End-to-End Gaze Grounding of a Person Pictured from Behind

Hayato Yumiya[1], Daisuke Deguchi[1], Yasutomo Kawanishi[2] and Hiroshi Murase[1]
[1] *Institute of Intelligent System, Nagoya University, Japan*
[2] *RIKEN GRP, Japan*

Keywords: 3D Human Posture, Gaze Grounding, Metric Learning, Person-to-Person Differences.

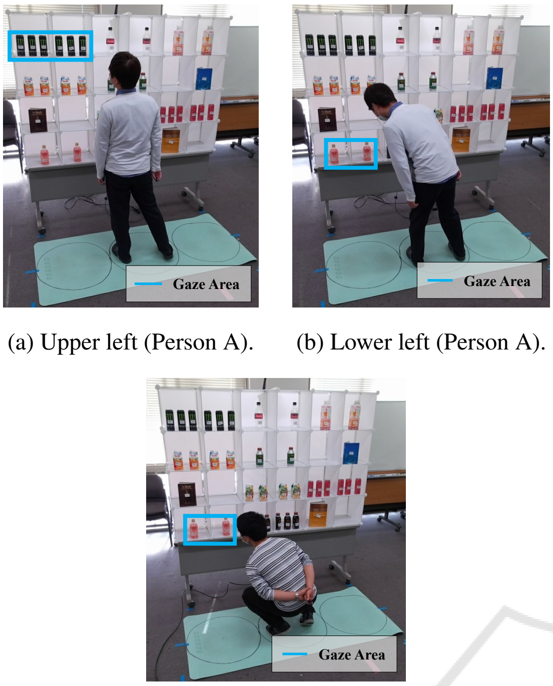
Abstract: In this study, we address a novel problem with end-to-end gaze grounding, which estimates the area of an object at which a person in an image is gazing, especially focusing on images of people seen from behind. Existing methods usually estimate facial information such as eye gaze and face orientation first, and then estimate the area at which the target person is gazing; they do not work when a person is pictured from behind. In this study, we focus on individual's posture, which is a feature that can be obtained even from behind. Posture changes depending on where a person is looking, although this varies from person to person. In this study, we proposes an end-to-end model designed to estimate the area at which a person is gazing from their 3D posture. To minimize differences between individuals, we also introduce the Posture Embedding Encoder Module as a metric learning module. To evaluate the proposed method, we constructed an experimental environment in which a person gazed at a certain object on a shelf. We constructed a dataset consisting of pairs of 3D skeletons and gazes. In an evaluation on this dataset,HEREHEREHEREwe confirmed that the proposed method can estimate the area at which a person is gazing from behind.

1 INTRODUCTION

Understanding the objects to which a person directs their gaze plays an important role in understanding human actions and intentions. The more a person is attracted to an object, the more they may tend to focus their gazed on it. This information is important in various applications, such as analyzing degrees of interest in commercial products. For example, suppose a customer is gazing at a certain product for an extended period. We can therefore predict that the customer may be interested in the product and may be likely to do so. In this research, we refer to the task of associating the area at which a person appears to be gazing with an actual object in the real world as gaze grounding. Many methods have been proposed to estimate the area at which a person in an image is gazing (Jha and Busso, 2018; Fridman et al., 2016). Fridman et al. proposed a method to accurately estimate the area of a person's attention by extracting face orientations from images. However, these methods are typically ineffective in practice because cameras may have a viewpoint situated behind the target person. Also, in realistic situations such as in retail stores, installing cameras to capture a person from a view may be awkward or impractical. Therefore, people in such camera images often stand with their backs to the camera. In these cases, because their face cannot be extracted from the image, the area at which they are gazing cannot be estimated by existing methods. Additionally, most existing methods estimate a heat map of the gaze area in the image, which does not directly correspond to the actual object.

