Scalable Face Image Retrieval with Identity-Based Quantization and Multireference Reranking

Zhong Wu, Qifa Ke, *Member*, *IEEE*, Jian Sun, *Member*, *IEEE*, and Heung-Yeung Shum, *Fellow*, *IEEE*

Abstract—State-of-the-art image retrieval systems achieve scalability by using a bag-of-words representation and textual retrieval methods, but their performance degrades quickly in the face image domain, mainly because they produce visual words with low discriminative power for face images and ignore the special properties of 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 and 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 first retrieved from the inverted index of visual words. We then use a new multireference distance to rerank 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 multireference reranking 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.

Index Terms—Face recognition, content-based image retrieval, inverted indexing, image search.

1 Introduction

Given a face image as a query, our goal is to retrieve images containing faces of the same person appearing in the query image from a web-scale image database containing tens of millions face images. In this paper, we assume face images are frontal with up to about 20 degrees of pose changes, such that the five face components (e.g., eyes, nose, and mouth) are visible in a given face image. Fig. 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.

- *Z. Wu is with Microsoft Bing, City Center 8341, One Microsoft Way, Redmond, WA 98052. E-mail: zhowu@microsoft.com.*
- *Q. Ke is with Microsoft Research, 1065 La Avenida, Mountain View, CA 94043. E-mail: qke@microsoft.com.*
- *J. Sun is with Microsoft Research Asia, Building 2, No. 5 Danling Street, Haidian District, Beijing 100080, P.R. China. E-mail: jiansun@microsoft.com.*
- *H.-Y. Shum is with Microsoft Corporation, City Center 24888, One Microsoft Way, Redmond, WA 98052. E-mail: hshum@microsoft.com.*

Manuscript received 16 Mar. 2010; revised 27 Nov. 2010; accepted 15 Apr. 2011; published online 26 May 2011.
Recommended for acceptance by G. Hua, E. Learned-Miller, M. Turk, Y. Ma, T. Huang, M.-H. Yeh, and D. Kriegman.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMISI-2010-03-0189.
Digital Object Identifier no. 10.1109/TPAMI.2011.111.

A straightforward approach is to use the bag-of-visual-words representation that has been used in state-of-the-art scalable image retrieval systems [8], [21], [23], [28]. 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 in achieving both high discriminative power (to differentiate different persons) and invariance (to tolerate the variations of the same person). Second, 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 [6], [11], [12], [13], [22], [29], [33], [35]. However, these features are typically high-dimensional and global and 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 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 identity-based quantization can better handle intraclass variation

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

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 good recall in this index traversing stage. Then, the Hamming signatures (derived from global features), which are as small as 40 KB for 1,000 images, are used to rerank the candidate images. By using a new multireference 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 Difference of Gaussian (DoG) [15], Maximally Stable Extremal Region (MSER) [18] in the image, from which discriminative feature descriptors [15], [20], [32] are then computed. These descriptors are subsequently quantized into visual words with a visual vocabulary [28] which is trained by the unsupervised clustering algorithms [21], [23]. To further improve the scalability, Jegou aggregates partial information of the standard bag-of-features vector to build the "miniBOF" (mini-Bag-of-Features) vector [9], which is more compact in the index. On the other hand, to improve the precision, some compact information can be embedded [8] for each visual word in the index, which compensates for 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, Local Binary Pattern (LBP) [22] feature, variations of LBP [30], [34], [38], and V1-like feature [25] are designed to capture the micropatterns of the face. Besides these "low level" features mentioned above, Kumar et al. [11] incorporate 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 [35], Wright proposes an Rp-tree (Random Projection tree)-based approach to implicitly encode the geometric information into the feature. It is nontrivial to make these global feature-based methods scalable. One might consider using k-d tree [4] or Locality Sensitive Hashing (LSH) [5], [10] 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.

Our multireference reranking approach is closely related to Pseudo-Relevance Feedback (**PRF**) approaches [2], [3], [17], [26], [36], [37] originated from the query expansion techniques in text information retrieval. Standard **PRF** methods assume the top k initial retrieval results are relevant documents. The additional information obtained from these top k documents are then used in the Relevance Feedback [17], [27] step to expand the original query and to improve the retrieval precision/recall. Chen et al. [3] proposed PRF in content-based image retrieval by reweighting the visual words using the initial top k retrieval results. Both **RF** and **PRF** try to better understand and represent the query by incorporating textual features and other information from the selected relevant documents. In our multireference approach, we focus on improving the ranking precision of the top 1,000 face candidates without issuing a new expanded query. We try to obtain a more comprehensive and robust representation of the query face to account for appearance variations and information loss in Hamming signatures. We also assume that the reference images of the same identity have been retrieved to the top 1,000 candidates in the first stage using local features. However, instead of directly choosing the top k candidates ($k \ll 1{,}000$) as relevant documents as has been done in **PRF**, we use the global Hamming signatures in an iterative algorithm to robustly select the references from the top 1,000 initial return candidates, i.e., some of the top k candidates may not be in our reference set, while some candidates beyond top k may be in our reference set.

