RSMG2 Report
Gaze Estimation in the Wild

Wang Hengfei

April 21, 2021

1 Administrative Information

Name of student	Hengfei Wang
Name of supervisor	Dr. Hyung Jin Chang, Prof. Ales Leonardis
Other Thesis Group members	Dr. Mohan Sridharan, Dr. Jinming Duan
Date of report	21 April 2021
Working title of thesis	Gaze Estimation in the Wild

2 Introduction

2.1 Background

The research work is about gaze estimation in the wild. Gaze estimation is a task to predict where a person is looking at given the person's eye or face. In general, gaze is represented by pitch and yaw angles (Fig.1) of eyes. To be convenient, we usually use 3D gaze direction in camera coordinate system as the target of gaze estimation in practice.

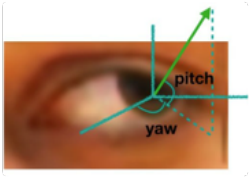

Figure 1: Pitch and yaw angles

Eye is one of the most salient features of the human face. We get most information through vision in our daily life and our gaze can also reflect a lot of personal information like desires, needs, cognitive processes, emotional states and interpersonal relations. Usually, we can know what people want to do and their emotions through observing their eyes even though they wear a mask. I believe it is quite useful in this hard time. However, it is quite hard to predict people's attention or emotion when only seeing other parts of people's face without their eyes. That means eye is the main window through which people convey most information.

Gaze estimation has a variety of applications. Most of them fall into two main fields, diagnostic and interactive. In diagnostic part, gaze can be used to analyze human attention. For example, we can analyze customers' attention to product when they are shopping, then we can decide what kind of product is most popular and what information people care about. And gaze can also be used to analyze human cognitive state like autism diagnosis. For interactive part, we can use gaze for gaming and driving. Actually, you can control anything if they can recognize your gaze. That is interesting. It's also much more

convenient than tradition methods. It can definitely improve our efficiency.

In short, we can have a better understanding of people and even know what they are thinking if we know what they are seeing, how often they see an item and so on. To achieve this goal, gaze estimation in the wild is the first key problem we have to solve. In this project, we plan to use computer vision technology (mainly deep learning methods) to do accurate gaze estimation from images or videos.

2.2 Literature Review

There are roughly two kinds of methods to realize gaze estimation[1].

The first one is model based methods, they model the common physical structures of the human eye geometrically so as to calculate a 3D gaze direction vector (see Fig.2). The advantage of model based methods is that they are usually accurate and stable under different domains. The disadvantage is the high requirement for image quality because they need high resolution images to extract accurate features used for gaze calculation.

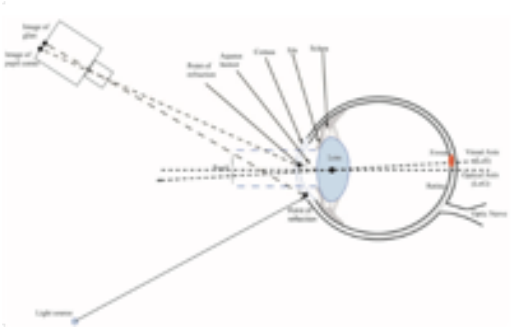

Figure 2: The structure of a traditional model based method[1]

The other one is appearance based methods. Appearance based methods for gaze estimation do not explicitly extract features, but rather use the image contents as input with the intention of mapping these directly to 3D gaze estimation. In recent yeas, deep learning methods are dominating in this field. The advantage of appearance based methods is that they are insensitive to image quality. This advantage is huge because we cannot guarantee the quality of most images from our daily life. If model based methods cannot extract accurate features from images of low quality, they would generate great errors. However, appearance based methods still have some vital problems. One is that most of them need a large amount of data to train. Another is overfitting problem. Overfitting means the model is trained in one dataset and have a good performance in this dataset, but the trained model cannot performed well in other datasets.

In this project, we mainly focus on deep learning method as it has better performance than other methods and is the most promising way in this field.

According to the information used to estimate gaze information, we can separate deep learning method into three parts: eye based method, semantic based method, and face based method.