In this study, we address the problem of gaze grounding, which estimates the gaze object area of a person in an image, especially the person pictured from behind. Humans can estimate where the person is looking from the posture, even from behind. This is the case because we know that posture changes depending on where a person is looking. For example, Figs. 1 (a) and (b) show examples of a person standing with their back to the camera. They show that the posture changes depending on where the person directs their gaze, with their head turning in different directions or bending forward to observe lower objects. This inspired us to estimate the gaze area by focusing on differences in postures. However, by analyzing the postures more deeply, we found that considerable differences between individuals in postures even when looking at the same object from the same location, as shown in Figs. 1 (b) and (c). This indicates that these differences need to be compensated

Yumiya, H., Deguchi, D., Kawanishi, Y. and Murase, H.
End-to-End Gaze Grounding of a Person Pictured from Behind.
DOI: 10.5220/0011899900003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 898-905
ISBN: 978-989-758-634-7; ISSN: 2184-4321



(a) Upper left (Person A). (b) Lower left (Person A).

(c) Lower left (Person B).

Figure 1: Looking at one of the objects on a shelf.

for to estimate the area of the object at which a person is gazing from their posture.

In this study, we propose an end-to-end method to estimate the area of the object at which a person is gazing from their posture while differences between individuals. Here, human posture is defined as a set of 3D locations of body joints calculated by 3D pose estimation. In the proposed method, a posture is embedded into a posture feature space. The feature space is trained by deep metric learning to differentiate poses if the gazed objects different while bringing even different persons close together in the feature space if the objects of their gaze are the same. This emphasizes posture differences depending on the gaze target while minimizing person-to-person differences. Then, a likelihood map associated with the scene is generated from the embedded posture feature. By referring to the object location in the scene, the method aggregates likelihoods within each target object region. Finally, the region with the highest likelihood is selected as the region of the object of the person's gaze.

The contributions of this paper are as follows.

- We define the problem of gaze grounding for a person pictured from behind.
- We propose a method to estimate the area of an object at which a person is gazing from a camera viewpoint located behind them. The method generates a likelihood map from posture information and associates the map with regions of objects.
- We propose a deep feature embedding method that can compensate for differences between individuals. This makes the distances between posture pairs of different persons looking at the same target object close together.
- We provide a method to aggregate likelihood by referring to object locations. This enables us to estimate gazed objects with good robustness to object locations.
- We also propose an end-to-end training method that jointly trains the deep feature embedding and the likelihood-map generation model.

2 RELATED WORK

2.1 Gaze Estimation

Kellnhofer et al. (Kellnhofer et al., 2019) have proposed a method of gaze estimation with a model trained on images captured under various situations and camera viewing directions. They constructed a dataset called Gaze360 comprising indoor and outdoor videos captured by an omnidirectional camera and annotated with 3D gaze directions. Because this dataset contains a large number of individual persons, it can be widely used to evaluate 3D gaze estimation methods. The estimation accuracy was also improved using multiple consecutive frames as input to a long short-term memory (LSTM). However, this method cannot be applied to estimate the gaze area of a person with their back to the camera.

Nonaka et al. (Nonaka et al., 2022) focused on the cooperativeness between gaze, head, and body, and proposed a gaze estimation method using temporal information of head position and posture. They constructed a dataset with 3D annotations of gaze direction on videos of multiple situations captured by surveillance cameras, and modeled the gaze direction likelihood distribution representing the relationships between head and body postures. Here, a neural network was used to represent the conditional distribution of gaze direction. This shows that 3D gaze can be estimated even for scenes with considerable occlusion. However, this method cannot estimate the 3D gaze direction from a single frame because it requires temporal information.

2.2 Gaze Estimation from the Behind a Person

Bermejo et al. (Bermejo et al., 2020) proposed a method to estimate gaze direction from the back of a person's head. Their method estimates the gaze direction using the head region detected by YOLO (Redmon and Farhadi, 2018) from a single frame captured by a third-person view camera. In addition, they created 3D models of various people and virtually generated images of a person pictured from behind in various environments (varting elements such as light source location, angle, camera distance, and so forth). By using these images for training, they reduced the estimation error caused by camera placement, angle, lighting conditions, resolution, and so forth. Finally, they achieved an estimation error of about 23 degrees in the horizontal direction and 26 degrees in the vertical direction, which is relatively accurate for estimating gaze direction from behind. In contrast, it is difficult to estimate the gaze area because the target object cannot be accurately determined only by the gaze direction.

