Pattern Recognition Letters

Extracting Discriminative Features using Task-oriented Gaze Maps Measured from Observers for Personal Attribute Classification

Masashi Nishiyama[a,b,**], Riku Matsumoto[a], Hiroki Yoshimura[a], Yoshio Iwai[a,b]

[a] *Graduate School of Engineering, Tottori University, 101 Minami 4-chome, Koyama-cho, Tottori, 680-8550, Japan*
[b] *Cross-informatics Research Center, Tottori University, 101 Minami 4-chome, Koyama-cho, Tottori, 680-8550, Japan*

ABSTRACT

We discuss how to reveal and use the gaze locations of observers who view pedestrian images for personal attribute classification. Observers look at informative regions when attempting to classify the attributes of pedestrians in images. Thus, we hypothesize that the regions in which observers' gaze locations are clustered will contain discriminative features for the classifiers of personal attributes. Our method acquires the distribution of gaze locations from several observers while they perform the task of manually classifying each personal attribute. We term this distribution a task-oriented gaze map. To extract discriminative features, we assign large weights to the region with a cluster of gaze locations in the task-oriented gaze map. In our experiments, observers mainly looked at different regions of body parts when classifying each personal attribute. Furthermore, our experiments show that the gaze-based feature extraction method significantly improved the performance of personal attribute classification when combined with a convolutional neural network or metric learning technique.

1. Introduction

Personal attributes such as gender, clothing, and carried items, which are of interest in the field of soft-biometrics [7, 27, 32, 6], help the collection of statistical data about people in public spaces. Furthermore, personal attributes have many potential applications, such as video surveillance and consumer behavior analysis. In general, pedestrians captured on video or in still images are used for personal attribute classification. Researchers have proposed several methods for automatically classifying personal attributes in pedestrian images; for example, techniques involving convolutional neural networks (CNNs) [30, 25, 29, 22] and metric learning [21, 41] have been proposed. The existing methods can extract discriminative features for personal attribute classification and obtain high accuracy when many training samples containing diverse pedestrian images are acquired in advance. However, the collection of a sufficient number of training samples is very time consuming. Unfortunately, the performance of the existing methods has been found to decrease when the number of training samples is small.

People correctly and quickly classify personal attributes. We believe that people have the visual ability to extract features from an individual. For instance, people correctly classify gender from facial images [3, 4]. In the research field of cognitive science, Yarbus [38] reported that human observers can recognize personal attributes in a scene image with high accuracy when they are given different tasks such as remembering the clothes worn by the individuals or estimating their ages. In this interesting research, he noticed that the observers paid attention to different regions in the scene when they tackled a different task even though they viewed the same image. Recently, researchers have made some efforts to analyze the role of task in various applications [19, 14, 13]. Based on these observations, we hypothesize that people pay attention to different informative regions in pedestrian images while tackling various tasks of personal attribute classification.

It may be possible to reproduce human visual abilities via an algorithm on a computer with a small number of training samples such that the classification performance is equivalent to that of humans. With respect to object recognition, several existing methods for mimicking human visual abilities have been proposed [33, 12, 40]. To mimic human visual ability, the exist-

**Corresponding author:
e-mail: nishiyama@tottori-u.ac.jp (Masashi Nishiyama)

ing methods exploited a saliency map computed from low-level features in a given image using techniques such as those described in [17, 39, 42]. However, the use of the saliency map does not sufficiently represent human visual abilities because of the deep mechanisms of human vision.

An increasing number of pattern recognition studies, specifically those attempting to mimic human visual ability, have measured the gaze locations of observers [37, 11, 36, 31, 18]. These gaze locations have great potential for the collection of informative features during various recognition tasks. Very recently, state-of-the-art techniques [28, 26] have demonstrated that gaze locations can help to extract informative features for the attribute classification of fashion clothing and face images. However, these existing methods do not consider how to treat the case in which observers tackle different tasks for body attributes in the same pedestrian image. We believe that the informative region of the body for each classifier is significantly different for each task of personal attribute classification.

In this paper, we consider the challenging case in which participants in an experiment are given different tasks of personal attribute classification while viewing the same pedestrian images. We confirm whether or not test participants look at different regions when tackling each task. We determine whether or not the gaze locations measured from the participants play an important role in the personal attribute classification. To this end, we generated a task-oriented gaze map from the distribution of gaze locations recorded while participants viewed images to complete each task of manually classifying personal attributes. The high values in a task-oriented gaze map correspond to regions that are frequently viewed by participants. We assume that these regions contain discriminative features for each classifier of a personal attribute because they appear to be useful when the participants are tackling each task of personal attribute classification. When extracting features to learn the classifier, larger weights are given to the regions of the pedestrian images that correspond to the attention regions of the task-oriented gaze maps. The experimental results indicate that our method improves the accuracy of feature extraction when using a CNN or metric learning technique with a small number of training samples.

This paper is organized as follows. Section 2 describes related work, Section 3 describes the generation of task-oriented gaze maps, and Section 4 describes feature extraction using the maps. Our concluding remarks are given in Section 5.