2.2.1 Eye Based Method

Eye based method only takes eye images as input. Zhang et al. [2] train a simple LeNet network using the images of one single eye. The output of the model is gaze vector which can be converted to pitch and yaw angles (see Fig.3). This work proposes to concatenate head pose and eye features extracted from eye images to estimate gaze vector in camera coordinate system. And another important contribution of this work is that they collected MPIIGaze dataset, one of the most often used datasets in gaze estimation, and made it public. In 2017, Zhang et al. [3] replaced LeNet network with VGG16 network and the accuracy of the model is improved greatly.

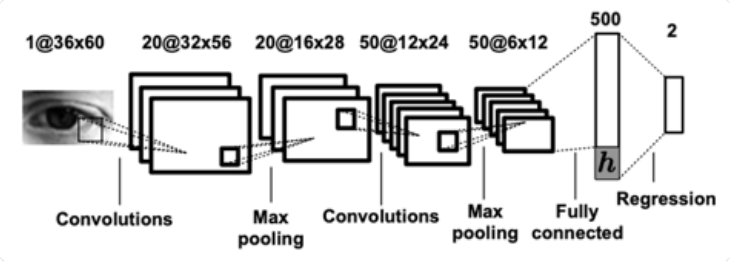

Figure 3: MPIIGaze[2]

Both of the work above only take the images of one single eye as the input of model. Cheng et al. [4] argue that the gaze information of the other eye is also helpful to gaze estimation. So they take images of two eyes as the input of model and tune the weights of losses regarding to left and rights according to the quality of images. The weight of the eye with high quality should be larger.

2.2.2 Semantic Based Method

Not all information of eye images is useful to gaze estimation. Sometimes, the information on our face is harmful to gaze estimation. For example, if we draw some dots on our face, the dots would confuse the estimation model. We can try to extract the information we need for this specific task. And models should be much easier to learn useful information in this way.

Park et al. [5] propose to represent gaze only using the locations of eyeball and iris. They name the simplified representation as gazemap. In the model, they first extract gazemap from a single eye image using a stacked hourglass network and then estimate gaze vector using a regression network (see Fig.4). This approach gets a better result than eye based method.

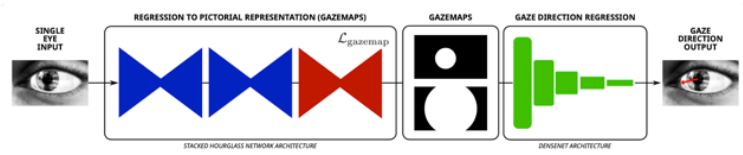

Figure 4: Semantic based method[5]

Yu et al. [6] propose to apply multitask learning to gaze estimation. They estimate gaze vector and landmarks of eyes at the same time. These two tasks can improve each other and get a better performance.

2.2.3 Face Based Method

The methods mentioned above require eye images as input. However, they usually need other models to locate eyes and estimate head poses. It would be more convenient if model can estimate gaze directly from a full face image which means the model can be trained in end-to-end way.

It is hard to learn gaze estimation since eyes are relatively small compared to full face. Zhang et al. [7] try to solve this problem using attention mechanism. Their model learns an average weight map which tends to enlarge the weight of eye region. Then they combine the image features with the weight map to estimate gaze vector (see Fig.5).

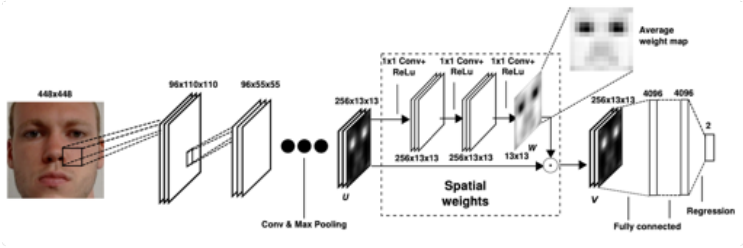

Figure 5: Face based method[7]

2.2.4 Challenges in Gaze Estimation

Even though we have several models for gaze estimation, the performance of the models still cannot satisfy the requirement of our daily use. There are three reasons for this: dataset collection, personalization, and domain adaptation.

