FACE IMAGE RETRIEVAL USING ATTRIBUTE - ENHANCED SPARSE CODEWORDS

E.Sakthivel[1] , M.Ashok kumar[2]

[1]PG scholar, Communication Systems, Adhiyamaan College of Engineeing,Hosur,(India)
[2]Asst. Prof., Electronics And Communication Engg., Adhiyamaan College of Engg.,Hosur,(India)

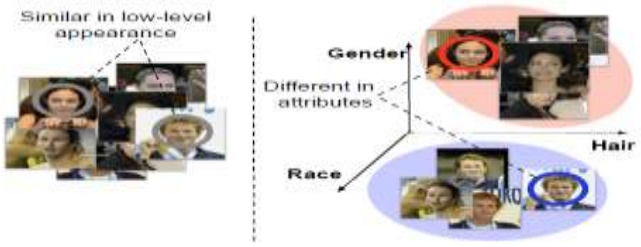
ABSTRACT

Photos with people (e.g., family, friends, celebrities, etc.) are the major interest of users. Thus, with the exponentially growing photos, large-scale content-based face image retrieval is an enabling technology for many emerging applications. In this work, we aim to utilize automatically detected human attributes that contain semantic cues of the face photos to improve content-based face retrieval by constructing semantic codewords for efficient large-scale face retrieval. By leveraging human attributes in a scalable and systematic framework, we propose two orthogonal methods named attribute-enhanced sparse coding and attribute-embedded inverted indexing to improve the face retrieval in the offline and online stages. We investigate the effectiveness of different attributes and vital factors essential for face retrieval. Experimenting on two public datasets, the results show that the proposed methods can achieve up to 43.5% relative improvement in MAP compared to the existing methods.

Index Terms—Face image, human attributes, content-based image retrieval

I. INTRODUCTION

Due to the popularity of digital devices and the rise of social network/photo sharing services (e.g., Facebook, Flickr), there are largely growing consumer photos availablein our life. Among all those photos, a big percentage of them are photos with human faces (estimated more than 60%). The importance and the sheer amount of human face photos make manipulations (e.g., search and mining) of large-scale human face images a really important research problem and enable many real world applications [1], [2].

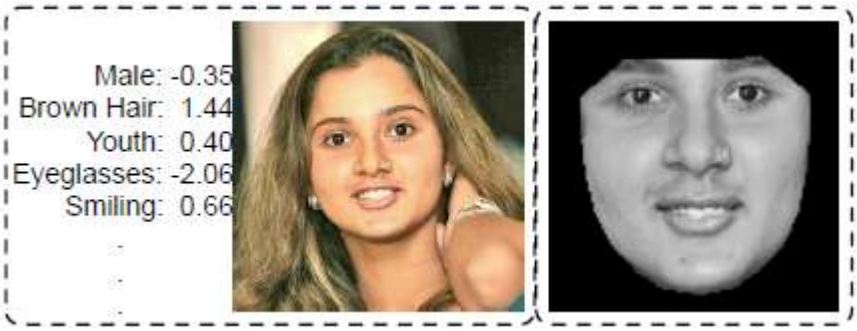


(a) (b)

Fig 1 (a) Because low-level features are lack of semantic meanings, face images of two different people might be close in the traditional low-level feature space. (b) By incorporating high-level human attributes (e.g., gender) into feature representations, we can provide better discriminability for face image retrieval. (Best seen in color)

Our goal in this paper is to address one of the important and challenging problems - large-scale content-based face image retrieval. Given a query face image, content-based face image retrieval tries to find similar face images from a large image database. It is an enabling technology for many applications including automatic face annotation [2], crime investigation [3], etc.

In this work, we provide a new perspective on content-based face image retrieval by incorporating high-level human attributes into face image representation and index structure. As shown in Figure 1, face images of different people might be very close in the low-level feature space. By combining low-level features with high-level human attributes, we are able to find better feature representations and achieve better retrieval results. The similar idea is proposed in [6] using fisher vectors with attributes for large-scale image retrieval, but they use early fusion to combine the attribute scores. Also, they do not take advantages of human attributes because their target is general image retrieval.



