FaceFetch : An Efficient and Scalable Face Retrieval System that uses your Visual Memory

Harsh Shrivastava, P V N S Rama Krishna, Karmanya Aggarwal, Meghna P Ayyar, Yifang Yin, Rajiv Ratn Shah and Roger Zimmermann

November 19, 2019

FACEFETCH : AN EFFICIENT AND SCALABLE FACE RETRIEVAL SYSTEM THAT USES YOUR VISUAL MEMORY

Harsh Shrivastava[1], *Rama Krishna P V N S*[1], *Karmanya Aggarwal*[1], *Meghna P Ayyar*[1], *Yifang Yin*[2], *Rajiv Ratn Shah*[1], *Roger Zimmermann*[2]

MIDAS Lab, IIIT-Delhi[1], National University of Singapore, Singapore[2]

ABSTRACT

Often in many situations in our life, we wish to envision the person we met but we could not recall what the person exactly look like except for a slight impression of the face. Yugo Sato et.al. introduced a face retrieval system for this problem, which utilises visual inputs from the users and attempts to retrieve the target face. The major drawback of their approach was that their system was slow and only applicable for small databases like Chicago Face Database. In this paper, we introduce a robust and scalable face retrieval system that is capable of retrieving the envisioned face from a large-scale database. Furthermore, instead of information specific to the target, our face retrieval system asks the user to select common face attributes that they remember their target face had, using which the system filters out the irrelevant faces thus speeding up the search process. Then our system asks the user to select several images that resembles with the envisioned face. On the basis of this selection, our system automatically reduces the "semantic gap" between human description and the computer based description of the target image. In order to evaluate our system, We conducted user studies on a large-scale database and established that our framework succeeded in beating the state of the art results in this particular task and thus proved itself to be very effective for retrieving the envisioned face image in approximately half the total number of search iterations and taking one-third of the overall search time thereby putting much less burden on the user.

***Index Terms*—** Face Retrieval, Relevance Feedback, Deep Convolutional Neural Network, Active Learning, User Interaction

1. INTRODUCTION

Due to the recent advancements in technology, increasing popularity of various digital applications like cameras, video recorders, *etc.* and photo sharing social media networks (such as Facebook, Instagram and Flickr), a large number of photos are being generated and circulated across the internet. A large fraction of these photos have human faces. The sheer amount of these large-scale human face images and their importance make it really crucial to develop systems that are capable of retrieving (*e.g.,* searching and mining) them from large-scale networks. Since these social media networks has large-scale databases of the images, the problem of retrieving a desired image from them has risen as a very critical research area. The research in this area has enabled many useful applications as well [1, 2, 3, 4, 5, 6, 7, 8]. Hence, there is a rapidly growing need for fast and efficient systems that can retrieve these images from such huge volumes of data. Internet companies like Google, Baidu and Bing utilise numerous meta-data information such as file names, file formats, text on web pages and image captions. Though these systems might have succeeded in the tasks relating to document retrieval, they largely rely on the textual information. Also, these associated tags often fail in distinguishing the image. Moreover, these hand-coded tags can negatively impact the retrieval performance [9, 10, 11]. In addition to that, if the user is seeking an image with visual attributes that is impossible to be easily represented by the tags or the keywords, the user would be forced to go through a large number of retrieved images, in search of the target image. Thus, these text based query methods are not efficient and effective.

Visual features-based retrieval systems have performed better than the tags-based systems [1]. Various primitive features *i.e.,* colour, structure, shape, texture *etc.* can be used to extract the similarities between the images that are already stored in the database. Content Based Image Retrieval (CBIR), which makes use of visual contents which are primitive features, can also be used. Several classical features like Gabor [12] and HOG [13] can be used as these low-level features.

Recently, the world has seen the rise of deep learning, a set of advanced machine learning algorithms called Artificial Neural Networks (ANN). These are deep architectures of non-linear transformations stacked one over other, which attempt to capture the highly complex patterns in data. Convolutional Neural Networks are one of classes of ANNs which models high level features in images. Given a large-scale database for training, a deep convolutional neural networks can produce generic image representations and

