ITE Trans. on MTA Vol. 7, No. 2, pp. 68-79 (2019)

Invited Paper

Interactive Face Retrieval Framework for Clarifying User's Visual Memory

Yugo Sato †, Tsukasa Fukusato ††, Shigeo Morishima (member)†††

Abstract This paper presents an interactive face retrieval framework for clarifying an image representation envisioned by a user. Our system is designed for a situation in which the user wishes to find a person but has only visual memory of the person. We address a critical challenge of image retrieval across the user's inputs. Instead of target-specific information, the user can select several images that are similar to an impression of the target person the user wishes to search for. Based on the user's selection, our proposed system automatically updates a deep convolutional neural network. By interactively repeating these process, the system can reduce the gap between human-based similarities and computer-based similarities and estimate the target image representation. We ran user studies with 10 participants on a public database and confirmed that the proposed framework is effective for clarifying the image representation envisioned by the user easily and quickly.

Key words: Face retrieval, User interaction, Deep convolutional neural network, Relevance feedback, Active learning.

1. Introduction

In recent years, a large number of photos that include a variety of unconstrained subjects, such as generic objects and human faces, have been uploaded to social networks or photo-sharing services. Hence, efficient systems to retrieve images from such a large volume of data are in demand. Web image search systems such as Google, Yahoo!, and Bing utilize several items with embedded information, such as filenames, image captions, and text on web pages[1)2)3)]. While text-based search techniques have achieved success in document retrieval tasks, these embedded tags are often unreliable for describing image contents, and the quality of manually defined tags can affect the performance of the image retrieval process[4)5)6)]. In addition, if a user seeks an image with visual characteristics that cannot be easily expressed by keywords, the user would generally have to scroll through large numbers of image results retrieved by using keywords, in search of the desired image.

Computer-vision-based studies generally analyze contents of images; for example, they compute the similarity between a query and each image of a database with image descriptors such as color histograms[7)] or Gabor texture features[8)]. Using these similarities, the system enables a user to retrieve a set of images easily without text queries, which is a process called content-based image retrieval. However, there is a well-known challenging problem called "semantic gap" between low-level visual features and the high-level intention of the user, which makes it difficult to search for user-desired images[9)10)11)].

Recently, highly accurate image recognition methods with deep learning have been reported (e.g., image classification tasks). Dense data of raw images are abstracted into high-dimensional sparse representations via convolution and pooling layers. By learning from a large-scale database, deep convolutional neural networks can generate generic representations and classifiers adapted to a given task. By extending its features, Donahue et al. introduced the deep convolutional activation feature (DeCAF), which utilizes the representation layers as image descriptors and can compute image representations with more semantic information compared to low-level visual features[12)]. The image representations obtained through deep learning consistently outperform conventional hand-crafted features and boost the image retrieval performance[13)14)15)]. However, the image representations are calculated fully automatically, and it is difficult to reflect the intention of a user in the retrieval process.

To ensure that a user's intention is reflected in the retrieval process, many studies have typically utilized

Received August 19, 2018; Revised November 18, 2018; Accepted January 6, 2019
† Waseda University
(Tokyo, Japan)
†† The University of Tokyo
(Tokyo, Japan)
††† Waseda Research Institute for Science and Engineering
(Tokyo, Japan)

the "relevance feedback" approach, which allows the user to interactively refine retrieval results[16)17)18)]. The main process includes three steps: the system (i) provides initial results of queries provided by the user; (ii) gathers user feedback according to his/her subjective judgment; and (iii) updates the retrieval results based on the user's feedback on whether those results are relevant[16)]. However, these systems require an additional user-task of finding text or image queries related to the target in advance because they assume that the user has some specific queries.

In this study, we propose a framework that belongs to a general category of content-based image retrieval but is different from existing techniques in that it clarifies an obscure target image that a user envisions by relying on his/her visual memory. A usage case is provided in Fig.1. The system enables the user to find a person whose name or affiliation is unknown by selecting similar people on a search window. To achieve such a system, we extend the concept of DeCAF and propose an interactive image descriptor based on online learning with multiple feedback instances. Fig.2 shows a flowchart of the proposed retrieval framework. Our search process includes the following steps: (i) gathering a user selection based on relevance feedback (including images that are similar to the target envisioned by the user); (ii) online learning based on the user-selected images (modifying a search area according to the relevant images and fine-tuning an image descriptor); (iii) selection of images presented to the user (re-ranking initial retrieval results based on the fine-tuned image representation and sample compression with active learning). By interactively repeating the user's input and the system's search process, we can estimate the target image representation envisioned by the user.

This paper is the extended version of the recent conference report[19)] and more detailed experiments and discussions are newly added.

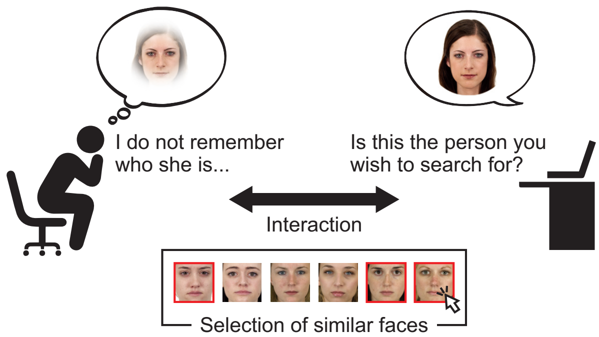

Fig. 1 Face retrieval relying on the user's visual memory. The user selects several faces that are similar to their impression of the target face. Based on the selection, the search system can estimate the image representation of the target envisioned by the user and retrieve it from a database.

2. Related Work

2.1 Face Image Retrieval

In general, because face images are taken under different photographic conditions, such as pose, expression, illumination, and occlusion, many stable and highly accurate image retrieval systems for changing environmental conditions have been studied[20)21)22)23)]. Among them, facial contour points are mainly used to compute geometric facial attributes[24)25)]. In this system, a user can manipulate facial landmark positions to retrieve various expressions. However, because these systems focus only on the sparse facial shape, it is difficult to determine or quantify facial characteristics such as gender or impression. Kemelmacher-Shilzerman et al. proposed a real-time system that finds a photograph with a similar facial expression to a given query for application to puppetry[26)]. In this system, the query is the user's own facial expression, such as a smile or frown, and the system automatically retrieves photographs of different persons who have a similar facial expression. After it aligned faces by using 3D template models, it extracted LBP histograms[27)] from face regions for face representations. On the other hand, Kumar et al. employed simple text queries such as "a smiling man with blonde hair and mustache"[28)]. This system learned correspondences between image features and manually defined tags, such as "smiling man" or "blonde hair," by using a support vector machine (SVM). However, because the system can only retrieve face images that have some specific attributes defined in the pre-training process, it is necessary to reconstruct face image descriptors to quantify various facial attributes.

