RGBD Based Gaze Estimation via Multi-task CNN

Dongze Lian[1]*, Ziheng Zhang[1]*, Weixin Luo[1], Lina Hu[1], Minye Wu[1]
Zechao Li[2], Jingyi Yu[1], and Shenghua Gao[1]
[1]ShanghaiTech University [2]Nanjing University of Science and Technology

AAAI-19: Thirty-Third AAAI Conference on Artificial Intelligence
January 27 – February 1, 2019, Hilton Hawaiian Village, Honolulu, Hawaii, USA

Introduction:

➢ Gaze estimation:

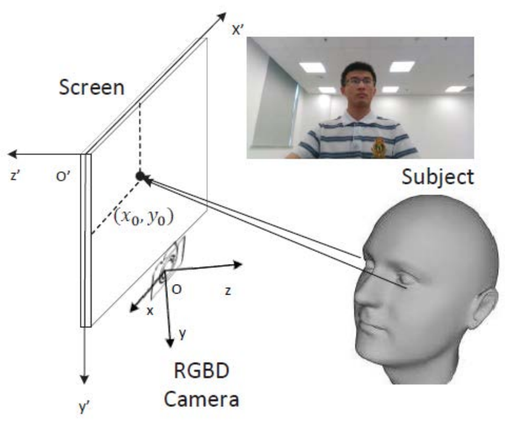
- Estimate which direction or which point on the screen target one person is looking at.

➢ Applications:

- Human-computer interaction;
- Eye tracker;
- Attention analysis.

➢ Challenges:

- Head pose sensitivity;
- Illumination inconsistencies;
- Occlusions;
- Low image quality;
- Accuracy varies across subjects.



➢ Contributions:

- Decompose gaze point estimation into eyeball pose, head pose, and 3D eye position estimation;
- CNN-based multi-task learning network to simultaneously refine depth map and predict gaze point;
- A large-scale RGBD gaze estimation dataset;
- State-of-the-art performance.

Our approach:

➢ Motivation:

- Improve head pose and 3D eye position representation with depth information.

➢ Method:

- Depth complements head pose and 3D eye position information, however, the raw depth images contain noises. We apply GAN to refine the depth maps and predict gaze points simultaneously via multi-task CNN.

Network architecture:

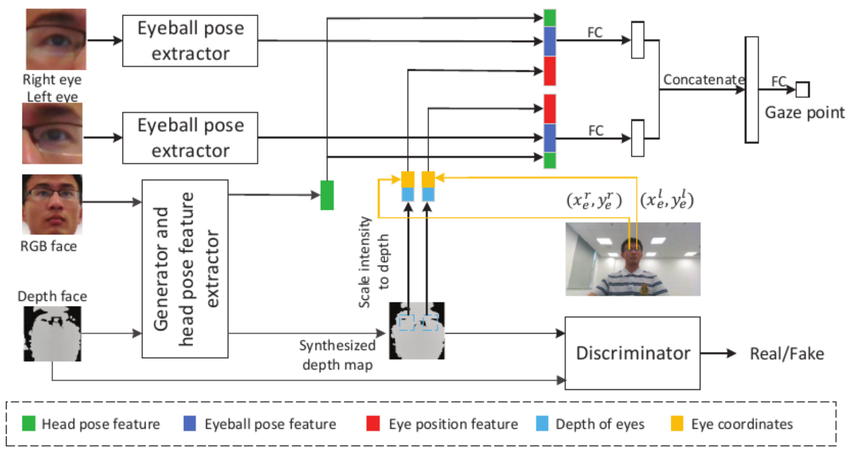


Figure 4: The network architecture for gaze point prediction. Eyeball pose features are extracted from two single-eye images. Head pose features are obtained from RGB and depth images. 3D eye positions are determined by eye coordinates and depth of eyes. Finally, all features are combined to predict gaze point (Best viewed in color).

