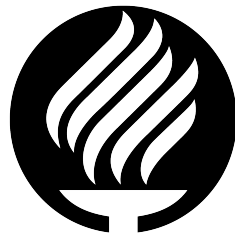


INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE
MONTERREY
CAMPUS QUERÉTARO
DEPARTAMENTO DE COMPUTACIÓN Y MECATRÓNICA



**Tecnológico
de Monterrey**

Gaze-Tracking with Convolutional Neural Networks

by

[Alberto Castañeda Arana](#)

Proyecto Integrador para el Desarrollo de Soluciones Empresariales

*Ingeniería
en
Sistemas Computacionales*

Asesor: Dr. José Antonio Cantoral Ceballos

Santiago de Querétaro, Querétaro, México
15/06/2022

List of Figures

1.1	Overview of Gaze-Estimation setups	1
1.2	Overview classification of Gaze-Tracking methods	2
2.1	Generalized architectures of state-of-the-art models	4
2.2	Samples from different datasets	6

List of Tables

2.1	2D Gaze-Estimation Benchmark of state-of-the-art	4
2.2	Dataset comparison	7

Contents

List of Figures	i
List of Tables	ii
1 Introduction	1
1.1 Gaze-Tracking	1
1.2 Model-based methods	2
1.3 Appearance-based methods	3
2 State-of-the-art	4
2.1 Related Work	4
2.1.1 Multimodal Convolutional Neural Network	5
2.1.2 Spatial Weight CNN	5
2.1.3 iTracker	5
2.1.4 Geometric & Texture-based Networks	5
2.1.5 RT-GENE	6
2.1.6 Hybrid Two-step Training	6
2.2 Datasets	6
2.2.1 GazeCapture	7
2.2.2 MMPIIGaze	7
2.2.3 Columbia Eye Gaze / CAVE	8
2.2.4 EyeDiap	8
2.2.5 ETH-XGaze	8
2.2.6 Discussion	8
3 Development	9
3.1 Objectives	9
3.2 Evaluation Metrics	9
3.3 Proposed Architecture	10
3.4 Training	10
4 Results	11
5 Conclusion	12

Bibliography

13

Chapter 1

Introduction

1.1 Gaze-Tracking

Applications of estimating the human eye gaze direction, or Gaze-Tracking, can be found in multiple study areas such as health care [4], virtual-reality [11], and human-computer interaction [10]. Devices that utilize gaze-tracking methods for human-computer interaction are often referred to as Eye Trackers. While there is commercial use for Eye Trackers, they have not reached the accuracy needed to allow users with fine-motor disabilities to interact with a computer by using their own eyes as a replacement for a traditional mouse device.

Conventional methods for gaze-tracking require additional devices such as infrared light in order to make the gaze estimations. However, these methods are prone to error in poor light conditions.

Research in computer vision and Deep Learning have allowed improvement in Gaze-Tracking systems by allowing better predictions while also being less intrusive for users with no additional devices required.

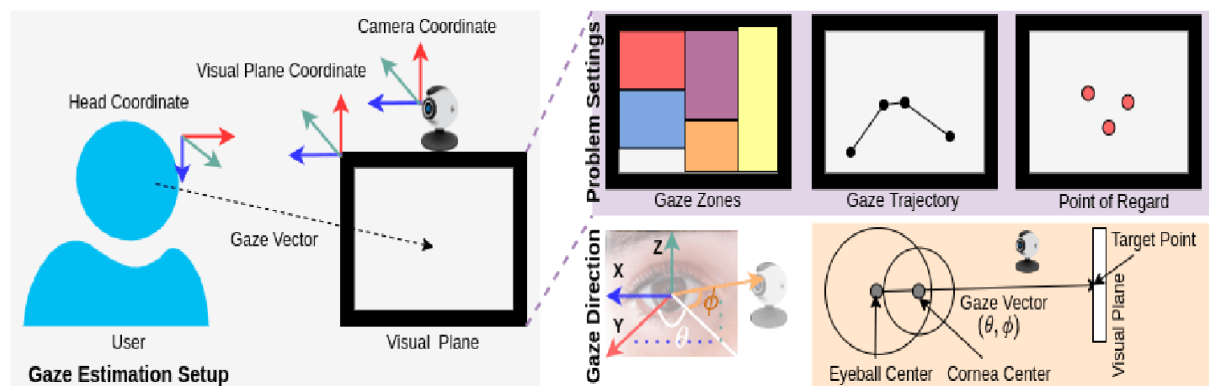


FIGURE 1.1: Overview of Gaze-Estimation setups. Recovered from [3]

