Gaze from Head: Gaze Estimation without Observing Eye

Jun'ichi Murakami[1] and Ikuhisa Mitsugami[1]

Hiroshima City University, Hiroshima, Japan
murakami@sys.info.hiroshima-cu.ac.jp, mitsugami@hiroshima-cu.ac.jp
http://www.sys.info.hiroshima-cu.ac.jp

Abstract. We propose a gaze estimation method not from eye observation but from head motion. This proposed method is based on physiological studies about the eye-head coordination, and the gaze direction is estimated from observation of head motion by using the eye-head coordination model trained by preliminarily collected data of gaze direction and head pose sequence. We collected gaze-head datasets of from people who walked around under real and VR environments, constructed the eye-head coordination models from those datasets, and evaluated them quantitatively. In addition, we confirmed that there was no significant difference between the models from the real and VR datasets in their estimation accuracy.

Keywords: gaze estimation · eye-head coordination · virtual realty.

1 Introduction

When we observe a person at a distance, we can often predict their gaze direction, regardless of whether their eyes are clearly observed or totally hidden. Thus, we can estimate gaze direction even without observing the eyes directly. Why is it possible? What do we observe in fact to do it?

When a person gazes in a particular direction, they rotate both the head and the eyeballs. Physiological research has revealed several types of coordination between the eyeballs and head. For example, Fang *et al.* reported a significant linear relationship between the rotation angles of the eyeballs and head [1, 2]. Besides, in a further study, Okada *et al.* reported that the same relationship is present even while walking [3]. The vestibulo-ocular reflex (VOR) [4] is another well-known type of eye-head coordination. VOR is an unconscious reflex that enables the stabilization of images on the retinas during head movement by producing eye movements in the direction opposite to head movement, maintaining the image at the center of the visual field. During natural vision, we unconsciously control the eyeballs and head using these coordination functions. Thus, implicit knowledge about eye-head coordination may also be unconsciously applied during the estimation of the gaze direction of others, utilizing information about the position of the head.

There are several existing methods for estimating gaze, which can be categorized into the following two approaches: model-based and appearance-based. The model-based methods use 2-D or 3-D geometric eye models. Studies based on corneal reflection-based methods [5–8] use 3-D models. In addition, several methods have been developed for extracting features, including techniques that extract pupil center or iris edges from 2-D eye images [9–12]. These methods generally require comparatively high-resolution images, and so are not robust for lower resolution images like those captured by commonly used surveillance cameras. On the other hand, appearance-based methods use eye images directly. Compared with the model-based methods, these approaches tend to be more robust at lower resolutions. Early appearance-based methods were based on several assumptions, such as a fixed head position [13–17]. Recently developed methods have less reliance on such assumptions by simultaneously estimating 3D head pose [18–21]. All those methods are based on the measurement or estimation of gaze by observing the eyes. However, the resolution and quality of images captured by commonly-used surveillance methods are insufficient for appearance-based methods to process accurately. Thus, to our knowledge, no previous gaze estimation method is appropriate for use with surveillance images.

In this paper, therefore, we propose a method for estimating gaze not from eye appearance but from head motion because usually the head pose can be estimated more easily than the gaze direction due to a difference in their physical sizes. Such an approach is valid also when the eyes cannot be observed since a person wears glasses or is captured by the camera from his/her back. Moreover, it would be used complementarily even when the eyes can be observed. To realize such a method, we utilize physiological findings; there is an unconscious coordination mechanism between the eye and head motions. The proposed method trains this eye-head coordination by collecting dataset of the eye and head motions and applying machine learning techniques. More concretely, the gaze direction at a particular moment is estimated from the head motion around that moment using the Gradient Boosting Regression [22]. To collect the gaze-head datasets, we constructed two environments; real and VR environments. The real one is a corridor-like space, where a participant wore eye-tracker and repeatedly walked while gazing at several targets. As the VR one, we constructed an eye-tracking VR system that has a participant wearing a VR goggle feel as if he/she was in various scenes. We collected the gaze and head motion from multiple participants to construct datasets, and quantitatively evaluated the gaze estimation accuracy using the datasets and the proposed method. The experimental results revealed that the eye-head coordination actually existed and the proposed method relying on the coordination was effective for the gaze estimation.

