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

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

Abstract

We introduce a novel method and dataset for 3D gaze estimation of a freely moving person from a distance, typically in surveillance views. Eyes cannot be clearly seen in such cases due to occlusion and lacking resolution. Existing gaze estimation methods suffer or fall back to approximating gaze with head pose as they primarily rely on clear, close-up views of the eyes. Our key idea is to instead leverage the intrinsic gaze, head, and body coordination of people. Our method formulates gaze estimation as Bayesian prediction given temporal estimates of head and body orientations which can be reliably estimated from a far. We model the head and body orientation likelihoods and the conditional prior of gaze direction on those with separate neural networks which are then cascaded to output the 3D gaze direction. We introduce an extensive new dataset that consists of surveillance videos annotated with 3D gaze directions captured in 5 indoor and outdoor scenes. Experimental results on this and other datasets validate the accuracy of our method and demonstrate that gaze can be accurately estimated from a typical surveillance distance even when the person's face is not visible to the camera.

Figure 1. We introduce a novel method for estimating the gaze direction of people (orange arrows) in videos captured from afar by leveraging the temporal coordination of the gaze, head, and body orientations in a Bayesian framework. Our method does not rely on the appearance of the eyes, and can tell the gaze direction even when the person is facing away from the camera. We introduce a new dataset for gaze estimation in the wild with ground truth annotation. Note that the markers, eye tracker, and body worn cameras are only used for ground truth annotation.

1. Introduction

What if we could continuously trace the gaze direction of a person from a distance, for instance, with cameras fixed to room and street corners? If we can, the practicality of gaze estimation will be significantly increased and its utility will be greatly expanded. It will allow us to use already installed surveillance cameras or those monitoring elderlies to follow the dynamically changing gaze of a person, which will let us gauge much deeper into the person's internal state not just her whereabouts.

Despite the large advances in gaze estimation research, especially by leveraging deep neural networks [9, 16, 32, 34, 35], most appearance-based methods cannot be applied to videos taken from a distance. This is because they inherently require a clear and close-up view of the eyes. For instance, most leading methods assume a person sufficiently close to the camera (ranging from $10\mathrm{cm}$ to $1\mathrm{m}$), or they are only applicable to the frontal view (up to $90°$) of a person. We target typical surveillance and monitoring views which may range from a few meters to $10\mathrm{m}$.

The few methods that demonstrate gaze estimation from surveillance images approximate the gaze with the head or body orientation, which is too crude for most downstream tasks [25, 26]. A recent method [6] does estimate gaze direction from surveillance cameras by regressing it from human body keypoints detected with OpenPose. The method, however, only estimates gaze in 2D (*i.e.*, in the image plane)

and it is validated only on a limited number of surveillance videos (1 scene with 2 cameras).

In this paper, we introduce a 3D gaze estimation method for videos of freely moving people taken from a distance, for typical room-sized surveillance scenarios. Our key idea is to fully leverage the temporal coordination of the gaze, head, and body orientations of a person to estimate the person's dynamically changing 3D gaze direction just from the head and body orientations which can be estimated reliably from afar. We show that we can estimate the gaze direction from a temporal sequence of head and body orientation estimates even without seeing the eyes at all.

We formulate gaze estimation as Bayesian prediction that capitalizes on a learned angular-temporal prior of the gaze direction conditioned on the head and body orientations. The gaze, head, and body angular orientations have strong but complex temporal dependencies. For example, when we look for something in the room, our eyes first move into the desired direction, and then our head moves to follow the eyes. When our head is following the saccade, our eyes move in the opposite direction to stabilize the image on the retina during head movement. These seemingly simple angular-temporal relationships are cascaded one after another and temporally blended together resulting in a complex dependency that can no longer be captured with a simple analytical model. We model this complex gaze-head-body coordination with a cascade of two learned deep networks that encode the head and body orientation likelihoods and gaze orientation conditioned on them, respectively, to leverage their rich dependency to infer the gaze direction just from the head and body orientations.

Given a sequence of video frames of a person captured at a distance, we first estimate the head and body orientations by devising a network that uses both the body appearance and 2D trajectory. These orientations are estimated as von Mises-Fisher distributions to canonically represent their uncertainties. These head and body orientation likelihoods are then multiplied with a learned conditional prior of the gaze direction given the head and body orientations that encode their natural temporal coordination. We model this conditional prior with a network that encodes the temporal dependency of gaze direction in each frame on past and future head and body orientations. Optionally, we extend our method to opportunistically leverage the eye appearance when they happen to be visible and multiview head and body appearance when we have access to multiple cameras.

We introduce a new dataset of annotated surveillance videos of freely moving people taken from a distance in both indoor and outdoor scenes. The videos are captured with multiple cameras placed in eight different daily environments. People in the videos undergo large pose variations and are frequently occluded by various environmental factors. Most important, their eyes are mostly not clearly visible as is often the case in surveillance videos. We introduce the first rigorously annotated dataset of 3D gaze directions of freely moving people captured from afar. Through extensive evaluation using this new dataset, we show that our method enables accurate 3D gaze estimation from afar. We also demonstrate that our method generalizes well to different scenes and camera poses. All data and code can be downloaded from our project web page.