2.3 Gaze Area Estimation from Posture Information

Kawanishi et al. (Kawanishi et al., 2018) proposed a method for estimating a gaze target using the posture of a person in an image. Based on the idea that posture can vary relative to the gaze target, they estimated the target at which a person was looking as a classification problem into four areas on a book page based on the person's posture. Their results suggested that the human posture can be used to estimate gaze area. However, because this is a pre-defined classification problem, all the target locations should be fixed beforehand, and the system cannot estimate other targets.

2.4 Metric Learning

Metric learning is a method for constructing a feature space embedding that maps semantically identical data to nearby locations and semantically different data to distant locations. A typical approach is to learn a feature space embedding using anchor data, positive data of the same class, and negative data of a different class. Then, the model is trained so that the distance between the anchor data and the positive data is smaller than the distance between the anchor data and the negative data (Chopra et al., 2005; Wang et al., 2017). In this study, by using this framework, we obtain the embeddings that transform postures gazing at the same area into close features in a feature space.

3 ESTIMATING GAZE OBJECT AREA FROM BEHIND

To associate the gaze area with an actual object in the real world, we propose an end-to-end method that generates a likelihood map of the gaze area for a given posture and aggregates likelihoods within each object region to obtain object-wise likelihoods.

The method estimates the gazed object area from behind a person using their posture. As may be observed from seen in Fig. 1 (a), humans can easily estimate that Person A is looking at the object located at the upper left of the shelf. In addition, from the posture looking at the different areas (Figs. 1(a) and (b)), we can observe that they have different characteristics in terms of head orientation, bending of the hips and legs, and so forth. These indicate that we usually take a similar posture when looking at the same place and vice versa. From this characteristic, we consider estimating the gaze object area by focusing on differences in posture, even from behind.

When we analyze the postures more deeply, as shown in Figs. 1 (b) and (c), we can note some differences between individuals. In the figure, a different person is looking at an object placed at the lower left; they are in different postures even though they are looking at the same area. To compensate for these differences, we introduce a deep metric learning technique into the Posture Embedding Encoder module.

Fig. 2 shows an overview of the proposed method. The neural network model consists primarily of two parts, including a Posture Embedding Encoder module and a Likelihood Map Generator module, followed by a Likelihood Aggregation process. The neural network model is a combination of the Posture Embedding Encoder module and the Likelihood Map Generator module. The Posture Embedding Encoder module is trained to compensate for the person-to-person differences, while the Likelihood Map Generator module is trained to generate a likelihood map from a posture. It is trained in an end-to-end manner, which minimizes the sum L of the losses from the Posture Embedding Encoder module L_e and the Likelihood Map Generator module L_d as given below.

$$L = L_e + L_d. \tag{1}$$

The Likelihood Aggregation process calculates the object-wise gazed likelihood from the likelihood map in reference to object locations.