2.2 Deep Image Representation for Content-based Image Retrieval

Many researchers studied face representations generated by deep convolutional neural networks (CNN)[29)30)31)], and they achieved highly accurate verification methods for cropped, incomplete, or occluded face images with the generated face representations. In the face recognition field, using the deep learning has become a state-of-the-art approach, and how bringing out the abilities of them is the key issue[32)33)34)].

Donahue et al. proposed DeCAF[12)], which is a more robust and generic image descriptor compared to con-

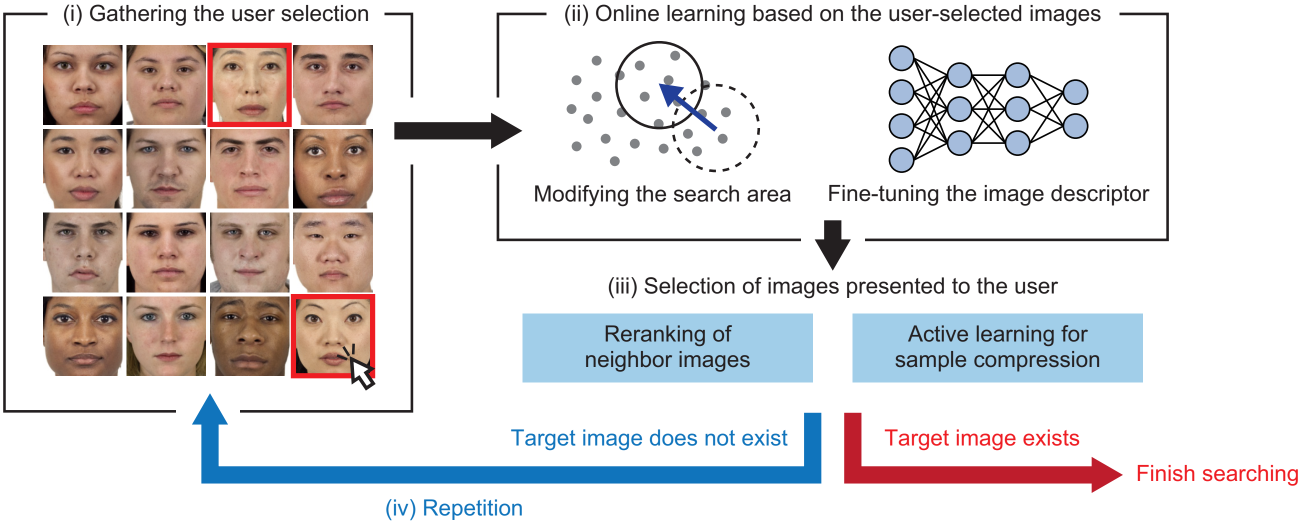

Fig. 2 Flowchart of the proposed retrieval framework. By interactively repeating the user's input and the system's search process, the framework can estimate the target image representation envisioned by the user.

ventional descriptors[35]. They also found that activation features closer to the output layer of the network can describe the semantics of an input image. Lin et al. extended the concept of DeCAF to the retrieval of images of clothes[36]. They utilized a pre-trained network model that had learned rich mid-level visual representations and fine-tuned it using their dataset. It has been reported that applying deep feature representations in a new domain, similarity learning, can significantly boost the retrieval performance. This performance boost is much better than the improvements achieved by "shallow" similarity learning with conventional hand-crafted features[15][37][38]. Therefore, inspired by these methods, we employ DeCAF as a face image descriptor for accurately computing semantic facial similarity.

Zhu et al. proposed the generative visual manipulation model (GVM)[39] to edit images on a natural image manifold and generate a new query image using generative adversarial nets (GAN)[40] for searching. In this search process, a user can manipulate the appearance of retrieval results through hand sketching, including coloring and warping. However, the retrieval performance significantly depends on the quality of the user's sketch as a search query.

2.3 Interactive User Feedback for Concept Learning

In content-based image retrieval, one challenging task is to reflect a user's search intention in retrieval results. To solve this problem, many researchers have attempted to utilize relevance feedback[41][42]. In the field of face retrieval, Wu et al. proposed identity-based quantization using a dictionary constructed using the identities of 270 peoples for large-scale face image retrieval[43]. They improved the precision of local ranking by updating the distance metrics of the top k face representations with user-annotated references.

Our framework is similar to that of CueFlik[44], WhittleSearch[45] and AMNet[46], which manipulate the attributes of retrieval results based a user's input. In a search process, the systems can interactively estimate a search concept by the user's editing of various attributes of the retrieval results based on a comparison with a target image envisioned by the user. However, this process requires a large number of annotated parameters to be provided by the user because of the massive number of items for evaluation. Additionally, these systems are based on the assumption that the user can input queries for initial searching. In contrast, in the present study, we assume a situation in which the user cannot input proper image or text queries, and our system can estimate the representation of a face that the user wishes to find by relying on his/her visual memory.

2.4 Deep Metric Learning

The deep metric learning has also recently been used to modify embedded feature space of a database. For the metric learning, many researchers commonly use a triplet loss model to minimize distance between related images in the embedded feature space and the effectiveness has been reported in some research field such as the person re-identification[47][48][49]. The person re-identification is similar to image retrieval in that both methods are based on similarities between images stored in the database. The key component of the person re-identification is to use a triplet loss for machine learning algorithm. The triplet loss optimization can learn the embedding space such that the data points with same identity are closer to each other than those with different identities[50].

The concept of minimizing the distance between related images is similar to our search process in a database. On the other hand, the conventional triplet mining is computationally expensive because these methods generally optimize the whole embedded feature space of the database. Therefore, for reducing the calculation cost, we propose a method to refer only images in a local search area, and interactively optimize the feature space.

3. Interactive Face Retrieval with Selection of Similar Faces

In this section, we describe a method to interactively retrieve a face image envisioned by a user by relying on the user's visual memory. In our retrieval process, instead of image or text queries, our system requires the user's decisions on whether each facial candidate is similar to the individual the user is searching for; the user's decisions are recorded when they click on images. After pre-processing (see Section 3. 1, 3. 2), based on this user interaction, a search area in a database is modified in the direction of interest and an image descriptor is fine-tuned automatically (see Section 3. 3), and the initial retrieval results are re-ranked based on the estimation of the target image representation (see Section 3. 4).

