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.[13] 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.[22], [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. 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). In the next step, 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, like MagicEyes[21] that includes 587 subjects with 80,000 images, MPIIGaze[23] which contains 213,569 full face images and EYEDIAP[7] in which data is captured by RGB-D camera.

B. 3D Gaze estimation

Conventional methods for 3D gaze estimation task can be generally categorized in model-based[20] and appearance-based approaches[22]. 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.

C. Object detection and tracking

Detection and Tracking of the objects plays a significant role in better understanding the 3D scene, since a scene consists of multiple objects and environments. For most of learning-based approaches, a bounding box of the corresponding object is firstly detected by applying some detectors like YOLO[14] and RetinaNet[9]. The bounding boxes are then encoded in the neural network and be classified and tracked in further frames if required.[19], [6]

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.

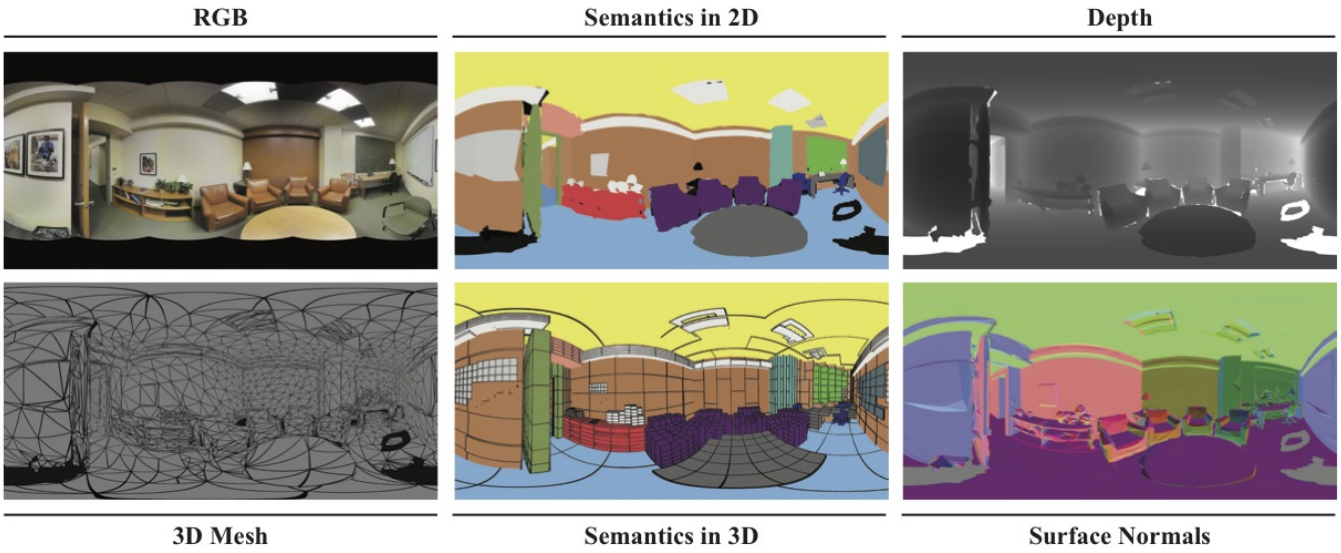


Fig. 1: Main representations of a scene in 2D and 3D frame

III. Project Description

The eye tracking dataset we will use for the project is captured by the device Pupil Invisible from Pupil labs, letting people looking around and collecting the scene and ground truth gaze data. For scanning the 3D scene or objects at which people are staring, a IPhone12 Pro Max with a 3D Scanner app is provided.

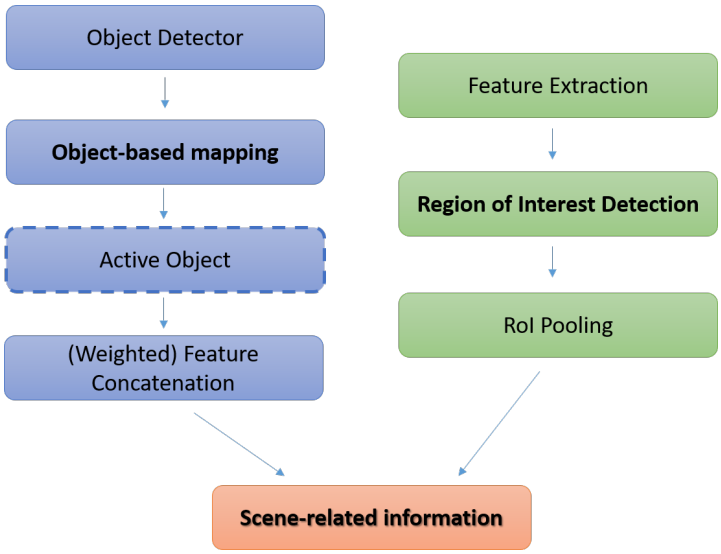
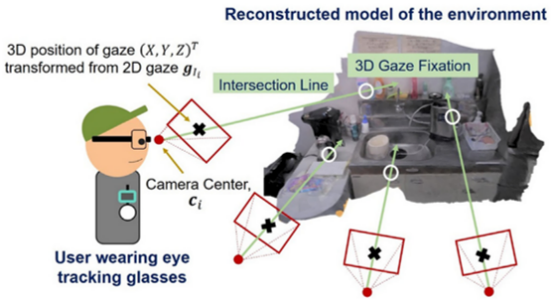
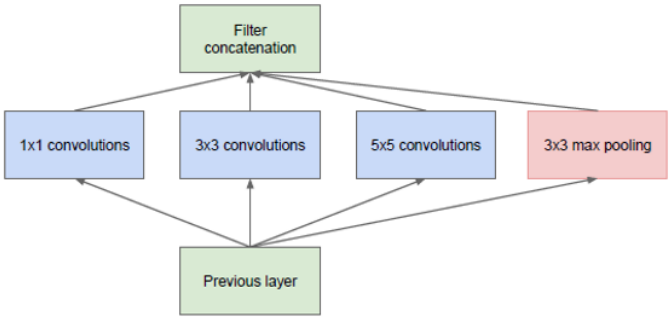