2. Related work

To mimic human visual ability, existing methods [33, 12, 40] involve the saliency maps of object images with representations of the regions that draw visual attention. Walther et al. [33] combined a recognition algorithm with a saliency map generated from low-level features of gradients of color and intensity using [17]. Researchers have developed techniques [12, 40] that use the object labels of images in addition to the low-level features of objects to generate saliency maps. Furthermore, existing methods [39, 42] add image boundary information in low-level features to generate saliency maps with high accuracy. However, the use of low-level features to generate a saliency map does not sufficiently represent human visual abilities. Our method exploits the use of gaze locations instead of a saliency map to increase the performance of personal attribute classification.

Existing methods [37, 11, 36, 31, 18] aim to design an algorithm that is close to the human visual ability by measuring gaze locations from observers. Xu et al. [37] generated saliency maps of facial images using prior gaze locations from participants who viewed the images. They reported that the generated saliency maps represented high-level features corresponding to the facial feature points of the eyes, nose, and mouth. Furthermore, gaze locations are used in applications involving action recognition or image preference estimation. Fathi et al. [11] classified actions by simultaneously inferring regions where gaze locations were gathered via an egocentric camera. Xu et al. [36] demonstrated that the use of gaze tracking information (such as fixation and saccade) significantly helps the task of egocentric video summarization. Sugano et al. [31] estimated more highly preferable images using gaze locations and low-level features. Karessli et al. [18] classified objects using only gaze features without object labels for zero-shot learning. Additionally, Sattar et al. [28] predicted the category and attribute of fashion clothing images by embedding gaze distributions in the pooling layers of a CNN. Murrugrra-Llerena et al. [26] classified the attributes of shoe and face images using a binary masking of gaze distributions. However, the existing methods do not consider the variation of gaze locations with respect to body regions when participants tackle several different tasks using pedestrian images. We attempt to observe the variation of gaze locations for different tasks of personal attribute classification. Based on the gaze locations measured with respect to body regions, we develop a method for extracting features to improve the performance of personal attribute classification.

3. Generating task-oriented gaze maps

3.1. Gaze locations in personal attribute classification

Here, we consider the regions of pedestrian images that are frequently looked at by observers when manually classifying personal attributes. For instance, Hsiao et al. [15] found that observers looked at a region around the nose when they identified individuals from a facial image. In the case of gender classification, we believe that the human face plays an important role. However, a pedestrian image contains not only a face but also a body. Yarbus [38] found that observers look at a different region in a scene image when tackling each task of personal attribute classification. However, he did not analyze the distributions of gaze locations in pedestrian images for personal attribute classification. Thus, we attempt to discern the regions of pedestrian images that tend to collect gaze locations from observers when given several different manual personal attributes classification tasks. Note that we assume that the alignment of the pedestrian images has already been completed using a pedestrian detection technique such as [9, 16]. The details of our method are also described below.

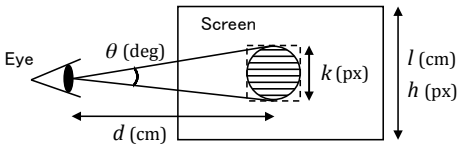

Fig. 1. Parameters used to determine kernel size.

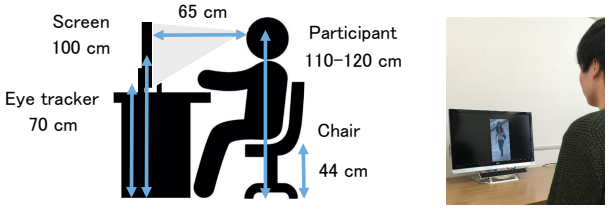

Fig. 2. Setup for capturing gaze locations.

3.2. Generation algorithm

To generate a task-oriented gaze map, we use a gaze tracker to acquire gaze locations while a test participant views a pedestrian image on a screen. We prepare T tasks, P participants, and N pedestrian images. Given a gaze location (x_f, y_f) in a certain frame f, gaze map $g_{t,p,n,f}(x, y)$ is labeled 1 when $x = x_f, y = y_f$; otherwise, it is labeled 0, where p is a participant, t is a task, and n is a pedestrian image. Note that the participant not only looks at point (x_f, y_f) on each pedestrian image, but also the region surrounding this point. Thus, we apply a Gaussian kernel to the measured gaze map $g_{t,p,n,f}(x, y)$. To determine the size k of the Gaussian kernel, we use the following equation:

$$k = \frac{2dh}{l} \tan \frac{\theta}{2}, \tag{1}$$

where d is the distance between the screen and the participant, θ is the angle of the region surrounding a measured gaze point, l is the vertical length of the screen, and h is the vertical resolution of the screen. Figure 1 illustrates the parameters used to determine the kernel size. We assume that each pixel on the screen is square. We aggregate each $g_{t,p,n,f}(x, y)$ to $g_{t,p,n}(x, y)$ to represent the distribution of gaze locations in a certain pedestrian image as

$$g_{t,p,n}(x, y) = \sum_{f=1}^{F_{t,p,n}} k(u, v) * g_{t,p,n,f}(x, y), \tag{2}$$

where $F_{t,p,n}$ is the time taken to classify personal attribute by a participant, $*$ is the convolution operator, and $k(u, v)$ is a Gaussian kernel of size $k \times k$. We apply L1-norm normalization as $\|g_{t,p,n}(x, y)\| = 1$ because $F_{t,p,n}$ is different for each measurement. Furthermore, we aggregate $g_{t,p,n}(x, y)$ into a single gaze map for all participants and all pedestrian images. An aggregated gaze map $g_t(x, y)$ representing the distribution of gaze locations is represented as

$$g_t(x, y) = \sum_{p=1}^{P} \sum_{n=1}^{N} g_{t,p,n}(x, y). \tag{3}$$

Note that we apply a scaling technique to the aggregated gaze maps as $\tilde{g}_t(x, y) = g_t(x, y)/\max(g_t(x, y))$. Finally, $\tilde{g}_t(x, y)$ is a task-oriented gaze map.