3. 1 Deep Face Representation

For facial image representations, we use a pre-trained CNN model of the VGG-Face CNN descriptor, which has been trained with a dataset containing 2.6 million face images[33]. This network architecture is based on the VGG-Very-Deep-16 CNN[51], which consists of 16 neural network layers (the first 13 are convolutional layers, and the remaining 3 are fully connected layers). Each convolutional layer includes convolution, rectified linear (ReLU) transform ($f(x) = max(x, 0)$), and max-pooling transform. An input image is abstracted into high-dimensional representations via the convolution layers and pooling layers alternately, and it is connected to the fully connected layers. The fully connected layers focus on the activation maps of the previous layer and determine the features with the strongest correlation to a particular class. To construct a face image database, we first detect the face area in images stored in the database[52] and normalize them to 224 $\times$ 224 pixels. Then, we use activations of the second fully connected layer to extract high-dimensional facial representation vectors (i.e., DeCAF; 4096-dimensional representation vectors) from all database images passed through the VGG-Face network.

3. 2 Indexing for Searching on Large-scale Database

Generally, as the amount of data increases, a retrieval system requires a greater amount of time for computing all similarities between images stored in the database. Therefore, we create search indexes of the facial representation vectors (see Section 3. 1) with the approximate k-nearest neighbor graph (ANNG)[53]. ANNG, which is incrementally constructed with approximate k-nearest-neighbors calculated on a partially constructed graph, is a method for indexing a large-scale database. In addition, the neighborhood graph and tree implementation used for indexing originate from a common library and can perform a similarity search using ANNG, and they have already been applied in several commercial services[54][55]. Given a centroid vector of a search area, ANNG can retrieve k-nearest neighbors based on the cosine distance between their facial representation vectors.

3. 3 Online Learning based on User-selected Images

Modifying Search Area.

In a search process, we first estimate a query vector (i.e., the centroid of a search area in Section 3. 2) based on the images selected by the user. In this process, based on the relevance feedback approach, we estimate the query vector that can retrieve more candidates that are similar to the target image in the feature space of the database constructed in Section 3. 1. For estimating the query vector, we utilize the Rocchio algorithm[41], which is generally used in the exploratory information searching. The algorithm is based on the assumption that most users have a general conception of which information is relevant or irrelevant. This algorithm modifies the vector to separate the relevant and irrelevant vectors maximally by calculating each of their centroids as follows:

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_k \in D_{nr}}\vec{d}_k, \quad (1)$$

where $\vec{q}_m$ is the modified vector, $\vec{q}_0$ is the original vec-

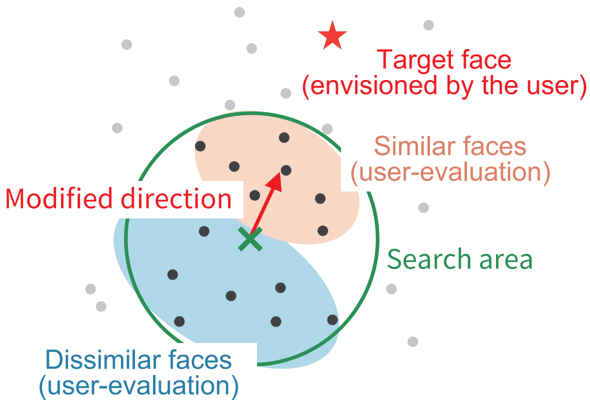

Fig. 3 Modifying a search area with user-evaluated faces and the Rocchio algorithm. The centroid of the search area is moved toward the centroid of faces selected as similar by the user.

tor, D_r is the set of relevant vectors, D_{nr} is the set of irrelevant vectors, and α, β, and γ are weight values (in this paper, we empirically set $\alpha = 1.0$, $\beta = 0.8$, and $\gamma = 0.1$ referring to the previous retrieval work[56]). The centroid of the search area is moved toward the relevant vectors, i.e., those including similar faces, and away from irrelevant vectors, i.e., those including dissimilar faces (see Fig.3). During a search process, by interactively repeating the user's input and the system's search process, we modify the search area and refer to the database images in an exploratory manner.

Fine-tuning Image Descriptor.

In general, the retrieval results depend on the feature representations obtained via the pre-training of the network and are uniquely determined. Thus, there may be a semantic gap between human-based image representations and computer-based image representations. To solve this problem, we dynamically fine-tune the representation parameters by using user feedback for every search iteration. The fine-tuning process is performed with the pre-trained VGG-Face model initialized in the facial representation extraction (see Section 3.1). The network architecture remains unchanged except for the last layer, which is replaced with a new classification layer (i.e., the 2 classes of similar and dissimilar). The activations of the last layer are given to a softmax function, which is expressed as

$$p_k = \frac{\exp(h_k)}{\sum_{j=1}^{K} \exp(h_j)}, \tag{2}$$

where h_k is the k-th activation of the last layer and K is the number of classes; p_k denotes the probability of the k-th class. In the training process, while all the convolutional layers' parameters are fixed, we fine-tune the fully connected layers by using back-propagation. We minimize the cross-entropy error of every training image set. The cross-entropy error is expressed as

$$E = -\sum_{n=1}^{N}\sum_{k=1}^{K} l_{nk} \log p_k, \tag{3}$$

$$l_{nk} = \begin{cases} 1 & \text{(if } n\text{-th image is similar to the target)} \\ 0 & \text{(otherwise)} \end{cases}, \tag{4}$$

where N is the number of the training image set and l_{nk} is the label vector of the n-th training image provided by the user. The error E is minimized by calculating its gradient and optimizing the network parameters by using AdaDelta[57].

3.4 Active Selection

For gathering detailed user feedback, relevance feedback systems generally present more neighbor samples of a search point in a ranking style to the user. However, as the number of proposed samples increases, the process becomes time-consuming and burdensome because the user is required to evaluate all of them. For example, in WhittleSearch[45], it is necessary for the user to observe approximately 50 images while evaluating 18 attributes. In this section, we propose a novel method to decrease the number of samples presented to the user by estimating the image representation of the target envisioned by the user. We call this method "active selection." Concretely, after performing two-class classification learning with the deep convolutional neural network (see Section 3.3), by re-extracting DeCAF, we re-rank the neighbor samples of the search point. Then, instead of presenting all the re-ranked results to the user, we apply an active learning method to low-ranked images for decreasing the number of images presented.

Re-ranking of Neighbor Images.

In a search process, the user can smoothly find a set of images that includes the target face and similar faces if they are placed at the top of the proposed results. Therefore, we re-rank the neighbors to place particular images having representations envisioned by the user at the top position of the retrieval results. After fine-tuning the image descriptor, in this section, we describe the re-ranking of the neighbor images retrieved by ANNG searching. We first re-extract the DeCAF features from k neighbor images with the fine-tuned VGG-Face model, which is the same procedure as in Section 3.1. Then, based on the fine-tuned image representations, we calculate the cosine distance between a query vector, i.e., the centroid of a search area, and each vector of the neighbors. Finally, we define the images that are close to the search point as the top-ranked images.

