Multimodal Image Retrieval Based on Eyes Hints and Facial Description Properties

Yuelong Li[1,2], Junyu Bi[1,2], Tongshun Zhang[1,2], and Jianming Wang[1,2](✉)

[1] Tianjin Key Laboratory of Independent Intelligent Technology and System,
Tianjin, People's Republic of China
wangjianming@tiangong.edu.cn
[2] School of Computer Science and Technology, Tiangong University,
Tianjin, People's Republic of China

Abstract. Eyes are the most prominent visual components on human face. Obtaining the corresponding face only by the visual hints of eyes is a long time expectation of people. However, since eyes only occupy a small part of the whole face, and they do not contain evident identity recognition features, this is an underdetermined task and hardly to be finished. To cope with the lack of query information, we enroll extra face description properties as a complementary information source, and propose a multimodal image retrieval method based on eyes hints and facial description properties. Furthermore, besides straightforward corresponding facial image retrieval, description properties also provide the capacity of customized retrieval, i.e., through altering description properties, we could obtain various faces with the same given eyes. Our approach is constructed based on deep neural network framework, and here we propose a novel image and property fusion mechanism named Product of Addition and Concatenation (PAC). Here the eyes image and description properties features, respectively acquired by CNN and LSTM, are fused by a carefully designed combination of addition, concatenation, and element-wise product. Through this fusion strategy, both information of distinct categories can be projected into a unified face feature space, and contribute to effective image retrieval. Our method has been experimented and validated on the publicly available CelebA face dataset.

Keywords: Image retrieval · Multimodal information fusion · Eyes hints · Deep neural network · Customized face query

1 Introduction

Image is a very convenient tool to store and demonstrate visual information. Hence, how to query and obtain wanted images from giant image datasets is

J. Wang—Supported by the National Natural Science Foundation of China (No. 61771340), the Tianjin Natural Science Foundation (No. 18JCYBJC15300), the Program for Innovative Research Team in University of Tianjin (No. TD13-5032), and the Tianjin Science and Technology Program (19PTZWHZ00020).

Y. Peng et al. (Eds.): PRCV 2020, LNCS 12306, pp. 151–163, 2020.
https://doi.org/10.1007/978-3-030-60639-8_13

an attractive research topic both academically and industrially [2,19,20]. But since the image belongs to a kind of unstructured information, image retrieval is never an easily conducted task. Furthermore, in the past few decades, accompanied by the rapid development of image capturing and collecting techniques, the number of available images is booming astonishingly. Hence how to effectively acquire wanted images from tremendous candidates is attracting more and more attentions.

Eyes are the most important facial visual features, and image retrieval based on eyes is a long history interesting research topic. It has wide practical value and application significance in the fields of public safety, blind date matching, beauty retouching, and so forth. However, due to their quite limited region area, the unique identity information that could be conveyed by eyes hints is seriously restricted, and hence, by this information source alone, accurate image retrieval is hardly achievable. In order to deal with this problem, in this paper, we designed a multimodal image retrieval method based on eyes hints combined with facial description properties. In our approach, both image and text information are unified as the query source, hence more effective identity information can be utilized to guide the searching procedure. On the other hand, the introduction of description properties can not only directly improve the accuracy of image retrieval, but also introduce more personality and customization. The users can perform customized retrieval by alternating the text facial descriptions, as shown in Fig. 1. Here it can be observed that the image query procedure is customized by the yellow and green background facial description properties respectively.

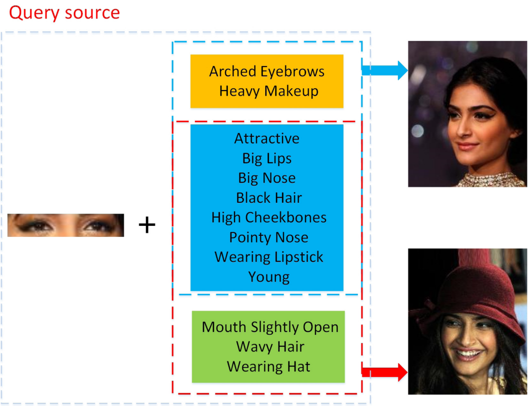

Fig. 1. A demonstration of the image retrieval based on both eyes hints and facial description properties. Here, both images of the rightmost column are the retrieval output, where the top one is based on the far left eyes image and the yellow and blue background facial description properties, while the bottom one is on the same eyes image and the blue and green background properties. (Color figure online)

In recent years, there have been some research findings involving the combination of vision and text information [11]. But most of them focus on visual question answering [4,9], cross-modal retrieval and image captioning [7,27]. The works that are directly related with the multimodal image retrieval specifically discussed in this paper are relatively rare.