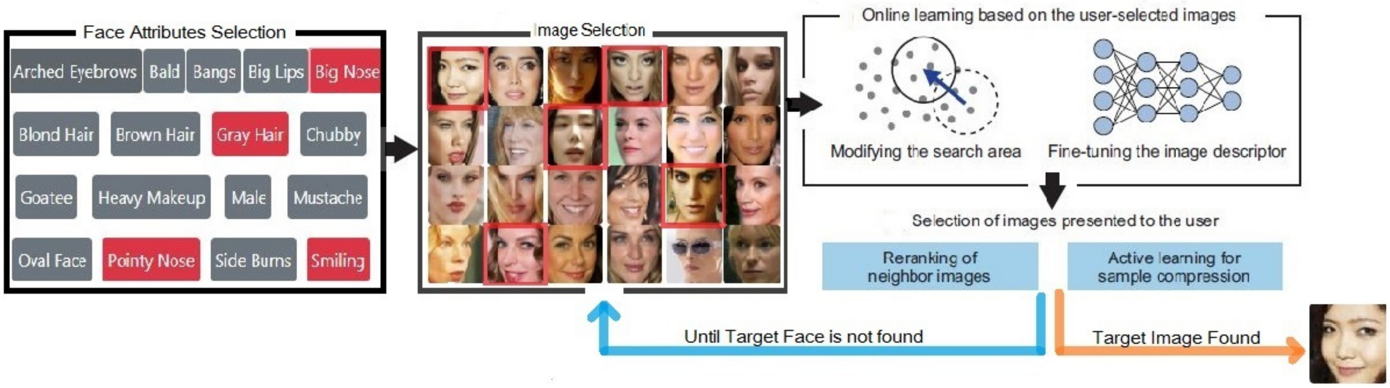

Face Attributes Selection
Arched Eyebrows
Bald
Bangs
Big Lips
Big Nose
Blond Hair
Brown Hair
Gray Hair
Chubby
Goatee
Heavy Makeup
Male
Mustache
Oval Face
Pointy Nose
Side Burns
Smiling
Image Selection
Online learning based on the user-selected images
Modifying the search area
Fine-tuning the image descriptor
Selection of images presented to the user
Reranking of neighbor images
Active learning for sample compression
Until Target Face is not found
Target Image Found

Fig. 1. The flowchart of the retrieval framework. By interactive repetition of the user's input/feedback and system's search process, the framework can find the target image.

later can be adapted to a domain specific task. The image representations which are obtained through deep learning based architectures consistently outperformed numerous conventional hand-crafted feature based methods. This has shown a huge boost in the overall image retrieval performance [14, 15, 16].

However, deep Learning based retrieval systems lacks in reflecting the user's intention in the retrieved results. The process of extracting image representation is fully automatic and it is difficult to reflect the intention of the user in the process of image retrieval [14, 17, 18].

Relevance feedback has been a successful algorithm which tries to capture the user's intention in search process [19, 20, 21]. Jiang et al. [22] proposed a RF-learning scheme where the outcomes of a search are presented to the user and they are permitted to choose related or different objects. According to the query, the user decides whether and how much this image looks similar or dissimilar to the query image. Our approach uses this concept of Relevance Feedback and updates the Deep Neural Network using Human-in-Loop optimisation [23].

Our system is different from the existing face retrieval systems in the following ways:

- It is fast. The state of the art system is slow because of its inability to filter out the irrelevant images. Our system employs a simple but quite efficient technique for this problem, which results in drastic improvement in the performance.
- Moreover, many existing frameworks extract 4096D (4096 dimensions) feature representation vectors which might be computationally expensive at times of frequent updating. Our framework uses 128D feature representation vectors that are obtained from the FC1 layer of the FaceNet [24] model.
- Our method is scalable to large-scale databases like the Celebrity Face Database on which many of the existing approaches fail to perform.
- Our proposed framework is robust. Our database has variations like illumination differences, images taken from different angles and multiple poses.

The rest of the paper has been organised as follows: related and previous works on face retrieval relying on visual memory is briefly discussed in Section 2. Section 3 discusses various components of our proposed framework. We discuss the framework's interface design, experimental setups and the research findings in Section 4. In Section 5, we discuss some applications of our system. Finally in Section 6, we conclude our paper with a summary of our work and future directions.

2. RELATED WORK

In this section, we discuss about the previous work that have been done in the area of Face retrieval. We also talk about the utilisation of deep learning techniques which proved to be successful in this task.

A large volume of work has been done in the Face Image Retrieval domain [25, 26, 27, 28]. Facial contour points are used in face image retrieval that are extracted by computing the geometric facial attributes [29]. However, the problem here is that, it is quite difficult to determine the facial characteristics such as impressions with these contour-points based face retrieval systems. 3D face templates are used for face alignment and after the alignment, LBP Histograms are extracted from the face regions for face representations in this work of Kemelmacher-Shilzerman [30]. Text-based query systems are also employed for face retrieval [30]. The major drawback in all these approaches is that, they could only extract specific face attributes but could not accurately learn the representation of a face as a whole.

With the introduction of Deep Neural Networks in the field of computer vision, many researchers started using various deep learning methods for learning good face representations [31, 32]. This resulted in devising highly accurate face verification and face recognition systems. DeCAF [33], a robust generic image descriptor was later proposed. This was better than the traditional LLC or GIST image descriptors. Lin et al. [34] used DeCAF for the retrieval of the images of clothes. Therefore, we also employ deep face representations in our framework.

Due to recent successes in Generative Adversarial Networks (GAN), researchers have also used them for face retrieval. Zhu et al. [35] proposed Generative Visual Manipulation Model (GVM). It was used to edit the images on a natural image manifold and generate new query image using the GAN for the search process. The retrieval performance in this effort heavily relies on the quality of user's sketch, which is a major drawback of this approach.

One of the most difficult tasks in content based image retrieval is to match the user's search intention with that of the retrieval system. Many researchers resorted to Relevance Feedback [36, 37] for the retrieval process. An identity-based quantization by using a dictionary which was built on the identities of 270 people was proposed [38]. They successfully brought an improvement in the precision of local ranking by making an update in the distance metrics of the top k face representations.

