PAPER

Gender recognition using a gaze-guided self-attention mechanism robust against background bias in training samples

Masashi NISHIYAMA[†], *Member*, Michiko INOUE[†], *Nonmember*, *and* Yoshio IWAI[†], *Member*

SUMMARY We propose an attention mechanism in deep learning networks for gender recognition using the gaze distribution of human observers when they judge the gender of people in pedestrian images. Prevalent attention mechanisms spatially compute the correlation among values of all cells in an input feature map to calculate attention weights. If a large bias in the background of pedestrian images (e.g., test samples and training samples containing different backgrounds) is present, the attention weights learned using the prevalent attention mechanisms are affected by the bias, which in turn reduces the accuracy of gender recognition. To avoid this problem, we incorporate an attention mechanism called gaze-guided self-attention (GSA) that is inspired by human visual attention. Our method assigns spatially suitable attention weights to each input feature map using the gaze distribution of human observers. In particular, GSA yields promising results even when using training samples with the background bias. The results of experiments on publicly available datasets confirm that our GSA, using the gaze distribution, is more accurate in gender recognition than currently available attention-based methods in the case of background bias between training and test samples.

key words: *Gaze distribution, attention mechanism, convolutional neural network, gender recognition, self-attention*

1. Introduction

Gender recognition [1], [2] is a technology adopted to classify pedestrians as a woman or man on the basis of the appearances of their whole-body regions in images. It offers promise for use in a variety of applications; e.g., video surveillance and marketing. Currently available methods of gender recognition [3]–[5] use deep learning techniques, which require the preparation of a large number of training samples comprising pairs of pedestrian images and their supervised gender labels. In general, these training samples are manually collected, and this requires considerable time and effort. It also involves identifying background bias not intended by the collectors of the training samples, as described in [6]. Background bias frequently occurs if pedestrian images are acquired at a specific location. Background bias present in training samples is erroneously extracted as features and reduces the accuracy of gender recognition. We thus need to extract the essential characteristics of physical appearance differences in the presence of background bias in the training samples.

To explore ways of extracting informative features for gender recognition from pedestrian images when using deep learning networks, we consider here a case in which human observers view images of pedestrians to judge their gender. The human observers can correctly judge the gender of people in images regardless of any bias in the background. We think this is the case because the observers do not see the background in the images and are thus unaffected by bias in the images when determining the gender. In addition, the observers are likely to focus on the body of a person in the images. Here, we aim to mimic the human visual attention represented by the distribution of gaze locations, measured when observers view pedestrian images. We introduce the gaze distribution to the training process of a deep learning network for gender recognition. The gaze distribution can be used to extract features representing only true appearance differences without incorrectly training for the background features, even when background bias features heavily in the training samples.

Researchers [7]–[12] have recently attempted to extract informative features for classification tasks by incorporating the gaze distribution (such as fixation and saccade), where this distribution is measured when observers view images. However, they have not considered the background bias of the training samples. We think, nonetheless, that the use of the gaze distribution can help solve the problem of background bias. In particular, Nishiyama et al. [12] proposed a gender recognition method using the gaze distribution for pedestrian images. Their method involves measuring the distribution of gaze locations when observers judge gender in pedestrian images and then using the distribution in prepossessing to extract discriminative features. However, their method uses the gaze distribution only to preprocess for feature extraction and does not introduce it to the end-to-end framework for deep learning.

We consider using convolutional neural networks (CNNs), which are representative algorithms of deep learning for pattern recognition tasks. Sattar et al. [13] proposed a method of incorporating the gaze distribution into a CNN, called gaze pooling (GP). This assigns uniform weights, as guided by the gaze distribution, to cells in a feature map entered into the pooling layer of the network. However, GP cannot adaptively set the weights of the cells for each input feature map when the appearances of pedestrians variously change. Thus, we do not expect to sufficiently improve the gender recognition accuracy using GP's uniform weights.

To determine appropriate weights using the gaze distribution for each input feature map, we focus on an attention mechanism for the CNN. We develop an automatic adjustment of the attention weights. To this end, we consider

Manuscript received May 22, 2021.

Manuscript revised September 20, 2021.

[†] The authors are with the Graduate School of Engineering, Tottori University, Japan.

E105.D.1

self-attention (SA) [14], which is a well-known attention mechanism. If we simply use the available SA, its performance is incorrectly affected by the presence of background bias in the training samples. The problem, in which gender recognition accuracy decreases, arises because the SA straightforwardly uses spatial correlations among all cells in a feature map to determine the weights of deep attention. The same problem arises even when we use other existing methods of attention mechanisms such as in [15], [16].

To solve this problem, we propose a novel method for incorporating the gaze distribution into an attention mechanism to adaptively compute the spatial attention weights of each input feature map. We call the proposed method gaze-guided self-attention (GSA). We conducted experiments on gender recognition using publicly available CUHK and RAP datasets to assess the accuracy of our GSA in comparison with the prevalent existing methods. The results confirmed that our method improves the accuracy of gender recognition on datasets with different background biases between the training samples and test samples. The remainder of this paper is organized as follows. Section 2 describes related work, Section 3 describes the background bias, Section 4 describes the proposed attention mechanism that uses the gaze distribution, and Section 5 presents the experimental results. Our concluding remarks are given in Section 6.

2. Related work

Attention mechanisms have been widely used to improve the performance of deep learning networks. In image recognition tasks [17]–[19], attention mechanisms are designed to assign large weights to cells that are informative in the feature map of a deep learning network. To perform a person re-identification task, Song et al. [20] proposed an attention mechanism that is learned from pedestrian images with body-masked images to extract discriminative features. Furthermore, to simultaneously perform attribute recognition and person re-identification tasks, the attention mechanisms [21], [22] are proposed to assign informative weights using training samples with both attribute and ID labels. However, prevalent methods [17]–[22] do not consider the spatial relationship between the attention mechanisms and gaze distribution.

Some recent methods [7], [8] do incorporate the gaze distribution into the attention mechanism in deep learning. For video captioning, Yu et al. [7] used the gaze distribution as supervised labels to design a spatial attention mechanism for deep learning. Qiao et al. [8] designed an attention mechanism for answering visual questions using an idea similar to that proposed in [7]. The attention mechanisms described in [7], [8] have been trained so that the spatial attention weights are close to the gaze distribution observed in humans. An idea in place of the attention mechanism has also been proposed; it combines gaze distribution with the feature extraction layer of a deep learning framework. To summarize video contents, Wu et al. [9] used the gaze distribution of locations preferred by users as a feature to input signals of a deep network. To estimate the behavioral scanpaths of the eyes, Yang et al. [23] introduced the gaze distribution to inverse reinforcement learning. However, the above methods [7]–[9], [21] are not designed for the gender recognition task and thus cannot be easily used for this task. Furthermore, these methods do not address the problem of background bias in the training samples.

The gaze distribution has also been used in machine learning techniques, instead of deep learning techniques, to perform image recognition tasks. In assessing image quality, Xia et al. [10] estimated the graphlets of local regions representing the order in which locations in a given image were viewed. They used non-negative matrix factorization to generate the graphlets. In classifying the attributes of images of fashion-related items, Murrugarra-Llerena et al. [11] used the gaze distribution to select a discriminative region in the preprocessing of a classifier. Furthermore, the gaze distribution has been used in various applications of image recognition. Xu et al. [24] showed that the use of the gaze distribution helps in performing egocentric video summarization tasks. Sugano et al. [25] estimated preferable images using gaze distributions. Karessli et al. [26] classified objects using only gaze-related features without object labels for zero-shot learning. However, these methods [10], [11], [24]–[26] have not been used for the gender recognition task in the presence of background bias. To solve the problem of background bias in gender recognition, we design a method of developing an attention mechanism in a deep learning network that effectively uses the gaze distribution.

3. Background bias in training samples

To solve the problem caused by background bias, we need to be aware of obstacles in the images when collecting training samples for gender recognition. We discuss the case where training samples showing a pedestrian belonging to one gender class contain a particular obstacle in the background, whereas training samples showing a pedestrian belonging to the other gender class do not. For example, as shown in Fig. 1, we assume that there is an obstacle in front of the pedestrian in certain images. In these examples, the images used as training samples featuring men also contain an obstacle at the bottom, whereas those showing women do not contain the obstacle. This case applies, for instance, to images acquired in places where many women or men are in the vicinity of a certain camera (e.g., near a women's cosmetics counter or around the menswear section of a shop). In both cases, parts of the bodies of some people are occluded in the images.