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

Fig. 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 [31] by a neural-network-based component detector [14]. 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 grid points from five components. From each grid point, we extract a square image patch. A T3hS2 descriptor (responses of steerable filters) [32] is then

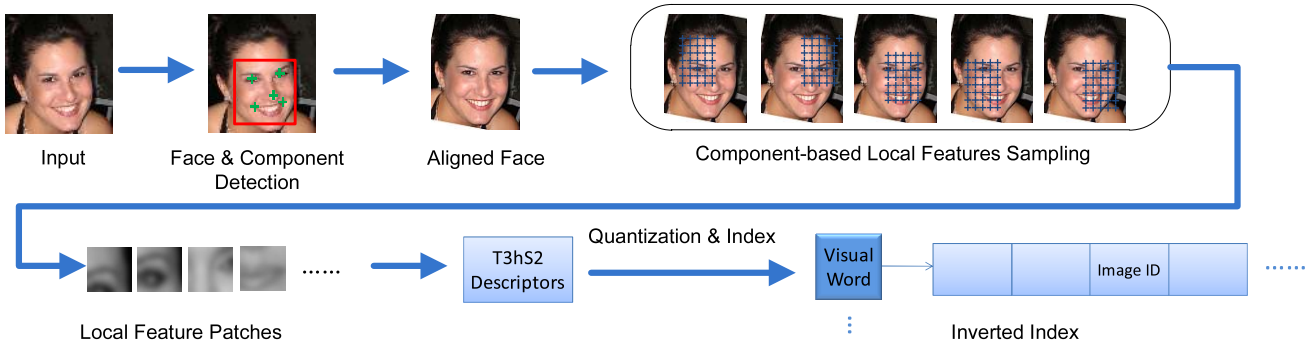

Fig. 2. Local feature extraction and indexing.

computed for each patch. All descriptors are quantized into visual words that are subsequently indexed. Notice that the existing interest point-based local feature detectors [19] are not suitable for face images. 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 grid points over the whole face [13], [35], 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 gird points 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 point a unique ID, which is called the "position id." The position id will be concatenated with the feature quantization id (described next) to form the "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 point in that component. This is in contrast to existing models that allow features to match even if they are coming from different grid points in the face, which performs worse in our task.

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) [21]. But unsupervised learning is not very good for training a vocabulary for face images, where intraclass variations are often larger than interclass 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 consist of P different people and each person has T face examples, at various poses, expressions, and illumination conditions. Fig. 3 shows example face images of one person and constructed visual words. Since each person has a unique "person id" and each grid point has a unique "position id," we define a visual word as the pair $<\text{person id}, \text{position id}>$ 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.

With the identity-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

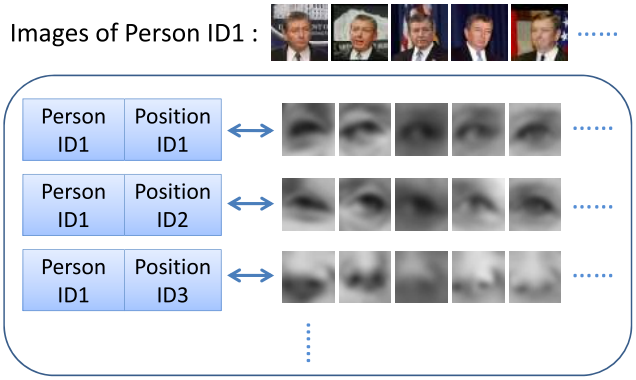

Fig. 3. Identity-based vocabulary from one person. A visual word is formed by concatenating two IDs: $<\text{person id}, \text{position id}>$. The final vocabulary is the combination of all visual words from all persons.

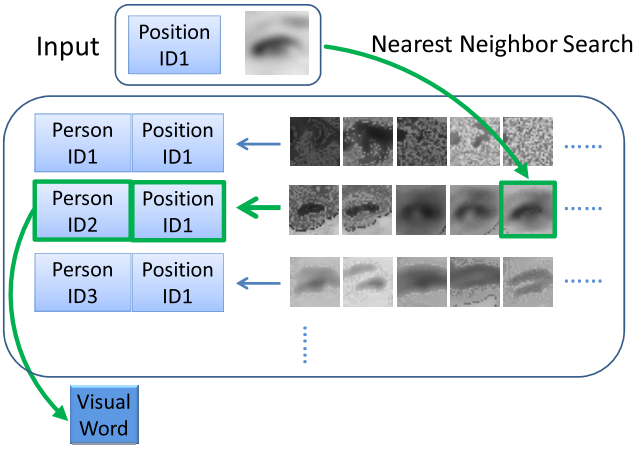

Fig. 4. Identity-based quantization of a local feature extracted at the "Position ID1."

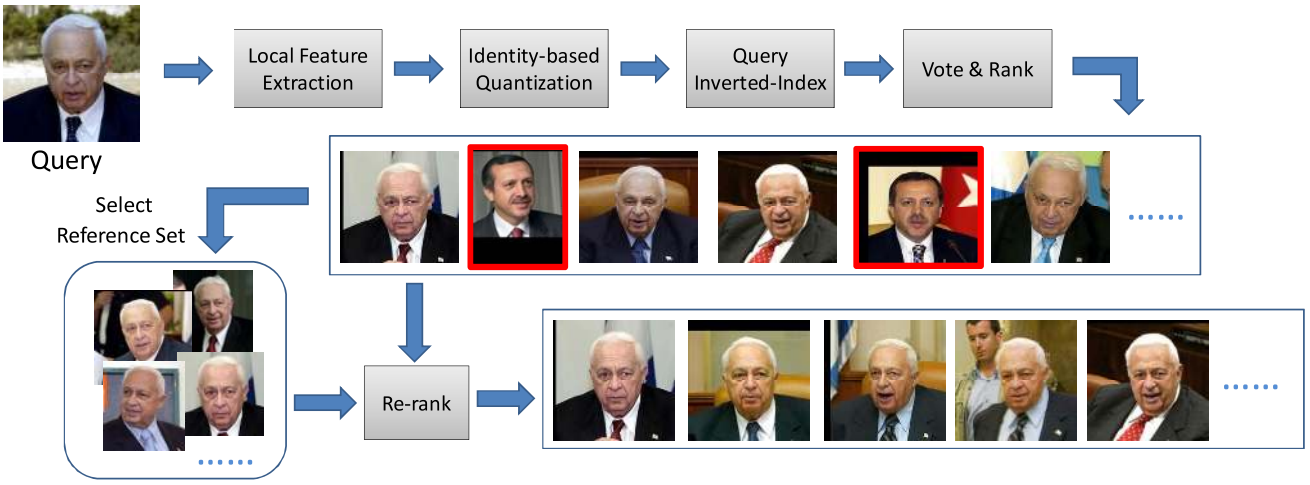

Fig. 5. Face query pipeline. There are two major steps: 1) using local features to traverse index to collect candidate images, and 2) multireference ranking of candidate images. False positives are shown in red boxes.

