IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 12, 2021, accepted March 30, 2021, date of publication April 6, 2021, date of current version April 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3071406

Visual Place Recognition From Eye Reflection

YUKI OHSHIMA, KYOSUKE MAEDA, YUSUKE EDAMOTO, AND ATSUSHI NAKAZAWA, (Member, IEEE)
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
Corresponding author: Atsushi Nakazawa (nakazawa.atsushi@i.kyoto-u.ac.jp)

This work was supported by Core Research for Evolutional Science and Technology under Grant JPMJCR17A5.

ABSTRACT The cornea in the human eye reflects incoming environmental light, which means we can obtain information about the surrounding environment from the corneal reflection in facial images. In recent years, as the quality of consumer cameras increases, this has caused privacy concerns in terms of identifying the people around the subject or where the photo is taken. This paper investigates the security risk of eye corneal reflection images: specifically, visual place recognition from eye reflection images. First, we constructed two datasets containing pairs of scene and corneal reflection images. The first dataset is taken in a virtual environment. We showed pre-captured scene images in a 180-degree surrounding display system and took corneal reflections from subjects. The second dataset is taken in an outdoor environment. We developed several visual place recognition algorithms, including CNN-based image descriptors featuring a naive Siamese network and AFD-Net combined with entire image feature representations including VLAD and NetVLAD, and compared the results. We found that AFD-Net+VLAD performed the best and was able to accurately determine the scene in 73.08% of the top-five candidate scenes. These results demonstrate the potential to estimate the location at which a facial picture was taken, which simultaneously leads to a) positive applications such as the localization of a robot while conversing with persons and b) negative scenarios including the security risk of uploading facial images to the public.

INDEX TERMS Corneal reflection, computer vision, deep learning, image recognition, biometrics, privacy, security.

I. INTRODUCTION

The cornea of the human eye acts as a mirror that reflects light from a person's environment, which means that visual information about the environment can be measured from the corneal reflections. The environmental information can include the environmental image [1], illumination conditions [2], and high-resolution reconstructions of the environment [3].

Since a lot of information can be retrieved from human eye reflections, and people are mostly unaware of the potential of information retrieval from eye images, we need to think carefully about the security concerns of publishing and sharing facial-eye images. Jenkins *et al.* were the first to point out the potential of human identification from an eye reflection in a facial image [4]. On the positive side, eye reflections have been used for crime scene investigations. On the negative side, they have been abused in stalker incidents [5]. Thus, assessing the security of eye reflections is a crucial step in the development of digital cameras and smartphones.

The associate editor coordinating the review of this manuscript and approving it for publication was Tomasz Trzcinski.

In this paper, we propose a method of visual place recognition from eye reflection images for the purpose of alerting users to the security risk of publically exposing facial images. Conceptual illustration is shown in Fig.1. Our objectives are to help evaluate whether a facial image is secure or not and to develop methods to remove the location information from faces shown in the image. To this end, we first develop two facial image datasets taken from 11 subjects and 100 scenes by using a publicly available scene image dataset and corresponding eye images. Namely, we show the scene images in a 180-degree virtual display and take the eye images of the subject who is located in the display. Secondly, we developed a dataset consisting of pairs of scene and eye reflection images taken in actual 104 outdoor scenes involving eight subjects.

We then develop a novel method of visual recognition from corneal reflections. While a considerable number of visual scene recognition algorithms already exist, the problem we examine is different from such algorithms due to a large amount of noise in eye images, such as iris texture contamination, eyelid and eyelash shadows, limited image resolution, and image blur.

We implemented and evaluated several approaches to combat this issue. Fig. 6 illustrates the general flow of the

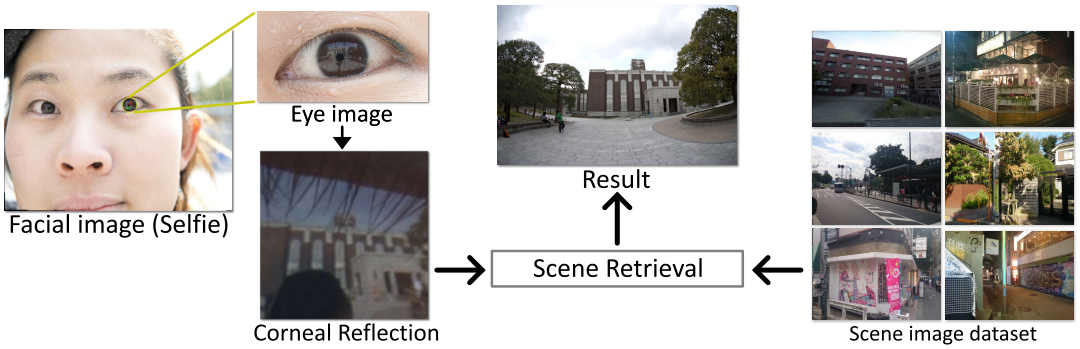


FIGURE 1. Conceptual illustration of the scene retrieval from a facial (selfie) image. First, the corneal reflection (eye-scene reflection) is retrieved from a facial image. The image is then matched with the scene image dataset by using a novel DNN-based image retrieval algorithm pretrained by our scene-eye reflection image dataset. In the final step, the scene where the facial image was taken is recognized.

