# Kappa Angle Regression with Ocular Counter-Rolling Awareness for Gaze Estimation
# Supplemental Material

Shiwei Jin, Ji Dai, Truong Nguyen
ECE Dept. UC San Diego.
{sjin, jid046, tqn001}@eng.ucsd.edu

## A. Dataset

### A.1. Synthetic Eye Images

Synthetic eye images are normalized from the synthetic face images generated by ST-ED [13]. ST-ED utilizes the 'face' gaze instead of the 'eye' gaze during gaze redirection. The gaze directions are defined by the gazing target point and source point. The main difference between the 'face' and 'eye' gaze comes from the source point's 3D locations. The source point of the 'face' gaze is the midpoint between two eye centers. Thus we only have one 'face' gaze direction for each face image. As for the 'eye' gaze, the corresponding source point is the center of the eye, which means we have two 'eye' gaze directions for each face image.

To make the redirection of the 'face' gaze consistent with our 'eye' gaze estimation task, we calculate the gazing target location instead of the gaze direction during the preprocessing and redirection processes. To be specific, we first normalize face images given the preprocessing requirements of ST-ED. After normalization, in addition to saving the normalized (rotated) gaze directions, we also keep the normalized gazing target location. During the process of redirecting the gaze directions, we assume that the gazing target maintains the same distance to the source point. Then we normalize [11] the redirected face images given the 'repositioned' gazing target to acquire normalized eye images with the 'eye' gaze directions. Fig. 1 shows several normalized real and synthetic eye images with the dataset provided (left three columns) or assigned 'eye' gaze directions (right three columns).

### A.2. Gaps between Real and Synthetic Data

The gap between real and synthetic data shown in Figure 4 does not provide conclusive evidence that the unobserved Kappa Angle is the cause. To investigate further, we mixed the real and synthetic data in training Diff-NN to determine if the unobserved person-dependent component was eliminated. However, as illustrated in Section 4.4, the mixture of real and synthetic data performs even worse than real data alone, providing further evidence of the absence of the Kappa Angle of the synthetic data.

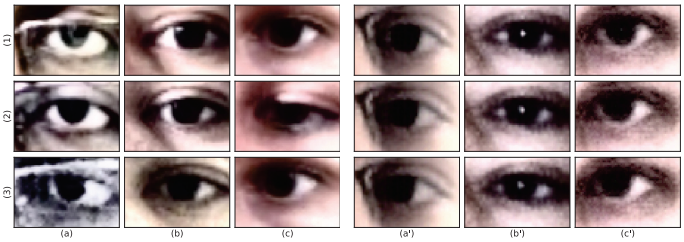


Figure 1. The comparison between real (left three columns) and synthetic (right three columns) eye images. Columns (a) to (c) are eye images from different persons in MPIIFaceGaze [12]. Columns (a') to (c') are synthetic images generated from the three previous columns, respectively. Rows (1) to (3) represent the gaze direction with the same pitch ($-5$ degrees) and the changed yaw of 5, 10, 15 degrees, respectively.

### A.3. Roll Distribution

With the leave-one-subject-out protocol, each subject's data was viewed as the test subset one time. Thus we calculated the distribution of the roll of the head pose before normalization given the whole dataset. Fig. 2 presents the distribution given MPIIFaceGaze [12] and EYEDIAP [1] (SP and DP, respectively). We can notice that in EYEDIAP (DP), the distribution of the roll of the head is wider compared to MPIIFaceGaze. On the other hand, EYEDIAP (SP) exhibits the smallest range of roll of the head.

## B. Implementation Details

### B.1. Network Architecture

The network only contains one single branch built with three convolutional layers and three fully connected layers. The convolutional part's structure inherits from the Diff-NN

