IMPROVING LARGE-SCALE FACE IMAGE RETRIEVAL USING MULTI-LEVEL FEATURES

Xiaojing Chen, Le An, Bir Bhanu

Center for Research in Intelligent Systems, University of California, Riverside, California, USA
xchen010@ucr.edu, lan004@ucr.edu, bhanu@cris.ucr.edu

ABSTRACT

In recent years, extensive efforts have been made for face recognition and retrieval systems. However, there remain several challenging tasks for face image retrieval in unconstrained databases where the face images were captured with varying poses, lighting conditions, *etc*. In addition, the databases are often large-scale, which demand efficient retrieval algorithms that have the merit of scalability. To improve the retrieval accuracy of the face images with different poses and imaging characteristics, we introduce a novel feature extraction method to bag-of-words (BoW) based face image retrieval system. It employs various scales of features simultaneously to encode different texture information and emphasizes image patches that are more discriminative as parts of the face. Moreover, the overlapping image patches at different scales compensate for the pose variation and face misalignment. Experiments conducted on a large-scale public face database demonstrate the superior performance of the proposed approach compared to the *state-of-the-art* method.

Index Terms— Face retrieval, large-scale, multi-level features

1. INTRODUCTION

An explosion of visual media such as images and videos on the Internet has been observed during the past decade. Among all the randomly downloaded images, a significant fraction of them (45%) contain faces [1]. With such gigantic quantities of faces, large-scale face retrieval in images or videos is gaining increased attention [2, 3]. The goal for an image-based retrieval system, given a face image as a query, is to return face images containing the same person as in the query image from a large-scale image database possibly containing millions of face images and most of which are obtained with various poses, lighting conditions, and/or image resolutions. Traditional face retrieval systems would hardly achieve scalability in a large-scale image database. The main reason is that, most of these systems essentially obtain the retrieval results by comparing the query image with each and every face image in the database. Such a scan of the entire database is quite time consuming and would directly cause the complexity of the system grow linearly with the size of the database. Although there are many potential applications for a large-scale face image retrieval system, very limited work has been done in this area [2].

In order to achieve scalability, the following techniques are adopted in the *state-of-the-art* image retrieval systems: the forest of *k-d* tree, hashing based method, and bag-of-words (BoW) based method. It is known that *k-d* tree is not good for indexing high dimensional feature vectors, hashing based method only performs well in a dense feature space using Euclidean distance as metric. On the other hand, many face descriptors are high dimensional and sparse feature vectors, and popular face descriptors such as Local Binary Pattern (LBP) [4] measure difference between feature vectors using Chi square distance. Based on the aforementioned reasons, we build the face image retrieval system using BoW based method as it is commonly done in text retrieval systems. A key component of such retrieval systems is to quantize local feature vectors into "visual words" according to a trained visual vocabulary. In this paper, we propose to use multi-level features to generate the visual words. The multi-level features take advantage of various scales of features to encode different texture information and address the image patches that contain more person-specific information. In addition, the multi-level features in this paper are extracted from overlapping image patches to mitigate the pose variation or misalignment among different face images.

2. RELATED WORK AND CONTRIBUTIONS

Although the BoW framework for image retrieval is shown to be efficient and simple, there are still several problems that prohibit us from obtaining better accuracy and efficiency. The problems are two-folds: *first*, the discriminative power of local feature descriptors is inevitably degraded due to both vector quantization and the large size of the databases; *second*, although it has been observed that the spatial information of local features is quite important for improving retrieval accuracy , it is ignored or only partially used. To overcome these limitations, many research attempts have been made. The approximate nearest neighbor method [5] and vocabulary tree [6] speedup the assignment of individual feature descriptor to visual words, thus make it possible to build a large vocabulary. The problem of information loss during visual word

quantization is addressed in soft assignment [7] and Hamming embedding [8]. Spatial verification [5] and query expansion [9] are introduced for re-ranking the initial return to increase the recall of the retrieval result.

However, the performance of most of these techniques degrades when applied on face images directly. On one hand, face images are basically smooth in texture. Therefore, the visual vocabulary constructed based on regions only with large contrast or rich texture is not able to sufficiently describe the detail of a human face. On the other hand, the specific spatial layout of the facial components is not utilized in most retrieval systems. In [2], a face image retrieval system using identity-based quantization and multi-reference reranking is proposed to address both issues.