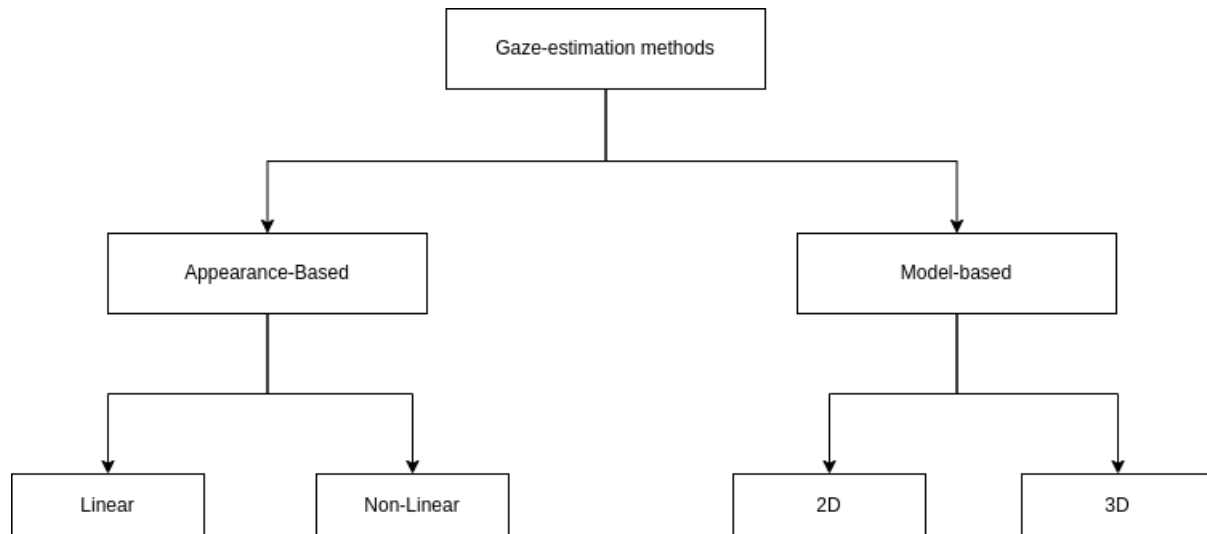


FIGURE 1.2: Overview classification of Gaze-Tracking methods

Gaze-Tracking with computer vision can be achieved with two different type of methods: Model-based methods and Appearance-based methods. These methods will be described in the following sections.

1.2 Model-based methods

Model-based methods, or sometimes referred to as feature-based methods, use the geometric form of the eyes to make the eye gaze prediction. As seen on [Figure 1.2](#), these can be interpreted as 3D models. Model-based 3D Gaze-Estimation methods use 3D eyeball models and estimate the gaze direction using geometric eye features, such as the iris center and the eye corners [5].

Furthermore, most these methods require an infrared light source (IR) in order to create a refraction in the human eye which is then used to calculate the eye gaze direction by comparing the refraction position with the pupil position. The main drawback of model-based methods is that they can have a lower accuracy with low-resolution images and poor-condition lighting environments. Additionally, all of them have limited working distance.¹

Some state-of-the-art commercial products that have adapted these techniques are Tobii Eye Tracker (<https://www.tobii.com/>), Eye Tribe (<https://imotions.com/hardware/the-eye-tribe-tracker/>) and EyeSee (<https://eyeseesearch.com/>). While these products allow very accurate predictions for the Eye Tracking, they suffer from the previously mentioned model based methods disadvantages while also requiring users to purchase additional expensive equipment. Most Eye-Tracking products also require wearable glasses peripherals that can be intrusive for daily usage.

With the increase of interest and usage of Gaze-Tracking, there is also a growing need for new Eye-Trackers that are cheap and easy to use. Gaze-Tracking methods that utilize built-in or external cameras included in almost every computer can be used as a way of making eye tracking more accessible. Appearance-based approaches are beginning to be researched to achieve the previous goal.

1.3 Appearance-based methods

Recently, investigations for appearance based methods with Deep Learning has surfaced which have demonstrated high accuracy while only using on-the-shelf web cameras to capture the human eye appearance hence the method name.