DNN-based scene retrieval algorithm. First, we train the DNN-based image descriptors by using aligned images of eye reflection and scene images. Here we used the idea of DNN-based metric learning, namely, we train a DNN that is trained to output 1) the same feature vector when the input eye reflection or scene images captures the same scene, and 2) a different feature vector when the images capture the different. Afterward, feature aggregation is conducted that output the entire image feature vector. Thanks to the metric learning-based feature descriptors, the algorithm robustly performs the scene similarity regardless of image noise. As a result, our algorithm is able to attain an accuracy of more than 73% for large numbers of scene datasets.

Our main contributions in this paper are as follows.

1) We developed image datasets including more than 1,000 pairs of scenes and eye reflections. As far as we know, this is the first dataset of pairs of eye reflection and scene images. Since these image pairs are accurately aligned, they can be used for evaluation and training.
2) We developed and evaluated several visual place recognition algorithms that use handcrafted and DNN-based image descriptors combined with a vector of locally aggregated descriptors (VLAD) [6] or NetVLAD [7].
3) Through comprehensive experiments on visual place recognition and comparison to the existing algorithms, we demonstrated that the proposed algorithm outperformed all others and reached an accuracy of more than 73% in the top-five scene candidates detection task.

In Section 2 of this paper, we provide an overview of related work. We explain the datasets and the algorithms in Sections 3 and 4. The experimental results are reported in Section 5. We conclude in Section 6 with a brief summary.

II. RELATED WORK

This section discusses related work on eye image analysis and image registration.

A. EYE IMAGE ANALYSIS

The iris region in an eye image is a mixture of the refracted iris texture and the corneal surface reflection of the scene illumination. As the iris texture is important for personal identification [8], [9] and iris biometrics [10], several works have investigated methods to separate iris texture and corneal reflection. He *et al.* obtained a reflection map from an iris region by using an adaptive thresholding approach and applied a bilinear interpolation to fill out the region [10]. Tan and colleagues used a labeling-based corneal reflection removal for the purpose of iris segmentation [11], and Wang *et al.* applied the color chromaticity of the iris texture for this task [12]. To estimate the scene illumination, they took the consensus of corneal reflections from the images of both eyes. As these approaches rely on heuristic rules, such as assuming bright scene reflections with sharp edges or consistent chromaticity in iris colors, they typically have weak performance in scenes, where the assumptions do not hold. We believe this problem can be solved easily and accurately when a pixel-wise correspondence between an eye and a scene image is available. Moreover, while the above approaches are purely image-based, we show that explicit geometric modeling of the eye and the light reflection at the corneal surface is beneficial for this task.

The first corneal imaging technique was developed by Nishino and Nayar [1], and then several research groups conducted extensive studies [13], [14]. In these works, a camera capturing an image of the spherical or aspherical eye that exhibits corneal reflections is modeled as a non-rigid catadioptric imaging system [15]. Applying this model enables the scene illumination to be reconstructed from an eye image, such as through geometric calibration between the eye and a computer display [16], optical see-through head-mounted displays (OST-HMD) [17], or a fish-eye camera [18]. In [19], an aspherical surface model was introduced for the cornea. Using an extensive model, they demonstrated an accurate image registration algorithm between scene and corneal reflection images. While these works have demonstrated application showcases of scene recognition from corneal reflections, the number of target scene images is limited and therefore the true applicability is unknown.

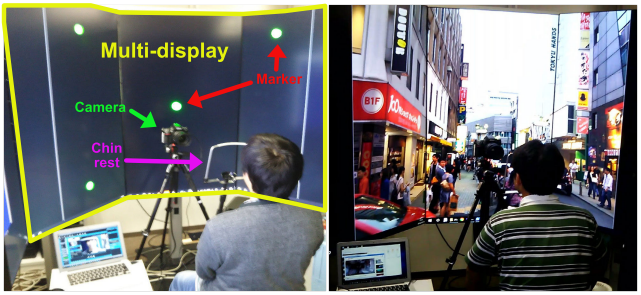
Since eye reflection captures a lot of information about the surrounding scene at a very wide angle, several studies have examined using eye/facial images for digital forensics, which is tricky because there is a security risk in terms of exposing such images. As a digital forensics application, Johnson *et al.* developed a method to reveal whether a group

photo is digitally composited or not by using the lighting conditions reconstructed from the persons' eye reflections [20]. Backets *et al.* demonstrated the recovery of the content presented in a computer display from the corneal reflection of a subject who was looking at [13]. More recently, Jenkins *et al.* pointed out the potential of human identification from an eye reflection in a facial image [4]. They found that people could identify human faces with an accuracy of about 71% and 84% for non-familiar and familiar faces, respectively. Corneal reflections have now started being used in actual crime scene investigations. In 2017, police analyzed the photo of a victim on a smartphone and uncovered a reflection of the suspect in the victim's eyes, and used it for trial evidence. Unfortunately, a similar technique was abused in a stalker incident where the criminal analyzed the corneal reflection in an uploaded facial picture to identify the place where the victim was living [5].