We consider using training samples containing background bias when learning a CNN using the prevalent attention mechanisms. In this case, the attention layer mistakenly learns the presence or absence of a specific obstacle as a gender feature, rather than the differences in physical appearances in the pedestrians' body regions. As an example, if the network was learned using the training samples shown in Fig. 1, a test sample featuring a woman with an obstacle

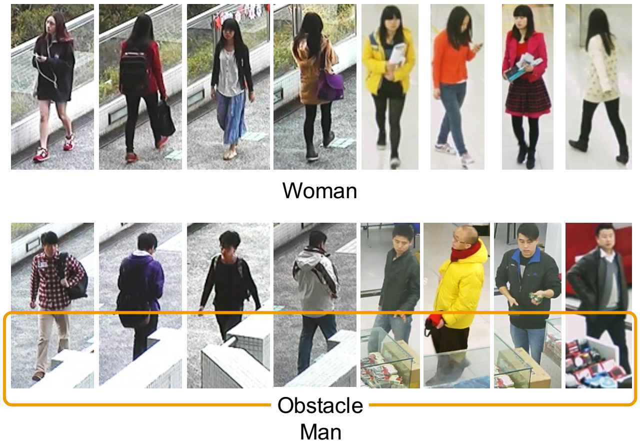

Fig. 1 Examples of background bias. We assume a scene in which an obstacle is present in front of a man subject and one in which no obstacle is present in front of a woman subject. In this case, the accuracy of gender recognition decreases when using training samples containing background bias. The subject in a test sample, featuring a woman with an obstacle in front of her, was incorrectly classified as a man.

in front was incorrectly classified as featuring a man.

In general, avoiding this problem requires a large number of training samples containing various backgrounds for both genders. When the background is clearly biased, we can modify camera settings during sample collection. In some cases, once the sample collection has been carried out, unexpected bias may occur in the training samples after analyzing the outputs of a gender classifier. In the worst case, it is necessary to repeat the sample collection process. Because collecting the training samples is time consuming and costly, researchers may sometimes need to contend with the collected training samples even if they contain bias.

We consider how to extract the characteristics of gender differences using training samples that have already been collected. By applying segmentation [27]–[29] and background subtraction [30]–[32], we can remove the background regions containing bias. However, it is difficult to accurately remove these regions with a limited number of training samples. Thus, instead of removing the background, we consider ways of extracting discriminative features from pedestrians' body regions. We develop a method for correctly classifying gender using the spatial attention mechanism of a network architecture that incorporates the gaze distribution.

4. Gaze-guided self-attention for gender recognition

4.1 Overview

We now summarize the prevalent SA [14], which is the basic idea behind our proposed method. In the SA layer, when computing an individual cell's attention weight in the input feature map, the system uses multiple cells with high correlations. The use of correlations allows the SA to consider the spatial relationship between neighboring cells and distant cells while computing the weight of the feature map. However, the simple use of the SA leads to the result being strongly affected by background bias in the training samples. This is because all cells of the feature map containing regions of both the body and background are directly used to compute the spatial attention weights.

Our method, the GSA, explicitly incorporates a weight computation using the gaze distribution in the spatial attention mechanism. We aim to mitigate the problem caused by background bias in the training samples using the spatial characteristics of the gaze distribution. Our GSA appropriately computes the spatial attention weights for each feature map corresponding to the body region. In the next section, we describe the details of our method.

4.2 Algorithm of the proposed method

We represent by $\boldsymbol{x} \in \mathbb{R}^{C \times H \times W}$ a feature map entered into the GSA layer. The number of channels of the feature map is C, the height of the map is H, and the width of the map is W. The gaze distribution pre-measured for human observers is represented by $\boldsymbol{g}$. (We describe how to measure $\boldsymbol{g}$ in Section 4.3.) The size of $\boldsymbol{g}$ is re-scaled to $C \times H \times W$ to fit the vector to the size of $\boldsymbol{x}$. Each cell at a certain height and width in $\boldsymbol{g}$ has the same value for different channels. We express $\hat{\boldsymbol{y}} \in \mathbb{R}^{C \times H \times W}$, a feature map generated by the GSA layer, as

$$\hat{\boldsymbol{y}} = \gamma \hat{\boldsymbol{o}} + \boldsymbol{g} \circ \boldsymbol{x}, \tag{1}$$

where γ is a learnable coefficient and $\hat{\boldsymbol{o}} \in \mathbb{R}^{C \times H \times W}$ is a GSA map. Note that the function $\circ$ represents the Hadamard product that takes an element-by-element product of cells between arrays of the same size. Figure 2 is an overview of the procedure adopted to compute a GSA map $\hat{\boldsymbol{o}}$. We describe details on how to compute $\hat{\boldsymbol{o}}$ below.

The point-wise convolution is applied to $\boldsymbol{x}$, where the purpose is to change the number of channels and thus reduce the computational complexity. Here, $\boldsymbol{x} \in \mathbb{R}^{C \times H \times W}$ is transformed into $\boldsymbol{x} \in \mathbb{R}^{C \times N}$ to simplify the computation of the arrays. We compute $\boldsymbol{q}(\boldsymbol{x}) = \mathbf{W}_q \boldsymbol{x}$, $\boldsymbol{k}(\boldsymbol{x}) = \mathbf{W}_k \boldsymbol{x}$, and $\boldsymbol{v}(\boldsymbol{x}) = \mathbf{W}_v \boldsymbol{x}$. We represent the weights of the point-wise convolution as $\mathbf{W}_q \in \mathbb{R}^{C' \times C}$ $(C' < C)$, $\mathbf{W}_k \in \mathbb{R}^{C' \times C}$, and $\mathbf{W}_v \in \mathbb{R}^{C \times C}$. As the training of our network progresses, $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are updated. The size of $\boldsymbol{q}(\boldsymbol{x})$ and $\boldsymbol{k}(\boldsymbol{x})$ is $\mathbb{R}^{C' \times N}$. The size of $\boldsymbol{v}(\boldsymbol{x})$ is $\mathbb{R}^{C \times N}$.

Our GSA assigns weights to $\boldsymbol{q}(\boldsymbol{x})$, $\boldsymbol{k}(\boldsymbol{x})$, and $\boldsymbol{v}(\boldsymbol{x})$ using the gaze distributions $\boldsymbol{g} \in \mathbb{R}^{C \times N}$ and $\boldsymbol{g}' \in \mathbb{R}^{C' \times N}$ as

$$\hat{\boldsymbol{q}}(\boldsymbol{x}) = \boldsymbol{g}' \circ \boldsymbol{q}(\boldsymbol{x}), \tag{2}$$

$$\hat{\boldsymbol{k}}(\boldsymbol{x}) = \boldsymbol{g}' \circ \boldsymbol{k}(\boldsymbol{x}), \tag{3}$$

$$\hat{\boldsymbol{v}}(\boldsymbol{x}) = \boldsymbol{g} \circ \boldsymbol{v}(\boldsymbol{x}), \tag{4}$$

where $\boldsymbol{g}'$ is an array in which the cell values are spatially the same as those in $\boldsymbol{g}$ and only the number of channels is changed. The main contribution of our algorithm is the incorporation of gaze distribution $\boldsymbol{g}$ into the prevalent spatial attention mechanism using Eq. (2), (3), and (4).

We then generate an array $\hat{\boldsymbol{s}} \in \mathbb{R}^{N \times N}$ as

$$\hat{\boldsymbol{s}} = \hat{\boldsymbol{q}}(\boldsymbol{x})^{\mathrm{T}} \hat{\boldsymbol{k}}(\boldsymbol{x}). \tag{5}$$