3.3. Experiments for task-oriented gaze map generation

To investigate the task-oriented gaze maps for personal attribute classification, we captured gaze locations for $P = 14$ test participants (seven males and seven females, with an average age of 22.6 ± 1.3 years old) using a standing eye tracker (GP3 Eye Tracker, sampling rate 60 Hz). All participants were Asian with Japanese nationality. We used a 24-inch display (size 53.1 × 29.9 cm, 1920 × 1080 pixels) as a screen. The vertical distance between the screen and the participant was 65 cm in the setting, as illustrated in Figure 2. The height from the floor to the eyes of the participant was between 110 cm and 120 cm. The participants sat on a chair in a room with no direct sunlight (illuminance 825 lx). We gave participants the following four tasks:

- t_1: Gender classification—to determine whether the pedestrian is male or female;
- t_2: Logo classification—to determine whether the pedestrian is wearing a top with or without logos on it;
- t_3: Short-sleeve classification—to determine whether the pedestrian is wearing short sleeved clothing or not;
- t_4: Backpack classification—to determine whether the pedestrian is carrying a backpack or not.

The gender attribute is a representative physical characteristic, as described in [7]. We attempted to observe which regions of the whole body were viewed by the participants. The logo and short-sleeve attributes are clothing attributes, and a backpack is a carried-item attribute. These attributes are categorized as adhered human characteristics, as described in [7]. We assumed that participants looked around the torso in these attribute classification tasks. We attempted to observe the range of the gaze-gathered region of each attribute.

We used 4,563 pedestrian images from the CUHK dataset included in the PETA dataset [8] with attributes labels. We selected $N = 8$ pedestrian images of Figure 3 from the dataset. To select the eight pedestrian images, we first checked the multi-attribute labels to determine if a pedestrian is wearing a top with a logo (Logo: yes), is wearing short-sleeved clothing (Short-sleeve: yes), and is carrying a backpack (Backpack: yes), simultaneously. The number of pedestrian images with these multi-attribute labels was 49 (Male: 46, Female: 3) in the CUHK dataset. We selected the four pedestrian images at the top of Figure 3 while keeping the ratio of directions (front, back, left, and right) equal. We also selected the remaining pedestrian images of Figure 3 in the same manner. We used the same pedestrian images for the four tasks for each participant because our aim was to investigate whether the gaze locations change depending on the tasks. We enlarged the pedestrian images from 80 × 160 pixels to 480 × 960 pixels to display the stimulus images on the screen. We used this image size because it is the maximum size that fits within the vertical resolution of the display. To avoid a center bias [2, 5] in which the gaze locations are grouped in the center of the screen, we changed the positions of the pedestrian images by randomly adding offsets in the range of ±720 pixels horizontally and ±60 pixels vertically.

Gender	Female	Male	Male	Male
Logo	Yes	Yes	Yes	Yes
Short-sleeve	Yes	Yes	Yes	Yes
Backpack	Yes	Yes	Yes	Yes

Gender	Male	Female	Female	Female
Logo	No	No	No	No
Short-sleeve	No	No	No	No
Backpack	No	No	No	No

Fig. 3. Pedestrian images for generating task-oriented gaze maps.

Fig. 4. Examples of displayed stimulus images.

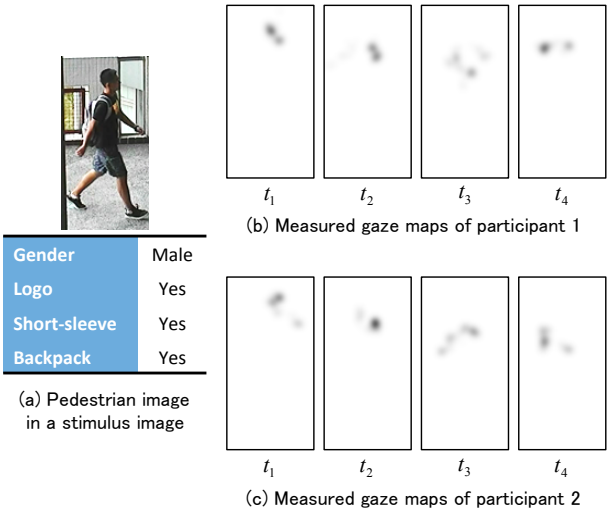

Fig. 5. Examples of measured gaze maps of each participant for different tasks. (a) Pedestrian image displayed in a stimulus image. (b) or (c) Gaze maps measured from two participants viewing the pedestrian image.

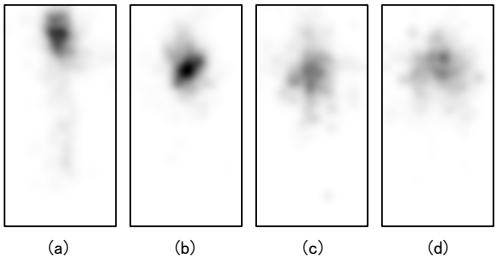

Fig. 6. Task-oriented gaze maps for (a) gender, (b) logo, (c) short-sleeve, and (d) backpack classification.

