Face Retrieval Framework Relying on User's Visual Memory

Yugo Sato
Waseda University
y.sato9372@fuji.waseda.jp

Tsukasa Fukusato
The University of Tokyo
tsukasafukusato@is.s.u-tokyo.ac.jp

Shigeo Morishima
Waseda Research Institute for Science and Engineering
shigeo@waseda.jp

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 (or a single image) 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 (human-in-the-loop optimization), 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 subjects on a public database and confirmed that the proposed framework is effective for clarifying the image representation envisioned by the user easily and quickly.

CCS CONCEPTS

• **Information systems** → *Users and interactive retrieval*;

KEYWORDS

User interaction; Deep convolutional neural network; Relevance feedback; Active learning

ACM Reference Format:
Yugo Sato, Tsukasa Fukusato, and Shigeo Morishima. 2018. Face Retrieval Framework Relying on User's Visual Memory. In *Proceedings of 2018 International Conference on Multimedia Retrieval (ICMR'18).* June 11–14, 2018, Yokohama, Japan, 9 pages. https://doi.org/10.1145/3206025.3206038

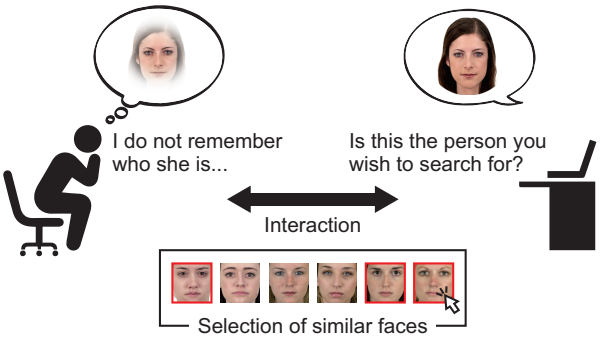

Figure 1: Face retrieval relying on the user's visual memory. The user selects several faces that are similar to their impression of the target face. Based on the selection, the search system can estimate the image representation of the target envisioned by the user and retrieve it from a database.

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 [22, 24, 31]. 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 [5, 20, 45]. 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 [9] or Gabor texture features [47]. 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 [2, 39, 48].

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 [4]. The image representations obtained

