Bag-of-Features Based Medical Image Retrieval via Multiple Assignment and Visual Words Weighting

Jingyan Wang, *Student Member, IEEE,* Yongping Li, *Member, IEEE,* Ying Zhang, Chao Wang, Honglan Xie, Guoling Chen, and Xin Gao

***Abstract*—In this paper, we investigate the bag-of-feature based medical image retrieval methods, which represent an image as a collection of local features, such as image patch and key points with SIFT descriptor. To improve bag-of-feature method, we first model the assignment of local descriptor as contribution functions, and then propose a new multiple assignment strategy. Assuming the local feature can be reconstructed by its neighboring visual words in vocabulary, we solve the reconstruction weights using QP problem and use them as contribution functions, resulting a new assignment methods, called QP assignment. At the same time, we also propose a novel visual weighting method. We first analysis each visual word by modeling the sub-similarity or sub-distance function comparing only one single bin corresponding to the visual word; then we treat each of them as a weak classifier for triplets and learn a strong classifier, the resulting weights will be used as visual weighting factors. We carry our experiments on three medical image datasets: the ImageCLEFmed dataset, the 304 CT Set and the Basal-Cell Carcinoma Image set. The vast experiments results show that our proposed methods have many advantages and work well for the bag-of-feature based medical image retrieval tasks.**

***Index Terms*—Medical Image Retrieval, Bag-of-Features, Multiple Assignment, QP Problem, Visual Words Weighting, Boosting.**

I. Introduction

THROUGHOUT the world, rapid growth of computerized medical imagery using picture archiving and communication systems (PACS) in hospitals has generated a critical need for efficient and powerful search engines. In addition, the growing workload on radiologists in recent years increases the need for computerized systems which could help the radiologist in prioritization and in the diagnosis of findings [1].

As an important complementary search approach, content-based image retrieval (CBIR) has been one of the most vivid research areas in the field of computer vision over the last 10 years. In the medical field, CBIR also draws extensive attention [2]. Traditional global features include color features, texture features, and shape features. Recently, along with the rapid progress in the application of local descriptor in pattern recognition, computer vision and image retrieval , the bag-of-features deriving from local features like keypoints or image patches has appeared as promising for object classification and image retrieval [1], [2]. Unlike text retrieval, image retrieval should create visual word first. Usually, k-means is adopted to cluster centers of features which are extracted from all images; then, these cluster centers are used as a vocabulary for all images to obtain word vector representations.

Avni Uri etc. once presented an X-ray image categorization and retrieval method using patch-based visual words representation [1], while Zhi Li-Jia etc. developed a medical image retrieval method using SIFT feature [2]. Juan C. Caicedo etc. raised an evaluation of different representations obtained from the bag of features approach to classify histopathology images, including both the image patch and SIFT local features [21]. All the above methods build the histogram for image representation by assigning the local image feature descriptors to the single nearest visual word in the vocabulary, which is called Nearest Neighbor (NN) Assignment in this paper. However, one inherent component of the transitional NN model is the assignment of discrete visual words to continuous image features, which shows a clear mismatch of this hard assignment with the nature of continuous features [9]. By explicitly modeling the ambiguity of visual word assignment , Jan C. van Gemert etc. improved classification performance compared to the hard assignment of the traditional codebook model based bag-of-feature methods [9]. However, the assignment is based on the usage of Gaussian kernel, which is very sensitive to the smoothing parameter σ. Herve Jegou increased the performance by using multiple assignment of descriptors to visual words at the cost of reduced efficiency [8]. The disadvantage of this method is it treats all the candidate nearest neighboring visual words equally without considering the neighborhood structure of descriptors and visual words. In [26], Yang et al. develop an extension of the spatial pyramid matching (SPM) method, by generalizing the NN assignment based vector quantization to sparse coding (SC) followed by multi-scale spatial max pooling, and propose a linear SPM kernel based on SIFT sparse codes. Yang et al. argued that The NN assignment may be too restrictive, giving rise to often a coarse reconstruction of local feature space. They relax the constraint by instead putting a L_1-norm regularization on cluster membership indicators, which enforces cluster membership indicators to have a small number

Jingyan Wang, Yongping Li ,Ying Zhang and Honglan Xie are with Shanghai Institute of Applied Physics, Chinese Academy of Science, 2019 Jialuo Road, Jiading District, Shanghai 201800, P. R. China. E-mail: {wangjingyan, ypli, zhangying1, xiehonglan}@sinap.ac.cn.