<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 jth position can be computed as:

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

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

To improve repeatability, we use soft quantization [24]—a descriptor is quantized to top R *identities* according to (1), where R is called a "soft factor." In our implementation, we first collect $3 \times R$ nearest neighbors for each descriptor. The top R distinct pairs of <person id, position id> in the resulting $3 \times R$ nearest neighbors are the soft quantization results. To avoid significantly increasing the storage of index, soft quantization is only applied to the query image.

The number of persons P and the number of examples T affect the effective vocabulary size ($P \times 175$), the discriminative power-invariance trade-off, and system complexity. Increasing P is equivalent to increasing the size of the vocabulary, thus increasing the discriminative power of visual words. In the meantime, 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 variations. 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" [11] 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 Fig. 5. In the next section, we present a reranking method to improve the precision.

3 Global Multireference Reranking

Our basic idea is to rerank 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 multireference reranking 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]. In the following, we briefly introduce the LE descriptor. For each detected face, a standard facial component [14] detector is used to extract fiducial points and align the face. A DoG filter (with $\sigma_1 = 2.0$ and $\sigma_2 = 4.0$) [6] is then applied to remove illumination variations. For each pixel p in the DoG-filtered image, its neighboring pixels are sampled from concentric rings centered at the pixel p. These sampled pixels are used to form a low level feature vector. We follow [1] to use a double-ring sampling pattern with an inner circle of radius $r_1 = 1$ and an outer circle of radius $r_2 = 2$. On each circle of radius r, we sample $r \times 8$ pixels at uniform intervals. These sampled pixels, as well as the pixel at the center, are normalized and then fed into a learning-based encoder to generate a discrete code. In our implementation, we use a random-projection tree-based encoder and set the code number to 256, which is reported in [1] to be a good trade-off of performance and computational cost. The input image is thus converted into a "code" image after the encoding. These codes are further aggregated in a grid of 5×7 cells, and a code histogram is computed for each cell. The result histograms are concatenated into a 8,960-dimensional vector ($256 \times 5 \times 7 = 8{,}960$), and then compressed by Principal Component Analysis (PCA) to form a 400D 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 ith random projection and h_i is the corresponding threshold in that projection.

The more projection directions we use, the better the approximation is [8]. 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 multireference distance metric, it can achieve better retrieval precision than the linear scan system using the original 400D LE descriptor.

3.2 Multireference Reranking

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 intraclass variations by using a set of reference images to rerank the candidate images. In particular, we rerank each candidate based on its average distance to the reference images.

In addition to being more robust to intraclass variations, the use of multiple references can also compensate for the information lost during Hamming embedding—while a false candidate image may be confused with the query image due to Hamming embedding, it can hardly be confused with the majority of the reference images.

We need to be careful in 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 (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 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 "multireference" reranking is robust to "wrong" images, i.e., with 50 percent "wrong" images in the reference set, our reranking algorithm still performs as well. Fig. 5 shows the basic process of the multireference reranking.

In our approach, we use a fixed size of reference set for all of the queries. One alternative is to set up some thresholds for choosing the reference set adaptively. However, in our experiments, the observation is this adaptive approach doesn't make significant improvement from our fixed reference size approach. One possible reason is that the face feature space is not uniform. Faces in a dense region have smaller distances among each other. Applying a threshold to decide the number of reference images may bring many unrelated images. On the other hand, faces in a sparse region of the feature space may result in insufficient reference images. Another variation of our multireference approach is to assign each reference image a weight while computing the distances from query images to the reference set. However, our experiments also show the change is minor by incorporating this weighting scheme.

4 Experiments

4.1 Data Sets

We use a face detector [31] to detect one million face images from the web to serve as our basic data set. We then use 40 groups of labeled face images from the Labeled Face in Wild (LFW) data set [7]. 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 data set to serve as our ground-truth data set. In order to evaluate the scalability and retrieval performance with respect to the size of the database, we also sample the basic data set to form three smaller data sets of 10, 50, and 200 K images, respectively.

4.2 Evaluations

In the following evaluation, we select 220 representative face images from the ground-truth data set to serve as our queries. Following existing image retrieval works [8], 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 [15] or MSER [18]) perform worse than the fixed-grid approach. The visual vocabulary is obtained by applying the hierarchical k-means clustering on 1.5 million feature descriptors. We call this the "non-component-based baseline quantization." We evaluate the baseline with two vocabulary sizes: 10 and 100 K visual words for each grid