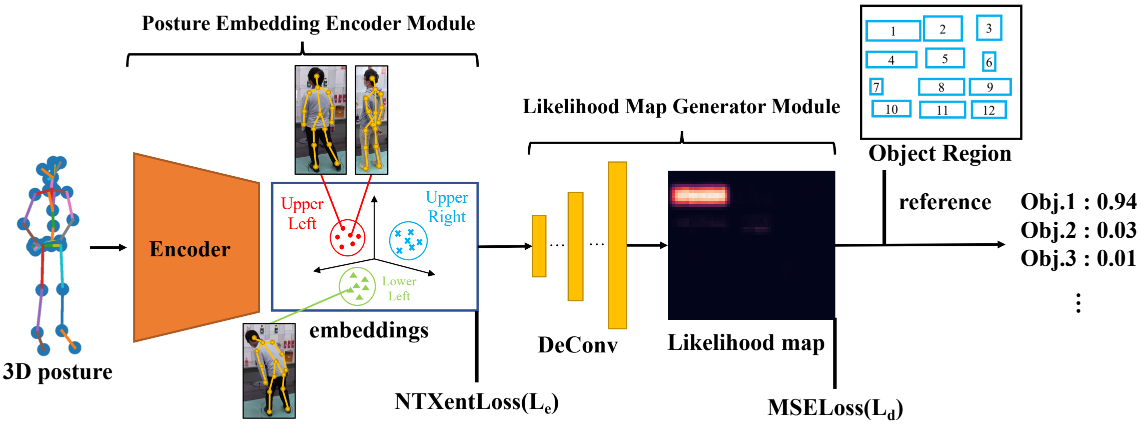


Figure 2: The architecture of the proposed model.

3.1 Posture Embedding Encoder Module

Based on the deep metric learning framework, we use posture and gazed object area labels to learn the Posture Embedding Encoder $h(\cdot)$ that embeds a i-th posture $\mathbf{p}_i$ into a posture feature space that can absorb the person-to-person differences.

The ground truth labels of the gaze object area are provided in the training data. Here, the labels are IDs of the gaze target object regions. The encoder is trained so that the distance between two posture features is close if the labels are the same and farther away if the labels are different. This enables us to project postures into the embedding space that controls for differences between individuals in the posture feature. The input of the encoder is 21 three-dimensional coordinates of human joints that is, it is a 63-dimensional vector $\mathbf{p}_i \in \mathbb{R}^{63}$. A posture is embedded into a posture feature $\mathbf{f}_i = h(\mathbf{p}_i)$ by the encoder. Here, $h(\cdot)$ is implemented as multiple fully-connected layers. In the scene, there are several objects that people might be expected to gaze at. We assign object IDs for each object and use them for the metric learning. Here, we use NTXentLoss, (Chen et al., 2020) which can consider the multiple labels simultaneously for L_e. NTXextLoss for i-th sample is defined as

$$L_e = -\log \frac{\exp(\mathbf{f}_i \cdot \mathbf{f}_+/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[i\neq k]} \exp(\mathbf{f}_i \cdot \mathbf{f}_k)/\tau)}, \qquad (2)$$

where $\mathbf{f}_i$, $\mathbf{f}_+$, and $\mathbf{f}_k$ are vectors in feature embedding space calculated from the input posture vectors by the encoder. Here, $\mathbf{f}_+$ is a vector of the same class as $\mathbf{f}_i$ in the mini-batch, and $\mathbf{f}_k$ is a vector in the mini-batch. This mini-batch is selected by Easy Positive Triplet Mining. (Xuan et al., 2020) $\mathbb{1}_{[i\neq k]} \in \{0,1\}$ is a function that outputs 1 if $i \neq k$ and 0 otherwise. The variable τ is a hyperparameter, and we use $\tau = 0.07$ in the experiment.

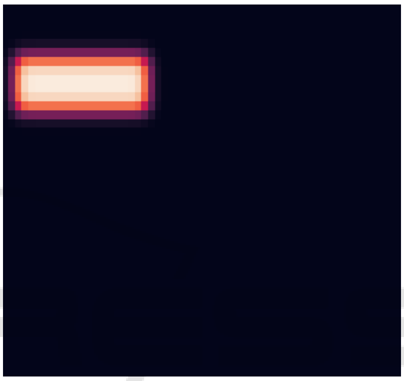


Figure 3: An example of a ground-truth likelihood map used in the training.

3.2 Likelihood Map Generator Module

This section describes the details of the likelihood map generator corresponding to the gaze area likelihood based on the embedded feature space.

The generator $g(\cdot)$ also consists of a neural network comprising, fully connected layers followed by inverse-convolutional (Deconvolution) layers. The input of the generator is an embedded feature $\mathbf{f}_i$ computed by the encoder described in the previous section, and the generator is trained to output a likelihood map $\mathbf{m}_i = g(\mathbf{f}_i)$ representing the gaze area likelihood over the target space.

