Scalable Face Image Retrieval with Identity-Based Quantization and Multi-Reference Re-ranking

Zhong Wu *
Tsinghua Univ., Ctr Adv Study

Qifa Ke[†1], Jian Sun[†2]
Microsoft Research
[1]Silicon Valley Lab, [2]Asia Lab

Heung-Yeung Shum[†]
Microsoft Corporation

Abstract

State-of-the-art image retrieval systems achieve scalability by using bag-of-words representation and textual retrieval methods, but their performance degrades quickly in the face image domain, mainly because they 1) produce visual words with low discriminative power for face images, and 2) ignore the special properties of the faces. The leading features for face recognition can achieve good retrieval performance, but these features are not suitable for inverted indexing as they are high-dimensional and global, thus not scalable in either computational or storage cost.

In this paper we aim to build a scalable face image retrieval system. For this purpose, we develop a new scalable face representation using both local and global features. In the indexing stage, we exploit special properties of faces to design new component-based local features, which are subsequently quantized into visual words using a novel identity-based quantization scheme. We also use a very small hamming signature (40 bytes) to encode the discriminative global feature for each face. In the retrieval stage, candidate images are firstly retrieved from the inverted index of visual words. We then use a new multi-reference distance to re-rank the candidate images using the hamming signature. On a one-millon face database, we show that our local features and global hamming signatures are complementary—the inverted index based on local features provides candidate images with good recall, while the multi-reference re-ranking with global hamming signature leads to good precision. As a result, our system is not only scalable but also outperforms the linear scan retrieval system using the state-of-the-art face recognition feature in term of the quality.

1. Introduction

Given a face image as a query, our goal is to retrieve images containing faces of the same person appeared in the query image, from a web-scale image database containing tens of millions face images. Figure 1 shows some example online celebrity face images with various poses, expressions, and illumination. Such a face retrieval system has many applications, including name-based face image search, face tagging in images and videos, copyright enforcement, etc. To the best of our knowledge, little work aims at web-scale face image retrieval.

Figure 1. Example online celebrity face images with variances in pose, expression, and illumination.

A straight-forward approach is to use bag-of-visual-words representation that has been used in state-of-the-art scalable image retrieval systems [5, 17, 19, 22]. However, the performance of such a system degrades significantly when applying on face images. There are two major reasons. On one hand, the visual word vocabulary, learned from local SIFT-like features detected from the face images, has difficulty to achieve both high discriminative power (to differentiate different persons) and invariance (to tolerate the variations of the same person). Secondly, existing systems ignore strong, face-specific geometric constraints among different visual words in a face image.

Recent works on face recognition have proposed various discriminative facial features [27, 8, 18, 10, 9]. However, these features are typically high-dimensional and global, thus not suitable for quantization and inverted indexing. In other words, using such global features in a retrieval system requires essentially a linear scan of the whole database in order to process a query, which is prohibitive for a web-scale image database.

In this paper, we propose a novel face image representation using both local and global features. First, we locate component-based local features that not only encode geometric constraints, but are also more robust to pose and

*This work was done while Zhong Wu was an intern at Microsoft Research. Email: wuz02@mails.tsinghua.edu.cn

†Email: {qke, jiansun, hshum}@microsoft.com

expression variations. Second, we present a novel identity-based quantization scheme to quantize local features into discriminative visual words, allowing us to index face images, a critical step to achieve scalability. Our identify-based quantization can better handle intra-class variation using multiple examples from multiple identities. Finally, in addition to the local features, we compute a 40-byte hamming signature for each face image to compactly represent a high-dimensional discriminative global (face recognition) feature.

Our face retrieval system takes advantages of the fact that local features and global features are complementary. Local features allow us to efficiently traverse the index of a large scale face image database, and return top candidate images (e.g., 1,000 candidates). While the precision may be low, we can achieve a good recall in this index traversing stage. Then, the hamming signatures (derived from global features), which is as small as 40KB for 1,000 images, are used to re-rank the candidate images. By using a new multi-reference distance metric, the precision can be significantly improved. Overall our face retrieval system is not only scalable, but also outperforms state-of-the-art retrieval or recognition systems in terms of both precision and recall, which is demonstrated by experiments with a database containing more than one million face images.