Active Learning for Sample Compression.

We decrease the number of images presented to the user to reduce the burden on the user for evaluating the proposed images. However, in general, as the number of labeled samples provided by the user decreases, the accuracy of estimation with primitive compression methods (e.g., simply cut the low-ranked images) decreases. Therefore, we propose a novel compression method considering this trade-off relationship. In this paper, we apply the key idea of active learning. The idea of active learning is that a machine learning algorithm can achieve a greater accuracy with fewer training labels if it is allowed to choose the data from which it learns[58]. Namely, active learning can choose images requiring labeling from non-labeled samples for high-accuracy estimation. In this paper, we define the top 30% of the re-ranked results as the top-ranked images and the rest of the images as the low-ranked images. We adopt active learning for low-ranked images, which can choose the images having their class estimated uncertainly by the current trained network model. The images satisfying the requirement defined as follows are chosen from the low-ranked images:

$$\arg\min_{x}(P(y_1|x) - P(y_2|x)), \tag{5}$$

where y_1 and y_2 are the most-probable and second-most-probable class labels (i.e., the similar class or dissimilar class), respectively, and P is the probability of x belonging to the class (so-called the margin sampling method). We pick as many low-ranked images as the number of top-ranked images with the procedure mentioned above. Therefore, we propose active selection, which proposes to the user a mixture of the top-ranked images and the uncertain low-ranked images chosen by active learning (see Fig.4).

4. Experimental Results

4.1 Interface Design

Here, we describe the interface used in the user studies mentioned in the subsequent sections (see Fig.5). To support intuitive browsing, a user can select images on our interface through a drag-and-drop operation. The user can select images that are similar or dissimilar to their impression of the target by dragging and dropping them from the search window (Fig.5: upper-right) to the labeling boxes (Fig.5: left). Note that it is not necessary for the user to select dissimilar images because images that are not selected as similar images are automatically treated as dissimilar. The search process requires at least one face image selected as “similar” so the Rocchio algorithm (equation (1)) can move the search area in the direction of interest (i.e, the face envisioned by the user). In addition, the interface enables the user to modify the labels of images evaluated in the past searching iterations by moving the images to another box in the search process. The bottom-left image is the nearest-neighbor face in the current search iteration (i.e., a top-ranked image among re-ranked neighbor images) used for supplemental information. Based on the top-ranked image, the user can intuitively understand the process of creating face representations via user-labeling.

In these experiments, our retrieval system ran on an Intel Xeon CPU E5-2687W 3.10 GHz with 32 GB RAM and an NVIDIA TITAN X GPU.

4.2 User Study Settings

Database.

For our experiment, we used the Chicago Face Database[59], which consists of 597 face images (290 male and 307 female), as the target database. It provides high-resolution photographs of the participants' frontal pose with neutral expressions. The participants have various nationalities and ethnicities and are between the ages of 17 and 65 years. Since previous works used small-scale databases for experiments (for example, WhittleSearch[45] used 772 images including only 8 persons), we assume that the number of images in the Chicago Face Database is sufficient for confirming the usefulness of our retrieval system.

Methodology.

To assess the utility of our face retrieval system, we conducted user studies. In the user studies, we invited 10 computer science students (20 to 27 years old; 7 male and 3 female). First, each participant was given a brief overview of our interface and a step-by-step tutorial for familiarization with our retrieval framework. Then, we asked them to perform search tasks using our face retrieval interface.

We evaluated the search task for a specific person as well as its features. In this experiment, the participants were shown a single face image that was randomly selected from the database. Then, they searched for the

Fig. 4 Active selection for decreasing the number of images presented to users. The selection includes a mixture of the top-ranked images, and low-ranked images chosen by active learning.

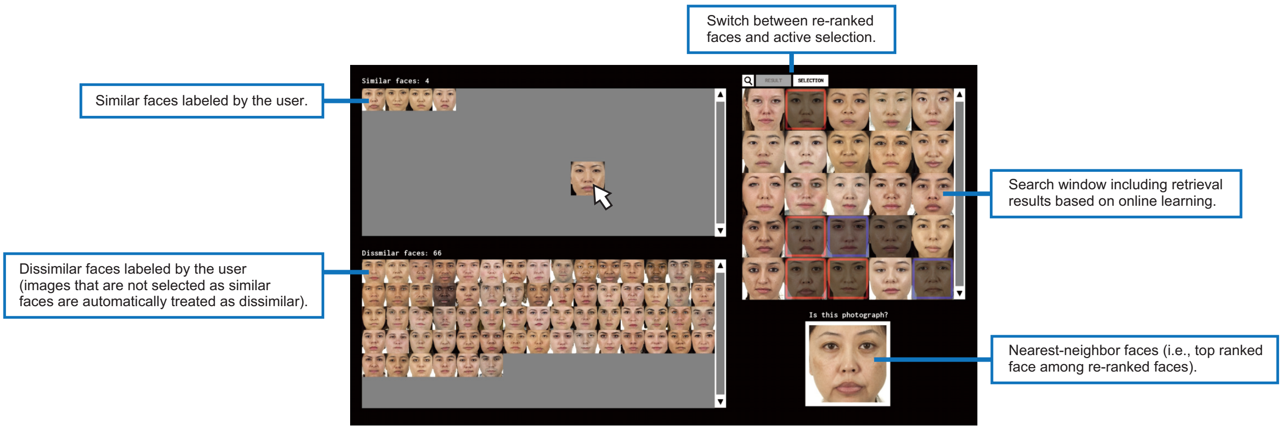

Fig. 5 Proposed user interface. A user can select images to retrieve the target image by interactively repeating the drag-and-drop operation.

person from visual memory without observing further examples (i.e., the participants repeatedly selected several face candidates that were similar to the target face until it was found). The experiment facilitator did not provide a time constraint or intervene unless the participant had difficulty in completing the task. In addition, each experiment was performed once for each participant.

Baseline Creation.

Our goal was to observe whether the participants could independently search for the target face using our system. In addition, since face retrieval relying on a user's visual memory is a new problem; to our knowledge, there has been no existing work on this problem. Thus, in this paper, we assess our framework by changing the contents of images proposed to the user in a search process as follows:

50 neighbors

50 neighbor images of a current search point, i.e., the top 50 images of the original retrieval results of ANNG searching.

25 neighbors

25 neighbor images of a current search point, i.e., the top 25 images of the original retrieval results of ANNG searching.