B. VISUAL PLACE RECOGNITION

Place recognition from an image is a task to find the location similar to the query image from a dataset of place images [21], [22]. Traditional methods used hand-crafted features such as SIFT or SURF [23] for scene images, and then aggregated them by using such as bag-of-words [24], [25], Fischer vector [26], and Vector of Locally Aggregated Descriptors (VLAD) [6], [27], and constructed a scene image database. The scene retrieval is performed by comparing an image feature of a query scene image with the database. Bag-of-words uses a histogram representing the number of features belonging to each cluster as a descriptor. Fisher vector and VLAD extend the idea namely, the descriptors represent the residuals between the local features and cluster centers. Fisher vector is a fixed-length vector representation based on the assumption that visual words are generated from the Gaussian mixture model (GMM), while VLAD aggregates all residuals of local features belonging to each cluster, and produces a $k \times d$-dimensional descriptor where k is the number of clusters and d is the dimension of the local feature. While most of the aggregated features omit the positional information of the local descriptors thus are potentially fair to the brute-force matching, they are robust against misalignment and require smaller computational power due to their compact feature size. Torii *et al.* [28] proposed a place recognition algorithm robust to changes of the conditions such as viewpoints and day/night using panoramic images combined by viewpoint changes. In this case, DenseVLAD is used as the descriptor.

In recent years, with the progress of CNN, considerable numbers of methods have been proposed to use the pre-learned CNN as a descriptor [22], [29], [30]. One of the early and major approaches is NetVLAD [7] which enables end-to-end learning by incorporating the computing step of VLAD into a DNN-layer. Local image features are directly computed from the 2D-CNN feature map, and then they are aggregated. Moreover, the anchors of the clusters in computing VLAD are also determined instead of using the predefined cluster center.



(a) Experimental setup (Virtual environment, Markers are displayed)

(b) Experimental scene (Virtual environment, A scene image is displayed)

FIGURE 2. Experimental setup of virtual environment. (a) Experimental setup. Markers are displayed on a large multi-display. (b) Experimental setup where a scene image is displayed.

III. DATASET

We used two scene datasets and two eye reflection datasets. The publicly available scene dataset (Tokyo24/7) was mainly used as the training data. We set up a large multi-display for projecting scene images and then obtained the eye reflections of a subject who was looking at the display.

A. SCENE DATASET

1) Tokyo24/7 DATASET

The Tokyo24/7 dataset [28] consists of 1,125 scene images taken by smartphones (Apple-iPhone5s and Sony Xperia). Images were taken from 125 distinct locations and three different viewing directions at three different times of the day. Since the actual location where the images were taken is unknown, we used this dataset to obtain the training data. Specifically, we presented the scene images to a subject using a 180-degree display and obtained eye reflection images (see section III-B1).

2) KYOTO SCENE DATASET

We collected 104 scene images taken in outdoor campus environments. The images were taken by a Nikon Z6 digital camera with an SIGMA 24-70mm lens. The resolution of the images was 6048 × 4024 pixels. We cropped the center area and resize to 256 × 256 pixels.

B. EYE REFLECTION DATASETS

1) EYE REFLECTIONS IN A VIRTUAL ENVIRONMENT (EyeVE DATASET)

We took 100 scenes from the Tokyo24/7 dataset and collected 1,053 eye reflection images from 11 subjects. The experimental setup is shown in Fig. 2. Scene images were shown to each subject using a multi-display environment consisting of three 70-in monitors. The subject's head was fixed on a chin-rest located about 80 cm from the display center. Facial images of the subjects were taken by a Nikon Z6 digital camera at a distance of 25 cm. Fig. 3 shows several examples of the scene and eye reflection images. The average diameter of the iris was about 350 pixels.

To obtain accurately aligned images of scenes and eye reflections, we performed the following steps when taking images. First, we sequentially showed an image of five grid

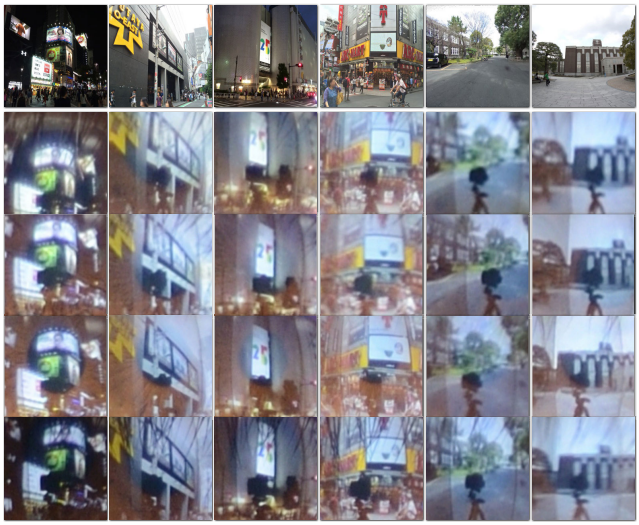


FIGURE 3. Examples of EyeVR dataset. Top row: Scene images. The four images on the left and two images on the right are taken from Tokyo24/7 and Kyoto scene datasets, respectively. Rows 2–6: Corneal reflection images taken from different subjects.