➢ Depth image refinement based on GAN:

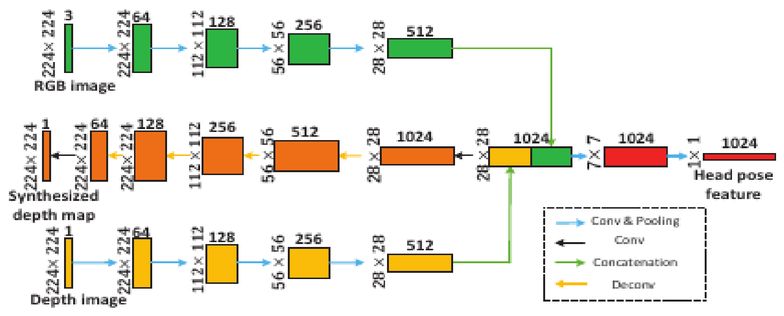


Figure 5: The network architecture of generator and head pose feature extractor.

- Adversarial loss:

$$l_g = \mathrm{E}\left[\log\left(\mathrm{D}(\mathrm{G}(\mathrm{I}^{\mathrm{d}},\ \mathrm{I}^{\mathrm{RGB}}))\right)\right]\sim \mathrm{G};$$

- Reconstruction loss:

$$l_{l1} = \frac{1}{M}\sum_i \left\| G\left(I_i^d(\Omega), I^{RGB}\right) - I_i^d(\Omega)\right\|_1;$$
$$\Omega = \left\{(\mathrm{x},\mathrm{y})\middle|\mathrm{I}_{\mathrm{i}}^{\mathrm{d}}(\mathrm{x},\mathrm{y}) \neq 0\right\}.$$
I_i^d, I^{RGB}: the i^{th} RGBD image pair.
M: the number of images.

- Discriminator: $\mathrm{l_d} = \mathrm{E}\left[\log\left(\mathrm{D}\left(\mathrm{I}^{\mathrm{d}}\right)\right)\right] + \mathrm{E}\left[\log\left(1-\mathrm{D}\left(\mathrm{G}\left(\mathrm{I}^{\mathrm{d}},\ \mathrm{I}^{\mathrm{RGB}}\right)\right)\right)\right]\sim \mathrm{D}.$

➢ Gaze point estimation network:

- Eyeball pose estimation: extract features from two single-eye images.
- Head pose estimation: fully encoded by RGBD image implicitly.
- 3D eye position extraction:
  1) the coordinates of the left and right eye centers $(\mathrm{x_e^l}, \mathrm{y_e^l})$ and $(\mathrm{x_e^r}, \mathrm{y_e^r})$.
  2) depth values of left and right eyes: $\mathrm{z_e^l}$ and $\mathrm{z_e^r}$.
- Gaze point estimation: $l_{gp} = \frac{1}{M}\sum_i^M \left\|\hat{p}^i - p^i\right\|_2^2$
- p^i: ground-truth for the i^{th} training image pair; $\hat{p}^i$: corresponding prediction.

ShanghaiTechGaze+ Dataset:

➢ Data collection:

- Screen target: 27-inch Apple iMac machine;
- RGBD camera: Intel RealSense SR300;

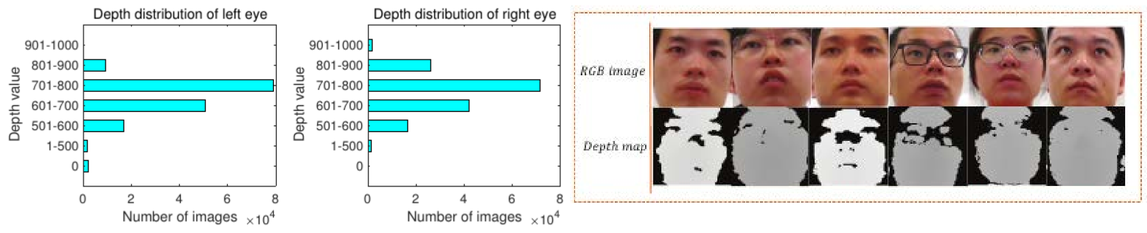
➢ Data acquisition:

- Clicking a white dot on the screen target and a blue dot will be generated after clicking;
- The clicking action makes the participant concentrate on the dot;
- The distance between two dots help us to judge reliability;
- 218 participants (141 males, 77 females), 165,231 RGB/depth pairs in total.