The ground truth likelihood map for the training is prepared as follows. First, an image of the target scene in 40 × 60 pixels, named object area map, is prepared. Here, we assume that the target space is a shelf containing multiple objects in a store. In an object area map, pixels corresponding to a gazed object in the space are filled with 1, and the remainder

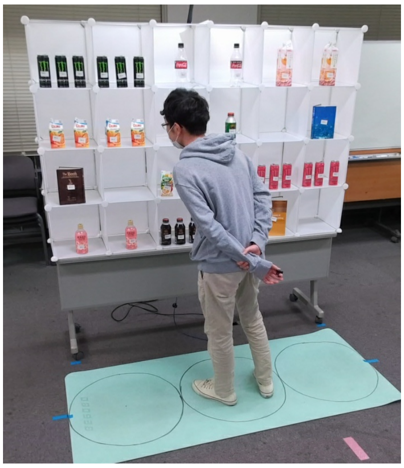


Figure 4: Example of dataset acquisition setting.

of the pixels are filled with 0. To make the output of the inverse-convolution (Deconvolution) network be a square size (equal height and width), the object area map is extended to 64 × 64 pixels with zero padding. A Gaussian filter ($\sigma = 3$) is then applied to the object area map to get smooth contours. Figure 3 shows an example of the heat map created as the target.

For the training, the loss L_d is a mean squared error (MSE) between an estimated map and a ground-truth map. The loss for i-th sample is calculated as

$$\begin{aligned} L_d &= \mathrm{MSE}(\mathbf{m}_i, \widehat{\mathbf{m}}_i) \qquad (3) \\ &= \frac{1}{K^2} \sum_{j=1}^{K^2} (y_{ij} - \hat{y}_{ij})^2 \\ & y_{ij} \in \mathbf{m}_i, \;\; \widehat{y}_{ij} \in \widehat{\mathbf{m}}_i, \end{aligned}$$

where $\widehat{\mathbf{m}}_i$ is the ground-truth map corresponding to the i-th input. Here, the size of a likelihood map was $K \times K$, and $K = 64$ is used in the experiment.

3.3 Likelihood Aggregation Process

In this section, we describe the details of the likelihood aggregation process from the estimated likelihood map.

The gaze target object region is determined from the likelihood map generated by Likelihood Map Generator Module described above, by referring to the object area map. For each object region, the average likelihood within the object region is computed from the likelihood map. Then, the area having the highest average likelihood is selected as the gazed object area.

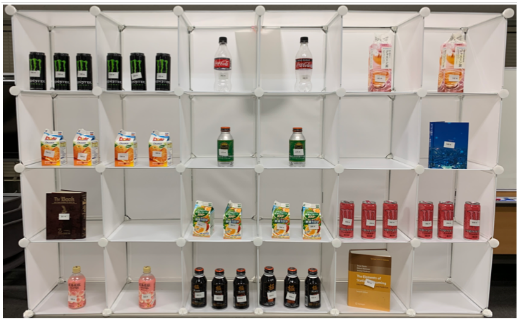


Figure 5: Example of object placement on the shelf.

4 EXPERIMENTS

We evaluated the performance of the proposed method. To do so, we constructed a new dataset as described below.

4.1 Dataset

The purpose of this study was to estimate the gaze area at which a person is gazing from behind using the 3D coordinates representing a human posture. However, no datasets are publicly available for this task. Therefore, we constructed a new dataset consisting of 3D human postures and their corresponding gaze area annotations.

First, we describe the details of data acquisition setting. In the dataset acquisition process, we simulated a situation in which a surveillance camera captures a person looking at one of the objects on a shelf in a convenience store. Figure 4 shows the data acquisition environment that we prepared for capturing a person looking at one of the objects on the shelf from a specified position. The size of the shelf is 120 cm height × 180 cm width, and it is divided into 12 areas where each size is 30 cm height × 60 cm width.

