Unsupervised Face Image Retrieval using Adjacent Weighted Component-based Patches

R. Banaeeyan1, M. H. Lye1, M. F. A. Fauzi1, H. A. Karim1, and J. See2

1Faculty of Engineering, 2Faculty of Computing and Informatics
Multimedia University
Cyberjaya 61000, Selangor, Malaysia
banaeeyan.r.b@ieee.org, {haris.lye, faizal1, hezerul, johnsee}@mmu.edu.my

Abstract—Face Image Retrieval (FIR) remains a challenging problem in many real word applications due to various pose and illumination alterations of face images. State-of-the-art systems attain good precision by utilizing Bag-of-Visual-Words (BoVW) retrieval model, but their average precision (AP) decline rapidly while retrieving face images, primarily because they disregard face-specific features, and generate low discriminative visual words, mainly at the quantization level. In this paper, we employ facial patch-based features to preserve more discriminative features at patch-level in order to achieve a higher precision. We take advantage of the TF-IDF voting scheme to give more weights to more discriminative facial features. First, features are extracted from facial components instead of the whole face which preserves more informative and person-specific features. Then, an adjacent patch-based comparison is performed to preserve more discriminative features at patch-level while scoring candidate face images. Finally, a weighting approach is implemented to give even more discrimination to different scoring candidate face images. Finally, a weighting approach is applied to the scoring scheme in order to achieve a higher precision. We take advantage of the TF-IDF voting scheme to give more weights to more discriminative facial features. First, features are extracted from facial components instead of the whole face which preserves more informative and person-specific features. Then, an adjacent patch-based comparison is performed to preserve more discriminative features at patch-level while scoring candidate face images. Finally, a weighting approach is implemented to give even more discrimination to different features from different face components. Experimental results on 1,000 face images from LFW (Labeled Faces in the Wild) indicate the superiority of proposed approach by means of higher mean average precision (mAP).

Key Words: Face Retrieval; Image Retrieval; Unsupervised; Component-based; Patch-based; Weighing

I. INTRODUCTION

Retrieving similar face images using a single query image remains a challenging problem in many real word applications, particularly in police investigations such as real time criminal face recognition. The goal of such content-based Face Image Retrieval (FIR) systems is to retrieve top most similar faces from a database of not relevant faces (distracting) which are mostly under different pose variations, facial expressions, geometrical deformation, and lighting conditions.

In order to achieve a better precision, several state-of-the-art FIR systems have been developed based on an adaptation of different feature descriptors such as Scale-invariant Feature Transform (SIFT) [1], Speeded-up Robust Features (SURF) [2], and Local Binary Pattern (LBP) [3]. Although these descriptors are proven to be robust under diverse geometrical deformations, they are not face-specific and do not take into account special features of facial components (e.g. eyes, nose, mouth, etc.). Already developed local feature descriptors [4] are not discriminative enough to be employed for extracting facial features. Face images mainly encompasses smooth textures while existing feature descriptors perform well on images with high contrast and rich textures.

Because of the aforementioned reasons, a FIR system is developed in this paper using BoVW model [5] inspired by text retrieval systems in Natural Language Processing (NLP). The key feature of BoVW model is its low computational cost in terms of system resources and complexity achieved by quantizing local similar feature vectors into “visual words”. Firstly, we extract facial features from face components (two eyes, nose, and two mouth corners) which preserve more informative features in the FIR system. Secondly, we propose to employ adjacent patch-based features within face components to preserve more discriminative information under geometrical deformation in facial components, which are mainly due to diverse head poses and facial expressions. In addition, facial features are extracted from patches of overlapped components to tolerate some degrees of error at patch-level while scoring candidate face images in the dataset. Finally, since it is empirically proved that different component-based features have unequal discriminative power [6], a weighting approach is applied to the scoring scheme in order to develop FIR system to take advantage of this fact.

This paper is organized as follows. The next section briefly discusses related works in face image retrieval. Section III presents the methodology of our proposed system while section IV discusses the experimental results. Finally, section V concludes the paper.