Fig. 2: Proposed Approach for the project



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



(b) Inception module in GoogleNet[17]

A general 3D scene from the first-person perspective can provide information what exactly the observer is seeing and may contribute to better gaze estimation that is robust to view change. Nonetheless, it normally contains too rich information, for instance, a person always stare at only one specific object and ignore the surroundings but the scene is normally unable to predict his intentions.[5] Inspired by the issue, we intend to first preprocess the 3D scene data instead of encoding the point cloud or multi-view frames directly. Two potential approaches is proposed here: object-based and region-based extraction of the 3D scene information. The Figure2 shows the overview of my proposal pipeline for data preprocessing.

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.

A. 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, [5] 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.[4] Besides, cross-attention module yielded in [8] 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.

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[6]. 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 4, 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 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.

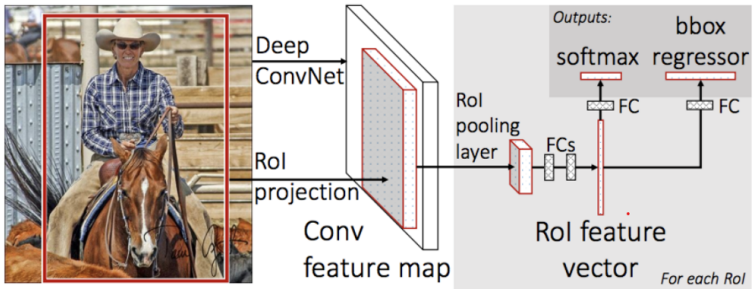


Fig. 4: Overview of Fast R-CNN[6]

B. 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. Most of related researches are CNN-based which means that feature extraction is the most critical in those methods.[22] A typical example of the deep-learning model for gaze estimation is shown in Figure 5. 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).

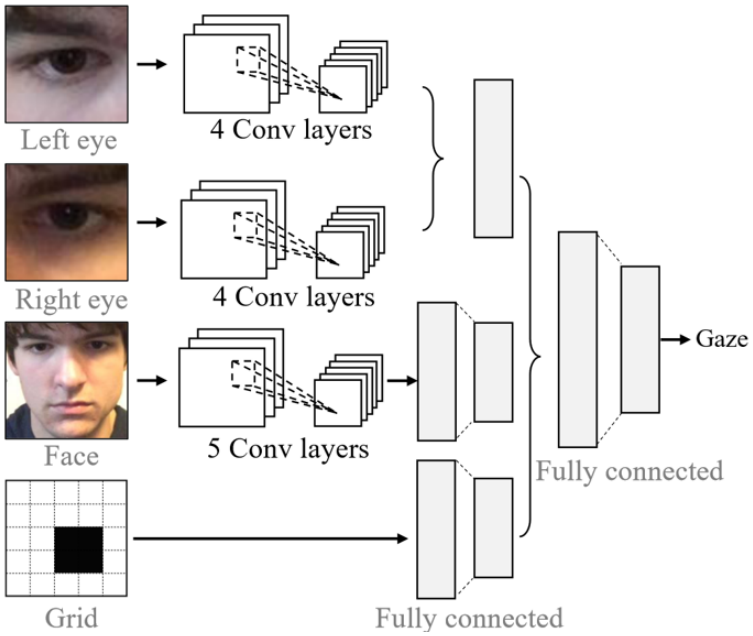


Fig. 5: Example of CNN-based approach for gaze estimation [24]

Similarly, we could also use such a branched architecture to take the advantage of features from different inputs (eye, face, scene). 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)
Environmental setup and code & literature review	1
Implementation of proposed methods	2
Modification of main network	3
Mid-term presentation	1
Experimental study	3
Buffer	2
Report Writing and final presentation	2

Environmental setup and code & literature review
We familiarize ourselves with the lab environment, reproduce gaze estimation results and review related works at the same time.
Implementation of proposed methods
We aims at implementing an algorithm to extract useful related information from 3D scene data for further work
Modification of main network
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 implement one.
Mid-term presentation
Review of previous work and preliminary results
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] 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.

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

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

[7] 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.

[8] 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.

[9] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[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] 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.

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[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] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. 11 2014.

[20] 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.

[21] 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.

[22] 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.

[23] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, 41(01):162–175, jan 2019.

[24] 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.