2. Related works

Gaze Estimation Models Gaze estimation methods can be roughly categorized into geometry-based and appearance-based. Geometry-based methods use 3D eye models to exploit geometric or optical characteristics of the eye [12, 14, 19, 36]. In exchange for high accuracy, these methods usually require detailed information of the eyes, which often necessitates hardware eye trackers. In contrast, appearance-based methods directly estimate the gaze direction from images of eyes. Recent methods often use deep neural networks to learn this mapping, and achieve high accuracies [9, 32–35]. These methods, however, naturally require a close-up frontal view of the target face or eyes, which are not available in images and videos taken from a distance (*e.g.*, more than a meter). Recently, Gaze360 [16] created a large-scale dataset that contains head poses with 360° of yaw, and showed that a model trained on their dataset can estimate gaze even when the person is facing backward. Although Gaze360 contains diverse appearances, their dataset only contains humans standing still and limited head poses in terms of pitch and roll. In contrast, our dataset is of freely moving people. Our dataset also contains images of people with a wide, natural variety of head poses reflecting realistic surveillance and monitoring scenarios.

Past methods for gaze estimation from surveillance images, in which eyes are basically not clearly or not at all visible, typically use head orientation as a surrogate of gaze direction [25, 26]. These methods are robust to low image quality but the head orientation is rarely the true gaze direction. Dias *et al*. used facial keypoints to estimate 2D gaze in still images, and evaluated the model with their dataset of surveillance images manually annotated with 2D gaze directions [6]. In contrast, we estimate dynamically changing 3D gaze direction in video. Gaze direction recovered in 3D has wide utility in down stream applications as it allows 3D reasoning of a person's attention in the environment.

Gaze, Head, and Body Relationship. Head and eye coordination has been studied extensively [1, 27, 31]. A linear relationship between gaze and head orientations can be observed, for example, during watching movies [7] or daily activities such as tea making [18]. Vestibulo-Ocular Reflex (VOR) characterizes the coordinated temporal movements of them; the eye moves in the opposite direction when the head is moving to fixate the retinal image [1]. Various

research have studied the relation between gaze and body as well as gaze and head orientations [8, 18]. Yamazoe *et al*. [30] reported that a linear relationship, similar to that of gaze and head orientations, was observed between the gaze and body orientations during free walking. Murakami *et al*. recently showed that the gaze direction can be estimated from head and body orientations when their true directions are known using a simple regression model [20], which corroborates our intuition. Estimating gaze in actual surveillance views without known head and body orientations, however, requires a significant leap, which we make by seamlessly integrating learned angular-temporal relation of the complex eye-head-body coordination in a canonical Bayesian framework that can be learned end-to-end. Our method also models and propagates estimation uncertainties in a principled manner.

Head and Body Orientation Estimation. Early head pose estimation methods used facial landmarks to align a geometric template [3]. Recent methods rely on deep neural networks and large image databases to directly estimate the orientation of the head from its appearance [25]. Zhou and Gregson [37] show that head pose can be estimated even when the subjects are looking away from the camera. These methods, however, can suffer from gimbal lock and are not suitable for images with extreme head or camera poses. We deal with this problem by estimating only the yaw and pitch of head orientation, since the range of roll is relatively small and hardly affects gaze estimation.

Human body orientation estimation has been widely studied particularly for behavior analysis such as movement prediction. A large body of work has demonstrated accurate 2D body orientation estimation from images [5, 13, 23]. As with other tasks, recent methods greatly improve the accuracy by using deep neural networks to directly estimate the orientation of the body from its appearance. For example, Wu *et al*. annotated 2D body orientations in 55K images from the COCO dataset, and showed that their method generalizes well across different camera poses and backgrounds [29]. Two-dimensional body orientation is, however, insufficient for us to estimate 3D gaze as the pitch can also vary greatly. For this, we estimate 3D body orientation in the same manner as we estimate head orientations.

3. Bayesian Gaze from Head and Body

Our goal is to estimate the gaze direction of a person without relying on clear appearance of her eyes but instead on the coordinated head and body movements. Figure 2 shows an overview of our framework. We formulate gaze estimation as Bayesian prediction where we estimate the likelihoods of head and body orientations given an input image, and then multiply a learned conditional temporal prior of gaze direction by cascading two neural networks.

3.1. Head and Body Network

We first estimate head and body orientations from an input video. Instead of having independent networks for each, we build a network that simultaneously estimates head and body orientations from whole body images to reduce the computational cost. The network also takes the binary mask of the head bounding box, which is used to determine the head position in the image. In addition, the network exploits the in-image 2D velocity of the person to better estimate the body orientation. The head bounding box and the velocity of the person are normalized with respect to the height of the person in the image.

Figure 2 left shows the architecture of the head and body orientation estimation network. First, the network processes multiple frames of whole body images to extract shared features (**Shared conv**). The mask of the head position is downsampled by an average pooling layer so that it aligns the size of the feature map. The shared features and the head mask are multiplied together and input to convolutional layers (**Head Conv**), or are directly input to another set of convolutional layers (**Body Conv**). The output of Head Conv and Body Conv layers and the body velocity are concatenated and fed into an LSTM layer to jointly estimate the head and body orientations. For the backbone network, we used the first two convolutional layers of EfficientNet-b0 [28] for Shared Conv, and later layers of EfficientNet-b0 for Head Conv and Body Conv layers.