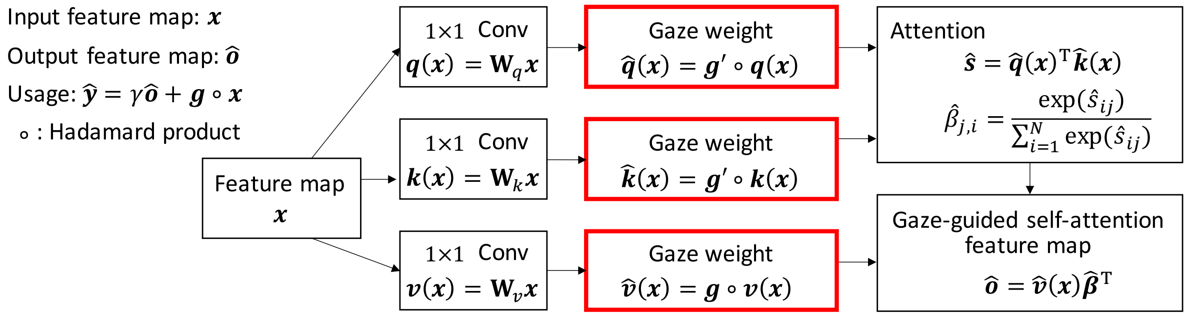
Input feature map: $\boldsymbol{x}$
Output feature map: $\hat{\boldsymbol{o}}$
Usage: $\hat{\boldsymbol{y}} = \gamma\hat{\boldsymbol{o}} + \boldsymbol{g} \circ \boldsymbol{x}$
$\circ$: Hadamard product
Feature map $\boldsymbol{x}$
1×1 Conv $\boldsymbol{q}(\boldsymbol{x}) = \mathbf{W}_q\boldsymbol{x}$
1×1 Conv $\boldsymbol{k}(\boldsymbol{x}) = \mathbf{W}_k\boldsymbol{x}$
1×1 Conv $\boldsymbol{v}(\boldsymbol{x}) = \mathbf{W}_v\boldsymbol{x}$
Gaze weight $\hat{\boldsymbol{q}}(\boldsymbol{x}) = \boldsymbol{g}' \circ \boldsymbol{q}(\boldsymbol{x})$
Gaze weight $\hat{\boldsymbol{k}}(\boldsymbol{x}) = \boldsymbol{g}' \circ \boldsymbol{k}(\boldsymbol{x})$
Gaze weight $\hat{\boldsymbol{v}}(\boldsymbol{x}) = \boldsymbol{g} \circ \boldsymbol{v}(\boldsymbol{x})$
Attention $\hat{\boldsymbol{s}} = \hat{\boldsymbol{q}}(\boldsymbol{x})^{\mathrm{T}}\hat{\boldsymbol{k}}(\boldsymbol{x})$
$\hat{\beta}_{j,i} = \frac{\exp(\hat{s}_{ij})}{\sum_{i=1}^{N}\exp(\hat{s}_{ij})}$
Gaze-guided self-attention feature map
$\hat{\boldsymbol{o}} = \hat{\boldsymbol{v}}(\boldsymbol{x})\hat{\boldsymbol{\beta}}^{\mathrm{T}}$

Fig. 2 Overview of our method for computing the GSA map $\hat{\boldsymbol{o}}$. We incorporate the gaze distribution into the weight computation of spatial attention. Our GSA computes appropriate attention weights for each feature map extracted from discriminative regions. We aim to mitigate the effect of background bias in training samples using gaze distribution $\boldsymbol{g}$.

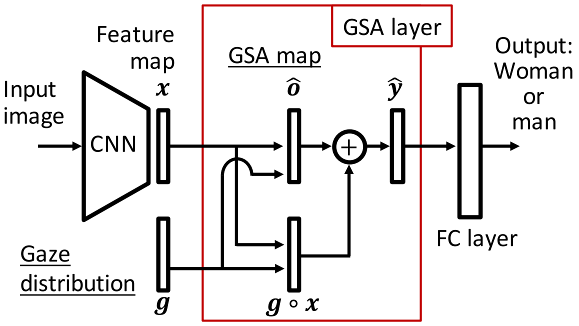

Fig. 3 Example of incorporating the GSA layer of our method into a network for gender recognition.

The element $\hat{s}_{ij}$ in the array $\hat{\boldsymbol{s}}$ indicates the spatial correlation between the cells at locations i and j. We generate an attention map $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{N\times N}$ by normalizing $\hat{\boldsymbol{s}}$ as

$$\hat{\beta}_{j,i} = \frac{\exp(\hat{s}_{ij})}{\sum_{i=1}^{N}\exp(\hat{s}_{ij})}. \tag{6}$$

The element $\hat{\beta}_{j,i}$ in array $\hat{\boldsymbol{\beta}}$ indicates the spatial correlation between the cells at locations j and i. Finally, our method generates the GSA map $\hat{\boldsymbol{o}}$ using $\hat{\boldsymbol{v}}(\boldsymbol{x})$ and $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{o}} = \hat{\boldsymbol{v}}(\boldsymbol{x})\hat{\boldsymbol{\beta}}^{\mathrm{T}}. \tag{7}$$

Note that $\hat{\boldsymbol{o}} \in \mathbb{R}^{C\times N}$ is reshaped to $\hat{\boldsymbol{o}} \in \mathbb{R}^{C\times H\times W}$ when computing Eq. (1).

Figure 3 shows an example of incorporating the GSA layer from our proposed method into a network for gender recognition. The GSA layer outputs a feature map $\hat{\boldsymbol{y}}$ from Eq. (1) by computing the GSA map $\hat{\boldsymbol{o}}$ using gaze distribution $\boldsymbol{g}$ and an input feature map $\boldsymbol{x}$.

4.3 Measurement of the gaze distribution

We now describe the method of measuring gaze distribution $\boldsymbol{g}$ when an observer views an image of a person to judge his/her gender. We think that the recognition accuracy can be improved if we measure the gaze distribution for each training sample. However, the gaze measurement for a large number of training samples is time consuming and costly. To avoid this expense, we measure a representative gaze distribution from a few stimulus images. In our experiments, the gaze measurement was performed according to the procedure described in [12]. The details of the gaze measurement are described below.

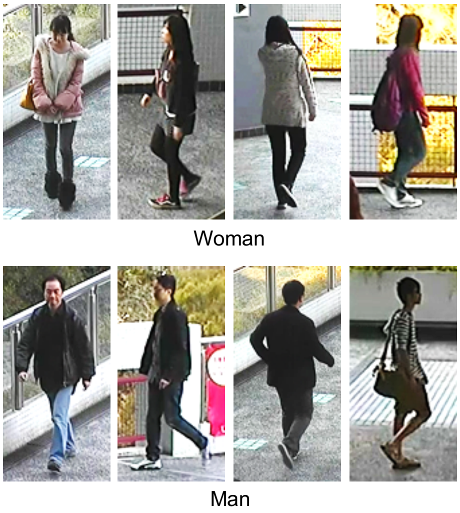

Fig. 4 Examples of stimulus images presented to observers to determine the gender of the subject during gaze measurement. We assume that the stimulus images were approximately aligned with the body regions.

To measure gaze locations, the observers were asked to complete a gender recognition task on the stimulus images while the locations of their gazes were recorded. We used pedestrian images from the CUHK dataset, which is part of the PETA dataset [33], as stimulus images. Figure 4 show examples of the stimulus images. We posed 16 stimulus images to each observer in our experiment. To control the experimental conditions, the numbers of man and woman subjects in the stimulus images were set equal. The proportions of all body orientations of subjects in the stimulus images (front, back, left, and right) were also equal. The same person did not appear more than once in the stimulus images. In addition, there was no specific object in the background of the pedestrian. The size of all stimulus images

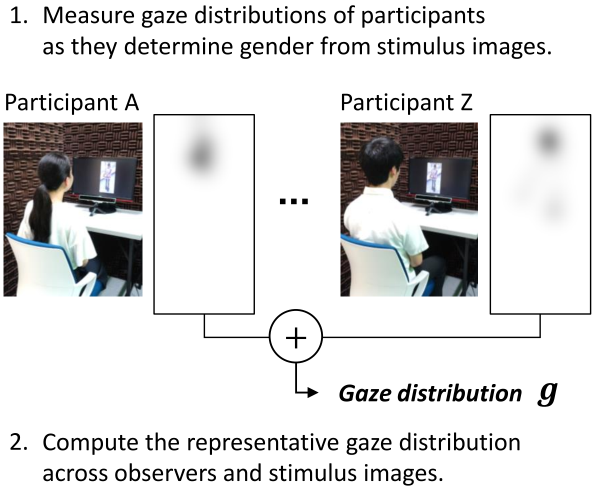
1. Measure gaze distributions of participants as they determine gender from stimulus images.
Participant A
Participant Z
...
+
Gaze distribution g
2. Compute the representative gaze distribution across observers and stimulus images.

Fig. 5 Overview of the measurement of the gaze distribution of the observers.

was 80 × 160 pixels.

Eighteen participants (nine men and nine women, average age of 22.6 ± 1.2 years) participated in the study. The participants were seated 65 cm from the display in the experiment. Each participant adjusted the height of the chair to maintain a gaze level between 110 and 120 cm from the ground. A 24-inch display (size of 53 × 30 cm and resolution of 1920 × 1080 pixels) was used to show the stimulus images. The display was set on a table that was 74 cm high. We used a GP3 gaze measurement device (Gazepoint HD), which has a sampling rate of 150 fps. Its angular resolution was between 0.5 and 1 degree. The stimulus image was enlarged to 480 × 960 pixels on the display. To avoid center bias [34] that causes the calculated gaze locations to converge to the center of the display during measurement, the stimulus images were set at random locations within a range ±720 pixels horizontally and a range ±60 pixels vertically from the center of the display.

