Dynamic 3D Gaze from Afar: Deep Gaze Estimation from Temporal Eye-Head-Body Coordination Supplementary Material

Soma Nonaka[†] Shohei Nobuhara[†‡] Ko Nishino[†]
[†]Graduate School of Informatics, Kyoto University [‡]JST, PRESTO
https://vision.ist.i.kyoto-u.ac.jp/

Evaluation of head/body orientation estimation Even though head and body orientation estimation is not our end goal, the proposed vMF network performs well as a standalone head and body orientation estimator. This can be attributed to the joint estimation architecture of head and body orientations, the use of body velocity, and the fact that they focus on the simpler tasks of 3D orientation estimation, in contrast to full head and body pose estimation (*i.e.*, we only estimate the yaw and pitch of the head orientation, and all three angles but for the whole body as body orientation).

We evaluated the head and body network on the validation set of the AGORA dataset [3]. We compare with the state-of-the art head pose estimation model, WHENet [5], which estimates the full head orientation (pitch, yaw, and roll) unlike our method which does not recover the roll as it is not necessary for gaze estimation. As far as we know, our model is the only model that can handle head poses with 360° degrees of yaw. The top rows in Tab. S1 show the MAE computed on the estimated pitch and yaw. Our model shows slightly better accuracy than WHENet.

We also tested our vMF network for body orientation estimation as a standalone estimator. We compare with the state-of-the-art 3D pose estimation model (SPIN [2]). For SPIN, we defined the body orientation as the outer product of the line connecting both shoulders and the line connecting the neck and the pelvis. As bottom rows in Tab. S1 show, our vMF network achieves better accuracy for whole body orientation estimation.

These results demonstrate the accuracy of our head and body orientation estimates. They are at least comparable to state-of-the-art methods and sufficient for estimating the gaze. These head and body orientation estimates may find applications in other tasks beyond gaze estimation.

	Method	MAE
Head	Fixed bias	85.5
	WHENet [5]	20.1
	Ours (Head)	17.6
Body	Fixed bias	90.0
	SPIN [2]	49.8
	Ours (Body)	17.2

Table S1. Comparison of head and body orientation estimation on the AGORA dataset. 3D mean angular errors (MAEs) are shown. Albeit only for the pitch and yaw for the head orientation and all three angles but only for the whole body (not its full pose), our vMF networks show slightly better accuracy when compared with state-of-the-art full head pose and body pose estimation methods. These results demonstrate the advantage of limiting the estimation to only those angles necessary for gaze estimation.

Uncertainties We examine the relationship between the estimated uncertainty and angular error. We use the reciprocal of estimated concentration (κ) as a measure of uncertainty. We compute the angular error and the uncertainty estimate for each test sample, and evaluate the mean uncertainty and angular errors for every 5 degrees in the ascending order of angular errors. As clearly shown in Fig. S1, there is a positive correlation between the estimated uncertainty and actual estimation errors in gaze directions ($r = 0.26$). These values would be useful for downstream tasks that use our gaze estimates as we demonstrate with our multiview extensions.

Effects of camera distance between camera and person To examine how camera distance affects the performance of our model, we evaluated estimation accuracy with respect to the camera distance. The estimation error was the smallest at 12.1° when the person is closest to the camera (below 1 m), and the largest at 29.0° when the person is farthest from the camera (above 7 m). For a typical in-room distance of 3 to 5 m, the mean error was 22.6°.

Annotation acquisition from the AGORA dataset When training, we compute the head and body orientations

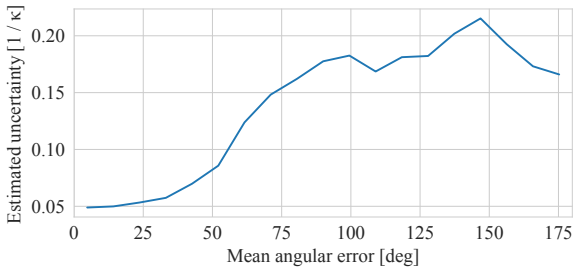

Figure S1. Estimated uncertainty versus angular error. The uncertainty estimates clearly positively correlate with the angular errors, suggesting that they faithfully quantify gaze direction uncertainty for downstream tasks.

from the 3D keypoints provided in the AGORA dataset and use them as ground truth. We define the body orientation as $B = l_s \times l_m$, where l_s is the line connecting the left and right shoulders, and l_m is the line along the torso. Specifically, l_s and l_m are computed by $l_s = x_{rs} - x_{ls}$, $l_m = x_n - x_{mh}$ where x_{rs}, x_{ls}, x_n, x_{mh} are the 3D coordinates of the right and left shoulders, neck, and mid-hip, respectively.

The head orientations for training are obtained from the 3D coordinates of facial keypoints following [5]. First, a reference camera matrix (R_{ref}) and reference keypoints (x_{ref}) are manually defined so that the camera looks at the face from the front. Then, the reference camera matrix is transformed so that the reference keypoints align with the actual keypoints, which produces the transformed camera matrix (R_{virt}). This R_{virt} is a virtual camera matrix that is looking at the front of the face. Finally, a rigid transform (T) from R_{virt} to the actual camera matrix (R_{real}) is computed. The head orientation is defined by $H = T \cdot [0, 0, -1]$.

Implementation details Given an input sequence of images, we use OpenPose [1] to detect 2D keypoints of a person, and crop the images so that they contain all head or body keypoints. For the backbone, we use EfficientNet-b0 [4] up to its final average pooling layer, and obtain its output feature of size = 1280×1. The extracted features from head and body images are concatenated to produce a vector of size = 2560×1 which is input to a GRU layer. The hidden size of the LSTM layer is 128. For the gaze estimation module, we use bidirectional LSTM with two layers of hidden size = 512. We use 7 video frames as inputs.

Runtime analysis Inference for 1 frame was 5.5 ms (180 fps) on a NVIDIA GTX 1080Ti GPU, Inference including head/body orientation estimation and gaze estimation for 1 frame takes about 5.5 ms (180 fps) on a NVIDIA GTX 1080Ti GPU. For this, we believe the computational speed is sufficient for real-time usage.

References

[1] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser A Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *TPAMI*, 2019. 2

[2] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to Reconstruct 3d Human Pose and Shape via Model-fitting in the Loop. In *Proc. ICCV*, 2019. 1

[3] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in Geography Optimized for Regression Analysis. In *Proc. CVPR*, June 2021. 1

[4] Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proc. ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. 2

[5] Yijun Zhou and James Gregson. WHENet: Real-time Fine-Grained Estimation for Wide Range Head Pose. In *Proc. BMVC*, 2020. 1, 2