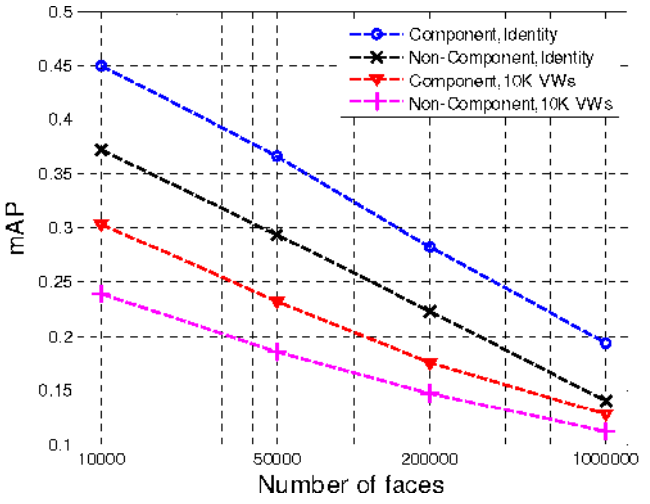

Fig. 6. Comparison of "component"-based and "non-component"-based local feature extraction approaches with different quantization methods. "Identity" means the identity-based quantization and "10K VWs" is the baseline quantization with vocabulary size equal to 10 K.

point position, respectively. We have also experimented with 1 K visual words for each grid point position, but it performs worse than 10 or 100 K. 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 compare 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. Fig. 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. Fig. 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 quantization significantly outperforms the baseline quantization—with 1 M images in the database, a 50.8 percent improvement over baseline quantization without multireference reranking. With multireference reranking, both quantization schemes have significant mAP improvements, and the identity-based scheme still achieves a 58.5 percent improvement over baseline. Increasing the vocabulary size of the baseline quantization (from 10 to 100 K visual words per grid point position) slightly improves the mAP (by about 0.02), but it is still inferior to our identity-based quantization.

Multireference reranking. We evaluate multireference reranking with several parameter settings, including 1) the number of reference images $N_r = 1$ or 10, and 2) using either the 400D global feature (see Section 3.1) or our Hamming signature. Fig. 8 shows that our multireference reranking 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 Fig. 8. First, multireference is important when using Hamming signatures in the reranking. As we can see, with 1 M images in the database, the mAP of $N_r = 10$ has a 47.8 percent

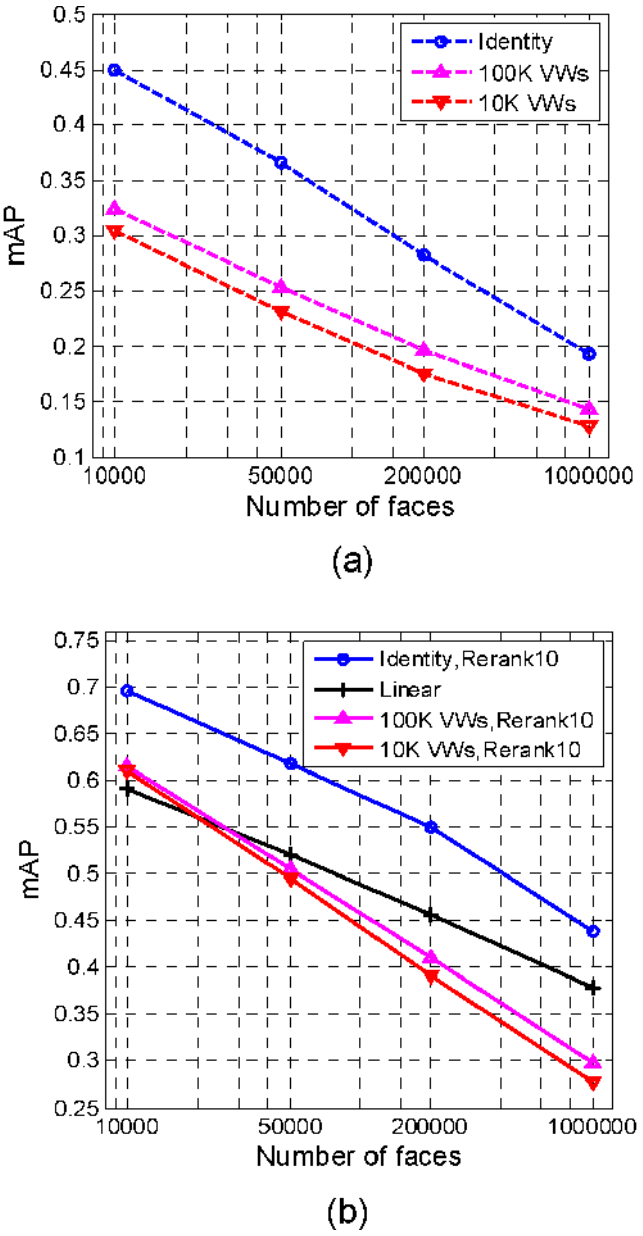

Fig. 7. (a) Comparison of "identity"-based quantization and baseline quantization. "Identity" is the result of our "identity"-based quantization approach. "10 K VWs" and "100 K VWs" are the baseline quantization with 10 and 100 K visual words, respectively. (b) Comparison of "identity"-based quantization and baseline quantization with multireference reranking. "Rerank10" is the result of 10-reference reranking using Hamming signatures. "Linear" is the linear scan approach with global features. In our implementation, the reranking is performed on the top-1,000 candidate images returned from the traversing index.

improvement over $N_r = 1$ when using Hamming signatures in reranking; the improvement is only 20.2 percent when using global features in reranking. For $N_r = 10$, using Hamming signatures achieves a mAP similar to using global feature reranking, but requires only about 10 percent storage space and is significantly faster.