Active selection

25 images compressed by applying active selection to 50 neighbor images.

The first two are simply cut low-ranked images presented in conventional relevance feedback studies. In this paper, we defined the number of images in active selection as 25 based on the number of images visible on a page of the search window. The images initially proposed to the user were randomly selected.

4.3 Search Cost to Find Target Face

In this section, we report the search cost for a participant to find a specified face using our retrieval interface. We recorded the total search time, total number of search iterations, and frequency of the participants' dragging and dropping until they found the specified face in a search window. Fig.6 shows the obtained scores. In these experiments, even though the proposed framework used only unstable inputs relying on the participants' visual memory, it achieved rapid searching for the target image within 1 min on average (50 neighbors: 118.6 sec (SD 64.0 sec); 25 neighbors: 80.3 sec (SD 44.0 sec); active selection: 58.5 sec (SD 33.0 sec)). Furthermore, the specified image was found with a small number of search iterations on average (50 neighbors: 5.9 (SD 3.1); 25 neighbors: 7.2 (SD 5.4); active selection: 4.5 (SD 3.0)). Because the number of dragging and dropping operations by the participant was reduced as well (50 neighbors: 11.2 (SD 6.9); 25 neighbors: 7.5 (SD 3.3); active selection: 7.0 (SD 2.5)), we also confirmed that active selection could reduce the burden on the user for searching. In summary, each chart provides credible evidence that the proposed active selection method outperforms the baseline method and is effective in searching for a specified face image intuitively and efficiently. This is because there is a critical trade-off relationship between the reduction of user burden by simply removing original retrieval results and the probability of presence of the target or similar faces. Active selection can reduce the strength of this trade-off relationship by re-ranking based on the participant's visual similarities and active learning for sample compression. In addition, active selection also contributes to decreasing the size of the scroll panel used in the search window because of the smaller number of images presented to a user.

4. 4 Efficiency of Exploratory Searching

Here, we report the efficiency of exploratory searching. In this paper, the distance between a search point and the target image point is used for evaluation. At every search iteration, we observed the cosine distance between the centroid vector of a search point and the representation vector of the specified image. Fig.7 shows that the cosine distance converges as the search progresses. At the end of a search process (i.e., at 10th search iteration), the average cosine distance of "50 neighbors" was 0.40 (SD 0.060), "25 neighbors" was 0.47 (SD 0.079), and "Active selection" was 0.37 (SD 0.070). We observed the convergence of the distance in both the baseline methods and active selection. The convergence speed of the distance was higher for active selection than for the baseline methods. The reason for such results was that the participants could effectively evaluate similarities between proposed faces because of a small number of images visible on a search window, and the system correctly modified a search area. In these experiments, some participants were confused in the evaluation of similarities when many images were presented simultaneously (for example, when 50 neighbor face images were presented). Therefore, decreasing the number of images proposed to a user with active selection could result in intuitive searching. Note that the performance of convergence when 25 neighbor images were presented was inferior to that of the other methods because the participants could not find similar images in the small number of proposed images, and the system inefficiently modified the search area.

4. 5 Convergence of Facial Characteristics

As a further extension, instead of a specific face searching, our retrieval framework can be applied to searching for a face cluster that a user prefers in a database. For evaluating convergence of our iterative search suited to the user's preference, we conducted a search task with characteristic keywords related to faces. First, participants were given a keyword that expresses a characteristic of a face. In this experiment, we selected two common keywords, "plump faces" and "pretty faces." Then, each participant searched for face images including the specified characteristic based on their subjective judgment.

During a search process, the participants were asked to count the number of the desired face images with "plump" or "pretty" characteristic in the top-25 retrieval results at every search iteration (see Fig.8). After our iterative search, the average number of "plump faces" was 22.2 (SD 4.17) and "pretty faces" was 14.2 (SD 5.31), respectively. Fig.9 shows an example of this user study results for retrieving "plump faces". Although the initial retrieved faces (1st search) satisfying the user's preference were less than half of the number of presented results, our fine-tuning approach gradually improved the results after the first retrieval (from 2nd to 5th search). These results indicate that our iterative approach can improve results after initial searching. A standard CNN model generates generic representations based on a whole face image (i.e., initial search results), so it is difficult to consider local semantics such as details of facial shape. In contrast, our fine-tuning process can modify and personalize the facial representations based on similarities of relevant faces so that we successfully re-rank retrieval results suited to the user's preference. In addition, at the end of this experiment, the participants were also asked to complete a survey on the degree of their satisfaction against the retrieval results' quality. The answers were scored using a seven-point Likert scale (1: "Extremely dissatisfied" to 7:

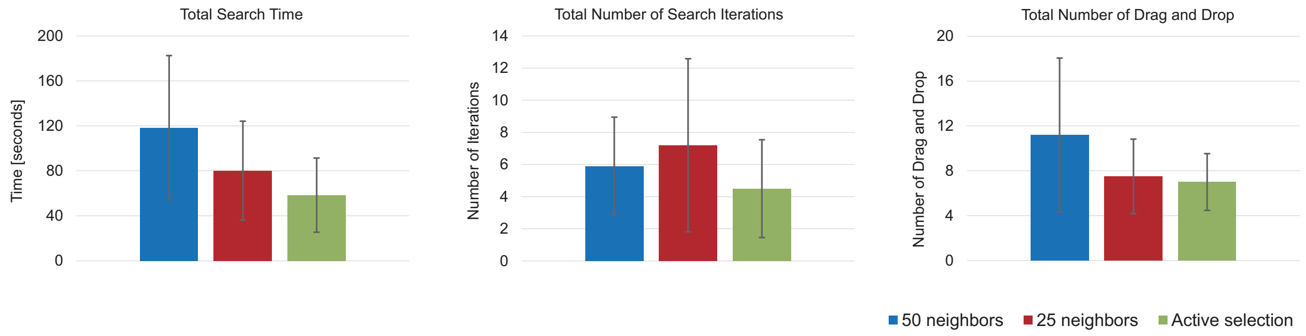

Fig. 6 Average search cost for participants to find a specified face (left: total search time; middle: total number of search iterations; right: frequency of drag and drop). Retrieval of the specified face by using active selection resulted in easy and quick searching.