Chao Wang is with OGI School of Science & Engineering, Oregon Health & Science University (OHSU), Beaverton, Oregon, 97006, US. E-mail: wangcha@csee.ogi.edu.

Guoling Chen is with Zhongshan Hospital, Fudan University, 180 Fenglin Road, XuHui District, Shanghai 200032, P. R. China. E-mail: elebell@163.com

Xin Gao is with Mathematical and Computer Sciences and Engineering Division, King Abdullah University of Science and Technology, Jeddah 21534, Saudi Arabia. E-mail: xin.gao@kaust.edu.sa.

Manuscript received April 19, 2010; revised June 2, 2011.

of nonzero elements. However, this method assume that a local feature is reconstructed by all the visual words in the vocabulary, bring a redundance and making the computation much complex. In this paper, we will develop a new multiple assignment method by assuming the local descriptor can be linearly constructed by its neighboring visual words. This is different from SC basically on the local reconstruction, while SC is based on a global reconstruction.

At the same time, the visual word weighting methods assign appropriate weights to the visual words to improve the performance of medical image retrieval and classification. For example, the weighting scheme provided by inverse document frequencies (IDF) is also employed by Juan C. Caicedo etc. [21]. However, in their experiment, it is not clear when IDF improves the classification performance. In [27], Cai et al. present a visual word weighting factors learning approach for image classification and retrieval. It corresponds to learning a weighted similarity metric to satisfy that the weighted similarity between the same labeled images is larger than that between the differently labeled images with largest margin. In this paper, we try to develop a new visual word weighting factor using the boosting methods, which will treat each visual word as a weak classifier and combine them together using these weighting factors.

The contributions of this paper are twofold:

1) Firstly, we model the local descriptor by proposing the contribution function, and the other assignment strategies can be described generally by this model using different contribution functions. Then we give the novel assignment strategy by using the linear construction weighting of neighboring visual words as contribution functions. The weighting can be solved by modeling a quadratic programming (QP) problem, so we call it QP assignment.
2) Secondly, given the similarity and distance measure for the comparison of histogram, we propose the sub-similarity and sub-distance to analyze the discriminating ability of the visual words. Then we construct the weak classifier for triplets of medical images using the sub-similarity or distance functions, and learn their weighting factors applying the Boostmap algorithm.

This paper is organized as follows: Section II reviews Bag-of-Features Based Medical Image Retrieval Framework. The newly proposed descriptor assignment algorithm is described in Section III. In Section IV-B, the algorithm for learning the weighting of the visual word is presented, which is based on the boosting algorithm. Experimental evaluation is presented in Section V. In Section VI, we conclude the paper by tracing back the origins of our work and point out major differences and improvements made here as compared to other well-known algorithms.

II. Bag-of-Features Based Medical Image Retrieval Framework

The basic steps of the bag-of-features image retrieval algorithm is illustrated in Fig. 1, where each block represents an operation on the objects.