In order to solve the mentioned problem, we propose a novel image and property information fusion mechanism. After obtaining the high-level semantic features of the eye images and the description properties respectively through the encoders, we do element wise addition and concatenation between different types of features, then product the results to achieve effective multimodal information fusion. This proposed method is named the Product of Addition and Concatenation (or PAC for short). We will give the detailed introduction in Sect. 3. The approach is experimented on the public available CelebA dataset, and satisfactory performance is achieved.

To summarize, the main contribution of this paper is threefold:

(1) We propose a multimodal image retrieval method based on eyes hints and facial description properties.
(2) A novel cross-modal feature fusion mechanism is introduced which could effective combine both vision and text information.
(3) The multimodal information fusion capacity of deep neural network is beneficially explored.

The rest parts of the paper are organized as follows: Sect. 2 reviews related work; The proposed PAC based image retrieval method will be presented in Sect. 3; Validation experiments are shown in Sect. 4; Sect. 5 concludes the paper.

2 Related Work

Vision Question Answering (VQA). VQA is a typical application that combines both image and text information. Its goal is to automatically generate the natural language answers according to the image and natural language question input. There are generally two ways to combine features in VQA, one of which is the direct combinations, such as, concatenation, element wise multiplication, and element wise addition [14]. Zhou et al. [28] introduce to use the bag-of-words to express the questions, the GoogleNet to extract visual features, and then direct connect the two features. Agrawal et al. [1] worked out to input the product of two feature vectors into a multi-layer perceptron of two hidden layers. Another way to combine features is using bilinear pooling or related schemes in a neural network framework [14]. Fukui et al. [8] proposed to the Multimodal Compact Bilinear (MCB) pooling to combine image features and text features. Due to the high computational cost of MCB, a multimodal low-rank bilinear pooling strategy (MLB) is worked out, where the Hadamard product and linear mapping are used to achieve approximate bilinear pooling [15].

Cross-Modal Retrieval. Cross-modal retrieval uses a certain modal sample to search for other modal samples with similar semantics. The traditional method generally obtains the mapping matrix from the paired symbiotic information of different modal sample pairs, and maps the features of different modalities into a common semantic vector space. Li et al. [17] introduced a cross-modal factor analysis method. Rasiwasia et al. [13,21] designed to apply the canonical correlation analysis (CCA) to the cross-modal retrieval between text and images. In recent years, many researches use deep learning to extract the effective representations of different modalities at the bottom layer, and establish semantic associations of different modalities at the top layer [6,16]. Wei et al. [26] worked out an end-to-end deep canonical correlation analysis method to retrieve text and images. Gu et al. [10] enrolled the Generative Adversarial Networks and Reinforcement Learning for cross-modal retrieval. In their work, the generation process is integrated into the cross-modal feature embedding. Here, not only the global features can be learned but also the local features. Wang et al. [24] believed that the previous methods rarely consider the interrelationship between image and text information during calculating the similarity, so they proposed the Cross-modal Adaptive Message Passing (CAMP) method.

Metric Learning. The goal of metric learning is to maximize the inter-class variations while minimize the intra-class variations, and it is quite common in pattern recognition applications. In neural network based approaches, LeCun et al. [12] designed the contrastive loss to increase inter-class variations. Schroff et al. [22] proposed the triplet loss. Then a large number of subsequent metric learning methods are worked out based on the triplet loss, such as the quadruplet loss [5].

3 Method

As mentioned in the introduction section, our goal is to achieve multimodal image retrieval based on both eyes hints and facial description properties. Here, how to effectively combine the query information coming from distinct categories is the most critical problem. Since both vision and text information are complex and comprehensive, we designed a neural network based information fusing and processing strategy. The main training pipeline is demonstrated in Fig. 2.

