3DG: Better 3D gaze estimation with the help of 3D scene information

Semester Project Proposal
Supervised by: Xi Wang (xi.wang@inf.ethz.ch)
September 23, 2022

BIN YANG

I. INTRODUCTION

From the sociological aspect, eye gaze is recognized as one of the most important communicative way in two-person interaction.[14] However, it's application has been already extended into wider range of areas in the past few decades, such as human-computer interaction[15], virtual reality[16] and robotics[11]. To automatically track and predict where a person exactly see, known as gaze estimation, various deep-learning based methods are proposed and explored.[21], [2], [3]

In the initial stage of previous work, most of researches focused on the 2D gaze tracking techniques, which map the movement of the sight on the 2D screen for the human-computer interaction, or simply detect the 2D gaze positions.[10] In the more recent approaches, 3D orientation of eye gaze is estimated for further applications, however, only using a 2D input reveals some limitation such as scale ambiguity due to the lack of depth information. Therefore, 3D contents are also encouraged to use to increase the accuracy of the prediction.[3]

In this project, our goal is to develop a better 3D gaze estimation framework by finding some approach about how to apply 3D scene information to improve the estimated 3D gaze locations. The input dataset we will apply consists of different 2D images taken by the front-view camera. Primarily, there exists various representations of the 3D scene information, e.g. 3D point clouds, multi-view frames or even those with additional category- or object-level semantic mapping (seeing Figure 1). However, the depth map comes to first since the input dataset still remains at 2D. It is the first prior work to study how depth information could be accurately extracted from a front-view image and corresponding 3D scene. Based on that, we could also derive other annotations like semantic label, object-based bounding boxes. In the mean time, a benchmark study is required for reviewing how previous work regarding to gaze-related dataset recorded the ground truth, especially for images from the front-view camera, so that we could get a reliable standard reference for further refinement by recording custom dataset with our own device. Then, the architecture of the deep-learning model needs to be extended and suited for different inputs. It is worth mentioning that different experiments with different representations can be done for finding the best choice or combination that results in the highest accuracy of estimated 3D gaze positions.

II. RELATED WORKS

A. Datasets

There already exists a bunch of datasets available for solving the eye gaze estimation task. However, they do not specifically have the same structures. For example, MagicEyes[20] includes 80,000 images from 587 subjects and their ground truth label is the semantic map of human's eye ball. EYEDIAP[8] collected images containing object's faces captured by RGB-D camera and its ground truth label is the 3D location of visual target.

B. Depth map

A depth map is an image or image channel that contains information relating to the distance of the surfacs of scene objects from a viewpoint. The depth information contributes to scene understanding of 3D real world including segmentation of objects, perspective projection and augumented reality. Except estimating depth using a neural network, depth map can also be created in a traditional way, for instance, a depth map is constructed from a 3D point cloud given the camera parameters, camera's Field of View and position in the world frame in [4] and [13]

C. 3D Gaze estimation

Conventional methods for 3D gaze estimation task can be generally categorized in model-based[19] and appearance-based approaches[21]. For instance, model-based method intends to reconstruct the 3D eyeball models and estimate gaze direction with the help of geometric eye features. The appearance-based methods extracts the eye feature from 2D images and use a regression model or deeper neural network to predict the eye gaze. In contrast to the model-based approach, it does not require dedicated devices with capturing eye movements precisely. However, most of reference researches still focused on 2D visual cues but pursuing the accurate 3D estimated results.

D. Understanding of 3D scene in gaze-related task

As mentioned above, the main purpose of the project is to improve the estimation of 3D gaze locations. Conventional methods like FAZE-Net[12] and L2CS-Net[1] are only focusing on 2D visual cues. As motivated by the limitation of 2D inputs, it is expected that understanding and introducing the 3D scene data can effectively optimize the performance of deep-learning model. According to some research work, ESCNet[3] proposed to explicitly reconstruct the given scene with a 3D point cloud from estimated depth maps and reference objects (active object that the person might look at). Under assumption that the person can not see through the occuler, the scene geometry is simply represented by some front-most points w.r.t the person. The prediction of gaze target is further incorporated with 3D gaze and preprocessed scene contextual cues. Their results already showcased some advantages over most of existing approaches.