Figure 5 is an overview of the process of measuring the gaze distribution of the participants as they performed the gender identification task. To acquire the gaze distribution of each participant, we presented randomly selected stimulus images to the participant, each for 2 s. We repeated this for each participant until all stimulus images had been presented. We used only the gaze locations of the participants, which were output by the gaze measurement device as fixations. We summed the locations of the fixations from all the participants and all stimulus images to generate a representative gaze distribution $\boldsymbol{g}$. The range of values for each cell in $\boldsymbol{g}$ was normalized to $(0, 1]$.

We checked the alignment using the average image computed from all images of subjects in the CUHK dataset. Figure 6(a) shows the average images of the subjects. The black circle at the top corresponds to the head region, the black ellipse near the center of the image corresponds to the torso region, and the light-gray part at the bottom of the image corresponds to the foot region. Figure 6(b) shows the representative gaze distribution $\boldsymbol{g}$ measured for the 18 participants. In the figure, the dark region in the gaze map represents the most frequent gaze locations gathered from the observers. We see that the participants mainly viewed the head regions of the subjects in the stimulus images. This tendency was identical to that reported in a past study [12]. We think that regions with large values in gaze distribution $\boldsymbol{g}$ contained informative features because these regions were attended to by the participants when judging gender. On the basis of this observation, we assume that assigning large weights to such regions is useful for a spatial attention mechanism in deep learning networks. Note that the gaze locations of the participants did not converge near regions of the feet of the subjects in the stimulus images.

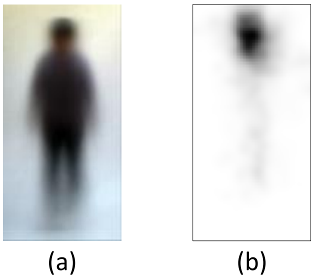

Fig. 6 Average images of people in the CUHK dataset (a) and gaze distribution $\boldsymbol{g}$ measured for 18 participants and 16 stimulus images (b).

5. Experiments

5.1 Experimental conditions

We evaluated the accuracy of gender recognition using our method on the CUHK dataset, which is part of the PETA dataset [33]. We used images of people with and without a background bias, as described in Section 3, where this bias was created using an obstacle in the background of the pedestrians. We chose the training samples T and test samples P according to the condition

- T (Woman, Man with an obstacle),
- P (Woman with an obstacle, Man).

Figure 7 shows examples of the training samples and test samples. There was no obstacle in front of women in the images used as training samples, whereas an obstacle was present in front of woman subjects in images used as test samples. We assessed the degree to which the accuracy of gender recognition was affected by the background bias in the training samples and test samples.

The CUHK dataset in the original PETA dataset contains some images of the same people with no person ID labels associated with them. We did not allow the same subject to appear in images of people in the training samples and the test samples. Attribute labels were used to identify images that could have featured the same person. The CUHK dataset comprised 476 images of men without an obstacle

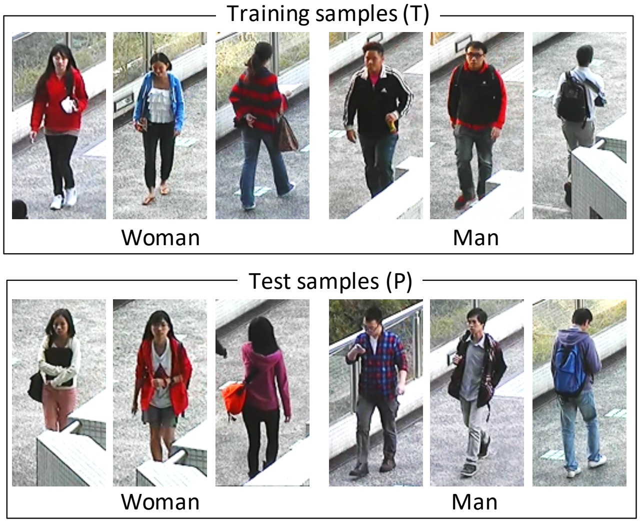

Fig. 7 Examples of the training samples T and test samples P used in our evaluation of gender recognition performance on the CUHK dataset.

in front of them, 426 of men with an obstacle in front, 419 images of women without an obstacle in front of them, and 355 of women with an obstacle. The size of the images was 80 pixels (width) by 160 pixels (height).

We generated five test sets by randomly selecting samples from the CUHK dataset. Each training set comprised images of 300 men and 300 women, and each test set comprised images of 50 men and 50 women. We made sure that there was no overlap of images of the same person between the training and test samples. The average and standard deviation of the accuracy of gender recognition was calculated from the five sets of the test samples and training samples.

5.2 Comparison with prevalent existing methods

The proposed GSA was designed by taking advantage of both the GP method and prevalent spatial attention mechanism methods. We thus expected the accuracy in terms of gender recognition to be superior to that of any of the original methods. To confirm this expectation, we assessed the accuracy of gender recognition when using the following methods.

- **B1**: We used a conventional CNN. A miniCNN [35] with two convolution layers and two pooling layers was used. The network excluding the GSA layer and gaze distribution $\boldsymbol{g}$, shown in Figure 3, was used.
- **B2**: We used the ResNet50 model with binary cross entropy loss between target gender labels and predicted output logits. We used part of the implementation provided by [36].
- **GP**: We used gaze pooling proposed by Sattar et al. [13]. The GP layer was added to the network of B1 between the last pooling layer and the FC layer. The output of the GP layer $\boldsymbol{y} = \boldsymbol{g} \circ \boldsymbol{x}$ was computed, where $\boldsymbol{x}$ is an input feature map and $\boldsymbol{g}$ is the gaze distribution of size $128 \times (10 \times 5)$.
- **SA**: We used the self-attention method proposed by Zhang et al. [14], which computes spatial attention weights. The SA layer was added after the pooling layer and placed at the end of the network of B1. The output of the SA layer $\boldsymbol{y} = \gamma \boldsymbol{o} + \boldsymbol{x}$ was computed, where $\boldsymbol{o}$ is an SA map.
- **CBAM**: We used the convolutional block attention module (CBAM) proposed by Woo et al. [15], which computes not only spatial attention weights but also channel attention weights. The CBAM was added after the pooling layer and placed at the end of the network of B1.
- **EA**: We used the efficient attention (EA) module proposed by Shen et al. [16], the implementation of which requires less memory and has a lower computational cost while maintaining the accuracy of the spatial attention mechanism. The EA was added after the pooling layer and placed at the end of the network of B1.
- **GSA**: We used our gaze-guided self-attention expressed by Eq. (1) in the network shown in Fig. 3. The GSA was added after the pooling layer and placed at the end of the network of B1.

Table 1 Accuracy (%) of gender recognition on the samples T and the samples P of the CUHK dataset.

B1	MiniCNN [35]	28.8±11.6
B2	ResNet50 [36]	35.8±4.4
GP	Gaze pooling [13]	56.0±1.9
SA	Self-attention [14]	29.8±10.6
CBAM	Convolutional block attention [15]	33.8±5.6
EA	Efficient attention [16]	24.2±7.2
GSA	Gaze-guided self-attention (ours)	**64.2**±**4.6**

We set the number of epochs to 50 and the batch size to 64 during the training of the CNN. A stochastic gradient descent method with momentum was used for optimization.

Table 1 shows the accuracy of gender recognition using the B1, B2, GP, SA, CBAM, EA, and GSA methods. We confirmed that the accuracy of our GSA was superior to those of B1 and B2. These baseline methods did not account for the background bias and were thus strongly affected by it. We also confirmed that the accuracy of our GSA was better than that of GP, which also uses the gaze distribution. Furthermore, the accuracy of our GSA was higher than those of the SA, CBAM, and EA, which are variations of spatial attention modules. We believe that the use of the gaze distribution helps solve the problem of the prevalent spatial attention mechanisms. Thus, the proposed method, by incorporating the gaze distribution into the attention mechanism, can improve recognition accuracy of the gender of a subject in an image even when the background biases of the training and test samples are different.

We evaluated the gender recognition accuracy when using the backbone network B2 instead of B1. Table 2 gives the accuracy for GP, SA, CBAM, EA, and GSA methods. We confirmed that the GSA method obtained better accuracy compared with the other methods. However, it is seen that the

Table 2 Accuracy (%) when using the backbone network B2 on the samples T and the samples P of the CUHK dataset.