We put several types of target objects on the shelf such as plastic bottles, cans, books, and paper cartons. There are several kinds of cans and bottles; one type of object is placed in each area. Fig. 5 shows the shelf that we used in the dataset acquisition. We added annotations of the object area as follows. First, the objects are grouped into 12 groups. The 12 regions corresponding to 12 groups of objects, on the shelf were annotated as segmented regions. Even for the multiple objects of the same type is in each segment, we consider these segments as object areas. These 12 object areas were used as the ground truth(GT) corresponding to the postures of the training data.

The subjects were standing 0.5 m away from the shelf to see each object. The data were collected from seven participants in the experiment (one female and

Table 1: Results of correct answer rate and estimation error.

Method	Correct answer rate			Estimation Error (m) ↓
	Top-1 (%) ↑	Top-2 (%) ↑	Top-3 (%) ↑	
Proposed	**34.26**	**55.02**	**66.13**	**0.33**

six males). Here, we used Azure Kinect to capture images, and their resolutions were $1,280 \times 720$ pixels and the frame rate was 15 fps. The 3D posture in Azure Kinect is originally composed of 32 3D skeletal coordinates. When the target person image is taken from behind, the nose, eyes, thumbs, and ears are difficult to estimate accurately by occlusion, so these joints were not used. In this dataset, 3d postures were composed of 21 3D skeletal coordinates.

Through this data acquisition process, a total of 15,228 frames were collected as a dataset.

4.2 Experimental Settings

A summary of the proposed method is proposed as follows.

Propose Method

First, the proposed encoder module is applied to posture (3D coordinates of human joints) to obtain posture features in the feature space. Then, deconvolutional neural network is applied to the embedded features to reconstruct a likelihood map corresponding to the subject's gaze area. Finally, object-wise likelihoods are aggregated for every object region referring to object locations.

Here, we used five fully-connected layers for the encoder, which output a 4-dimensional vector $\mathbf{f} \in \mathbb{R}^4$ from a 63-dimensional vector $\mathbf{p}$. As a metric learning framework, Easy Positive Triplet Mining (Xuan et al., 2020) was used for sampling triples from the training data.

For the Likelihood Map Generator, we used 6 fully connected layers and 3 convolutional transpose layers, and the sigmoid activation function was applied to the output layer to restrict the output values within the range [0, 1].

For training the entire network, we used the AdamW (Loshchilov and Hutter, 2017) optimizer with the loss L defined in equation (1).

Experiments were conducted in a cross-validation scheme and the dataset was split with six of the seven participants as training data and one as testing data. An evaluation was performed using the following two evaluation metrics. The first was the correct answer rate corresponding to how much the method was able to correctly estimate the gazed object from the 12 areas on the shelf. In the proposed method, the average likelihood for each object was calculated, and the highest one is selected. We evaluated whether the object with the highest value was the same as the GT, which we refer to as the Top-1 correct answer rate. Also, we evaluated whether the correct answer can be achieved within the 2nd highest and 3rd highest areas, referred to as Top-2 and Top-3 rates of correct answers, respectively.

The second evaluation metric is an estimation error that is the average of Euclidean distances between a center point of the area with the highest value and the GT area.

4.3 Results and Discussions

Fig. 6 (c), Fig. 7 (c) show the likelihood map generated by the proposed method from postures captured by Azure Kinect. Table 1 shows the average correct answer rate evaluated via cross-validation by the proposed method. From the Table 1, Top-1 correct answer rate was 34.26%. This is a better result than 8% chance rate of considering this problem as a 12-class classification problem. In addition, the estimation error was 0.33 m, and it may be considered that even when the estimation failed, it is often estimated in the neighborhood of the correct answer. From these results, it may be considered say that it is possible to end-to-end method was able to estimate the gaze area from the posture.

4.4 Ablation Study

To investigate the effectiveness of the Posture Embedding Encoder module, an ablation study was conducted with a model designed to estimate the likelihood map from the posture without the Encoder.

A summary of the model's characteristics is provided below.