ICMR'18, June 11–14, 2018, Yokohama, Japan

ACM ISBN 978-1-4503-5046-4/18/06…$15.00
https://doi.org/10.1145/3206025.3206038

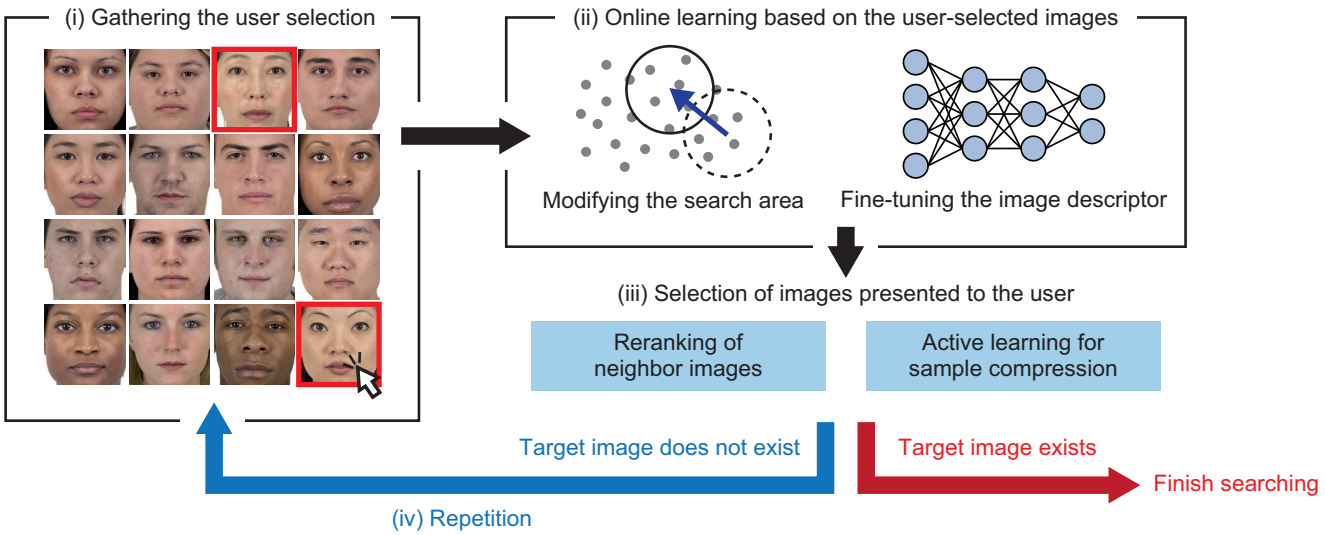

Figure 2: Flowchart of the proposed retrieval framework. By interactively repeating the user's input and the system's search process, the framework can estimate the target image representation envisioned by the user.

through deep learning consistently outperform conventional hand-crafted features and boost the image retrieval performance [1, 38, 51]. 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 the "relevance feedback" approach, which allows the user to interactively refine retrieval results [3, 14, 49]. 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 [3]. 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 Figure 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. Figure 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 the 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.

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 [10, 15, 27, 43]. Among them, facial contour points are mainly used to compute geometric facial attributes [35, 53]. 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 [16]. 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 [25] 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" [19]. 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

Deep learning for image analysis has been mainly studied in the field of computer vision. Many researchers studied face representations generated by deep convolutional neural networks (CNNs) [32, 37, 42], and they achieved highly accurate verification methods for cropped, incomplete, or occluded face images with the generated face representations. Donahue et al. proposed DeCAF [4], which is a more robust and generic image descriptor compared to conventional descriptors such as GIST [26] and LLC [41]. 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 [21]. 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 [8, 38, 40]. 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) [52] to edit images on a natural image manifold and generate a new query image using generative adversarial nets (GAN) [7] 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 [29, 30]. 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 [44]. 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 [6], WhittleSearch [18] and AMNet [50], which manipulate the attributes of retrieval results based a user's input. In the 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.

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, the 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 [28]. This network architecture is based on the VGG-Very-Deep-16 CNN [34], 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 [17] and normalize them to 224×224 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) [13]. 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 [12, 36]. Given a centroid vector of a search area, ANNG can retrieve k-nearest neighbors based on the cosine similarity between their facial representation vectors.

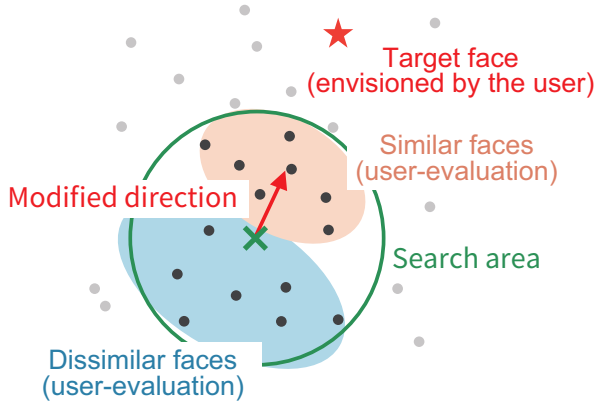

Figure 3: Modifying a search area with user-evaluated faces and the Rocchio algorithm. The centroid of the search area is moved toward the centroid of faces selected as similar by the user.

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 [29], 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:

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

where $\vec{q}_m$ is the modified vector, $\vec{q}_0$ is the original vector, 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 set $\alpha = 1.0$, $\beta = 0.8$, and $\gamma = 0.1$). 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 Figure 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 [46].