Second, even with $N_r = 1$ (i.e., the reference image is the query image itself), reranking the top-1,000 candidates outperforms exhaustive linear scan of the whole database using global features, as indicated by the curve "*global, rerank 1*" and the curve "*linear*" in Fig. 8. This indicates local features are complementary to global features—the candidate images

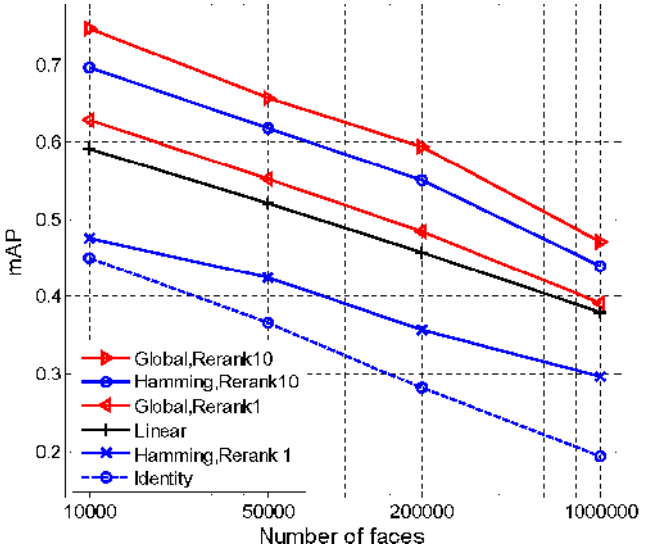

Fig. 8. Comparison of different reranking methods. "Identity" is the result of identity-based quantization but *without* reranking. "Hamming, rerank 1" means reranked by Hamming signatures with one reference image (the query image). "Hamming, rerank10" means reranked with 10 reference images. "Global, rerank 1" and "global, rerank10" means reranked using global features with one and 10 reference images, respectively. For reference purpose, we also include "linear," the result of the linear scan approach.

chosen by the local features can be more easily differentiated and ranked by the global features.

Impact of Hamming signature length. A larger Hamming signature gives better reranking accuracy, but with a larger storage and computational cost. From Fig. 9, we can see that in both cases of "Hamming, rerank 10" and "Hamming, rerank 1," the mAP consistently improves as the lengths of Hamming signatures increase. By using a Hamming signature with more than 400 bytes, the mAP is similar to or even slightly better than using the original 400D global features. With our multireference algorithm, with a smaller number of bytes we achieve an accuracy similar to that of using the global features. With $N_r = 10$, the 80-byte Hamming signatures achieve the same performance as "global, rerank 10." However, with $N_r = 1$ it requires 400 bytes to achieve the same performance as "global, rerank 1." This shows the importance of our multireference algorithm in compensating for the information loss during Hamming embedding, which is also consistent with our observation from Fig. 8.

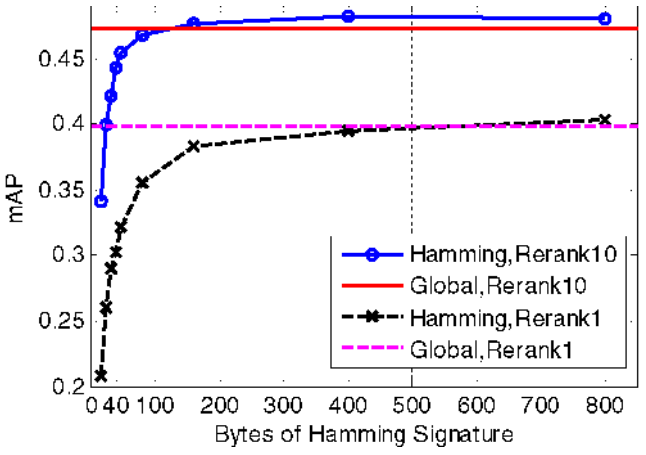

Fig. 9. Comparison of reranking methods using different lengths of Hamming signatures. "Hamming, rerank 1" means reranked by Hamming signatures with one reference image (the query image). "Hamming, rerank10" means reranked with 10 reference images. "Global, rerank 1" and "Global, rerank10" means reranked using 400D global features with one and 10 reference images, respectively.

TABLE 1
Runtime of Multireference Reranking Using Different Features

	Global	H-16	H-40	H-80
Time (ms)	70.1	5.5	9.8	18.6
	H-160	H-400	H-800	
Time (ms)	33.6	69.1	127.1	

"Global" means reranked using 400D global features with 10 reference images. "H-16" means reranked using 16-byte Hamming signatures with 10 reference images, and so on.

Table 1 compares the runtime of reranking methods using both original global feature and Hamming signatures with different lengths. We can see that with a proper length of code, the Hamming signature-based reranking approach is also significantly faster than the global feature-based approach while achieving similar performance.

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

Impact of multireference selection range and reranking range. Another two important parameters affecting the overall performance are: 1) the selection range of the reference images S, and 2) the reranking range M. Our multireference algorithm selects reference images in the top-S candidates using (3) and then reranks the top-M candidates. The selection range S should be large enough to cover the true positive candidates. However, a too large S can bring in more false positives to the reference set, which decreases the retrieval accuracy. In the other hand, the choice of the reranking range M is a trade-off of the recall and computational cost. In a practical image retrieval system, reranking more candidates may involve more network or I/O operations. The impact of these two parameters is shown in Fig. 11, empirically using our 1 M data set. We fix $M = 1,000$ while varying S and fix $S = 1{,}000$ while varying M. From Fig. 11a, we can see $S = 1{,}000$ is the optimal setting. And in Fig. 11b, we can also observe a mAP drop from $M = 10^5$ to $M = 10^6$, which again shows the complementarity between local features and global features. In our system, we set M to 1,000 as a trade-off of the retrieval accuracy and the computational cost of reranking.