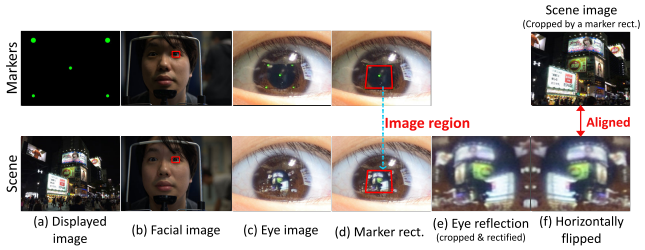


FIGURE 4. Image alignment steps in collecting EyeVE dataset that is mainly used for training. (a) A marker and scene image are sequentially displayed in the multi-display environment. (b) Facial and (c) eye images. (d) Marker positions are found from the eye image and then an image region is obtained. (e) The image region in the eye reflection is cropped and rectified. (f) The image region is flipped horizontally. An aligned pair of the scene and eye reflection images is obtained in the end.

markers and a scene image ((Fig. 4(a)). An eye image was taken for each display content ((Fig. 4(b),(c)). By choosing images where the subject's eye did not move while a marker and scene image were shown, we can obtain a pair of eye reflection and scene images that are accurately aligned by cropping and rectifying the eye reflection image according to the marker locations ((Fig. 4(d)–(f)).

2) EYE REFLECTION DATASET IN OUTDOOR ENVIRONMENTS (EyeKyoto DATASET)

We collected 104 eye reflection images in real outdoor environments. Each scene corresponds to the images in the Kyoto scene dataset. Fig. 5 shows several examples of the scene and eye reflection images. We used a Chin-rest to fix the subject's face, and the camera is located at about 25cm from the subject's eye. Assuming the actual application scenario that finding a scene from an eye reflection, we did not performed fine alignment but manually cropped the scene regions in the eye reflection images and used them for evaluation.

IV. SCENE RECOGNITION ALGORITHMS

To recognize scenes from noisy eye reflection images, we implemented a novel DNN-based image recognition

FIGURE 5. Examples of Eye-Kyoto dataset that is used for testing. Top row: Scene images. Bottom row: Corneal reflection images taken from different subjects.

algorithm that is robust to image noises. Fig. 6 shows the algorithm overview. First, we train a feature descriptor network to evaluate the similarities of image patches. From the pairs of scene and eye images, we take the patches corresponding to the ground truth and use them to train the network. We implement and evaluate Siamese network-based [31] and AFD-Net-based [32] descriptors. In the end, the network is trained to output similar feature values for corresponding patches and different values for non-corresponding ones. Using the trained network, we take dense features from every scene image and construct VLAD features that represent the images. In recognition, dense features are obtained from an eye reflection image, and then a VLAD feature vector is computed. The VLAD feature is matched to the scene VLAD features. In the following subsections, we introduce networks to evaluate patch similarity and VLAD descriptors.

A. TRAINING FEATURE DESCRIPTOR NETWORKS

We use two patch-based deep image descriptors to represent local image features: a CNN-Siamese network and AFD-Net.

1) CNN-SIAMESE NETWORK

The Siamese network is a metric learning method that is often used in tasks such as face verification. It is a model that learns the projection from the feature vector to the L2 space. Fig. 7 shows the structure of our network. It uses three convolutional neural networks with shared weights. The inputs of these three networks are an anchor, a positive sample with the same label as the anchor, and a negative sample with a different label from the anchor, respectively. This model needs triplet loss to make the L2 distance between the anchor feature vector and the positive feature vector close and to make the L2 distance between the anchor feature vector and the negative feature vector move away. In our method, we use a network formed by concatenating a convolutional layer, an instance normalization layer, an activation layer by a hyperbolic tangent (tanh), and average pooling as convolutional neural networks. The inputs to this network are the anchor, which is a patch from the eye image, the patch of the scene image corresponding to that patch as the positive sample, and the patch at different random positions as the negative sample. We manually pick up these corresponding points from both types of images.

2) AFD-NET

Another local image descriptor is the AFD-Net which is another metric learning algorithm for extracting more

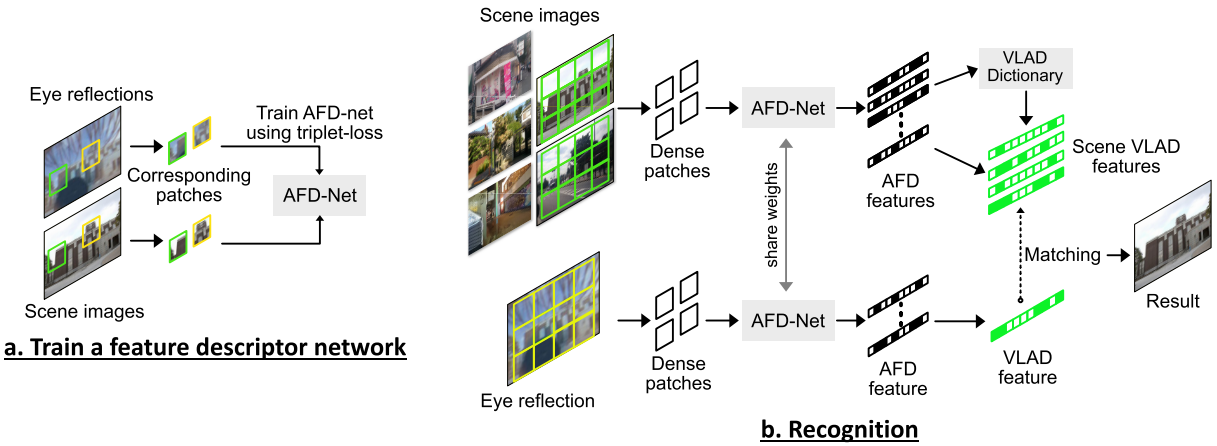


FIGURE 6. Overview of the proposed AFD-VLAD-based image retrieval algorithm. (a) First, a feature descriptor network (Siamese network or AFD-Net) is trained by using the corresponding image patches between scene and eye reflection images. The feature descriptor network then outputs the same feature vectors for the corresponding patches. (b) In recognition, scene VLAD features is generated from the scene images and the trained feature descriptor network. Similarly, local features are obtained from image patches of an eye reflection image, and then a VLAD feature vector is computed. The VLAD feature vector is matched to the scene VLAD features and the corresponding scene is retrieved.

