From A Glance to "Gotcha": Interactive Facial Image Retrieval with Progressive Relevance Feedback

Xinru Yang
Carnegie Mellon University, Google LLC
xinruy@cs.cmu.com

Haozhi Qi
UC Berkeley
hqi@berkeley.edu

Mingyang Li
University of Pennsylvania
myli@alumni.upenn.edu

Alexander Hauptmann
Carnegie Mellon University
alex@cs.cmu.edu

Figure 1: When a user is trying to identify a character by providing relative descriptions for multiple rounds. Images are property of 2019 Home Box Office, Inc., reproduced under fair use.

ABSTRACT

Facial image retrieval plays a significant role in forensic investigations where an untrained witness tries to identify a suspect from a massive pool of images. However, due to the difficulties in describing human facial appearances verbally and directly, people naturally tend to depict by referring to well-known existing images and comparing specific areas of faces with them and it is also challenging to provide complete comparison at each time. Therefore, we propose an end-to-end framework to retrieve facial images with relevance feedback progressively provided by the witness, enabling an exploitation of history information during multiple rounds and an interactive and iterative approach to retrieving the mental image. With no need of any extra annotations, our model can be applied at the cost of a little response effort. We experiment on CelebA [20] and evaluate the performance by ranking percentile and achieve 99% under the best setting. Since this topic remains little explored to the best of our knowledge, we hope our work can serve as a stepping stone for further research.

AIIS'20, July 30, 2020, Virtual Event, China

ACM ISBN TBA…$TBA
https://doi.org/TBA

CCS CONCEPTS

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

KEYWORDS

interactive retrieval, dialog, relevance feedback

ACM Reference Format:
Xinru Yang, Haozhi Qi, Mingyang Li, and Alexander Hauptmann. 2020. From A Glance to "Gotcha": Interactive Facial Image Retrieval with Progressive Relevance Feedback. In *Proceedings of AIIS'20*. ACM, New York, NY, USA, 9 pages. https://doi.org/TBA

1 INTRODUCTION

Facial image retrieval is an interesting yet challenging task for its practical use in computational forensics. It is even more difficult when the target image is not known to the system but only exists in the user's mind (mental image). Therefore, the descriptions of the target image from the user are necessary for a reasonable retrieval results. However, compared to describing the perceived face with absolute depictions, people naturally feel it easier to refer to an existing image and provide the descriptions of the their difference. For example, in Figure 1, when a user is trying to identify a character without knowing the character's name but only an impression of the character's face, a reference image is shown to the user and the user responds with differences between the reference image and the mental image. Upon receiving these responses, another reference image will be shown to the user for further feedback. Ideally, the reference images shown to the user should be refined and closer

to the mental image over time. This process might last for limited rounds or till the mental image is retrieved.

In reality, when people are asked to describe facial appearances of another person, the descriptions can be roughly categorized into two aspects, basic descriptions and advanced descriptions. Basic descriptions contain objective measurements on facts such as the color of hair, whether wearing hat or eyeglasses, which can be conveniently mapped into attributes with certain values. Advanced descriptions involve with subjective opinions people sense from the facial appearances such as beauty, aging, friendliness etc. Previous studies have demonstrated the instability in advanced descriptions due to the ambiguities in human perception [2, 6, 23, 27]. Moreover, researchers have expressed concerns that verbally attending to facial differences might alter witness's memory of the original face, which can be detrimental to forensic applications [3]. Therefore, we take an attribute-aware approach (e.g. "with or without glasses") where users are able to describe and response easily and efficiently. An appropriate dataset for this research goal is `CelebA`, which will be described in Subsection 4.1.

Recent research [7, 19, 30] shows that interactive image retrieval has the advantages of integrating user feedback and improving retrieval performances by relevance feedback. Therefore, we build an interactive retrieval system that takes multiple rounds of collecting user's feedback and refining the retrieval results. Apart from the interactive framework, our model considers an extra mechanism, progressiveness, during the interactions.

Considering the difficulty for real people in describing and comparing facial images in mind, it is hard for them to provide thorough feedback at each round. Thus, we design a mechanism that progressively discloses the feedback. Specifically, in each round of retrieval, the system only provides partial relevance by masking the rest of it. In the following rounds, the ratio of masked relevance feedback is gradually decreased, allowing for more information to be disclosed. This setting mimics a progressive disclosure that might better reflect the functionality of human memory. Essentially, it outperforms all other settings in our experiments and might share a similar fundamentality as dropout.

With the target image unknown to the retrieval system but only its attributes, one naive way would be annotating the attributes of every image in the database and seeking for the closest one. However, annotating a dataset is highly expensive and exhaustive. Thus, our model retrieves the image by cooperating image features and instant feedback without prior annotations. When the system returns a candidate image in each round of retrieval, the user is only required to provide relevance feedback between it and the mental target image. We believe this instantaneous responding process is light and feasible since our system has only a few rounds with only 1 image each round.