2 Eye-Head Coordination

When a person shifts his/her gaze direction, he/she typically moves the head as well as the eyeballs. Physiological researchers have reported a strong relationship between head and eyeball motions. Thus, in natural behavior, the head and

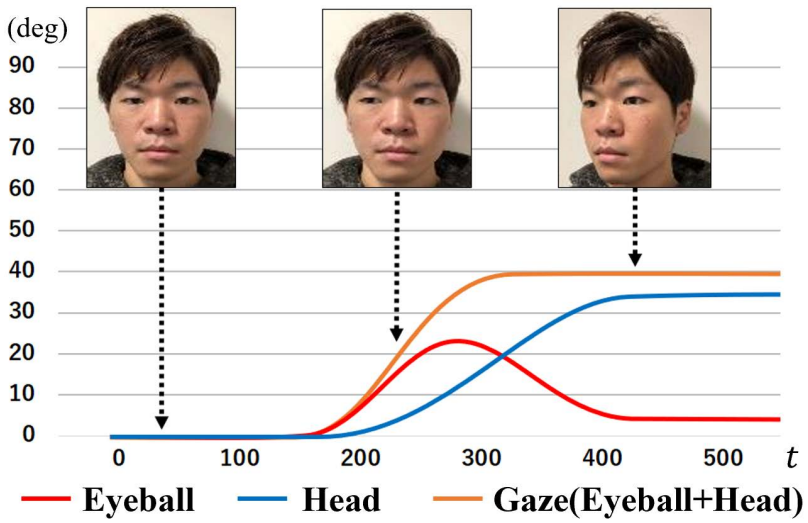

Fig. 1. Eye-head coordination

eyeballs do not move independently, but in a coordinated way, known as human eye-head coordination [23, 24].

Fang *et al.* reported that the rotation angles of the head and eyes exhibited a significant linear relation [1, 2]. In a further study, Okada *et al.* reported that the same relationship was present while subjects walked [3].

The vestibulo-ocular reflex (VOR) [4] is another well-known coordination system. VOR is an unconscious reflex that stabilizes images on the retinas during head movement by producing eye movements in the opposite direction to head movement, maintaining the image at the center of the visual field. Fig. 1 shows a typical example of VOR. The red and blue curves denote the motion of the eyeball and head, respectively, and the orange curve denotes the gaze direction, which corresponds to the summation of the eyeball and head rotation angles. When a person changes their attention to the right, the eyeballs initially move quickly to the right. This motion (i.e., a saccade) captures an image of the target in the center of the retinas. The head then follows the direction of movement, moving less quickly than the eyeballs. During the head motion to the right, the eyeballs turn to the left for stabilizing images on the retinas. This phenomenon can be seen as a crossing of the red (eyeball) and blue (head) curves in the figure. As a result, the head movement and gaze exhibit sigmoid-like curves.

3 Methods

To estimate gaze from the head movement, we need to define a model that describes the eye-head coordination. While several studies have described such models as control diagrams, however, they are too complicated to deduce a

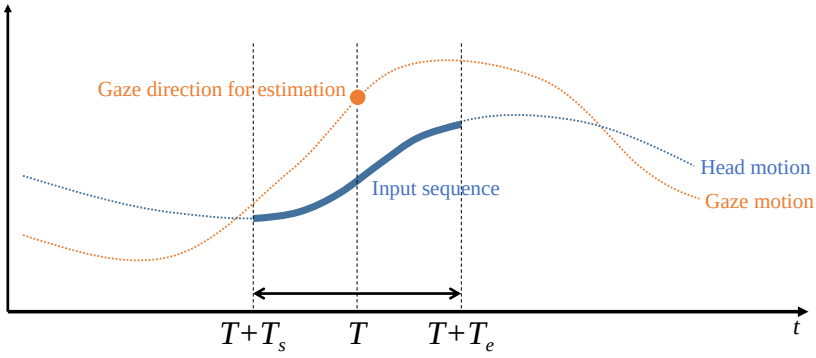

Fig. 2. The proposed method.

formulation analytically or are solely conceptual. Besides, as a fundamental limitation of this approach, even complicated control diagrams cannot cover every possible movement of the eyes and head, because of the high degree of variance within and between participants.

Considering the drawbacks mentioned above of those top-down modeling, it would be more useful to obtain the eye-head coordination using a bottom-up approach that does not rely on examining the mechanisms of the eye-head coordination. The proposed method thus trains the coordination as the relationship between gaze direction at a particular moment T and a sequence of head direction $[T + T_s : T + T_e]$ around that moment, as shown in Fig. 2. Gradient Boosting Regression [22] is then applied to construct a model that estimates the gaze direction from the head motion.