(a) (b)

Fig. 2. (a) A face image contains rich context information (hair color, skin color, race, gender, etc.). Using automatic human attribute detection (b)Same image after pre-processing steps for face image retrieval or recognition..

Human attributes (e.g., gender, race, hair style) are high level semantic descriptions about a person. Some examples of human attributes can be found in Figure 2. These results indicate the power of the human attributes on face images. In Table I, we also show that human attributes can be helpful for identifying a person by the information-theoretic measures.

TABLE I

Entropy And Mutual Information Computed From Two Different Datasets. X Is A Random Variable For The Identity Of A Person. Y Is The Attribute. The Conditional Entropy, Given The Attribute (E.G., Gender), Drops. It Suggests That Using Human Attributes Can Help Identify A Person.

Dataset	H (X)	H (X |Y)	I (X ; Y)
LFW	11.21	10.45	0.77
Pubfig	5.43	4.46	0.97

In order to evaluate the performance of the proposed methods, we conduct extensive experiments on two separate public datasets named LFW [7] and Pubfig [8]. These two datasets contain faces taken in unconstrained environment and are really challenging for content-based face image retrieval. Some examples of the datasets can be found in Figure 6. During the experiments, we show that the proposed methods can leverage the context

information from human attributes to achieve relative improvement up to 43.55% in mean average precision on face retrieval task compared to the existing methods using local binary pattern (LBP) [9] and sparse coding [5]. We also analyse the effectiveness of different human attributes across datasets and find informative human attributes.

The rest of the paper is organized as follows. Section II describes our observations on the face image retrieval problem and the promising utilities of human attributes. Section III introduces the proposed methods including attribute-enhanced sparse coding and attribute-embedded inverted indexing. Section IV gives the experimental results, and section V concludes this paper.

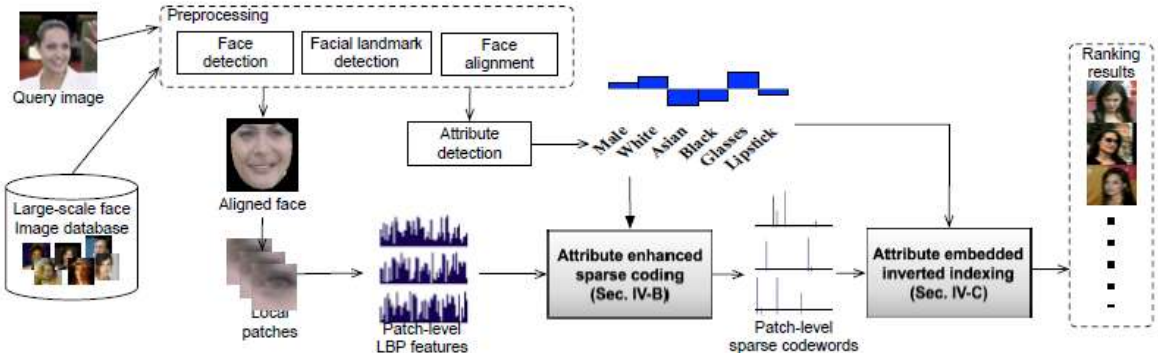
II. OBSERVATIONS

When dealing with face images, prior works [2], [4], [5] usually crop only the facial region and normalize the face into the same position and illumination to reduce intra-class variance caused by poses and lighting variations. Doing these pre-processing steps, they ignore the rich semantic cues for a designated face such as skin color, gender, hair style. To illustrate, Figure 2 shows the face image before and after the common reprocessing steps. After pre-processing steps, the information loss causes difficulty in identifying attributes (e.g., gender) of the face. In [6], the authors conducted human experiments to support similar points. When using a cropped version of face images, the face verification performance will drop comparing with using the original uncropped version. Interestingly, their experiments also show that human can achieve salient verification performance using only the surrounding context of face images. The experiments suggest that the surrounding context indeed contains important information for identifying a person. Therefore, we propose to use automatically detected human attributes to compensate the information loss.