The proposed multi-level feature in this paper is inspired by the multi-scale methods [10] [11]. However, our method differs from the previous work in that in addition to the information at different spatial levels being explored, we extract the local features on overlapping image patches to specifically address the face pose variation or misalignment. This allows the information for a specific part of the face to be encoded by at least one of the image patches in a local neighborhood. As suggested in the experiments, the proposed multi-level features provide improvements over the *state-of-the-art* for large-scale face retrieval. To the authors' best knowledge, no previous work has been done using multi-level features for large-scale face retrieval in unconstrained face databases.

3. TECHNICAL DETAILS

3.1. Overview of Our Approach

The flowchart of the proposed approach is shown in Fig. 1, the framework is similar to that in [2]. Given an aligned face image, the face is cropped, then normalization is carried out to remove illumination changes. The normalization technique in [12] is adopted in this paper. After normalization, five facial components (two mouth corners, nose tip, and two eyes) are located. A 5×4 grid is defined on each facial component, thus every face image provides 100 cells in total, and each cell is a 14×14 image patch. The widely used face descriptor LBP [4] is employed to extract feature from each cell. Higher level features can be generated by features extracted from single cells, in the following section, the multi-level feature generation is presented in detail.

In order to encode spatial information, a unique position ID is assigned to each feature according to the region from which that feature is extracted. Two features are matched if and only if they have the same quantization result and position ID. Because intra-class variations could be larger than inter-class variations in face image dataset where the same person's face undergoes pose, illumination and expression changes, an identity-based quantization [2] using supervised learning is adopted to avoid quantization errors. P different people and

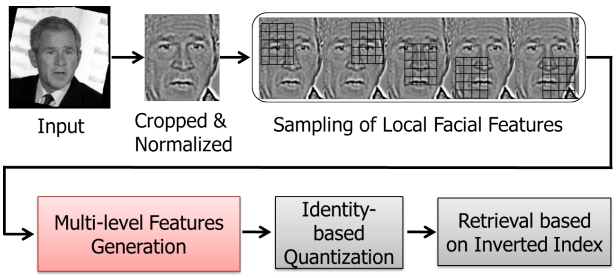


Fig. 1. Overview of our approach.

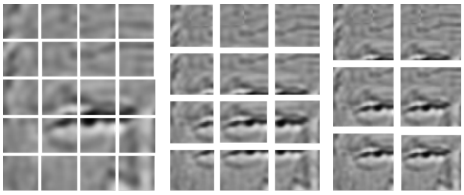
each with T various face examples are used as the training set to generate the visual vocabulary. A visual word in the vocabulary contains features extracted from regions with the same person ID and position ID. With the visual vocabulary, the quantization of a local feature is simply the process of finding its nearest-neighbor from all the training features with the same position ID. When only features extracted from single cells are used, each face image is described by 100 visual words after quantization. To improve recall of the retrieval result, soft quantization [7] is adopted - a local feature is quantized into R visual words which correspond to the R different identities after searching the top k nearest neighbor of the original feature.

The inverted index structure used in text retrieval is employed here for the purpose of reducing storage cost and maintaining retrieval efficiency. Therefore, instead of storing feature descriptors (visual word) of each image independently, we can use inverted index to store the mapping from each visual word in the vocabulary to the occurrences of that visual word in all images. During the retrieval, all potential candidates (images that contain common visual words with the query image) are returned after searching the inverted index. These candidates are ranked according to the similarity (number of common words) between the candidate image and the query image.

3.2. Multi-level Features and Visual Words

In [2], only local feature (or visual word) extracted from the image patch covered by a single cell is used. However, such a small image patch merely contain any information that can be used to identify the face. Therefore, instead of using a single cell to define the image patch of interest, we use multiple cells that are close to each other to define a higher level image patch. As can been seen in Fig. 2 (a), due to the small size of each image patch, the person specific appearance is not well encoded. As the patch size goes larger in Fig. 2 (b), the salient parts of the face (*e.g.*, eye, eyebrow) are better sampled such that each image patch contains more person specific information. In Fig. 2 (c), the further enlarged image patch does not only incorporate the person specific clue, but also the co-occurrence pattern of the person specific facial components

(*e.g.*, eye and nose together), thus providing more discriminative power.