Our framework does manipulations in the results on the basis of the user's input. Unlike the systems of Yugo Sato et al. [38] and WhittleSearch et al. [39], our framework works for large-scale databases and also performs very well on images with variations in terms of illumination, angle and pose.

3. PROPOSED FRAMEWORK

In this section, we describe our robust and scalable face image retrieval framework. Our retrieval framework takes the visual inputs from the users the following way: the user selects the images that they think are similar to the one they have envisioned. We do not ask users for any text or image queries. In the zeroth iteration of the process, we ask users to select some of the facial attributes that they remember their envisioned image has (see Section 3.1). Based on these selected attributes, we propose a number of face images to the user and the user decides whether the candidate facial image is similar to the one they are searching for. The images clicked by the user are stored for further processing. These images are then processed (see Section 3.2) to obtain a semantic vector representation and on the basis of the user interaction (see Section 3.3), search space is modified in the direction of user's interest and the face image descriptor networks weights are updated (see Section 3.4). The new representation from the fine-tuned network is obtained and using a distance metric the images are re-ranked (see Section 3.5).

3.1. Face Attributes Selection

In the search process proposed by Yugo Sato et al. [38], in the first search iteration, the framework presents a set of **random images** for the evaluation. This is a major drawback of their approach. These images may or may not be relevant to the image envisioned by the user (the target image) thus resulting in the burden on the user to evaluate irrelevant images. Also if the user labels an image as dissimilar (that is the user does not select the image), then the framework should not repeat that image in further iterations which however does not happen in Yugo Sato's [38] method.

These issues can be resolved by filtering out the irrelevant images from the database and thereby presenting only the relevant images to the user. We call this "Face Attributes Selection" step. So In this step, we propose to ask user the facial attributes that they remember that the target image has. We do not force user (and thus helps in saving time) to think about the features and tell us through the text queries instead we present a list of some common facial attributes that are very easy to remember. From the list shown on the screen, the user has to select attributes that they think the target image has. Then the selected attributes can then be used to filter out the irrelevant faces in the search process.

This makes the process easier by removing the burden from the users of remembering specific details of the target face. In addition, this speeds up the search process by presenting only the relevant faces to the users.

3.2. Face Representation

The rise of deep learning has closed the "semantic gap" in unconstrained face recognition and it is on the level of human accuracy in some of the benchmarks. For the representation of face images, we used the Convolutional Neural Network which is trained on the dataset of 100M-200M face thumbnails using a triplet loss which allowed it learn an efficient image representation with only 128D vector [24]. The network is based on GoogleLeNet styled inception models [40]. The model is trained with Stochastic Gradient Descent and with standard backprop and adagrad with a learning rate of 0.05. This network was built by stacking up inception modules on top of each other and the detailed architecture is described in this work [40]. Each convolutional layer includes convolution, rectified linear (ReLU) transform

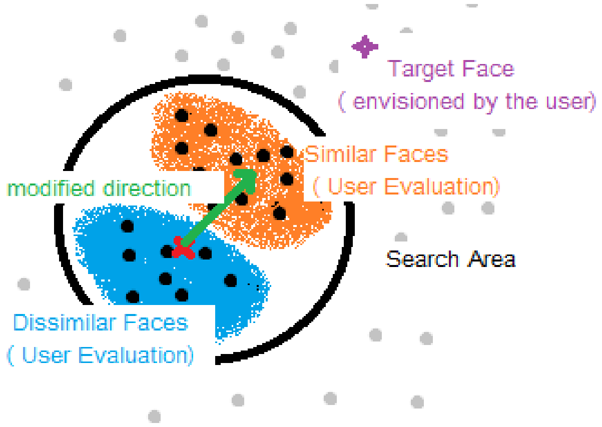


Fig. 2. Search Area Modification

($f(x) = max(x, 0)$), and max-pooling transformation. An input image is transformed into high-dimensional representations via the convolution layers and pooling layers (each is called a convBlock) alternately, and is connected to the fully-connected layers.

The two main reasons for selecting this network are: (i) this was trained using the triplet loss which made it to inherently learn a semantically rich image representation and (ii) this network produces very small *i.e.,* 128-dimensional vector representation for the image which helped us in searching and indexing the images in the database to a large extent. As an image pre-processing, we first detect the face in images stored in the database [41] and normalise them to 96x96 pixels. Then, we use activations of the last fully connected layer to extract high-dimensional facial representation vectors (*i.e.,* 128-dimensional representation vectors) from all database images each fed to the network.

3.3. Searching on the Large-scale Database

In general, as the size of the database increases, an image retrieval system takes a huge amount of time in computing all the similarities between the images stored in the database. For the resolution of this problem for some extent, we create search indexes for the facial image representations (Section 3.2) with the help of Approximate Nearest Neighbour Graph (ANNG) [42]. ANNG is a large-scale database indexing method. It is built in a incremental manner with the k-Nearest Neighbours calculated on a partially complete graph. This has been implemented in the form of libraries in many programming languages. For our purposes, we have used Python based ANNG library call pyNGT. Given a centroid vector of a search area, pyNGT can retrieve k-nearest neighbours based on the cosine similarity between their facial representation vectors.