To canonically model uncertainties, we estimate head and body orientations as 3D von Mises-Fisher (vMF) distributions. A vMF distribution is a spherical directional statistics distribution represented by two parameters $\{\mu, \kappa\}$

$$\mathrm{vMF}(x; \mu, \kappa) = \frac{\kappa}{4\pi \sinh \kappa} \exp\left[\kappa x^T \mu\right] , \qquad (1)$$

where μ is a 3D unit vector that represents the mean direction, and κ is the concentration parameter (the larger κ, the higher confidence). This vMF distribution gives us a natural interpretation of the directional estimate and its uncertainty.

To constrain the output to become a valid vMF distribution, we introduce a final layer to the network by extending the method of [2, 22] to 3D. This layer consists of two branches to estimate the mean direction (μ) and the concentration (κ), respectively. The mean direction branch (f_μ) makes the output a unit vector by performing normalization on a fully connected layer and the concentration branch (f_κ) makes the output positive with a Softplus function

$$f_\mu(x; W, b) = \frac{Wx + b}{\|Wx + b\|} = \begin{pmatrix} \sin\phi\cos\theta \\ \sin\phi\sin\theta \\ \cos\phi \end{pmatrix} = \hat{\mu} \qquad (2)$$

$$f_\kappa(x; W, b) = \mathrm{Softplus}(Wx + b) = \hat{\kappa} , \qquad (3)$$

where x, W, b are the input, weight, and bias, respectively.

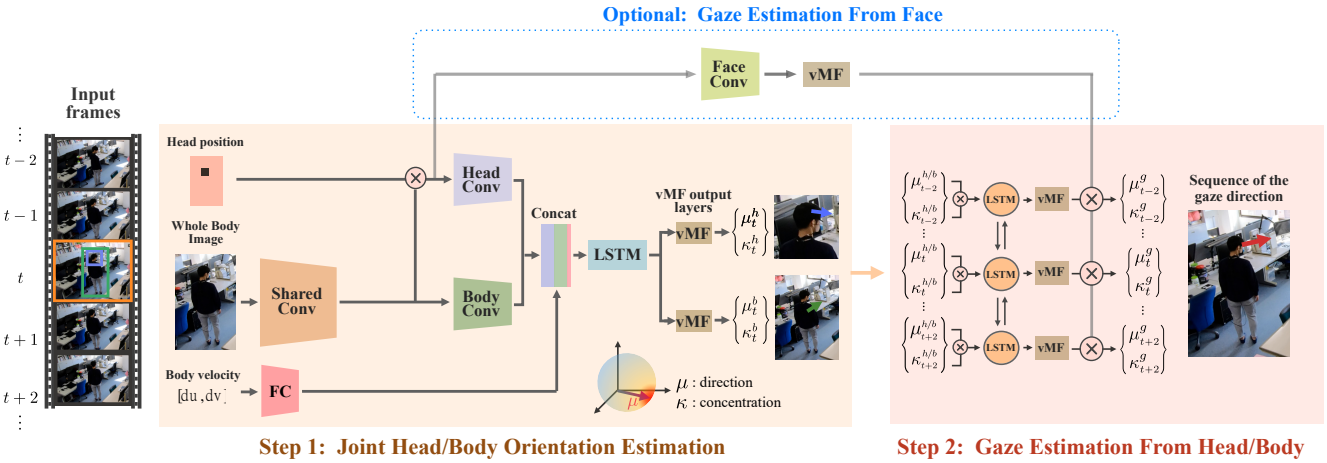

Figure 2. Model architecture. The head and body estimation network receives whole body images, head position, and body velocity, and outputs directional statistics (vMF) distributions as their orientation estimates. Both distributions for multiple frames are fed into the gaze estimation network that outputs the sequence of gaze direction estimates as vMF distributions. Optionally, we add a network that can opportunistically estimate the gaze direction directly from the face appearance. The estimated gaze directions are integrated to produce the final gaze estimate as vMF distributions.

We learn the vMF network parameters with maximum likelihood estimation. For this, we define the loss function as the negative log-likelihood of the vMF distribution

$$L_{\text{vMF}} = -\ln \hat{\kappa} + \ln \sinh \hat{\kappa} - \hat{\kappa} \mu^T \hat{\mu} \,. \tag{4}$$

We optimize both the direction ($\hat{\mu}$) and the concentration ($\hat{\kappa}$) with this loss while alternatingly fixing each other.

We train the vMF network using a 3D human pose dataset. In particular, we use the training set of the AGORA dataset [21], which contains realistic human models placed in various rendered 3D scenes. To deal with low image resolution, we randomly reduce the resolution of input images between $\times 0.1$ to $\times 0.9$. Please see the supplemental material for visualization.

3.2. Head–Body Conditional Temporal Gaze Prior

The second step of our Bayesian formulation is to estimate the sequence of gaze directions from the distributions of estimated head and body orientations. That is, given the head and body orientation likelihoods, we now want to multiply the conditional temporal prior of the gaze direction given those orientations. Our key idea is to learn the complex angular-temporal dependency between gaze and head, and gaze and body orientations with a recurrent neural network. Note that it is a temporal prior, not an instantaneous static one, that embodies the gaze coordination with the head and the body as a dynamically changing system. For this, we use a bidirectional LSTM which consists of two bidirectional LSTM layers and a final layer to output the parameters of a vMF distribution. It takes head and body orientation estimates as inputs, and outputs the parameters of the gaze vMF distribution (Fig. 2 right).