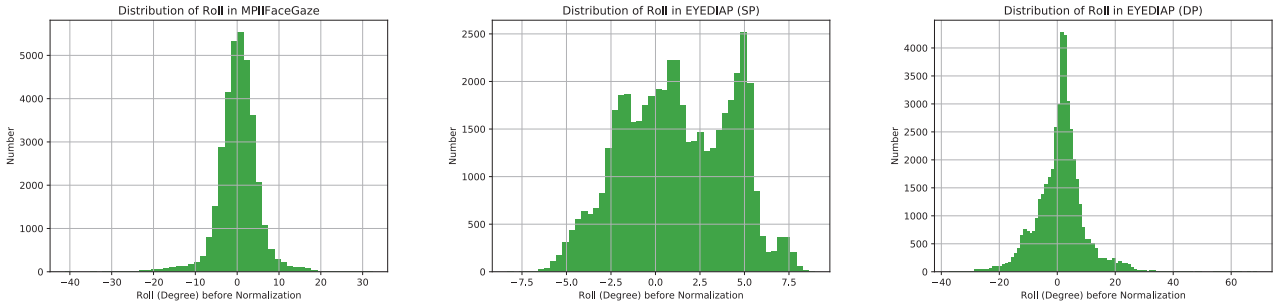


Figure 2. Distribution maps of the roll of the head pose before normalization in two benchmark datasets. These distribution maps revealed the presence of the OCR response when the roll of the head is not in a zero position. Since we utilized 'leave-one-subject-out' protocol, each sample in MPIIFaceGaze and EYEDIAP was included in the test subset.

[6]. Each time, we feed only one eye image $\boldsymbol{I}^e \in \mathbb{R}^{H\times W\times C}$ into the network where $(H, W, C) = (48, 72, 3)$. The extracted features from the convolutional part are fed into the fully connected part. The corresponding head pose is concatenated with the output from the first fully connected layer. Then the last two fully connected layers are applied to the concatenated output to estimate the optical axis direction with yaw and pitch components.

The proposed network architecture has several advantages. The first point is the simple network structure. Instead of accompanying several inputs and branches for one output, our proposed network only takes one input but still has comparable prediction performance. The second point is overfitting avoidance. The differential method needs a dropout layer to avoid overfitting, which can cause bad performance on the new participant data. However, the proposed KAComp-Net could proactively prevent this obstacle by eliminating dropout layers.

### B.2. Training Parameters

We train KAComp-Net with 12 epochs and a batch size of 128. The initial learning rate is set as 0.01. After each epoch, the learning rate is divided by 2. The optimizer is Adam [4], with a weight decay coefficient of 0.09. We use the default momentum value of $\beta_1 = 0.9, \beta_2 = 0.999$.

### B.3. Gaze Inference

In the testing phase, we randomly pick a certain number $(M)$ of calibrated samples $F^e$ with the gaze directions. We first feed these samples into the KAComp-Net to estimate their optical axis directions. Then given the provided gazes and the OCR response, we can derive several $\{\underline{\widehat{\boldsymbol{\kappa}}}_i, i \in [1, M]\}$ from $M$ calibrated images. We then utilize their average to represent the estimated Kappa Angle,

$$\underline{\widehat{\boldsymbol{\kappa}}} = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{R}_{OCR,i}^{-1} \cdot \left[\boldsymbol{g}^{gt}(\boldsymbol{F}_i^e) - \psi(\boldsymbol{F}_i^e)\right]. \quad (1)$$

Given the estimated Kappa Angle from calibrated samples, we can predict the gaze. The predicted gaze error of eye image $\boldsymbol{I}_j^e$ is

$$\boldsymbol{g}^{err}(\boldsymbol{I}_j^e) = \boldsymbol{g}^{gt}(\boldsymbol{I}_j^e) - \left[\psi(\boldsymbol{I}_j^e) + \boldsymbol{R}_{OCR,j} \cdot \underline{\widehat{\boldsymbol{\kappa}}}\right], \quad (2)$$

where the $\underline{\widehat{\boldsymbol{\kappa}}}$ is derived from Eq. (1).

## C. Discussion

### C.1. Ambiguity from the Camera Pose

Head poses in images can be modified due to different camera poses, even if the subject's head pose remains invariant. When using benchmark datasets, we can assume that the camera was placed horizontally based on data collection settings and clues from upper torsos, as discussed in Section 4.1. To ensure accurate estimation of head roll motion in practical scenarios, it is crucial to determine the camera's roll pose with respect to the horizontal level. If the camera can be placed statically, it can be manually calibrated to ensure it is positioned horizontally.

A possible alternative is capturing high-resolution iris images. Then the OCR response can be directly estimated from these images without relying on the derivation from head roll motion, as demonstrated in [7].

### C.2. Why Rotation

In the normalization step, when the roll of the head is normalized to an upright status, the ground truth gaze is transformed by a rotation matrix instead of an affine transformation matrix, as illustrated in [11]. This process is similar to our OCR compensation process. When OCR occurs, the eyeball has an undesired roll after normalization, which redistributes the pitch and yaw of the Kappa Angle. To counteract this redistribution caused by various roll statuses within the same subject's data, we apply rotation matrices to compensate, as shown in Eq. 4.