3.4. Online Machine Learning with relevance feedback based on Image Selection

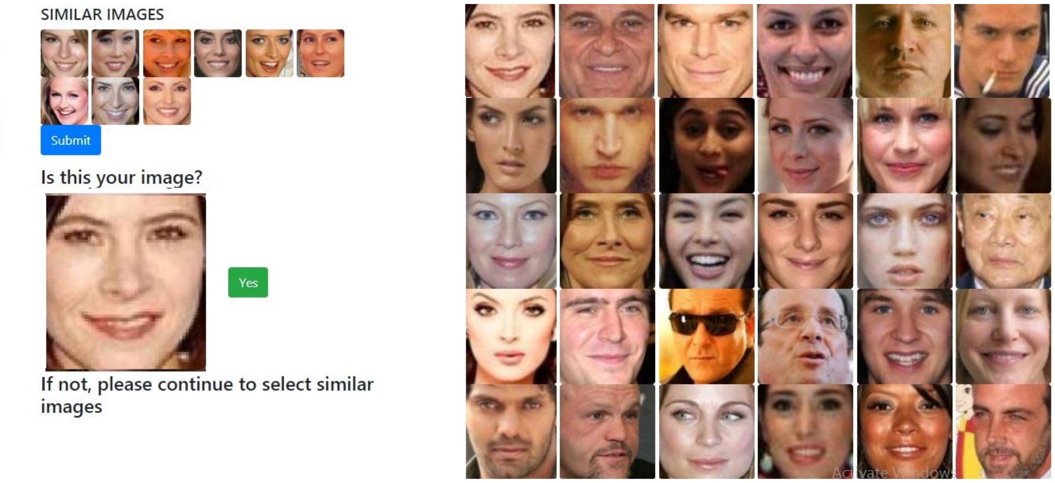
The search process begins with the estimate of the query vector (or the centroid of the search area as in Section 3.3). The query vector is calculated based on the selected images. Here the relevance feedback approach is used. In this approach, based on the feedback, the vector which returns more vectors which are similar to the target image vectors in the database (in Section 3.1) is selected as the query vector. Concretely, we used a well-established algorithm called as Rocchio Algorithm [36] for estimating the query vector. This algorithm is commonly used in exploratory information searching. The fundamental assumption on which this algorithm is built upon is that all users have some similar conception of what is relevant or irrelevant information (or images) to the target image. What this algorithm simply does is that it returns a modified query vector by a maximal separation of the relevant and irrelevant vectors. Mathematically, the operation is shown in Equation 1.

$$\vec{q_m} = \alpha\vec{q_0} + \beta\frac{1}{|D_r|}\sum_{\vec{d_j}\in D_r}\vec{d_j} - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d_k}\in D_{nr}}\vec{d_k} \quad (1)$$

where $\vec{q_0}$ 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 weights and the $\vec{q_m}$ is the modified vector.

Since we want query vector to be moving in the direction of only similar images, we decided to weigh the set of similar vectors much higher than the set of irrelevant vectors, hence we chose the following values for the weights: α = 1.0, β = 0.9, and γ= 0.1 (see Equation 1). The centroid of the search area will now be moved towards the direction of the interest that is towards the relevant faces in the search space and away from the irrelevant faces. This is how the search process goes on and the search area is modified by interactive repetition of the users input and we refer to the database in an exploratory manner. The modified search area takes the user more closer to the relevant images and away from the irrelevant images and thus user sees many similar images as the search process progresses (see Figure 2).

Fine-tuning of Image Descriptor Network
Generally, the retrieval systems use image representations generated by the pretrained neural networks and their performance solely depends on the extent to which these representations could fill up the semantic gap. In real world, human perception may vary with human to human. To resolve this problem and in order to give a personalised result, for each user the system dynamically fine-tune the weights of the convolutional neural network using the users feedback in each iteration in the process. This fine-tuning is performed on the same model which is used for extraction of image embeddings (see Section 3.1). There is no change


SIMILAR IMAGES

Is this your image?

If not, please continue to select similar images

Fig. 3. User Interface: A user can select similar images (top-left) from the presented images (right) to retrieve the target image (bottom-left) by interactively repeating the selection process.

in the network architecture except that a classification layer with two logistic units (for binary classification, i.e., similar and dissimilar classes) has been added to the network. This layer is now the output layer of the network. The activations of the last layer are passed to a Softmax function, which is expressed in Equation 2.

$$p_r = \frac{\exp(h_r)}{\sum_{j=1}^{R} \exp(h_j)} \tag{2}$$

where h_r is the r-th activation of the last layer and R is the number of classes; p_r denotes the probability of the r-th class.

In the process of fine-tuning, while keeping the parameters of convolutional neural network freezed, with the help of back-propagation, we fine-tune the fully connected layers of the network. Conventionally, we minimise the traditional cross-entropy cost function for each training image set. The cross-entropy cost function is given in Equation 3 and 4.

$$E = -\sum_{n=1}^{N} \sum_{r=1}^{R} l_{nr} \log p_r \tag{3}$$

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

where N represents the size of training image set and l_{nr} is the label vector of the n-th training image provided by the user.