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

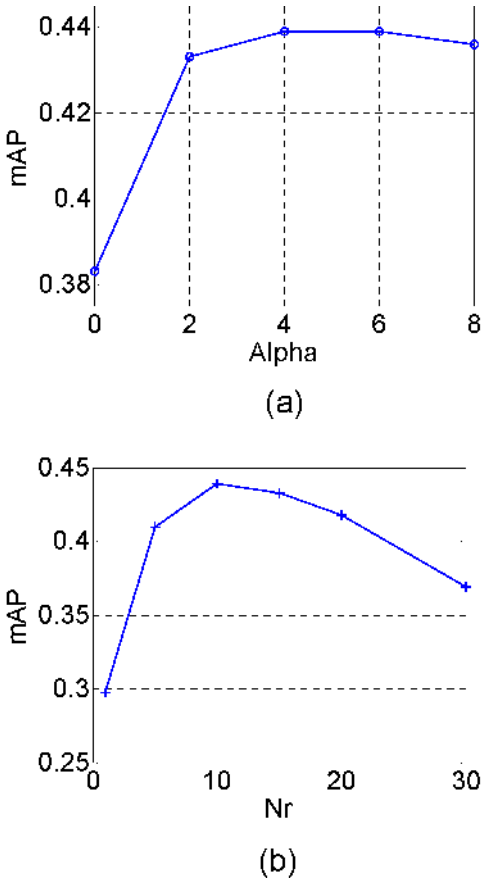

Fig. 10. Comparison of different settings of N_r and α using 1 M data set: (a) mAP of various α values while fixing $N_r = 10$. (b) mAP of various N_r values while $\alpha = 6.0$.

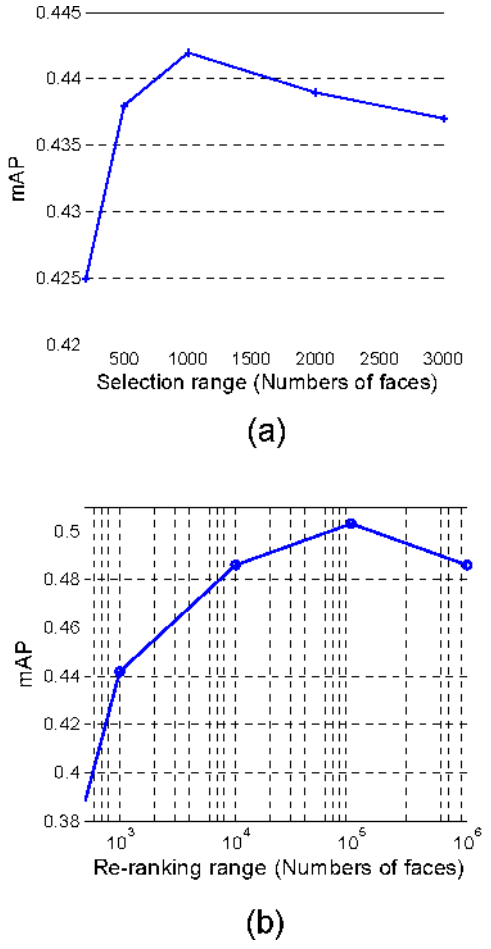

Fig. 11. Comparison of different settings of selection range (S) and reranking range (M) using 1 M data set: (a) mAP of various S values while $M = 1,000$. (b) mAP of various M values while fixing $S = 1,000$.

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.

Fig. 12 shows the query processing time w.r.t. the number of images in the database. We perform our experiments with a single 2.6 GHz CPU on a desktop with 16 G 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. A 1:4 compression ratio can be easily achieved by compressing the index [16]. 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 400D 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. Fig. 13 shows the results of different approaches. We can see that there are seven false positives in the top-10 images in Fig. 13a, the baseline quantization approach without reranking. By using our identity-based quantization approach, the number of false positives is reduced to three, as shown in Fig. 13b. Our multireference reranking approach further improves the accuracy as shown in Fig. 13d, which does not have any false positives in the top-10. We also present the results by linear scan with global features in Fig. 13c, which are better than Fig. 13a and Fig. 13b, but still

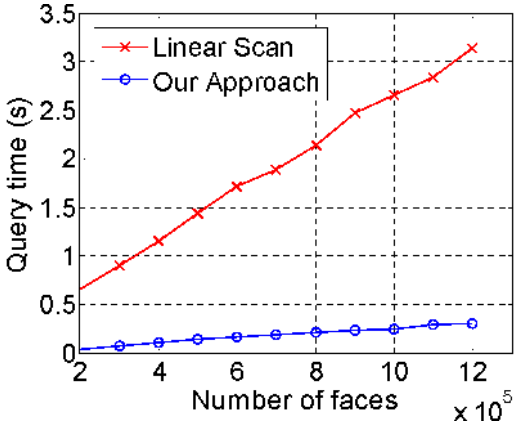

Fig. 12. Query time per image (not including fixed time cost for face detection and feature extraction).

Fig. 13. Example query results using different approaches. The 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 reranking using Hamming signatures. False positives are shown in red boxes.

inferior to our approach using reranking. This is also consistent to the mAP results shown in Fig. 8.

Note that the first image in the multireference ranking case (Fig. 13d) is different from the first image of the other three approaches. This is because multireference reranking 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. 14 gives more example results of our approach with identity-based quantization and multireference reranking.