4 Datasets

For evaluation of the proposed method, we collected two types of gaze-head datasets.

4.1 Real dataset

Considering the aim and expected applications of the proposed method, it would be appropriate to evaluate our approach using surveillance images. It is difficult to obtain ground truth information about the gaze direction of a person without an eye-tracking device. However, if a person is wearing an eye-tracker, surveillance images may look unnatural and affect the estimation of head pose. In this paper, therefore, for quantitative evaluation we used a wearable eye-tracker, and head pose was obtained not from surveillance images but using a camera position estimation technique with a head-mounted camera that was part of the eye-tracker.

Fig. 3. EMR-9

Fig. 4. Sight of EMR-9 camera.

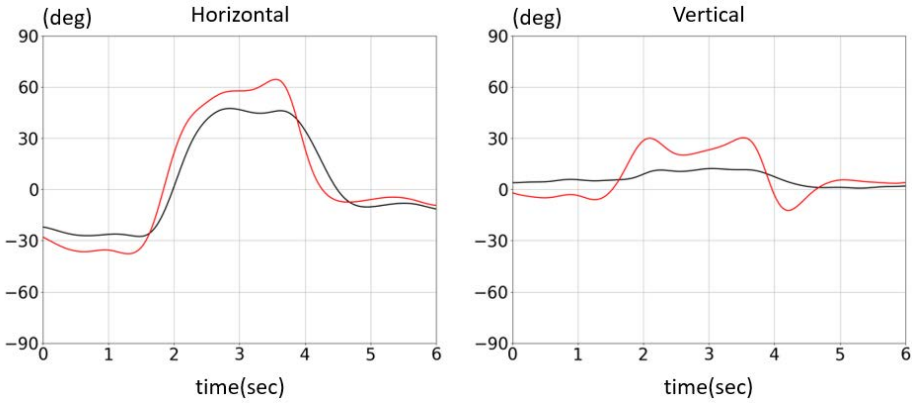

Fig. 5. Real dataset. The black lines denote the measured head motion. The red lines denote the measured gaze.

We collected data of gaze and head motion from people who walk around under the real environment. We used NAC Inc. EMR-9 eye-tracker [25]. This device is a binocular eye-tracker based on a pupil corneal reflection method [26]. With this system, binocular eye directions are described as points on the image plane of the egocentric camera. Details of calibration for this device have been described by Mitsugami *et al.* [27]. Fig. 4 shows an image captured by the camera. Each participant walked back and forth along a straight 8m corridor, where 60 AR markers were located at various positions and orientations, enabling us to obtain positions and poses of the cameras located anywhere in the environment. There were also several gaze targets, and the participant was asked to repeatedly gaze at each of them in accord with verbal instructions.

Fig. 6. Experimental environment using VR system [28].

The camera positions in the global coordinate system were estimated from captured videos using AR marker-based position estimation. Using the camera position and its intrinsics, we transformed the gaze directions measured by the eye-tracker into the global coordinate system. The gaze/head angles in the global coordinate system were then transformed again into the person coordinate where Z axis denotes their walking direction, which was obtained as a differential of the camera trajectory. The videos had a frame-rate of 30 frames per second.

Fig. 5 shows examples of the gaze and head motions captured in this system. We observed the gaze direction changes faster than head movement. We also confirmed that this dataset contained more gaze and head changes in the horizontal direction than in the vertical direction due to the locations of the targets. Using this system, we collected the gaze and head motions from 8 participants.

4.2 VR Dataset

On the other hand, to evaluate our method with more datasets, we collected data of gaze and head motion from people who walk around under the VR environment. We used an eye-tracking VR system that can simultaneously measure gaze direction and head position/poses from a user who experiences 6-DoF VR scenes [28]. As shown in Fig. 6, a participant worn an eye-tracking VR goggle and two controllers. The goggle is connected to a PC, and is calibrated using software offered by its providers. Once the calibration finished, the position and pose of the head, positions of both shoulders, and gaze direction were obtained in the global coordinate system. The gaze and head pose were then transformed again into the person coordinate system where Z axis denotes their chest direction, which was obtained from positions of both shoulders.

We designed user experience so as that we could efficiently construct a large dataset containing frequent gaze shifts and adequately long sequences of gaze behaviors and head motions of many participants in various scenes. For maintaining the scene variation, we implemented three realistic CG scenes, as shown

(a) Library (b) Shop (c) Class Room

Fig. 7. Variation of CG scenes.