3.5. Sample Compression

It is beneficial to have the users feedback as detailed as possible. Generally, the existing relevance feedback systems present the user a greater number of samples that are similar to the query point. This imposes a severe problem. As the number of samples increases, the process becomes tedious and burdensome for the user as the user has to evaluate all the presented images. For example in this work [39], the user has to evaluate all 50 images and for all 18 attributes. In this paper, we use the technique proposed by Yugo Sato et al. [38] called as Active Selection. Concretely, after human-in-loop optimisation (Section 3.4) and extracting 128D image embeddings with updated weights, the framework re-ranks the images and presents them to the user. But instead of presenting all the images we apply Active Selection (see Figure 1) on the images for decreasing the number of images presented. We found that Active Selection doesn't bring much change in the performance of our system (see section 4). We describe Active Selection in the following paragraphs.

In the search process, on the basis of the images selected, the convolutional neural network is fine-tuned with human in loop. New image representations are extracted for all the images. Then these images are re-ranked using the cosine similarity metric. Using the KNN graph, k-nearest neighbours are obtained. Top nearest neighbours are defined as the "top-ranked" images and the remaining as "low-ranked" images.

The idea of Active Selection is inspired from Active Learning. The basic concept of Active Learning is that a learning algorithm can have better performance with lesser training labels if it can make choice of the data from which it learns [43]. In this paper, top 30% re-ranked results are defined as top-ranked images and remaining as low-ranked images. Active learning is adopted for low-ranked images, which can choose the images that have their class determined uncertainly by the currently trained network. The images that satisfies this requirement as mathematically formulated in Equation 5, are taken from the low-ranked images:

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

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 to belong to that class.

4. EVALUATION

In this section, we describe our evaluation strategy. We discuss about the database we used (see Section 4.1), the interface of our retrieval system (see Section 4.2) and our evaluation methodology (see Section 4.3). We then present our research findings (in Section 4.4) and talk about the use-cases of our framework (see Section 4.5). We then conclude our work in Section 4.6.

4.1. Database

For the effective evaluation of our framework, we needed a face dataset which is larger by a big margin than all the previously used datasets in the area of CBIR and has images in various face postures, illumination and clutter in the background. Therefore for our experiments, we used a subset of 20,000 images of the Large-scale Celebrity Face Dataset [44] [which is significantly larger because, many existing techniques used small databases like Yugo Sato [38] used Chicago Face database (597 images 290 male and 307 female), WhittleSearch [39] used 772 images of only 8 persons, etc.].

The 20K subset of Celebrity Face Dataset [44] is created by sampling the images according to the normal distribution for each of the facial attribute in the dataset, thus ensuring sufficient images in the datset corresponding to all the attributes. Celebrity Face Dataset is a large-scale face attributes dataset with more than 200K celebrity images, each with approximately 40 attribute annotations. The images in this dataset cover numerous pose variations and background clutter (see Figure 3). CelebA has large diversities, large quantities, and rich annotations, including 10,177 number of identities, 202,599 number of face images, along with 5 landmark locations and 40 binary attribute annotations per image [45]. It is a very good choice for evaluating the performance of our framework. Unlike the dataset in the previous work [38], this dataset is comparatively very big and has lots of variations in terms pose, illuminations, background clutter etc. and also the facial attributes annotations which makes it perfect choice for executing our idea.

4.2. Interface Design

In this section, we describe the user interface that is used in the user studies mentioned in the subsequent sections (see Figure 3). This application can be opened in the browser on any operating system. In the zeroth iteration of the search process, the user selects the facial attributes that they think their envisioned image has (see Figure 1). For all the next iterations, the following process is repeated until the user finds the target image.

Some number of images are presented in the search window for the user to evaluate whether each of these are similar or dissimilar to the target image. The user then selects the similar image(s) by clicking on them. As soon as the user clicks on the image, the image disappears from the search window and appears on the left side of the interface. After the user has selected the similar images, the user will have to click on the ‘Submit’ for submitting the images and ending the current iteration. As soon as the current iteration ends, a pop up image appears asking the user ‘Is this the image?’, if the shown image is the image in the user’s memory. This bottom-left image is the top nearest-neighbour face in the current search iteration (i.e., a top-ranked image after re-ranking the neighbour images). Based on the top-ranked image, the user can intuitively understand the process of creating face representations through image-labelling. If it is not the image that the user envisioned, the process moves ahead to the second iteration of similar image(s) selection. This process continues until the user successfully finds the target image. We ran this framework on NVIDIA TITAN XP GPU for conducting our experiments.

4.3. Methodology

We performed user studies in order to assess the utility of our face retrieval system. In the user studies, we invited undergraduate students of which 5 were male and 5 were female each 18 to 24 years old. We gave a detailed overview of our system interface and made them comfortable in using the system for searching the image. After they became familiar with the system, we asked them to search the image using our retrieval framework. In the experiment, we showed each subject a randomly selected face from the database and asked them to closely look at it and search for it using our framework. Next, they were asked to select the face attributes that they remember that the random image had. Based on the selected attributes, the framework filtered out the irrelevant images. The users repeated the search as described to them until they found the target image with the help of their visual memory. We also ensured that the user was not allowed to see any more examples. We measured the total search time for each user and the total number of iterations the system took in order to search for the target image.

