We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,300

Open access books available

172,000

International authors and editors

190M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Feature Extraction Using Observer Gaze Distributions for Gender Recognition

Masashi Nishiyama

Abstract

We determine and use the gaze distribution of observers viewing images of subjects for gender recognition. In general, people look at informative regions when determining the gender of subjects in images. Based on this observation, we hypothesize that the regions corresponding to the concentration of the observer gaze distributions contain discriminative features for gender recognition. We generate the gaze distribution from observers while they perform the task of manually recognizing gender from subject images. Next, our gaze-guided feature extraction assigns high weights to the regions corresponding to clusters in the gaze distribution, thereby selecting discriminative features. Experimental results show that the observers mainly focused on the head region, not the entire body. Furthermore, we demonstrate that the gaze-guided feature extraction significantly improves the accuracy of gender recognition.

Keywords: gaze distribution, region of interest, feature extraction, pedestrian image, gender recognition

1. Introduction

Gender recognition, which is of interest in the field of soft-biometrics, is part of the collection of statistical data about people in public spaces. Furthermore, gender recognition has many potential applications, such as video surveillance and consumer behavior analysis. Often, gender recognition experiments are conducted on pedestrians captured on video. Researchers have proposed several methods for automatically recognizing gender in pedestrian images; many of these techniques use convolutional neural networks (CNNs) [1]. The existing methods can extract discriminative features for gender recognition and obtain highly accurate results 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, deep learning methods typically require these large training sets to maintain suitable recognition performance.

People quickly and correctly recognize gender; thus, we believe that people effectively extract visual features from subjects in images. For instance, people correctly

recognize gender from facial images [2, 3]. It may be possible to reproduce human visual abilities in a computer algorithm with a small number of training samples and achieve a recognition performance equivalent to that of humans. Existing methods [4, 5] have been proposed to mimic human visual abilities for object recognition tasks. These methods used a saliency map generated from low-level features [6–8]. However, these saliency maps does not sufficiently represent human visual abilities because they are not directly measured from human observers. We thus consider that the existing methods disregard 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 distribution of observers [9–12]. This gaze distribution has great potential in the collection of informative features for various recognition tasks. Several techniques [13, 14] have demonstrated that the gaze distribution facilitates the extraction of informative features. Sattar et al. [13] applied the gaze distribution to analyze fashion in images. Murrugarra-Llerena and Kovashka [14] applied the gaze distribution for attribute prediction in facial images. However, the existing methods using observer gaze distribution do not study gender recognition from pedestrian images. We consider that the region of interest measured from observers’ gaze is also effective for gender recognition.

Here, we conduct a gaze measurement experiment for observers performing a gender recognition task on images of subjects. We investigate if the gaze distribution measured from the observers facilitates gender recognition. **Figure 1** shows the overview of our gaze-guided feature extraction. We generate a task-oriented gaze distribution from the gaze locations recorded while observers manually determined the genders of subjects in images. High values in a task-oriented gaze distribution correspond to regions that observers frequently view. We assume that these regions contain discriminative features for gender recognition because they appear to be useful when

Figure 1. Overview of our gaze-guided feature extraction. We consider that the regions gathering the gaze distribution contain discriminative features for gender recognition because they appear to be useful when the observers are tackling the gender recognition task.

the observers are determining the subject gender. When extracting features to train the gender classifier, larger weights are assigned to the regions of the pedestrian images corresponding to the attention regions of the task-oriented gaze distribution. The experimental results indicate that our gaze-guided feature extraction improves the gender recognition accuracy when using a CNN technique with a small number of training samples.

2. Generating a task-oriented gaze distribution for gender recognition

2.1 Observer gaze distribution in gender recognition

We discuss which body regions of subjects in images are frequently viewed for gender recognition by observers. With respect to the analytical study of facial images, Hsiao et al. [15] reported that people looked at the nose region when they recognized others. We consider that the human face is a key factor in gender recognition. Furthermore, we consider that the entire body, including the chest, waist, and legs, is also helpful. Thus, we aim to reveal the body regions that tend to collect the observer gaze distribution during a gender recognition task. Note that we assume that the pedestrian images have been pre-aligned using pedestrian detection techniques. The details of our method are described below.

