Depth-Based Gaze Target Detection in Video

Yarden Bakish
yardenbakish@mail.tau.ac.il

Abstract

The objective of this project is to construct a novel model for the purpose of gaze detection in videos. This Project proposes an innovative model architecture via taking advantage of distinctive solutions found in existing literature while notably extending it to include LSTM layers and depth-estimation data to predict where people are looking. To the best of my knowledge, it the only project to incorporate temporal and 3D spatial features for gaze detection in videos. Results show that incorporating depth information achieves competitive and consistent results across multiple evaluation metrics.

1 Introduction

Gaze Detection is the task of following people's gaze in a scene and inferring where they are looking. Given track of a scene, depth map, head location and eyes' locations for each person (bounding boxes), the proposed model predicts where they are looking, including identifying out-of-frame targets and locating inside-frame targets (see Figure 1).

Most research have been predominantly geared towards physically constrained applications such as smartphone gaze tracking due to the lack of sufficiently large-annotated datasets. Furthermore, existing work on unconstrained gaze detection focus on 2D gaze and 2D saliency but fail to exploit 3D contexts.

Inspired by state-of-the-art papers [Fang et al., 2021, Lian et al., 2018, Chong et al., 2020], this project utilizes distinctive solutions found in recent literature as a leverage point to compete with familiar benchmarks, as described below:

- Encoding depth channel features in the scene, thus better reducing loss derived from candidate objects at different depths existing along the subject's planar gaze direction.
- Learn dependency between eyeball orientation and head pose for better gaze direction estimation and coping with fixation inconsistency (e.g., facing forward but looking downward).
- Bidirectional ConvLSTM layers which provide a means of modeling sequences where the output for one element is dependent on spatial features of both past and future inputs.
- Complex interaction between the head and scene feature maps.

Based on the above solutions, the following approach was adopted for gaze inference in 3d space from video data (see figure 2)):

1. Head and eyes crops from each image, as well as the entire image and its depth map, are individually processed by pre-trained convolutional neural networks (backbone), which produces high-level features.
2. Subjects' field of view in an image (without considering the scene contents) is encoded via features extracted from head orientation, depth image and head location which are fused and processed to generate an attention map.
3. The attention map is multiplied with image scene features. Features extracted from the head orientation and depth map are passed through a conv network for Field of View feature map extraction.
4. Field of View feature map is fused with the weighted scene feature map and fed to subsequent Conv and LSTM blocks ultimately producing two outputs for the following tasks:

identifying out-of-frame targets as a binary classification problem and locating in-frame targets as a heatmap regression problem.

The project's contributions are summarized below:

- Extend existing work and introduce a novel architecture that implicitly embodies the person's field of view regulated by temporal and depth information for gaze detection in videos. To the best of my knowledge, this is the only paper which experimented with combining LSTMs and depth maps for gaze detection in videos.
- Demonstrate that the proposed method achieves relatively high accuracy against other recent gazing benchmarks.

2 Related Works

2.1 2D Gaze Target Prediction

Recent relevant works [Recasens* et al., 2015, Lian et al., 2018, Chong et al., 2020, and Chong et al., 2018] on the field of unconstrained gaze detection typically develop a two-branch-based model where one branch is for gaze direction prediction and the other for saliency map of the scene. The two are then fused to infer gaze target. These works are based on 2D visual cues and lack scene depth understanding and depth-channel gaze supervision, resulting in ambiguity in fore/background points.

2.2 3D Gaze Target Prediction

Existing methods [Fang et al., 2021,Senarath et al., 2022,Al-Hindawi et al., 2022] which incorporated scene depth understanding in 2D gaze target detection, all relied on the state-of-the-art model implemented by Fang et al., 2021. Both Fang et al., 2021 and Lian et al., 2018 have used an independently trained gaze direction estimation model to predict head pose vector (yaw and pitch) and generated an estimated field of view of the subjects. This field of view is concatenated with the original image and passed through a backbone network to predict gaze target. Specifically, Fang et al., 2021 generated the field of view by using depth-map data and the gaze direction estimation and thereafter analyzed the 3D geometry of the scene along the gaze direction. This calculation is done explicitly without any learnable parameters and strictly relies upon a general estimated field of view.