Figure 4 shows examples of the stimulus images displayed on the screen.

We asked participants to complete the task of personal attribute classification and measured gaze locations according to the following procedure:

- P1: We gave the participant a single attribute classification task to complete.
- P2: We displayed a flat gray image on the screen for 1 s.
- P3: We displayed a stimulus image that included a pedestrian image on the screen for 2 s. Prior to the trial, the participants had been instructed to keep looking at the image.
- P4: We displayed a flat black image on the screen for 2 s and the participant verbally reported whether or not the attribute for that task appeared in the image.
- P5: We repeated P2 to P4 until all the eight pedestrian images (in random order) had been displayed.

Note that each participant tackled the procedure P1 to P5 for all the four tasks of attribute classification. We also gave the tasks in random order. In our preliminary experiment, we observed that participants first assessed the position of the pedestrian image on the screen and then, after establishing the position of the image, attempted to determine the answer of the given task. To determine $F_{t,p,n}$, we set the start time as the point at which the gaze first stopped on the pedestrian image for more than 440 ms, and the end time as the point at which the pedestrian image disappeared. In this scenario, the average $F_{t,p,n}$ between the start and end times was 1.56 ± 0.38 s. The accuracy of gender classification by the participants was 100.0%, the accuracy of logo classification was 91.1%, the accuracy of short-sleeve classification was 97.3%, and the accuracy of backpack classification was 94.6%.

We set $\theta = 3^\circ$ in Equation (1) by considering the range of the fovea, which is about two degrees, as described in [10] and the error of the eye tracker, which is about one degree. We used a kernel size of $k = 125$ for the enlarged pedestrian images (480×960 pixels). Further, the size of the gaze maps was downsized by 80×160 after adjustment from the original size of the pedestrian images.

Figure 5 shows examples of the measured gaze maps $g_{t,p,n}(x, y)$ of different tasks for a pedestrian image. We selected gaze maps measured from two participants. The dark regions in the gaze maps represent the gaze locations gathered from the participants. The minimum intensities in Figure 5 represent the maximum values of all $g_{t,p,n}(x, y)$. We observed that participants frequently concentrated their gaze on a certain region according to each task even though they viewed the same pedestrian image. For instance, both participants looked around the head when tackling gender classification task t_1 and they looked around the chest when tackling logo classification task t_2.

Figures 6 (a)–(d) shows the task-oriented gaze maps before scaling for gender, logo, short-sleeve, and backpack classification, respectively. To consider the properties of the task-

oriented gaze map, we check how the gaze maps align with the pedestrian images of Figure 3. We can see that the rough positions of the body parts in the pedestrian images are well aligned. Given these results, we consider the task-oriented gaze maps to include the following regions:

- The region around the head gathered a large number of gaze locations for task t_1 (gender classification).
- The region around the chest gathered a large number of gaze locations for task t_2 (logo classification).
- The region around the upper body gathered a moderate number of gaze locations for tasks t_3 and t_4 (short-sleeve and backpack classification, respectively).
- The regions around the lower body and background gathered few gaze locations for all tasks.

4. Extracting features using task-oriented gaze maps

4.1. Overview of our method

Here, we describe our method for extracting features using task-oriented gaze maps. The regions that obtain high values in the maps appear to contain informative features for participants because these regions are given attention when the participants manually classified the personal attribute in the pedestrian images for each task. We assume that these regions contain discriminative features for the classifiers of personal attributes. Based on this assumption, we extract these features by assigning large weights to the regions that obtain high values in the task-oriented gaze map of each personal attribute. Importantly, we assign weights for both the test and training samples using a gaze map generated in advance. Thus, our method does not require gaze measurements for test samples. After extracting the weighted features, we can apply machine learning techniques. The details of our method are described below.

4.2. Feature extraction algorithm

Given a task-oriented gaze map $\tilde{g}_t(x, y)$, the weight $w_t(x, y)$ for each pixel in a pedestrian image is given by $w_t(x, y) = \mathrm{C}(\tilde{g}_t(x, y))$, where C() is a correction function that emphasizes or weakens values when a moderate level of gaze locations is gathered for each task. We will show the efficacy of the correction function in Section 4.3.1.

A weighted pedestrian image $i_w(x, y)$ is calculated from pedestrian image $i(x, y)$ using $i_w(x, y) = w_t(x, y)i(x, y)$. After applying a weight function, we generate a feature vector for a personal attribute classifier using raster scanning $i_w(x, y)$. Note that we transform the RGB images to CIE L*a*b* color space, weight the L* values, and do not change the a*b* values. We do this because a numerical change in the L* channel corresponds to the same amount of change in human perception.