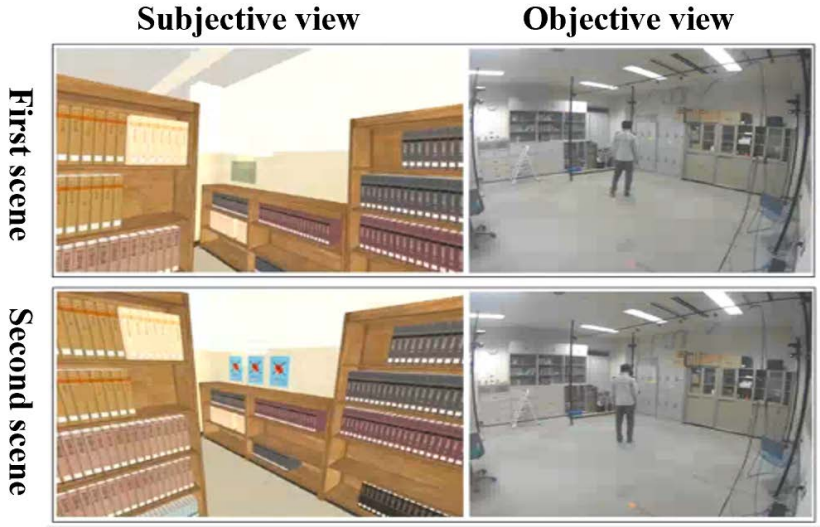

Fig. 8. The spot-the-difference trials.

in Fig. 7, where each participant was able to walk freely within the field of $2m$ square. To collect unconstrained and natural data of the gaze and head motion, we asked each participant to do the spot-the-difference tasks, as shown in Fig. 8. In the first stage, a participant browsed a base scene walking freely to remember the scene for a minute. In the second stage, a participant then browsed the same scene but containing just a difference (e.g., a poster pasted on a wall, or a book eliminated from a bookshelf), and tried to find the difference as quickly as possible. Although his/her chest, head, and gaze directions were measured in both the first scene and the second scene, we used only the first scene data for training and evaluation of the eye-head coordination modeling. This is because, in the case of the second scene, the participant tends to move his/her eyeballs and head faster than usual by the intention to search for a different part quickly, and were

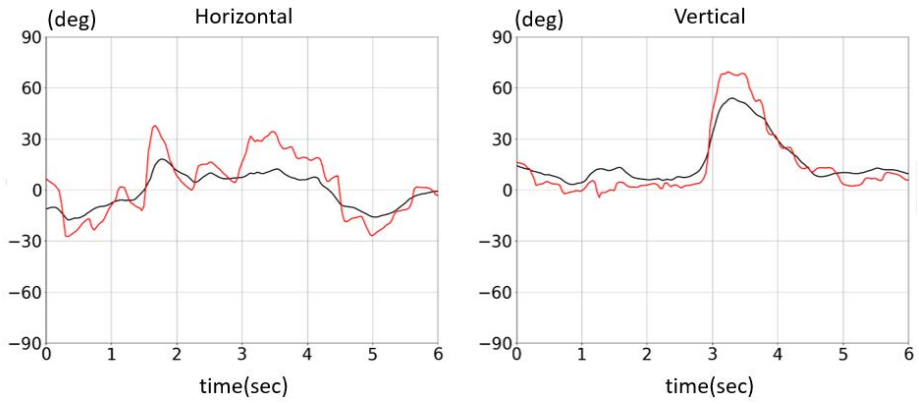

Fig. 9. VR dataset. The black lines denote the measured head motion. The red lines denote the measured gaze.

considered unsuitable for acquiring natural coordination. The chest, head, and gaze directions were measured with a frame-rate of 90 frames per second (fps) by this system. They were resampled to 30fps to match the real dataset.

Fig. 9 shows examples of the captured gaze-head data in the VR system. In addition, we confirmed that the VR dataset contained data of greater gaze/head direction changes in the vertical direction than the real dataset. Thanks to this system, it is easy to have participants experience various scenes, so that it is easier to increase the number of participants, which is vital to make the coordination model more robust and general. We collected data from 16 participants.