Dataset Collection

Accurate gaze label is hard to get. Different with other tasks, accurate gaze label is invisible. So additional equipment is needed to get gaze label which makes data collection complex and time-consuming. In [8], participants

were asked to stare at the ping pong ball at all times, and the depth camera recorded the entire process. After the data is collected, they used algorithms or manual methods to label the eye center position and the ping-pong ball position in the RGB video. They mapped these two positions to the 3D point recorded by a depth camera to obtain the corresponding 3D position. After subtracting these two three-dimensional positions, the gaze direction is obtained. Zhang et al. [2] used public parameters of a camera to transform the gaze target and eye position (obtained through a three-dimensional 6-key point model) to the camera coordinates through algorithms, and then gaze is calculated. This method is not only complicated, but also inaccurate. There are also many metheods based on synthesis [9, 10, 11, 12]. But synthetic images are not realistic which causes large bias when training neural network.

In addition, existing datasets lack data continuity which means they cannot cover head poses and gaze angles in a large range (see Fig. 6). People with different age, gender, and race have totally different eye shapes and eye shape would also change with the change of the surrounding environment like illumination. However, few dataset in gaze estimation can handle all the factors above which causes trained models perform well in some specific datasets.

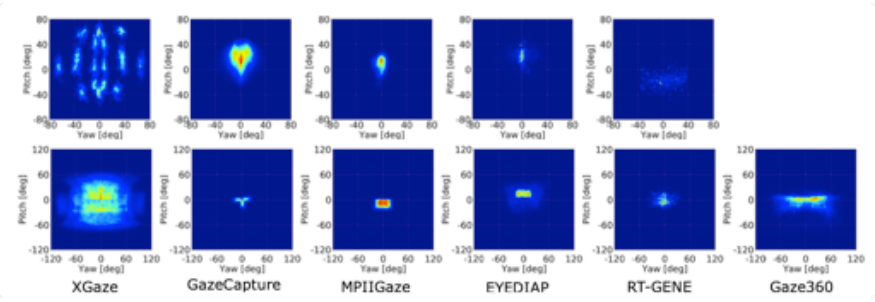

Figure 6: Head pose (top row) and gaze direction (bottom row) distributions of different datasets.[13]

Personalization

There are some differences in eye's structure among people which cause personalization problem. As we can see in Fig. 7, our eyes have optical axis($\mathbf{O}$) and visual axis($\mathbf{V}$). Optical axis is through pupil center and eyeball center while visual axis is through fovea and node N which has a distance from eyeball center. N and fovea positions are different regarding different people and invisible in images. Which means we can not infer these information from image. This cause errors when we want to estimate gaze of different people.

Researchers propose three ways to solve this problem. The first is using differential approach to Learn the differential gaze from the images of the same person [14]. The second is using mixed approach to Model the person-specific parameters using random effects. The fixed effects in the formula is similar to different people [15, 16]. The third is using several images of the same person to finetune generic gaze estimator [17, 18]. But these approaches above only

slightly improve the performance and do not solve this problem completely. In addition, personalization should be real-time to be used in our daily life.

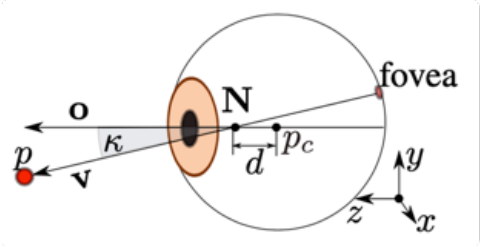

Figure 7: Structure of Eye[19]

Domain Adaptation

Domain adaptation is not only for gaze estimation, but also for almost all the tasks using deep learning. We have so many factors which can influence the domain of dataset like illumination, age, gender, race, resolution of images, makeup, etc. We can use larger dataset including various data to train model, but dataset cannot be large enough to cover all domains and the training process would be too long to use. Therefore, making the models trained in one domain work in others domains is an important topic in deep learning.

