# FEATURE EXTRACTION USING GAZE OF PARTICIPANTS FOR CLASSIFYING GENDER OF PEDESTRIANS IN IMAGES

*Riku Matsumoto, Hiroki Yoshimura, Masashi Nishiyama, and Yoshio Iwai*

Department of Information and Electronics, Graduate School of Engineering, Tottori University

## ABSTRACT

Human participants look at informative regions when attempting to identify the gender of a pedestrian in images. In our preliminary experiment, participants mainly looked at the head and chest regions when classifying gender in these images. Thus, we hypothesized that the regions in which participants gaze locations were clustered would contain discriminative features for a gender classifier. In this paper, we discuss how to reveal and use gaze locations for the gender classification of pedestrian images. Our method acquired the distribution of gaze locations from various participants while they manually classified gender. We termed this distribution a gaze map. To extract discriminative features, we assigned large weights to regions with clusters of gaze locations in the gaze map. Our experiments show that this gaze-based feature extraction method significantly improved the performance of gender classification when combined with either a deep learning or a metric learning classifier.

***Index Terms***— Gender, Gaze, Feature

## 1. INTRODUCTION

Gender classification of individuals captured on video or still images has many potential applications, such as video surveillance and consumer behavior analysis. Researchers have proposed several methods for classifying gender in images of pedestrians; for example, techniques involving metric learning [1] and deep learning [2]. Using existing methods, it is possible to extract discriminative features for gender classification and to 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.

Humans have the visual capability to extract features from an individual and identify them as male or female. For instance, people correctly classify gender from facial images [3, 4]. We believe that people are also able to correctly distinguish males and females in pedestrian images. It may be possible to reproduce this visual capability 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 visual capability have been proposed [5, 6, 7]. The existing methods involve saliency maps of object images with representations of the regions that draw visual attention. Walther et al. [5] combined a recognition algorithm with a saliency map generated from low-level features of gradients of color and intensity. Further, researchers have developed techniques [6, 7] that use object labels of images in addition to low-level features of objects to generate saliency maps. However, the use of low-level features and object labels does not sufficiently represent human visual capability.

Recently, an increasing number of pattern recognition studies, specifically those attempting to mimic human visual capability, have measured gaze locations from human participants [8, 9, 10]. Xu et al. [8] 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 tasks involving action recognition or image preference estimation. Fathi et al. [9] classified actions by simultaneously inferring regions where gaze locations were gathered via an egocentric camera. Additionally, Sugano et al. [10] estimated more highly preferable images using gaze locations and low-level features. As just described, gaze locations measured from participants have great potential for the collection of informative features during various recognition tasks.

In this paper, we sought to demonstrate that gaze locations play an important role in the gender classification of pedestrian images. If we measured gaze locations for both test and training samples and compared the data between these, as in [10], we expect that we would find a significant improvement in the accuracy of gender classification. However, we cannot measure gaze locations for test samples in real-world applications. Thus, it is necessary to develop a method for representing an alternative to the gaze locations of pedestrian images. To this end, we generated a gaze map from the distribution of gaze locations recorded while participants viewed images, some of which were selected from training samples, and completed a gender classification task. The high values in a gaze map correspond to regions that are frequently viewed

by participants. We assumed that these regions contained discriminative features for gender classification because they appeared to be useful during the gender classification task. When extracting features from both the test and training samples, larger weights were given to the regions of the pedestrian images that corresponded to the attended regions of the gaze map. The experimental results indicated that our method improved the accuracy when using representative classifiers with a small number of training samples.

## 2. GENERATING A GAZE MAP

### 2.1. Gaze locations in gender classification

Here, we consider the regions of pedestrian images that are frequently attended to by participants when manually classifying gender. For instance, Hsiao et al. [11] found that participants looked at a region around the nose when identifying individuals from a facial image. We believe that the human face also plays an important role in gender classification. However, a pedestrian image contains not only a face but also a body. Thus, we attempted to discern the regions of pedestrian images that tended to collect gaze locations from participants while they completed a gender classification task. Note that we assumed that the alignment of the pedestrian images had already been completed using a pedestrian detection technique. The details of our method are described below.