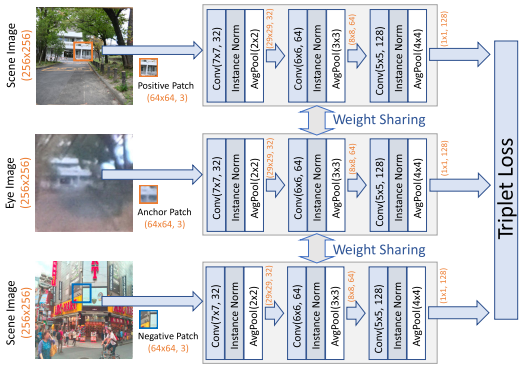


FIGURE 7. The structure of CNN-Siamese network.

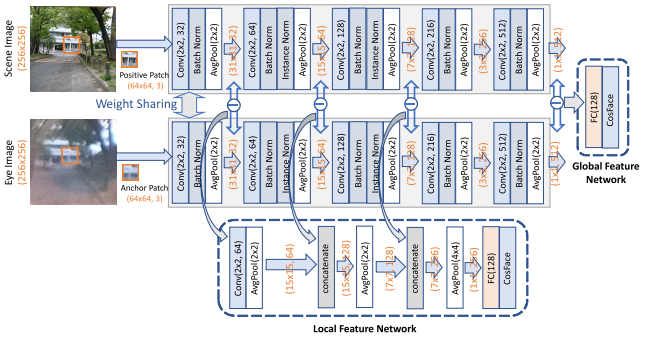


FIGURE 8. The structure of AFD-Net.

effective features for scene identification [32]. The structure of AFD-Net is illustrated in Fig. 8. First, we respectively input the patch from the eye image and the patch from the scene image to the two weight-sharing CNNs, similar to the method using triplet loss. Our CNN is a network consisting of five convolution layers. After all the convolutional layers, a batch regularization layer, an activation layer by the *tanh* function, and a pooling layer using average pooling are connected. In addition, instance normalization layers are added after the second and third batch normalization layers. The difference of L1 norm between the two 512-dimensional feature vectors, which are the outputs of the final layer, becomes the input of the global feature network and the input of the large margin cosine loss (LMCL) of Cosface [33] via a fully connected layer. In addition, in order to use the local features that appear in the shallow layers of the convolutional network, we construct a local feature network using the differences between the convolutional layer's output tensors and perform learning with multitasking loss. Specifically, the difference is taken for the output from the 1st layer to the 3rd layer in the same way as the global feature, and the difference between the previous layer with average pooling applied and the difference between the next layers are concatenated and calculated. Then, the process of average pooling to reduce the image size is repeated, and finally, a 512-dimensional feature vector is obtained and input to Cosface. We set the scale parameter s of Cosface as 20 and the margin parameter m as 0.05.

B. FEATURE AGGREGATION USING VLAD/NetVLAD

We then compute aggregated features to represent the entire scene or eye reflection images. Given trained local feature descriptor networks, dense local features are taken from images. Namely, we uniformly sample patches from an entire image and obtain multiple feature descriptors, and then VLAD or NetVLAD image features are computed. In the recognition step, a VLAD or NetVLAD feature vector is obtained from a query eye reflection image and matched to stored scene features obtained from a scene image dataset.

V. EXPERIMENTS

We evaluated the proposed and existing algorithms using the datasets. We used the Tokyo24/7 and EyeVE datasets for training and the Kyoto and EyeKyoto datasets for testing. The input image size was 256 × 256 pixels.

TABLE 1. Experimental result.

	SIFT +VLAD	NetVLAD (Entire img)	CNN-Siamese (Entire img)	CNN-Siamese +NetVLAD	CNN-Siamese +VLAD	AFD-Net (Entire img)	AFD-Net(Patch) +NetVLAD	AFD-Net(Patch) +VLAD
Top 1	11 (10.58%)	3 (2.88%)	17 (16.35%)	2 (1.92%)	**49 (47.12%)**	16 (15.38%)	2 (1.92%)	**49 (47.12%)**
Top 1-2	15 (14.42%)	5 (4.81%)	25 (24.04%)	2 (1.92%)	56 (53.85%)	22 (21.15%)	4 (3.85%)	**62 (59.62%)**
Top 1-3	17 (16.35%)	7 (6.73%)	28 (26.92%)	2 (1.92%)	63 (60.58%)	27 (25.96%)	7 (6.73%)	**71 (68.27%)**
Top 1-4	21 (20.19%)	9 (8.65%)	32 (30.77%)	3 (2.88%)	67 (64.42%)	34 (32.69%)	9 (8.65%)	**76 (73.08%)**
Top 1-5	24 (23.08%)	10 (9.62%)	35 (33.65%)	5 (4.81%)	69 (66.35%)	40 (38.46%)	9 (8.65%)	**76 (73.08%)**