Another key idea is to modulate the estimated head and body orientations by their estimated concentrations, which lets us deal with highly uncertain situations such as when either the head or the chest is not observable. For this, before inputting the head and body orientations, we modulate them with their estimated uncertainty to make our model robust to occlusions. The weighted direction of body $\kappa^b(\mu_x^b, \mu_y^b, \mu_z^b)$ and head $\kappa^h(\mu_x^h, \mu_y^h, \mu_z^h)$ are concatenated to produce a 6D vector and input to the bidirectional LSTM model.

3.3. Opportunistic Eye Appearance Integration

When a freely moving person is seen from afar, *e.g.*, in a surveillance view, the eyes are rarely (in our dataset, only less than 6% of all the frames) clearly visible. Nevertheless, when they are visible, we may leverage their appearance. For this, we extend our framework to integrate (but not rely on) the eye appearances.

As depicted in the upper part of Fig. 2, we add a set of convolutional layers that are the same as the Head Conv layer but directly estimates gaze direction from the appearance of the eyes. The gaze direction estimated from eye appearance and from gaze-head-body coordination are combined by weighting the estimated directions with their associated uncertainties. This enables us to exploit eye appearance only when the associated uncertainty is low, *i.e.*, opportunistically when the eyes are clearly visible.

3.4. Multi-view Gaze Estimation

Multiple surveillance or monitoring cameras are often installed in a single location. In such cases, we may opportunistically leverage the multiview observations to gain further gaze estimation accuracy. We propose two extensions for integrating multiview video feeds.

The difficulty of gaze estimation varies depending on the appearance of the person in each camera. For example, the person may be clearly visible from one of the cameras but occluded from another. When occluded, the estimated gaze has low certainty. We therefore combine the estimated gaze directions from each view with their associated uncertainties after converting them into the world coordinate. This is naturally done by maximizing the sum of the log likelihoods of the vMF distributions

$$\overline{\mu} = \mathrm{argmax}_{\mu} \left[\Sigma_i \{-\ln \hat{\kappa}_i + \ln \sinh \hat{\kappa}_i - \hat{\kappa}_i \mu^T \hat{\mu}_i\}\right] \quad (5)$$

$$= \frac{r}{\|r\|}, r = \Sigma_i \hat{\kappa}_i \hat{\mu}_i \,, \quad (6)$$

where i denotes the camera index.

In addition, we also test the use of 3D body velocity of the person in the head and body network. We obtain the 3D position of the body center by triangulating the 2D body centers from the multiple views, and compute the 3D body velocity from them. We input the 3D velocity instead of the 2D velocity to our head and body network and integrate the head and body orientations from multiple cameras by weighted averaging. The gaze direction is estimated from the same network as introduced in Sec. 3.2.

4. Gaze from Afar Dataset

Study on gaze estimation from afar, particularly from typical surveillance views, necessitates a thorough dataset of videos capturing people in their natural settings with various postures but with frame-wise annotation of their gaze directions. In particular, videos of people freely roaming around in daily environments would be preferable. Moreover, we are interested in 3D gaze estimation, not just 2D estimates that would be view-dependent. Previous gaze datasets [9, 10, 16, 17, 35] contain only close-up images of faces or images of people standing still. Although a surveillance image dataset [6] for gaze estimation has been introduced in the past, it only contains 1 scene with 2 cameras with manually annotated 2D gaze directions.

We introduce Gaze from Afar (GAFA) Dataset which contains videos of freely moving people taken in 5 different daily environments including a kitchen, library, laboratory, living room, and courtyard. In contrast to previous datasets, our dataset contains long-term rich gaze behaviors guided by different environments, and surveillance videos that have challenging head poses (*e.g.*, back view or high pitch), which is typical of people in unconstrained settings.

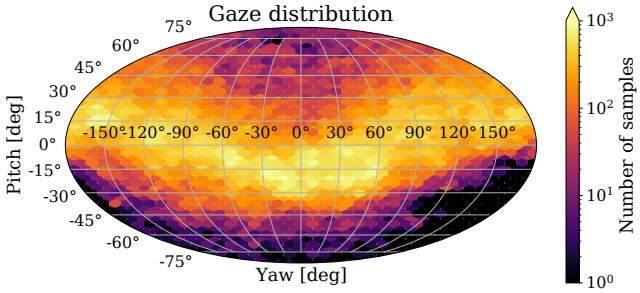

Figure 3. Distribution of 3D gaze directions in GAFA dataset shown with Mollweide projection. A wide range of gaze directions are uniformly captured in the dataset.

It consists of 882,000 frames of video that capture various gaze behaviors. We automatically annotated them with accurate 3D gaze directions as well as head and body orientations. Our dataset is the first publicly available dataset of its kind and can serve as a common platform for advancing 3D gaze estimation in the wild.

We set up our data collection to capture natural gaze behavior in daily environments. We chose five different environments for our recordings, and asked participants to freely walk around in the environments and to look for objects in accord with verbal instructions. We chose target objects that are commonplace in each environment. The behavior of participants during the experiment was recorded with 3 to 9 cameras positioned at high places in each environment. The distance between the camera and the participants varies from 50 cm to 7 m, where the size of the person's head region ranges from about 240×240 to 10×10 pixels.

We used wearable cameras to automatically compute ground-truth gaze, head, and body orientations for each frame. Each participant wore eye-tracking glasses (Pupil Core [15]). The glasses are minimally visible as they only have an upper frame. They have two infrared eye cameras that capture binocular gaze direction. They also have a camera pointing outwards which we used to compute the head orientation. Participants also wore a camera (GoPro HERO 7) on their chests to obtain body orientation.