Specifically, first, the query eyes image $\boldsymbol{x}$ is encoded by a Light CNN [25]. Light CNN is a light-weight, noise-removable network proposed for face recognition. Here the query eyes image is transformed into 2D spatial feature vector $f_{\text{img}}(\boldsymbol{x}) = \phi_{\boldsymbol{x}} \in \mathbb{R}^{W \times H \times C}$, where W is the width, H is the height, and $C = 512$ is the number of feature channels. Note that we modify the size of the last fully connected layer of Light CNN from 256 to 512 to make the number of channels of image and text features the same. Second, we encode the facial description properties $\boldsymbol{t}$ with LSTM [28]. We define $f_{\text{text}}(\boldsymbol{t}) = \phi_{\boldsymbol{t}} \in \mathbb{R}^{L \times S \times d}$ to be the hidden state at the final time step, where L is the sequence length, S is the batch size, and $d = 512$ is the hidden layer size. Finally, both $\phi_{\boldsymbol{x}}$ and $\phi_{\boldsymbol{t}}$ are combined into

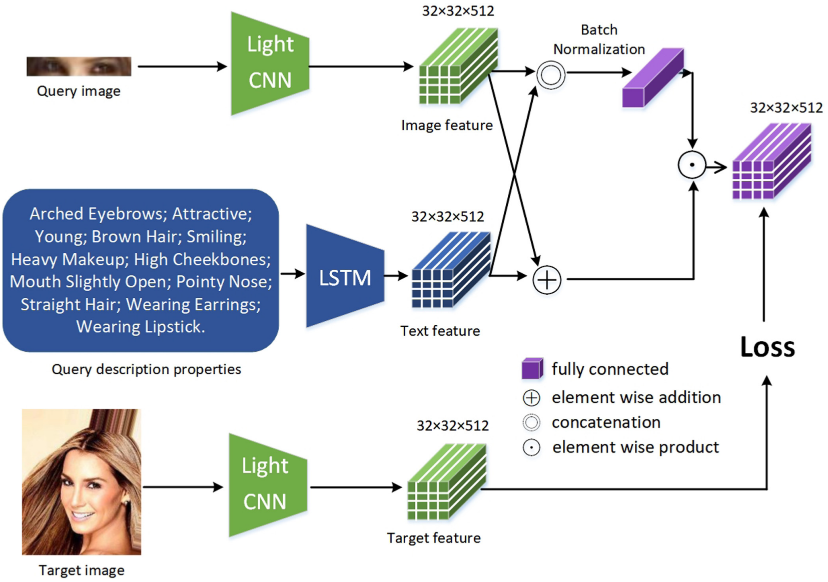

Fig. 2. The training pipeline of our multimodal image retrieval method based on eyes hints and facial description properties.

$\phi_{\boldsymbol{xt}} = f_{\text{combine}}(\phi_{\boldsymbol{x}}, \phi_{\boldsymbol{t}})$ with the proposed PAC method, which will be introduced in Sect. 3.1 in details.

On the other hand, during image retrieval, we calculate the similarity of the fused feature and that of candidate images by cosine distance, and then sort to get the face images that best meets the query conditions.

3.1 Feature Fusion by PAC

In order to effectively achieve multimodal information fusion, we explored a comprehensive combination strategy. Since the element wise addition and catenation are the most common direct way of fusion, our first glance is to contain both operation continuously. But because the dimensions of both information may be different, a co-dimensionalization approach is worked out. In detail, a convolution operation is enrolled to adjust the latitude of the concatenated feature matrix. At the same time, the sigmoid function is introduced to avoid taking too large a value. After the co-dimensionalization, the two types of combination methods are fused again by bitwise multiplication to obtain the final fusion feature. The whole mentioned processing is named the Product of Addition and Concatenation. Specifically,

$$\phi_{\boldsymbol{xt}} = f_{\text{add}}(\phi_{\boldsymbol{x}}, \phi_{\boldsymbol{t}}) \odot f_{\text{concat}}(\phi_{\boldsymbol{x}}, \phi_{\boldsymbol{t}}), \tag{1}$$

where $\odot$ is element wise product, $\phi_{\boldsymbol{x}}$ denotes image feature, $\phi_{\boldsymbol{t}}$ is text feature, $\phi_{\boldsymbol{xt}}$ is the fused feature.