1.1. Related Work

State-of-the-art large scale image retrieval systems have relied on bag-of-words image representation and textual indexing and retrieval schemes for scalability. In these systems, feature detectors first detect distinctive and repeatable points or regions such as DoG [12], MSER [14] in the image, from which discriminative feature descriptors [12, 15, 25] are then computed. These descriptors are subsequently quantized into visual words with a visual vocabulary [22], which is trained by the unsupervised clustering algorithms [17, 19]. To further improve the scalability, Jegou aggregates partial information of the standard bag-of-features vector to build the "miniBOF" vector [6], which is more compact in the index. On the other hand, to improve the precision, some compact information can be embedded [5] for each visual word in the index, which compensates the information loss in the quantization. However, the performance of these traditional image retrieval systems degrades significantly when applied to face images.

In recent years, many effective features have been proposed for face recognition. For example, LBP [18] feature, variations of LBP [23, 26, 28], and V1-like feature [21] are designed to capture the micro-patterns of the face. Besides these "low level" features mentioned above, Kumar *et al* [8] incorporates the traits information with the attribute and simile classifiers. Efforts are also made to tackle the face alignment and matching problem in face recognition. In [27], Wright proposes a Rp-tree based approach to implicitly encode the geometric information into the feature. It is non-trivial to make these global feature based methods scalable. One might consider using k-d tree [2] or LSH [7, 3] to avoid scanning every image. But we have found these approximated nearest neighbor search methods do not scale or work well with high dimensional global face features.

2. Local Features for Indexing

In this section, we describe the details of the local features, and a novel identity-based quantization for inverted indexing.

2.1. Component-Based Local Features

Figure 2 shows our local feature extraction and indexing pipeline. First, five facial components (two eyes, nose tip, and two mouth corners) are located on a detected face [24] by a neural-network based component detector [11]. The face is then geometrically normalized by a similarity transform that maps the positions of two eyes to canonical positions.

We define a 5×7 grid at each detected component. In total we have 175 grids from five components. From each grid we extract a square image patch. A T3hS2 descriptor (responses of steerable filters) [25] is then computed for each patch. All descriptors are quantized into visual words that are subsequently inverted indexed. Notice that the existing interest-point based local feature detectors [16] are not suitable for the face image. Such detectors tend to detect features in regions with rich textures or high contrast. They do not perform as well on face images since they contain mostly smooth textures.

Compared to defining grids over the whole face [27, 10], our features are localized to the components, which allows flexible deformation among the components and are more robust to face pose and expression variations. Also note that girds from different components have some overlaps, this, together with the histogram-based T3hS2 descriptors, allows our system to tolerate some degree of errors in the component localization.

To enforce geometric constraints among features, we assign each grid a unique ID, which is called "position id". The position id will be concatenated with the feature quantization id (described next) to form the a "visual word". By doing so, each visual word carries strong geometric information - two features can be matched only if they come from the same component and are extracted from the same grid in that component. This is in contrast to existing models that allow features to match even they are coming from different grids in the face, which performs worse in our task.

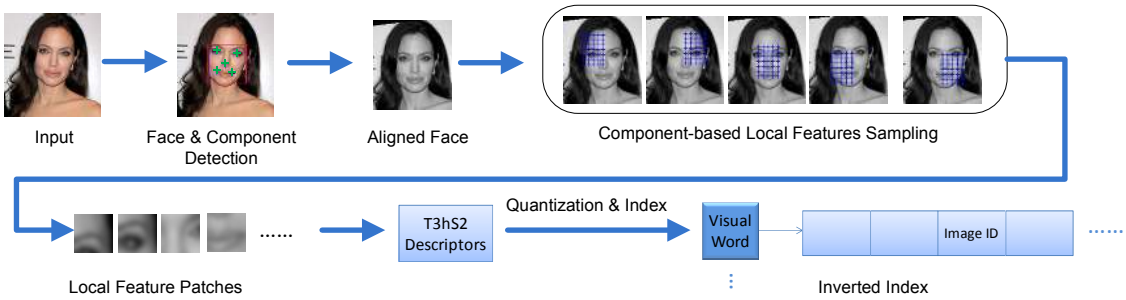

Figure 2. Local feature extraction and indexing.