### 2.2. Generation algorithm

To generate a gaze map, we used a gaze tracker to acquire gaze locations while displaying a pedestrian image on a screen. We prepared $P$ participants, and $N$ pedestrian images. Given a gaze location $(x_t, y_t)$ in a certain time $t$, the gaze map $g_{p,n,t}(x, y)$ was 1 when $x = x_t, y = y_t$ otherwise 0, where $p$ is a participant, and $n$ is a pedestrian image. Note that the participant not only looked at point $(x_t, y_t)$ on each pedestrian image, but also the region surrounding the point. Thus, we applied a Gaussian kernel to the measured gaze map $g_{p,n,t}$. To determine the size $k$ of the Gaussian kernel, we used the following equation $k = \frac{2dh}{l} \tan \frac{\theta}{2}$, where $d$ is the distance between the screen and the participant, $\theta$ 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 assumed that each pixel on the screen was a square. We aggregated each $g_{p,n,t}(x, y)$ to $g_{p,n}(x, y)$ to represent the distribution of gaze locations in a certain pedestrian image as $g_{p,n}(x, y) = \sum_{t=1}^{T_{p,n}} k(u, v) * g_{p,n,t}(x, y)$, where $T_{p,n}$ is the time taken to classify gender by a participant, $*$ is the convolution operator, and $k(u, v)$ is a Gaussian kernel of size $k \times k$. We applied L1-norm normalization as $\|g_{p,n}(x, y)\| = 1$ because $T_{p,n}$ is different for each measurement. Furthermore, we aggregated $g_{p,n}(x, y)$ to a single gaze map for all participants, and all pedestrian images. An aggregated gaze map $g(x, y)$ representing the distribution of gaze locations is represented as $g(x, y) = \sum_{p=1}^{P} \sum_{n=1}^{N} g_{p,n}(x, y)$. Note that we applied a scaling technique to aggregated gaze maps as $g(x, y)/\max(g(x, y))$.

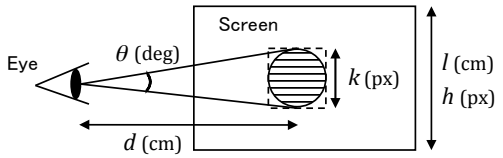


**Fig. 1**. Parameters used to determine kernel size.

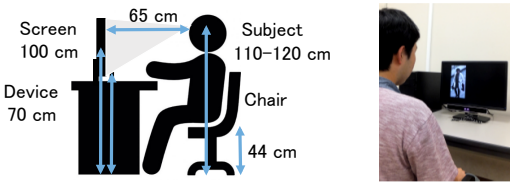


**Fig. 2**. Setting for capturing gaze locations.

### 2.3. Evaluating a gaze map

To evaluate the constructed gaze maps, we captured gaze locations for $P = 14$ participants (average age $22.4 \pm 0.8$, 7 males) using a standing eye tracker (GP3 Eye Tracker, sampling rate 60 Hz). We used a 24-inch display (size $51.7 \times 32.3$ mm, $1920 \times 1200$ 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 used 4563 pedestrian images from the CUHK dataset included in the PETA dataset [12] with gender labels. We randomly selected $N = 30$ images of pedestrians in frontal, sideways, and back poses (15 male and 15 female images). Note that we used the same pedestrian images between the participants. We enlarged the pedestrian images from $80 \times 160$ pixels to $480 \times 960$ pixels to display the images on the screen. To avoid a center bias 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 $\pm 720$ pixels vertically and $\pm 120$ pixels horizontally.

We asked participants to complete the gender classification task and measured gaze locations according to the following procedures (P1 to P3). **P1**: We displayed a flat gray image on the screen for 1 second. **P2**: We displayed a pedestrian image on the screen for 2 seconds. Prior to the trial, the participants had been instructed to keep looking at the image. **P3**: We displayed a flat black image on the screen for 2 seconds and the participant verbally reported the inferred gender of the pedestrian. In our preliminary experiment, we observed

**Fig. 3**. Gaze maps of each pedestrian image. (Left: a pedestrian image, Right: a measured gaze map)

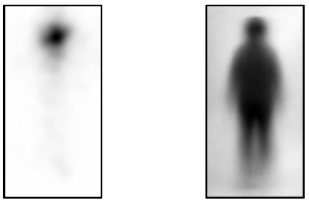


**Fig. 4**. (a) Aggregated gaze map for gender classification and (b) Average intensities of pedestrian images.

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 whether the pedestrian was male or female. To determine $T_{p,n}$, we set the start time as the point at which the gaze first stopped on the pedestrian image for more than 150 msec, and the end time as the point at which the pedestrian image disappeared. In this scenario, the average $T_{p,n}$ between the start and end times was $1.66 \pm 0.24$ seconds. The accuracy of gender classification by the participants was $100\%$. We set $\theta = 2^\circ$ by considering the range of the fovea, which is about two degrees, as described in [13]. We determined a kernel size of $k = 81$ against the enlarged pedestrian images ($480 \times 960$ pixels). We determined that the size of the gaze map was downsized by $80 \times 160$ after adjustment from the original size of the pedestrian images.