The ground-truth head and body orientations were obtained by using an AR marker-based positioning system (ArUco) [11, 24] from videos of the eye-tracking glasses' outward pointing camera and the chest camera, respectively. Thirty to fifty AR markers were placed at various locations in each environment. All AR markers in the environment were scanned beforehand to calculate their 3D positions in the world coordinate system. We obtained head and body orientations by solving a PnP problem from videos taken with head- and body-mounted cameras. The gaze direction relative to the head pose was obtained by the eye tracker and was transformed into the world coordinate system.

We collected videos from 17 sessions of 8 subjects. The

GAFA dataset contains 882K frames in total (789K for training and 93K for testing). To evaluate the generalizability across subjects, we excluded data from 3 participants for testing, whose data are never used in the training set. Figure 3 shows the distribution of the gaze direction in the camera coordinate. The yaw evenly spreads out over 360° degrees, which clearly shows that the GAFA dataset contains wide range of head poses including back-views.

5. Experimental Results

To our knowledge, our method is the first to realize 3D gaze estimation from surveillance views that works even when the eyes are not visible. There are no other methods that we can directly compare with nor are there any other dataset than our GAFA dataset that can be used to fully examine the accuracy of our and existing gaze estimation models in the expected context (*e.g.*, surveillance views of people with 3D gaze annotations).

We thoroughly evaluate our method and compare it against existing methods with a number of carefully designed experiments. First, to evaluate how well our and existing gaze estimation methods work in realistic surveillance videos, we train and test these models with our GAFA dataset. We conduct an ablation study to examine the validity of the key components of our method, in terms of the effectiveness of each architectural component.

Next, we perform cross-dataset evaluation on the MoDiPro dataset [6] which contains surveillance videos of freely moving people in a post-hospitalization facility. Although the MoDiPro dataset contains only one scene with two cameras and only manually annotated 2D gaze annotations, we use it to examine how well our method trained with other datasets works on real surveillance videos. Note that we cannot use the MoDiPro dataset for training our method as the dataset only contains 2D gaze annotations.

Finally, we examine the effectiveness of our optional integration of eye appearance and multiview cameras on the GAFA dataset. We also include quantitative validation of estimated uncertainty and evaluation of our head and body orientation estimates in the supplementary materials.

Models for comparison. We experimentally compare the accuracy of representative gaze estimation methods with ours. As far as we know, there are few methods applicable to surveillance videos.

The method by **Dias *et al.*** [6] estimates 2D gaze in the image plane from facial keypoints detected by OpenPose [4]. We also compare with two appearance-based method. **Gaze360** [16] takes successive whole head images as input and outputs the 3D gaze direction. Note that, in contrast, our method estimates gaze from head and body orientations, not appearance. **X-Gaze** [32] takes face images as input. X-Gaze assumes high-resolution facial images as input, and thus is fundamentally not applicable to gaze estimation from afar. For fair comparison, we also train them on back-facing images in addition to regular frontal-views when fine-tuning. As a simple baseline, we compute the mean gaze direction in the training set, and evaluate the angular errors (MAEs) when that dataset mean is used as the gaze estimate in the test set (**Fixed bias**). This baseline shows the lower bound of the estimation accuracy.

For our models, we tested four variations: the model using temporal head-body coordination to estimate gaze direction (**Ours**), opportunistic use of eye appearances (**Ours + Face**), multiview integration by weight averaging (**Ours Multiview-WA**), and multiview integration using 3D trajectory (**Ours Multiview-3DTraj**).

We also tested our model with different training strategies. We either trained our model on GAFA dataset in an end-to-end manner, or separately trained the head and body network with the AGORA Dataset and trained the Gaze LSTM network with the GAFA dataset. Throughout all experiments, each model was trained with Adam with learning rate = 10^{-4} and batch size = 32.

Results on the GAFA dataset To test our method on realistic, if not truly real, surveillance videos, we first evaluate the accuracy on our GAFA dataset. The input video is rescaled to 480p so that the quality of images matches typical surveillance images. Table 1 shows the mean angular error (MAE) in 3D and 2D for each scene. Our method achieves significantly higher accuracy compared with Dias *et al.*'s method. Because the number of video frames with clear eye appearances is limited, appearance-based models (Gaze 360 and X-Gaze) perform worse. Among the 5 scenes, Office has the largest space in which subjects tend to walk without stopping exhibiting natural gaze-head-body coordination, where our method shows higher accuracy. Accuracy becomes low in LR and outdoor scenes, which contain furniture or trees that cause frequent occlusions. These results demonstrate the effectiveness of our method, *i.e.*, estimating gaze just from the head and body orientations which can be robustly estimated from surveillance views by leveraging the angular-temporal dependency of gaze-head-body orientations.

The second row of Tab. 1 shows the results of the four variations of ablation study. "No temporal" has the same architecture as our proposed model, but receives only one input frame, thus the model does not consider the temporal relationship of gaze-head-body orientations. In "No uncert.," the bidirectional LSTM layer does not receive the estimated uncertainty of head and body orientations. "No body input" and "No head input" do not receive either the estimated body direction or head direction as input. The results indicate that all of these components are essential for accurate estimation.