5 Experiment

We constructed the eye-head coordination models by applying the proposed method to the real and VR datasets and evaluated their performances. In this experiment, $T_s = -30$(frame) (i.e., -1(sec)) and $T_e = 30$(frame) (i.e., 1(sec)).

To confirm the gaze estimation performance of those models, we adopted leave-one-subject-out cross-validation; we selected a participant for testing, and used the other participants for training. We used Gradient Boosting for the regression.

5.1 Qualitative evaluation

Fig. 10 show examples of the gaze estimation using the real and VR datasets; the left and right graphs show the real and VR datasets, respectively. In both graphs, the dotted blue curves (the estimated gaze direction) are close to the red curves (the measured (ground truth) gaze direction).

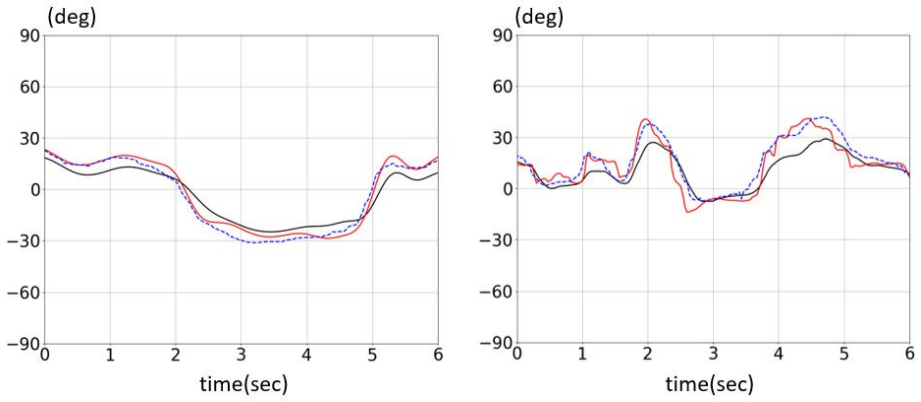

Fig. 10. Examples of the gaze estimation using the real and VR datasets (Left: real, Right: VR). The black lines denote head motion. The dotted blue and red lines denote the estimated gaze and ground truth, respectively.

5.2 Quantitative evaluation

To quantitatively evaluate the gaze estimation performance, we adopted leave-one-subject-out cross-validation; we used a participant for testing, and the others for training.

Table 1 and Table 2 show the mean absolute error (MAE) of the estimated gaze direction. Table 1 and Table 2 show the mean absolute error (MAE) of the estimated gaze direction using only the real and VR datasets, respectively. The MAE values in the table are similar among different people, with no extremely large or small values. Thus, it was not necessary to adjust parameters for each person, enabling us to use a consistent trained model. Our method was thus able to estimate gaze direction across different people, which is a favorable characteristic considering the use in real-world surveillance applications.

In addition, for evaluating the effectiveness of the proposed method, Table 3 and Table 4 show the performance of the case in which head direction was simply regarded as the estimated gaze direction, referred to as the baseline method. As shown in both tables, the proposed method outperformed the baseline.

5.3 Compatibility of the real and VR datasets

To confirm compatibility between the real and VR datasets, we used different datasets for training and testing; estimating the real data using a model trained by the VR data, and vice versa.

Table 5 shows the results when using the real dataset for testing and the VR dataset for training, and Table 6 vice versa. We confirmed that the errors in the vertical direction are larger than those in Table 1 and 2. This fact indicates that the coordination models in the vertical direction obtained from the real and VR datasets are not consistent. We consider, however, it is mainly because of

Table 1. The performance of model using the real dataset.

	MAE (deg)	
ID	Horizontal	Vertical
01	8.0	8.2
02	6.5	6.6
03	7.4	8.6
04	8.7	8.6
05	6.9	6.9
06	7.6	7.9
07	8.3	8.0
08	6.4	6.4
Average	7.5	7.7

Table 2. The performance of model using the VR dataset.

	MAE (deg)	
ID	Horizontal	Vertical
09	7.8	6.6
10	8.0	6.9
11	9.8	7.9
12	8.3	8.6
13	8.5	7.7
14	10.0	8.9
15	8.1	6.7
16	7.7	6.8
17	8.2	7.3
18	7.7	7.6
19	8.4	7.5
20	7.5	5.2
21	7.3	6.5
22	9.1	7.6
23	7.6	9.8
24	8.1	7.5
Average	8.3	7.4

the difference in location of the gaze targets; in the real environment the targets were located at similar heights, while in the VR environment the participant saw widely in the horizontal and vertical directions. As for the horizontal direction, on the other hand, there is no large difference between the errors in Table 1, 2, 5 and 6, which are summarized in Table 7. We then applied the t-test to check whether there was a significant difference between the real and VR datasets. As a result, there was no significant difference ($p > 0.05$).