2.2. Identity-based Quantization

For scalability, the extracted local features need to be quantized into a set of discrete visual words using a visual vocabulary which is often obtained by an unsupervised clustering algorithm (e.g., k-means) [17]. But the unsupervised learning is not very good for training a vocabulary for face images, where intra-class variations are often larger than inter-class variations when the face undergoes pose and expression changes. Quantization errors will degrade the retrieval performance.

In this section, we propose an identity-based quantization scheme using supervised learning. Our training data consists of P different people and each person has T face examples, at various poses, expressions, and illumination conditions. Figure 3 shows example face images of one person and constructed visual words. Since each person has a unique "person id" and each grid has a unique "position id", we define a visual word as the pair <person id, position id> and associate it with T local feature descriptors computed from the training samples of the "person id". In other words, each visual word is an example-based representation - containing multiple examples. That is the strength of our identity-based quantization - the features under various pose/lighting/expression conditions have a chance to be quantized to the same visual word.

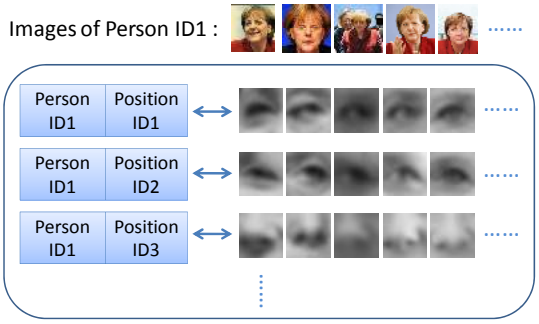

Figure 3. Identity-based vocabulary from one person. A visual word is formed by concatenating two IDs: <person id, position id>. The final vocabulary is the combination of all visual words from all persons.

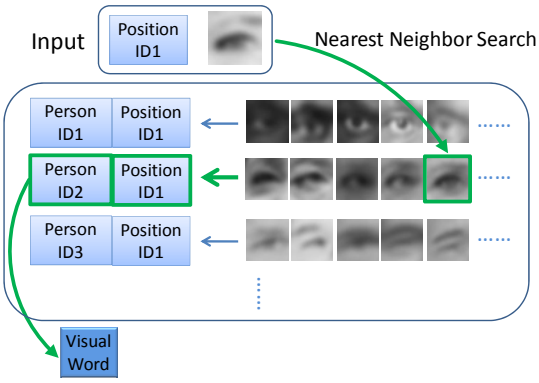

Figure 4. Identity-based quantization of a local feature extracted at the "Position ID1".

With the identify-based visual vocabulary, the quantization is simply performed by the nearest-neighbor search using k-d trees. For each position id, we build a k-d tree on all training features ($T \times P$ descriptors associated with the visual words, see Fig. 4) at the given position. Given a new face image, we extract 175 descriptors and find their nearest neighbors independently. The resulting pair <person id, position id> is the quantization result. Fig. 4 illustrates the quantization process. Mathematically, let S_i^j be the set of T feature descriptors associated with the visual word <person id=i, position id=j>. The obtained person id $ID(q^j)$ of a feature descriptor q^j at the j_{th} position can be computed as:

$$ID(q^j) = \arg\min_i \{\min_p \{d(q^j, p_i^j),\ p_i^j \in S_i^j\}\} \quad (1)$$

where $d(q, p)$ is the L_2 distance between features q and p.

To improve repeatability, we use soft quantization [20] - a descriptor is quantized to top R *identities* according Equation (1), where R is called "soft factor".

The number of persons P and the number of examples T affect the effective vocabulary size ($P \times 175$), the discriminative power-invariance tradeoff, and system complexity. Increasing P is equivalent to increasing the size of the vocabulary, thus increasing the discriminative power of visual words. At the mean time, increasing the example number T will lead to a more comprehensive face representation of the person, which will help reduce quantization errors. However, there is a trade-off between the discriminative power and the invariance to noises and appearance varia-

tions. Moreover, large P and/or T also increases the memory and computational cost for performing quantization. In our implementation, we choose P and T empirically, and find $P = 270$ and $T = 60$ performs best given a fixed budget of memory consumption.