In addition, we also tested our model using eye appear-

Method	Office	LR	Kitchen	Library	Courtyard	Front 180°	Back 180°	Mean
Fixed bias	88.0/76.0	85.5/76.7	86.0/82.4	89.0/85.1	89.7/87.8	86.3/99.4	90.3/55.0	88.1/79.7
Dias *et al.* [6]	-/27.2	-/25.2	-/19.8	-/24.9	-/36.1	-/22.89	-/34.8	-/27.1
Gaze 360 [16]	24.0/19.2	41.1/31.3	32.4/21.2	27.5/20.7	28.2/28.3	21.8/19.6	36.3/26.7	30.4/24.5
X-Gaze [32]	24.2/23.0	42.0/40.9	23.3/22.9	24.6/22.3	30.2/31.9	26.2/23.5	31.5/31.7	29.2/28.4
No temporal	20.0/18.1	25.6/25.5	21.5/18.6	21.9/20.1	28.4/30.5	22.9/20.0	25.0/25.7	24.1/23.3
No uncert.	17.5/17.6	23.9/26.3	20.2/19.6	20.6/18.5	23.9/25.6	20.9/18.1	23.6/24.1	22.1/21.6
No body input	17.3/15.2	31.3/28.0	22.0/19.4	21.7/19.5	25.7/27.1	21.9/19.6	26.5/24.7	24.1/22.5
No head input	20.5/21.9	31.7/30.8	24.0/23.8	23.2/22.0	24.7/27.2	23.2/20.6	27.0/28.3	24.9/24.8
Ours (AGORA)	24.9/22.8	25.7/24.2	23.4/20.8	27.7/27.1	30.1/32.2	27.3/22.2	28.3/28.2	27.3/26.8
Ours (GAFA)	**14.4**/14.3	25.1 /**22.6**	20.4/19.6	19.8/**18.4**	25.4/26.9	20.7/17.4	23.2/21.9	21.7/20.9
Ours + Face (GAFA)	15.3/**14.3**	**24.0**/24.2	**19.1/17.2**	**18.2**/19.0	**24.4/26.0**	**19.3/16.8**	**21.7/21.5**	**20.4/20.0**

Table 1. Quantitative evaluation in comparison with existing methods on our GAFA dataset (mean angular errors (MAE) on test data). The last column shows the mean MAE for all scenes. All models were trained on our training dataset. The MAEs are shown in 3D/2D. Our method consistently outperforms past methods on these challenging real-world data.

Figure 4. Example 3D gaze estimates on the GAFA dataset. Estimated (orange) and true gaze directions (blue) are shown with mean angular errors (MAE). Top, middle, and bottom rows show the results with high, average, and low accuracy, respectively.

ances (Ours + Face) and our model with the head and body network trained on the AGORA dataset. As the bottom rows of Tab. 1 shows, the model trained on the AGORA dataset is slightly less accurate compared with that trained on the GAFA dataset. The accuracy of our model slightly improves when using the appearance of eyes, especially for front facing images (Front 180°).

Figure 4 shows example results of our gaze estimation on the GAFA dataset. As shown in the top row, our model tends to show higher accuracy when the person is looking at a certain object. The estimation becomes more challenging when the person is walking (middle row). The bottom row shows samples from failure cases, in which the head or body part is often out of sight. Although gaze estimation becomes challenging in these cases, our model returns best guesses based on the head or body orientations whenever either of the two is observable. Please see the supplemental video for results in image sequences.

Results on the MoDiPro dataset Next, to understand the generalization performance of our and existing models in a different scene, we conducted a cross-dataset experiment on the MoDiPro dataset. The first row of Tab. 2 shows the results when the models are trained on the GAFA dataset and tested on the MoDiPro dataset. Note that our estimates are in 3D, which is projected onto the image plane to evaluate against the 2D ground truth of the MoDiPro dataset. Our model performs the best. The estimation accuracy can be increased by fine-tuning on any new scene if some 3D gaze supervision can be prepared.

Our model has a two-stage architecture that enables us to separately train the head and body network with other large-scale datasets such as for human pose estimation. This is particularly useful because the first stage of the model takes images as input, and a larger image variation leads to better generalization. This is also the case for Dias *et al.*'s method which is built on OpenPose [4]. To make most of this point, we separately trained the first-stage network (the head and body network) on the AGORA dataset, and trained the second-stage network with the GAFA dataset. In addition to the Dias *et al.*'s original model with OpenPose trained on COCO dataset, we also retrained OpenPose on the AGORA dataset for fair comparison. The bottom rows of Tab. 2 shows the results. Our model trained on the AGORA dataset shows better accuracy than Dias *et al.*'s model regardless of the training dataset. This result indicates that the temporal relationship of gaze-head-body orientations is a more robust and generalizable cue for estimating the gaze direction compared to static facial keypoints used in Dias *et al.*'s method. Figure 5 shows example out-

Method	Training data	MAE (2D)
Gaze 360 [16]	GAFA	52.5
X-Gaze [32]	GAFA	51.4
Ours	GAFA	46.3
Dias *et al*. [6]	COCO (OpenPose) + GAFA	28.1
Dias *et al*. [6]	AGORA + GAFA	32.1
Ours	AGORA + GAFA	**25.6**

Table 2. Cross-dataset evaluation. Each model is trained on the specified dataset and tested on MoDiPro dataset. 2D MAEs on MoDiPro dataset are shown. Our model partly trained on the AGORA dataset shows the highest accuracy.