The key contributions of our work are as follows:

- A new retrieval problem setting on human facial images under interactive search where users are allowed to convey their mental images to the system and iteratively refine the retrieved results.
- An end-to-end interactive Content-Based Image Retrieval (CBIR) framework to address the above problem setting by employing supervised learning approach.
- A novel progressive disclosure mechanism in collecting relevance feedback from users during multiple rounds of interaction. The mechanism reaches the best performance while mimicking human behaviors.
- An instant feedback setting for interactive applications. The setting can help to reduce workload of manual annotations necessary for learning about the annotator.

The paper is organized in 5 sections. After reviewing related work in Section 2, we describe the structure of our framework (in Subsection 3.1) and the algorithm it runs by (in Subsection 3.2) in Section 3. We then continue to Section 4, where we demonstrate the validity with a baseline experiment and the robustness with a series of ablation studies. Finally in Section 5, we make final remarks about our work in terms of its applications and limitations.

2 RELATED WORK

In this section, we introduce related work in the fields of image retrieval and facial recognition.

2.1 Image Retrieval Researches

Overview. With countless images generated everyday, efficient navigation demands intuitive approaches that are aware to image content, giving rise to the field of Content-Based Image Retrieval (CBIR) [11]. To further facilitate retrieval, interactive querying methods are being developed over the past decades. A prominent approach in this area is relevance feedback (RF) [26]. In traditional RF settings, users evaluate how relevant one retrieved image is to their desired result, report this perceived relevance as a numerical value, and expect a refined result from the next retrieval.

Relative Attributes. However, for complex images, a single relevance value can be ambiguous and thus misleading [9]. In order to improve specificity in feedback, relative attributes (RA) has been proposed as a new mechanism [18]. In RA-enabled CBIR, user dictates which attribute(s) of candidate image(s) should be tweaked, and optionally by how much. Users may also tune the parameters of the system with an emphasis on specific attributes and image features [8, 16, 21].

Interactive Feedback. Feedback mechanism implies that each retrieval involves multiple rounds of information exchange, or a *dialog*. Each round provides the CBIR system with extra information to refine its results. This refinement process can take the form of a decision tree [22] or a neural network (NN) [28]. In NN-based implementations, reinforcement learning (RL) can be employed to reduce training-time supervision [5]. Owing to the descriptive nature of feedback, CBIR experience can be further enhanced with natural language processing (NLP) [13]. Moreover, a combination of RL and NLP is proved useful in a setting of shoe-shopping [12].

2.2 Facial Image Researches

Use Case. In our work, we study the retrieval of human face images instead of footwear. Face image retrieval has socially critical applications in industries such as forensics [24]. A typical use case would be having a witness identify the appearance of some specific personnel from Closed-circuit television (CCTV) recordings. The footage repository can be huge, impossible for untrained eyes to

examine thoroughly. This workload demands facial recognition technology combined with CBIR techniques.

Feature Extractor. Plenty of CBIR research involving RF has focused on algorithmically-generated visual descriptors, such as `MPEG-7` [29]. These low-level features (hue, angle, slope, etc.) are infamously difficult to map to high-level concepts (glasses on, oval face, etc.) [25]. To bridge over this gap, our pipeline employs a pre-trained Convolutional Neural Network (CNN) for extracting facial features.

3 METHOD

3.1 Model Architecture

Our model, shown in Figure 2, consists of four components for different purposes. They will be demonstrated in detail respectively as follows:

3.1.1 The User Simulator. The user simulator mimics a human user who has a target image in mind and provides feedback at each round.

A human user 1) annotates and attributes from a given candidate image, 2) compares specific attributes of the candidate image with those of the target image (which only exists in the user's mind), and 3) reports to the encoder model. Note that the target image is not known to the system because it symbolizes the image in the memory of the witness and its attributes are not explicitly defined either.

In training process, we utilize the existing annotations of attributes in `CelebA` to avoid additional inaccuracies if a new annotator is introduced. In testing, we assume that the online annotations by a person should be the same as the existing annotations under ideal circumstances. Therefore we again utilize the existing annotations to examine our model in this case.

3.1.2 The Encoder Model. This encoder model encodes signals from different spaces into one unified representation. Besides the candidate image annotations and the relevance feedback provided by the user simulator, the candidate image itself is also referenced so the encoder model can learn the correspondences between image features and attributes to produce a shared representation of them.

3.1.3 The Aggregator. The aggregator aggregates the history representation during multiple rounds to exploit the information of previous rounds. Specifically, a gated recurrent unit (GRU) is implemented followed by a linear transformation.

3.1.4 The Retrieval Model. The retrieval model searches the database for a new candidate image that best matches the representation and returns it back to the user.

Upon receiving, it computes the distance between the representations of each image in the database and the aggregated representations from aggregator and selects the nearest neighbors.

During training, it returns a random image in these nearest neighbors for the sake of robustness; In testing, we use greedy approach that returns exactly the nearest one.