Given a face image, let X be a random variable for the identity of a person, and Y is the attribute (e.g., gender).In information-theoretic perspective, knowing attributes can reduce the entropy for identifying a person and the information gain can be computed as,

$$I(X;Y) = H(X) - H(X|Y), \qquad (1)$$

where $H(X)$ denotes the Shannon entropy of X, which is used to measure the uncertainty of the random variable X. $H(X|Y)$ is the conditional entropy of X given Y and shows the uncertainty of X after knowing the value of Y. The larger mutual information indicates more help coming from Y for predicting X. Table I shows the entropy and mutual information computed from two different public datasets using only gender as the human attribute.



III. PROPOSED METHOD

In this section, we first describe the overview of our scalable content-based face image retrieval system, and then we explain the proposed methods: attribute-enhanced sparse coding and attribute-embedded inverted indexing in details.

A. System overview

For every image in the database, we first apply Viola-Jones face detector [10] to find the locations of faces.We then use the framework proposed in [6] to find 73 different attribute scores. Active shape model is applied to locate 68 different facial landmarks on the image. Using these facial landmarks, we apply barycentric coordinate based mapping process to align every face with the face mean shape [3]. For each detected facial component, we will extract7×5 grids, where each grid is a square patch [4]. In total we have 175 grids from five components including two eyes, nose tip, and two mouth corners. on the aligned image using similar methods proposed in [4]. From each grid, we extract an image patch and compute a 59-dimensional uniform LBP feature descriptor as our local feature. After obtaining local feature descriptors, we quantize every descriptor into codewords using attributeenhanced sparse coding described in section III-B. Attributeembedded inverted index described in section III-C is then built for efficient retrieval. When a query image arrives, it will go through the same procedure to obtain sparse codewords and human attributes, and use these codewords with binary attribute signature to retrieve images in the index system. Figure 3 illustrates the overview of our system.

B. Attribute-enhanced sparse coding (ASC)

In this section, we first describe how to use sparse coding for face image retrieval. We then describe details of the proposed attribute-enhanced sparse coding. Note that in the following sections, we apply the same procedures to all patches in a single image to find different codewords and combine all these codewords together to represent the image.

1) Sparse coding for face image retrieval (SC): Using sparse coding for face image retrieval, we solve the following optimization problem:

$$\min_{D,V} \sum_{i=1}^{n} \left\| x^{(i)} - Dv^{(i)} \right\|_2^2 + \lambda \left\| v^{(i)} \right\|_1 \qquad (2)$$

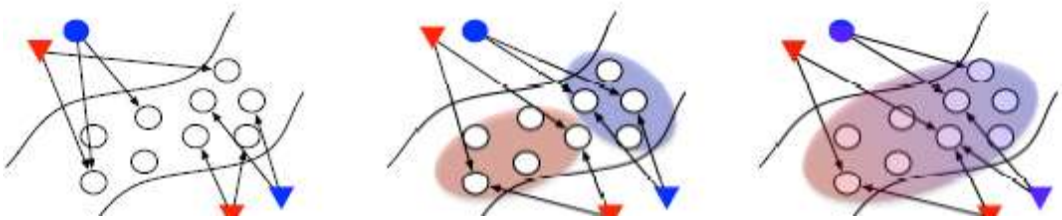
subject to $\|D_{*j}\|^2 = 1, \forall_j$

where $x^{(i)}$ is the original features extracted from a patch of face image i, $D \in R^{d \times K}$ is a to-be-learned dictionary contains K centroids with d dimensions. $V = [v^{(1)}, v^{(2)}, \ldots, v^{(n)}]$ is the sparse representation of the image patches. The constraint on each column of D $(D{*}\,j)$ is to keep D from becoming arbitrarily large. Using sparse coding, a feature is a linear combination of the column vectors of the dictionary.

Note that the Equation (2) actually contains two parts: dictionary learning (find D) and sparse feature encoding (find V). In [11], Coates et. al. found that using randomly sampled image patches as dictionary can achieve similar performance as that by using learned dictionary (< 2.7% relative improvement in their experiments) if the sampled patches provide a set of over complete basis that can represent input data. Because learning dictionary with a large

vocabulary is time-consuming (training 175 codebooks with 1600 dimension takes more than two weeks to finish), we can just use randomly sampled image patches as our dictionary and skip the time-consuming dictionary learning step by fixing D in the Equation (2) and directly solve V. When D is fixed, the problem becomes a L1 regularized least square problem, and can be efficiently solved using LARS algorithm [12]. After finding $v^{(i)}$ for each image patch, we consider nonzero entries as codewords of image i and use them for inverted indexing. Note that we apply the above process to 175 different spatial grids separately, so codewords from different grids will never match. Accordingly, we can encode the important spatial information of faces into sparse coding.