$$f_{\text{add}}(\phi_{\boldsymbol{x}}, \phi_{\boldsymbol{t}}) = W_{\text{img}}\phi_{\boldsymbol{x}} + W_{\text{text}}\phi_{\boldsymbol{t}}, \tag{2}$$

where W_{img}, W_{text} are learnable weights to balance both components.

$$f_{\text{concat}}(\phi_{\boldsymbol{x}}, \phi_{\boldsymbol{t}}) = \sigma(W_g \circ [\phi_{\boldsymbol{x}}, \phi_{\boldsymbol{t}}]), \tag{3}$$

where $[\phi_{\boldsymbol{x}}, \phi_{\boldsymbol{t}}]$ is matrix concatenate, and σ denotes the sigmoid function. We define $W_g \circ$ to be a series of convolution operations: first, the concatenated matrix is normalized in batches, then goes through the ReLU activation function, and finally its the number of channels is reduced from 1024 to 512 through the fully connected layer.

3.2 Loss Function

Clearly, the goal of our training is to make the fused features of faces in the same identity and state closer, while pulling apart the features of distinct images. For this task, we employ a triplet loss, which contains anchors, positive samples, and negative samples. When selecting the triples, we choose negative samples that are the same identity but different states as the anchor in order to enable that the network can distinguish the differences in face states. We define this triple as $T_{\text{state}}(f(x_i^a), f(x_i^p), f(x_i^n))$, where x_i^a is anchor, x_i^p is positive sample, and x_i^n is negative sample. $f(x)$ is embedding constrained to live on the d-dimensional hypersphere [22], where $d = 512$, i.e. $\|f(x)\|_2 = 1$. Similarly, we define $T_{\text{identity}}(f(x_j^a), f(x_j^p), f(x_j^n))$ to denote a triple whose negative sample and anchor have different identities but similar status. We then use the following triplet loss:

$$\begin{aligned} L = &\sum_{i}^{N_{\text{state}}} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha\right]_+ \\ &+ \sum_{j}^{N_{\text{identity}}} \left[\left\|f(x_j^a) - f(x_j^p)\right\|_2^2 - \left\|f(x_j^a) - f(x_j^n)\right\|_2^2 + \alpha\right]_+, \end{aligned} \tag{4}$$

where α is a margin that is enforced between positive and negative pairs. N_{state} is the number of T_{state}, and N_{identity} is the number of T_{identity}. We believe that splitting loss into two parts, status and identity, is beneficial for the network to combine two types of data.

4 Experiments

In this section, the proposed multimodal image retrieval approach based on eyes hints and facial description properties will be experimented both quantitatively and qualitatively.

4.1 Experiment Configurations

Datasets. The experiments are conducted on a widely used face attribute dataset CelebA [18], which contains 202,599 images of 10,177 celebrity identities, and each of its image contains 40 attribute tags. Our experiments utilize 35 of them which do not related with eyes. As shown in Fig. 1, the query eyes images are in a single rectangle shape. We adopt the same subset division introduced in [18], where 40,000 images constitute the testing set, while all the left images form the training set.

Implementation Details. The experiments are realized by PyTorch code. The training is run for 210k iterations with a start learning rate 0.01.

Evaluation Metrics. As to the performance evaluation, we use the most commonly used evaluation metric R@K, which is the abbreviation for Recall at K and is defined as the proportion of correct matchings in top-k retrieved results [24]. Specifically, we adopt the R@1, R@5, R@10, R@50 and R@100 as our evaluation metrics.

4.2 Quantitative Results

In order to objectively evaluate the method performance, several classical information fusion approaches are enrolled as comparison. The MLB is a very classic multimodal fusion method in the field of VQA, which is based on the Hadamard product [15]. The MUTAN is a fusion method based on tensor decomposition applied in the field of VQA [3]. The TIRG method converts multimodal features into two parts called gating and the residual features, where the gating connection uses the input image features as the reference for the output composite features, and the residual connection represents the modifications in the feature space [23].