3.2 Algorithm

Initialization. At the beginning of each retrieval, the user simulator randomly selects an image from the database as its target image. The user simulator then annotates and stores the target image for the convenience of calculating relevance feedback in each round. For simplification, we use the existing annotations in the dataset which is a 40-dimensional Boolean vector $\vec{q} \in \mathbb{R}^{40}$ where $\mathbb{R} = \{-1, 1\}$. Before the first round of retrieval, retrieval model randomly returns a candidate image from the database as a starting point.

Loop. After initialization, the system executes the following steps iteratively until termination.

In the User Simulator. At the t-th round of retrieval, the user simulator first annotates the t-th candidate image and stores as a 40-dimensional Boolean vector $\vec{a}_t \in \mathbb{R}^{40}$. Then it calculates the relevance $\vec{o}_t$ between the t-th candidate image and the target image: $\vec{o}_t = \vec{a}_t \circ \vec{q} \in \mathbb{R}^{40}$, where $\circ$ denotes elementwise multiplication. The attributes, together with all other Boolean values, takes -1 as `False` and 1 as `True`. This enables the user simulator to **calculate the relevance** between the attributes of candidate image and those of target image and the relevance is also binary using the same numbers as attributes: A term in the relevance is -1 if the corresponding attribute in candidate image is different than that in the target image and 1 if they are identical. To realize the progressiveness during the retrieval, the relevance feedback will be replaced by 0 in accordance with certain proportion p_t representing the masked part. As the t increases, p_t will gradually decrease indicating more and more disclosure in the relevance feedback. In our implementation, we set $p = \{0.5, 0.3, 0.2, 0.1, 0.0\}$. The computed relevance $\vec{o}_t$ is then fed to the encoder model along with the annotated attributes $\vec{a}_t$.

In the Encoder Model. Firstly, $\vec{o}_t$ and $\vec{a}_t$ are concatenated together and embedded by a linear transformation named *indication layer*, with the intuition that some attributes (such as gender, compared to nose size) are more indicative than the others: $\vec{x}_t = W_A(\vec{o}_t \bigoplus \vec{q})$. Here, $\bigoplus$ denotes concatenation, and $W_A \in \mathbb{R}^{80\times256}$ is our first linear transformation. Meanwhile, a CNN $Conv$ is employed to extract the features of the candidate image which is passed through another linear transformation: $\vec{f}_t = W_I(Conv(CandidateImage))$. In our implementation, $W_I \in \mathbb{R}^{256\times256}$, and $Conv$ is a pre-trained SE-ResNet [14]. Outputs from the these two linear transformations are concatenated together and fused in a multi-layer perceptron (MLP): $\vec{r}_t = W_M(\vec{x}_t \bigoplus \vec{f}_t)$, where $W_M \in \mathbb{R}^{512\times256}$ is the MLP.

In the Aggregator. Historical information is then referenced in a GRU followed by a third linear transformation, $W_G \in \mathbb{R}^{256\times256}$: $\vec{g}_t, \vec{h}_t = GRU(\vec{r}_t, \vec{h}_{t-1}), \vec{s}_t = W_G\vec{g}_t$, where hidden state $\vec{h}_t \in \mathbb{R}^{256}$ and the output of GRU $\vec{g}_t \in \mathbb{R}^{256}$. The final representation, $\vec{s}_t \in \mathbb{R}^{256}$, consists of history representations and information of the current round.

In the Retrieval Model. Next, $\vec{s}_t$ is sent to the retrieval model. For $\forall i = 1, 2, ..., N$ where N denotes the size of the database, it calculates the $L2$ distance between $\vec{s}_t$ and $\vec{m}_i = Conv(I_i)$, the feature representation the i-th image I_i: $d_i = \|\vec{s}_t - \vec{m}_i\|_2$. Using the $L2$ distances, the top-K nearest neighbors of $\vec{s}_t$ can be found, denoted