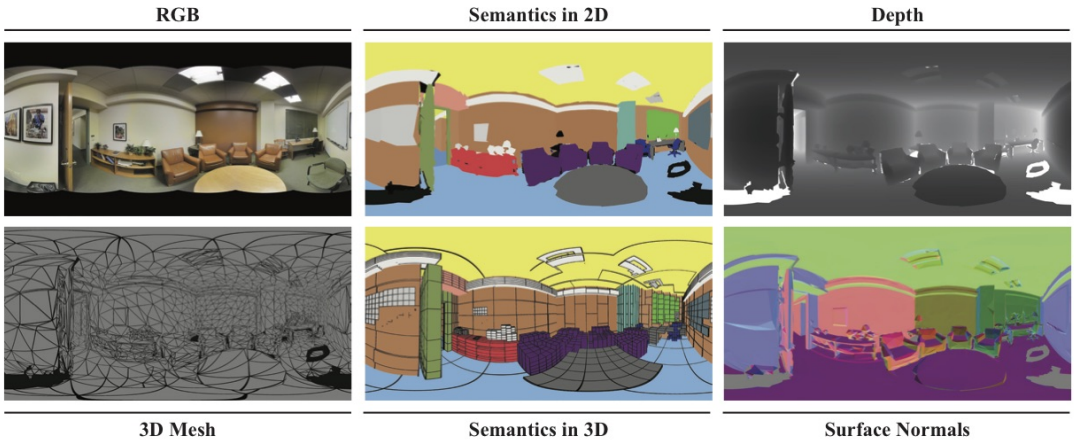


Fig. 1: Main representations of a scene in 2D and 3D frame

III. Project Description

An overview of project pipeline is given in Figure2. We will get into details of each part in the following subsections.

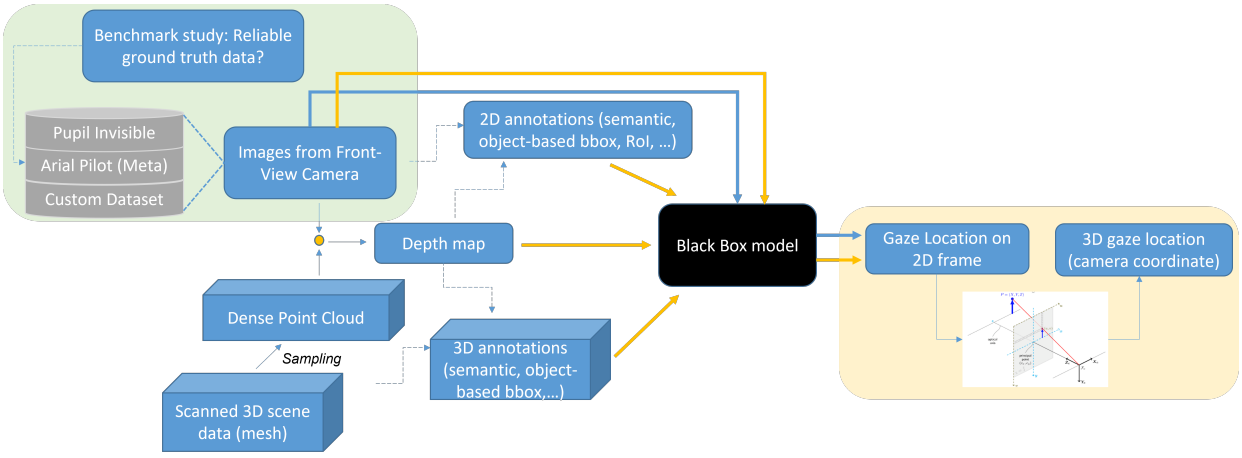


Fig. 2: Overview of the project pipeline

A. Datasets

It is always costly to create a reliable large-scale dataset to develop different approaches for some task. Eye gaze is one of those fundamental tasks. As mentioned above, there exists even different kinds of datasets regarding to eye gaze estimation task. Specifically, we will focus on the image taken by a front-view camera. As an illustration, the Figure3 shows a typical example. With the known 3D camera paramters, the 2D gaze center can be projected onto the 3D frame and convert to the 3D location which is the final output of our project (seeing the yellow part in the Figure2) and it will be estimated and evaluated by different approaches. To reach that, there are some open-source datasets collected by other institutions available, e.g. the Aria Pilot Dataset from Meta and Eye tracking data from Pupil Invisible (two-eye cameras), in addition, their estimated results can be used for the refinement of our outcome, but the 3D scene data corresponding to the front view is not particularly provided. From that reason, we have to use our own device to record the custom dataset and scan the environment to produce the 3D scene data. (seeing the green part in the figure2) Besides, a benchmark study is required for preparing the datasets. It helps to learn a relible way to label the ground truth data comprehensively when collecting data with our own devices.