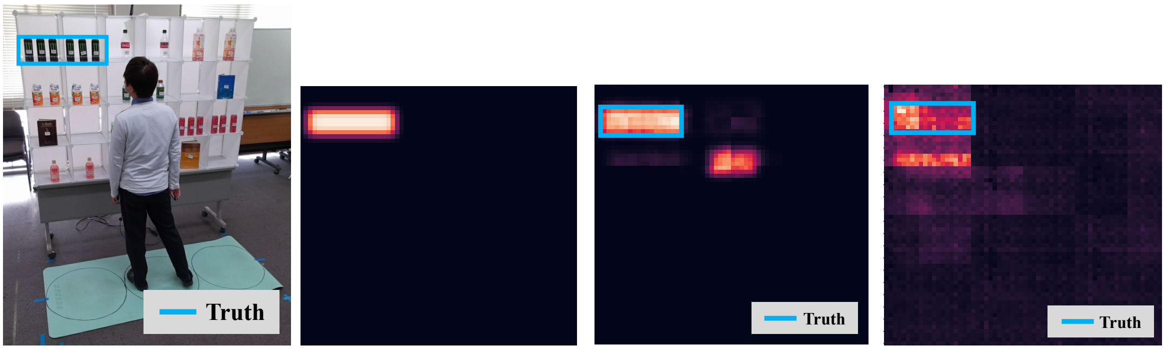
Ablated Model

A deconvolutional neural network was directly applied to the posture feature (3D coordinates of human joints) to reconstruct a likelihood map corresponding to the subject's gaze area. The number of parameters in this model was adjusted as in the proposed model.

Table 2 shows the estimation results between the proposed method and the ablated model. We observed

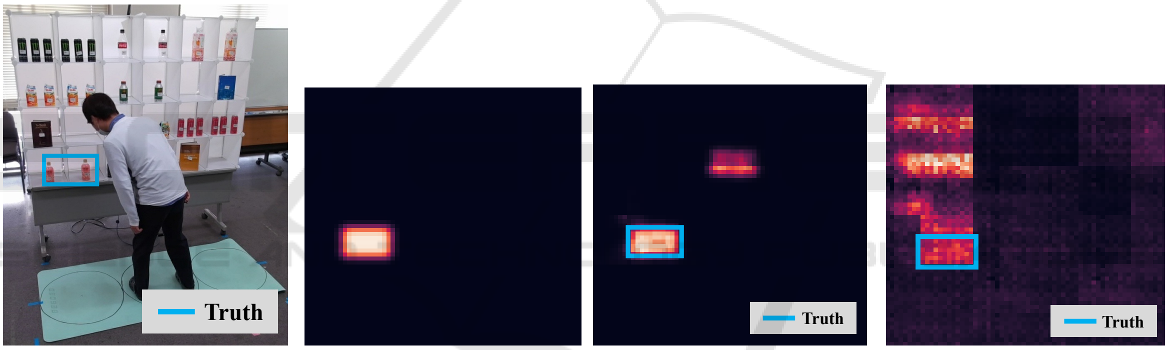
Table 2: Results of correct answer rate and estimation error.

Model	Encoder Module	Correct answer rate			Estimation Error (m) ↓
		Top-1 (%) ↑	Top-2 (%) ↑	Top-3 (%) ↑	
Ablated model	-	20.94	39.02	52.38	0.47
Full model	✓	**34.26**	**55.02**	**66.13**	**0.33**



(a) looking upper-left (b) Ground Truth (c) Full model (d) Ablated model

Figure 6: Result of a likelihood map estimated for person A looking at upper-left.



(a) looking lower-left (b) Ground Truth (c) Full model (d) Ablated model

Figure 7: Result of a likelihood map estimated for person A looking at lower-left.

that the ablated model that estimated the likelihood map without posture embedding showed a reduced Top-1 correct answer rate by 13.32 points and estimation error by 0.14 meters compared to the proposed method, and Top-2 and Top-3 correct answer rates were also greatly decreased. This indicates that the area with the highest average likelihood selected by the ablated model was located far from the GT area.