Conventional appearance-based methods used regression in order to learn the mappings from appearance to human gaze. However, since these methods use the appearance of the eye, the main challenges are head motion and subject differences that alter the eye appearance on use.

To overcome the previous challenges, appearance-based Gaze-Estimation with Deep Learning has become a trend in research. Deep Learning based methods can extract high-level abstract gaze features from high-dimensional images, making them more robust and accurate than conventional methods. The mapping between eye-images and gaze direction can be achieved with various regression techniques, such as neural networks, local interpolation, or Gaussian process regression. Convolutional Neural Networks (CNNs) models are commonly used in computer vision tasks for its high accuracy and performance. Current state-of-the-art solutions have begun to adopt CNN based architectures, adopting the advantages of appearance-based methods and achieving high accuracy. CNN models used for Gaze-Tracking learn from RGB images which are preprocessed to segment the human eyes in training.

Some of the drawbacks that appearance-based methods have been that they require more images for training than model-based methods, which are not easily retrieved. Additionally, since most appearance-based methods require deep-learning architectures such as Convolutional Neural Networks, they require higher computational power for training. Several works apply CNNs for gaze estimation, as they have shown to outperform traditional model-based approaches [13].

Chapter 2

State-of-the-art

2.1 Related Work

TABLE 2.1: 2D Gaze-Estimation Benchmark of state-of-the-art

Models	Datasets	MPIIGaze	EyeDiap	GazeCapture
Multimodal Convolutional Neural Network		13.9 °	10.5 °	N/A
iTracker		7.67 cm	10.13 cm	2.81 cm
RT-Gene		5.36 cm	7.19 cm	N/A
Spatial Weights		4.2 cm	8.56 cm	N/A

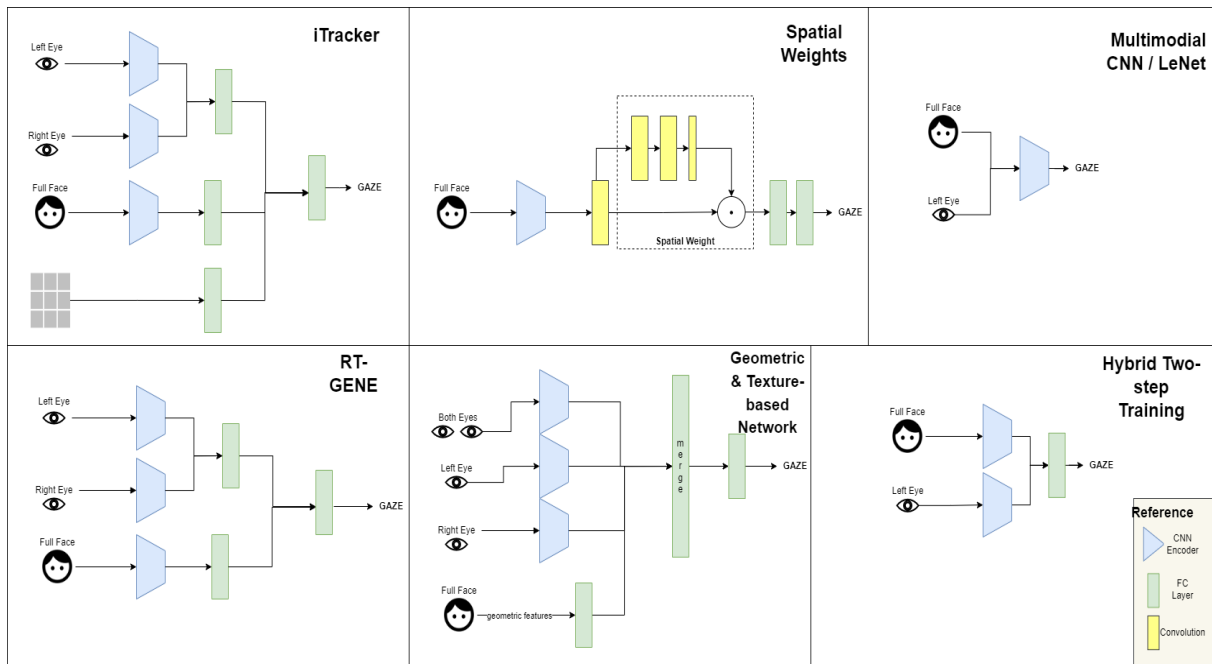


FIGURE 2.1: Generalized architectures of state-of-the-art models

2.1.1 Multimodal Convolutional Neural Network

Zhang et al. proposed a six-layered CNN based on LeNet that takes the eye image as input and combines the head pose in the last fully connected layer of the network [13]. The results demonstrated a large performance gap between person-specific training results (RF model) with the proposed CNN model and revealed the potential of appearance-based gaze estimation.