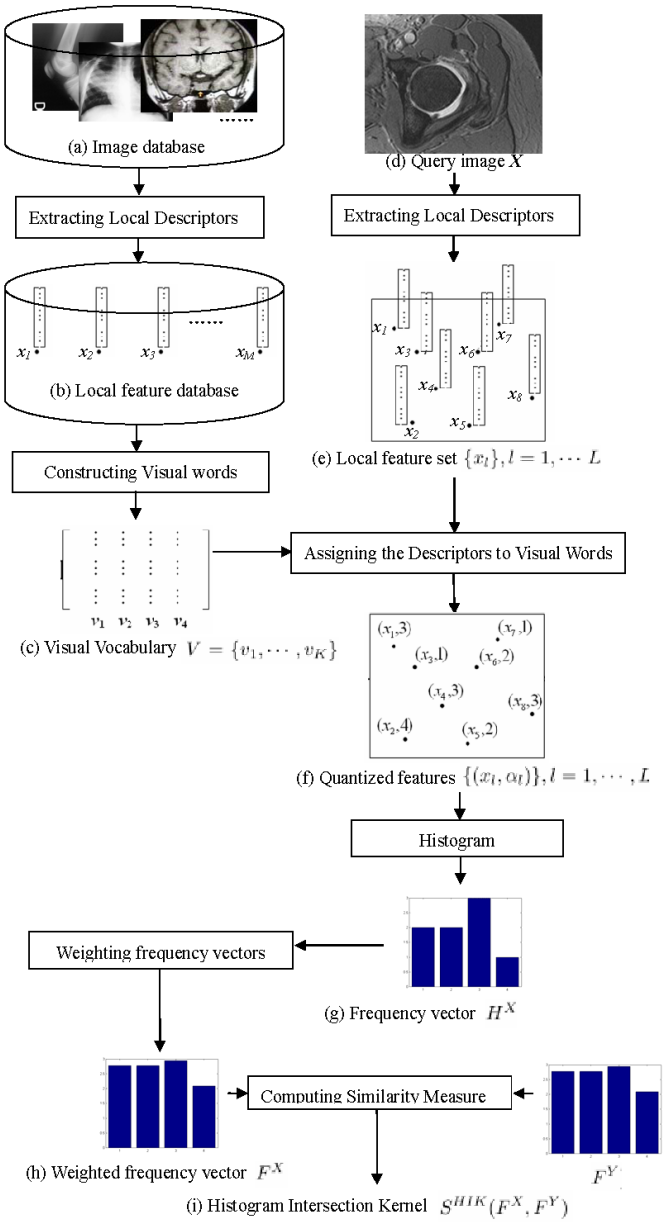


Fig. 1. Bag-of-Features Based Medical Image Retrieval Framework. For the compactness of the figure, we only shown the quantization procedure of query image X. Note that the database images in (a) also should be processed as X.

- **Extracting Local Descriptors**. We start by representing an input image as a collection of local feature descriptors. Let X be a bag of features of an image and $\{x_l\}, l = 1, \cdots L$ be a collection of local features extracted from X, as is shown in Fig. 1 (e). Typically, the local features or regions x_l are detected interest regions with SIFT descriptors [5], [6], or densely sampled image patches [23].
- **Constructing Visual words**. In the learning phase, we construct a Visual Vocabulary V using a clustering algorithm. Usually, k-means is used to cluster centers of features which are extracted from all images in the

database; then, these cluster centers are used as a vocabulary (codebook) $V = \{v_1, \cdots, v_K\}$ with K visual words for all images to get word vector representation, as is shown in Fig. 1 (c).

- **Assigning the Descriptors to Visual Words**. Each local descriptor x_l of a given image X is assigned to the closest visual word v_k. So, in this first coding stage, the image X is represented by the local feature $\{(x_l, \alpha_l)\}, l = 1, \cdots, L$, with each α_k identified with the integers $k = 1, \cdots, K$.

$$\alpha_l = \arg\min_k (D(v_k, x_l)) \tag{1}$$

where x_l is an image region l, and $D(v_k, x_l)$ is the distance between a codeword v_k and region x_l, as is shown in Fig. 1 (f). The histogram of visual word occurrences is subsequently normalized with the L_1 norm, generating a frequency vector $H^X = [h_1^X \ h_2^X \ \cdots \ h_K^X]$, as is shown in Fig. 1 (g).

As a variant, instead of choosing the nearest neighbor, a given local descriptor is assigned to the several nearest visual words. This variant will be referred to as multiple assignment (MA). We will improve MA in this paper by developing a new assignment method.