2.2 Generating a task-oriented gaze distribution

To generate a task-oriented gaze distribution, we use a gaze tracker to acquire gaze locations while the observer views a pedestrian image on a screen. We briefly describe our method in **Figure 2**. We work with P participating observers and N pedestrian images. Given a gaze location (x_f, y_f) in a certain frame f , the gaze distribution $g_{p,nf}(x, y)$ is computed as

$$g_{p,nf}(x, y) = \begin{cases} 1 & (x = x_f, y = y_f), \\ 0 & (\text{otherwise}), \end{cases} \quad (1)$$

where p is an observer, and n is a pedestrian image. Note that the observer 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 distribution $g_{p,nf}(x, y)$. **Figure 3** illustrates the parameters used to determine the size k of the Gaussian kernel. We compute the following equation:

Figure 2.
Overview of our method for generating a gaze distribution $\tilde{g}(x, y)$.

Figure 3. Parameters used to determine the kernel size for generating the gaze distribution.

$$k = \frac{2dh}{l} \tan \frac{\theta}{2}, \quad (2)$$

where θ represents the angle of the region surrounding (x_f, y_f) , l represents the screen's vertical length, h represents the screen's vertical resolution, and d represents the distance from the participant to the screen. We aggregate each $g_{p,nf}(x, y)$ to $g_{p,n}(x, y)$ to represent the gaze distribution in a particular pedestrian image as

$$g_{p,n}(x, y) = \sum_{f=1}^{F_{p,n}} k(u, v) * g_{p,nf}(x, y), \quad (3)$$

where $F_{p,n}$ is the time taken by an observer to recognize the gender of the subject in the image. Function $k(u, v)$ represents a Gaussian kernel of size $k \times k$ and operator $*$ represents the convolution. Our method performs L1-norm normalization as $\|g_{p,n}(x, y)\| = 1$. We aggregate $g_{p,n}(x, y)$ into a single gaze distribution across all observers and all pedestrian images. The aggregated gaze distribution $g(x, y)$ is represented as

$$g(x, y) = \sum_{p=1}^P \sum_{n=1}^N g_{p,n}(x, y). \quad (4)$$

Note that we apply a scaling technique to the aggregated gaze distributions as follows: $\tilde{g}(x, y) = g(x, y) / \max(g(x, y))$. $\tilde{g}(x, y)$ is the final task-oriented gaze distribution.

3. Experiments to generate a task-oriented gaze distribution

3.1 Setup

We evaluated the task-oriented gaze distributions for gender recognition. We acquired the gaze locations for $P = 14$ participating observers (seven men and seven women, with an average age of 22.6 ± 1.3 years old, Japanese students). We used a display screen (size 53.1×29.9 cm, 1920×1080 pixels). The vertical distance between the screen and the participant was set to 65 cm, as illustrated in **Figure 4**.

Figure 4.
Setup used to acquire the gaze distribution in a gender recognition task.

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 use a standing eye tracker (GP3 Eye Tracker, sampling rate 60 Hz). We asked the participants to perform a gender recognition task to determine if the pedestrian in an image is a man or a woman. We determined which regions of the entire body were viewed by the participants to complete this task.

We used 4563 pedestrian images from the CUHK dataset included in the PETA dataset [16] with gender labels (woman or man). From this dataset, we use the $N = 8$ pedestrian images in **Figure 5** to use in the observer experiment to generate the gaze distribution map. We selected the four pedestrian images at the top of **Figure 5** keeping the ratio of directions (front, back, left, and right) equal. We selected the remaining pedestrian images in **Figure 5** in the same manner. When displaying the stimulus images on the screen, the pedestrian images were enlarged from 80×160 pixels to 480×960 pixels. We simply changed the stimulus images' positions by adding random offsets to avoid a center bias [17, 18].