4.3. Evaluation of the performance of personal attribute classification

4.3.1. Comparison of gaze map-based weight correction functions.

We evaluated the accuracy of personal attribute classifications for various correction functions. We used the task-oriented gaze maps in terms of t_1 to t_4, as shown in Figures 6 (a)–(d). We randomly selected images from the CUHK

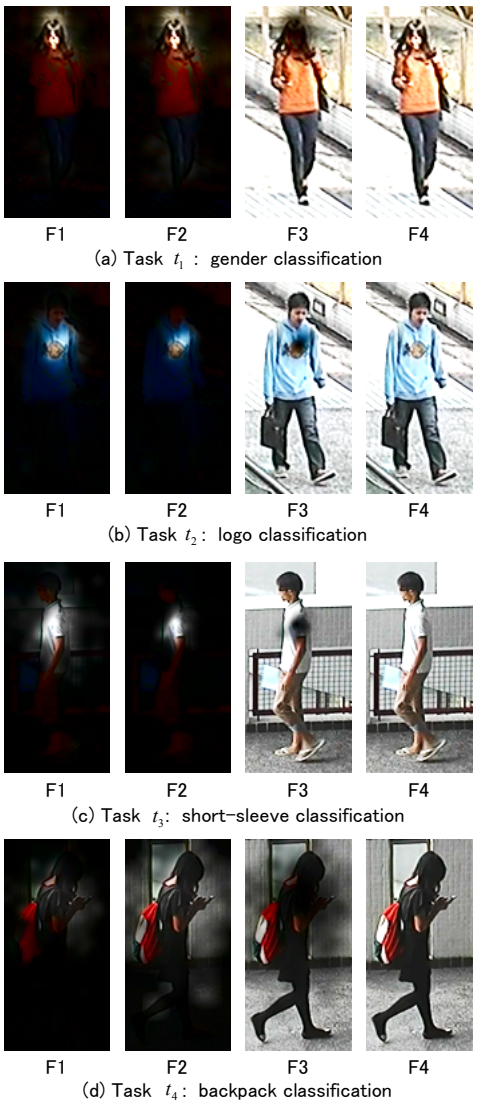

Fig. 7. Examples of test pedestrian images after applying correction functions. We used the task-oriented gaze maps in F1 to F3. The results of our gaze-based feature extraction are F1 and F2.

dataset, which is included in the PETA dataset [8]. There are substantially fewer positive samples than negative samples for each attribute in the CUHK dataset. To avoid the problems associated with imbalanced data, we equalized the number of samples for each attribute label in the test and training sets. We also did not allow the same individual to appear in the pedestrian images in the training and test sets. Attribute labels were used to identify pedestrian images that could be of the same individual. We used 2,720 pedestrian images as training samples and test samples for learning a gender classifier, 560 pedestrian images for learning a logo classifier, 1,200 pedestrian images for learning a short sleeve classifier, and 2,480 pedestrian images for learning a backpack classifier. We applied 10-fold cross-validation for each classification task. We added the eight images for generating a gaze map used in Section 3.3 as training samples. We set an equal ratio of positive and negative labels. Both the training and test samples contained not only frontal poses, but also side and back poses. The metric of the performance of personal attribute classification was the accu-

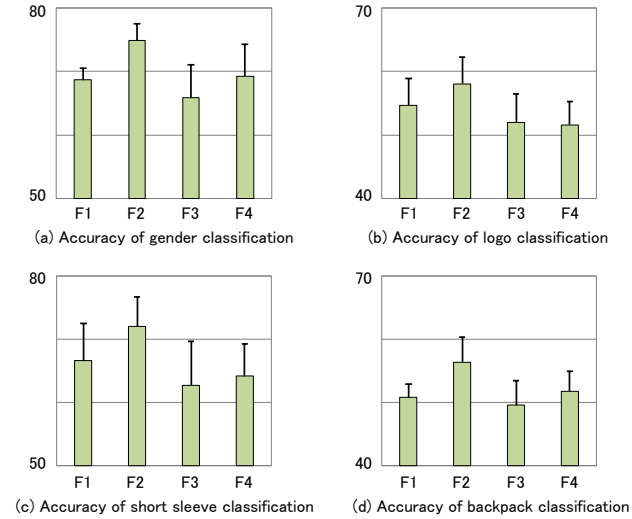

Fig. 8. Comparison of accuracy for different gaze map-based weight correction functions using nearest neighbor.

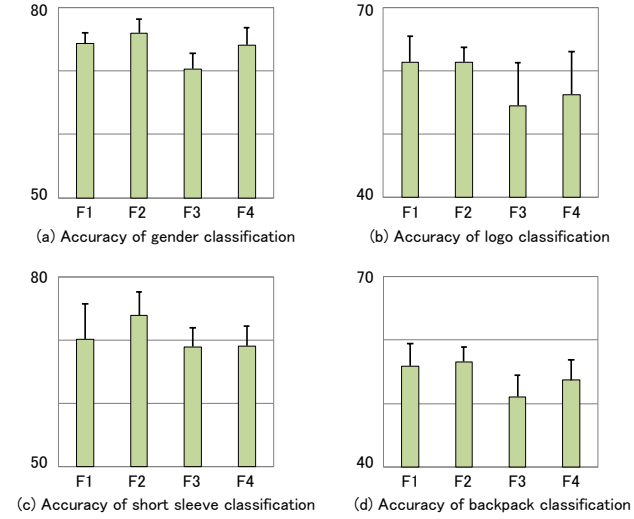

Fig. 9. Comparison of accuracy for different gaze map-based weight correction functions using support vector machine.

racy of classification for each attribute label. We generated feature vectors by raster scanning RGB values with down sampling ($40 \times 80 \times 3$ dimensions) from weighted pedestrian images. We used the k-nearest neighbor technique for all attribute classifiers ($k = 40$) to confirm the baseline performance of personal attribute classification. We compared the accuracies of the following correction functions:

F1: $C(z) = z$,
F2: $C(z) = \min\{1, z^a + b\}$,
F3: $C(z) = 1 - \min\{1, z^a + b\}$, and
F4: $C(z) = 1$.