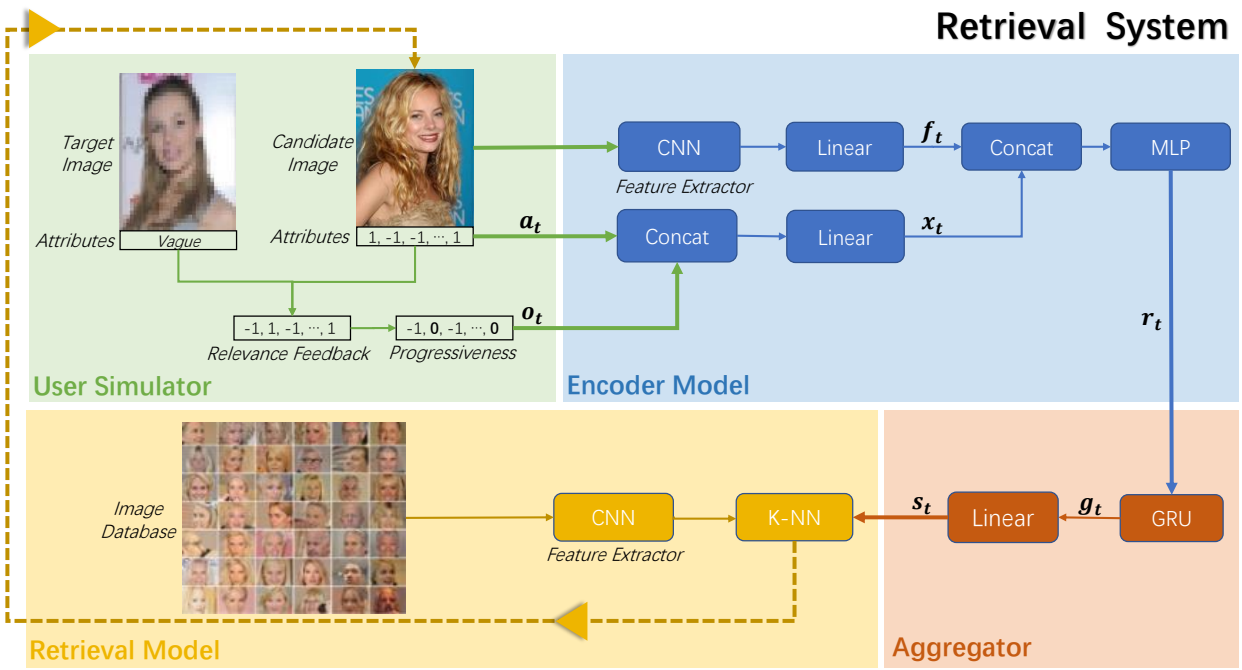

Figure 2: Our proposed end-to-end interactive facial image retrieval framework with progressive disclosure.

by $\vec{n} \in \mathbb{N}^K$. We model the sampling probability with a softmax distribution over the top-K nearest neighbors:

$$\pi(j) = \mathrm{e}^{-d_j} / \sum_{k=1}^{K} \mathrm{e}^{-d_{n_k}}, j = 1, 2, \ldots, K. \tag{1}$$

Two approaches can be adopted to choose the $(t+1)$-th candidate image I'_j:

- In training, we choose a random image I'_j where $j' \sim \pi$.
- During testing, we choose the nearest image I'_j where $j' =_j \pi(j)$.

Termination. The loop terminates when user simulator reports that candidate image *is* target image, or when the maximum number of rounds (default 5) is reached.

3.3 End-to-end Training

In practice, the system might return multiple candidate images for the user in each turn and collect their relevance feedback respectively for better retrieval performance. While in our work, we simplify the scenario by returning a single image in each turn. It is also available to extend our framework to the practical case by enabling the user to choose one preferred image out of multiple candidate images to obtain the relevance feedback.

Aiming at improving the ranking position of the target image, We train the model by a supervised learning objective. In the beginning, all the parameters of the network are randomly initialized. For loss function, we refer to [12] where it uses triplet loss objective.

$$\mathcal{L} = \mathbb{E}[\sum_{t=1}^{T} max(0, \left\|\vec{s_t} - \vec{x^+}\right\|_2 - \left\|\vec{s_t} - \vec{x^-}\right\|_2 + m)] \tag{2}$$

where $\vec{x^+}$ is the features of the target image and $\vec{x^-}$ is the features of a random image sampled from the database as a negative sample. m is a hyper-parameter and constant representing the margin. $\|.\|$ means $L2-$norm. Even though the ranking position is not available to learn directly since it is not differentiable, we can exploit the advantage of triplet loss objective that the rank of the target image can be improved by ensuring the proximity of the target image and candidate images.

As for evaluation, we report the average ranking percentile of all the image in the training or testing set. More details on ranking percentile will be described later.

4 EXPERIMENTS

All experiments are conducted on a NVIDIA 1080 Ti GPU and it takes about 14 hours for training 14 epochs. We implement the framework partially based on [12].

4.1 Dataset

4.1.1 Statistics. We employ the CelebA dataset for benchmark purposes. CelebA contains 202, 599 facial images from 10, 177 identities. It is about 13 times larger than the Shoes dataset (14, 765) studied in [12]. Each face image is labeled with 40 binary attributes, such as "big nose" and "bald".The dataset contains 115, 114 unique combinations of attributes and the top 10 frequency of identical set of

Figure 3: Sample images from the CelebA dataset.

attributes are 358, 203, 200, 184, 182, 181, 147, 143, 142, 137. However, due to the different poses and angles the same person might have on different images, we use the whole dataset to train and test our model.

4.1.2 Reasons to Use CelebA. Using attributes can help avoid ambiguity in human perception as discussed in Subsection 2.2. CelebA also covers various ethnicities and genders, making it popular among facial image researchers [10, 31]. Its binary attributes, massive coverage, and wide acceptance made the dataset a sufficient choice for our purposes.