Table 1. Image retrieval performance on the CelebA dataset.

Methods	R@1	R@5	R@10	R@50	R@100
Image only (light CNN) [25]	3.3	7.6	10.5	20.0	25.7
Text only (LSTM) [28]	1.9	5.6	9.0	24.3	34.8
MLB [15]	3.1	9.1	14.3	32.2	44.7
MUTAN [3]	3.0	9.5	13.2	31.0	42.4
TIRG [23]	3.5	10.4	15.4	33.4	45.3
PAC (ours)	**4.1**	**11.4**	**16.3**	**34.4**	**45.8**

In order to fairly compare their performance with that of ours, during experiments, only the feature fusion part is distinct, while all other components are exactly the same as that of ours.

Table 1 presents the detailed performance. It can be observed that our method is evidently better than other methods on each evaluation indicator. Here, one thing to be mentioned is that, since facial description is not a kind of unique identity information, while the ground truth image used as the evaluation benchmark is unique, the overall performance may not be quite impressive. But this is caused by the nature of this task, rather than the method adopted.

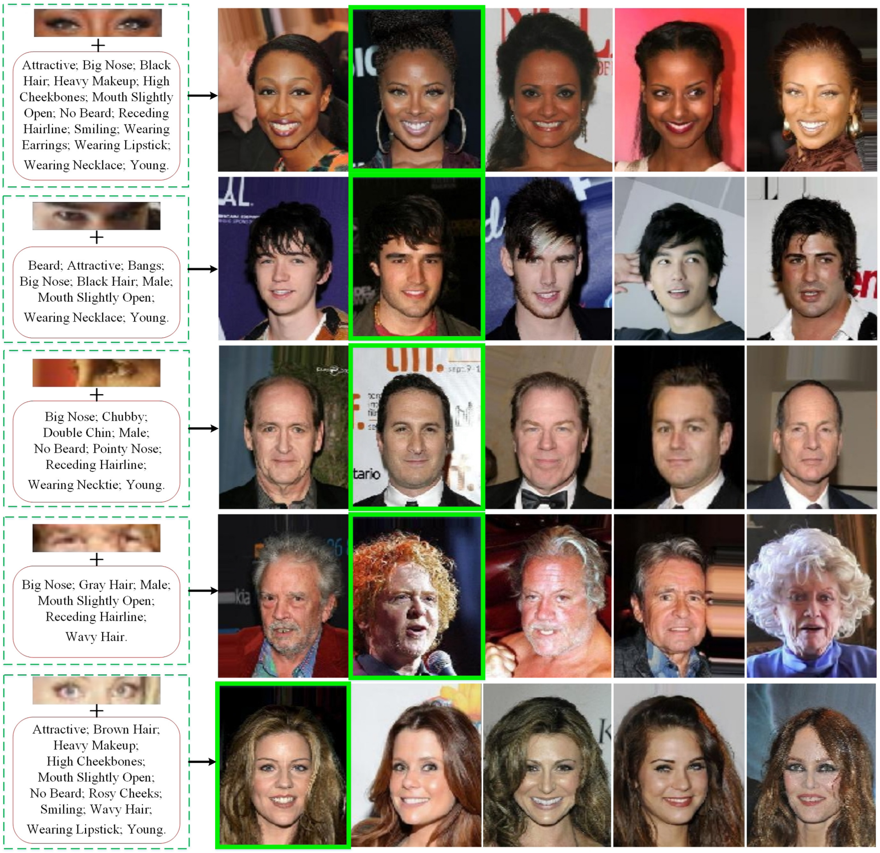
+
Attractive; Big Nose; Black Hair; Heavy Makeup; High Cheekbones; Mouth Slightly Open; No Beard; Receding Hairline; Smiling; Wearing Earrings; Wearing Lipstick; Wearing Necklace; Young.
+
Beard; Attractive; Bangs; Big Nose; Black Hair; Male; Mouth Slightly Open; Wearing Necklace; Young.
+
Big Nose; Chubby; Double Chin; Male; No Beard; Pointy Nose; Receding Hairline; Wearing Necktie; Young.
+
Big Nose; Gray Hair; Male; Mouth Slightly Open; Receding Hairline; Wavy Hair.
+
Attractive; Brown Hair; Heavy Makeup; High Cheekbones; Mouth Slightly Open; No Beard; Rosy Cheeks; Smiling; Wavy Hair; Wearing Lipstick; Young.

