overfitting, we used certain patient values for early stopping. We trained the model with different backbones such as EfficientNet [17], ResNest [18], ResNet [19], VGG-16 [16] and selected the best-performed backbone by comparing each model's performance. First, our model was trained using the gazefollow dataset and then transfer learned using the GOO-Real dataset to optimize the model performance on GOO-Real dataset.

IV. Results and Analysis

A. Gaze Model Results

In this section we discuss the experimental results of our presented gaze estimation model. We have experimented with six CNN based architectures namely EfficientNet-b0 [17], EfficientNet-b6 [17], ResNest-50 [18], ResNest-101 [18], ResNet-18 [19], and VGG-16 [16]. Table I, presents a comparison of the performance metrics of the model for each backbone.

TABLE I
Results Comparison with Different backbones

Backbone	***AUC***	***Dist.***	***Ang.***
EfficientNet-b0	0.859	0.199	37.09
EfficientNet-b6	0.877	0.200	33.37
ResNest-50	0.911	0.163	31.56
ResNest-101	0.906	0.167	32.09
ResNet-18	0.887	0.189	34.82
VGG-16	0.909	0.163	30.16

We evaluate the performance of our gaze model based on the three performance measures: AUC, L2-distance (Dist.), and Angular error (Ang.). The AUC criteria proposed by Judd et al. [20] is used as the first performance metric. The AUC is defined as the area under the ROC curve where the saliency maps are thresholded by categorizing pixels as fixated and unfixed. The mean Euclidean distance between the ground-truth gaze annotation and the gaze prediction is defined as L2-distance. Angular error is defined as the angular difference between the ground truth gaze vector and the predicted gaze vector in gaze estimation literature. It can be seen that models with ResNest-50 and VGG-16 backbones are the best-performed models with the highest AUC, lowest L2-distance, and lowest angular error. We selected VGG-16 as our model backbone due to its lower angular error than ResNest-50.

Qualitative results of the model are presented in Fig. 3, where Fig. 3(a) shows correctly predicted images and Fig. 3(b) shows few incorrectly predicted images. The top incorrect prediction shows a scenario where the model predicts a point outside the shelf. This can be reduced by training the model