We determined the parameters of each personal attribute classification task via a grid search using validation sets. These validation sets consisted of the remaining pedestrian images not selected in test and training sets on the CUHK dataset. Parameters $\{a, b\}$ were {0.75, 0.21} for task t_1, {1.75, 0.02} for task t_2, {2.75, 0.04} for task t_3, and {0.25, 0.10} for task t_4.

Figure 7 shows examples of pedestrian images after applying correction functions with the task-oriented gaze maps. Function F1 directly uses values from the maps. If the correction function parameter is $0 < a < 1$, F2 emphasizes the values from the maps, otherwise, it weakens the values from the maps. Function F3 inversely emphasizes the values from the maps. Using F3, we confirmed that the accuracy decreases when we assign small weights to the regions to which the participants paid attention. Function F4 was equal to the original pedestrian images.

Figure 8 shows the average accuracies for each gaze map-based weight correction function for each task of personal attribute classification. The error bars denote the standard deviations of the accuracies. We found that the accuracy of F2 was superior to that of F4 for each personal attribute. Thus, the use of a gaze map appears to increase the performance of attribute classifications. Given that F2 is superior to F1, it appears that this correction function improves accuracy. The inverse weights of F3 decreased the accuracy compared with those of F2. We believe that the regions in which gaze locations were measured

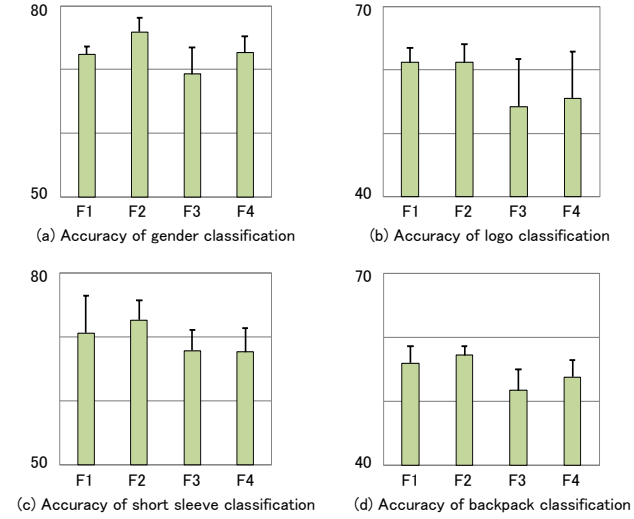

Fig. 10. Comparison of accuracy for different gaze map-based weight correction functions using logistic regression.

from participants for each task may contain discriminative features for the classifiers of personal attributes.

Additionally, we investigated the classification performance of our method by combining it with conventional classifiers. Figure 9 shows the average accuracies using a linear support vector machine classifier (the penalty parameter was $C = 1$) and Figure 10 shows those using a logistic regression classifier (the regularization parameter was 1). We again found that F2 was superior to F1, F3, and F4. We believe that our gaze map-based feature extraction can be used as preprocessing for various conventional classifiers to improve the performance of personal attribute classification.

4.3.2. Combining gaze maps with existing classifiers

We evaluated the performance of attribute classification by combining our gaze-based feature extraction technique with representative classifiers. We exploited deep learning and metric learning techniques as representative classifiers because

Table 1. Accuracy (%) of personal attribute classification by combining the task-oriented gaze map with existing classifiers.

Classification task	Task-oriented gaze map	Accuracy CNN	Accuracy LMNN
Gender (t_1)	with	**79.6 ± 2.2**	**78.5 ± 1.1**
	without	75.3 ± 3.1	76.0 ± 2.7
Logo (t_2)	with	**60.0 ± 3.5**	**56.1 ± 5.3**
	without	57.9 ± 4.0	52.5 ± 4.9
Short sleeve (t_3)	with	**74.0 ± 3.2**	**71.9 ± 3.1**
	without	66.9 ± 4.2	68.1 ± 4.2
Backpack (t_4)	with	**56.9 ± 4.0**	**57.9 ± 2.6**
	without	53.5 ± 3.8	54.7 ± 2.4

other state-of-the-art methods of person attribute classification used CNNs [30, 25, 29, 22] and metric learning [21, 41]. We used the following classifiers: a CNN with a layer architecture of *Mini*-CNN, as described in [1], and large margin nearest neighbor (LMNN) classifier [34], which is a metric learning technique. To avoid overfitting due to the small number of training samples, we used a small network with few convolutional layers. We used the training and test samples described in Section 4.3.1 for learning each classifier. We applied 10-fold cross-validation for each classification task. Table 1 shows the averages and standard deviations for each attribute classification with and without the task-oriented gaze maps. The observed significant improvement demonstrates the efficacy of our gaze-based feature extraction method for personal attribute classification.