II. RELATED WORKS AND CONTRIBUTIONS

The proposed FIR method in this paper is inspired by the facial component-based method reported in [8]. In recent researches on face retrieval and recognition ([8]–[10]), more discriminative information are preserved by comparing features of query face with dataset faces at facial component level (left eye, right eye, nose, left mouth corner, and right mouth corner). To achieve more person-specific facial features, component-based patches, which define a grid on each face component, is applied. This enables the retrieval system to compare query features with their corresponding feature among face images in the dataset, which are within the same component and the same patch (Fig.1).
Due to geometrical deformation in different face images of an individual, the defined facial patches are highly likely to be shifted. The already proposed component-based methods do not take into account relocation of informative patches within facial components, which is mainly due to variation in head poses and facial expressions. Our proposed method aims to preserve the informative patches at component level by not only comparing corresponding patches of two face images, but also their adjacent patches as illustrated in Fig. 2.

III. METHODOLOGY

An overall pipeline of our FIR approach is presented in Fig. 3. Given a face image as query, the system begins by extracting the low-level features at facial component level as shown in Fig. 4, followed by quantizing the feature vectors into clusters for indexing and constructing visual words. Then, component-based and patch-based features are given different weights. Adjacent patches of each unique patch are also taken into account for scoring candidate images in the database but with lower weights. Finally, after applying the mentioned models, images in the dataset are scored and most similar faces are retrieved and ranked.

A. Image Voting Scheme

Given a face image as query \( y \) and database images \( x_{i,j}, 1 \leq j \leq N \), their local descriptors are represented by \( y_{i} \) and \( x_{i,j} \) respectively, where \( N \) is number of database images. The voting scheme [7] is summarized as follows:

1. Scores of all images in dataset \( s_{j} \) are set to 0.
2. For each descriptor in query image \( y_{i} \) and for each descriptor in database \( x_{i,j} \), the score of corresponding image \( s_{j} \) is increased by
   \[
   s_{j} = s_{j} + f(x_{i,j}, y_{i})
   \]
   where \( f \) is the function indicating similarity between \( x_{i,j}, y_{i} \). Since in this research we use SIFT features [1], similarity function \( f \) is based on Euclidean distance symbolized as \( d(., .) \), and it is computed by
   \[
   f(x_{i,j}, y_{i}) = \begin{cases} 
   1 & \text{if } d(x_{i,j}, y_{i}) < \varepsilon \\
   0 & \text{otherwise}
   \end{cases}
   \]
3. The final image score \( s_{j}^{\prime} \) reserved for voting is achieved based on \( s_{j} \) by
   \[
   s_{j}^{\prime} = \sum_{i=1}^{m} \sum_{j=1}^{m_{i}} f(x_{i,j}, y_{i})
   \]
   The final image score \( s_{j}^{\prime} \) indicates the number of multiple matches between query image descriptors and database image descriptor.

B. Quantization & Bag of Visual Word (BoVW) Construction

BoVW model [5] applies quantization on all database descriptors; a quantization function \( q \) maps descriptors \( x \) from a \( d \)-dimensional space \( \mathbb{R}^{d} \) to an integer index \( q(x) \) between 1 to \( k \).
   \[
   q: \mathbb{R}^{d} \rightarrow [1, k] \\
   x \mapsto q(x)
   \]

The quantization is typically performed by \( k \) – means clustering on database descriptors, which results in \( k \) centroids or visual words. \( q(x) \) is the index of visual word
closest to descriptor \( x \). Therefore, two descriptors \( x \) and \( y \) that belongs to same visual word satisfy \( q(x) = q(y) \).

To obtain final image scores of the database, descriptors are efficiently compared based on their corresponding quantized indexes. This is proposed in [7] as:

\[
s_j^* = \sum_{i=1}^{k} \frac{m_i m_{ij}}{m_j}
\]

(5)

Where \( m_i \) and \( m_{ij} \) represents the number of descriptors in query image and database image \( j \), which are in the same cluster \( l \).