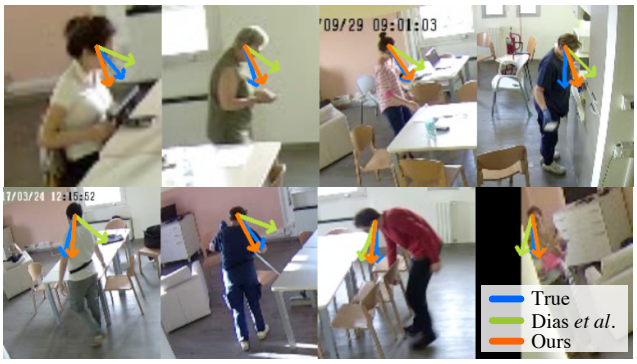

Figure 5. Example gaze estimates on the MoDiPro dataset. The output of the proposed model trained with AGORA, Dias *et al*.'s model trained with COCO, and ground-truth directions are shown.

puts of our and Dias *et al*.'s model on the MoDiPro dataset.

Cross-scene analysis We also conducted cross-scene experiments, in which the model was trained on 1 of 5 scenes and tested on the remaining 4 scenes in the GAFA dataset. Table 3 summarizes the results. Training on the library data achieves the best generalization accuracy. This is likely due to the fact that the library data contains a wide variety of gaze behaviors, for example, looking up and down to scan bookshelves. On the other hand, training on kitchen data leads to low generalization performance. This is because the viewpoints of the cameras are limited (most of the data is backward-looking images of people).

Multiview experiment We tested our multiview models on the GAFA dataset. We discard frames from camera views where no one is detected. On average, the multiview models uses 4.6 camera views. Table 4 shows the results of our monocular model, the weighted average from the output of monocular models, and the model using 3D body velocity trained and tested on the GAFA dataset. The simple weighted averaging produces slightly better accuracy than the monocular method. The model using 3D body velocity performs even better.

To validate our multiview method, we split the test set into the case where a part of the person is occluded and the case where the person is clearly visible from all cameras based on the number of detected keypoints from OpenPose [4]. While the performance of the monocular model significantly drops when the person is occluded, the multiview models are less affected, because they can appropriately assign a low weight to occluded cameras based on the estimated uncertainty.

		Test sets				
		Office	LR	Kitchen	Library	Courtyard
Training set	Office	-	23.3	36.4	31.2	33.5
	LR	36.6	-	32.9	29.8	43.4
	Kitchen	39.2	32.9	-	49.1	45.1
	Library	28.6	32.4	27.7	-	29.9
	Courtyard	33.4	35.6	37.0	32.4	-

Table 3. Cross-scene evaluation for different combinations of training/test sets. 3D MAEs for each scene are shown.

Method	Occluded	No occlusion	Mean
Baseline (Monocular)	24.8	20.2	21.6
Multiview -WA-	21.6	20.5	20.8
Multiview -3DTraj-	19.9	18.2	**18.9**

Table 4. Evaluation of our monocular and multiview models on GAFA dataset. Mean 3D MAEs across all scenes are shown. Combining multiple views with our model leads to better accuracy.

Limitation Since our method uses the temporal coordination of gaze-head-body orientations to estimate gaze direction, the accuracy of gaze estimation is limited when the person does not move at all (*e.g*., while reading a book). We plan to address this by exploiting other cues such the saliency of the scene in our future work.

6. Conclusion

In this paper, we introduced a novel method for 3D gaze estimation from a distance without relying on the appearance of eyes. We formulate gaze estimation as a Bayesian prediction that leverages learned angular-temporal dependency between gaze, head, body orientations. The experimental results clearly show that our method can robustly estimate gaze direction from afar, from typical surveillance views. We also introduced an extensive dataset for gaze estimation from afar which will be made public. We believe these results open new avenues of research on gaze estimation and human behavior analysis.

Acknowledgement This work was in part supported by JSPS 20H05951, 21H04893, JST JPMJCR20G7, JPMJPR1858, and RIKEN GRP. We also thank Ikuhisa Mitsugami and Yusuke Nishii for their help in the early stage of this work.

References

[1] Gareth R Barnes. Vestibulo-ocular function during coordinated head and eye movements to acquire visual targets. *J. Physiol.*, 287:127–147, Feb. 1979. 2

[2] Lucas Beyer, Alexander Hermans, and Bastian Leibe. Biternion Nets: Continuous Head Pose Regression from Discrete Training Labels. In *Pattern Recognition*, volume 9358, pages 157–168. Springer International Publishing, Cham, 2015. 3

[3] Patrick Burger, Martin Rothbucher, and Klaus Diepold. Self-initializing Head Pose Estimation With a 2D Monocular USB Camera. *arXiv:2005.10353*, 2014. 3

[4] 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. 6, 7, 8

[5] Jinyoung Choi, Beom-Jin Lee, and Byoung-Tak Zhang. Human Body Orientation Estimation using Convolutional Neural Network. *arXiv:1609.01984*, Sept. 2016. 3

[6] Philipe A Dias, Damiano Malafronte, Henry Medeiros, and Francesca Odone. Gaze Estimation for Assisted Living Environments. In *Proc. WACV*, 2020. 1, 2, 5, 6, 7, 8

[7] Yu Fang, Masaki Emoto, Ryoichi Nakashima, Kazumichi Matsumiya, Ichiro Kuriki, and Satoshi Shioiri. Eye-Position Distribution Depending on Head Orientation when Observing Movies on Ultrahigh-Definition Television. *ITE Trans. MTA*, 3(2):149–154, 2015. 2