Our approach is related to recent "simile classifier" [8] which also uses the traits information of a set of reference persons. In their work, a number of binary classifiers are trained from the selected regions using the reference persons, while we use the reference persons for the purpose of quantization of the local feature.

As demonstrated in the experiment later, our identity-based quantization can give good recall in the top 1,000 candidate images, even comparable with the exhaustively linear scan system using leading global face recognition feature. But the precision of the top candidate images may be lower due to the unavoidable quantization errors, as demonstrated in Figure 5. In the next section, we present a re-ranking method to improve the precision.

3. Global Multi-Reference Re-Ranking

Our basic idea is to re-rank the top candidate images using a very light and compact global signature so that we can improve the precision but without losing the scalability. In this section, we first describe a hamming signature and then present a multi-reference re-ranking method.

3.1. Hamming signature for global feature

Our compact hamming signature is based on a leading face recognition feature, called Learning-based (LE) Descriptor [1]. For the completeness, we brief the LE descriptor here. For the detected and aligned face, a DoG filter is first applied to remove illumination variations. Then, at each pixel a discrete code is computed by a learning-based encoder. These codes are further aggregated in a grid of cells to compute a code histogram within each cell. The result histograms are then concatenated and compressed with PCA (Principal Component Analysis) to form a 400-dimensional LE descriptor.

To create the hamming signature, we first randomly sample N_p projection directions in the original LE descriptor space. For each direction, the LE descriptors from a set of training face images are projected onto it. The median of the projected values is chosen as the threshold for that direction. Thus, the global LE descriptor G can be approximated by the following N_p-bit hamming signature:

$$\mathbf{B} = [b_1, b_2, \ldots, b_{N_p}], \quad b_i = \begin{cases} 1, & G \cdot P_i \geq h_i \\ 0, & G \cdot P_i < h_i \end{cases} \tag{2}$$

where P_i is the i_{th} random projection and h_i is the corresponding threshold in that projection.

The more projection directions we use, the better the approximation is [5]. In our implementation, we choose 320 random projections, i.e., $N_p = 320$. This results in a 40-byte compact hamming signature, which is an order of magnitude smaller than the original global LE descriptor in terms of both storage and computation (hamming distance can be efficiently computed by XOR operation).

Although hamming signature is an approximation, we will show that, by combined use of a multi-reference distance metric, it can achieve better retrieval precision than the linear scan system using the original 400-dimension LE descriptor.

3.2. Multi-reference re-ranking

The candidate images returned from traversing index are initially ranked based on the number of matched visual words, i.e., it is solely based on the query image. Images of one face contain variations induced by changes in pose, expression, and illumination. We account for such intra-class variations by using a set of reference images to re-rank the candidate images. In particular, we re-rank each candidate based on its average distance to the reference images.

In addition to be more robust to intra-class variations, the use of multiple references can also compensate the information lost during hamming embedding—while a false candidate image may confuse with the query image due to hamming embedding, it can hardly confuse with the majority of the reference images.

We need to be careful on selecting the reference images–inappropriate or incorrect references may hurt the system. In this paper, we use an iterative approach to select reference images from the returned top candidates. At each iteration, we select a reference image that is close to both the query image and the reference images from the previous iteration. More specifically, at each iteration we select an image I that minimizes the following cost:

$$D = d(Q, I) + \alpha \cdot \frac{1}{|\mathbf{R}|} \sum_i d(R_i, I), \tag{3}$$

where Q is the query image, $\mathbf{R} = \{R_i\}$ is the current reference set, $d(\cdot, \cdot)$ is the hamming distance between two faces, and α is a weighting parameter. I is then added to $\mathbf{R}$. The iterative process stops when the expected number of reference images are chosen, or the distance D is larger than a threshold.