2.3 Scientific Questions to Solve

- How to apply generative model to dataset collection with most factors controllable for saving time and huma resource - the factors includes head pose, gaze, data continuity, illumination, age, gender, race, etc
  - How to disentangle these factors from images or 3D models?
  - How to make these factors controllable?
  - Which generative model is appropriate to do this job?
  - How to guarantee the reality of the generated data?
  - How to get accurate head pose label and gaze label after changing some factors?
  - Taking model speed into consideration
- How to improve gaze estimation when subjects wear glasses or makeup which make part of eyes invisible or change the shape of eyes?
  - How to extract useful information and recover eye shape for gaze estiamtion?
  - Is it possible to use other information in test images as constraints of gaze estiamtion and how?

- How to improve gaze estimation cross subject?
  - Since collecting images with accurate gaze labels of one subject is time- consuming and expersive, is it possible to use a short video of the same subject to finetune gaze estimator? Video includes our eye-moving habit and other personal information which may be helpful to gaze estiamtion.
  - Can we generate more images of one subject using genetative model for finetuning gaze estiamtor in realtime?
- How to specify the contributions of different factors to gaze estiamtion in an image and integrate the factors reasonably for more accurate gaze estimation?
  - How to define and measure contributions of different factors to gaze estimation?
  - How to combine these factors to improve gaze estiamtion?
  - Since people may make inferences according to different factors in different scenarios, the contributions of different factors in gaze estimation may change in different images. How to dynamically tune the contributions of these factors?
- How to make gaze estiamtion from sequential frames?

3 Review of Activities

3.1 Course Work

As required by the school, I have completed *LM Research Skills, Evaluation Methods, and Statistics*. Through the training of the course, I learn how to analyze data and how to make a concise persentation. Besides, I also completed the course *CS231n* from Stanford University and *Robot Vision* which my supervisor teach. These two courses help me learn the basic structure of deep learning and computer vision which is fundamental to my PhD research. Now I am taking *Machine Learning 2021* from Taiwan University which introduces new approaches of deeping learning.

As for programming skill, I have completed reading *Automate the boring stuff with Python: practical programming for total beginners* and the PyTorch tutorial online. In addition, I also read some books about Linux system and Numpy and Pandas tutorials. I have reimplemented GAN, DCGAN code and run the demos successfully. In gaze estimation, I have reimplemented MPIIGaze code from [2], NeRF from [20], and Nerfies from [21].

The following are the coursework I have completed and ongoing.

- Completed
  - LM Research Skills, Evaluation Methods, and Statistics: via zoom

  - CS231n: via Youtube
  - Robot Vision: PPT supplied by my supervisor
  - Automate the boring stuff with Python: practical programming for total beginners: ebook
  - PyTorch Tutorial: online
  - Linux Tutorial: ebook
  - Numpy and Pandas Tutorial: ebook
  - Programming: GAN, DCGAN, MPIIGaze, NeRF, Nerfies

- Ongoing
  - Machine Learning: via Youtube
  - Pattern Recognition & Deep Learning: ebook

3.2 Research Progress

I have read most papers in gaze estimation from 2014 and the paper list is in https://github.com/cvlab-uob/Awesome-Gaze-Estimation. The paper list is also maintained by myself. Besides, my supervisor and I reviewed three papers from CVPR2021 together. I have learned a lot about how to write a CVPR paper and how to write a convincing rebuttal by reading the papers and giving some comments. In addition, I actively participate in IRLab seminar every week and the members in my lab will also watch a lecture about deep learning in group meeting every week. IRLab seminar and group meeting continuously bring the information of the most advanced research work all around the world to me. I truely benefit a lot from them.

Over past few months, I have done some preliminary work in create gaze dataset with continuous head pose and gaze direction.