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 [18], 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 the 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

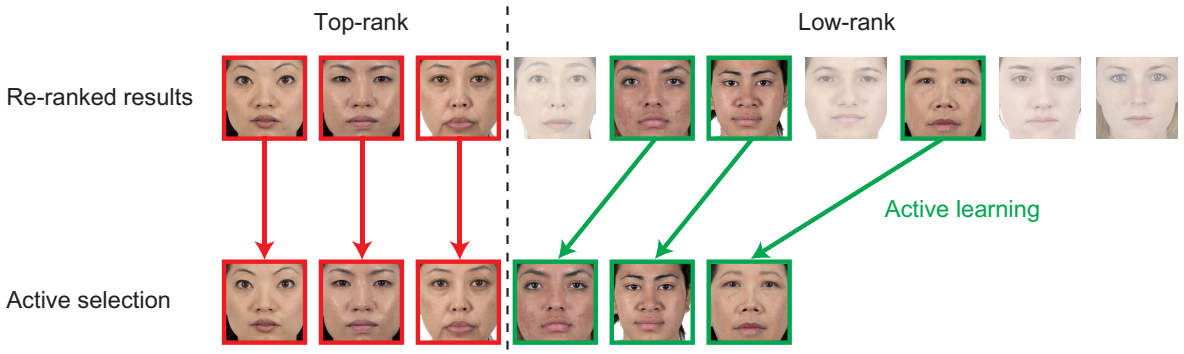

Figure 4: Active selection for decreasing the number of images presented to users. The selection includes a mixture of the top-ranked images, and low-ranked images chosen by active learning.

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 [33]. 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 Figure 4).

4 EXPERIMENTAL RESULTS

4.1 Interface Design

Here, we describe the interface used in the user studies mentioned in the subsequent sections (see Figure 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 (Figure 5: upper-right) to the labeling boxes (Figure 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. 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 [23], which consists of 597 face images (290 male and 307 female), as the target database. It provides high-resolution photographs of the subjects' frontal pose with neutral expressions. The subjects 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 [18] 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 performed user studies. In the user studies, we invited 10 computer science

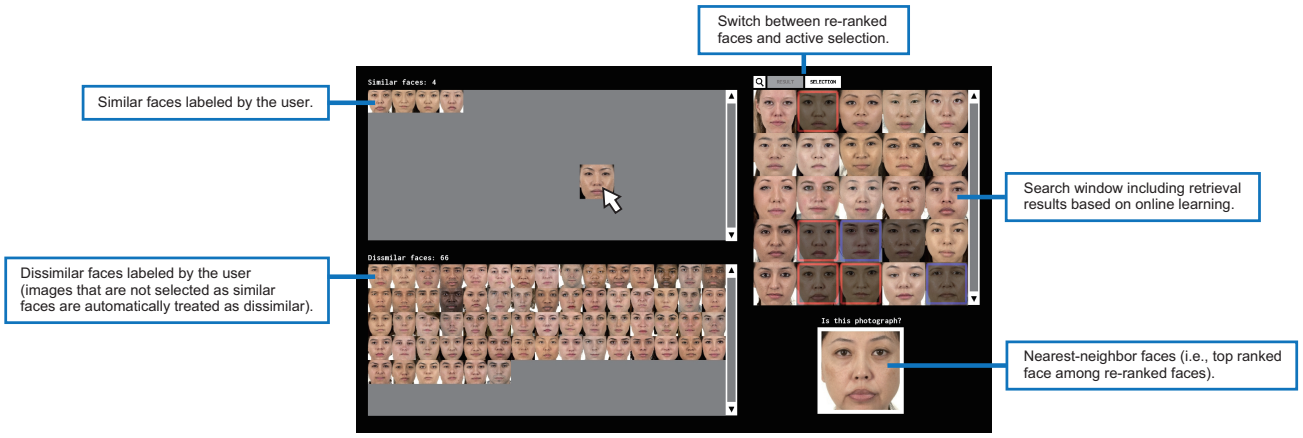

Figure 5: Proposed user interface. A user can select images to retrieve the target image by interactively repeating the drag-and-drop operation.

students (20 to 27 years old; 7 male and 3 female). First, each subject 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 subjects were shown a single face image that was randomly selected from the database. Then, they searched for the person from visual memory without observing further examples (i.e., the subjects 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 subject had difficulty in completing the task.