GP	Gaze pooling [13]	42.6±1.2
SA	Self-attention [14]	36.8±2.6
CBAM	Convolutional block attention [15]	33.0±1.4
EA	Efficient attention [16]	39.2±1.3
GSA	Gaze-guided self-attention (ours)	**46.2±1.7**

accuracy of the B1 network with GSA was higher than that of the B2 network by comparing Table 1 with Table 2. We consider that the simple B1 network is suitable relative to the rich B2 network when using the training samples containing background bias.

5.3 Evaluation using training and test samples with/without an obstacle

We evaluated the gender recognition accuracy using the training and test samples with/without an obstacle in the CUHK dataset. We chose the training samples T_k and the test samples P_k according to condition 1

- T_1 (Woman, Man),
- P_1 (Woman with an obstacle, Man with an obstacle),

condition 2

- T_2 (Woman with an obstacle, Man),
- P_2 (Woman, Man with an obstacle),

and condition 3

- T_3 (Woman, Man),
- P_3 (Woman, Man).

All other experimental conditions except the training and test samples were the same as those in Section 5.2.

Table 3 gives the accuracy of gender recognition of the B1, B2, GP, SA, CBAM, EA, and GSA methods using the training and test samples with/without an obstacle. When using T_1 and P_1, the GSA method was more accurate than the other methods. When using T_2 and P_2, the accuracy of the GSA method was again better than that of the other methods. We believe that the gaze-guided attention mechanisms are suitable for the training and test samples containing background bias. In contrast, when using T_3 and P_3, the GSA method did not have the best performance. We consider that the conventional spatial attention mechanisms of the SA and EA methods are suitable for the training and test samples not containing background bias.

5.4 Discussion of attention maps

Our method is robust against the background bias of the training samples, whereas this bias is problematic for the conventional SA method. The proposed method can also assign adaptive spatial attention weights to features of the images, whereas GP is incapable of this. We visualized the arrays of the attention maps obtained using the compared methods to highlight the superiority of our method. We compared the following arrays:

Table 3 Accuracy (%) when using the training and test samples with/without obstacles on the samples T_k and the samples P_k of the CUHK dataset.

Training	T_1	T_2	T_3
Test	P_1	P_2	P_3
B1	74.8±1.2	24.6±6.2	86.2±1.3
B2	71.4±3.9	21.6±1.5	86.0±1.1
GP	76.0±1.1	53.4±2.6	85.2±1.5
SA	76.0±0.9	21.2±1.0	87.8±1.0
CBAM	73.2±2.0	31.8±3.1	81.2±1.2
EA	75.4±1.2	31.4±14.4	**88.0±1.8**
GSA	**77.8±0.4**	**60.6±1.4**	86.6±1.0

- the feature maps $\boldsymbol{x}$ entered into the GP layer and the arrays $\boldsymbol{g}\circ\boldsymbol{x}$ weighted by the gaze distribution, shown in Fig. 8(a);
- the feature maps $\boldsymbol{x}$ entered into the SA layer and the SA maps $\boldsymbol{o}$ shown in Fig. 8(b); and
- the feature maps $\boldsymbol{x}$ entered into the GSA layer, the arrays $\boldsymbol{g}\circ\boldsymbol{x}$ weighted by the gaze distribution, and the GSA maps $\hat{\boldsymbol{o}}$ shown in Fig. 8(c).

In the figure, the red regions represent large values and the blue regions small values of the arrays. To visualize the arrays, we averaged the values of all cells along the direction of the channel. The size of the channel of each array for the visualization was $C = 128$, with a height $H = 10$ and width $W = 5$. We used test samples that had been incorrectly classified by both the SA and GP methods but correctly classified by our GSA.

The results in Fig. 8(a) show that the background bias strongly affected the feature maps $\boldsymbol{x}$ of the GP method. In particular, the test samples featuring woman subjects responded strongly to the lower-right obstacle area. In the arrays $\boldsymbol{g}\circ\boldsymbol{x}$ weighted by the gaze distribution, the effect of the background was suppressed. However, the values of the head regions of men in the images were close to those of the women. The results of the prevalent SA method in Fig. 8(b) show that the background bias strongly affected both the feature maps $\boldsymbol{x}$ and SA maps $\boldsymbol{o}$. Figure 8(c) shows that the feature maps $\boldsymbol{x}$ were also affected by background bias. The effect of the background was suppressed in the arrays $\boldsymbol{g}\circ\boldsymbol{x}$. However, the values of the head regions in the images of men in $\boldsymbol{g}\circ\boldsymbol{x}$ were close to those of the head regions in the images of women. Our GSA map $\hat{\boldsymbol{o}}$ took different values near the head regions for each test sample. We think that these different values of $\hat{\boldsymbol{o}}$ helped in classifying the contents of the images correctly. Our method thus achieved a higher accuracy of gender recognition by adding adaptive weights $\hat{\boldsymbol{o}}$ to $\boldsymbol{g}\circ\boldsymbol{x}$ for each input feature map.

5.5 Evaluation of variations of our method

In the proposed method, the gaze distribution was incorporated into the SA mechanism using Eq. (1), but other variations are possible. We thus evaluated the accuracy of gender recognition using variations of our method, and compared them.

- **GSA**: Our method, $\hat{\boldsymbol{y}} = \gamma\hat{\boldsymbol{o}} + \boldsymbol{g} \circ \boldsymbol{x}$,

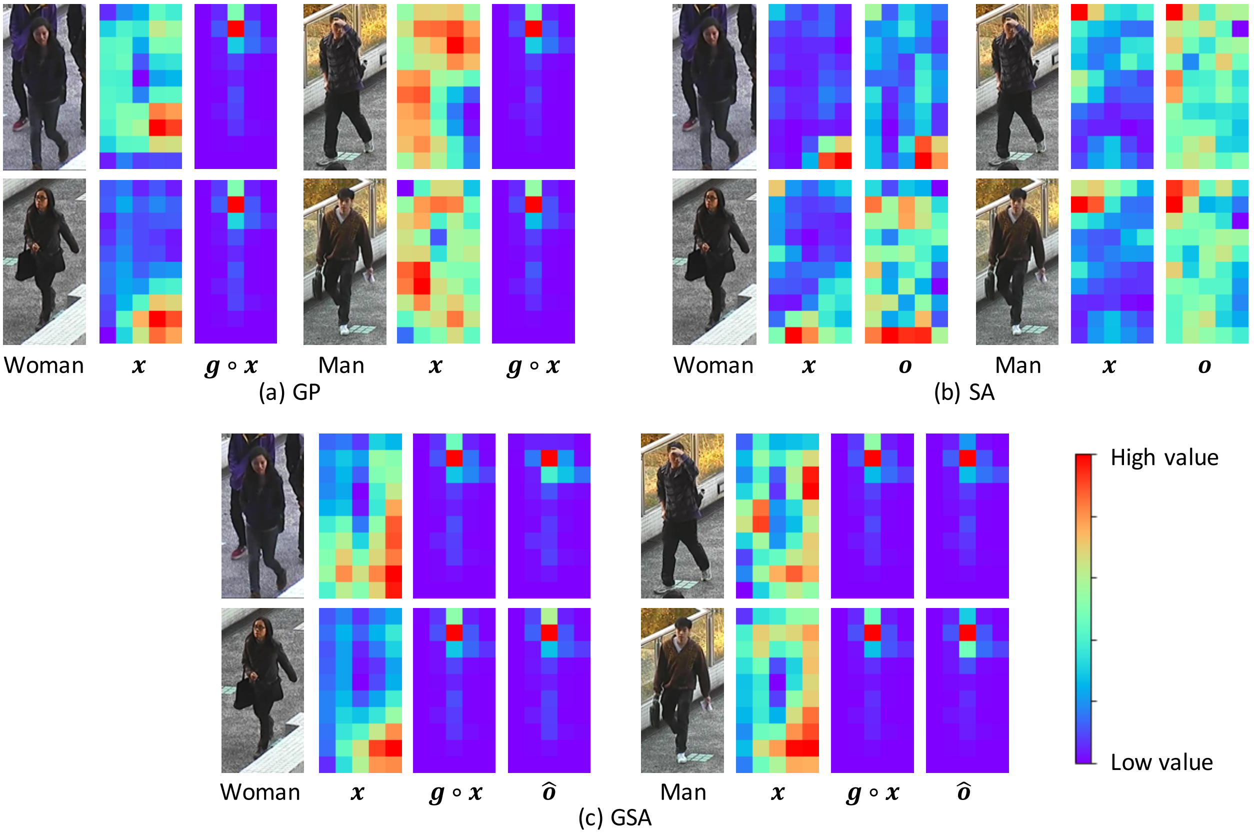

Fig. 8 Visualization of the arrays computed using different methods: GP, SA, and GSA. $\boldsymbol{x}$ represents an input feature map, $\boldsymbol{g}$ the gaze distribution, $\boldsymbol{o}$ the SA map, and $\hat{\boldsymbol{o}}$ the GSA map. To visualize the arrays, we averaged the values of all cells in the direction of the channel.