We acquired the gaze distribution when participants performed the gender recognition task according to the following procedure:

- P1. A gray image is shown on the screen for one second.
- P2. A pedestrian stimulus image is shown on the screen for two seconds.
- P3. A black image is shown on the screen for two seconds, and the participant replied whether the pedestrian was a woman or a man.
- P4. We repeated P1 to P3 until all eight pedestrian images had been displayed 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 complete the gender recognition task. To determine $F_{p,n}$, we set the start time at the point when the gaze first stopped on the pedestrian image for more than 440 ms, and the end time corresponds to the pedestrian image no longer appearing on screen. In this scenario, the average $F_{p,n}$ between the start and end times was 1.56 ± 0.38 s. The participating observers achieved a gender recognition accuracy of 100.0%.

We set $\theta = 3^\circ$ in Eq. (2) by considering the range of the fovea, which is approximately two degrees (as described in [19]), and the error of the eye tracker, which is about one degree (as described in the tracker's specification sheet). We used a kernel size of $k = 125$ for the enlarged pedestrian images (480×960 pixels). The size of the

Figure 5.
Pedestrian images for generating task-oriented gaze distributions during the gender recognition task.

gaze distribution images was downsized by 80×160 , adjusting according to the original size of the pedestrian images. This standardized the size of the test samples and training samples to input into the gender classifier.

3.2 Results

Figure 6 shows examples of the measured gaze distributions $g_{p,n}(x,y)$ for the gender recognition task for a pedestrian image. We show the gaze distribution map from two participants for the pedestrian image shown in **Figure 6(a)**. The dark regions in the gaze distribution maps represent the gaze locations recorded from the participants by the eye tracker. The minimum (black) and maximum (white) intensities in **Figure 6** represent the maximum and minimum values of the measured $g_{p,n}(x,y)$, respectively. We observed that participants frequently concentrated their gaze on the head region to complete the gender recognition task.

Figure 7 shows the overall task-oriented gaze distribution $\tilde{g}(x,y)$ for gender recognition synthesized from all of the participating observers. To study the properties of the task-oriented gaze distribution, we verify how the gaze distributions align with the pedestrian images of **Figure 5**. We see that the region corresponding to the head

Figure 6.
Examples of measured gaze distributions $g_{p,n}(x,y)$ from two participants. (a) Stimulus image of pedestrian. (b) and (c) Gaze distributions measured from each participant viewing the pedestrian image in (a).

Figure 7.
Task-oriented gaze distribution $\tilde{g}(x,y)$ for the gender recognition task.

gathered a large number of gaze locations, while regions around the lower body and background gathered few gaze locations.

4. Feature extraction algorithm using the task-oriented gaze distribution for gender recognition

4.1 Overview of our gaze-guided feature extraction

Here, we describe our method to extract features using the task-oriented gaze distribution for gender recognition. The regions corresponding to high values in the distribution $\tilde{g}(x,y)$ appear to contain informative features because participants focus on these regions to manually recognize gender in the pedestrian images. Thus, we assume that these regions contain discriminative features for the gender classifiers. Based on this assumption, we extract these features by assigning higher weights to the regions corresponding to high values in the task-oriented gaze distribution.

Figure 8 provides an overview of our method. Our methods assign weights using $\tilde{g}(x,y)$ for both the test samples and training samples. Therefore, we do not need to

Figure 8. Overview of our gaze-guided feature extraction using the gaze distribution $\tilde{g}(x, y)$.

acquire gaze distributions on the test samples. Our method extracts the weighted features and applies deep learning and machine learning techniques to obtain the final classification.

4.2 Procedure

Given a gaze distribution $\tilde{g}(x, y)$, our method computes the weight $\tilde{w}(x, y)$ for each pixel as

$$\tilde{w}(x, y) = C(\tilde{g}(x, y)). \quad (5)$$

We use a correction function $C()$ that weakens or emphasizes values according to the density of the gaze distribution.