- **Weighting frequency vectors**. The components of the frequency vector are then weighted with a strategy similar to the one in [3]. Denoting by t_k the weighting factor for k-th bin, corresponding to visual word v_k, the weighted component f_k^X associated with image X is given by

$$f_k^X = t_k \cdot h_k^X \tag{2}$$

A common used weighting factor is inverse document frequency (idf) factor

$$t_{idf} = log\frac{n}{n_k} \tag{3}$$

where n is the number of images (bags) in the database and n_k is the number of images containing the k-th visual word v_k [15]. The resulting visual word frequency vector $F^X = [f_1^X \ f_2^X \ \cdots \ f_K^X]$ or simply visual word vector, is a compact representation of the image, as is shown in Fig. 1 (h).

In this paper, we proposed a new frequency vectors weighting strategy, which weight each visual word according to its contribution to the classification of the query images.

- **Computing Similarity Measure**. Frequency vectors is a kind of histogram, according to [7], Histogram Intersection Kernel (HIK) is suitable to be a measure to compute similarity between query image X's histogram F^X and the image Y's histogram F^Y in database:

$$S^{HIK}(F^X, F^Y) = \sum_{k=1}^{K} min(f^X k, f_k^Y) \tag{4}$$

Then the the images in the database will be ranked according to their similarities to the query image, and several most similar image will be returned as retrieval results.

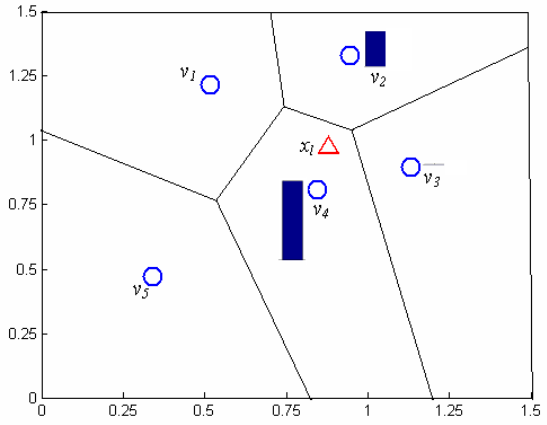


Fig. 2. An example illustrating assigning local feature descriptors x_l to visual words $v_i, i = 1, \cdots, 5$ in the Visual Vocabulary V, where $v_1 = (0.5180, 1.2140)$, $v_2 = (0.9460, 1.3260)$, $v_3 = (1.1320, 0.8920)$, $v_4 = (0.8460, 0.8060)$, $v_5 = (0.3420, 0.4700)$ and $x_l = (0.8780, 0.9660)$. Beside x_l's neighboring visual words v_2, v_3 and v_4, we also plot their weights that QP finds to reconstruct x_l.

III. QP Multiple Assignment

In this section, we generalize the bag-of-features based image presentation as a visual word contribution form. First of all, for a local feature descriptor x_l in an image X, we define it's contribution in the constructing of an statistical presentation (histogram) as Visual Word Contribution Function.

Here, we define the Visual Word Contribution Function of x_l for v_k as its contribution in accumulating the k-th bin as $C(v_k, x_l)$, so that for visual vocabulary based histogram can be rewritten as

$$h_k^X = \sum_{j=1}^{m_i} C(v_k, x_l) \tag{5}$$

With a 2D visual vocabulary $V = \{v_i\}, i = 1, 2, \cdots, 5$ of size of 5, we give an illustrative examples of assigning local feature descriptor x_l in Fig. 2. The Visual Word Contribution Functions $C(v_i, x)$ of different assignment strategies are shown in Fig. 3.

At query time, we assign a local feature x to its nearest visual word $v_k \in V$ to construct the frequency vector H^X. So for the transitional visual word assignment, we define the Nearest Neighbor (NN) contribution function as follows,

- **Nearest Neighbor Assignment**.

$$C_{NN}(v_i, x_l) = \begin{cases} 1, & if \ v_k = NN(x_l); \\ 0, & else. \end{cases} \tag{6}$$

where $NN(x_l)$ is the nearest neighboring visual word in visual vocabulary V. The NN assignment's contribution function $C_{NN}(v_i, x_l)$ is shown in Fig. 3 (a).