Key differences between (Fang et al., 2021, Lian et al., 2018) and my approach are as follows:

- Field of view estimation will be implicitly incorporated through feature extraction, allowing to deal with diverse fields of view. Encoding the field of view will be carried out with a Fully Connected network instead of geometric calculations (further detailed in section 3.2).
- No component of my model will be trained separately

3 Approach

Following the notations of Fang et al., 2021, I aim to solve the following optimization problem:
$L = \gamma_1 \cdot L_{bce} + \gamma_2 \cdot L_{reg}$

Where L_{reg} is the heatmap loss computed with pixel-level MSE loss when the target is in frame per ground truth and L_{bce} is the In-frame loss which is computed with binary cross entropy loss, thus optimizing the model by its prediction of gaze target pixels and binary classification to whether the target is out-of-frame.

This section presents the architecture of the model, as shown in Figure 2. The overall workflow is comprised of Field of View estimation, scene feature extraction and a gaze target detection.

3.1 Dataset Preprocessing

Following Fang et al., 2021, data preprocessing was carried out algorithmically using state-of-the-art pretrained estimators for the entire dataset, for it to include the following:

- *Eyes bounding boxes for each subject in each Image* [1]– Extending the annotation labels of the dataset with each annotated person's left and right eye bounding boxes if the person is facing the camera. That way the model can deal with eye occlusion and the possible large gap between eye orientation and head orientation. To make less false positives, I used a pretrained head pose estimator and considered self-occluded eyes when extreme poses were in order.

[1]Implemented using dlib library http://dlib.net/python/

(a) (b) (c) (d)

Figure 1: The blue line is the ground truth, and the green line is the final prediction prediction. (a) and (b) indicate success, where in (a) the ground truth and predicted gaze is almost the same, and in (b), the model was able to label the target gaze as out-of-frame. Failure cases can be seen at (c) and (d). The model achieves lower accuracy when eyes are occluded as in (d).

- *Priori depth map estimator for each image (Ranftl et al., 2022)* – depth maps are provided by a well generalized model which is trained across diverse datasets and 3D movies.

3.2 Field of View Estimation Pathway

Field of View ("FOV") pathway takes head image, head position and depth map as inputs for gaze attention map.

feature vectors are extracted from the cropped head patch and eyes' patches of the subject of interest in the image. If eyes are invisible, the gaze will be coarsely approximated by the head pose only. The cropped head patch is fed to a ResNet-50, and the cropped eye patches are fed to two parallel ResNet-18 separately to generate feature vectors. These vectors are concatenated and passed to a Fully connected layer to produce high-level features of gaze direction.

Depth map is concatenated with a binary image of the head (black pixels designating head bounding box) and fed to a ResNet-50, producing high-level feature vector of the depth map, constrained by the location of the subject in the image.

Head location of the subject is encoded using a MaxPool layer on a binary image of the head and then flattened.

Depth map feature, gaze direction feature and head location feature are concatenated and fed to a one-hidden-layer FC for a final output of 7x7 spatial soft-attention weights. Similar to the attention map generated by Fang et al., 2021, This layer filters candidate targets over depth and field of view simultaneously.

Features extracted from the head orientation and depth map are passed through a Conv network for Field of View feature map extraction.

3.3 Scene Feature Extractor

Scene feature extraction is done by computing feature map from the scene image with a ResNet-50.

Scene feature map is then multiplied by the spatial soft-attention weights generated from the Field of View Estimation branch. This enables the model to learn to pay more attention to the scene features that are more likely to be attended to, based on the properties of the head, the depth difference between subject and target, and the saliency of the target.

3.4 Gaze Target Detection

Following [Fang et al., 2021,Chong et al., 2020], FOV feature map and weighted scene feature map are concatenated and passed through a backbone to perform feature extraction. They are shared across the Binary Classification Head and the Heatmap Regression Head.