Fig. 3: An example of front-view image with the calibrated gaze center

B. Front-view image and 3D scene

After the data preparation, depth information needs to be extracted with the help of 3D scene data. Given the 3D scene data and parameters of the front-view camera, it is straightforward to build a perspective projection model. As introduced and described in [4], a depth map from a certain point of view can be constructed if the camera position and field of view are fixed. However, the quality of 3D scene data could affect the accuracy of depth completion, because we only use a mobile device and manually scan the surroundings while most of related researches focused on completing a depth map using the Lidar-based point cloud with higher resolution.

C. Extraction of other features

Once the depth map is extracted, we could use the original input image including 3D scene data and depth map to derive other most common features, e.g. semantic labels, bounding boxes and region of interests. Inspired by that 3D scene data is large-scale and human has inherently some intention to interact with specific objects and regions, two approaches are approached from the research work at the initial stage. The figure4 shows the pipeline of how two methods work.

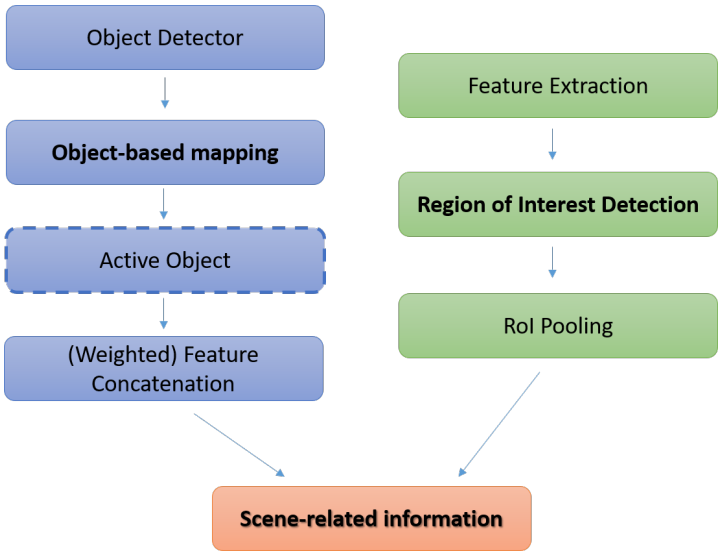


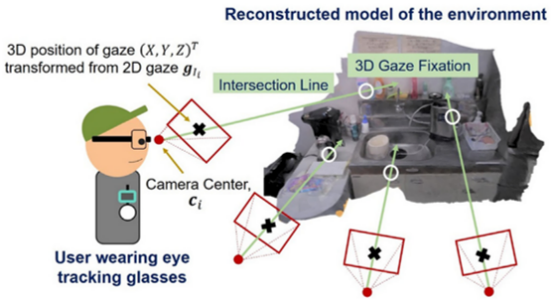
Fig. 4: Proposed Approach for the project

In general, we will try two approaches in the project. They have the same intention to extract the most important information that could represent the 3D scene for eye gaze estimation.

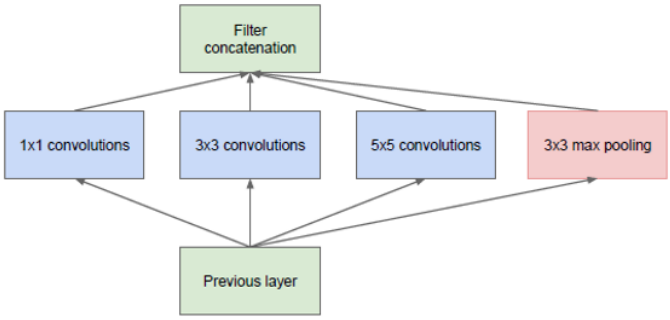
1) Object-based feature extraction: For the object-based approach, it firstly detects objects in the scene and a 3D bounding box is predicted and associated to each object. As mentioned above, a person might pay more attention to one specific object than any other in the environment. So it might be important to take the fact into consideration. It will be a challenging part in the project that a mathematical implementation is needed to compute the interaction between human's eye gaze and multiple objects. Some recent work has proposed related formulation, for instance, [6] introduced an active score to track the dynamics of the scene to recognize the continuous interacted object. The parameter shows how the person is looking at the object. What's more, it is also possible to visualize and record the user's attention on objects in 3D environment. [18] illustrates a model-based methodology that records the user's 3D gaze fixation by calculating and tracking the intersection point between 3D eye gaze ray and reconstructed 3D scene.