2.1.2 Spatial Weight CNN

After the previous proposal, Zhang et al. proposed a spatial weights CNN based on the AlexNet architecture [9] that takes the full face image as input [15]. The spatial weights classify the importance of different facial areas in order to perform the Gaze-Estimation. This mechanism includes three additional 1 x 1 convolutional layer followed by a ReLU activation. Compared to other approaches only use eye regions, this method was demonstrated to be more robust against appearance variation such as head pose as well as illumination. The results achieved an error of 4.8-6.0°.

2.1.3 iTracker

Kannan et al. proposed a CNN architecture for an iOS based Gaze-Tracking system called iTracker [7]. The model's large neural network was constructed with multiple smaller neural networks, with four total inputs: a right eye image, a left eye image, a face image, and a face-grid. The face-grid is a 25x25 binary grid that indicates where the user face is placed relatively to the camera view.

This model achieved 1.66 centimeter error with no calibration needed. The GazeCapture dataset which is described in [section 2.2](#) was used to train this model.

While achieving good results, this model is biased with mobile usage which makes it unusable for laptop or computer desktop usage. Additionally, the authors describe there being room for improvement for better segmentation of eyes. However, by adjusting the model's face-grid dimensions to simulate webcam dimensions and with a different dataset, the model could potentially be used for desktop/laptop use.

2.1.4 Geometric & Texture-based Networks

Jyoti et al. proposed an ensemble of networks that uses the full face, left and right eye [6]. A Deep Neural Network (DNN) is trained using the face geometric features. Secondly, three CNNs are trained for the texture based features i.e. the segment of the left eye, right eye,



FIGURE 2.2: Samples from different datasets. Recovered from [8], [13], [12] and [14] respectively.

and the combined eyes. The model was trained with the publicly available Columbia Eye Gaze and TabletGaze datasets for experiments. When trained with the Columbia Eye Gaze dataset, experiments resulted in an error of 2.22° . Additionally, the authors emphasize on the importance of choosing different activation functions for different results by comparing the model outcome with ReLU vs Swish, where the former achieved better performance for geometric network while Swish showed better performance on the rest.

2.1.5 RT-GENE

Fischer et al. proposed a two-stream VGG network that takes the left and right as inputs [2]. This approach considers large camera-to-user distances and high variations in head pose found in natural environments. This approach allows automatic annotation of subject's ground truth gaze and head pose regardless of large camera-user distances, achieving an error of 7.7° . However, the approach setup requires additional devices such as mobile eye-tracking glasses (Kinect v2 RGB-D camera).

2.1.6 Hybrid Two-step Training

Deng and Zhu proposed a two-step training architecture, where a head CNN and eye CNN are independently trained and then together transformed with a "gaze transform layer" [1]. The proposed model does not suffer from head-gaze correlation overfitting which makes it possible to use existing datasets for training. The solution achieves an error of 5.6° when trained with a large dataset, achieving state-of-the-art gaze tracking accuracy.

2.2 Datasets

Until recently, datasets for Gaze-Estimation models training were built manually by taking pictures of different users' faces by placing their head firmly in a stand. Training models require

TABLE 2.2: Dataset comparison

Dataset	Total #	Size	Sub- jects #	Gaze Points #	Full Face	2D Gaze	3D Gaze	Ref
GazeCapture 2016 images	2.4M	136 GB	1,474	203	✓	✓	✗	[8]
MPIIGaze 2015 images	213K	2.1 GB	15	203	✗	✓	✓	[13]
CAVE 2013 images	6K	2.2 GB	56	105	✓	✗	✓	[12]
EyeDiap 2014 videos	94	11.6 TB	16	105	✓	✓	✓	[12]
ETH-XGaze 2020	1.1M	130 GB	110	N/A	✓	✓	✓	[14]

multiple high-quality images, which makes the previous methods slow and not ideal. To improve the previous methods' problem, multiple data sets containing images for eye tracking have been compiled with crowd founding to train new models. Figure 2.2 represents an overview of different state-of-the-art datasets, while Table 2.2 provides an analytical comparison of each of them.