- **V1**: Variation 1, $\hat{\boldsymbol{y}} = \gamma \boldsymbol{o} + \boldsymbol{g} \circ \boldsymbol{x}$,
- **V2**: Variation 2, $\hat{\boldsymbol{y}} = \gamma \hat{\boldsymbol{o}} + \boldsymbol{x}$,
- **V3**: Variation 3, $\hat{\boldsymbol{y}} = \gamma_1 \boldsymbol{o} + \gamma_2 \boldsymbol{g} + \boldsymbol{x}$,
- **V4**: Variation 4, $\hat{\boldsymbol{y}} = (\boldsymbol{1} - \boldsymbol{g}) \circ \hat{\boldsymbol{o}} + \boldsymbol{g} \circ \boldsymbol{x}$,
- **SA**: Existing method, $\boldsymbol{y} = \gamma \boldsymbol{o} + \boldsymbol{x}$,

where $\hat{\boldsymbol{o}}$ is our GSA map, $\boldsymbol{g}$ is the gaze distribution, and $\boldsymbol{o}$ is the map generated using the prevalent SA method. All other experimental conditions except the computation of $\hat{\boldsymbol{y}}$ were the same as those in Section 5.1.

Table 4 shows the accuracy of gender recognition using V1 to V4, which are the variations of our proposed method. The GSA in this table is identical to the results of 5.2. The GSA obtained a higher recognition accuracy than all its variants. A comparison of the accuracy of SA and V1 makes it clear that $\boldsymbol{g} \circ \boldsymbol{x}$ had substantially increased accuracy. A comparison of the accuracy of SA and V2 shows that $\hat{\boldsymbol{o}}$ only slightly improved the accuracy. When comparing the accuracy of V1 and V2, $\boldsymbol{g} \circ \boldsymbol{x}$ was found to have contributed to increasing the accuracy more than $\hat{\boldsymbol{o}}$. Note that the effect of $\hat{\boldsymbol{o}}$ in GSA could not be ignored. We think that the synergistic effect of $\hat{\boldsymbol{o}}$ and $\boldsymbol{g} \circ \boldsymbol{x}$ in our method improved the accuracy of gender recognition in comparison with the other methods. By contrast, no improvement in accuracy was noted in V3. We checked the accuracy of V4 when the weighting manner for the attention map $\hat{\boldsymbol{o}}$ was changed from GSA. V4 has better accuracy than SA but worse accuracy than GSA.

Table 4 Accuracy (%) of gender recognition using V1 to V4, variations of our proposed method, on the samples T and the samples P of the CUHK dataset.

GSA	$\hat{\boldsymbol{y}} = \gamma \hat{\boldsymbol{o}} + \boldsymbol{g} \circ \boldsymbol{x}$	**64.2±4.6**
V1	$\hat{\boldsymbol{y}} = \gamma \boldsymbol{o} + \boldsymbol{g} \circ \boldsymbol{x}$	58.4±5.8
V2	$\hat{\boldsymbol{y}} = \gamma \hat{\boldsymbol{o}} + \boldsymbol{x}$	35.6±6.4
V3	$\hat{\boldsymbol{y}} = \gamma_1 \boldsymbol{o} + \gamma_2 \boldsymbol{g} + \boldsymbol{x}$	23.4±6.1
V4	$\hat{\boldsymbol{y}} = (\boldsymbol{1} - \boldsymbol{g}) \circ \hat{\boldsymbol{o}} + \boldsymbol{g} \circ \boldsymbol{x}$	51.0±0.9
SA	$\boldsymbol{y} = \gamma \boldsymbol{o} + \boldsymbol{x}$	29.8±10.6

5.6 Comparison with the case in which body regions are segmented

We evaluated the accuracy of proposed method for gender recognition in the case when body regions had been segmented before being entered as input images into a deep learning network. We used Mask R-CNN [27] as the segmentation technique. Figure 9 shows the training samples T_m and test samples P_m. The background obstacle that had been present at the bottom of the images was removed. The foot, which had been hidden by the obstacle, was not present in these images. Some of the legs that had been visible in the original images were removed. We used B1, B2, GP, SA, CBAM, and EA described in Section 5.2 as the com-

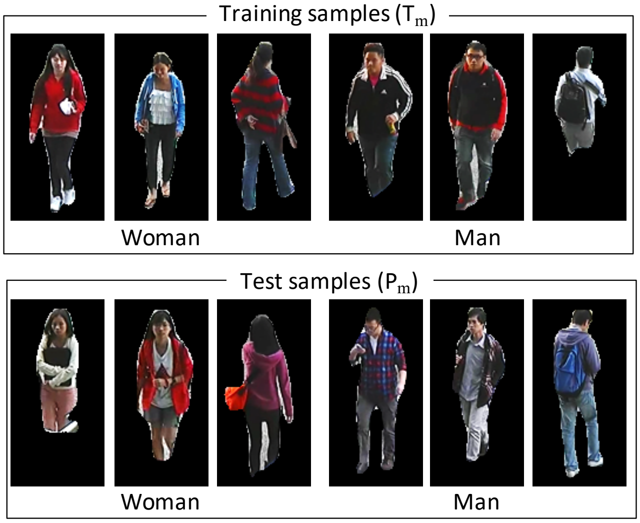

Fig. 9 Examples of the training samples T_m and test samples P_m with the segmented regions of pedestrian images.

parative methods. All other experimental conditions, except for the training samples, were the same as those described in Section 5.2.

The comparative methods using the segmented samples of T_m and P_m obtained the following gender recognition accuracy. The B1 method achieved an accuracy of 57.8±7.5%, the B2 method an accuracy of 59.8±0.7%, the GP method an accuracy of 60.8±1.2%, the SA method an accuracy of 59.0±0.9%, the CBAM method an accuracy of 52.0±2.1%, and the EA method an accuracy of 59.4±1.6%. In contrast, our GSA method using the non-segmented samples of T and P achieved an accuracy of 64.2±4.6% as shown in Table 1. We thus believe that our method with non-segmented body regions, which uses the gaze-guided self-attention mechanism, is more effective than the comparative methods with segmented body regions.

Additionally, we evaluated the accuracy when using individual segmentation masks instead of the gaze distribution $\boldsymbol{g}$ in the GSA layer. The pixel value in the segmentation mask is 1 for the body region and 0 for the background region. We interpolated the size of the segmentation mask from 80 × 160 pixels to 5 × 10 pixels for fitting to the size of the feature map in the GSA layer. All other experimental conditions except for the segmentation masks were the same as those described in Section 5.2. This comparative method using the segmentation masks achieved an accuracy of 47.2±1.6%. Our GSA method using $\boldsymbol{g}$ achieved an accuracy of 64.2±4.6% as described in Table 1. We believe that the use of the gaze distribution improves the accuracy relative to the use of segmentation masks.

5.7 Evaluation of a case featuring a different dataset

5.7.1 RAP dataset

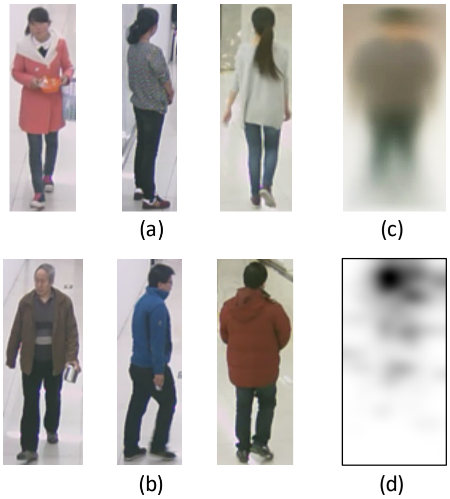

Fig. 10 Examples of stimulus images of the RAP dataset: (a) woman samples and (b) man samples. The average image of the pedestrians (c) and gaze distribution $\boldsymbol{g}$ (d).