Table 1. Experimental Setups

Experimental Setup 1 (baseline)	Experimental Setup 2	Experimental Setup 3
25 images	25 images	25 images
Active Selection	Face Attributes Selection	Both

There has not been much work done in face retrieval relying on user's visual memory. To the best of our knowledge, only Yugo Sato et al. [38] has worked on this problem so far. We take the results of their approach as our baseline results (Experimental Setup 1). Experimental Setup 2 is our approach and Experimental Setup 3 is both the approaches combined. Thus, in this paper, we evaluate our retrieval system as follows (see Table 1):

- **Experimental Setup 1 (baseline)**
  Our framework presents 25 images that are obtained on applying only the Active Selection to the 50 neighbour images. We refer to it as a 'baseline' as this step was proposed in [38].

- **Experimental Setup 2**
  25 images are presented after the irrelevant images are filtered out in the beginning following the 'Face Attributes Selection' step. We do not apply Active Selection in this setup.

- **Experimental Setup 3**
  In this setup, we apply the 'Face Attributes Selection' as well as Active Selection. The purpose of this setup is to see how much change does the Active Selection bring when applied along with our proposed technique.

The choice of 25 images in the experimental setups is inspired from [38].

4.4. Search Cost

In this section we present our research findings. We report the search cost incurred for a subject to find a target celebrity face image in the 20K subset of the CelebA dataset using our face retrieval interface. The total search time and the total number of search iterations (until the target face is successfully found) are recorded. The results including the baseline setup are shown in Figure 4 and 5. In spite of the unstable facial memory inputs, our framework successfully found the envisioned face image efficiently and rapidly. The obtained scores for all three experimental setups are given as: 485.6 seconds; 198.3 seconds; 169.9 seconds (see Table 2). On an average, our system took a time of approximately (less than or equal to) 3 min in searching for the target image while the baseline approach took over 8 min. In addition to that, the search iterations for all three Experimental settings: 7.6; 4.1; 3.6 (see Table 2). On this metric too, our approach outperformed the

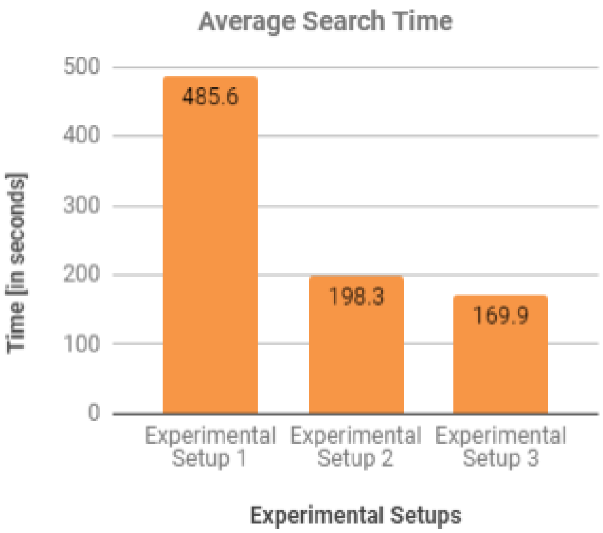


Fig. 4. Average target image search time for three experimental setups.

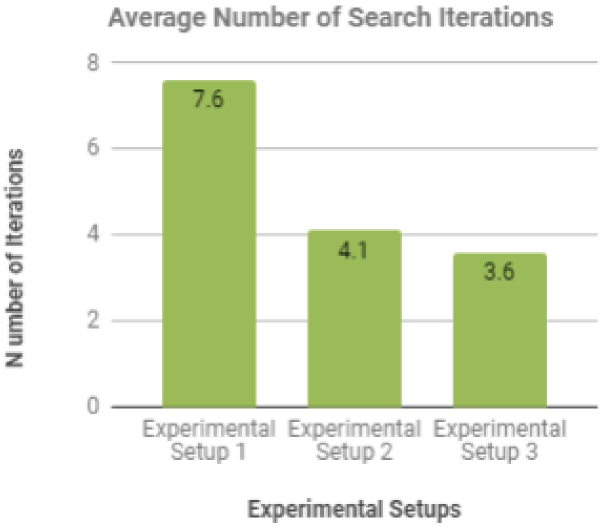


Fig. 5. Average number search iterations for different experimental setups.

Table 2. Experimental Results

	Experimental Setup 1	Experimental Setup 2	Experimental Setup 3
Average Number of Search Iterations	7.6	**4.1**	3.6
Average Search Time	485.6 seconds	**198.3 seconds**	169.9 seconds

Yugo Sato's [38] approach by a big margin of 4 search iterations. Hence the experiments confirm that our framework outperformed all the existing face retrieval systems that relies on the concept of visual memories. As the results show, the facial attributes selection step brought a very large difference resulting in a large reduction of search time as well as the number of search iterations. Active selection when combined with our approach gave us the state of the art results. Though Active Selection did not bring drastic changes.