4.3.3. Comparison with a method using saliency maps

We compared the accuracy of our method with that of a method that exploits saliency maps. We generated saliency maps from each pedestrian image using existing methods proposed by Zhang et al. [39] and Zhu et al. [42]. We then used the saliency map instead of the task-oriented gaze map in the classification task. We assigned test and training samples large weights in regions with high saliency before applying CNN. Figure 11 shows examples of the saliency maps. Note that we scaled the saliency maps to normalizing the range of the intensities to [0,1]. We evaluated the performance using the same experimental conditions as in Section 4.3.2. The accuracy of the existing method using Zhang et al.'s saliency map is 66.9 ± 2.5%, 56.4 ± 4.4%, 68.2 ± 2.3%, and 53.4 ± 2.8% for tasks t_1, t_2, t_3, and t_4, respectively, and the accuracy of the existing method using Zhu et al.'s saliency map is 66.8 ± 2.8%, 57.2 ± 3.6%, 65.4 ± 3.3%, and 53.9 ± 2.4%, respectively. In contrast, the accuracy of our method is 79.6 ± 2.2%, 60.0 ± 3.5%, 74.0 ± 3.2%, and 56.9 ± 4.0%, respectively. Thus, for personal attribute classification, our method outperforms some methods using saliency maps.

4.3.4. Comparison with methods using regions of body parts

We evaluated the accuracy of a method using head–shoulder or torso regions introduced in [35] instead of the task-oriented gaze maps. For instance, Li et al. [20] reported improved gender classification performance using the head–shoulder region.

Fig. 11. Examples of saliency maps. (a) Test pedestrian images. (b) and (c) Generated saliency maps.

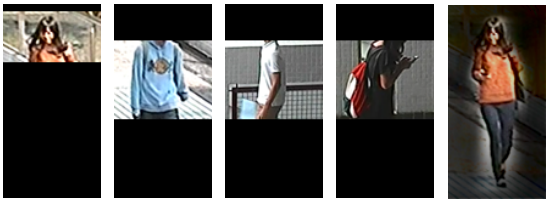

Fig. 12. Examples of head–shoulder or torso regions for comparison with our gaze-based feature extraction.

We used the head–shoulder region for task t_1 and the torso region for tasks t_2 to t_4. Figures 12 (a)–(d) show examples of head–shoulder and torso regions. In addition, we evaluated the accuracy of a method using whole body regions generated from the average intensities of the test and training pedestrian images. Figure 12 (e) shows an example of a whole body region. We used the same experimental conditions as in Section 4.3.3.

The accuracy of the existing method using the head–shoulder or torso regions is 76.0 ± 1.9%, 59.9 ± 3.9%, 64.7 ± 3.2%, and 52.8 ± 3.9% for tasks t_1, t_2, t_3, and t_4, respectively, and the accuracy of the existing method using the whole body region is 77.1 ± 2.0%, 56.2 ± 4.1%, 70.2 ± 3.2%, and 56.6 ± 3.3%, respectively. In contrast, the accuracy of our method is 79.6 ± 2.2%, 60.0 ± 3.5%, 74.0 ± 3.2%, and 56.9 ± 4.0%, respectively. The methods using head–shoulder, torso, or whole body regions performed worse than our method. We believe that not only does the task-oriented gaze map ignore lower-body parts and background regions, but that it also contains meaningful cues for classifying the personal attributes of individuals in pedestrian images.

5. Conclusions

We hypothesized that gaze locations measured from observers performing a classification task contain informative features and help to extract discriminative features for classifiers of personal attributes. We demonstrated that the measured gaze locations tended to concentrate on specific regions of the human body according to the manual personal attribute classification task. We represented the informative region as a task-oriented gaze map for each personal attribute classifier. Owing to the efficacy of the task-oriented gaze maps for feature extraction, our personal attribute classification method was more accurate than representative existing classifiers. As part of our future work, we intend to evaluate the classification performance of

this approach with various pedestrian image datasets and generate gaze maps for various personal attributes such as the physical characteristics described in [23]. We also intend to develop a method for inferring the ambiguity of attribute labels [24] using gaze maps. We would like to expand our research to investigate whether or not there are differences among the nationality and ethnicity of the participants when measuring gaze maps.

Acknowledgment This work was partially supported by JSPS KAKENHI Grant No. JP17K00238 and MIC SCOPE Grant No. 172308003.

References

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

[2] Bindemann, M., 2010. Scene and screen center bias early eye movements in scene viewing. Vision Research 50, 2577 – 2587.

[3] Bruce, V., Burton, A., Hanna, E., Healey, P., Mason, O., Coombes, A., Fright, R., Linney, A., 1993. Sex discrimination: how do we tell the difference between male and female faces? Perception 22, 131–152.

[4] Burton, A., Bruce, V., Dench, N., 1993. What's the difference between men and women? evidence from facial measurement. Perception 22, 153–176.

[5] Buswell, G.T., 1935. How people look at pictures: A study of the psychology of perception of art. University of Chicago Press.

[6] Dantcheva, A., Elia, P., Ross, A., 2016. What else does your biometric data reveal? a survey on soft biometrics. IEEE Transactions on Information Forensics and Security 11, 441–467.

[7] Dantcheva, A., Velardo, C., Dfangelo, A., Dugelay, J., 2011. Bag of soft biometrics for person identification. Multimedia Tools and Applications 51, 739–777.