The above iterative approach will select a cluster of reference images that are not only close to the query image but also close to each other. In particular, the second term in Equation (3) prevents selecting faces far from the center of the current reference images. By using a conservative threshold, the majority of the reference images are expected to be from the same person in the query image. Even though there might be some face images different from the

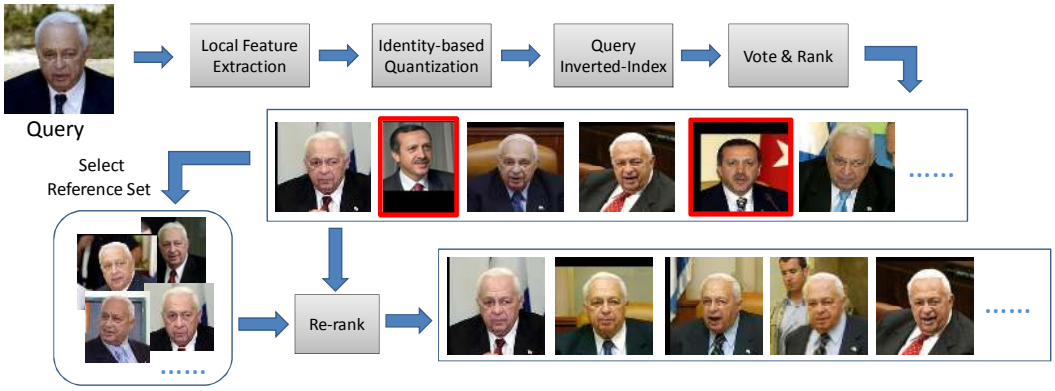

Figure 5. Face query pipeline. There are two major steps: 1) using local features to traverse index to collect candidate images, and 2) multi-reference ranking of candidate images. False positives are shown in red boxes.

person in the query, such "wrong" faces are still close (i.e., similar) to the query face image. As a result, they do not affect the performance much since the majority of the references are correct. Experiments showed that our "multi-reference" re-ranking is robust to "wrong" images, i.e., with 50% "wrong" images in the reference set, our re-ranking algorithm still performs as well. Figure 5 shows the basic process of the multi-reference re-ranking.

4. Experiments

4.1. Datasets

We use a face detector [24] to detect one million face images from the web to serve as our basic dataset. We then use 40 groups of labeled face images from the LFW (Labeled Face in Wild) dataset [4]. Each group consists of 16 to 40 face images of one person, and there are 1,142 face images in total. These 1,142 labeled images are added to the basic dataset to serve as our ground-truth dataset. In order to evaluate the scalability and retrieval performance with respect to the size of the database, we also sample the basic dataset to form three smaller datasets of 10K, 50K, and 200K images, respectively.

4.2. Evaluations

In the following evaluation, we select 220 representative face images from the ground-truth dataset to serve as our queries. Following existing image retrieval works [5], we use mean average precision (mAP) as our retrieval performance metric. For each query we compute its average precision from its precision-recall curve. The mAP is the mean value of all average precisions across 220 queries.

Baseline We use a fixed-grid approach as our baseline. The local features are sampled from a regular 16×11 grid over the face. As we mentioned before, interest-point detectors (such as DoG [12] or MSER [14]) perform worse than fixed-grid approach. The visual vocabulary is obtained by

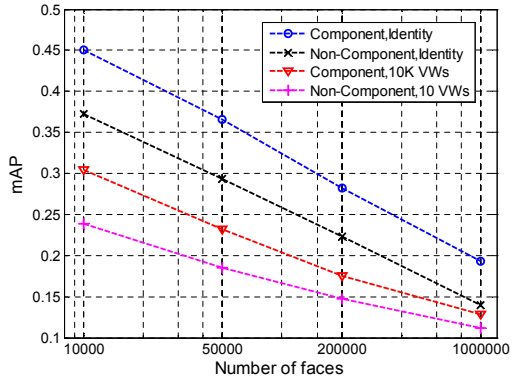

Figure 6. Comparison of "Component"-based and "Non-Component"-based local feature extraction approaches with different quantization methods. "Identity" means the identity-based quantization, and "10K VWs" is the baseline quantization with vocabulary size equal to 10K.

applying the hierarchical k-means clustering on 1.5 million feature descriptors. We call this "non-component-based baseline quantization". We evaluate the baseline with two vocabulary sizes: 10K and 100K visual words for each position in the grid, respectively. We have also experimented with 1K visual words for each grid position, but it performs worse than 10K or 100K. We set the soft quantization factor to 30, which performs best for baseline approaches. Note that we do not use soft quantization during indexing for baseline or our approaches.