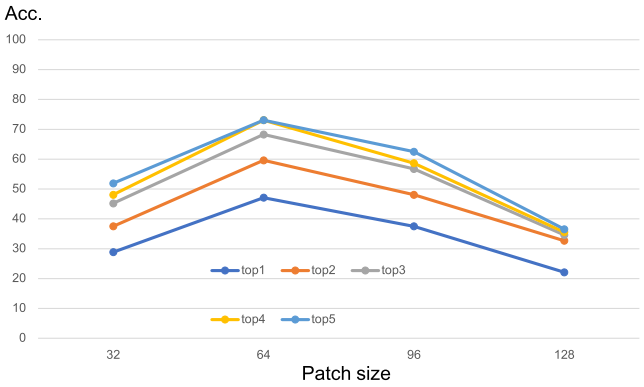


FIGURE 9. Patch-size v.s accuracy in AFD-Net(Patch)+VLAD. It performed best when patch size = 64.

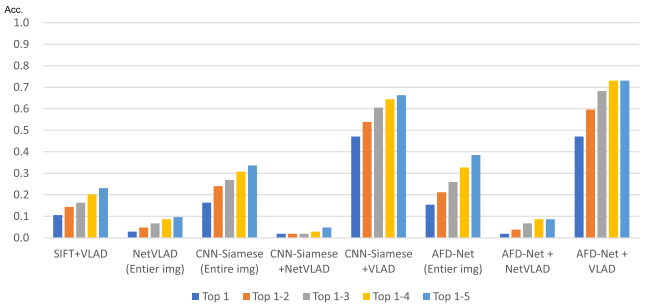


FIGURE 10. Experimental result (Accuracy comparison).

A. METHODS

We implemented the following algorithms and compared the results. The baseline algorithm was the combination of a handcrafted feature (SIFT) and VLAD. For the naive DNN-based image retrieval, we used NetVLAD and the CNN-Siamese network, both of which use the enter image as their inputs. The others are combinations of two DNN-based local image descriptors (CNN-Siamese and AFD-Net) and two entire image descriptors (VLAD and NetVLAD). For the DNN-based local descriptors, the size of the input patch was 64×64 pixels and the output feature dimension was 512. For the entire image descriptors, the number of clusters was set to 64, so the output feature dimension was 512×64.

Experiments were performed on a PC environment (Ubuntu 16.04, Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz, 64GB Memory, Nvidia GeForce1080Ti-11GB). Matlab 2019a + VLFeat 0.9.21 [34] were used for the SIFT-VLAD-based algorithms, and pytorch-1.4.0 + tensorflow were used for the DNN-based algorithms.

B. RESULTS

Table 1 and Fig.10 show the experimental results and Fig. 11 shows several examples of the Top-1 retrieval results of each method. The best performance in the retrieval of the top five candidates was by AFD-Net(Patch)+VLAD, which accurately determined the scene in 73.08% of the top-five candidate scenes. We also evaluated the effect of the patch size of the local descriptor on the performance. Fig. 9 shows the results of AFD-Net(Patch)+VLAD. As the result, it performed best when the patch size is 64.

Overall, methods using patches (local descriptors) performed better than those using entire images. This is because the methods using the entire image suffer from 1) the misalignment of the query and target images and 2) partial occlusions and noise in the eye images, while patch-based ones could avoid the issues. A method using a hand-crafted feature (SIFT) could retrieve the images that have high-contrast structures e.g. buildings (Fig.11(l)), however, performed poorer for the scenes with natural objects. A method using CNN-Siamese (Entire image) could retrieve images having similar color distributions, therefore, it worked when the eye reflection had a similar color and texture features to that of scene images (Fig.11(c)).

Another advantage of patch-based features over entire-image features is the ability to utilize local structures for the decision. As seen in the results of Fig.11(c), CNN-Siamese (Entire Image) retrieved the images whose whole structure in an image – such as the perspective of the buildings or boundaries between roads and other objects – were similar to the ground-truths. However, they failed to use local structures such as colors or textures. Specially in NetVLAD (Entire image) retrieved particular images very frequently because the method confused with the ground surface (in scene images) and iris (in eye reflection images) since they had similar colors.

Regarding the comparison of AFD-Net and CNN-Siamese local features, AFD-Net seems to be much robust to image noises (Fig.11(d)-(g)). Since AFD-Net is trained to minimize the outputs of intermediate layers between scene and eye reflections, it becomes robust to multi-level eye-related image noises such as eyelash or iris textures. Inversely, when the scene and eye reflection images are quite similar, the local feature network does not contribute to increasing performances since the differences of intermediate layers' outputs are very close (Fig.11(h)-(l)).

Fig.11(m) shows the hardest case. In the image, the upper half is covered by eyelash shadows and the lower half is iris texture, therefore, scene reflections cannot be observed from the eye reflection, therefore, all methods failed to recognize.