4.2 Model Setup

We first experiment with different selections of hyper-parameters in Table 2, and we use the best selection where the constant margin m is set to 2.0. Then, we experiment with data pre-proscessing and reshape the images from 218×178 to 308×256 and zero-center them before extracting their features. We use the first 180,000 images sorted by name of the files as training set and the rest as testing set. There are 852 (8.37% of all 10,177 identities) individuals whose images will appear in both training set and testing set. The number of rounds is fixed and set to 5. The learning rate is set to 0.001. We use adam [17] as optimizer and the value of weight decay is set to 0.0.

4.3 Metrics

The metric is the ranking percentile directly referred from [12]. The reason why we do not use precision is that for each query (target image), there is only one relevant answer in a huge pool (200,000). Unlike QA system or search engines where multiple items can be labeled as relevant, this task is challenging because it asks for highly refined retrieval to get the single relevant answer. Therefore, we calculate the ranking percentile of the target image in the whole search space by their $L2$ distance of representations (which are processed offline using SE-ResNet model). Note that even if there are some images have exactly the same attributes with the target images and we rank them top, we only count the ranking percentile of the target image, which might be lower. The higher percentile is, the more accurate the model is, and the more likely the model can retrieve the correct target image (even though there might be some images with the same attributes that are not target image). We don't do per-attributes evaluation. And more images share the same attributes will only degrade the model performance rather than cheating or make the number higher.

4.4 Baseline

Attribute-based Retrieval: To retrieve a target image in mind, we can refer to its attributes and search in our database and return an image with closest description. Since in our scenario, the attributes of images are not known in advance, so we split the dataset to train a forty-dimensional classifier on the training set and test its performance on the test set. Note that the training set and testing set are the same for baseline experiment for fair comparison. The evaluation metrics are a bit different that instead of calculating $L2-$ distance, we sort the images in the database based on the number of matched attributes between them and the target image and report the ranking position of it.

4.5 Ablation Studies

4.5.1 Different Types of Input.

- Full disclosure with attributes of candidate image: Unlike progressive disclosure, this setting reflects an extreme situation where the thorough disclosure is provided by a complete comparison between the target image and candidate image at each round.
- Full disclosure without attributes of candidate image: Instead of encoding the attributes of candidate image and the relevance feedback together, we experiment with the absence of attributes information of reference image. This setting can reduce the cost of providing feedback for real users since they only need to say "unlike" or "like" rather than "unlike the big nose" or "like the curly hair".
- Progressive disclosure: As described in our work.

4.5.2 Different Features of Images. Instead of training and learning the features dynamically, we employ the following pre-trained models to extract the features of images to save us a lot of computational cost.

- OpenFace [1]: A face recognition model trained on Labeled Faces in the Wild (LFW) dataset [15]. The features will be extracted to a 128-dimensional representations.
- SE-ResNet on VGGFace2 [4]: A face recognition model trained on VGGFace2 dataset. The features will be extracted to a 256-dimensional representations.
- SE-ResNet on CelebA: Based on SE-ResNet on VGGFace2, this model is fine-tuned on CelebA to classify 40 attributes. Apart from designing it as our baseline experiment, we extract the features from its last layer in our model which is a 256-dimensional representations.

4.6 Results and Analysis

4.6.1 Baseline and Our Model. Realistically, there might be multiple samples that share the same combination of attributes, they will be sorted into a consecutive sequence in the database. In this

case, retrieving any of them would be considered as a valid operation for the system. Thus, we report in Table 1 the position of the head and tail of the sequence as the upper bound and lower bound respectively and calculate the expectation by mean value. Note that all other settings are the same and the best.

Method	Upper Bound	Lower Bound	Expectation
Baseline	98.79%	97.09%	98.19%
Ours	N/A	N/A	**98.66%**

Table 1: Ranking percentile for baseline and our model.

The results demonstrate a better performance from our model. Though the numbers might look close to each other, the absolute difference, 0.47%, will be amplified by the enormous number of image pool. For large dataset such as CelebA which contains more than 200, 000 images, this absolute difference means more than 940 images are examined and excluded as irrelevant samples in our model. We believe this is of significant importance for users to save their efforts of looking at more than 940 images unnecessarily.

4.6.2 Explorations of Hyper-parameters. We experiment on the value of margin m in the loss objective. Also we experiment with different ways of pre-processing the images to obtain the best performance. Note that the feature extractor we use here is SE-ResNet on VGGFace2 and full disclosure is provided in each round. When experimenting with margin, we set reshaped size as $214 * 178$. When experimenting with reshaped size, we set margin as 2.0.

Margin	Percentile	Reshaped Size	Percentile
2.0	**92.82%**	214*178 (raw)	92.82%
1.0	92.43%	385*320	93.21%
0.5	92.21%	308*256	**93.40%**
0.1	88.16%	224*224 (crop)	91.57%

Table 2: Ranking percentile is reported on testing set on the final round of the best epoch.