5. USE CASES

Our face retrieval framework is useful in all the situations in which a person tries to remember the face of some other person that they do not exactly remember.

Our retrieval framework can be used for the purpose of police investigation. A victim can use our system to retrieve the the criminal or suspect from the criminal database easily. This can help get rid tedious sketch drawing which is inaccurate several times.

Our framework can also be used in person re-identification, which has various applications like identifying the lost person or patient identification in health-care.

6. CONCLUSION AND FUTURE WORK

In conclusion, we proposed a novel method that improved the efficiency, scalability and robustness of the recently proposed state-of-art face retrieval framework based on user's visual inputs [38]. We chose highly representative 128D feature representation vectors that are given by the FC1 layer of the FaceNet [24] over highly dimensional 4096D feature representation vectors which reduced the time of database indexing [46]. We added a simple yet effective 'step-0' of face attributes selection to the framework thus allowing the framework to reduce the target image search space dramatically. We considered the images that are cluttered, occluded, posed in different angles and ones that are subjected to varying lighting conditions. Our method performed well on them which shows that our proposed method is robust. Moreover, our method is scalable to large-scale databases which many existing methods suffered to work on. In our case, we considered using a subset of 20K images from the existing 200K images of the Celebrity Face Database. However, there is certainly a lot of scope in future, for improving the proposed framework further.

7. ACKNOWLEDGEMENT

Rajiv Ratn Shah is partly supported by the Infosys Center for AI, IIIT Delhi and ECRA Grant by SERB, Government of India. This research is also supported by Singapore Ministry of Education Academic Research Fund Tier 1 under MOE's official grant number T1251RES1820.

8. REFERENCES

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

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

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

[4] Yi Yu Suhua Tang Kiyoharu Aizawa and Akiko Aizawa, "Category-based deep cca for fine-grained venue discovery from multimodal data," *IEEE Transaction on Neural Network and Learning System (TNNLS)*, vol. 30, 2019.

[5] Yi Yu Suhua Tang Francisco Raposo and Lei Chen, "Cross-modal correlation learning for audio and lyrics in music retrieval," *ACM Transaction on Multimedia Computing Communication and Applications (TOMCCAP)*, vol. 15.

[6] Rajiv Ratn Shah Yi Yu and Roger Zimmermann, "Advisor - personalized video soundtrack recommendation by late fusion with heuristic rankings," *ACM international conference on Multimedia (ACM MM14)*, 2014.

[7] Rajiv Ratn Shah Yi Yu Akshay Verma Suhua Tang Anwar Dilawar Shaikh and Roger Zimmermann, "Leveraging multimodal information for event summariza-

tion and concept-level sentiment analysis," *Journal of Knowledge-Based Systems*, 2016.

[8] Rajiv Shah and Roger Zimmermann, "Multimodal analysis of user-generated multimedia content," *ACM international conference on Multimedia (ACM MM14)*, 2017.

[9] Faezeh Ensan and Ebrahim Bagheri., "Document retrieval model through semantic linking.," *Proc. of the Tenth ACM International Conference on Web Search and Data Mining. ACM,*, p. 181190., 2017.

[10] Xirong Li Tiberio Uricchio Lamberto Ballan Marco Bertini Cees GM Snoek and Alberto Del Bimbo., "Rapid clothing retrieval via deep learning of binary codes and hi- erarchical search.," *Proc. of the 5th Conference on Multimedia Retrieval. ACM,*, pp. 499–502, 2015.

[11] Jun Yu Xiaokang Yang Fei Gao and Dacheng Tao., "Deep multimodal dis- tance metric learning using click constraints for image ranking.," *IEEE Trans. on Cybernetics.*, 2016.

[12] Ju Han and Kai-Kuang Ma., "Fuzzy color histogram and its use in color image retrieval.," *IEEE Trans. on Image Processing.*, pp. 944–952, 2002.

[13] Dengsheng Zhang Aylwin Wong Maria Indrawan and Guojun Lu., "Content-based image retrieval using gabor texture features.," *IEEE Trans. on Pat- tern Analysis and Machine Intelligence.*, pp. 13–15, 2000.

[14] DayongWang Steven Chu Hong Hoi PengchengWu Jianke Zhu Yong-dong Zhang JiWan and Jintao Li., "Deep learning for content-based image re- trieval: A comprehensive study," *Proc. of the 22nd ACM international conference on Multimedia. ACM.*, p. 157166, 2014.

[15] Artem Babenko Anton Slesarev Alexandr Chigorin and Victor Lempitsky., "Neural codes for image retrieval.," *Proc. of the European Conference on Computer Vision. Springer.*, p. 584599, 2014.

[16] Liang Wang Fang Zhao Yongzhen Huang and Tieniu Tan, "Deep seman- tic ranking based hashing for multilabel image retrieval.," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition.*, p. 15201528., 2015.

[17] Bor-Chun Chen Yan-Ying Chen Yin-Hsi Kuo and Winston H Hsu., "Scal- able face image retrieval using attribute-enhanced sparse codewords.," *IEEE Trans. on Multimedia.*, vol. 15, pp. 1163–1173, 2013.