We calculate a weighted intensity $i_w(x, y)$ from an original intensity $i(x, y)$ as follows:

$$i_w(x, y) = \tilde{w}(x, y)i(x, y). \quad (6)$$

We generate a feature vector for gender recognition using raster scanning $i_w(x, y)$. The RGB images are converted to CIE $L^*a^*b^*$ color space. Note that our method weights the L^* values and does not change the a^*b^* values. We consider only the lightness changes without any color changes because a numerical change in the L^* channel corresponds to the lightness change in human perception.

5. Evaluation of the gender recognition performance using the gaze distribution

5.1 Comparison of weight correction functions for feature extraction

We evaluated the accuracy of gender recognition using various correction functions. We used the gaze distribution $\tilde{g}(x, y)$, as shown in **Figure 7**. We randomly picked up pedestrian images from the CUHK dataset, which is included in the PETA dataset [16]. We equalized the ratio of women and men samples in the test sets and training sets to avoid problems associated with imbalanced data. The same individual did not appear in both training and test samples. We used 2720 pedestrian images as training samples and test samples. We applied 10-fold cross-validation for gender recognition. Both the training and test samples contained not only frontal poses, but also side and back poses. We evaluate the gender recognition performance as the accuracy of the woman or man classification labels. We generated feature vectors by

raster scanning RGB values with down sampling ($40 \times 80 \times 3$ dimensions) from weighted pedestrian images. We used a linear support vector machine classifier (the penalty parameter was $C = 1$) to confirm the baseline performance of gender recognition. For the other classifiers, we show experimental results in Section 5.2. We compared the accuracy 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. $C(z) = 1$.

Figure 9 shows a visualization of the correction functions $C(z)$. We determined the parameters of the gender classifier using a grid search over the validation sets. These validation sets consisted of the remaining pedestrian images not used in the test sets and training sets from the CUHK dataset. Parameters $\{a, b\}$ were set to $\{0.75, 0.21\}$.

Figure 10(a) shows pedestrian images after applying $C(z)$. Function F1 outputs an intensity weighted by the gaze distribution for each pixel. Function F2 emphasizes an intensity around a face using gaze distribution. In contrast, function F3 weakens the intensity. Function F4 directly outputs the intensity of the original pedestrian image.

Figure 9.
Visualization of correction functions $C(z)$.

Figure 10.
Gender recognition accuracy. (a) Examples of test pedestrian images after applying correction functions. F1 and F2 show the results of our gaze-based feature extraction. (b) Comparison of gender recognition accuracy using each gaze-guided weight correction function with a linear support vector machine classifier.

Figure 10(b) shows the gender recognition accuracy of each gaze-guided weight correction function for gender recognition. We confirmed that the accuracy of F1 and F2 was superior to that of F4. Thus, the use of the gaze distribution $\tilde{g}(x, y)$ appears to increase the performance of gender recognition. F2 yields superior performance compared with F1, indicating that this correction function improves gender recognition accuracy. The inverse weights of F3 decreased the accuracy compared with the other correction functions. Thus, we demonstrate that the regions corresponding to observer gaze distribution $\tilde{g}(x, y)$ measured from participants completing a gender recognition task contain discriminative features for the gender classifier.

5.2 Combining our gaze-guided feature extraction with existing classifiers

We investigated the gender recognition performance by combining our gaze-based feature extraction technique with representative classifiers. We used *Mini*-CNN architecture [20], which is a small network with few convolutional layers. We also used a large margin nearest neighbor (LMNN) classifier [21], which is a metric learning technique. The test samples and training samples described in Section 5.1 were used in the evaluation. We applied 10-fold cross-validation. **Table 1** shows the accuracy for gender recognition with and without our gaze-guided feature extraction. Our gaze-based feature extraction method leads to improved gender recognition for both classifiers.