Fig. 3. A few image retrieval outputs on the CelebA dataset. In the green dotted frame of the first column, the eyes image and the description are the query condition. The next five columns are the top five search results obtained by our method, where the ground truth image is surrounded by a solid green border. (Color figure online)

4.3 Qualitative Results

A few of retrieved images are shown in Fig. 3. It can be observed that generally the top five worked out candidates all conform to the query eyes and facial properties. Let's take the first row as an example. It can be easily found that all the five images are "Attractive, Big Nose, Black Hair, Heavy Makeup, High Cheekbones, Mouth Slightly Open, No Beard, Receding Hairline, Smiling, Wearing Earrings, Wearing Lipstick, Wearing Necklace, Young", while their eyes are similar with the query eyes to some extent.

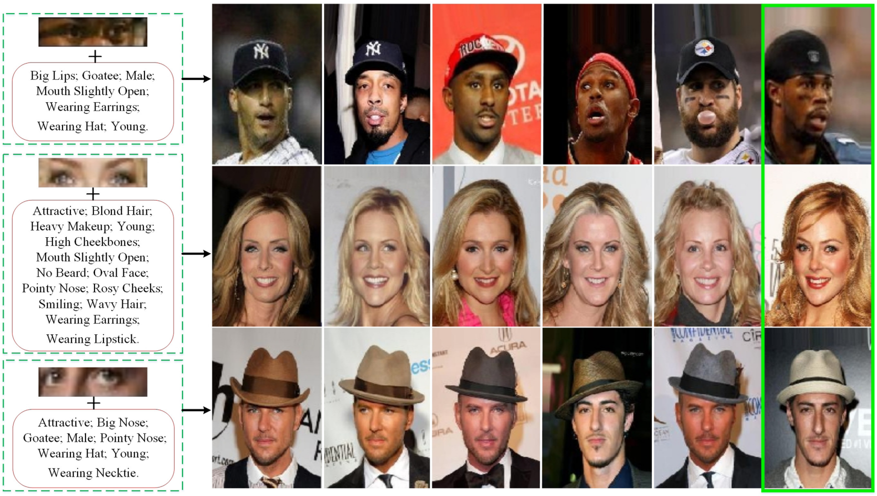
+
Big Lips; Goatee; Male; Mouth Slightly Open; Wearing Earrings; Wearing Hat; Young.
+
Attractive; Blond Hair; Heavy Makeup; Young; High Cheekbones; Mouth Slightly Open; No Beard; Oval Face; Pointy Nose; Rosy Cheeks; Smiling; Wavy Hair; Wearing Earrings; Wearing Lipstick.
+
Attractive; Big Nose; Goatee; Male; Pointy Nose; Wearing Hat; Young; Wearing Necktie.

Fig. 4. A few of "unsuccessful" retrieval output on the CelebA dataset. In the green dotted frame of the first column, the eyes image and the description are the query condition. The next five columns are the top five search results obtained according to our method. The images surrounded by solid green border in the rightmost column are the corresponding ground truth images. (Color figure online)

Figure 4 shows some "unsuccessful" retrievals, which means the ground truth image is not in the top five worked out images. It can be observed that, even in those "unsuccessful" retrievals, the obtained images are still compatible with the query eyes and description properties.

On the other hand, Fig. 5 shows some retrieval output according to the same eyes but distinct description properties. It can be seen that the text query condition can directly influence the retrieval output. Hence, it can be claimed that customized image retrieval is achievable by our method.