Figure 3 shows examples of gaze maps $g_{p,n}(x, y)$ for each pedestrian image. The gaze maps on the right side were acquired when the pedestrian images on the left side were shown. The dark regions in the gaze maps represent the gathered gaze locations from the participants. The minimum intensities in Figure 3 represent the maximum values of all $g_{p,n}(x, y)$. We observed that participants frequently concentrated their gaze on the head regions and sometimes looked at chest regions.

Figure 4 (a) shows an aggregated gaze map $g(x, y)$. To consider the properties of the gaze maps, we checked how the gaze maps were aligned with the pedestrian images. Figure 4 (b) shows the average gray-scaled intensities calculated from 4563 pedestrian images. We can see a silhouette of a whole body, indicating that the pedestrian images were well aligned. From the results in Figure 4 (a) and (b), we infer that the aggregated gaze map to include the region around the head gathered a large number of gaze locations, the region around the chest gathered a moderate amount of gaze locations, and the region around the whole body gathered few gaze locations.

## 3. EXTRACTING FEATURES USING A GAZE MAP

### 3.1. Overview of our method

Here, we describe our method for extracting features using a gaze map. The regions that obtained high values in the gaze maps appeared to contain informative features for participants because these regions were attended to while the participants manually inferred the gender of the pedestrians. We assumed that these regions also contained discriminative features for a gender classifier. Based on this assumption, we aimed to extract these features by giving large weights to the regions that obtained high values in the gaze map for each pedestrian image. Importantly, in our method, we gave weights for both the test and training samples using a gaze map that was generated in advance. Thus, our method does not require gaze measurements for test samples. After extracting weighted features, we can apply machine learning techniques. The details of our method are described below.

### 3.2. Feature extraction algorithm

Given gaze map $g(x, y)$, weight $w(x, y)$ for each pixel in a pedestrian image is given by $w(x, y) = \mathrm{C}(g(x, y))$, where $\mathrm{C}()$ is a correction function that emphasizes values when gaze locations are somewhat gathered. We will show the efficacy of the correction function in Section 3.3.1.

A weighted pedestrian image $i_w(x, y)$ is determined from a pedestrian image $i(x, y)$ as $i_w(x, y) = w(x, y)i(x, y)$. After applying a weight function, we generated a feature vector for a gender classifier using raster scanning $i_w(x, y)$. Note that if the pedestrian images were in color, we transformed RGB color space to CIE L*a*b* color space, gave a weight to L* values only, and did not change a*b* values.

### 3.3. Evaluations of gender classification

*3.3.1. Comparison of correction functions for weights given using a gaze map.*

We evaluated the accuracy of gender classification by changing the correction functions. We used a gaze map, as shown in Figure 4 (a). We used 2,000 pedestrian images as training samples for learning a gender classifier (1,970 images randomly selected from the CUHK dataset included in the PETA dataset [12] and 30 images for generating a gaze map used in Section 2.3). For test samples, we used 400 pedestrian images randomly selected from the CUHK dataset, except for the training samples. Note that we eliminated images that had the same attribute labels between training and test to avoid including the same individual. We generated five test sets by


**Fig. 5**. Examples of pedestrian images after applying correction functions. We used the gaze map in (a) to (d) and the average intensities in (e)

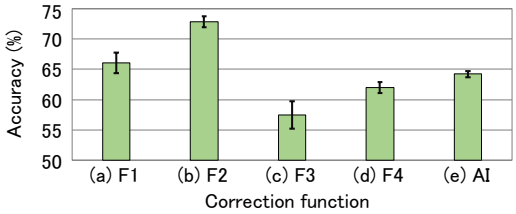


**Fig. 6**. Comparison of accuracy for different correction functions with weights given using a gaze map or average intensities.