5.3 Evaluation of assigning weights using saliency maps

We evaluated the gender recognition accuracy of a method that uses saliency maps. We used the existing methods of Zhang et al. [7], and Zhu et al. [8] to generate saliency maps. **Figure 11** shows the saliency maps used in the evaluation of gender recognition. We scaled the intensity in the saliency map to fit the intensity range to [0,1]. We performed feature extraction using the saliency map instead of the task-

Condition	Accuracy using CNN	Accuracy using LMNN
With our gaze-guided feature extraction	79.6 ± 2.2	78.5 ± 1.1
Without our gaze-guided feature extraction	75.3 ± 3.1	76.0 ± 2.7

Table 1. Accuracy (%) of gender recognition by combining our gaze-guided feature extraction with existing classifiers.

Figure 11. Examples of saliency maps used in gender recognition. (a) Test pedestrian images. (b), (c) generated saliency maps.

Our gaze distribution	Zhang et al.'s saliency map	Zhu et al.'s saliency map
79.6 ± 2.2%	66.9 ± 2.5%	66.8 ± 2.8%

Table 2.
Gender recognition accuracy (%) using our task-oriented gaze distribution compared with using the existing saliency maps.

oriented gaze distribution $\tilde{g}(x, y)$. Our method assigned the test samples and training samples large weights in regions corresponding to high saliency values before using a CNN classifier. We evaluated the accuracy using the same conditions of Section 5.2. **Table 2** shows the gender recognition accuracy obtained when using our task-oriented gaze distribution compared with the accuracy obtained using the existing saliency map approaches. The results indicate that our gaze-guided feature extraction method outperforms the use of saliency maps for gender recognition.

5.4 Visualization of the regions of focus when using CNNs

We conducted an experiment to visualize the regions of focus in a pedestrian image during gender recognition. To this end, we used gradient-weighted class activation mapping (Grad-CAM) [22]. **Figure 12** shows the visualization results of the regions of focus of the CNN method. In (a), we show the pedestrian test images for gender recognition. In (b), we show the visualization results without our gaze-guided feature extraction. We only used the conventional CNN of the VGG16 model with fine-tuning. In the woman test samples, the model emphasized the leg and waist regions. In the man test samples, the model emphasized the shoulder and head regions. This indicates that the conventional CNN emphasizes various body part regions for gender recognition but in a different manner than used by the participating observers in the experiments of Section 3.2. In (c), we show the visualization results using our gaze distribution maps for gender recognition. We used our gaze-guided feature extraction with *Mini*-CNN, as described in Section 5.2. We confirmed that our method mainly emphasizes the head region, mimicking the human observers' gaze behavior. In particular, we consider that our method recognizes gender by focusing on the hairstyle of the subject in an image because it emphasized the regions containing the boundary between the head and the background.

6. Conclusions

We hypothesized that the gaze distribution measured from observers performing a gender recognition task facilitates the extraction of discriminative features. We demonstrated that the gaze distribution measured during a manual gender recognition task tended to concentrate on specific regions of the pedestrian's body. We represented the informative region as a task-oriented gaze distribution for a gender classifier. Owing to the efficacy of the task-oriented gaze distribution for feature extraction, our gender recognition method demonstrated increased accuracy compared with representative existing classifiers and saliency maps.

As part of our future work, we will expand our analytical study to explore the differences in gaze distributions with respect to observer nationality and ethnicity. Furthermore, we intend to generate gaze distributions for various tasks beyond gender recognition, such as evaluating impressions of subjects' clothing in images.

Figure 12. Regions of focus of the gender classifier when performing gender recognition. We used CNNs and Grad-CAM. (a) Test pedestrian images. (b) Results without the use of the gaze distribution $\tilde{g}(x, y)$. (c) Results with our gaze-guided feature extraction.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant No. JP20K11864.