Baseline.
Our goal was to observe whether the subjects 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 subject to find a specified face using our retrieval interface. We recorded the total search time, total number of search iterations, and frequency of the subjects' dragging and dropping until they found the specified face in a search window. Figure 6 shows the obtained scores. In these experiments, even though the proposed framework used only unstable inputs relying on the subjects' visual memory, it achieved rapid searching for the target image within 1 min on average (50 neighbors: 118.6 s; 25 neighbors: 80.3 s; active selection: 58.5 s). Furthermore, the specified image was found with a small number of search iterations on average (50 neighbors: 5.9; 25 neighbors: 7.2; active selection: 4.5). Because the number of dragging and dropping operations by the subject was reduced as well (50 neighbors: 11.2; 25 neighbors: 7.5; active selection: 7.0), 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 subject'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. Figure 7 shows

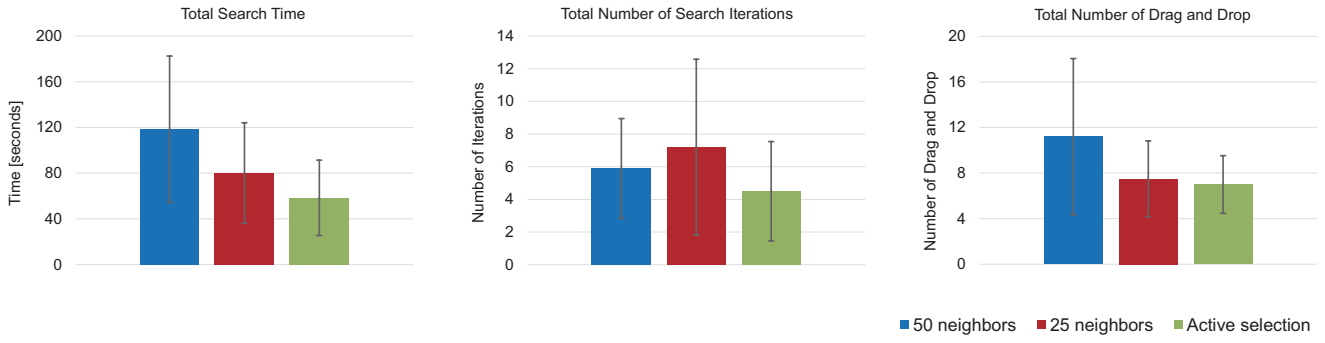

Figure 6: Average search cost for a subject to find a specified face (left: total search time; middle: total number of search iterations; right: frequency of drag and drop). Retrieval of the specified face by using active selection resulted in easy and quick searching.

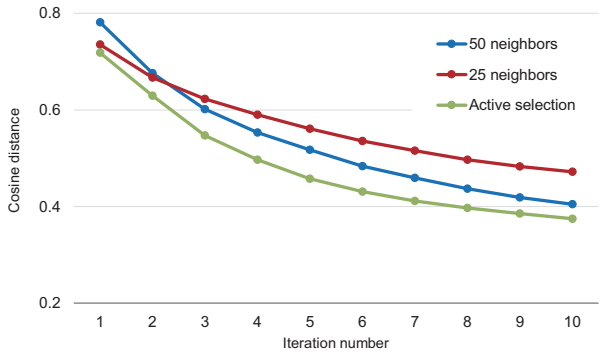

Figure 7: Cosine distance between the centroid of a search point and a target face image at every search iteration. With active selection, the distance converges more rapidly.

that the cosine distance converges as the search progresses. 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 subjects could effectively evaluate similarities between proposed faces because of the small number of images visible on a search window, and the system correctly modified the search area. In these experiments, some subjects 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 subjects could not find similar images in the small number of proposed images, and the system inefficiently modified the search area.