2.2.1 GazeCapture

Krafka et al. built and released the GazeCapture dataset, which crowdfunded a mobile-based data set of approximately 1500 participants from a wide variety of backgrounds captured by an iOS application [8]. However, given that this dataset is built with mobile users, training with this dataset may lead to a huge bias towards mobile users which could lead to poor results with laptop or desktop use. Additionally, the authors provide pre-trained models utilizing the dataset available as open source (<https://github.com/CSAILVision/GazeCapture>).

2.2.2 MMPIIGaze

Another state-of-the-art dataset that has been widely used is the MMPIIGaze dataset created by Zhang et al. [13]. The dataset contains a total of 213,659 images from 15 participants. For each participant, there is a varied number of images ranging from 34,745 to 1,498. The dataset contains multiple features, our target being the 3D gaze target position related to camera. The dataset is taken from the participant's laptops, contrasting GazeCapture's images taken from mobile devices.

2.2.3 Columbia Eye Gaze / CAVE

The CAVE dataset [12] consists of 5,880 images of 56 different participants. For each participant, there are 5 head poses with 21 gaze directions per head pose. The authors describe the dataset as having more fixed gaze targets than other publicly available gaze datasets. The participants are ethnically diverse and some of them are captured wearing glasses which makes this a strong candidate for learning. The dataset comes with already-segmented eye areas, so segmentation step can be skipped.

2.2.4 EyeDiap

The EYEDIAP dataset was designed to train and evaluate gaze estimation algorithms from RGB and RGB-D data. By using videos instead of pictures, it manages to be the biggest sized dataset of the state-of-the-art. Additionally, the recording methodology used is designed to systematically include, and isolate, most of the variables which affect gaze estimation algorithms: head pose and lightning variations.

2.2.5 ETH-XGaze

Finally, the ETH-XGaze dataset [14] consists of over 1,000,000 images of varying gaze under extreme head posers collected from 110 participants. While the extreme varied head poses provide good opportunities for learning, for the case of this project the head poses that are not looking into the screen would not be useful so they must be discarded.

2.2.6 Discussion

As seen on [Table 2.2](#), the dataset with the biggest scale is ETH-XGaze, having the biggest samples and gaze points available. While training with the ETH-XGaze dataset would create the biggest model, the huge size of its collection of images might require high computational power and training time in order to fully train the model. While GazeCapture is also a great candidate, its small number of gaze points is insufficient for the purpose of gaze-estimation in a computer screen.

Chapter 3

Development

3.1 Objectives

For this research, a new proposed model for Gaze-Estimation that outperforms current state-of-the-art systems is the goal. To fully realize this, the model must have lower accuracy and higher speed. Accuracy can be determined with evaluation metrics described in [section 3.2](#), while the speed of the model must be fast enough in order to potentially use this gaze-estimation system in real-time on the average user computer.

3.2 Evaluation Metrics

There are two commonly used metrics for performance evaluation in Gaze-Tracking Systems: Euclidean distance and the angular error. Angular error is used when referring to 3D gaze estimation, while euclidean distance is used for 2D gaze estimation.

Assuming the gaze direction is $\mathbf{g} \in \mathbb{R}^3$, and the estimated gaze direction is $\hat{\mathbf{g}} \in \mathbb{R}^3$, the angular error can be calculated with the formula:

$$\iota_{angular} = \frac{\mathbf{g} \cdot \hat{\mathbf{g}}}{\|\mathbf{g}\| \|\hat{\mathbf{g}}\|} \quad (3.1)$$

For 2D gaze estimation, the actual gaze position is denoted as $\mathbf{p} \in \mathbb{R}^2$ while the estimated gaze position as $\hat{\mathbf{p}} \in \mathbb{R}^2$. This would result in the following formula:

$$\iota_{Euclidean} = \|\mathbf{p} - \hat{\mathbf{p}}\| \quad (3.2)$$

For the purpose of this research, given that the gaze prediction desired is of the screen coordinates that the user is looking at, euclidean distance will be used as a performance evaluation metric for the proposed model's 2D gaze estimation.