In detail, the Binary Classification Head consists of two convolutional layers followed by a fully connected layer to classify whether the target is in-frame or not. For the Heatmap Regression Head, two convolutional layers are applied followed by two bidirectional Conv-LSTM [2], and four deconvolutional layers to predict where the target person is looking and output a full-sized heatmap. The point of the maximum value in this heatmap is considered the predicted gaze point.

[2]https://github.com/kamo-naoyuki/pytorch$_{c}onvolutional_{r}nn$

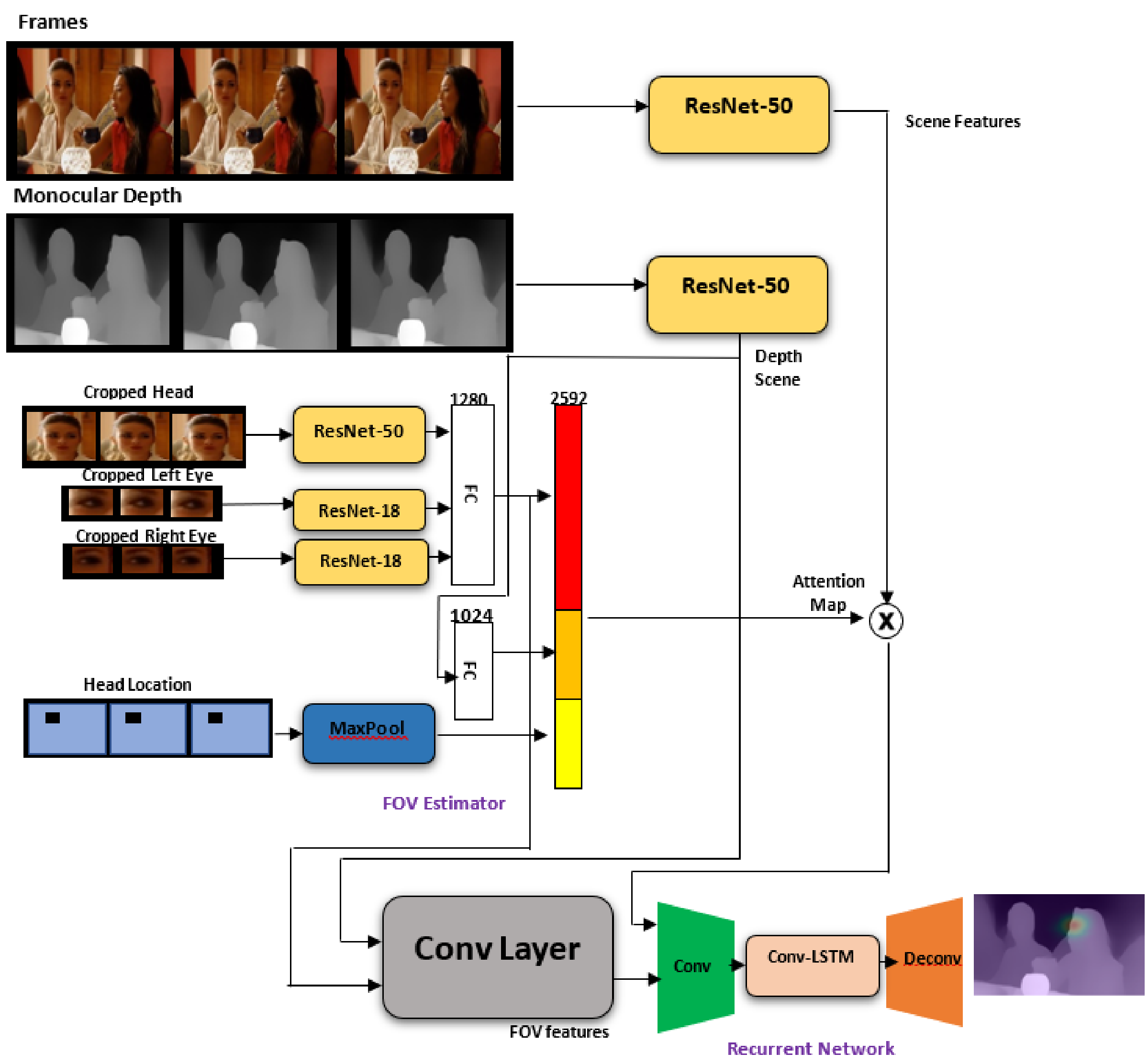

Figure 2: Architecture of the model.