It is also possible to assign a descriptor to not only one but several nearest visual words (using approximate visual word assignment) [8]. Our strategy is similar to the multiple descriptor assignment proposed in [8] or the soft quantization method proposed in [9]. By defining the two different Contribution Functions, we can formula the multiple assignment and soft assignment as follows:

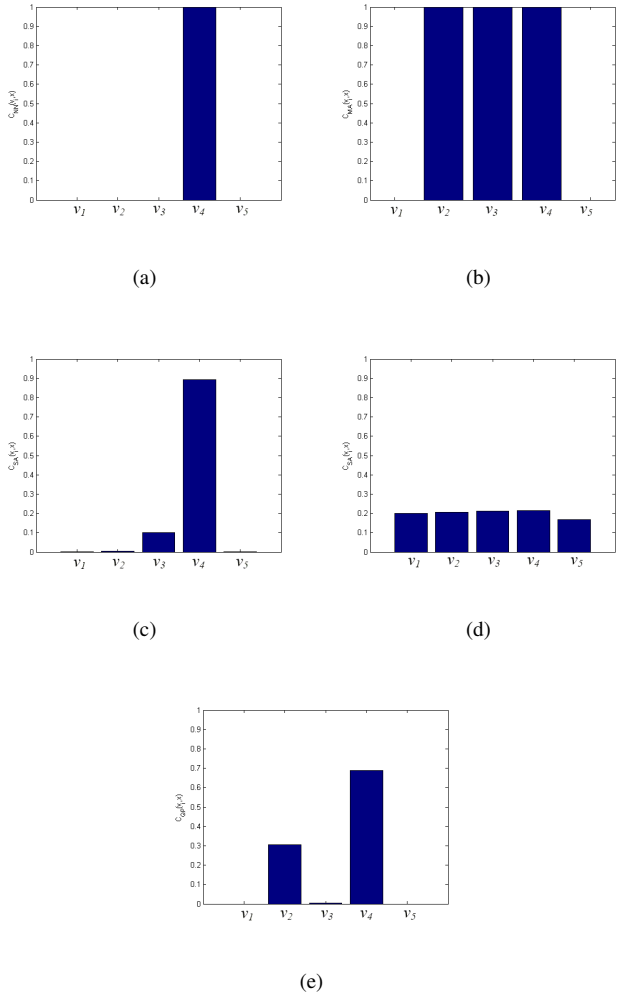


Fig. 3. Contribution functions $C(v_i, x)$ for different assignment strategies (with $N(x) = \{v_2, v_3, v_4\}$): (a) NN assignment $C_{NN}(v_i, x)$; (b) MA assignment $C_{MA}(v_i, x) with |N(x_l)| = 3$; (c) SA $C_{SA}(v_i, x)$ with $\sigma = 0.1$; (d) SA $C_{SA}(v_i, x)$ with $\sigma = 1$; (e) QP assignment $C_{QP}(v_i, x) with |N(x_l)| = 3$.

- **Multiple Assignment**. Multiple Assignment is proposed in [8] which can be modeled with a contribution function as follows,

$$C_{MA}(v_i, x_l) = \begin{cases} 1, & if\ v_k \in N(x_l); \\ 0, & else. \end{cases} \tag{7}$$

where $N(x_l) \subset V$ is x_l's neighboring visual words in visual vocabulary, and $|N(x_l)| < |V|$. The MA assignment contribution function $C_{MA}(v_i, x_l)$ is shown in Fig. 3 (b). The neighbor size $|N(x_k)|$ is a very important parameter for MA. In [3], experiments are carried out to valuate the impact of $|N(x_k)|$ to the retrieval performance, and $|N(x_k)|$ is set as 2 and 3 separately. We also test the performance in our experiments using the same neighbor size for MA.