repeating this procedure to avoid the bias of random selection. We used an equal ratio of male and female pedestrians. Both the training and test samples contained not only frontal poses, but also sideways and back poses. The metric of the performance of gender classification was the accuracy of classified gender labels. We generated feature vectors by the raster scanning of RGB values with down sampling ($40 \times 80 \times 3$ dimensions) from weighted pedestrian images. We used a $k$-nearest neighbor technique for a gender classifier ($k = 20$). We compared the accuracies of the following correction functions, **F1**: $\mathrm{C}(z) = z$, **F2**: $\min\{1, z^a + b\}$, **F3**: $\mathrm{C}(z) = 1 - \min\{1, z^a + b\}$, and **F4**: $\mathrm{C}(z) = 1$. We determined the parameters $a = 0.7, b = 0.1$ via a grid search technique using the test sets. Figure 5 (a) to (d) shows examples of pedestrian images after applying correction functions with the gaze map. F1 directly used values from the gaze map, and so consists of primarily head and chest regions. F2 emphasized the values from the gaze map, and so features the upper body regions. F3 inversely emphasized the values from the gaze map, such that the head regions disappeared. Using F3, we confirmed that the accuracy would decrease when we gave small weights to the regions to which the participants attended. F4 was equal to the original pedestrian images.

Figure 6 (a) to (d) shows the average accuracies for each correction function with the gaze map. The error bars denote standard deviations of the accuracies evaluated on the five test sets. We found that the accuracies of F1 and F2 were superior to that of F4. Thus, the use of a gaze map appears to increases the performance of gender classification. Given that F2 is superior to F1, it appears that correction function improves accuracy. The inversed weights of F4 decreased the accuracy compared with those of F2. We believe that the regions in which gaze locations were measured from participants may contain discriminative features for a gender classifier.

We also evaluated the accuracy when using the average intensities of the pedestrian images, as shown in Figure 4 (b), instead of using the gaze map. We applied a scaling technique to normalize the range of the intensities to [0,1]. We used the F1 correction function. Figure 5 (e) shows examples of pedestrian images after applying the correction function with the average intensities (AI). We obtained lower performance in Figure 6 (e) than (a) and (b). We believe that not only does the gaze map ignore background regions, but that it also contains meaningful cues to classify the gender of the pedestrian images.

### *3.3.2. Combining a gaze map with existing classifiers*

We evaluated the performance of gender classification by combining our gaze-based feature extraction technique with representative classifiers. We used 2,000 training samples, as described in Section 3.3.1. The test samples were also the same. We used the following classifiers: large margin nearest neighbor (LMNN) [14], which is a metric learning technique (neighbors parameter in the training process was 20); and a convolutional neural network (CNN) [15], which is a deep learning technique (the layer architecture was *Mini*-CNN described in [2]).

Table 1 shows the averages and the standard deviations of the accuracies of gender classification. The observed significant improvement in gender classification performance demonstrates the efficacy of our gaze-based feature extraction method.

**Table 1**. Accuracy of gender classification by combining a gaze map with existing classifiers

| Classifier | Gaze map | Accuracy (%) |
|---|---|---|
| CNN | with | $\mathbf{75.2 \pm 1.4}$ |
| | without | $69.7 \pm 1.1$ |
| LMNN | with | $\mathbf{72.1 \pm 1.0}$ |
| | without | $68.0 \pm 1.2$ |

## 4. CONCLUSIONS

We hypothesized that gaze locations measured from participants would contain informative features and help to extract discriminative features for a gender classifier. Owing to the efficacy of our gaze-based feature extraction approach, our method was highly accurate for gender classification compared with representative existing classifiers. As part of our future work, we intend to evaluate the classification performance with various datasets of pedestrian images.

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

## 5. REFERENCES


[1] J. Lu, G. Wang, and P. Moulin, "Human identity and gender recognition from gait sequences with arbitrary walking directions," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 1, pp. 51–61, 2014.

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

[3] V. Bruce, A.M. Burton, E. Hanna, P. Healey, O. Mason, A. Coombes, R. Fright, and A. Linney, "Sex discrimination: how do we tell the difference between male and female faces?," *Perception*, vol. 22, no. 2, pp. 131–152, 1993.

[4] A.M. Burton, V. Bruce, and N. Dench, "What's the difference between men and women? evidence from facial measurement," *Perception*, vol. 22, no. 2, pp. 153–176, 1993.

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

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

[7] J. Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 4, pp. 862–875, 2015.

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

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

[10] 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.

[11] J.H. Hsiao and G. Cottrell, "Two fixations suffice in face recognition," *Psychological Science*, vol. 19, no. 10, pp. 998–1006, 2008.

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

[13] M.D. Fairchild, *Color Appearance Models*, WILEY, 3rd edition, 2013.

[14] K.Q. Weinberger and L.K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, June 2009.

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.