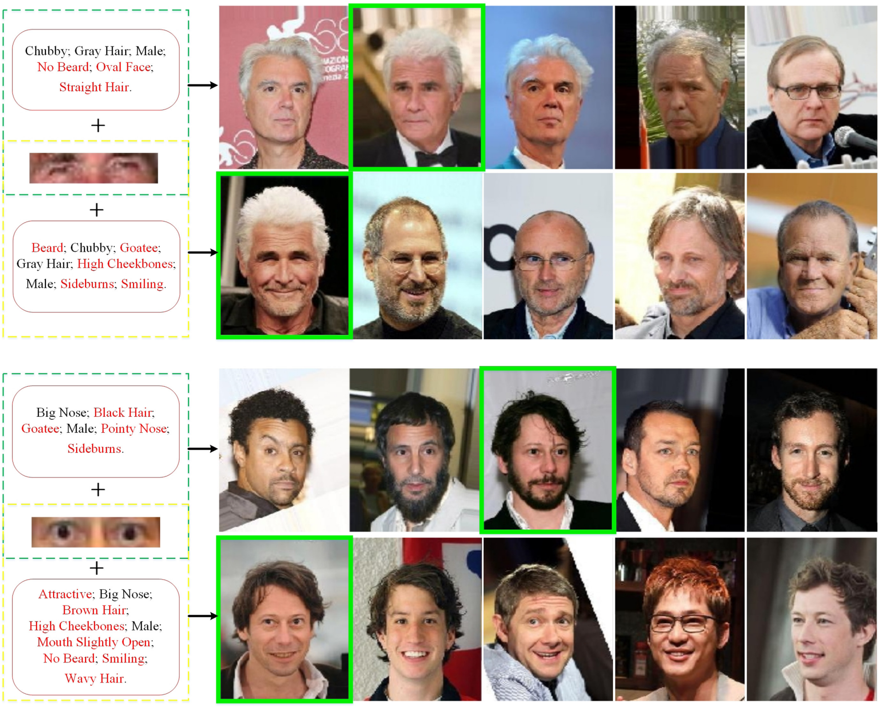
Chubby; Gray Hair; Male; No Beard; Oval Face; Straight Hair.
+
+
Beard; Chubby; Goatee; Gray Hair; High Cheekbones; Male; Sideburns; Smiling.
Big Nose; Black Hair; Goatee; Male; Pointy Nose; Sideburns.
+
+
Attractive; Big Nose; Brown Hair; High Cheekbones; Male; Mouth Slightly Open; No Beard; Smiling; Wavy Hair.

Fig. 5. Image retrieval outputs with the same eye but distinct description properties as the input. The image of the eyes in the dotted frame in the first column is the query image, and the description is the query text. Note that attribute words marked in red are unique. The next five columns are the top five search results obtained according to our method. The images surrounded by solid green border in the rightmost column are the corresponding ground truth images. (Color figure online)

In addition, we calculated the R@1 of each description properties as well, see Fig. 6. Among them, the average is 0.6874. It can be seen that “Male”, “No Beard”, “Mouth Slightly Open” have the highest recall rate in our model, while “Wearing Necktie”, “Blurry”, “Bald” have relatively lower recall rate. This phenomenon is identical with our intuition because those latter properties are relatively rare in the candidate dataset.