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

[19] En Cheng Feng Jing and Lei Zhang., "A unified relevance feedback frame- work for web image retrieval.," *IEEE Trans. on Image Processing.*, pp. 1350–1357., 2009.

[20] Zhong Ji Yanwei Pang and Xuelong Li., "Relevance preserving projection and ranking for web image search reranking.," *IEEE Trans. on Image Processing.*, p. 41374147., 2015.

[21] Yongdong Zhang Xiaopeng Yang and Tao Mei., "Image search reranking with query-dependent click-based relevance feedback.," *IEEE Trans. on Image Processing.*, p. 44484459, 2014.

[22] Wei Jiang, Guihua Er, Qionghai Dai, Lian Zhong, and Yao Hou, "Relevance feedback learning with feature selection in region-based image retrieval," *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 2, pp. ii/509–ii/512 Vol. 2, 2005.

[23] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, pp. 5:1–5:60, May 2008.

[24] Florian Schroff Dmitry Kalenichenko and James Philbin, "Facenet: A unified embedding for face recognition and clustering.," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[25] Christian Herrmann and Jrgen Beyerer., "Face retrieval on large-scale video data.," *Proc. of the 12th Conference on Computer and Robot Vision (CRV). IEEE.*, pp. 192–199, 2015.

[26] Subhradeep Kayal., "Improved hierarchical clustering for face images in videos: Integrating positional and temporal information with hac.," *Proc. of the International Conference on Multimedia Retrieval. ACM.*, p. 455, 2014.

[27] Enrique G Ortiz Alan Wright and Mubarak Shah., "Face recognition in movie trailers via mean sequence sparse representation-based classification.," *IEEE Conference on Computer Vision and Pattern Recognition.*, p. 35313538, 2013.

[28] BaoyuanWu Yifan Zhang Bao-Gang Hu and Qiang Ji, "Constrained clus- tering and its application to face

clustering in videos.," *Proc. of the IEEE Con- ference on Computer Vision and Pattern Recognition.*, p. 35073514, 2013.

[29] Brandon M Smith Shengqi Zhu and Li Zhang., "Face image retrieval by shape manipulation.," *Proc. of the IEEE Conference on Computer Vision and Pat- tern Recognition. IEEE.*, p. 769776, 2011.

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

[31] Hailong Liu, Baoan Li, Xueqiang Lv, and Yue Huang, "Image retrieval using fused deep convolutional features," *Procedia Comput. Sci.*, vol. 107, no. C, pp. 749–754, Apr. 2017.

[32] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li, "Deep learning for content-based image retrieval: A comprehensive study," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, New York, NY, USA, 2014, MM '14, pp. 157–166, ACM.

[33] Jeff Donahue Yangqing Jia Oriol Vinyals Judy Hoffman Ning Zhang Eric Tzeng and Trevor Darrell., "Decaf: A deep convolutional activation fea- ture for generic visual recognition.," *Proc. of the 31st International Conference on Machine Learning.*, vol. 32, pp. 647–655, 2014.

[34] Kevin Lin Huei-Fang Yang Kuan-Hsien Liu Jen-Hao Hsiao and Chu-Song Chen, "Rapid clothing retrieval via deep learning of binary codes and hierarchical search," *ACM international conference on Media Retrieval (ICMR15)*, 2015.

[35] Jun-Yan Zhu Philipp Krhenbhl Eli Shechtman and Alexei A Efros., "Gen- erative visual manipulation on the natural image manifold.," *Proc. of the Euro- pean Conference on Computer Vision. Springer*, p. 597613, 2016.

[36] Joseph Rocchio., "Relevance feedback in information retrieval.," *The Smart Retrieval System-Experiments in Automatic Document Processing*, 1971.

[37] Yong Rui Thomas S Huang Michael Ortega and Sharad Mehrotra., "Rele- vance feedback: a power tool for interactive content-based image retrieval.," *IEEE Trans. on Circuits and Systems for Video Technology*, 1998.

[38] Yugo Sato Tsukasa Fukusato and Shigeo Morishima, "Face retrieval framework relying on users visual memory.," *ICMR*, June 2018.

[39] Adriana Kovashka Devi Parikh and Kristen Grauman., "Whittlesearch: Image search with relative attribute feedback," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE*, 2012.

[40] Christian Szegedy Wei Liu Yangqing Jia Pierre Sermanet Scott Reed Dragomir Anguelov Dumitru Erhan Vincent Vanhoucke Andrew Rabinovich, "Going deeper with convolutions.," *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[41] Davis E King., "Dlib-ml: A machine learning toolkit.," *Journal of Machine Learning Research*, 2009.

[42] Masajiro Iwasaki, "Ngt: Neighborhood graph and tree for indexing.," *http://research-lab.yahoo.co.jp/software/ngt/.*, 2015.

[43] Burr Settles, "Active learning literature survey.," *University of Wisconsin, Madison*, 2010.

[44] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[45] Ziwei Liu Ping Luo Xiaogang Wang Xiaoou Tang, "http://mmlab.ie.cuhk.edu.hk/projects/celeba.html," .

[46] J.A. Jorge M.J. Fonseca, "Indexing high-dimensional data for content-based retrieval in large databases," .