6 Conclusion

In this paper we proposed a novel approach for gaze estimation from head motion, by exploiting the eye-head coordination function revealed by physiological

Table 3. Quantitative evaluation: the model using the real dataset.

Methods	MAE (deg)	
	Horizontal	Vertical
Baseline (Assuming gaze dir. = head dir.)	9.3	9.2
Proposed (Gradient Boosting Machine)	7.5	7.7

Table 4. Quantitative evaluation: the model using the VR dataset.

Methods	MAE (deg)	
	Horizontal	Vertical
Baseline (Assuming gaze dir. = head dir.)	9.9	8.9
Proposed (Gradient Boosting Machine)	8.3	7.4

Table 5. The performance of models. Test data was the real dataset and training data was the VR dataset.

	MAE (deg)	
ID	Horizontal	Vertical
01	8.4	11.2
02	6.4	9.8
03	7.4	9.3
04	8.5	8.8
05	7.0	9.7
06	7.3	12.0
07	6.9	5.5
08	6.5	9.9
Average	7.3	9.5

research. To analyze eye-head coordination, we propose a learning-based method that implicitly learns the eye-head coordination from collected gaze and head trajectories. By experiments using the datasets collected from people who walked around under real and VR environments, we quantitatively confirmed that our approach was effective for gaze estimation. This result indicated that the eye-head coordination actually existed and thus the proposed method relying on the coordination was effective for the gaze estimation. It was also important findings that there was no significant difference between the models from the real and VR datasets in their estimation accuracy.

As future work, we plan to increase the number of participants and variations of scenes/tasks to analyze consistency or differences of the eye-head coordination in order to construct a general coordination model. In addition, we need to find the optimum values for T_s and T_e. It is also an interesting topic to extend consumer VR goggles to those equipped with gaze estimation function. If the gaze information can be obtained in those goggles, it gets possible to subjectively present high-quality scenes by allocating resources preferentially to the scene ahead of his/her gaze direction even in environments with limited graphic resources, as described in [29].

Table 6. The performance of models. Test data was the VR dataset and training data was the real dataset.

	MAE (deg)	
ID	Horizontal	Vertical
09	8.6	9.8
10	8.2	10.2
11	10.1	10.0
12	8.8	10.7
13	9.1	9.0
14	10.0	10.5
15	8.6	10.8
16	8.3	8.4
17	8.8	10.5
18	8.8	9.0
19	9.0	9.4
20	7.8	10.1
21	7.9	10.1
22	9.6	11.1
23	7.9	10.9
24	8.9	11.2
Average	8.8	10.1

Table 7. Comparison of real and VR datasets (horizontal)

		Test data	
		Real	VR
Training data	Real	7.5	8.8
	VR	7.3	8.3

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP18H03312 and JP19H00635.

References

1. Y. Fang, M. Emoto, R. Nakashima, K. Matsumiya, I. Kuriki, S. Shioiri, "Eye-position distribution depending on head orientation when observing movies on ultrahigh-definition television," ITE Transactions on Media Technology and Applications, Vol.3, No.2, pp.149–154, 2015.
2. Y. Fang, R. Nakashima, K. Matsumiya, I. Kuriki, S. Shioiri, "Eye-head coordination for visual cognitive processing," PLoS ONE, Vol.10, No.3, e0121035, 2015.
3. T. Okada, H. Yamazoe, I. Mitsugami, Y. Yagi, "Preliminaly analysis of gait changes that correspond to gaze direcions," 2nd IAPR Asian Conference on Pattern Recognition, pp.788–792, 2013. pp. 788–792, Nov. 2013.