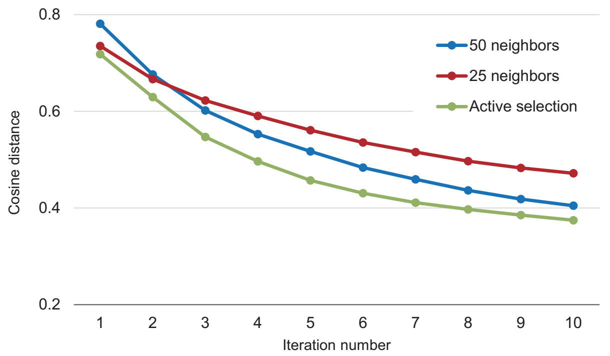

Fig. 7 Average cosine distance between the centroid of a search point and a target face image at every search iteration. With active selection, the distance converges more rapidly.

Table 1 Average score of surveys on the degree of satisfaction against retrieval results' quality (seven-point Likert scale).

	Plump faces	Pretty faces
Average score	6.0	5.2
Standard deviation	0.80	0.76

"Extremely satisfied"). Table 1 shows the average score of the survey results. The questionnaires show highly positive results. The average score of "plump faces" was 6.0 (SD 0.80) and "pretty faces" was 5.2 (SD 0.76) for both "Moderately satisfied" and "Slightly satisfied." Based on the above results, we conclude that the participants were satisfied with the quality of our retrieval results and our iterative search process had converged on their preference.

5. Conclusion and Future Work

In this paper, we proposed a novel framework for clarifying an image representation envisioned by a user. Our retrieval system enables the user to find a target image by relying on his/her visual memory easily and quickly. In addition, we proposed active selection for decreasing the number of presented images. We confirmed that active selection contributed to reducing the burden on the user for efficient exploratory searching. However, some future works are required for improving the proposed framework further.

Initial Presentation.

First, the total number of search iterations and search time may depend on the images presented initially to the user. In this paper, because the images initially presented to the user were randomly selected, the system might provide images that are not similar to the user-desired image. Our framework can flexibly handle such a case by repeating a search process, but further modification of a search area may be required. We plan to solve this problem by initially providing a simple database map showing images in a search space constructed by DeCAF features and ANNG so that the user can easily browse the database overview.

Number of Presented Images.

In our experiment, following previous works, we used a small-scale database and confirmed our retrieval system's usefulness. However, it is necessary to assess the proposed system on a large-scale database such as the LFW Face Database[60] for practical use in many applications. In addition, our framework requires the fine-tuning of the relationship between the scale of the database and the number of presented images.

Individual Differences of Facial Perception.

Our retrieval system strongly depends on a user's ability to recognize and remember human face because the system is based on his/her visual memory only. The variation of each retrieval result shows that there is a difference in facial perception between participants. It is difficult for each participant to remember and evaluate unfamiliar faces. In addition, there is a possibility

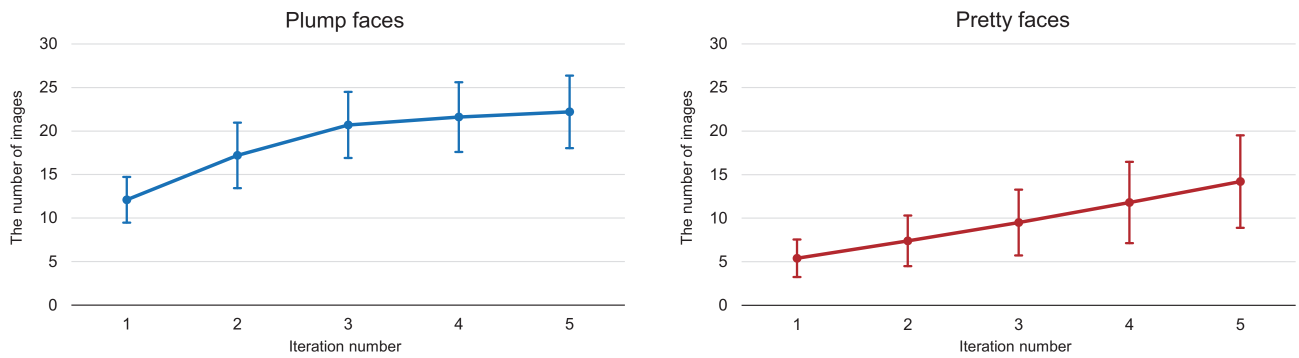

Fig. 8 Average convergence of facial characteristics in a search process for “plump faces” and “pretty faces.” The vertical axis shows the number of results a participant subjectively identified as “plump faces” and “pretty faces,” respectively. The results show that the number of faces suited to the participant’s preference increase each time search iteration.

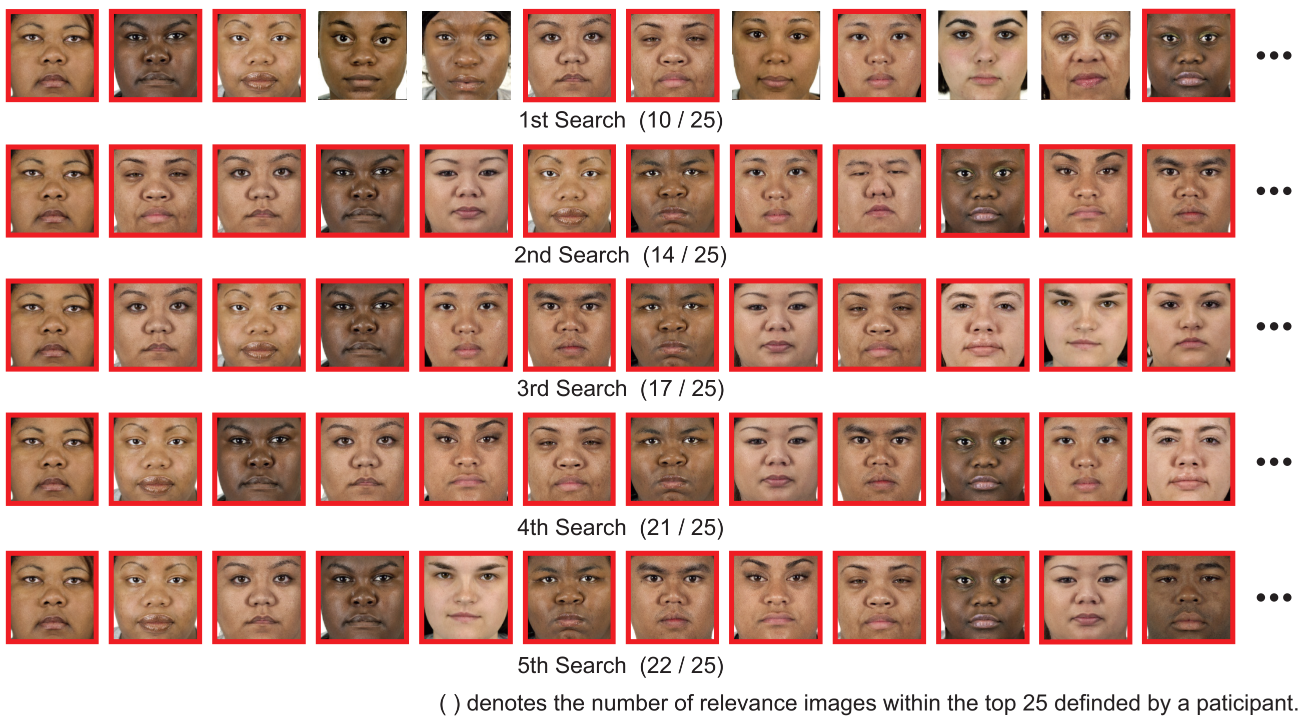

Fig. 9 Example retrieval results for “plump faces” (top-12 face images from left to right). The images with the red frames were selected by participants as “plump.”