[8] Deng, Y., Luo, P., Loy, C., Tang, X., 2014. Pedestrian attribute recognition at far distance, in: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 789–792.

[9] Dollar, P., Wojek, C., Schiele, B., Perona, P., 2012. Pedestrian detection: An evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 743–761.

[10] Fairchild, M., 2013. Color Appearance Models. 3rd ed., WILEY.

[11] Fathi, A., Li, Y., Rehg, J., 2012. Learning to recognize daily actions using gaze, in: Proceedings of the 12th European Conference on Computer Vision, pp. 314–327.

[12] Gao, D., Vasconcelos, N., 2004. Discriminant saliency for visual recognition from cluttered scenes, in: Proceedings of Neural Information Processing Systems, pp. 481–48.

[13] Hayhoe, M., Ballard, D., 2005. Eye movements in natural behavior. Trends in Cognitive Sciences 9, 188–194.

[14] Hayhoe, M.M., Shrivastava, A., R., M., B., P.J., 2003. Visual memory and motor planning in a natural task. Journal of Vision 3, 49–63.

[15] Hsiao, J., Cottrell, G., 2008. Two fixations suffice in face recognition. Psychological Science 19, 998–1006.

[16] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., Murphy, K., 2017. Speed/accuracy trade-offs for modern convolutional object detectors, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7310–7319.

[17] Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 1254–1259.

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

[19] Land, M., Mennie, N., J., R., 1999. The roles of vision and eye movements in the control of activities of daily living. Perception 28, 1311–1328.

[20] Li, M., Bao, S., Dong, W., Wang, Y., Su, Z., 2013. Head-shoulder based gender recognition, in: Proceedings of IEEE International Conference on Image Processing, pp. 2753–2756.

[21] Lu, J., Wang, G., Moulin, P., 2014. Human identity and gender recognition from gait sequences with arbitrary walking directions. IEEE Transactions on Information Forensics and Security 9, 51–61.

[22] Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., Feris, R., 2017. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1131–1140.

[23] MacLeod, M.D., Frowley, J.N., Shepherd, J.W., 1994. Whole body information: Its relevance to eyewitnesses, in: Adult eyewitness testimony: Current trends and developments, pp. 125–143.

[24] Martinho-Corbishley, D., Nixon, M.S., Carter, J.N., 2016. On categorising gender in surveillance imagery, in: Proceedings of IEEE 8th International Conference on Biometrics Theory, Applications and Systems, pp. 1–6.

[25] Matsukawa, T., Suzuki, E., 2016. Person re-identification using cnn features learned from combination of attributes, in: Pattern Recognition 23rd International Conference on, pp. 2428–2433.

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

[27] Nixon, M.S., Correia, P.L., Nasrollahi, K., Moeslund, T.B., Hadid, A., Tistarelli, M., 2015. On soft biometrics. Pattern Recognition Letters 68, 218 – 230.

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

[29] Schumann, A., Stiefelhagen, R., 2017. Person re-identification by deep learning attribute-complementary information, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1435–1443.

[30] Sudowe, P., Spitzer, H., Leibe, B., 2015. Person attribute recognition with a jointly-trained holistic cnn model, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 87–95.

[31] Sugano, Y., Ozaki, Y., Kasai, H., Ogaki, K., Sato, Y., 2014. Image preference estimation with a data-driven approach: A comparative study between gaze and image features. Eye Movement Research 7, 862–875.

[32] Tome, P., Fierrez, J., Vera-Rodriguez, R., Nixon, M.S., 2014. Soft biometrics and their application in person recognition at a distance. IEEE Transactions on Information Forensics and Security 9, 464–475.

[33] Walther, D., Itti, L., Riesenhuber, M., Poggio, T., Koch, C., 2002. Attentional selection for object recognition – a gentle way, in: Proceedings of the Second International Workshop on Biologically Motivated Computer Vision, pp. 472–479.

[34] Weinberger, K., Saul, L., 2009. Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research 10, 207–244.

[35] Wu, B., Nevatia, R., 2005. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors, in: Proceedings of Tenth IEEE International Conference on Computer Vision, pp. 90–97.

[36] Xu, J., Mukherjee, L., Li, Y., Warner, J., Rehg, J., Singh, V., 2015a. Gaze-enabled egocentric video summarization via constrained submodular maximization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2235–2244.

[37] Xu, M., Ren, Y., Wang, Z., 2015b. Learning to predict saliency on face images, in: Proceedings of IEEE International Conference on Computer Vision, pp. 3907–3915.

[38] Yarbus, A., 1967. Eye movements during perception of complex objects. Springer.

[39] Zhang, J., Sclaroff, S., Lin, X., Shen, X., Price, B., Mech, R., 2015. Minimum barrier salient object detection at 80 fps, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1404–1412.

[40] Zhu, J.Y., Wu, J., Xu, Y., Chang, E., Tu, Z., 2015. Unsupervised object class discovery via saliency-guided multiple class learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 37, 862–875.

[41] Zhu, P., Zhang, L., Zuo, W., Zhang, D., 2013. From point to set: Extend the learning of distance metrics, in: Proceedings of IEEE International Conference on Computer Vision, pp. 2664–2671.

[42] Zhu, W., Liang, S., Wei, Y., sun, J., 2014. Saliency optimization from robust background detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2814–2821.