The heatmap generated by the recurrent network is modulated by the binary classification Head, which quantifies whether the gaze target is in-frame. The modulation is performed by an element-wise subtraction from the full-sized heatmap from a scalar (the output of the binary classification head). This yields the final heatmap which quantifies the location and intensity of the predicted attention target in the frame.

Like most existing methods, the heatmap of ground truth gaze point is generated by centering a Gaussian kernel at the position of gaze point.

3.5 Implementation

The model is implemented on PyTorch.

For FOV feature extraction, the two detected eye patches will be cropped from the head image. If not detected, they will be replaced with black images.

The images, depth maps, and head crops were all resized to 224x224. Eye patches are resized to 36X60. The heatmap regression outputs a heatmap with size 64×64.

Two models are trained in this project, a single-image gaze prediction model, denoted by *Static Model*, and a full model, *"Complete Model"*. The recurrent network is removed from The *Static Model*. I first trained the *Static Model* on the *GazeFollow* dataset until convergence. Thereafter, the complete model is finetuned on the *VideoAttentionTarget* dataset, while freezing the layers up to the generation of the soft spatial weights to prevent overfitting. Evaluation for both models and further details on the datasets are presetned at section 4.

The *Static Model* is optimized by Adam, with learning rate of 0.00025 and batch size of 48. The *Complete Model* is optimized with Adam, with learning rate of 0.00005 and batch size of 8.

Random flip, color jitter, and crop augmentations were used as a mean of regularization during training.

The Scene and Depth Map ResNet-50 networks are initialized with CNN for scene recognition, and the Head Conv with CNN for gaze estimation.

4 Experiments

Please see the GitHub repository for the full results of the experiments, as well as Jupyter notebooks. It can be accessed at:
https://github.com/YardenBakish/Deep-Learning-Workshop
Colab Demo:
https://colab.research.google.com/drive/14ua6sTa-IgaPJZPLxqAvN3LhQx2SmI1J#scrollTo=G6PLExr7G728

4.1 Datasets

The *GazeFollow* (Recasens* et al., 2015) and *VideoAttentionTarget* (Chong et al., 2020) datasets were employed to evaluate my proposed method. The *GazeFollow* dataset, includes 122,143 images, with 130,339 annotations of head locations and corresponding gaze points. The *VideoAttentionTarget* dataset consists of 1,331 video clips collected from various sources on YouTube.

4.2 Evaluation Metrics

The following evaluation metrices were adopted for the *Static Model*:

- *AUC* – Each cell in the spatially discretized image is classified as gaze target or not. The ground truth comes from thresholding a Gaussian confidence mask centered at the human annotator's target location. The final heatmap provides the prediction confidence score which is evaluated at different thresholds in the ROC curve. The area under curve of this ROC curve is reported.
- *Average Distance* – The average Euclidean distance between predicted gaze point and the ground truth annotations.
- *Minimum Distance* – The minimum Euclidean distance between predicted gaze point and all ground truth annotations.

The following evaluation metrices were adopted for the *Complete Model*:

- *AUC* – detailed above
- *Average Distance* – detailed above.
- *Out of frame AP* – average precision to assess the accuracy of out-of-frame identifying.

AUC and Distance are computed whenever there is an in-frame ground truth gaze target.