To confirm the effectiveness of our GSA method, we evaluated the accuracy of gender recognition using the RAP dataset [37], which is a dataset different from the CUHK dataset used in Section 5.2. We measured gaze distribution $\boldsymbol{g}$ using stimulus images selected from the RAP dataset. Figure 10(a) and (b) show examples of the stimulus images given to the participants. We used 32 stimulus images. The size of each stimulus image differed, and the mean size of the height was 315.5 pixels and the mean size of the width was 130.4 pixels. We set the same proportions of man and woman pedestrians and the same proportions of different body orientations in the stimulus images. We used the measurement settings described in Section 4.3. Sixteen participants (six women and ten men, average age of 22.4±1.0 years) participated in this study. The accuracy of gender recognition performed by the participants was 96.5%.

Before describing the results of the gaze distribution, we checked the alignment of the pedestrian regions. For this purpose, we computed the average of the pedestrian images in the RAP dataset, as shown in Figure 10(c). We rescaled the height and width of the pedestrian images to 160 and 80 pixels, respectively. We see that the head, torso, and legs appeared in the upper, middle, and lower regions of the average image, respectively. Figure 10(d) shows the gaze distribution measured using the stimulus images in the RAP dataset. The figure shows that the participants' gazes are strongly gathered on the head region. This result has the same tendency as the gaze distribution measured for the CUHK dataset shown in Figure 6(b).

5.7.2 Case of training samples without an obstacle

To evaluate the accuracy of gender recognition when using the training samples without an obstacle and the test samples with an obstacle, we used the subset CAM25 included in the RAP dataset as the training sample and the subsets CAM18, CAM27, and CAM31 as the test samples. Fig-

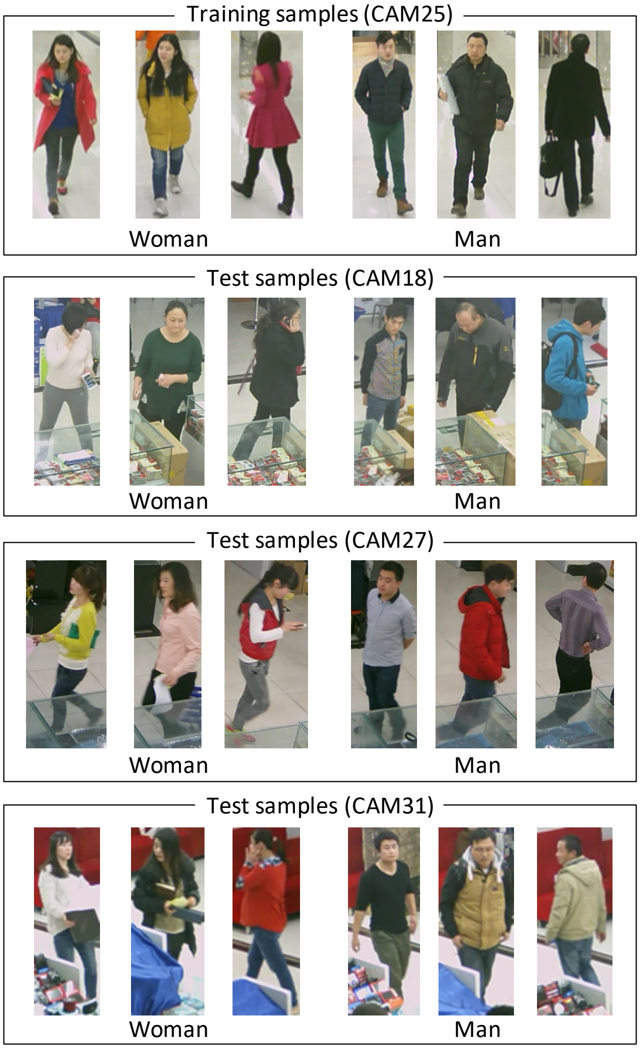

Fig. 11 Examples of training samples and test samples for evaluation of recognition performance on the RAP dataset.

ure 11 shows examples of training and test samples. We see background biases between training and test samples. There was no obstacle in front of both woman and man pedestrians in CAM25, whereas an obstacle was present in CAM18, CAM27, and CAM31. The RAP dataset comprised 3317 images (961 women, 2355 men) in CAM25, 964 images (293 women, 667 men) in CAM18, 2100 images (695 women, 1402 men) in CAM27, and 1321 images (476 women, 843 men) in CAM31. To maintain a 1:1 proportion for the number of images between women and men, we randomly selected the man images to match the number of the woman images. This random image selection was repeated three times, and the average accuracy was computed.

Table 5 shows the accuracy of gender recognition using the B1, B2, GP, SA, CBAM, EA, and GSA methods when using the test subsets CAM18, CAM27, and CAM31 with the training subset CAM25 in the RAP dataset. We confirmed that the accuracy of our GSA was superior to those of the B1, B2, and GP methods. We also confirmed that our GSA obtains higher accuracy than the prevalent spatial attention mechanisms of SA, CBAM, and EA. We believe that the proposed method using the gaze-guided self-attention mechanism has promising effectiveness on the RAP dataset as well as the CUHK dataset.

Table 5 Accuracy (%) of gender recognition on the RAP dataset when using the test subsets CAM18, CAM27, and CAM31 with the training subset CAM25.

Training	CAM25	CAM25	CAM25	Average
Test	CAM18	CAM27	CAM31	
B1	72.4±0.2	67.6±0.4	72.9±0.2	71.0±2.4
B2	69.0±0.2	73.7±0.4	73.5±0.7	72.1±2.2
GP	73.2±1.1	72.9±0.2	77.8±0.3	74.7±2.3
SA	71.9±0.3	67.8±0.1	72.8±0.2	70.8±2.2
CBAM	72.6±0.7	70.1±0.6	74.7±0.3	72.5±2.0
EA	72.9±0.2	68.0±0.2	74.6±0.5	71.8±2.8
GSA (ours)	**77.0±0.5**	**75.7±0.3**	**80.0±0.1**	**77.6±1.8**

5.7.3 Case of training samples with an obstacle

Furthermore, we evaluated the accuracy of gender recognition when using the training samples with an obstacle in the RAP dataset. We used the subsets of CAM18, CAM27, and CAM31. One subset was used as the training samples and another one as the test samples. We evaluated the accuracy using the permutation ${}_3P_2$ of the three subsets. All other experimental conditions except the training samples were the same as those in Section 5.7.2.

Table 6 gives the accuracy when using the training samples with an obstacle on the RAP dataset. We see that the accuracy of our GSA method was also superior to those of the B1, B2, GP, SA, CBAM, and EA methods. We believe that our GSA method has improved performance when using the training samples with an obstacle.

6. Conclusions

To improve the accuracy of gender recognition, we proposed here a method of designing a spatial attention mechanism for deep learning networks based on the gaze distribution that can be trained on samples featuring background bias and yet deliver high accuracy. The prevalent GP has the problem that it assigns only a uniform attention weight to each input feature map. Moreover, prevalent attention mechanisms, such as SA, CBAM, and EA, are affected by background bias in the training samples. To solve these problems, we proposed the GSA, which assigns adaptive attention weights to each input feature map while suppressing the effect of the background bias using the gaze distribution. We confirmed that our method, inspired by the gaze distribution representing human visual attention, improves the accuracy of gender recognition compared with prevalent methods, based on the GP, SA, CBAM, and EA, in cases involving different backgrounds in training and test samples on CUHK and RAP datasets. We also showed that incorporating the gaze distribution into the spatial attention mechanism of deep learning

Table 6 Accuracy (%) of gender recognition on the RAP dataset when using the training samples with an obstacle. We evaluated the permutation of the CAM18, CAM27, and CAM31 subsets for the training samples and test samples.

Training	CAM18	CAM18	CAM27	CAM27	CAM31	CAM31	Average
Test	CAM27	CAM31	CAM18	CAM31	CAM18	CAM27	
B1	66.5±0.5	60.0±0.7	68.8±0.6	68.6±0.6	64.1±0.6	67.0±0.4	65.7±3.3
B2	61.8±0.6	56.8±0.5	70.0±0.9	70.3±0.1	61.9±0.6	67.9±0.6	64.4±5.2
GP	66.7±0.3	66.9±1.5	71.2±0.6	71.4±0.3	66.6±0.8	70.7±0.3	68.7±2.4
SA	65.7±0.7	60.2±1.3	69.0±0.4	69.5±0.6	64.5±0.5	67.3±0.5	65.9±3.4
CBAM	64.2±0.7	61.8±1.1	70.0±0.6	70.2±0.9	65.1±0.7	68.2±0.2	66.4±3.3
EA	65.3±0.4	62.8±0.4	68.5±0.8	70.0±0.3	64.6±0.8	67.0±0.8	66.3±2.7
GSA (ours)	**68.5±0.7**	**71.3±0.4**	**73.8±0.4**	**73.2±0.7**	**69.4±0.1**	**72.6±0.3**	**71.4±2.1**

has the potential to solve the background bias problem.