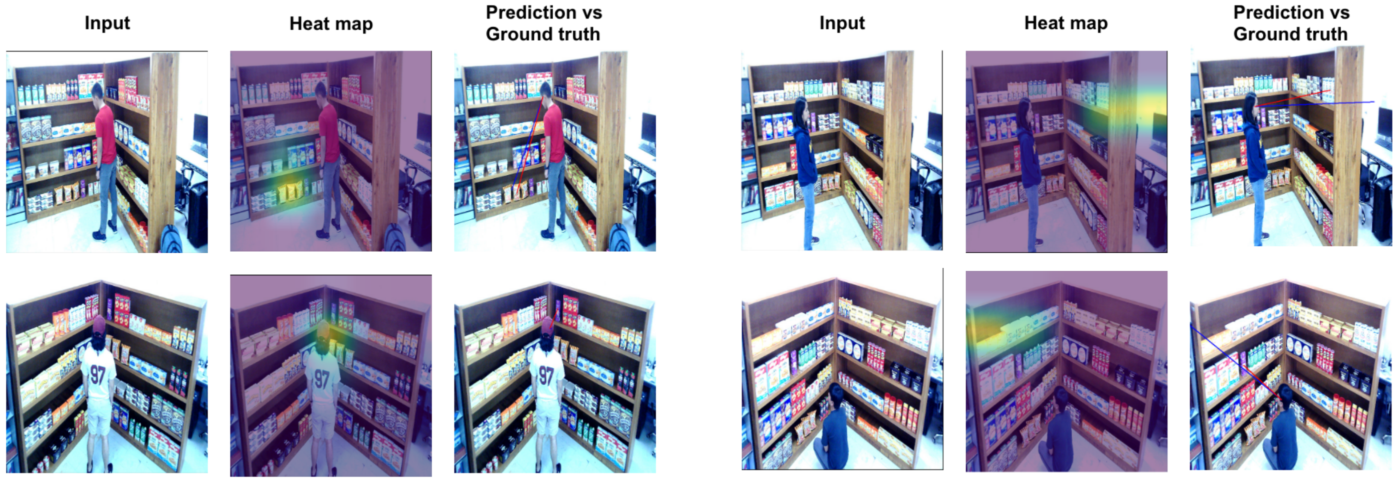

Fig. 3. Model Results

with out-of-the-shelf images and penalizing the model. The incorrect bottom prediction shows a scenario where the model cannot estimate the correct depth of the gaze point. This can be reduced by incorporating a depth channel of the scene into the model.

On the GazeFollow and GOO-Real datasets, we provide the benchmarks of our transfer learning algorithms. Experiments with feature extraction backbones, loss functions, and model optimizers are also discussed. The gaze model performance was measured separately for the training with the GOO-Real dataset as well as the subsequent transfer learning approach with the GazeFollow dataset. The results achieved with the optimal feature extraction backbone and other hyper-parameters on the GOO-Real test set are shown in Table II. The model achieved 33.512° Angular error when trained with the GOO-Real dataset for five epochs. The model achieved 25.224° Angular error when trained with GOO-Real dataset with GazeFollow pre-training. The proposed model showed the best performance when trained with the GazeFollow dataset until convergence and then transfer learned with the GOO-Real dataset for 13 and 25 epochs, respectively.

TABLE II
MODEL RESULTS WITH DIFFERENT DATASETS

Dataset	No. of Epochs	*AUC*	*Dist.*	*Ang.*
GOO-Real	5	0.897	0.175	33.512
GazeFollow + GOO-Real	13, 25	0.942	0.143	25.224

Furthermore, Fig. 4 shows the training learning curve for the model training with the GazeFollow dataset. Clear overfitting of the model is visible in the graph at epoch 13. Hence the model pre-training was early stopped at 13 epochs.

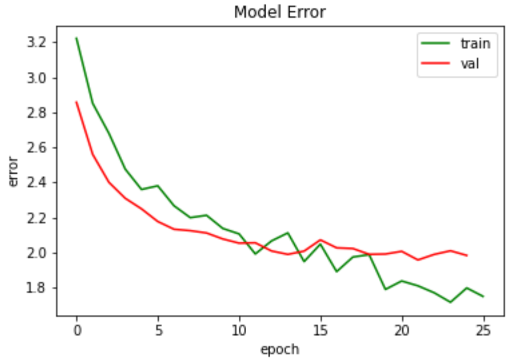

Fig. 4. Model Error

B. Comparison with Existing Studies

Table III, states a result comparison of the proposed approach with existing studies. We eliminate the studies that do not directly resemble our application domain and problem scope in this comparison. Hence we compare our work with the best results presented in [9] using the benchmark GOO dataset. It can be seen that our model surpasses the performance results of all other three models in the AUC, L2 distance, and Angular error criteria. We found that providing the object channel helped the model to learn faster and get more accurate gaze fixation from the scene. The hand-designed object channel feature helped the model to narrow down its gaze estimation point search space. Hence it increased the accuracy of the prediction. This is one of the main contribution of this study compared to the related work.

TABLE III
RESULTS COMPARISON WITH EXISTING STUDIES

Model	*AUC*	*Dist.*	*Ang.*
Recasense [10]	0.903	0.195	39.8
Lian [12]	0.890	0.168	32.6
Chong [14]	0.889	0.150	29.1
Our study	0.942	0.143	25.2

C. Future Research Directions

The presented novel deep learning model for gaze estimation in a retail environment using back-head images surpassed the existing benchmark baselines for AUC, L2 distance, and Angular error standard metrics. However, this research can be extended to improve the gaze estimation performance and the real-world applicability of the model.