### C.3. Listing's Law and OCR

Ocular counter-roll (OCR) is a vestibulo-ocular reflex characterized by torsional rotations of the eye in response to lateral tilt of the head [8]. Listing's law states that when the head is fixed, there is an eye position called primary position, such that the eye assumes only those orientations that can be reached from primary position by a single rotation about an axis in a plane called Listing's plane [9]. Listing's law holds during fixation, saccades, smooth pursuit, and vergence, but fails during sleep and vestibulo-ocular reflex [10], including OCR.

When the head tilts to the side, OCR occurs, causing the eye to rotate around the roll axis that is out of Listing's plane. This means that the orientation of Listing's plane changes when the head is tilted, as shown in [2]. However, if the head maintains a static tilted posture, Listing's Law still applies, and eye rotation vectors are still confined to a plane. This plane is shifted along the torsional axis in relation to the upright position, proportional to the roll-tilt angle [3]. In this case, the eye orientation can still be represented by pitch and yaw components with a constant torsional bias. Our proposed KAComp-Net considers the bias caused by the OCR, and the remaining estimation processes are consistent with the cases where OCR doesn't happen.

### C.4. Limitations and Plans

Our proposed method has several limitations, both from the structural design perspective and the data perspective. These limitations are viewed as research directions for future work.

1) **Synthetic Data:** KAComp-Net requires synthetic data to aid in the learning process of the optical axis direction. Compared with the real data, synthetic data has less unobserved person dependent components of gaze directions, as shown in Section 3.2 and Section 4.4. Although we took advantage of this property regardless of the gap, we still need to get rid of the dependence on synthetic data. In the future, the network can learn unified features directly from real data without Siamese learning between data from different domains. This could potentially improve the estimation accuracy.

2) **Static / Dynamic OCR:** KAComp-Net only considers static OCR response, which is related to the roll of the head. However, during head tilt, dynamic OCR occurs with slow phases away from and quick phases toward the head tilt [5]. With a sustained head tilt, the static OCR occurs, resulting in a static change in torsional eye position in the direction away from the head tilt [8]. In future work, if we have access to consecutive frames, we can model the process by considering both static and dynamic OCR.

3) **High-Resolution Iris Images:** In the KAComp-Net pipeline, OCR needs to be derived from the roll motion of the head, which is normally abandoned after normalization due to the low resolution of eye images. However, if we have high-resolution eye images, we don't need to derive the OCR response. Instead, we can measure the OCR directly given the high-resolution iris images [7] for a more accurate gaze estimation.

## References


[1] 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. 1

[2] Joseph M Furman and Robert H Schor. Orientation of listing's plane during static tilt in young and older human subjects. *Vision research*, 43(1):67–76, 2003. 3

[3] Th Haslwanter, D Straumann, BJM Hess, and V Henn. Static roll and pitch in the monkey: shift and rotation of listing's plane. *Vision research*, 32(7):1341–1348, 1992. 3

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[5] R John Leigh and David S Zee. *The neurology of eye movements*. Contemporary Neurology, 2015. 3

[6] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation with calibration. In *BMVC*, volume 2, page 6, 2018. 2

[7] Kwang-Keun Oh, Byeong-Yeon Moon, Hyun Gug Cho, Sang-Yeob Kim, and Dong-Sik Yu. Measurement of ocular counter-roll using iris images during binocular fixation and head tilt. *Journal of International Medical Research*, 49(3):0300060521997329, 2021. 2, 3

[8] Jorge Otero-Millan, Carolina Treviño, Ariel Winnick, David S Zee, John P Carey, and Amir Kheradmand. The video ocular counter-roll (vocr): a clinical test to detect loss of otolith-ocular function. *Acta oto-laryngologica*, 137(6):593–597, 2017. 3

[9] Hermann Von Helmholtz. *Handbuch der physiologischen Optik*, volume 9. Voss, 1867. 3

[10] Agnes MF Wong. Listing's law: clinical significance and implications for neural control. *Survey of ophthalmology*, 49(6):563–575, 2004. 3

[11] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–9, 2018. 1, 2

[12] 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. 1

[13] Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. Self-learning transformations for improving gaze and head redirection. *Advances in Neural Information Processing Systems*, 33:13127–13138, 2020. 1