4. J. Steen, "Vestibulo-Ocular Reflex (VOR)," Encyclopedia of Neuroscience, pp.4224–4228, 2009.
5. C. H. Morimoto, A. Amir, M. Flickner, "Detecting eye position and gaze from a single camera and 2 light sources," International Conference on Pattern Recognition, pp.314–317, 2002.
6. S.-W. Shih, J. Liu, "A novel approach to 3-d gaze tracking using stereo cameras," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol.34, No.1, pp.234–245, 2004.
7. D. H. Yoo, M. J. Chung, "A novel non-intrusive eye gaze estimation using cross-ratio under large head motion," Computer Vision and Image Understanding, Vol.98, No.1, pp.25–51, 2005.
8. C. Hennessey, B. Noureddin, P. Lawrence, "A single camera eye-gaze tracking system with free head motion," ACM International Symposium on Eye Tracking Research & Applications, pp.87–94, 2006.
9. T. Ishikawa, S. Baker, I. Matthews, T. Kanade, "Passive driver gaze tracking with active appearance models," 11th World Congress on Intelligent Transportation Systems, 2004.
10. J. Chen, Q. Ji, "3d gaze estimation with a single camera without ir illumination," International Conference on Pattern Recognition, pp.1–4, 2008.
11. H. Yamazoe, A. Utsumi, T. Yonezawa, S. Abe, "Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions," ACM International Symposium on Eye Tracking Research & Applications, pp.245–250, 2008.
12. R. Valenti, N. Sebe, T. Gevers, "Combining head pose and eye location information for gaze estimation," IEEE Transactions on Image Processing, Vol.21, No.2, pp.802–815, 2012.
13. S. Baluja, D. Pomerleau, "Non-intrusive gaze tracking using artificial neural networks," Technical report, DTIC Document, 1994.
14. K.-H. Tan, D. J. Kriegman, N. Ahuja, "Appearance-based eye gaze estimation," Winter Conference on Applications of Computer Vision, pp.191–195, 2002.
15. O. Williams, A. Blake, R. Cipolla, "Sparse and semisupervised visual mapping with the S3GP," IEEE Conference on Computer Vision and Pattern Recognition, pp.230–237, 2006.
16. W. Sewell, O. Komogortsev, "Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network," ACM CHI Conference on Human Factors in Computing Systems, pp.3739–3744, 2010.
17. F. Lu, Y. Sugano, T. Okabe, Y. Sato, "Adaptive linear regression for appearance-based gaze estimation," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.36, No.10, pp.2033–2046, 2014.
18. F. Lu, T. Okabe, Y. Sugano, Y. Sato, "Learning gaze biases with head motion for head pose-free gaze estimation," Image and Vision Computing, Vol.32, No.3, pp.169–179, 2014.
19. X. Zhang, Y. Sugano, M. Fritz, A. Bulling, "Appearance-Based Gaze Estimation in the Wild," IEEE Conference on Computer Vision and Pattern Recognition, pp. 4511–4520, 2015.
20. C. D. McMurrough, V. Metsis, J. Rich, F. Makedon, "An eye tracking dataset for point of gaze detection," ACM International Symposium on Eye Tracking Research & Applications, pp.305–308, 2012.
21. J. Choi, B. Ahn, J. Parl, and I. S. Kweon, "Appearance-based gaze estimation using kinect," 10th International Conference on Ubiquitous Robots and Ambient Intelligence, pp.260–261, 2013.

22. J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, pp.1189–1232, 2001.
23. T. Maesako, T. Koike, "Measurement of coordination of eye and head movements by sensor of terrestrial magnetism," Japanese Journal of Physiological Psychology and Psychophysiology, Vol.11, No.2 pp.69–76, 1993.
24. G. M. Jones, D. Guitton, A. Berthoz, "Changing patterns of eye-head coordination during 6h of optically reversed vision," Experimental Brain Research, Vol.69, No.3, pp.531–544, 1988.
25. Eye Mark Recorder EMR-9, NAC Inc., http://www.nacinc.com/
26. C. Jian-nan, Z. Peng-yi, Z. Si-yi, Z. Chuang, H. Ying "Key Techniques of Eye Gaze Tracking Based on Pupil Corneal Reflection," Proc. the 2009 WRI Global Congress on Intelligent Systems, Vol.02, pp.133–138, 2009.
27. I. Mitsugami, N. Ukita, M. Kidode, "Estimation of 3D Gazed Position Using View Lines," 12th International Conference on Image Analysis and Processing, 2003.
28. J. Murakami, I. Mitsugami, "VR-based Eye and Head Motion Collection for Modeling Their Coordination," IEEE 8th Global Conference on Consumer Electronics, 2019.
29. L. Xiao, A. Kaplanyan, A. Fix, M. Chapman, D. Lanman, "DeepFocus: Learned Image Synthesis for Computational Displays," ACM SIGGRAPH Asia, 2018.