The choice of K is investigated in section V-A. We use K = 1600 in the experiments, so the final vocabulary size of the index system will be 175 × 1600 = 280, 000.



(a)SC (b)ASC-D (c)ASC-W

Fig. 4. Comparison between attribute enhancing coding methods: SC, ASC-D and ASC-W. Colors denote the attribute of the images (patches) and the shapes (e.g., triangle, circle) indicate the identity.

2) Attribute-enhanced sparse coding (ASC): In order to consider human attributes in the sparse representation, we first propose to use dictionary selection (ASC-D) to force images with different attribute values to contain different codewords. For a single human attribute, as shown in Figure 4 (b), we divide dictionary centroids into two different subsets, images with positive attribute scores (blue ones in Figure 4) will use one of the subset and images with negative attribute scores will use the other. For example, if an image has a positive male attribute score, it will use the first half of the dictionary centroids. If it has a negative male attribute score, it will use the second half of the dictionary centroids. By doing these, images with different attributes will surely have different codewords. The above goal can be achieved by solving the following optimization problem modified from Equation (2):

$$\min_{v} \sum_{i=1}^{n} \left\| x^{(i)} - Dv^{(i)} \right\|_2^2 + \lambda \left\| z^{(i)} \circ v^{(i)} \right\|_1$$

$$z_j^{(i)} = \begin{cases} \infty, if\ (1) j \geq \left\lfloor \frac{K}{2} \right\rfloor\ and\ f_a(i) \geq 0, (2) j < \left\lfloor \frac{K}{2} \right\rfloor\ and\ f_a(i) < 0 \\ 1, \qquad otherwise, \end{cases} \tag{3}$$

where "∘" denotes the pairwise multiplication between two vectors, f_a (i) is the attribute score for i^{th} image, and $z^{(i)}$ is a mask vector for deciding which codewords are allowed to be used by image i.

For example, if there are two attributes and image i contains positive scores for both attributes, $z^{(i)}$ will become $[1, \ldots, 1, \infty, \ldots, \infty, 1, \ldots, 1, \infty, \ldots, \infty]^T$. we further propose to integrate the relative scores of human attributes to relax Equation (3) into a soft weighted version (ASC-W) by defining $z^{(i)}$ in Equation (3) based

on the attribute scores of images. Motivated by [34], we first assign a half of the dictionary centroids to have +1 attribute score and use them to represent images with the positive attribute; the other half of the dictionary centroids are assigned with −1 to represent images with the negative attribute. Figure 4 (c) illustrates the above method, images with similar attributes will be assigned with similar centroids, but images with erroneous attributes might still be able to retrieve correct images if their original features are similar (cf. Figure 4 (c) bottom-right blue triangle). In details, we first define an attribute vector $a \in \{1, -1\}K$, where aj contains the attribute scores of the j_{th} centroid as follows,

$$a_j = \begin{cases} +1, if\ j \geq \left\lfloor \frac{K}{2} \right\rfloor \\ -1, \quad otherwise, \end{cases} \qquad (4)$$

then we change the and $z^{(i)}$ in Equation (3) to become, $z_j^{(i)} = \exp\left(\frac{d(f_a(i), a_j)}{\sigma}\right)$ (5)

where $d(f_a(i), a_j)$ is the distance between the attribute score of the ith image patch and that of the j_{th} dictionary centroid, and σ is used to adjust the decaying weights. To solve the above problem, we use the modified version of the LARS algorithm [12] by adjusting the weights according to $z_j^{(i)}$.

C. Attribute embedded Inverted Indexing (AEI):

The methods described in Section III-B aim to construct codewords enhanced by human attributes. In this section we describe the second method that can utilize human attributes by adjusting the inverted index structure.