To compared with state-of-the-art global face features, we also present the results using exhaustive linear scan of global features to retrieve top face images.

Local features evaluation Figure 6 shows the advantage of the component-based feature. Features extracted at the component level perform better for both baseline quantization and identity-based quantization. Figure 7 gives the mAP results using different quantization methods. Here the soft factor of identity-based quantization is set to 10, which performs the best. We can see that the identity-based

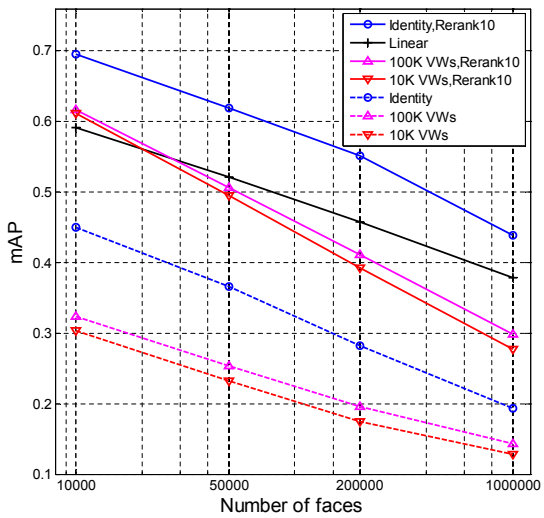

Figure 7. Comparison of "Identity"-base quantization and baseline quantization, with or without multi-reference re-ranking. "Identity, Rerank10" is the result of 10-reference re-ranking using hamming signatures. "10K VWs" and "100K VWs" are the baseline quantization with 10K and 100K visual words, respectively. "Linear" is the linear scan approach with global features. In our implementation, the re-ranking is performed on the top-1000 candidate images returned from traversing index.

quantization significantly outperforms the baseline quantization - with 1M images in the database, a 50.8% improvement over baseline quantization without multi-reference re-ranking. With multi-reference re-ranking, both quantization schemes have significant mAP improvements, and identity-base scheme still achieves a 58.5% improvement over baseline. Increasing the vocabulary size of the baseline quantization (from 10K to 100K visual words per grid position) slightly improves the mAP (by about 0.02), but it is still inferior to our identity-based quantization.

Multi-reference re-ranking We evaluate multi-reference re-ranking with several parameter settings, including 1) the number of reference images $N_r = 1$ or 10, and 2) using either the 400-dimensional global feature (see Section 3.1) or our hamming signature. Figure 8 shows that our multi-reference re-ranking significantly improves the mAP performance using either the compact hamming signature or the global feature. With $N_r = 10$ and hamming signature, our system, while having significantly less computational and storage cost, outperforms the approach of exhaustive linear scan using state-of-the-art global features.

We also have two interesting observations from Figure 8. First, multi-reference is important when using hamming signatures in the re-ranking. As we can see, with 1M images in the database, the mAP of $N_r = 10$ has a 47.8% improvement over $N_r = 1$ when using hamming signatures in re-ranking; the improvement is only 20.2% when using global features in re-ranking. For $N_r = 10$, using hamming signatures achieves a mAP similar to using global feature

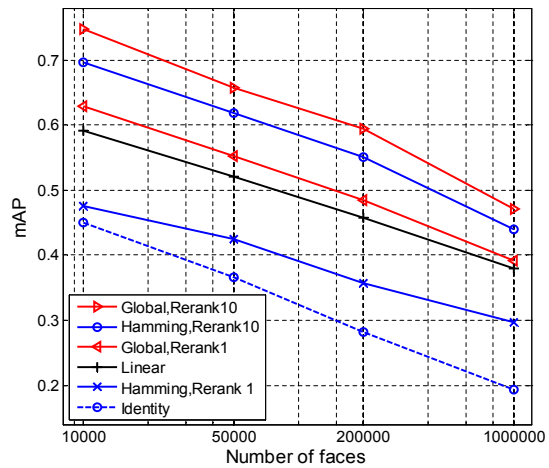

Figure 8. Comparison of different re-ranking methods. "Identity" is the result of identity-based quantization but *without* re-ranking. "Hamming, Rerank 1" means re-ranked by hamming signatures with one reference image (the query image). "Hamming, Rerank10" means re-ranked with 10 reference images. "Global, Rerank 1" and "Global,Rerank10" means re-ranked using global features with one and ten reference images, respectively. For reference purpose, we also include "Linear", the result of the linear scan approach.