- **Soft Assignment**. Soft Assignment is proposed in [9] which can be modeled with a contribution function as follows,

$$C_{SA}(v_k, x_l) = \frac{exp(-\frac{1}{2}\frac{\|x_l - v_k\|^2}{\sigma^2})}{\sum_{j=1}^{K} exp(-\frac{1}{2}\frac{\|x_l - v_j\|^2}{\sigma^2})} \tag{8}$$

where σ is the length scale parameter. The MA assignment contribution function $C_{SA}(v_i, x_l)$ is shown in Fig. 3 (c) and (d), with different length scale parameter $\sigma = 0.5$ and $\sigma = 1$ separately.

As we can see from Fig. 3 (b), the Multiple Assignment (MA) strategy assigns x_l to its neighboring visual words $N(x_l)$ equally with $C_{MA}(v_i, x_l) = 1$ in (7). An important assumption is that a discrete visual word is a characteristic representative of an image feature. To the contrary, based on the with the nature of continuous local features, the Soft Assignment strategy assigns x_l to all the visual word with different weights , which is given with a Gaussian-shaped kernel as in (8). One shortage of this strategy is that the contribution function is very sensitive to the kernel size σ, as is shown in Fig. 3 (c) and (d), as also can be seen from the experiments in [9].

To overcome this shortage, we propose a new multiple assignment strategy, with a new contribution function based on the QP problem. Following [3] and [8], [10], we also try to assign a local feature x_l to its neighboring visual words $N(x_l)$ in Visual Vocabulary V. To learn a robust contribution function, we assume the local feature x_l can be linearly constructed by its neighboring visual words $N(x_l)$ as

$$x_l = \sum_{k \in N(x_l)} w_{lk} v_k \tag{9}$$

where w_{lk} are construction weights. We estimate the linear construction weights by minimizing

$$E_l = \|x_l - \sum_{k \in N(x_l)} w_{lk} v_k\|^2 \tag{10}$$

This objective function is similar to the one used in locally linear embedding [11]. To avoid the negative contribution, we further constrain $\sum_{k \in V} w_{lk} = 1$ and $w_{lk} \geq 0$. Usually, the more similar x_l is to v_k, the larger w_{lk} will be. Thus, w_{lk} can be used to measure the similarity degree from x_l to v_k.

It can be easily inferred that

$$\begin{aligned} E_l &= \|x_l - \sum_{k \in N(x_l)} w_{lk} v_k\|^2 \\ &= \|\sum_{k \in N(x_l)} w_{lk}(x_l - v_k)\|^2 \\ &= \sum_{k,j \in N(x_l)} w_{lk}(x_l - v_k)\top(x_l - v_j) w_{lj} \\ &= \sum_{k,j \in N(x_l)} w_{lk} g^l_{kj} w_{lj} \end{aligned} \tag{11}$$

where $g^l_{kj} = (x_l - v_k)\top(x_l - v_j)$ at the local feature x_l. Thus, the reconstruction weights of each data point can be estimated by solving the following L standard quadratic programming

(QP) problems:

$$\min_{w_{lk}} \sum_{k,j \in N(x_l)} w_{lk} g_{kj}^{l} w_{lj} \quad s.t. \sum_{k \in N(x_l)} w_{lk} = 1 \; and \; w_{lk} \geq 0 \tag{12}$$

There exist many standard algorithms to solve these QP problems, as introduced in [12]. To solve them efficiently, we adopt the active set algorithm [12] with a warm start, which usually converges in several iterations. We first compute a relaxed solution without involving nonnegative constraints using the algorithm, as presented in [11]. Then, we compute a warm start by replacing the negative elements of this relaxed solution with 0. Finally, we run the active set algorithm on this warm start.

Then we define our QP based multiple assignment contribution function $C_{QP}(v_k, x_l)$ as following

- **QP Multiple Assignment**.

$$C_{QP}(v_k, x_l) = \begin{cases} w_{lk}, & if \; v_k \in N(x_l); \\ 0, & else. \end{cases} \tag{13}$$

The QP assignment contribution function $C_{QP}(v_i, x_l)$ is shown in Fig. 3 (e).