C. Term Frequency – Inverse Document Frequency (TF-IDF)

In order to give weighting to the computed final image scores \( s_j^* \), we apply the TF-IDF weighting proposed in [11]. TF-IDF is known as ‘term frequency-inverse document frequency’ and is computed as in equation (6):

\[
T_{jk} = \frac{n_{jk}}{n_j} \log \frac{N}{n_k}
\]

(6)

where \( n_{jk} \) and \( n_k \) are frequency of visual word \( k \) in image \( j \), and total number of visual word frequencies in image \( j \), respectively. \( N \) is the total number of images in database, and \( n_k \) is total frequencies of visual word \( k \) in database.

Suppose the BoVW includes \( k \) visual words, then each image is indicated by a \( k \)-dimensional vector of weighted visual word frequencies, where \( T_{jk} \) represents TF-IDF of image \( j \) of database and its corresponding visual word frequency \( k \).

Using \( T_{jk} \) weighting while voting database image scores \( s_j^* \), proportionally increases the number of times the visual word \( k \) appears in the corresponding image \( j \), but is offset by the visual word frequency in the database, helping to adjust for the fact that some words appear more frequently in database.

D. Component-based Weighting

Our proposed FIR system starts by detecting candidate’s faces and detecting five face components (two eyes, nose, and two mouth corners), after alignment of detected faces (Fig.4). We use the technique in [12] for face detection and a tree-structured model [13] for facial component detection. Then, each extracted face component is divided to a grid size of \( 5 \times 7 \) (35 patches), so the total defined facial patches are 175 \((35 \times 5)\), following the work performed in [8]. The component-based weighting scheme is inspired by the way Human Visual System (HVS) works while recognizing different faces. As stated in [6], human brain extracts important information for face recognition principally from eyes than other face components. Therefore, we treat extracted component-based features differently in order to improve the FIR results. To this end, in our research we give a score of 4.00 to extracted features from both left and right eyes while others (mouth and two mouth corners) are scored 1.00. These values are empirically proved to perform best in our experiment. Therefore, we use variable \( W \) which indicates the given weight value. This approach is applied to scoring eyes only, therefore, we compute final image scores by equation (7) while \( W \) is set to 4.00.

\[
s_j = s_j^* + (T_{jk} \times W)
\]

(7)

E. Adjacent Patch-based Weighting & Comparison

As explained earlier, due to geometrical deformation in different face images of an individual, the defined facial patches are highly likely to be shifted. To minimize the effect of this possible error, unlike former works [8], [10], we not only search for similar visual words in corresponding patches (Fig.1), but also local adjacent patches (green boxes in Fig.2). Furthermore, we weight these adjacent patches in a way to give more votes to central corresponding patch (red box in Fig.2). Therefore, adjacent patches are empirically proved in our experiments to be weighted 0.4 to achieve best results in our approach. This tolerates some degree of face poses alterations in scoring candidate faces while retrieving images in our FIR system. This approach is applied to scoring adjacent patches only, therefore, we compute final image scores by equation (7) while \( W \) is set to 0.40.

IV. EXPERIMENTS

Our experiments follow the same procedures as in [8], [10].

A. Database

The LFW face database (Labeled Faces in the Wild) [13], a database of face photographs designed for studying the problem of unconstrained face recognition, is used in our experiments. The database includes over 13,000 images of faces collected from the WWW. From LFW database, 1000 face images representing more than 300 individuals is used as
our basic dataset (distractors). From the same database, 100 face images representing 10 unique individuals, is added to the basic database (distractors) to construct our ground-truth database (totally 1,100 face images). The number of face images per person is 10. All faces, whether in query sets or ground-truth are resized to 256×256.

B. Evaluation

For evaluation purpose, we select 10 representative face images of unique individuals from LFW database, to be served as our query faces, in order to retrieve corresponding face images from our ground-truth database. A sample of relevant images in the ground-truth database is shown in Fig. 5.

After faces are extracted using Viola-Jones face detector [14], five face components (left eye, right eye, nose, left mouth corner, and right mouth corner) are localized and extracted using the algorithm described in [15]. Then, each extracted face component is divided to a grid size of 5×7 (35 patches), so the total defined facial patches are 175, following the work by [8].