It is highly advantageous to create an extended retail environment gaze dataset with out-of-frame gaze target annotations. This would improve the model's performance, and it is necessary to remove the bias in the dataset, hence reducing model outliers. The real-world scenario of retail gaze estimation requires efficient multi-user gaze estimation. The current model can be improved to predict multi-user gaze estimations with improved throughput. Furthermore, the model architecture can be improved as a Spatio-temporal architecture to gain the advantage of retail shoppers temporal nature gaze [21]. For these improvements, a video dataset annotated with multi-person gaze annotations is required.

V. Conclusion

We have presented a novel deep learning model for single-user 2D gaze estimation in a retail environment. Most of the existing studies have not addressed the product object boundaries that help the model to understand product items in the environment. We designed our deep learning model to specifically model the parameters in a retail object store and optimized it for back-head images. We improved the gaze estimation task using the gaze follow dataset rather than pre-training on the GOO-Synth dataset. Gaze follow dataset enhanced the face and scene feature extractors. The introduced model surpassed the existing benchmark AUC and Angular error baselines on the GOO dataset. Extending the dataset with out-of-frame gaze targets and estimating the gaze of multiple retail users at once can be seen as future directions.

References

[1] 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.

[2] G. E. Raptis, C. Katsini, M. Belk, C. Fidas, G. Samaras, and N. Avouris, "Using eye gaze data and visual activities to infer human cognitive styles: Method and feasibility studies," in *Proc. of the 25th Conference on User Modeling, Adaptation and Personalization*. New York, USA: ACM, 2017, p. 164–173, doi: 10.1145/3079628.3079690.

[3] X. Zhang, Y. Sugano, and A. Bulling, "Evaluation of appearance-based methods and implications for gaze-based applications," in *Proc. of the CHI Conference on Human Factors in Computing Systems*. New York, USA: ACM, 2019, p. 1–13, doi: 10.1145/3290605.3300646.

[4] S. De Silva, S. Dayarathna, G. Ariyarathne, D. Meedeniya, S. Jayarathna, A. M. P. 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.

[5] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, pp. 1–45, 2006.

[6] D. Meedeniya and A. Ratnaweera, "Enhanced face recognition through variation of principle component analysis (PCA)," in *Proceedings of International Conference on Industrial and Information Systems (ICIIS)*, Peradeniya, SriLanka, 2007, pp. 347–352, doi: 10.1109/ici-infs.2007.4579200.

[7] G. Gamage, I. Sudasingha, I. Perera, and D. Meedeniya, "Reinstating dlib correlation human trackers under occlusions in human detection based tracking," in *Proc. of the 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka, 2018, pp. 92–98, doi: 10.1109/ICTER.2018.8615551.

[8] D. Cazzato, M. Leo, C. Distante, and H. Voos, "When i look into your eyes: A survey on computer vision contributions for human gaze estimation and tracking," *Sensors (Switzerland)*, vol. 20, no. 13, pp. 1–42, 2020, doi: 10.3390/s20133739.

[9] 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.

[10] 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.

[11] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. M. Rehg, "Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency," in *Proc. of the 15th European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 397–412.

[12] 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.

[13] Y. Sugano, X. Zhang, and A. Bulling, "Aggregaze: Collective estimation of audience attention on public displays," in *Proc. of the 29th Annual Symposium on User Interface Software and Technology (UIST)*. New York, USA: ACM, 2016, p. 821–831, doi: 10.1145/2984511.2984536.

[14] 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.

[15] C. Bermejo, D. Chatzopoulos, and P. Hui, "Eyeshopper: Estimating shoppers' gaze using cctv cameras," in *Proc. of the 28th ACM International Conference on Multimedia*. New York, USA: ACM, 2020, p. 2765–2774, doi: 10.1145/3394171.3413683.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, USA, 2015, pp. 1–14.

[17] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. of the 36th International Conference on Machine Learning*, California, USA, 2019, pp. 6105–6114.

[18] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z.-L. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "Resnest: Split-attention networks," *ArXiv*, vol. abs/2004.08955, 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 2016, pp. 770–778.

[20] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. of the IEEE 12th International Conference on Computer Vision*, Kyoto, Japan, 2009, pp. 2106–2113.

[21] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, 2019, pp. 6911–6920.