One important property of the QP Multiple Assignment is that it estimates the contribution to different visual words jointly with the solution of QP problem. Other multiple assignment estimate the weights for different visual words independently, not considering the structure with in the $N(i)$, while QP assignment does.

One important phenomenon we can observe from Fig. 3 (e) is that, although v_2 is further away from x_l than v_3, its weight is much larger than v_2. This means v_2 contribute more than v_3 in reconstruction of x_l, as in (10). Intuitively, if v_3 is more close to x_l than v_2, it should contribute more than v_3. However, if we consider v_2, v_3 and v_4 jointly, and analyze the graph structure of v_2, v_3, v_4 and x_l, we can see that x_l lies on the line connecting v_2 and v_4, making it easy to be reconstructed by weighted averaging of v_2 and v_4, without v_3. In fact, we can see that to reconstruct x_l, v_2 is a complementary to v_4, while v_3 is a redundancy to v_4, since v_3 lies so close to v_4. We can find this relationship by using the QP assignment strategy, which give a characterization of the local feature x_l from a view of joint distribution of the visual words and their relationship to x_l. In contrary, the traditional soft multiple assignment strategy, computing the pairwise similarity between local feature x_l and visual words v_2, v_3 independently, neglects the influence of v_4. The readers should note that, a high pairwise similarity between visual word v_3 and local feature x_l does mean a large weight in QP strategy, which is determined by the local contextual neighborhood structure of the visual word set.

Compared with traditional soft multiple assignment, our QP assignment strategy provides an insight into the graph structure among the local features and the neighboring visual words by learning the similarity using context information, instead of only consider the pairwise relationship. We find it quite similar to the Contextual Dissimilarity Measure (CDM) algorithm proposed by Jegou Herve et al. [3], which assumes that the dissimilarity between a query image and database image should be different with different context (the other neighboring images). In fact, a common idea shared by QP assignment and CDM is that, the similarity between two objects should consider the neighborhood distribution, while pairwise similarity doesn't. However, there are two major differences:

1) The objective in this paper is to learn the local neighborhood similarity between a local feature x_l and visual words v_k for the bag-of-based based representation of an medical image, while theirs is to learn the similarity between a query and a database object in image retrieval task.
2) We proposed a new and direct way to solve the contextual similarity, which solve the weights of neighboring visual words in reconstruction of local features, while their measure takes into account the local distribution of the vectors and iteratively estimates distance update terms in the spirit of Sinkhorn's scaling algorithm, thereby modifying the neighborhood structure [3].

Recently, some contextual similarity/dissimilarity learning algorithm have been proposed for image/shape retrieval tasks [3], [24], [25]. But there are still no such kind of contextual local feature assignment methods. Our QP assignment provides a novel view for this task.

IV. Boosted Visual Word Weighting

An important procedure in bag-of-feature based image retrieval is to weight the frequency vectors as (2) according to its discriminant ability. Here, we propose a novel weighting strategy using boosting method in a supervised classification scene. We first try to analysis the discriminating power analysis of each visual word v_k, according to the k-th bin in H, and then select the most powerful visual words and the weights in a iterative way. Each visual words will be considered as a weak classifier and the final strong classifier will be learned using BoostMap framework [13], [14].

A. Visual Word's Discriminating Power Analysis

Visual word frequency h_k^X represents a close relationship between the visual word v_k and the image X, which contains this visual word. It is observed that if high-frequency visual word are spread widely over a large number of images, we may not retrieve the relevant images from the whole collection.