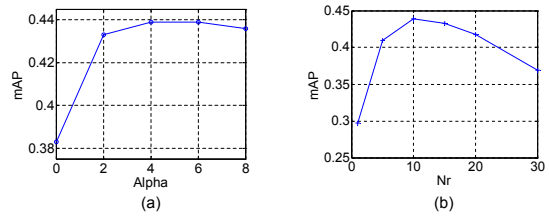

Figure 9. Comparison of different settings of N_r and α using 1M dataset: (a) mAP of various α values while fixing $N_r = 10$; (b) mAP of various N_r values while $\alpha = 6.0$.

re-ranking, but requires only about 10% storage space and is significantly faster.

Second, even with $N_r = 1$ (i.e., the reference image is the query image itself), re-ranking top-1000 candidates outperforms exhaustive linear scan the whole database using global features, as indicated by the curve "*Global,Rerank 1*" and the curve "*Linear*" in Figure 8. This indicates local features is complementary to global features—the candidate images chosen by the local features can be more easily differentiated and ranked by the global features.

Impact of N_r & α There are two parameters in our multi-reference algorithm: 1) the number of reference images N_r and 2) the value α in Equation 3 for selecting reference images. We set these two parameters empirically using our 1M dataset. To simplify the searching of N_r and α, we fix N_r while varying α, and vice versa. From Figure 9, we can see $(N_r = 10, \alpha = 6.0)$ is the optimal setting.

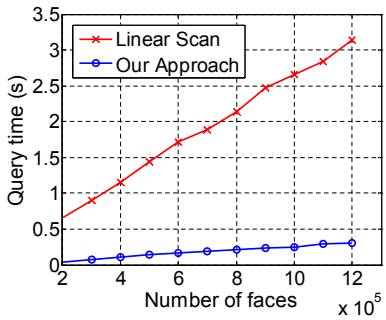

Figure 10. Query time per image (not including fixed time cost for face detection and feature extraction).

4.3. Scalability

To evaluate the scalability, we analyze the computational and storage cost with respect to the number of images in the database. We ignore fixed costs that are independent of database size as they do not affect the scalability.

Computational cost Let N be the number of images and D the dimension of the global feature. The computational cost of the linear scan approach is $N \times D$ of addition/minus/absolute operations. In our approach, for each local feature in the query, only a small portion of the index needs to be traversed. Denote C the percentage of the index need to be traversed, which is related to the vocabulary size and the soft quantization factor. Let N_F be the number of local features extracted from each face. The computational cost of our approach is $C \times N_F \times N$ voting operations. The value of $C \times N_F$ is one or two orders of magnitude smaller than D, also the voting operation is faster than L_1-norm computation. In other words, using indexing has significantly better scalability than the linear scan approach in terms of computational cost.

Figure 10 shows the query processing time w.r.t. the number of images in the database. We perform our experiments with a single 2.6GHz CPU on a desktop with 16G memory. Our approach scales well w.r.t. to database size.

Storage Cost In our implementation, we extract 175 visual words for each face image. An 1:4 compression ratio can be easily achieved by compressing the index [13]. On average, each visual word costs about 1 byte in the index. For each image, we store a 40-byte hamming signature instead of the original 400-dimensional global feature. Thus, in our system, the total storage cost for each face image is only about 200 bytes, which is easily scalable.