3.3 Proposed Architecture

3.4 Training

Chapter 4

Results

Chapter 5

Conclusion

References

- [1] Haoping Deng and Wangjiang Zhu. “Monocular Free-Head 3D Gaze Tracking with Deep Learning and Geometry Constraints”. In: Oct. 2017, pp. 3162–3171. DOI: [10.1109/ICCV.2017.341](https://doi.org/10.1109/ICCV.2017.341).
- [2] Tobias Fischer, Hyung Chang, and Yiannis Demiris. “RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments”. In: Sept. 2018. DOI: [10.1007/978-3-030-01249-6_21](https://doi.org/10.1007/978-3-030-01249-6_21).
- [3] Shreya Ghosh et al. “Automatic Gaze Analysis: A Survey of Deep Learning based Approaches”. In: *arXiv:2108.05479 [cs]* (Aug. 2021). arXiv: 2108.05479. URL: <http://arxiv.org/abs/2108.05479> (visited on 03/31/2022).
- [4] Alessandro Grillini et al. “Towards Using the Spatio-Temporal Properties of Eye Movements to Classify Visual Field Defects”. In: *Proceedings of the 2018 ACM Symposium on Eye Tracking Research Applications*. ETRA ’18. Warsaw, Poland: Association for Computing Machinery, 2018. ISBN: 9781450357067. DOI: [10.1145/3204493.3204590](https://doi.org/10.1145/3204493.3204590). URL: <https://doi.org/10.1145/3204493.3204590>.
- [5] E.D. Guestrin and M. Eizenman. “General theory of remote gaze estimation using the pupil center and corneal reflections”. In: *IEEE Transactions on Biomedical Engineering* 53.6 (2006), pp. 1124–1133. DOI: [10.1109/TBME.2005.863952](https://doi.org/10.1109/TBME.2005.863952).
- [6] Shreyank Jyoti and Abhinav Dhall. “Automatic Eye Gaze Estimation using Geometric & Texture-based Networks”. In: Aug. 2018, pp. 2474–2479. DOI: [10.1109/ICPR.2018.8545162](https://doi.org/10.1109/ICPR.2018.8545162).
- [7] Harini D. Kannan. “Eye tracking for the iPhone using deep learning”. eng. Accepted: 2018-01-12T20:59:18Z. Thesis. Massachusetts Institute of Technology, 2017. URL: <https://dspace.mit.edu/handle/1721.1/113142> (visited on 03/27/2022).
- [8] K. Krafka et al. “Eye Tracking for Everyone”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2016, pp. 2176–2184. DOI: [10.1109/CVPR.2016.239](https://doi.org/10.1109/CVPR.2016.239). URL: <https://doi.ieee.org/10.1109/CVPR.2016.239>.

- [9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. URL: <https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html> (visited on 04/15/2022).
- [10] Raphael Menges et al. “GazeTheWeb: A Gaze-Controlled Web Browser”. In: *Proceedings of the 14th International Web for All Conference*. W4A ’17. Perth, Western Australia, Australia: Association for Computing Machinery, 2017. ISBN: 9781450349000. DOI: [10.1145/3058555.3058582](https://doi.org/10.1145/3058555.3058582). URL: <https://doi.org/10.1145/3058555.3058582>.
- [11] Benjamin Outram et al. “Anyorbit: orbital navigation in virtual environments with eye-tracking”. In: June 2018, pp. 1–5. DOI: [10.1145/3204493.3209579](https://doi.org/10.1145/3204493.3209579).
- [12] B.A. Smith et al. “Gaze Locking: Passive Eye Contact Detection for Human?Object Interaction”. In: *ACM Symposium on User Interface Software and Technology (UIST)*. Oct. 2013, pp. 271–280.
- [13] Xucong Zhang et al. “Appearance-based gaze estimation in the wild”. en. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, June 2015, pp. 4511–4520. ISBN: 978-1-4673-6964-0. DOI: [10.1109/CVPR.2015.7299081](https://doi.org/10.1109/CVPR.2015.7299081). URL: <http://ieeexplore.ieee.org/document/7299081/> (visited on 04/06/2022).
- [14] Xucong Zhang et al. “ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation”. In: *European Conference on Computer Vision (ECCV)*. 2020.
- [15] Xucong Zhang et al. “It’s Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation”. In: *arXiv:1611.08860 [cs]* (May 2017). arXiv: 1611.08860. URL: <http://arxiv.org/abs/1611.08860> (visited on 03/31/2022).