1) Image ranking and inverted indexing: For each image, after computing the sparse representation using the method described in Section III-B, we can use codeword set $c^{(i)}$ to represent it by taking non-zero entries in the sparse representation as codewords. The similarity between two images are then computed as follows,

$$S(i, j) = \|c^{(i)} \cap c^{(j)}\| \qquad (6)$$

2) Attribute-embedded inverted indexing: To embed attribute information into index structure, for each image, in addition to sparse codewords $c^{(i)}$ computed from the facial appearance, we use a d_b dimension binary signature to represent its human attribute, $b^{(i)}$:

$$b_j^{(i)} = \begin{cases} 1, if\ f_a^{(i)}(j) > 0 \\ 0, \quad otherwise, \end{cases} \qquad (7)$$

The similarity score is then modified into,

$$S(i,j) = \begin{cases} \|c^{(i)} \cap c^{(j)}\|, if\ h(b^{(i)}, b^{(j)}) \leq T \\ 0, \quad otherwise, \end{cases} \qquad (8)$$

where h(i, j) denotes hamming distance between i and j, and T is a fixed threshold such that $0 \leq T \leq d_b$. As shown in Figure 5, attribute-embedded inverted index is built using the original codewords and the binary attribute signatures associated with all database images. As mentioned in [13], since XOR operation is faster than updating scores, by skipping images with high hamming distance in attribute hamming space, the overall retrieval time significantly decreases.

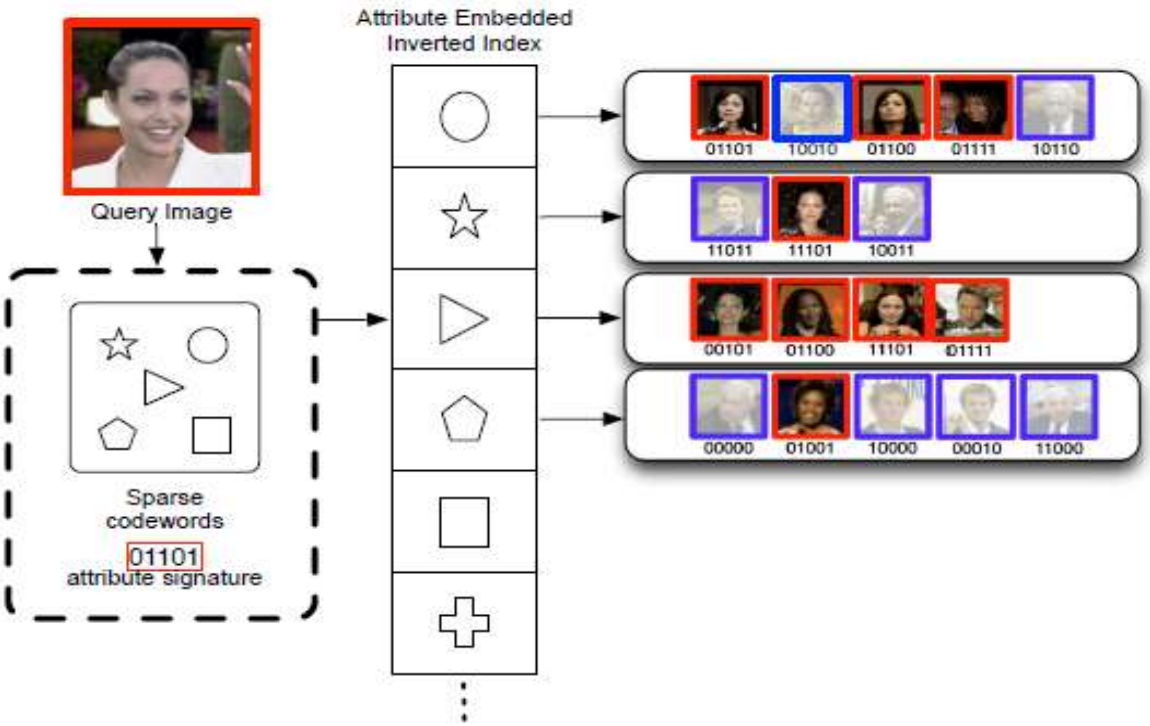


Fig. 5. Illustration of attribute-embedded inverted indexing. By considering binary attribute signature in the index system, we can skip images with large hamming distance in attribute hamming space and improve