After features of objects are extracted, it is expected that those features need to be fused in some way for each 3D scene. The reason is that scene-related feature is required to have the unique fixed dimension for further learning-based task for eye gaze estimation, however, the number of objects can be random in different scenes. This is another bottleneck since we did not get actual related work regarding to the topic yet. Nonetheless, a bunch of feature fusion methods have been proposed and widely used in object detection, so it might be a good start that we could do in the similar way: for example, GoogLeNet[17] uses an Inception module to extract a multi-scale feature, in other words, it utilizes convolutional layers with multiple size to extract features at multiple scale and merge them. The features with different receptive fields are aggregated and a large amount of experiments have proved that such an approach can effectively increases the network's robustness to scale.[5] Besides, cross-attention module yielded in [9] for object detection and tracking could also be a feasible option. The module aims at using attention masks from one modality (one object) to highlight the extracted features in another modality (another object). It could not only fuse the features but also contribute to learn the spatial correspondence to derive better alignment of important details from different objects in the scene.

2) Region-based feature extraction: For the region-based methods, it is more intuitive and simple than object-based one to go for. Here we mainly refer to the prior work of Fast R-CNN[7]. The unsupervised network is a boosted version of R-CNN that effectively alleviates the issue of heavy computation of proposal regions and saved a huge amount of cost. That could also play a key role in this project because the 3D scene is a large-scale data which has the high-level memory demand. As shown in figure 6, the image is fed into a CNN to get the feature map. From the feature map, we could identify the region of proposals and warp them into multiple squares. Most Importantly, all those region-based features would be reshaped into a fixed size by RoI pooling so that it could be



(a) Scheme of 3D gaze determination on the environmental model[18]



(b) Inception module in GoogleNet[17]

further processed in classification and regression tasks. Based on the knowledge, we could also extract the region of interest from a feature map (for example encoded by a 3D CNN) and refine the proposals by learning a Fast R-CNN network. Finally, we could add a interface instantly after the RoI pooling and regard the output as the feature of 3D scene.

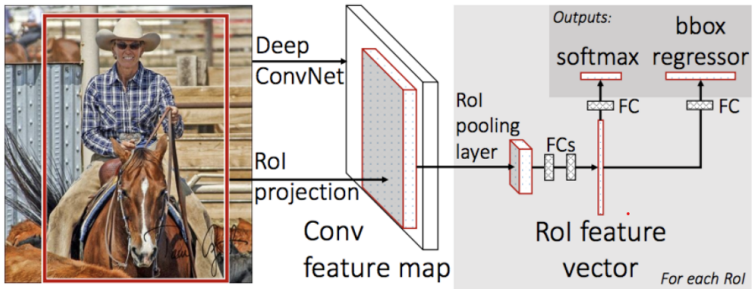


Fig. 6: Overview of Fast R-CNN[7]

D. Black Box Model for Eye gaze estimation

Another important work of the project is to build the network that could use the scene-related information to estimate the 3D eye gaze (Also for deriving the baseline without use of 3D scene). Most of related researches are CNN-based which means that feature extraction is the most critical in those methods.[21] A typical example of the deep-learning model for gaze estimation is shown in Figure 7. The network uses both eye images and face images as input. The advantage is that face images contain the head pose information that could help to improve the robustness to head motion, however, the face image also has the redundant information that might negatively affect the estimation results. In order to filter it out, the network introduces a spatial weighting mechanism to efficiently encode the location of the face into a standard CNN architecture (seeing the last row of the network).

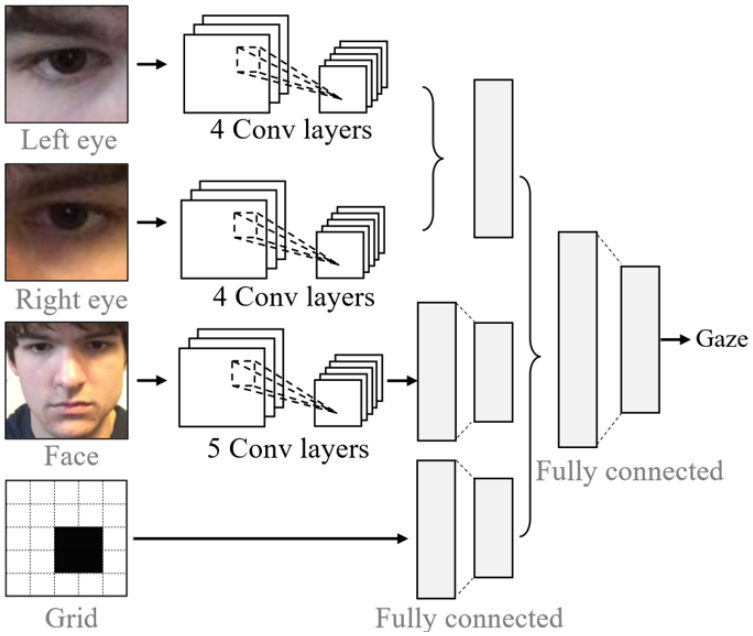


Fig. 7: Example of CNN-based approach for gaze estimation [22]

Similarly, we could also use such a branched architecture to take the advantage of features from different inputs (front-view image, depth, semantic, ...). In general, we need to study how 3D scene information could help to improve the estimation results, so the comparison experiments will be done during the progress of projects.