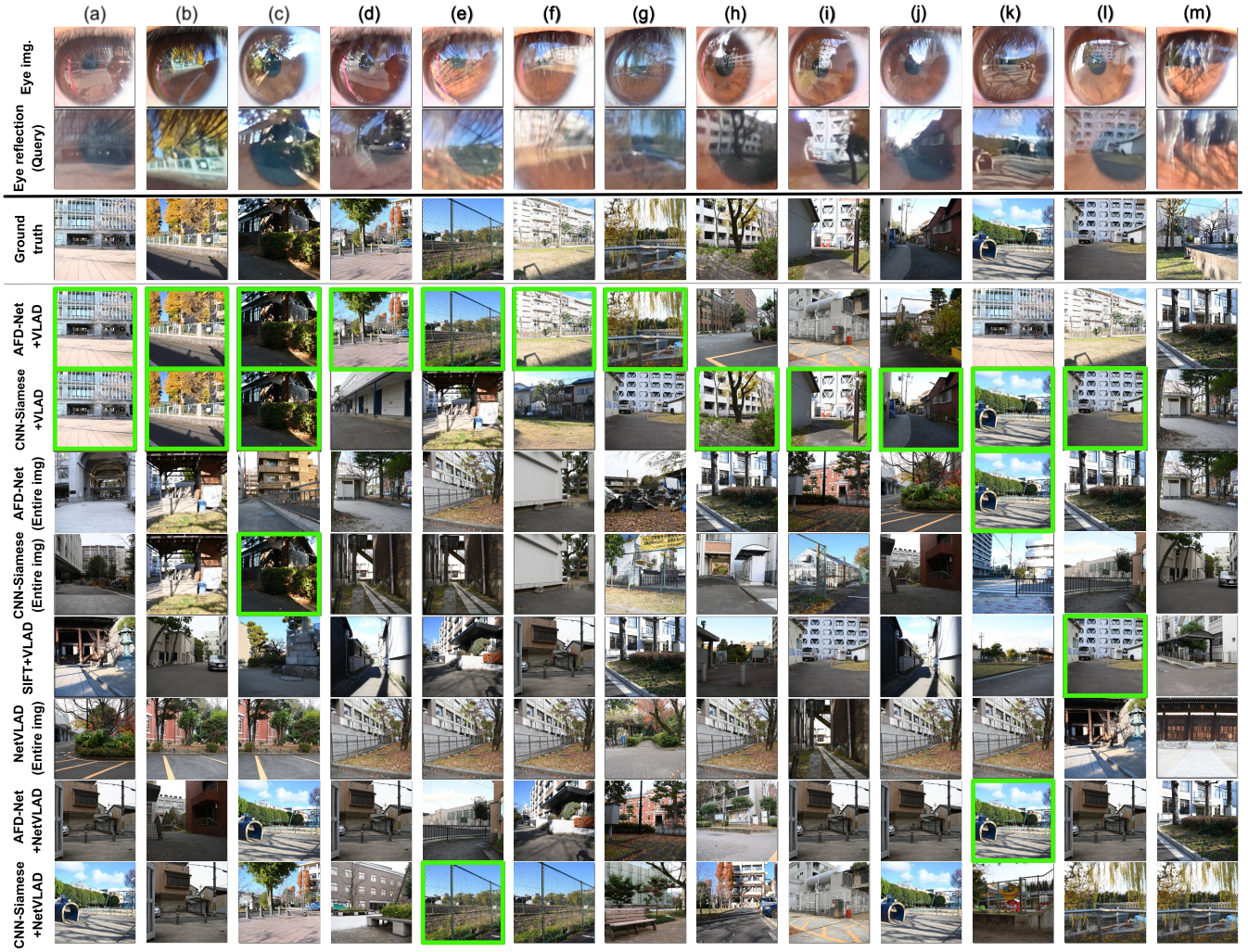


FIGURE 11. Experimental results (Top-1 retrievals). Thumbnails with green squares indicate correct retrievals.

VI. CONCLUSION

In this paper, we have presented our method of visual place recognition from eye reflection images. This task is much more difficult than naive image-based scene recognition tasks since eye reflections include a lot of image noise (e.g., iris textures, eyelid and eyelash shadows, and occlusions). However, it is a very socially important task due to inherent security concerns. We developed and evaluated algorithms that use handcrafted and DNN-based image descriptors combined with aggregated image descriptors (VLAD and NetVLAD) and found that the combination of AFD-Net+VLAD had an accuracy of more than 73% in the top-five scene recognition tasks. As far as we know, this is the first work to tackle this task by using a large image dataset.

Our findings also revealed the impact and limitations of eye reflection-based scene recognition. First, it becomes clear there is a considerable level of security risk in taking/sharing facial images due to a leakage of information where the image was taken. Considerable care must be taken to address the potential privacy issues with these kinds of images. The images in this study were taken with a high-end consumer camera under a controlled setup in which both the camera and face were fixed by mounts. Therefore, the accuracy reported here would not realistically be achieved in everyday photo-shooting conditions. However, considering the future performance improvements expected of consumer/smartphone cameras, we should keep an eye on the potential security risks.

REFERENCES

[1] K. Nishino and S. K. Nayar, "Corneal imaging system: Environment from eyes," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 23–40, Oct. 2006.

[2] K. Nishino, P. N. Belhumeur, and S. K. Nayar, "Using eye reflections for face recognition under varying illumination," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 519–526.

[3] C. Nitschke and A. Nakazawa, "Super-resolution from corneal images," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 22.1–22.12.

[4] R. Jenkins and C. Kerr, "Identifiable images of bystanders extracted from corneal reflections," *PLoS ONE*, vol. 8, no. 12, Dec. 2013, Art. no. e83325.