4.6.3 Different Choices of Feedback. As shown in Figure 4, without using attributes of candidate image, the results is limited. While applying progressive disclosure fails to utilize information completely in the beginning, as the disclosure is progressively enhanced, it ultimately outperforms slightly better than full disclosure with attributes of candidate image. When using full disclosure at each round, the growth in ranking percentile soon stagnates at the second round. On the contrary, using progressive disclosure continuously climb and does not converge till the fourth round.

While the ranking percentiles are very close between full disclosure and progressive disclosure with image attributes, we can further investigate their performances by referring to their loss curves in Figure 5. It is obvious that in the beginning, using progressive disclosure incurs more loss than the other two which matches with their performances at round 1. However, at round 5, the loss for progressive disclosure drops to 0.0297 while it is 0.0589, about 1.98 times as much as the former, for full disclosure with image attributes. Combining the loss with their performances at

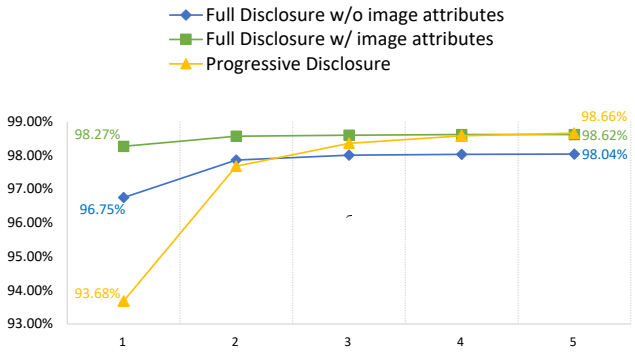

Figure 4: Ranking percentile on testing set in 5 rounds under different settings: full disclosure with and without attributes of candidate image and progressive disclosure.

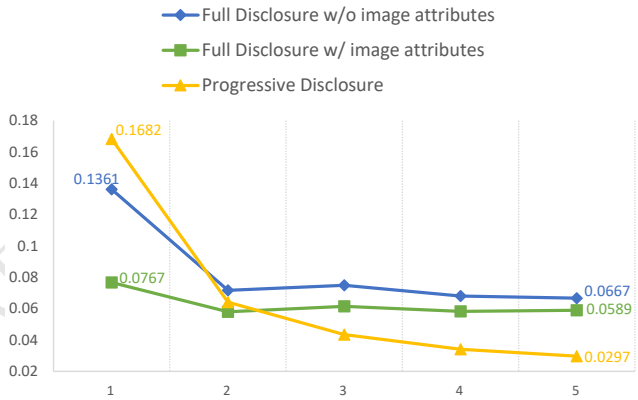

Figure 5: Loss on testing set in 5 rounds under different settings: full disclosure with and without attributes of candidate image and progressive disclosure.

round 5, we can conclude that using progressive disclosure is more capable of fitting the data.

4.6.4 Different Ways of Feature Extractions. Except the different features extracted from various networks, we keep all other settings the same as our best one. In Table 3, there is an apparent superiority in using SE-ResNet on CelebA. The results reveal that the features extracted might have more similar representations with those of attributes, enabling the model to learn their connections well afterwards.

Features	Percentile	Loss
OpenFace	87.02%	0.1225
SE-ResNet on VGGFace2	95.80%	0.0434
SE-ResNet on CelebA	**98.66%**	**0.0297**

Table 3: Ranking percentile and loss are reported on testing set on the final round of the best epoch.