IV. Work Package and timeline

Work Package	Duration(weeks)
Literature review for benchmark study	2
Depth map construction	1
Implementation of the feature extraction pipeline	3
Mid-term presentation	1
Model selection and modification	3
Experimental study	1
Buffer	1
Report Writing and final presentation	2

Literature review for benchmark study
We collect and review the documentation and literature regarding to data collection, particularly the part for ground truth labeling
Depth map construction
We implement a method to derive the depth map with given front-view image and corresponding 3D scene
Implementation of the feature extraction pipeline
We will extract other features that is expected to help to refine the estimation of eye gaze location **Mid-term presentation**
Review of previous work and preliminary results
Model selection and modification
The architecture of the network need to be adapted to the scene-related input. It is expected to modify the existing method instead of newly implementing one.
Experimental study
Design and Output different training experiments
Report Writing and final presentation
Summarize the work in the project

REFERENCES

[1] Aly Khalifa Ayoub Al-Hamadi Ahmed A.Abdelrahman, Thorsten Hempel. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In *IEEE*, 2022.

[2] Oncel Tuzel Josh Susskind-Wenda Wang Russ Webb Apple Inc Ashish Shrivastava, Tomas Pfister. Learning from simulated and unsupervised images through adversarial training. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] Jun Bao, Buyu Liu, and Jun Yu. Escnet: Gaze target detection with the understanding of 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14126–14135, June 2022.

[4] Pavel Chmelar, Ladislav Beran, and Lubos Rejfek. The depth map construction from a 3d point cloud. 2016.

[5] Jiang Deng, Sun Bei, Su Shaojing, and Zuo Zhen. Feature fusion methods in deep-learning generic object detection: A survey. In *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, volume 9, pages 431–437, 2020.

[6] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *ArXiv*, abs/1904.05250, 2017.

[7] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[8] Florent Monay Kenneth Alberto Funes Mora and Jean-Marc Odobez. Eyediap, a database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. *ACM Symposium on Eye Tracking Research and Applications*, 2014.

[9] Kejie Li, Daniel DeTone, Steven Chen, Minh Vo, Ian Reid, Hamid Rezatofighi, Chris Sweeney, Julian Straub, and Richard Newcombe. Odam: Object detection, association, and mapping using posed rgb video. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5978–5988, 2021.

[10] Mohsen Mansouryar, Julian Steil, Yusuke Sugano, and Andreas Bulling. 3d gaze estimation from 2d pupil positions on monocular head-mounted eye trackers. *CoRR*, abs/1601.02644, 2016.

[11] Oskar Palinko, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5054, 2016.

[12] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *International Conference on Computer Vision (ICCV)*, 2019.

[13] Digvijay Patil, Rushikesh Patade, Shraavya R. Srinivasarao, Shruti Hanchate, and Rajat Puri. Three to Two - A Novel Approach for Converting 3D Point Clouds into 2D Depth Maps. 6 2022.

[14] Anto R.Canigueral. The role of eye gaze during natural social interactions in typical and autistic people. *Typical and Atypical Processing of Gaze*, 10, 2019.

[15] Rafael Santos, Nuno Santos, Pedro M. Jorge, and Arnaldo Abrantes. Eye gaze as a human-computer interface. *Procedia Technology*, 17:376–383, 2014. Conference on Electronics, Telecommunications and Computers – CETC 2013.

[16] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1633–1642, 2018.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[18] Hiromasa Suzuki Ting-Hao Li and Yutaka Othake. Visualization of user's attention on objects in 3d environment using only eye tracking glasses. *Journal of Computational Design and Engineering*, 2020.

[19] Erroll Wood and Andreas Bulling. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proc. ACM International Symposium on Eye Tracking Research and Applications (ETRA)*, pages 207–210, 2014.

[20] Srivignesh ; van As-Tarrence ; Zimmermann Joelle Badrinarayanan Vijay ; Rabinovich Andrew Wu, Zhengyang ; Rajendran. Magiceyes: A large scale eye gaze estimation dataset for mixed reality. *CVPR*, 2020.

[21] Yiwei Bao Feng Lu Yihua Cheng, Haofei Wang. Appearance-based gaze estimation with deep learning: A review and benchmark. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[22] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2299–2308, 2017.