[5] *Stalker Found Japanese Singer Through Reflection in Her Eyes*. Accessed: Apr. 12, 2021. [Online]. Available: https://www.bbc.com/news/world-asia-50000234

[6] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.

[7] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5297–5307.

[8] K. W. Bowyer, K. Hollingsworth, and P. J. Flynn, "Image understanding for iris biometrics: A survey," *Comput. Vis. Image Understand.*, vol. 110, no. 2, pp. 281–307, May 2008.

[9] L. Ma, T. Tan, Y. Wang, and D. Zhang, "Personal identification based on iris texture analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1519–1533, Dec. 2003.

[10] Z. He, T. Tan, Z. Sun, and X. Qiu, "Toward accurate and fast iris segmentation for iris biometrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1670–1684, Sep. 2009.

[11] T. Tan, Z. He, and Z. Sun, "Efficient and robust segmentation of noisy iris images for non-cooperative iris recognition," *Image Vis. Comput.*, vol. 28, no. 2, pp. 223–230, 2010.

[12] H. Wang, S. Lin, X. Liu, and S. B. Kang, "Separating reflections in human iris images for illumination estimation," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Oct. 2005, pp. 1691–1698.

[13] M. Backes, T. Chen, M. Duermuth, H. P. A. Lensch, and M. Welk, "Tempest in a teapot: Compromising reflections revisited," in *Proc. 30th IEEE Symp. Secur. Privacy*, May 2009, pp. 315–327.

[14] C. Nitschke, A. Nakazawa, and H. Takemura, "Corneal imaging revisited: An overview of corneal reflection analysis and applications," *IPSJ Trans. Comput. Vis. Appl.*, vol. 5, pp. 1–18, Jan. 2013.

[15] P. Sturm, S. Ramalingam, J.-P. Tardif, S. Gasparini, and J. Barreto, "Camera models and fundamental concepts used in geometric computer vision," *Found. Trends Comput. Graph. Vis.*, vol. 6, pp. 1–183, Jan. 2011.

[16] C. Nitschke, A. Nakazawa, and H. Takemura, "Display-camera calibration using eye reflections and geometry constraints," *Comput. Vis. Image Understand.*, vol. 115, no. 6, pp. 835–853, Jun. 2011.

[17] A. Plopski, Y. Itoh, C. Nitschke, K. Kiyokawa, G. Klinker, and H. Takemura, "Corneal-imaging calibration for optical see-through head-mounted displays," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 4, pp. 481–490, Apr. 2015.

[18] T. Ogawa, A. Nakazawa, and T. Nishida, "Point of gaze estimation using corneal surface reflection and omnidirectional camera image," *IEICE Trans. Inf. Syst.*, vol. 101, no. 5, pp. 1278–1287, 2018.

[19] A. Nakazawa, C. Nitschke, and T. Nishida, "Registration of eye reflection and scene images using an aspherical eye model," *JOSA A*, vol. 33, no. 11, pp. 2264–2276, 2016.

[20] M. K. Johnson and H. Farid, "Exposing digital forgeries through specular highlights on the eye," in *Proc. Int. Workshop Inf. Hiding*. Heidelberg, Germany: Springer, 2007, pp. 311–325.

[21] S. Lowry, N. Sunderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[22] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognit.*, vol. 113, May 2021, Art. no. 107760.

[23] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[24] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 3921–3926.

[25] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, p. 1470.

[26] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3384–3391.

[27] R. Arandjelovic and A. Zisserman, "All about VLAD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1578–1585.

[28] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1808–1817.

[29] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Computer Vision—ECCV 2014* (Lecture Notes in Computer Science). Heidelberg, Germany: Springer, 2014, pp. 584–599.

[30] A. B. Yandex and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1269–1277.

[31] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 118–126.

[32] D. Quan, X. Liang, S. Wang, S. Wei, Y. Li, N. Huyan, and L. Jiao, "AFD-Net: Aggregated feature difference learning for cross-spectral image patch matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3017–3026.

[33] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[34] A. Vedaldi and B. Fulkerson. (2008). *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. [Online]. Available: http://www.vlfeat.org/

YUKI OHSHIMA was born in Okazaki, Aichi, Japan, in 1995. He received the B.S. degree in computer science from Kyoto University, in 2019, where he is currently pursuing the master's degree with the Department of Informatics. His interests include computer vision and biometrics.

KYOSUKE MAEDA was born in Nagoya, Aichi, Japan, in 1997. He received the B.S. degree in computer science from Kyoto University, in 2020, where he is currently pursuing the master's degree with the Department of Informatics. His interests include computer vision and human interaction.

YUSUKE EDAMOTO was born in Kyoto, Japan, in 1996. He received the B.S. degree in computer science from Kyoto University, where he is currently pursuing the master's degree with the Department of Informatics.

ATSUSHI NAKAZAWA (Member, IEEE) received the Ph.D. degree in systems engineering from Osaka University, in 2001. He worked with the Institute of Industrial Science, University of Tokyo and then in Cybermedia Center, Osaka University. Since 2013, he has been with Kyoto University, where he is currently an Associate Professor with the Department of Informatics. From 2007 to 2008, he joined the GVU Center, Georgia Institute of Technology, as a Visiting Researcher. In 2010, he was awarded the Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Agency (JST). In October 2017, he became a Program Investigator (PI) of the JST CREST project "Computational and cognitive neuroscientific approaches for understanding the tender care."