In future work, we intend to evaluate the effectiveness of our method in classifying attributes other than gender, such as age and clothing. We will also examine an image-by-image gaze distribution instead of the averaged, single-gaze distribution used in our GSA. We also intend to incorporate the gaze distribution into conventional attention mechanisms other than SA; e.g., pairwise self-attention and patchwise self-attention [38].

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant No. JP20K11864. We would like to thank Ms. Sayaka YAMAGUCHI for her cooperation in our experiments.

References

[1] C.B. Ng, Y.H. Tay, and B. Goi, "Recognizing human gender in computer vision: a survey," Proceedings of the Pacific Rim International Conference on Artificial Intelligence, pp.335–346, 2012.

[2] S.A. Khan, M. Nazir, S. Akram, and N. Riaz, "Gender classification using image processing techniques: A survey," Proceedings of the IEEE 14th International Multitopic Conference, pp.25–30, 2011.

[3] M. Fayyaz, M. Yasmin, M. Sharif, and M. Raza, "J-ldfr: joint low-level and deep neural network feature representations for pedestrian gender classification," Neural Computing and Applications, pp.1–31, 2020.

[4] C. Tang, L. Sheng, Z. Zhang, and X. Hu, "Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.4996–5005, 2019.

[5] H. Zeng, H. Ai, Z. Zhuang, and L. Chen, "Multi-task learning via co-attentive sharing for pedestrian attribute recognition," Proceedings of the IEEE International Conference on Multimedia and Expo, pp.1–6, 2020.

[6] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry, "Noise or signal: The role of image backgrounds in object recognition," Proceedings of the International Conference on Learning Representations, pp.1–28, 2021.

[7] Y. Yu, J. Choi, Y. Kim, K. Yoo, S. Lee, and G. Kim, "Supervising neural attention models for video captioning by human gaze data," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.6119–6127, 2017.

[8] T. Qiao, J. Dong, and D. Xu, "Exploring human-like attention supervision in visual question answering," Proceedings of the 32nd AAAI Conference on Artificial Intelligence, pp.7300–7307, 2018.

[9] J. Wu, S. Zhong, Z. Ma, S.J. Heinen, and J. Jiang, "Gaze aware deep learning model for video summarization," Proceedings of the Pacific Rim Conference on Multimedia, pp.285–295, 2018.

[10] Y. Xia, Z. Liu, Y. Yan, Y. Chen, L. Zhang, and R. Zimmermann, "Media quality assessment by perceptual gaze-shift patterns discovery," IEEE Transactions on Multimedia, vol.19, no.8, pp.1811–1820, 2017.

[11] N. Murrugarra-Llerena and A. Kovashka, "Learning attributes from human gaze," Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp.510–519, 2017.

[12] M. Nishiyama, R. Matsumoto, H. Yoshimura, and Y. Iwai, "Extracting discriminative features using task-oriented gaze maps measured from observers for personal attribute classification," Pattern Recognition Letters, vol.112, pp.241–248, 2018.

[13] H. Sattar, A. Bulling, and M. Fritz, "Predicting the category and attributes of visual search targets using deep gaze pooling," Proceedings of the IEEE International Conference on Computer Vision Workshops, pp.2740–2748, 2017.

[14] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," Proceedings of the 36th International Conference on Machine Learning, pp.7354–7363, 2019.

[15] S. Woo, J. Park, J.Y. Lee, and I.S. Kweon, "Cbam: Convolutional block attention module," Proceedings of the European Conference on Computer Vision, pp.3–19, 2018.

[16] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li, "Efficient attention: Attention with linear complexities," Proceedings of the Winter Conference on Applications of Computer Vision, 2021.

[17] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.6450–6458, 2017.

[18] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.7794–7803, 2018.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.7132–7141, 2018.

[20] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.1179–1188, 2018.

[21] J. Yang, J. Fan, Y. Wang, Y. Wang, W. Gan, L. Liu, and W. Wu, "Hierarchical feature embedding for attribute recognition," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.13052–13061, 2020.

[22] K. Han, J. Guo, C. Zhang, and M. Zhu, "Attribute-aware attention model for fine-grained representation learning," Proceedings of the 26th ACM International Conference on Multimedia, pp.2040–2048, 2018.

[23] Z. Yang, L. Huang, Y. Chen, Z. Wei, S. Ahn, G. Zelinsky, D. Samaras, and M. Hoai, "Predicting goal-directed human attention using inverse reinforcement learning," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.193–202, 2020.

[24] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. Rehg, and V. Singh, "Gaze-enabled egocentric video summarization via constrained submodular

maximization," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.2235–2244, 2015.

[25] Y. Sugano, Y. Ozaki, H. Kasai, K. Ogaki, and Y. Sato, "Image preference estimation with a data-driven approach: A comparative study between gaze and image features," Eye Movement Research, vol.7, no.3, pp.862–875, 2014.

[26] N. Karessli, Z. Akata, B. Schiele, and A. Bulling, "Gaze embeddings for zero-shot image classification," Proceedings of the IEEE conference on computer vision and pattern recognition, pp.4525–4534, 2017.

[27] H. Kaiming, G. Georgia, D. Piotr, and G. Ross, "Mask r-cnn," Proceedings of the IEEE International Conference on Computer Vision, pp.2980–2988, 2017.

[28] D. Novotny, S. Albanie, D. Larlus, and A. Vedaldi, "Semi-convolutional operators for instance segmentation," Proceedings of the European Conference on Computer Vision, 2018.

[29] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.39, no.12, pp.2481–2495, 2017.

[30] S.K. Choudhury, P.K. Sa, S. Bakshi, and B. Majhi, "An evaluation of background subtraction for object detection vis-a-vis mitigating challenging scenarios," IEEE Access, vol.4, pp.6133–6150, 2016.

[31] I. Setitra and S. Larabi, "Background subtraction algorithms with post-processing: A review," Proceedings of the 22nd International Conference on Pattern Recognition, pp.2436–2441, 2014.

[32] M. Babaee, D.T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," Pattern Recognition, vol.76, pp.635–649, 2018.

[33] Y. Deng, P. Luo, C.C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," Proceedings of the 22nd ACM international conference on Multimedia, pp.789–792, 2014.

[34] M. Bindemann, "Scene and screen center bias early eye movements in scene viewing," Vision Research, vol.50, no.23, pp.2577 – 2587, 2010.

[35] G. Antipov, S. Berrani, N. Ruchaud, and J. Dugelay, "Learned vs. hand-crafted features for pedestrian gender recognition," Proceedings of the 23rd ACM International Conference on Multimedia, pp.1263–1266, 2015.

[36] J. Jia, H. Huang, W. Yang, X. Chen, and K. Huang, "Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method," CoRR, abs/2005.11909, 2020.

[37] D. Li, Z. Zhang, X.Chen, H.Ling, and K.Huang, "A richly annotated dataset for pedestrian attribute recognition," CoRR, abs/1603.07054, 2016.

[38] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10073–10082, 2020.

Masashi Nishiyama is an associate professor in Graduate School of Engineering at Tottori University, Japan. He received his M.S. degrees in Engineering from Graduate School of Natural Science and Technology, Okayama University, Japan, in 2002. He joined in Corporate Research & Development Center, TOSHIBA Corporation, during 2002–2015. He received his Ph.D. degrees in Interdisciplinary Information Studies from Graduate School of Interdisciplinary Information Studies, University of Tokyo, Japan, in 2011. His recent research has focused on developing novel principles for representing identities, attributes, and behaviors of humans in video sequences.

Michiko Inoue received her M.S. degrees in Engineering from Graduate School of Sustainability Science, Tottori University, in 2019. She currently attends the doctor's course at Graduate School of Engineering, Tottori University. She is engaged in studies relating to person attribute recognition using gaze distribution.

Yoshio Iwai graduated from Osaka University in 1992 and completed the M.S. and doctoral programs in 1994 and 1997, respectively. He was then appointed a research associate at the university, subsequently becoming an associate professor. From 2004 to 2005, he was a visiting researcher at the University of Cambridge. He is currently a professor in the Graduate School of Engineering at Tottori University. He is engaged in studies relating to computer vision and pattern recognition. He is a member of IEEE, the Information Processing Society. He holds a D. Eng. Degree.