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 target. Our framework can flexibly handle such a case by repeating the search process, but further modification of the 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 [11] 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.

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. 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 memories) 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.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. 2014. Neural codes for image retrieval. In *Proc. of the European Conference on Computer Vision*. Springer, 584–599.

[2] Bor-Chun Chen, Yan-Ying Chen, Yin-Hsi Kuo, and Winston H Hsu. 2013. Scalable face image retrieval using attribute-enhanced sparse codewords. *IEEE Trans. on Multimedia* 15, 5 (2013), 1163–1173.

[3] En Cheng, Feng Jing, and Lei Zhang. 2009. A unified relevance feedback framework for web image retrieval. *IEEE Trans. on Image Processing* 18, 6 (2009), 1350–1357.

[4] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.. In *Proc. of the 31st International Conference on Machine Learning*, Vol. 32. 647–655.

[5] Faezeh Ensan and Ebrahim Bagheri. 2017. Document Retrieval Model Through Semantic Linking. In *Proc. of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 181–190.

[6] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: interactive concept learning in image search. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 29–38.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proc. of the Advances in neural information processing systems*. 2672–2680.

[8] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. 2016. Deep image retrieval: Learning global representations for image search. In *Proc. of the European Conference on Computer Vision*. Springer, 241–257.

[9] Ju Han and Kai-Kuang Ma. 2002. Fuzzy color histogram and its use in color image retrieval. *IEEE Trans. on Image Processing* 11, 8 (2002), 944–952.

[10] Christian Herrmann and Jürgen Beyerer. 2015. Face Retrieval on Large-Scale Video Data. In *Proc. of the 12th Conference on Computer and Robot Vision (CRV)*. IEEE, 192–199.

[11] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Technical Report 07-49, University of Massachusetts, Amherst* (2007).

[12] Masajiro Iwasaki. 2015. NGT: Neighborhood Graph and Tree for Indexing. http://research-lab.yahoo.co.jp/software/ngt/. (2015).

[13] Masajiro Iwasaki. 2016. Pruned Bi-directed K-nearest Neighbor Graph for Proximity Search. In *Proc. of the International Conference on Similarity Search and Applications*. Springer, 20–33.

[14] Zhong Ji, Yanwei Pang, and Xuelong Li. 2015. Relevance preserving projection and ranking for Web image search reranking. *IEEE Trans. on Image Processing* 24, 11 (2015), 4137–4147.

[15] Subhradeep Kayal. 2014. Improved Hierarchical Clustering for Face Images in Videos: Integrating positional and temporal information with HAC. In *Proc. of the International Conference on Multimedia Retrieval*. ACM, 455.

[16] Ira Kemelmacher-Shlizerman, Eli Shechtman, Rahul Garg, and Steven M Seitz. 2011. Exploring photobios. In *ACM Trans. on Graphics (TOG)*, Vol. 30. ACM, 61.

[17] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10, Jul (2009), 1755–1758.

[18] Adriana Kovashka, Devi Parikh, and Kristen Grauman. 2012. Whittlesearch: Image search with relative attribute feedback. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2973–2980.

[19] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. 2011. Describable visual attributes for face verification and image search. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 33, 10 (2011), 1962–1977.

[20] Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees GM Snoek, and Alberto Del Bimbo. 2016. Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Computing Surveys (CSUR)* 49, 1 (2016), 14.

[21] Kevin Lin, Huei-Fang Yang, Kuan-Hsien Liu, Jen-Hao Hsiao, and Chu-Song Chen. 2015. Rapid clothing retrieval via deep learning of binary codes and hierarchical search. In *Proc. of the 5th Conference on Multimedia Retrieval*. ACM, 499–502.

[22] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. 2007. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition* 40, 1 (2007), 262–282.

[23] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods* 47, 4 (2015), 1122–1135.

[24] Lyndon Nixon and Raphaël Troncy. 2014. Survey of semantic media annotation tools for the web: towards new media applications with linked media. In *Proc. of the European Semantic Web Conference*. Springer, 100–114.