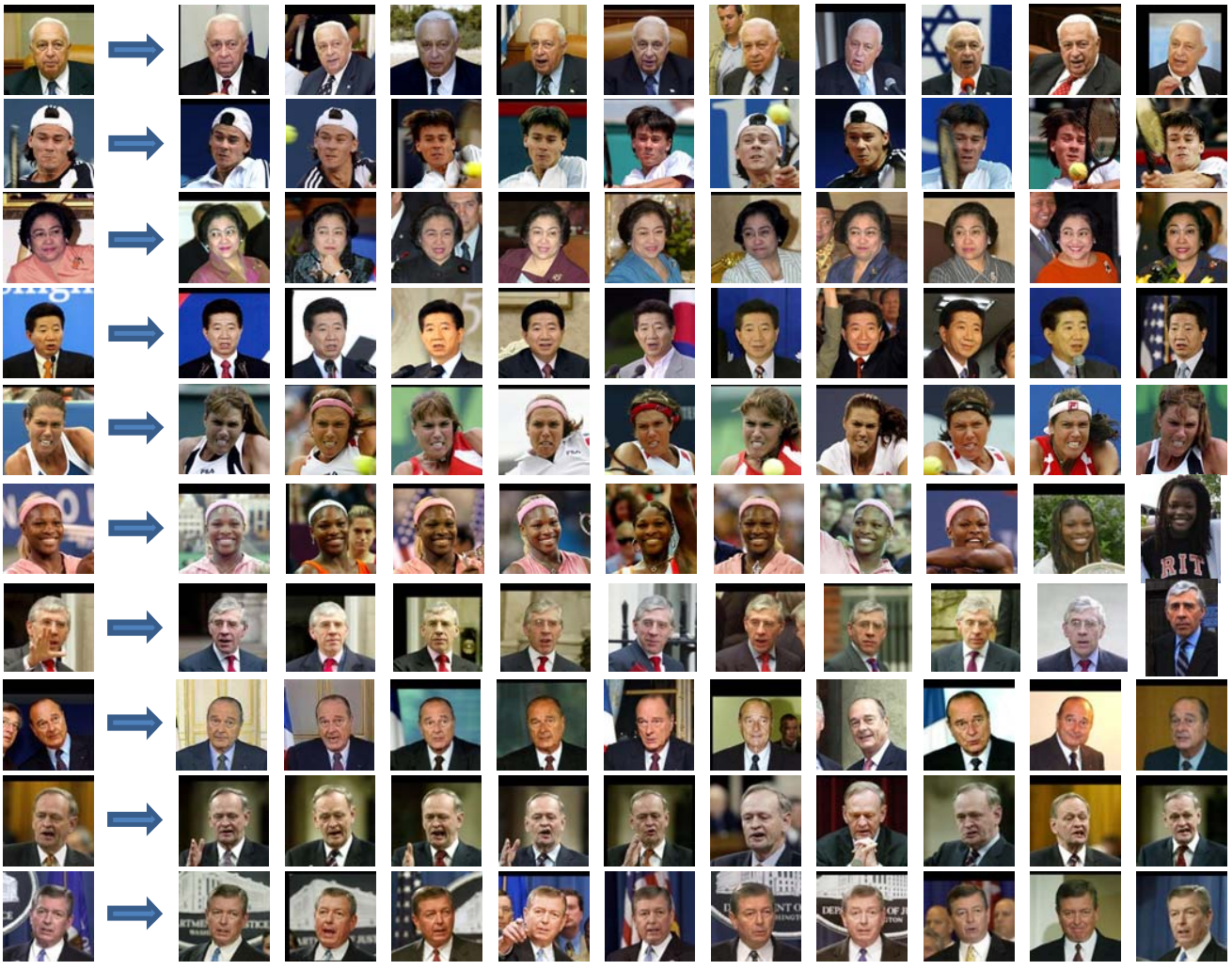

Fig. 14. Example results of our approach—identity-based quantization with 10-reference reranking using Hamming signatures. The left column are the queries and the top-ranked images are shown on the right.

Fig. 15. Challenging cases of our approach.

Fig. 15 shows two challenging cases of our approach. Faces of babies and faces with occlusion are generally similar and much easier to confuse with each other. However, our results show that even if our approach cannot find the faces of the same person, it can still retrieve a lot of other faces shared with some common attributes (baby, with sunglasses, etc.). This shows the potential of our approach in the applications of face attribute tagging, classification, etc.

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 point positions equally. In the future, we plan to learn a weight for each grid point 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 Descriptorand Pose-Adaptive Matching," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2010.

[2] J.G. Carbonell, Y. Yang, R.E. Frederking, R.D. Brown, Y. Geng, and D. Lee, "Translingual Information Retrieval: A Comparative Evaluation," *Proc. 15th Int'l Joint Conf. Artificial Intelligence,* pp. 708-714, 1997.

[3] J. Chen, R. Ma, and Z. Su, "Weighting Visual Features with Pseudo Relevance Feedback for cbir," *Proc. ACM Int'l Conf. Image and Video Retrieval,* 2010.

[4] J. Friedman, J. Bentley, and R. Finkel, "An Algorithm for Finding Best Matches in Logarithmic Expected Time," *ACM Trans. Math. Software,* vol. 3, pp. 209-226, 1977.

[5] A. Gionis, P. Indyk, and R. Motwani, "Similarity Search in High Dimensions via Hashing," *Proc. 25th Int'l Conf. Very Large Data Bases,* 1999.

[6] G. Hua and A. Akbarzadeh, "A Robust Elastic and Partial Matching Metric for Face Recognition," *Proc. IEEE 12th Int'l Conf. Computer Vision,* 2009.

[7] 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," *Proc. European Conf. Computer Vision,* 2008.

[8] H. Jegou, M. Douze, and C. Schmid, "Hamming Embedding, and Weak Geometric Consistency for Large Scale Image Search," *Proc. 10th European Conf. Computer Vision,* 2008.

[9] H. Jegou, M. Douze, and C. Schmid, "Packing Bag-of-Features," *Proc. IEEE 12th Int'l Conf. Computer Vision,* 2009.

[10] B. Kulis and K. Grauman, "Kernelized Locality-Sensitive Hashing for Scalable Image Search," *Proc. IEEE 12th Int'l Conf. Computer Vision,* 2009.

[11] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar, "Attribute, and Simile Classifiers for Face Verification," *Proc. IEEE 12th Int'l Conf. Computer Vision,* 2009.