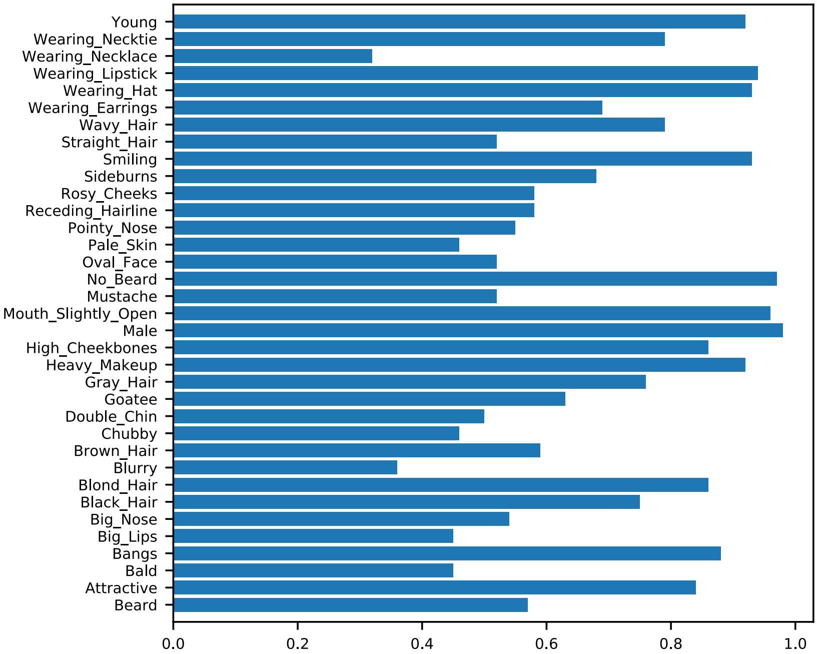

Fig. 6. The accuracy of identifying each properties.

5 Conclusion

In this paper, we propose to use the description properties as supplementary information for eye images in image retrieval tasks. And a novel multimodal information fusion method is worked out for the mission. The effectiveness of the proposed method is verified on a public available dataset, the CelebA. Generally the performance is identical with our expectation. In addition, personalized and customized image retrieval is achievable by the proposed approach. In the near future, we would like to try to extend this method to more general image retrieval problems.

References

1. Antol, S., et al.: VQA: visual question answering. In: International Conference on Computer Vision (2015)
2. Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
3. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: MUTAN: multimodal tucker fusion for visual question answering. In: IEEE International Conference on Computer Vision (2017)

4. Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y.: Counterfactual samples synthesizing for robust visual question answering. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
5. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: IEEE International Conference on Computer Vision and Pattern Recognition (2017)
6. Eisenschtat, A., Wolf, L.: Linking image and text with 2-way nets. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
7. Feng, Y., Ma, L., Liu, W., Luo, J.: Unsupervised image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
8. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)
9. Gao, D., Li, K., Wang, R., Shan, S., Chen, X.: Multi-modal graph neural network for joint reasoning on vision and scene text. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
10. Gu, J., Cai, J., Joty, S., Niu, L., Wang, G.: Look, imagine and match: improving textual-visual cross-modal retrieval with generative models. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
11. Gu, J., Joty, S., Cai, J., Zhao, H., Yang, X., Wang, G.: Unpaired image captioning via scene graph alignments. In: IEEE International Conference on Computer Vision (2019)
12. Hadsell, R., Chopra, S., Lecun, Y.: Dimensionality reduction by learning an invariant mapping. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2006)
13. Hotelling, H.: Relations between two sets of variates (1992)
14. Kafle, K., Kanan, C.: Visual question answering: datasets, algorithms, and future challenges. Comput. Vis. Image Underst. **163**, 3–20 (2017)
15. Kim, J.H., Kim, J., Ha, J.W., Zhang, B.T.: TrimZero: a torch recurrent module for efficient natural language processing. In: Proceedings of KIIS Spring Conference (2016)
16. Kiros, R., Salakhutdinov, R., Zemel, R.: Unifying visual-semantic embeddings with multimodal neural language models. In: International Conference on Machine Learning (2014)
17. Li, D., Dimitrova, N., Li, M., Sethi, I.K.: Multimedia content processing through cross-modal association. In: ACM International Conference on Multimedia (2003)
18. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: International Conference on Computer Vision (2015)
19. Liu, Z., Ping, L., Shi, Q., Wang, X., Tang, X.: DeepFashion: powering robust clothes recognition and retrieval with rich annotations. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
20. Ng, Y.H., Yang, F., Davis, L.S.: Exploiting local features from deep networks for image retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
21. Rasiwasia, N., et al.: A new approach to cross-modal multimedia retrieval. In: ACM International Conference on Multimedia (2010)
22. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)

23. Vo, N., et al.: Composing text and image for image retrieval-an empirical odyssey. In: IEEE International Conference on Computer Vision and Pattern Recognition (2019)
24. Wang, Z., Liu, X., Li, H., Sheng, L., Yan, J., Wang, X., Shao, J.: CAMP: cross-modal adaptive message passing for text-image retrieval. In: International Conference on Computer Vision (2019)
25. Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. IEEE Trans. Inf. Forensics Secur. **13**, 2884–2896 (2018)
26. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
27. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
28. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. Comput. Sci. (2015)