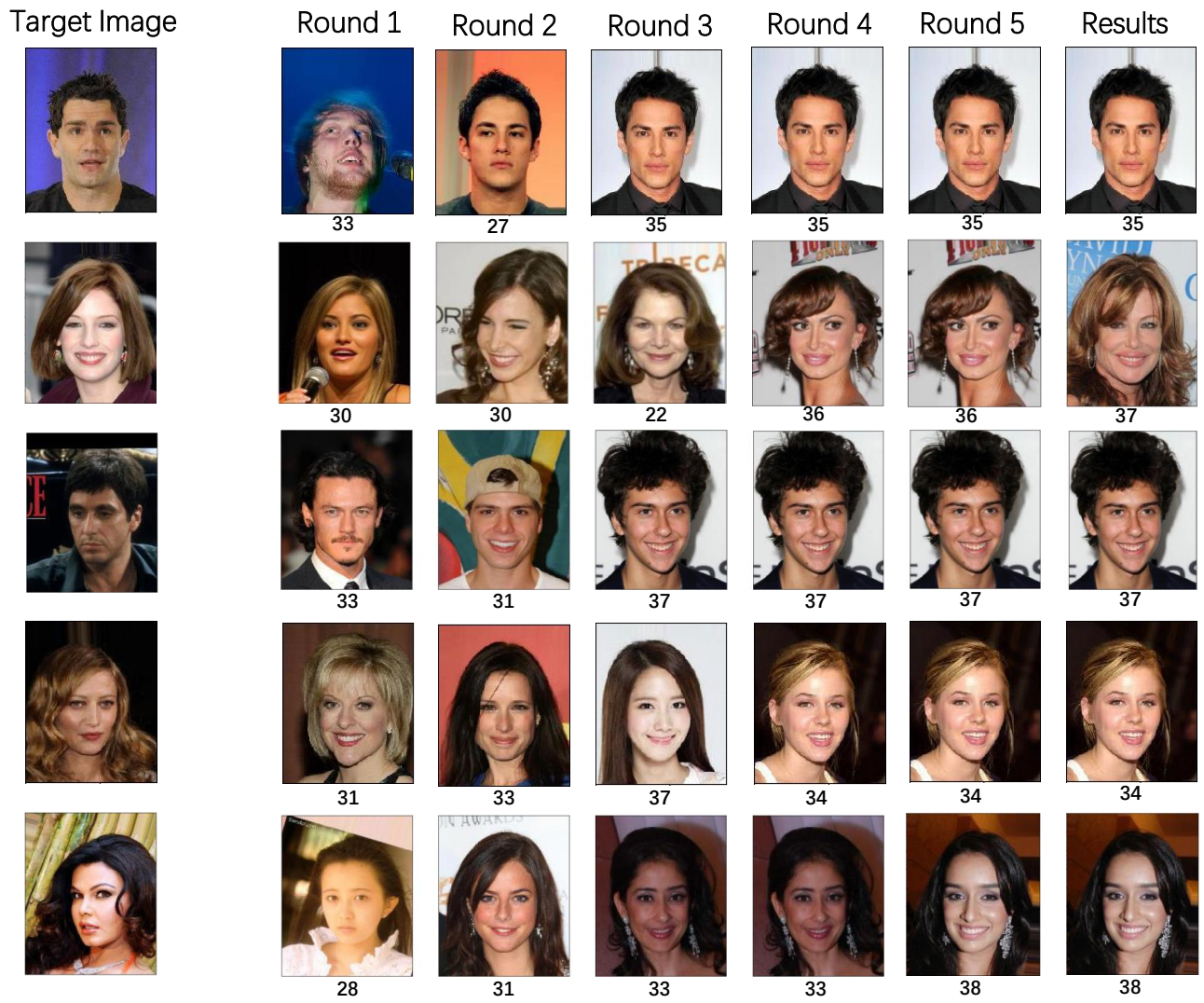

Figure 6: Examples of interactions between users and the proposed facial image retrieval framework with progressive relevance feedback on CelebA.

4.7 Visualizations

Apart from exhibiting the target image and candidate images during interaction between the system and users, we also calculate the number of attributes that are the same between the target image and the candidate image at each round to provide further investigation from another perspective.

From the visualization results, there are some interesting discoveries. We first find out that for target images of male, the system is more likely to converge at an early stage like at round three (see the first and the third row). However, when it comes to female target images, the system could still change at the last minute after all rounds for a better result (see the second row). These differences might come from the distribution of the dataset, where there are more diversity in female images than male images and it enables the system to refine the retrieval results with finer granularity.

Another interesting thing is that, we might assume that the increase in the number of matched attributes is consistent with the improvements in performance. However, although the number of matched attributes is indeed increasing in most cases, it is not always true. For example, from round 3 to round 4 at row four, the number of attributes drops from 37 to 34. Despite that the image from round 4 has 3 less matched attributes, it definitely seems to be a better match with the target image. At least they all share a western look with blond hair and high cheekbones while the image from round 3 is a typical East Asian face. Likewise, for images from round 2 to round 3 at row two and for images from round 1 to round 2 at row one and row three, the decline in the number of matches attributes essentially leads to a visible better results. This may indicate that our model has the advantage of combining

the image features together with attributes information for better retrieval performances.

5 CONCLUSION

In our work, we shed light on facial image retrieval problem and propose an end-to-end interactive framework with progressive disclosure. We also explore different settings in various scenarios and applications. Though not perfect, our work is sufficient to deal with many cases with over 98% ranking percentile. In the future, this retrieval problem can be upgraded to a conditional generation task that helps suspect sketch. Moreover, the forms of feedback can be expanded and enriched to capture the subtlety, such as fine-grained attributes or even verbal descriptions which enables smoother and more nature approach. The abundant ambiguity in the perception of facial images makes it particularly difficult but crucial to have an intelligent and accurate method to bridging the semantic gap and we hope our work can be a stepping stone for interested researchers.

6 ACKNOWLEDGEMENT

I am grateful for all the support from Yuwen Xiong on this work. Though not an official author, his encouragements and guidance to me on this work is one of the most important reasons of me finishing this paper.

REFERENCES

[1] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *OpenFace: A general-purpose face recognition library with mobile applications*. Technical Report. CMU-CS-16-118, CMU School of Computer Science.