(a) (b)

Fig. 6. Sample images from LFW (a) and Pubfig (b). Images in the same column are of the same person. These images contain large variances in expression, pose and illuminance which are very challenging for face retrieval

IV. EXPERIMENTAL RESULTS

In this work, we would like to highlight what improvements we can bring in as exploiting face attributes for semantic-rich sparse code-word representations. That is why we need to compare with LBP (one of the state-of-the-art low-level features), human attributes alone (ATTR), and the conventional sparse coding method.

Figure 7 (a) and (b) shows two query examples using SC and ASC-W respectively. The red oxes indicate the false positives, and the number below each image is its rank in the retrieval results and the number in parentheses represents the rank predicted by SC. The improvement of ASC-W compared with SC is probably contributed by the attributes such as eye wears and hair colors. Note that even though some attributes are not always consistent with the same person (cf. Figure 7(b) top 1 in SC), we can still retrieve the images using ASC-W. Figure 7 (c) and (d) show two examples using the proposed methods. Figure 7 (c) shows an example of occlusions. Using human attributes like hair colors we can gather information from not only face regions, therefore we can still achieve good performance under the occlusion. Figure 7 (d) shows a failure case, because the quality of the query image is poor, we cannot correctly predict the human attributes and sparse codewords, therefore the performance is not improved using ASC-W.

Fig. 7. Example retrieval results using the proposed methods. The red boxes indicate the false positives, and the number below each image is its rank in the retrieval results and the number in a parenthesis represents the rank by SC only.

V. CONCLUSION

We propose and combine two orthogonal methods to utilize automatically detected human attributes to significantly improve content-based face image retrieval. To the best of our knowledge, this is the first proposal of combining low-level features and automatically detected human attributes for content-based face image retrieval. Attribute-enhanced sparse coding exploits the global structure and uses several human attributes to construct semantic-aware codewords in the offline stage.

The experimental results show that using the codewords generated by the proposed coding scheme, we can reduce the quantization error and achieve salient gains in face retrieval .The proposed indexing scheme can be

easily integrated into inverted index, thus maintaining a scalable framework. During the experiments, we also discover certain informative attributes for face retrieval across different datasets and these attributes are also promising for other applications (e.g., face verification). Current methods treat all attributes as equal. We will investigate methods to dynamically decide the importance of the attributes and further exploit the contextual relationships between them.

REFERENCES

[1] Y.-H. Lei, Y.-Y. Chen, L. Iida, B.-C. Chen, H.-H. Su, and W. H. Hsu,"Photo search by face positions and facial attributes on touch devices,"ACM Multimedia, 2011.

[2] D. Wang, S. C. Hoi, Y. He, and J. Zhu, "Retrieval-based face annotation by weak label regularized local coordinate coding," ACM Multimedia,2011.

[3] U. Park and A. K. Jain, "Face matching and retrieval using soft biometrics," IEEE Transactions on Information Forensics and Security,2010.

[4] Z. Wu, Q. Ke, J. Sun, and H.-Y. Shum, "Scalable face image retrieval with identity-based quantization and multi-reference re-ranking," IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[5] B.-C. Chen, Y.-H. Kuo, Y.-Y. Chen, K.-Y. Chu, and W. Hsu, "Semi-supervised face image retrieval using sparse coding with identity constraint," ACM Multimedia, 2011.

[6] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Describable visual attributes for face verification and image search," in IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Special Issue on Real-World Face Recognition, Oct 2011.

[7] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech.Rep. 07-49, October 2007.

[8] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," International Conference on Computer Vision, 2009.

[9] T. Ahonen, A. Hadid, and M. Pietikainen, "Face recognition with local binary patterns," European Conference on Computer Vision, 2004.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," IEEE Conference on Computer Vision and Pattern Recognition, 2001.

[11] A. Coates and A. Y. Ng, "The importance of encoding versus training with sparse coding and vector quantization," ICML, 2011.

[12] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," Annals of statistics, 2004.

[13] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," European Conference on Computer Vision, 2008.