Figure 5: Sample of challenging relevant face images used in our ground truth database.

In order to quantize the extracted dense-SIFT feature vectors and subsequently constructing the visual words, an unsupervised clustering technique, k-means [16], is used. The number of cluster centers is set to 1,700 as this is experimentally proved to give the best retrieval performance.

C. Software and Hardware

For both dense-SIFT extraction and k-means clustering, VLFeat Toolbox (0.9.20) [17] is employed, and other programming parts are done using Matlab R2015a. Our experiments are performed by a single 3.40 GHz CPU on a desktop computer with 4 GB installed memory.

D. Performance Metrics

We have used three metrics to assess the performance of the proposed FIR system: mAP (mean average precision), and precision at different recalls of 10% and 20%, respectively.

mAP is defined as the mean of all average precisions eventuated from each unique face query image, and is calculated by:

\[ mAP = \frac{1}{N} \sum_{i=1}^{N} \int_{0}^{1} p_i(r)dr \]  

where \( N \) represents the number of face queries, and \( p_i(r) \) represents the precision at recall \( r \) for face query \( i \).

E. Retrieval Results

The performance of the proposed FIR system, according to different measurement metrics, is shown in Table 1. It is quite evident that when only BoVW model is used along with dense SIFT sampling, the FIR performs worst indicating features extracted from the whole face image are not discriminative enough to retrieve similar faces.

Table 1: Comparison of proposed approaches with baseline method (DSIFT+BoVW). DSIFT: Dense SIFT extraction, BoVW: Bag of Visual Words model, CB: Component-based feature extraction, APW: Adjacent patch-based weighting and comparison, CW: Component-based weighting.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP (%)</th>
<th>P@10 (%)</th>
<th>P@20 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSIFT+BoVW</td>
<td>14.59</td>
<td>49.41</td>
<td>32.34</td>
</tr>
<tr>
<td>DSIFT+BoVW+CB</td>
<td>24.16</td>
<td>69.31</td>
<td>53.03</td>
</tr>
<tr>
<td>DSIFT+BoVW+CB+APW</td>
<td>24.70</td>
<td>69.79</td>
<td>54.02</td>
</tr>
<tr>
<td>DSIFT+BoVW+CB+CW</td>
<td>25.24</td>
<td>67.40</td>
<td>55.56</td>
</tr>
</tbody>
</table>

The performance by employing component-based features is much better, and after that, adjacent patch-based weighting comparison and component-based weighting are the bests for face image retrieval with only 0.54% difference in mAP (Table 1). Regarding other performance metrics, precision at 10% recall is the best after applying adjacent patch-based weighting and comparison (APW), while precision at 20% recall belongs to component-based weighting (CB+CW). Sample retrieval results after applying all approaches (DSIFT+BoVW+CB+CW) are illustrated in Fig. 6.

V. CONCLUSION AND FUTURE WORKS

In this research, we proposed a model for retrieving face images, based on BoVW model, by employing component-based facial features which provide more discriminative person-specific features. We applied the TF-IDF voting scheme to take into account the contribution of each face descriptor. We also presented a model to weight adjacent patch-based features within each face component and compare them with the central patch in order to tolerate some degrees of error in patch localization due to geometrical deformation. We also applied the weighting to facial features at the component level to further improve the retrieval performance. Experimental results indicate an increase of 10.65% in mAP compared with baseline method (dense SIFT sampling + BoVW). By applying component-based feature extraction (CB) along with component-based weighting (CW), we achieved the best mAP. In order to further improve the precision of the presented FIR system as well as achieving scalability for larger face databases, we are currently applying the re-ranking approach to the top retrieved faces and using more discriminative facial features while constructing visual
words, as well as taking into account other low/high level face-specific features.

ACKNOWLEDGMENT

This work is supported by Ministry of Science, Technology, and Innovation (MOSTI), Malaysia.

Figure 6: Example top 10 retrieved face images after applying the proposed approach (DSIFT+BoVW+CB+CW). Query images are in the left column and top-ranked are in the right. True positives are indicated by green boxes.

REFERENCES