Experimental results are summarized in Table 1 and Table 2. The following findings are as follows:

- Competitive results in all evaluation metrices, however the model is not SOTA, as results did not managed to surpass any of the results presented by Chong et al., 2020, or all of the results presented by Chong et al., 2020.
- Results demonstrate that my method outperformed Chong et al., 2020 for AP, thus surpassing the second-best competitor on both datasets and assuring depth-channel supervision.
- Results demonstrate a setback for AUC. A potential reason lies in the idea that more weight went to distinguish between objects in different depths and accurate 3D gaze from distinguishing two or more meaningful objects close together

5 Discussion

5.1 Limitations

The proposed method in this paper which relies on implicit Field of View generation still performs worse than existing state-of-the-art method which relied on explicitly generate subjects' field of view. It can be inferred that while having a noticeable

Method	AUC ↑	Avg Dist ↓	Min Dist ↓
Recasens* et al., 2015	0.878	0.190	0.113
Chong et al., 2018	0.896	0.187	0.112
Lian et al., 2018	0.906	0.145	0.081
Chong et al., 2020	0.921	0.137	0.077
Fang et al., 2021	0.922	0.124	0.067
Workshop	0.912	0.141	0.099

Table 1: Evaluation on the *GazeFollow* dataset

Method	AUC ↑	Dist ↓	AP ↑
Chong et al., 2020	0.860	0.134	0.853
Fang et al., 2021	0.905	0.108	0.896
Workshop	0.833	0.145	0.877

Table 2: Evaluation on the *VideoAttentionTarget* dataset

effect, implicit Field of view generated by conv layers still did not surpass the attention map generated by Chong et al., 2020

In addition, as mentioned, pretrained estimators for eyes' bounding boxes of subjects of interest in images for a refined gaze estimation. Although several steps were taken to ensure a small number of false positives (e.g. bounding boxes not aligned properly on subjects' eyes), the estimator still had some noisy results. In addition, although the depth map generator is considered state-of-the-art, it still sometimes hard to infer the depth of scene with the monocular images generated by the estimator. Monocular depth estimation is an ill-posed problem for a single RGB image in general.

5.2 Future Work

The following issues could be addressed in future research:

- *AUC improvement* - while this work provided competitive results for AP metric, it had a setback with AUC. I believe performance for AUC was reduced due to it is possible to come up with a modified architecture based on what I have provided in this project, to strengthen the representation power of this model, keeping high performance for both of these metrics.
- *Object-channel Papers* - Few papers have experimented with gaze detection in retail environments [Senarath et al., 2022, Tomas et al., 2021] and specific objects gaze detection. I believe that the implicit gaze-based architecture proposed in this project, can be used as a solid infrastructure for incorporation of more channels' supervision.

5.3 Conclusion

In this project, I proposed an innovative model architecture via taking advantage of distinctive solutions found in existing literature while notably extending it to include LSTM layers and depth-estimation data for gaze detection in videos.

Extensive evaluations demonstrated that the proposed method performs favorably against existing approaches when identifying out-of-frame targets (binary classification) with a trade-off over identifying in-frame targets (heatmap regression).

References

Ahmed Al-Hindawi, Marcela P Vizcaychipi, and Yiannis Demiris. 2022. What is the patient looking at? robust gaze-scene intersection under free-viewing conditions. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2430–2434.

Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. 2018. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Computer Vision – ECCV 2018 - 15th European Conference, 2018, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture

Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 397–412, Germany. Springer. Funding Information: This study was funded in part by the Simons Foundation under Funding Information: This study was funded in part by the Simons Foundation under grant 247332. Publisher Copyright: © 2018, Springer Nature Switzerland AG.; 15th European Conference on Computer Vision, ECCV 2018 ; Conference date: 08-09-2018 Through 14-09-2018.

Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. 2020. Detecting attended visual targets in video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. 2021. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11390–11399.

Dongze Lian, Zehao Yu, and Shenghua Gao. 2018. Believe it or not, we know what you are looking at! In *ACCV*.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2022. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637.

Adria Recasens*, Aditya Khosla*, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*. * indicates equal contribution.

Shashimal Senarath, Primesh Pathirana, Dulani Meedeniya, and Sampath Jayarathna. 2022. Customer gaze estimation in retail using deep learning. *IEEE Access*, 10:64904–64919.

Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard Guinto. 2021. Goo: A dataset for gaze object prediction in retail environments. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3119–3127.