4.4. Example Results

In this section we compare and visualize results using real examples. Figure 11 shows the results of different approaches. We can see that there are 7 false positives in the top-10 images in (a), the baseline quantization approach without re-ranking. By using our identity-based quantization approach, the number of false positives is reduced to 3 as shown in (b). Our multi-reference re-ranking approach further improves the accuracy as shown in (d), which does not have any false positives in the top-10. We also present the results by linear scan with global features in (c), which are better than (a) & (b), but still inferior to our approach using re-ranking. This is also consistent to the mAP results shown in Fig. 8.

Note that the first image in the multi-reference ranking case (d) is different from the first image of the other three approaches. This is because multi-reference re-ranking uses the distance to the reference set rather than just the query image. As a result, the image sharing the most common appearance with the reference set will be ranked first. Fig. 12 gives more example results of our approach with identity-based quantization and multi-reference re-ranking.

5. Conclusion

We have designed a face image retrieval system with novel components that exploit face-specific properties to achieve both scalability and good retrieval performance, as demonstrated by experiments with a one-million face database. In our component-base local feature sampling, we currently treat 175 grid positions equally. In the future we plan to learn a weight for each grid position. In our identity-based quantization, we currently construct the visual word vocabulary by manually selecting 270 people and 60 face images for each person. An interesting future work is to design a supervised learning algorithm to automate this process to further improve the visual word vocabulary for face. Our system is highly scalable, and we plan to apply it on a web-scale image database using a computer cluster.

References

[1] Z. Cao, Q. Yin, J. Sun, and X. Tang. Face recognition with learning-based descriptor. In *CVPR*, 2010. 4

[2] J. Friedman, J. Bentley, , and R. Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software*, 3:209226, 1977. 2

[3] A. Gionis, P. Indyk, and R.Motwani. Similarity search in high dimensions via hashing. In *Int. Conf. on Very Large Data Bases*, 1999. 2

[4] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *ECCV*, 2008. 5

[5] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008. 1, 2, 4, 5

[6] H. Jegou, M. Douze, and C. Schmid. Packing bag-of-features. In *ICCV*, 2009. 2

[7] B. Kulis and K. Grauman. Kernelized locality-sensitive hashing for scalable image search. In *ICCV*, 2009. 2

[8] N. Kumar, A. C.Berg, P. N.Belhumeur, and S. K.Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 1, 2, 4

[9] P.-H. Lee, G.-S. Hsu, and Y.-P. Hung. Face verification and identification using facial trait code. In *CVPR*, 2009. 1

Figure 11. Example query results using different approaches. Left column is the query image and the top-ranked images are shown on the right. (a) baseline quantization approach; (b) identity-based quantization approach; (c) linear scan approach with the global features; (d) identity-based quantization with 10-reference re-ranking using hamming signatures. False positives are shown in red boxes.

Figure 12. Example results of our approach—identity-based quantization with 10-reference re-ranking using hamming signatures. Left column are the queries, and the top-ranked images are shown on the right. The first two query images are from the LFW dataset, and the last two queries are from Web.

[10] Z. Lei, S. Li, R. Chu, and X. Zhu. Face recognition with local gabor textons. *Lecture Notes in Computer Science*, 4642:49, 2007. 1, 2

[11] L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component-based discriminative search. In *ECCV*, 2008. 2

[12] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 20:91–110, 2003. 2, 5

[13] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. 7

[14] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002. 2, 5

[15] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005. 2

[16] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*, 65:43–72, 2005. 2

[17] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR'2006*. 1, 2, 3

[18] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 1, 2

[19] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 1, 2

[20] 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. 3

[21] N. Pinto, J. Dicarlo, and D. Cox. How far can you get with a modern face recognition test set using only simple features. In *CVPR*, 2009. 2

[22] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, Oct. 2003. 1, 2

[23] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Lecture Notes in Computer Science*, 4778:168, 2007. 2

[24] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 2, 5

[25] S. Winder and M. Brown. Learning local image descriptors. In *CVPR*, 2007. 2

[26] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Faces in Real-Life Images Workshop ECCV*, 2008. 2

[27] J. Wright and G. Hua. Implicit elastic matching with random projections for pose-variant face recognition. In *CVPR*, 2009. 1, 2

[28] L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Li. Face detection based on multi-block lbp representation. *Lecture Notes in Computer Science*, 4642:11, 2007. 2