[12] P.-H. Lee, G.-S. Hsu, and Y.-P. Hung, "Face Verification, and Identification Using Facial Trait Code," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2009.

[13] Z. Lei, S. Li, R. Chu, and X. Zhu, "Face Recognition with Local Gabor Textons," *Proc. Int'l Conf. Biometrics,* pp. 49-57, 2007.

[14] L. Liang, R. Xiao, F. Wen, and J. Sun, "Face Alignment via Component-Based Discriminative Search," *Proc. 10th European Conf. Computer Vision,* 2008.

[15] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision,* vol. 20, pp. 91-110, 2003.

[16] C.D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval.* Cambridge Univ. Press, 2008.

[17] C.D. Manning, P. Raghavan, and H. Schütze, "Relevance Feedback and Query Expansion," *Introduction to Information Retrieval,* pp. 177-194, Cambridge Univ. Press, 2008.

[18] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions," *Proc. British Machine Vision Conf.,* 2002.

[19] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A Comparison of Affine Region Detectors," *Int'l J. Computer Vision,* vol. 65, pp. 43-72, 2005.

[20] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 10, pp. 1615-1630, Oct. 2005.

[21] D. Nister and H. Stewenius, "Scalable Recognition with a Vocabulary Tree," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2006.

[22] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 7, pp. 971-987, July 2002.

[23] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object Retrieval with Large Vocabularies, and Fast Spatial Matching," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.

[24] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[25] N. Pinto, J. Dicarlo, and D. Cox, "How Far Can You Get with a Modern Face Recognition Test Set Using Only Simple Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2009.

[26] S. Rudinac, M. Larson, and A. Hanjalic, "Exploiting Visual Reranking to Improve Pseudo-Relevance Feedback for Spoken-Content-Based Video Retrieval," *Proc. Workshop Image Analysis for Multimedia Interactive Services,* 2009.

[27] G. Salton, "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer," *Proc. Int'l Symp. Mobile Agents,* 1989.

[28] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," *Proc. IEEE Ninth Int'l Conf. Computer Vision,* Oct. 2003.

[29] Y. Taigman, L. Wolf, T. Hassner, and I. Tel-Aviv, "Multiple One-Shots for Utilizing Class Label Information," *Proc. British Machine Vision Conf.,* 2009.
[30] X. Tan and B. Triggs, "Enhanced Local Texture Feature Sets for Face Recognition under Difficult Lighting Conditions," *Proc. Third Int'l Conf. Analysis and Modeling of Faces and Gestures,* pp. 168-182, 2007.
[31] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2001.
[32] S.A.J. Winder and M. Brown, "Learning Local Image Descriptors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[33] L. Wiskott, J. Fellous, N. Kruger, and C. Von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. Pattern Analysis and Machine Intellignece,* vol. 19, no. 7, pp. 775-779, July 1997.
[34] L. Wolf, T. Hassner, and Y. Taigman, "Descriptor Based Methods in the Wild," *Proc. Faces in Real-Life Images Workshop European Conf. Computer Vision,* 2008.
[35] J. Wright and G. Hua, "Implicit Elastic Matching with Random Projections for Pose-Variant Face Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2009.
[36] R. Yan, A. Hauptmann, and R. Jin, "Multimedia Search with Pseudo-Relevance Feedback," *Proc. Int'l Conf. Image and Video Retrieval,* 2003.
[37] R. Yan, A.G. Hauptmann, and R. Jin, "Negative Pseudo-Relevance Feedback in Content-Based Video Retrieval," *Proc. ACM 11th Int'l Conf. Multimedia,* pp. 343-346, 2003.
[38] L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Li, "Face Detection Based on Multi-Block LBP Representation," *Proc. Int'l Conf. Biometrics,* pp. 11-18, 2007.

Zhong Wu received the BS degree in automation from Tsinghua University, Beijing, China, in 2006 and the PhD degree in computer science and technology from the Institute for Advanced Study, Tsinghua University, Beijing, China, in 2011. Currently, he is a software development engineer in the Microsoft Bing Multimedia team. His research interests include computer vision, Internet image/video search, and machine learning.

Qifa Ke received the BS degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, in 1994, the MS degree in pattern recognition and artificial intelligence from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1997, and the PhD degree in computer science from Carnegie Mellon University, Pittsburgh, Pennsylvania, in 2003. From 2003 to 2006, he was a systems scientist in the Computer Science Department of Carnegie Mellon University. He is now a researcher at Microsoft Research Silicon Valley. His research interests include Internet image/video search, large scale data management and analysis, computer vision, and machine learning. He is a member of the IEEE.

Jian Sun received the BS, MS, and PhD degrees from Xian Jiaotong University in 1997, 2000, and 2003, respectively. He joined Microsoft Research Asia in 2003. His research interests include the fields of computer vision and computer graphics, with particular interests in interactive computer vision (user interface + vision), and Internet computer vision (large image collection + vision). He is also interested in stereo matching and computational photography. He is a member of the IEEE.

Heung-Yeung Shum received the doctorate degree in robotics from the School of Computer Science at Carnegie Mellon University in Pittsburgh, Pennsylvania. He is the corporate vice president responsible for search product development at Microsoft Corporation, www.bing.com. He joined Microsoft Research in 1996 as a researcher based in Redmond, Washington. He moved to Beijing as one of the founding members of Microsoft Research China (later renamed Microsoft Research Asia). His tenure there began as a research manager, subsequently moving on to become assistant managing director, managing director of Microsoft Research Asia, distinguished engineer, and corporate vice president. He has published more than 100 papers about computer vision, computer graphics, pattern recognition, statistical learning, and robotics. He holds more than 50 US patents. He is a fellow of the IEEE and the ACM for his contributions on computer vision and computer graphics.