(a) 1st level (b) 2nd level (c) 3rd level

Fig. 2. Image patches extracted from different levels.

Specifically, the feature extracted from a single cell is referred as the $1st$ level feature, and its corresponding quantization result is the $1st$ level visual word. The $2nd$ level features are extracted from the image patches defined by four neighboring cells, and each image patch is extracted in a sequential manner, as shown in Fig. 3. With this method, cells at the center where the texture is rich will be sampled more than once. Therefore, the regions that encode more characteristic information of the face can be emphasized. In addition, the overlapping of image patches help to compensate for the face component misalignment (*e.g.*, a mouth in an image patch at a certain location will correspond to the mouth in the query image at a shifted location). Similarly, the $3rd$ level features are extracted from the image patches, each of which covers 9 neighboring cells. Each face image yields 60 $2nd$ level features and 30 $3rd$ level features. In this way, the higher level image patches contain more discriminative information and important regions of a face are sampled multiple times.

It is worth noting that the size of feature grows as its level increases, which raises the computational cost for quantization. Therefore, in order to balance between discriminative power and computational cost, we only add $2nd$ and $3rd$ level features and use an efficient approximate nearest-neighbor method (Locality-sensitive Hashing [13]) for quantization.

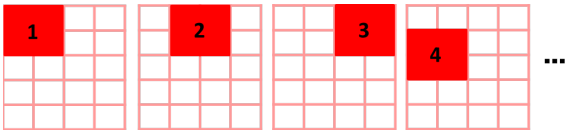


Fig. 3. Define image patches to extract $2nd$ level features.

3.3. Computational and Storage Cost Analysis

In order to demonstrate that with multi-level features (visual words) the proposed face image retrieval approach still achieves scalability, we analyze the extra computational and storage cost compared to the work in [2]. Fixed cost that is not affected by the size of the dataset is ignored.

Using multi-level features brings two additional computational costs: cost for higher level feature generation and cost for higher level feature quantization. For images in the testing set, generation and quantization of higher level features have no influence on the scalability, as they are independent of the query process and are carried out offline for only once.

Although higher level features cost more storage than lower level features, we save the visual words correspond to the features rather than the features themselves. Each visual word takes about 1 byte on average. Therefore, in our implementation, the total storage cost for each image is only about 190 bytes, which enables the proposed approach to maintain scalability.

4. EXPERIMENTAL RESULTS

4.1. Datasets and Evaluation Metric

The LFW (Labeled Faces in the Wild) database [14] is chosen as the basic testing set. It contains more than 13,000 face images from Internet and about 6000 people are included in the database. The training set is collected from Internet, which includes face images from 192 people and each with 30 face examples with various expressions, illumination conditions and poses. All the images both in the testing set and the training set are resized to 250×250 with color information removed. Moreover, each image is labeled with the name of the person that appears on the image, and aligned according to the position of the eyes.

In order to test the performance of the retrieval system, 200 sample face images from the testing set are selected and used as query images. Three testing sets with smaller sizes (4k, 7k and 10k) are generated by sampling images from the basic testing set (13k). Mean average precision (mAP) is adopted as the performance metric:

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

where N is the number of queries, and $p_i(r)$ is the precision at recall r for query i.

4.2. Retrieval Results

The performance of the retrieval system when different combinations of visual words are used is shown in Fig. 4. Similar performance trend is observed on testing sets with different sizes. When only the $1st$ level visual words are used, the system performs worst, which indicates that a small image patch dose not contain much useful information to describe a face. The performance by using the $2nd$ level visual words is much better, and the $3rd$ level visual words is the best for face identification using single-level features. By using different combinations of multi-level visual words, the best result is obtained when the $2nd$ and the $3rd$ level visual words are used. This demonstrates that using multi-level features instead of using single-level features improves the retrieval accuracy.

Fig. 5. Sample retrieval results. Query images are in the left column. The top-ranked images are shown on the right. The first two rows are queries with high mAP and the last two rows are queries with low mAP. False positives are shown in red boxes.