of the difference may result from age group and nationality. Therefore, in the future work, we plan to perform a large-scale user study and a task of retrieving familiar faces (e.g., the user’s family and friends) to investigate the difference between participants.

Since the proposed retrieval framework has demonstrated the potential to flexibly meet the demand of various users interactively, it may be useful for some applications such as criminal investigation (for example, a tool to identify criminal suspects based on eyewitness testimony). In addition, because our current system mainly focuses on facial images, the extension of the proposed human-in-the-loop framework (e.g., object recognition or an interface that can augment a user’s individual memory) may present interesting research opportunities, which we plan to explore in the future. We believe that our perception-based framework is a step toward the acceleration of research in the field of human computation.

Acknowledgment

This work was supported in part by JST ACCEL Grant Number JPMJAC1602, Japan.

References

1) Y. Liu, D. Zhang, G. Lu and W.-Y. Ma: "A survey of content-based image retrieval with high-level semantics", Pattern Recognition, 40(1), 262–282 (2007)
2) L. Nixon and R. Troncy: "Survey of semantic media annotation tools for the web: towards new media applications with linked media", In Proc. of the European Semantic Web Conference, pp. 100–114 (2014)
3) S. Sasikala and R. S. Gandhi: "Efficient content based image retrieval system with metadata processing", International Journal of Innovative Research in Science and Technology, 1(10), 72–77 (2015)
4) J. Yu, X. Yang, F. Gao and D. Tao: "Deep multimodal distance metric learning using click constraints for image ranking", IEEE Trans. on Cybernetics (2016)
5) F. Ensan and E. Bagheri: "Document retrieval model through semantic linking", In Proc. of the Tenth ACM International Conference on Web Search and Data Mining, pp. 181–190 (2017)
6) X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek and A. D. Bimbo: "Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval", ACM Computing Surveys, 49(1), 14 (2016)
7) J. Han and K.-K. Ma: "Fuzzy color histogram and its use in color image retrieval", IEEE Trans. on Image Processing, 11(8), 944–952 (2002)
8) D. Zhang, A. Wong, M. Indrawan and G. Lu: "Content-based image retrieval using gabor texture features", IEEE Trans. on Pattern Analysis and Machine Intelligence, pp. 13–15 (2000)
9) C. Wang, L. Zhang and H.-J. Zhang: "Learning to reduce the semantic gap in web image retrieval and annotation", In Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 355–362 (2008)
10) H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao and T.-S. Chua: "Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval", In Proc. of the 21st ACM International Conference on Multimedia, pp. 33–42 (2013)
11) B.-C. Chen, Y.-Y. Chen, Y.-H. Kuo and W. H. Hsu: "Scalable face image retrieval using attribute-enhanced sparse codewords", IEEE Trans. on Multimedia, 15(5), 1163–1173 (2013)
12) J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng and T. Darrell: "Decaf: A deep convolutional activation feature for generic visual recognition.", In Proc. of the 31st International Conference on Machine Learning, vol. 32, pp. 647–655 (2014)
13) A. Babenko, A. Slesarev, A. Chigorin and V. Lempitsky: "Neural codes for image retrieval", In Proc. of the European Conference on Computer Vision, pp. 584–599 (2014)
14) F. Zhao, Y. Huang, L. Wang and T. Tan: "Deep semantic ranking based hashing for multi-label image retrieval", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1556–1564 (2015)
15) J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang and J. Li: "Deep learning for content-based image retrieval: A comprehensive study", In Proc. of the 22nd ACM International Conference on Multimedia, pp. 157–166 (2014)
16) E. Cheng, F. Jing and L. Zhang: "A unified relevance feedback framework for web image retrieval", IEEE Trans. on Image Processing, 18(6), 1350–1357 (2009)
17) Z. Ji, Y. Pang and X. Li: "Relevance preserving projection and ranking for web image search reranking", IEEE Trans. on Image Processing, 24(11), 4137–4147 (2015)
18) Y. Zhang, X. Yang and T. Mei: "Image search reranking with query-dependent click-based relevance feedback", IEEE Trans. on Image Processing, 23(10), 4448–4459 (2014)
19) Y. Sato, T. Fukusato and S. Morishima: "Face retrieval framework relying on user's visual memory", In Proc. of the ACM International Conference on Multimedia Retrieval, pp. 274–282 (2018)
20) C. Herrmann and J. Beyerer: "Face retrieval on large-scale video data", In Proc. of the 12th Conference on Computer and Robot Vision, pp. 192–199 (2015)
21) S. Kayal: "Improved hierarchical clustering for face images in videos: Integrating positional and temporal information with hac", In Proc. of the International Conference on Multimedia Retrieval, p. 455 (2014)
22) E. G. Ortiz, A. Wright and M. Shah: "Face recognition in movie trailers via mean sequence sparse representation-based classification", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3531–3538 (2013)
23) B. Wu, Y. Zhang, B.-G. Hu and Q. Ji: "Constrained clustering and its application to face clustering in videos", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3507–3514 (2013)
24) B. M. Smith, S. Zhu and L. Zhang: "Face image retrieval by shape manipulation", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 769–776 (2011)
25) S. Zhu, B. M. Smith and L. Zhang: "Facesimile: A mobile application for face image search based on interactive shape manipulation", In Proc. of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 82–83 (2011)
26) I. Kemelmacher-Shlizerman, E. Shechtman, R. Garg and S. M. Seitz: "Exploring photobios", In ACM Trans. on Graphics (TOG), vol. 30(4), p. 61 (2011)
27) T. Ojala, M. Pietikainen and D. Harwood: "Performance evaluation of texture measures with classification based on kullback discrimination of distributions", Pattern Recognition, 1, 582–585 (1994)
28) N. Kumar, A. Berg, P. N. Belhumeur and S. Nayar: "Describable visual attributes for face verification and image search", IEEE Trans. on Pattern Analysis and Machine Intelligence, 33(10), 1962–1977 (2011)
29) Y. Sun, X. Wang and X. Tang: "Deeply learned face representations are sparse, selective, and robust", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2892–2900 (2015)
30) F. Schroff, D. Kalenichenko and J. Philbin: "Facenet: A unified embedding for face recognition and clustering", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
31) Y. Wen, K. Zhang, Z. Li and Y. Qiao: "A discriminative feature learning approach for deep face recognition", In Proc. of the European Conference on Computer Vision, pp. 499–515 (2016)
32) W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj and L. Song: "Sphereface: Deep hypersphere embedding for face recognition", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, p. 1 (2017)
33) O. M. Parkhi, A. Vedaldi and A. Zisserman: "Deep face recognition", In Proc. of the British Machine Vision Conference, vol. 1(3), p. 6 (2015)
34) Y. Taigman, M. Yang, M. Ranzato and L. Wolf: "Deepface: Closing the gap to human-level performance in face verification", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
35) C. Celik and H. S. Bilge: "Content based image retrieval with sparse representations and local feature descriptors: a comparative study", Pattern Recognition, 68, 1–13 (2017)
36) K. Lin, H.-F. Yang, K.-H. Liu, J.-H. Hsiao and C.-S. Chen: "Rapid clothing retrieval via deep learning of binary codes and hierarchical search", In Proc. of the 5th Conference on Multimedia Retrieval, pp. 499–502 (2015)
37) A. Gordo, J. Almazán, J. Revaud and D. Larlus: "Deep image retrieval: Learning global representations for image search", In Proc. of the European Conference on Computer Vision, pp. 241–257 (2016)
38) F. Wang, L. Kang and Y. Li: "Sketch-based 3d shape retrieval using convolutional neural networks", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1875–1883 (2015)
39) J.-Y. Zhu, P. Krähenbühl, E. Shechtman and A. A. Efros: "Generative visual manipulation on the natural image manifold", In Proc. of the European Conference on Computer Vision, pp. 597–613 (2016)
40) I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio: "Generative adversarial nets", In Proc. of the Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
41) J. Rocchio: "Relevance feedback in information retrieval", The Smart Retrieval System-Experiments in Automatic Document Processing, pp. 313–323 (1971)
42) Y. Rui, T. S. Huang, M. Ortega and S. Mehrotra: "Relevance feedback: a power tool for interactive content-based image retrieval", IEEE Trans. on Circuits and Systems for Video Technology, 8(5), 644–655 (1998)