IntechOpen

IntechOpen

Author details

Masashi Nishiyama
Graduate School of Engineering, Tottori University, Tottori, Japan

*Address all correspondence to: nishiyama@tottori-u.ac.jp

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

- [1] Fayyaz M, Yasmin M, Sharif M, Raza M. J-ldfr: Joint low-level and deep neural network feature representations for pedestrian gender classification. *Neural Computing and Applications*. 2021;33:361-391
- [2] Bruce V, Burton AM, Hanna E, Healey P, Mason O, Coombes A, et al. Sex discrimination: How do we tell the difference between male and female faces? *Perception*. 1993;22(2):131-152
- [3] Burton AM, Bruce V, Dench N. What's the difference between men and women? Evidence from facial measurement. *Perception*. 1993;22(2):153-176
- [4] Walther D, Itti L, Riesenhuber M, Poggio T, Koch C. Attentional selection for object recognition—A gentle way. In: *Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*. Berlin Heidelberg: Springer; 2002. pp. 472-479
- [5] Zhu JY, Wu J, Xu Y, Chang E, Tu Z. Unsupervised object class discovery via saliency-guided multiple class learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2015;37(4): 862-875
- [6] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1998;20(11):1254-1259
- [7] Zhang J, Sclaroff S, Lin X, Shen X, Price B, Mech R. Minimum barrier salient object detection at 80 fps. In: *Proceedings of the IEEE International Conference on Computer Vision*. IEEE Computer Society; 2015. pp. 1404-1412
- [8] Zhu W, Liang S, Wei Y, J. sun. Saliency optimization from robust background detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society; 2014. pp. 2814-2821
- [9] Xu M, Ren Y, Wang Z. Learning to predict saliency on face images. In: *Proceedings of IEEE International Conference on Computer Vision*. IEEE Computer Society; 2015. pp. 3907-3915
- [10] Fathi A, Li Y, Rehg JM. Learning to recognize daily actions using gaze. In: *Proceedings of the 12th European Conference on Computer Vision*. Berlin Heidelberg: Springer; 2012. pp. 314-327
- [11] Xu J, Mukherjee L, Li Y, Warner J, Rehg JM, Singh V. Gaze-enabled egocentric video summarization via constrained submodular maximization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society; 2015. pp. 2235-2244
- [12] Karessli N, Akata Z, Schiele B, Bulling A. Gaze embeddings for zero-shot image classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society; 2017. pp. 4525-4534
- [13] Sattar H, Bulling A, Fritz M. Predicting the category and attributes of visual search targets using deep gaze pooling. In: *Proceedings of IEEE International Conference on Computer Vision Workshops*. IEEE Computer Society; 2017. pp. 2740-2748
- [14] Murrugarra-Llerena N, Kovashka A. Learning attributes from human gaze. In: *Proceedings of IEEE Winter Conference on Applications of Computer Vision*. IEEE Computer Society; 2017. pp. 510-519

[15] Hsiao JH, Cottrell G. Two fixations suffice in face recognition. *Psychological Science*. 2008;**19**(10):998-1006

[16] Deng Y, Luo P, Loy CC, Tang X. Pedestrian attribute recognition at far distance. In: *Proceedings of the 22nd ACM International Conference on Multimedia*. Association for Computing Machinery; 2014. pp. 789-792

[17] Bindemann M. Scene and screen center bias early eye movements in scene viewing. *Vision Research*. 2010;**50**(23): 2577-2587

[18] Buswell GT. *How People Look at Pictures: A Study of the Psychology of Perception of Art*. Chicago, IL: University of Chicago Press; 1935

[19] Fairchild MD. *Color Appearance Models*. 3rd ed. New York City: Wiley; 2013

[20] Antipov G, Berrani SA, Ruchaud N, Dugelay JL. Learned vs. hand-crafted features for pedestrian gender recognition. In: *Proceedings of the 23rd ACM International Conference on Multimedia*. Association for Computing Machinery; 2015. pp. 1263-1266

[21] Weinberger KQ, Saul LK. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*. 2009;**10**: 207-244

[22] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of IEEE International Conference on Computer Vision*. IEEE Computer Society; 2017. pp. 618-626