[8] Yu Fang, Ryoichi Nakashima, Kazumichi Matsumiya, Ichiro Kuriki, and Satoshi Shioiri. Eye-head coordination for visual cognitive processing. *PLoS One*, 10(3):e0121035, Mar. 2015. 3

[9] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proc. ECCV*, pages 334–352, 2018. 1, 2, 5

[10] 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 *Proc. ETRA*, pages 255–258, 2014. 5

[11] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Rafael Medina-Carnicer. Generation of fiducial marker dictionaries using Mixed Integer Linear Programming. *Pattern Recognition*, 51, Oct. 2015. 5

[12] Craig Hennessey, Borna Noureddin, and Peter Lawrence. A single camera eye-gaze tracking system with free head motion. In *Proc. ETRA*, ETRA '06, pages 87–94, New York, NY, USA, 2006. Association for Computing Machinery. 2

[13] Duyeong Heo, Jae Yeal Nam, and Byoung Chul Ko. Estimation of Pedestrian Pose Orientation Using Soft Target Training Based on Teacher-Student Framework. *Sensors*, 19(5), Mar. 2019. 3

[14] Peiyun Hu and Deva Ramanan. Finding Tiny Faces. *arXiv:1612.04402*, abs/1612.04402, 2016. 2

[15] Moritz Kassner, William Patera, and Andreas Bulling. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proc. ACM UbiComp*, UbiComp '14 Adjunct, pages 1151–1160, New York, NY, USA, 2014. 5

[16] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically Unconstrained Gaze Estimation in the Wild. In *Proc. ICCV*, 2019. 1, 2, 5, 6, 7, 8

[17] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra M. Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye Tracking for Everyone. In *Proc. CVPR*, pages 2176–2184, 2016. 5

[18] Michael F Land. The coordination of rotations of the eyes, head and trunk in saccadic turns produced in natural situations. *Exp. Brain Res.*, 159(2):151–160, Nov. 2004. 2, 3

[19] Ji Woo Lee, Chul Woo Cho, Kwang Yong Shin, Eui Chul Lee, and Kang Ryoung Park. 3D gaze tracking method using Purkinje images on eye optical model and pupil. *Optics and Lasers in Engineering*, 50(5):736–751, May 2012. 2

[20] Junichi Murakami and Ikuhisa Mitsugami. Gaze from Head: Gaze Estimation Without Observing Eye. In *Pattern Recognition*, pages 254–267. Springer International Publishing, 2020. 3

[21] 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. 4

[22] Sergey Prokudin, Peter Gehler, and Sebastian Nowozin. Deep Directional Statistics: Pose Estimation with Uncertainty Quantification. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Proc. ECCV*, pages 542–559. Springer International Publishing, 2018. 3

[23] Mudassar Raza, Zonghai Chen, Saeed-Ur Rehman, Peng Wang, and Peng Bao. Appearance based pedestrians' head pose and body orientation estimation using deep learning. *Neurocomputing*, 272:647–659, Jan. 2018. 3

[24] Francisco J Romero-Ramirez, Rafael Muñoz-Salinas, and Rafael Medina-Carnicer. Speeded Up Detection of Squared Fiducial Markers. *Image Vis. Comput.*, 76, June 2018. 5

[25] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-Grained Head Pose Estimation Without Keypoints. *arXiv:1710.00925*, Oct. 2017. 1, 2, 3

[26] K Sankaranarayanan, M Chang, and N Krahnstoever. Tracking gaze direction from far-field surveillance cameras. In *Proc. WACV*, pages 519–526, 2011. 1, 2

[27] John S Stahl. Amplitude of human head movements associated with horizontal saccades. *Exp. Brain Res.*, 126(1):41–54, Apr. 1999. 2

[28] 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. 3

[29] Chenyan Wu, Yukun Chen, Jiajia Luo, Che-Chun Su, A Dawane, Bikramjot Hanzra, Z Deng, Bilan Liu, J Z Wang, and Cheng-Hao Kuo. MEBOW: Monocular Estimation of Body Orientation in the Wild. In *Proc. CVPR*, 2020. 3

[30] Hirotake Yamazoe, Ikuhisa Mitsugami, Tsukasa Okada, and Yasushi Yagi. Analysis of head and chest movements that

correspond to gaze directions during walking. *Exp. Brain Res.*, 237(11):3047–3058, Nov. 2019. 3

[31] Wolfgang H Zangemeister and Lawrence Stark. Types of gaze movement: Variable interactions of eye and head movements. *Exp. Neurol.*, 77(3):563–577, Sept. 1982. 2

[32] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. In *Proc. ECCV*, 2020. 1, 2, 6, 7, 8

[33] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-Based Gaze Estimation in the Wild. *Proc. CVPR*, June 2015. 2

[34] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *CVPRW, 2017*, pages 2299–2308. IEEE, 2017. 1, 2

[35] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *TPAMI*, 41(1):162–175, 2019. 1, 2, 5

[36] Zhiwei Zhu and Qiang Ji. Eye gaze tracking under natural head movements. In *Proc. CVPR*, volume 1, pages 918–923 vol. 1, 2005. 2

[37] Yijun Zhou and James Gregson. WHENet: Real-time Fine-Grained Estimation for Wide Range Head Pose. In *Proc. BMVC*, 2020. 3