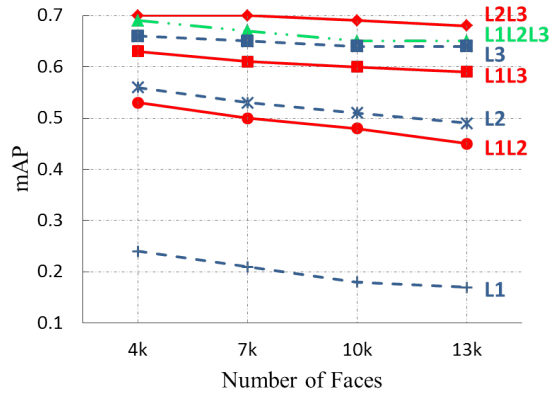


Fig. 4. Comparison of different combinations of visual words on testing sets with various sizes.

When our method is evaluated on the basic testing set (more than 13,000 images), using visual words combinations L2L3 and L1L3L3 give the best two retrieval mAP (0.68 and 0.65 respectively). The reranking method in [2] is employed here as a post processing step to further improve the retrieval performance. The reranking process uses the global features to refine the ranked list, where the retrieved images that are close to both the query image and a reference set would have higher ranks. Comparison of the proposed method before and after employing the reranking is shown in Table 1. The results demonstrate that although the reranking is more effective for face retrieval using single-level features (L1, L2, L3) than multi-level features (L2L3, L1L2L3), the reranking further improves our best results using multi-level features. Note that by using multi-level features (visual words) only the mAP is improved significantly (by 51%) compared to using $1st$ level features. On the other hand, the mAP of the method in [2] using complicated feature descriptors (T3hS2) on a similar dataset with fewer images (10,000 images) is 0.45, and after reranking the mAP increased to 0.7, which is comparable to our best result (0.71). This comparision indicates the superiority of using multi-level features (visual words), which achieve the *state-of-the-art* results with very simple LBP face descriptor. It is expected that by using more advanced feature descriptors, the performance of the proposed approach can be further improved. Sample retrieval results are shown in Fig. 5.

Table 1. Comparison of the proposed method before and after reranking on the basic testing set (13k).

	L1	L2	L3	L2L3	L1L2L3
mAP Before	0.17	0.49	0.64	0.68	0.65
mAP After	0.35	0.55	0.68	0.71	0.68

5. CONCLUSIONS

In this paper, we present a multi-level feature extraction scheme for large-scale face image retrieval system based on BoW framework. The multi-level features well describe the face image patches at different scales to robustly match the faces taken under different conditions. The proposed method is simple yet effective and improves over the *state-of-the-art* large-scale face image retrieval system. The multi-level feature extraction can be adopted in different face image retrieval frameworks with ease.

6. ACKNOWLEDGEMENTS

This work was supported in part by NSF grants 0641076 and 0905671. The contents and information do not reflect the position or policy of the U.S. Government.

7. REFERENCES

[1] N. Kumar, P. Belhumeur, and S. Nayar, "Facetracer: A search engine for large collections of images with faces," in *ECCV*, 2008.

[2] Z. Wu, Q. Ke, J. Sun, and H.-Y. Shum, "Scalable face image retrieval with identity-based quantization and multireference reranking," *IEEE T-PAMI*, 2011.

[3] H. T. Vu, T. D. Ngo, T. N. Nguyen, D-D. Le, S. Satoh, B. H. Le, and D. A. Duong, "Fast face sequence matching in large-scale video databases," in *ICIP*, 2011.

[4] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE T-PAMI*, 2006.

[5] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007.

[6] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *CVPR*, 2006.

[7] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *CVPR*, 2008.

[8] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV*, 2008.

[9] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *ICCV*, 2007.

[10] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

[11] C. Chan, J. Kittler, N. Poh, T. Ahonen, and M. Pietikainen, "(multiscale) local phase quantisation histogram discriminant analysis with score normalisation for robust face recognition," in *ICCV Workshop*, 2009.

[12] Xiaoyang Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE T-IP*, 2010.

[13] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," *Commun. ACM*, vol. 51, no. 1, pp. 117–122, Jan. 2008.

[14] 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," Tech. Rep., University of Massachusetts, Amherst, 2007.