➢ Dataset comparison:

Table 1: Statistics of our dataset with some publicly available datasets. Abbreviations: *cont.* for continuous, *illum.* for illumination.

Dataset	#Participants	#Poses	#Targets	Illum.	#Images	#Views	Modality
UT-Multiview	50	8 + synth.	160	1	64,000	8	RGB
OMEG	50	3 + cont.	10	1	45,000	1	RGB
MPIIGaze	15	cont.	cont.	cont.	213,659	1	RGB
TabletGaze	51	cont.	35	cont.	videos	1	RGB
iTracker	1474	cont.	cont.	cont.	2,445,504	1	RGB
Free-head	200	cont.	cont.	cont.	240,000	12	RGB
ShanghaiTechGaze	137	cont.	cont.	cont.	233,796	3	RGB
EYEDIAP	16	cont.	cont.	2	videos	1	RGBD
ShanghaiTechGaze+ (ours)	218	cont.	cont.	cont.	165,231 pairs	1	RGBD



- Occlusion, illumination, specularity of glasses, out-of-valid range issues.

Experiments:

➢ Datasets:

- ShanghaiTechGaze+ (ours): 159 subjects for training, 59 for validation.
- EYEDIAP: divide 14 subjects into 5 groups for cross-validation.

➢ Evaluation metrics:

- Gaze point: $d_e = \frac{1}{M}\sum_i^M \left\|p^i - \hat{p}^i\right\|_2$
- Gaze direction: $a_e = \frac{1}{N}\sum_i^N arccos\frac{\langle a^i, \hat{a}^i\rangle}{|a^i||\hat{a}^i|}$

➢ Performance comparison:

- Gaze point estimation results on ShanghaiTechGaze+ (ours):

Table 2: Performance comparison of gaze point estimation on our dataset. (unit: mm)

Methods	d_e
Multimodal CNN (Zhang et al. 2015)	67.2
iTracker (Krafka et al. 2016)	55.5
iTracker* (Krafka et al. 2016)	47.5
Spatial weights CNN (Zhang et al. 2017)	60.6
Our method	**38.7**

- Gaze direction estimation results on EYEDIAP:

Table 3: Performance comparison of gaze direction estimation on EYEDIAP. (unit: degree)

Methods	a_e
Multimodal CNN (Zhang et al. 2015)	10.2 (2.9)
iTracker (Krafka et al. 2016)	8.3 (1.7)
iTracker* (Krafka et al. 2016)	5.7 (1.1)
Spatial weights CNN (Zhang et al. 2017)	6.0 (1.2)
Ghiass *et al.* (Ghiass and Arandjelovic 2016)	7.2 (1.3)
Our method	**4.8 (0.7)**

➢ Analysis:

- Effective decomposition strategy;
- Depth improves head pose and 3D eye position representation;
- Even without pixel-wise accurate depth, GAN-based multi-task method can still simultaneously refine the depth and improves gaze estimation.

Future work:

- Consider to solve the reflection and occlusion of eyeglasses.
- Combine multi-view eye images.

Reference:

[1] Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2015. Appearance-based gaze estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4511–4520.

[2] K. A. F. Mora, F. Monay, and J.-M. 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. ACM, 2014, pp. 255–258.

[3] Lian, D.; Hu, L.; Luo,W.; Xu, Y.; Duan, L.; Yu, J.; and Gao,S. 2018. Multiview multitask gaze estimation with deep convolutional neural networks. IEEE Transactions on Neural Networks and Learning Systems 1–14.

[4] Zhu, W., and Deng, H. 2017. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3143–3152.

Code & dataset: https://github.com/svip-lab/RGBD-Gaze