[25] Timo Ojala, Matti Pietikainen, and David Harwood. 1994. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. *Pattern Recognition* 1 (1994), 582–585.

[26] Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42, 3 (2001), 145–175.

[27] Enrique G Ortiz, Alan Wright, and Mubarak Shah. 2013. Face recognition in movie trailers via mean sequence sparse representation-based classification. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 3531–3538.

[28] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. In *Proc. of the British Machine Vision Conference*, Vol. 1. 6.

[29] Joseph Rocchio. 1971. Relevance feedback in information retrieval. *The Smart Retrieval System-Experiments in Automatic Document Processing* (1971), 313–323.

[30] Yong Rui, Thomas S Huang, Michael Ortega, and Sharad Mehrotra. 1998. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Trans. on Circuits and Systems for Video Technology* 8, 5 (1998), 644–655.

[31] S Sasikala and R Soniya Gandhi. 2015. Efficient content based image retrieval system with metadata processing. *International Journal of Innovative Research in Science and Technology* 1, 10 (2015), 72–77.

[32] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 815–823.

[33] Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison* 52, 55-66 (2010), 11.

[34] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556* (2014).

[35] Brandon M Smith, Shengqi Zhu, and Li Zhang. 2011. Face image retrieval by shape manipulation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 769–776.

[36] Kohei Sugawara, Hayato Kobayashi, and Masajiro Iwasaki. 2016. On approximately searching for similar word embeddings. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 2265–2275.

[37] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2015. Deeply learned face representations are sparse, selective, and robust. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 2892–2900.

[38] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. 2014. Deep learning for content-based image retrieval: A comprehensive study. In *Proc. of the 22nd ACM international conference on Multimedia*. ACM, 157–166.

[39] Changhu Wang, Lei Zhang, and Hong-Jiang Zhang. 2008. 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*. ACM, 355–362.

[40] Fang Wang, Le Kang, and Yi Li. 2015. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 1875–1883.

[41] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. 2010. Locality-constrained linear coding for image classification. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3360–3367.

[42] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *Proc. of the European Conference on Computer Vision*. Springer, 499–515.

[43] Baoyuan Wu, Yifan Zhang, Bao-Gang Hu, and Qiang Ji. 2013. Constrained clustering and its application to face clustering in videos. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 3507–3514.

[44] Zhong Wu, Qifa Ke, Jian Sun, and Heung-Yeung Shum. 2011. Scalable face image retrieval with identity-based quantization and multireference reranking. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 33, 10 (2011), 1991–2001.

[45] Jun Yu, Xiaokang Yang, Fei Gao, and Dacheng Tao. 2016. Deep multimodal distance metric learning using click constraints for image ranking. *IEEE Trans. on Cybernetics* (2016).

[46] Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. *ArXiv Preprint ArXiv:1212.5701* (2012).

[47] Dengsheng Zhang, Aylwin Wong, Maria Indrawan, and Guojun Lu. 2000. Content-based image retrieval using Gabor texture features. *IEEE Trans. on Pattern Analysis and Machine Intelligence* (2000), 13–15.

[48] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2013. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *Proc. of the 21st ACM international conference on Multimedia*. ACM, 33–42.

[49] Yongdong Zhang, Xiaopeng Yang, and Tao Mei. 2014. Image search reranking with query-dependent click-based relevance feedback. *IEEE Trans. on Image Processing* 23, 10 (2014), 4448–4459.

[50] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. 2017. Memory-augmented attribute manipulation networks for interactive fashion search. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 1520–1528.

[51] Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan. 2015. Deep semantic ranking based hashing for multi-label image retrieval. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 1556–1564.

[52] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. 2016. Generative visual manipulation on the natural image manifold. In *Proc. of the European Conference on Computer Vision*. Springer, 597–613.

[53] Shengqi Zhu, Brandon M Smith, and Li Zhang. 2011. 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)*. IEEE, 82–83.