As mentioned above, most existing dataset only discrete head pose and gaze direction which cause a large bias of trained model. But collecting continuous head pose and gaze direction has a high requirement of equipment and is time-consuming. After looking for a new approach to solve this problem for a long time, my supervisor and I found that NeRF may be a good choice. NeRF stands for Neural Radiance Fields and it can accurately reconstruct 3D model only using several images (20-30) of the scene. It can be used to generate new images of the scene from novel views. Our idea is that we take several pictures of one person, then we can reconstruct 3D model of the person's head. After getting the 3D model, we can change view distance and direction continuously to generate images with continuous head pose. As for continuous gaze direction, we are trying to redirect gaze using GAN. If possible, we can explore futher to change the background and illumination of the new images. Recently, we have tried to train NeRF using xgaze dataset [13] which includes people's face images with high resolution and get a good result (see Fig.10).

Figure 8: Original Image

Figure 9: Novel Image

Figure 10: Train NeRF using xgaze dataset

4 Provisional 12 Month Time Table

Activity	Objectives	Date
Implement generative model based on NeRF	- Complete the structure design of generative model and implement the model - Can generate images with continuous head pose and gaze - Develop understanding of NeRF, generative model and deep learning coding	May 2021
Implement baseline methods and make comparisons	- Implement existing generative models of gaze - Experiment to make comparisons among different models - If necessary, modify our method to get a better performance	June 2021
Publish	- Write a Conference Paper on results of the experiments mentioned above	Jul 2021
Thesis Proposal	- Read and summarize the latest papers and make a detailed literature review - Formalise research questions - Write thesis proposal	Aug 2021
Investigate illumination control methods	- Find existing illumination control methods and implement some of them - Assess the possibility of applying existing methods to our generative model	Sep 2021
Implement illumination changing algorithm and combine it with our generative model	- Make our generative model capable to control illumination even the whole background - Prepare CVPR writing and submission	Oct/Nov 2021

Experiment using modified model and make comparison with baseline methods	- Train gaze estimators using the datasets generated by our model - Assess improvement of gaze estimators and make comparisions with baseline methods - If possible, make another Conference publish	Jan 2022
Investigate simulation methods of different factors to gaze estimation	- Define contributions of factors in gaze estimation - Find appropriate method to model contributions of factors in gaze estimation	Feb 2022
Implement contributions model	- Can model contributions of factors in one single image - Try to tune contributions' weights dynamically according to different scenarios	Mar/Apr 2022

References

[1] Dan Witzner Hansen and Qiang Ji. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):478–500, March 2010. ISSN 1939-3539.

[2] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-Based Gaze Estimation in the Wild. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, June 2015.

[3] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1): 162–175, 2017.

[4] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-Based Gaze Estimation via Evaluation-Guided Asymmetric Regression. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11218, pages 105–121. Springer International Publishing, Cham, 2018.

[5] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep Pictorial Gaze Estimation. *arXiv:1807.10002 [cs]*, 11217:741–757, 2018.

[6] Yu Yu, Gang Liu, and Jean-Marc Odobez. Deep Multitask Gaze Estimation with a Constrained Landmark-Gaze Model. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, volume 11130, pages 456–474. Springer International Publishing, Cham, 2019.

[7] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–60, 2017.

[8] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 255–258, 2014.

[9] E. Wood, T. Baltruaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling. Rendering of Eyes for Eye-Shape Registration and Gaze Estimation. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3756–3764, December 2015.

[10] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 131–138, Charleston South Carolina, March 2016. ACM.

[11] Kang Wang, Rui Zhao, and Qiang Ji. A hierarchical generative model for eye image synthesis and eye gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 440–448, 2018.

[12] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.

[13] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020.

[14] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[15] Erik Lindén, Jonas Sjostrand, and Alexandre Proutiere. Learning to personalize in appearance-based gaze tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[16] Yunyang Xiong, Hyunwoo J Kim, and Vikas Singh. Mixed effects neural networks (menets) with applications to gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7752, 2019.

[17] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11937–11946, 2019.

[18] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9368–9377, 2019.

[19] K. A. Funes Mora and J. Odobez. Geometric Generative Gaze Estimation (G3E) for Remote RGB-D Cameras. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1773–1780, June 2014.

[20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020.

[21] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo-Martin Brualla. Deformable neural radiance fields. *arXiv preprint arXiv:2011.12948*, 2020.