In addition, as shown in the generated likelihood maps of Figs. 6 (c) and 6 (d), it may be observed that the distribution in the generated likelihood map by the full model was smaller than that of the ablated model.

Fig. 8 show t-SNE visualization of the feature space embedding by the proposed encoder module. As shown in the figure, each class was clearly separated and embedded in the feature space.

These results suggest that the ablated model was strongly affected by the ambiguity of the gaze area caused by individual differences in posture. On the other hand, the proposed method can reduce this ambiguity by estimating a likelihood map using a feature embedding space constructed to compensate for differences between people, the proposed method can stably generate a likelihood map and improve the estimation accuracy by controlling for these variations.

5 CONCLUSIONS

In this study, we have addressed a problem of end-to-end gaze grounding, especially targeting a person pictured from behind. We have proposed an end-to-end method to estimate the gaze object from a posture of the person by referring to the object locations.

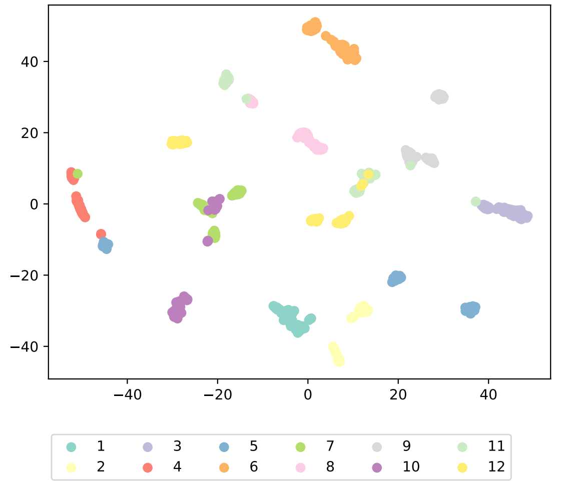


Figure 8: Visualization of embedded feature space by t-SNE. Each color corresponds to each gaze area.

In the proposed method, the 3D coordinates of body joints are first projected into a constructed feature embedding space to compensate for differences between individuals. The gaze area likelihood map is generated from the embedded features using a deconvolutional neural network. The likelihood is averaged within each object area by referring to the object locations, and object-wise likelihoods are calculated.

To confirm the effectiveness of the proposed method, we constructed a new dataset consisting of 3D coordinates of body joints and a target area to which a person directed their gaze, and experiments were conducted using this dataset. The experimental results showed that the proposed approach was able to estimate the gaze area from the posture, and the encoder module serves an important function in the performance role of the proposed model.

REFERENCES

Bermejo, C., Chatzopoulos, D., and Hui, P. (2020). Eyeshopper: Estimating shoppers' gaze using cctv cameras. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2765–2774.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International conference on machine learning*, pages 1597–1607. ICML.

Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Fridman, L., Langhans, P., Lee, J., and Reimer, B. (2016). Driver gaze region estimation without use of eye movement. *IEEE Intelligent Systems*, 31(3):49–56.

Jha, S. and Busso, C. (2018). Probabilistic estimation of the gaze region of the driver using dense classification. In *Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 697–702. IEEE.

Kawanishi, Y., Murase, H., Xu, J., Tasaka, K., and Yanagihara, H. (2018). Which content in a booklet is he/she reading? reading content estimation using an indoor surveillance camera. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1731–1736. IEEE.

Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., and Torralba, A. (2019). Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Nonaka, S., Nobuhara, S., and Nishino, K. (2022). Dynamic 3D gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2192–2201.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Wang, J., Zhou, F., Wen, S., Liu, X., and Lin, Y. (2017). Deep metric learning with angular loss. In *Proceedings of the 16th IEEE International Conference on Computer Vision*, pages 2593–2601.

Xuan, H., Stylianou, A., and Pless, R. (2020). Improved embeddings with easy positive triplet mining. In *Proceedings of the 2020 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2474–2482.