[2] Caroline Blais, Rachael E. Jack, Christoph Scheepers, Daniel Fiset, and Roberto Caldara. 2008. Culture Shapes How We Look at Faces. *PLOS ONE* 3, 8 (8 2008), 1–8. https://doi.org/10.1371/journal.pone.0003022

[3] Charity Brown and Toby J Lloyd-Jones. 2005. Verbal facilitation of face recognition. *Memory & Cognition* 33, 8 (2005), 1442–1456.

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 67–74.

[5] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 2951–2960.

[6] Jan B Engelmann and Marianna Pogosyan. 2013. Emotion perception across cultures: the role of cognitive mechanisms. *Frontiers in psychology* 4 (2013), 118.

[7] M. Ferecatu and D. Geman. 2007. Interactive Search for Image Categories by Mental Matching. In *2007 IEEE 11th International Conference on Computer Vision*. https://doi.org/10.1109/ICCV.2007.4409072

[8] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Qian Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. 1995. Query by image and video content: the QBIC system. *Computer* 28, 9 (Sep. 1995), 23–32. https://doi.org/10.1109/2.410146

[9] D Geman and R Moquet. 2000. A stochastic feedback model for image retrieval. In *Proc. RFIA*, Vol. 3. 173–180.

[10] Yağmur Güçlütürk, Umut Güçlü, Rob van Lier, and Marcel AJ van Gerven. 2016. Convolutional sketch inversion. In *European Conference on Computer Vision*. Springer, 810–824.

[11] V. N. Gudivada and V. V. Raghavan. 1995. Content based image retrieval systems. *Computer* 28, 9 (Sept. 1995), 18–22. https://doi.org/10.1109/2.410145

[12] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems*. 678–688.

[13] S. Harada, Y. Itoh, and H. Nakatani. 1997. Interactive image retrieval by natural language. *Optical Engineering* (Dec. 1997). https://doi.org/10.1117/1.601567

[14] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. *IEEE Conference on Computer Vision and Pattern Recognition*.

[15] 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.

[16] Q. Iqbal and J. K. Aggarwal. 2002. CIRES: a system for content-based retrieval in digital image libraries. In *7th International Conference on Control, Automation, Robotics and Vision, 2002. ICARCV 2002.*, Vol. 1. 205–210 vol.1. https://doi.org/10.1109/ICARCV.2002.1234821

[17] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (12 2014).

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

[19] Adriana Kovashka, Devi Parikh, and Kristen Grauman. 2015. WhittleSearch: Interactive Image Search with Relative Attribute Feedback. *International Journal of Computer Vision* 115, 2 (01 Nov 2015), 185–210. https://doi.org/10.1007/s11263-015-0814-0

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[21] Wei-Ying Ma and B.S. Manjunath. 1999. NeTra: A toolbox for navigating large image databases. *Multimedia Systems* 7, 3 (01 May 1999), 184–198. https://doi.org/10.1007/s005300050121

[22] Sean D MacArthur, Carla E Brodley, Avinash C Kak, and Lynn S Broderick. 2002. Interactive content-based image retrieval using relevance feedback. *Computer Vision and Image Understanding* 88, 2 (2002), 55–75.

[23] Sébastien Miellet, Luca Vizioli, Lingnan He, Xinyue Zhou, and Roberto Caldara. 2013. Mapping Face Recognition Information Use across Cultures. *Frontiers in Psychology* 4 (2013), 34. https://doi.org/10.3389/fpsyg.2013.00034

[24] David A Monroe. 2009. Method for incorporating facial recognition technology in a multimedia surveillance system. US Patent 7,634,662.

[25] Yong Rui, Thomas S Huang, and Sharad Mehrotra. 1997. Relevance feedback techniques in interactive content-based image retrieval. In *Storage and Retrieval for Image and Video Databases VI*, Vol. 3312. International Society for Optics and Photonics, 25–36.

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

[27] Piotr Sorokowski, Krzysztof Kościński, and Agnieszka Sorokowska. 2013. Is Beauty in the Eye of the Beholder but Ugliness Culturally Universal? Facial Preferences of Polish and Yali (Papua) People. *Evolutionary Psychology* 11, 4 (2013), 147470491301100400. https://doi.org/10.1177/147470491301100414 arXiv:https://doi.org/10.1177/147470491301100414

[28] Bing Wang, Xin Zhang, and Na Li. 2006. Relevance feedback technique for content-based image retrieval using neural network learning. In *2006 International Conference on Machine Learning and Cybernetics*. IEEE, 3692–3696.

[29] Ka-Man Wong, Kwok-Wai Cheung, and Lai-Man Po. 2005. MIRROR: an interactive content based image retrieval system. In *2005 IEEE International Symposium on Circuits and Systems*. IEEE, 1541–1544.

[30] Eric Zavesky and Shih-Fu Chang. 2008. CuZero: Embracing the Frontier of Interactive Visual Search for Informed Users. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval (MIR '08)*. ACM, 237–244. https://doi.org/10.1145/1460096.1460136

[31] Kaipeng Zhang, Lianzhi Tan, Zhifeng Li, and Yu Qiao. 2016. Gender and smile classification using deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 34–38.