43) Z. Wu, Q. Ke, J. Sun and H.-Y. Shum: "Scalable face image retrieval with identity-based quantization and multireference reranking", IEEE Trans. on Pattern Analysis and Machine Intelligence, 33(10), 1991–2001 (2011)
44) J. Fogarty, D. Tan, A. Kapoor and S. Winder: "Cueflik: interactive concept learning in image search", In Proc. of the SIGCHI Conference on Human Factors in Computing Systems, pp. 29–38 (2008)
45) A. Kovashka, D. Parikh and K. Grauman: "Whittlesearch: Image search with relative attribute feedback", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2973–2980 (2012)
46) B. Zhao, J. Feng, X. Wu and S. Yan: "Memory-augmented attribute manipulation networks for interactive fashion search", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1520–1528 (2017)
47) H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng and S. Z. Li: "Embedding deep metric for person re-identification: A study against large variations", In Proc. of the European Conference on Computer Vision, pp. 732–748 (2016)
48) E. Ustinova and V. Lempitsky: "Learning deep embeddings with histogram loss", In Proc. of the Advances in Neural Information Processing Systems, pp. 4170–4178 (2016)
49) W. Chen, X. Chen, J. Zhang and K. Huang: "Beyond triplet loss: a deep quadruplet network for person re-identification", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2 (2017)
50) A. Hermans, L. Beyer and B. Leibe: "In defense of the triplet loss for person re-identification", arXiv preprint arXiv:1703.07737 (2017)
51) K. Simonyan and A. Zisserman: "Very deep convolutional networks for large-scale image recognition", ArXiv Preprint ArXiv:1409.1556 (2014)
52) D. E. King: "Dlib-ml: A machine learning toolkit", Journal of Machine Learning Research, 10(Jul), 1755–1758 (2009)
53) M. Iwasaki: "Pruned bi-directed k-nearest neighbor graph for proximity search", In Proc. of the International Conference on Similarity Search and Applications, pp. 20–33 (2016)
54) M. Iwasaki: "Ngt: Neighborhood graph and tree for indexing", http://research-lab.yahoo.co.jp/software/ngt/ (2015)
55) K. Sugawara, H. Kobayashi and M. Iwasaki: "On approximately searching for similar word embeddings", In Proc. of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 2265–2275 (2016)
56) D. Markonis, R. Schaer and H. Müller: "Multi-modal relevance feedback for medical image retrieval", In Proc. of the MedIR@ SIGIR, pp. 20–23 (2014)
57) M. D. Zeiler: "Adadelta: an adaptive learning rate method", ArXiv Preprint ArXiv:1212.5701 (2012)
58) B. Settles: "Active learning literature survey", University of Wisconsin, Madison, 52(55-66), 11 (2010)
59) D. S. Ma, J. Correll and B. Wittenbrink: "The chicago face database: A free stimulus set of faces and norming data", Behavior Research Methods, 47(4), 1122–1135 (2015)
60) G. B. Huang, M. Ramesh, T. Berg and E. Learned-Miller: "Labeled faces in the wild: A database for studying face recognition in unconstrained environments", Technical Report 07-49, University of Massachusetts, Amherst (2007)

Yugo Sato received the B.E. degree in Department of Applied Physics and the M.E. degree in Department of Pure and Applied Physics from Waseda University, Tokyo, Japan, in 2016 and 2018, respectively. He joined Sony Corporation, Tokyo, Japan, in 2018, where he has been engaged in the research and development of vision system technology.

Tsukasa Fukusato received the B.E, M.E., and Ph.D. in the Department of Pure and Applied Physics from Waseda University in 2012, 2014, and 2017, respectively. He is currently an Assistant Professor at the Graduate School of Information Science and Technology in the University of Tokyo, where he is a member of User Interface Research Group.

Shigeo Morishima received the B.S., M.S. and Ph.D. degrees, all in Electrical Engineering from the University of Tokyo, Tokyo, Japan, in 1982, 1984, and 1987, respectively. From 1987 to 2001, he was an associate professor and from 2001 to 2004, a professor of Seikei University, Tokyo. Currently, he is a professor of School of Advanced Science and Engineering, Waseda University.