We analyze each visual word v_k's discriminating power by using individual bin h_k in H as a similarity function for classification of images in a supervised scene. Given a specific visual word v_k in the visual vocabulary V, we define a similarity subfunction for a pair of images X and Y using the concept of Histogram Intersection Kernel:

$$s_k^{HIK}(H^X, H^Y) = min(h_k^X, h_k^Y) \tag{14}$$

Now let's consider the relationship between the sub-HIK functions $s_k^{HIK}(H^X, H^Y)$ and the final HIK function $S^{HIK}(F^X, F^Y)$ between two images X and Y which is given

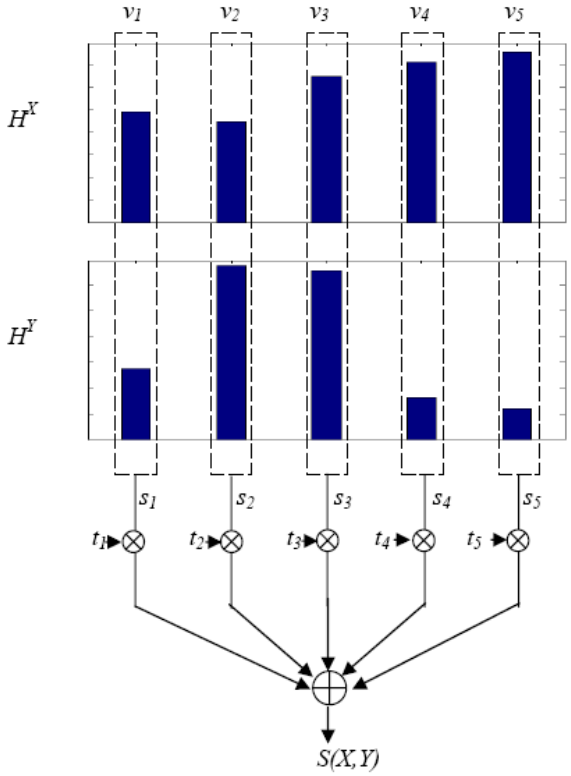


Fig. 4. Combining sub-similarities with visual weighting factors for comparison between histograms between a pair of image X and Y.

in (4). We can prove that, $S^{HIK}(F^X, F^Y)$ equals to the sum of weighted similarity sub-functions $s_k^{HIK}(H^X, H^Y)$ as

$$S^{HIK}(F^X, F^Y) = \sum_{k=1}^{K} t_k \cdot s_k^{HIK}(H^X, H^Y) \qquad (15)$$

Proof:

$$\begin{aligned} S^{HIK}(F^X, F^Y) &= \sum_{k=1}^{K} min(f_k^X, f_k^Y) \\ &= \sum_{k=1}^{K} min(t_k \cdot h_k^X, t_k \cdot h_k^Y) \\ &= \sum_{k=1}^{K} t_k \cdot min(h_k^X, h_k^Y) \\ &= \sum_{k=1}^{K} t_k \cdot s_k^{HIK}(H^X, H^Y) \end{aligned} \qquad (16)$$

■

This visual word weighting strategy combines the sub-similarity functions s_k with their weighting factors t_k in a linear way, as is shown in Fig. 4.

Some other works does not apply Histogram Intersection Kernel as a similarity measure [2]. According to Puzicha [16], Jeffrey divergence or Jensen-Shannon divergence(JSD) is also suitable to be a similarity measure to compute similarity between query image X and the images Y visual words histogram in database:

$$S^{JSD}(H^X, H^Y) = \sum_{k=1}^{K} [h_k^X log \frac{h_k^X}{h_k^X + h_k^Y} + h_k^Y log \frac{h_k^Y}{h_k^Y + h_k^X}] \qquad (17)$$

where H^X and H^Y are the histograms to be compared and h_k^X is the k-th bin of H^X. We can also define the sub-JSD-function for k-th visual word v_k as

$$s_k^{JSD}(H^X, H^Y) = h_k^X log \frac{h_k^X}{h_k^X + h_k^Y} + h_k^Y log \frac{h_k^Y}{h_k^Y + h_k^X} \qquad (18)$$

Here we do not weight the frequency vector directly as $f_k^X = t_k \cdot h_k^X$. Instead, we weight the sub similarity functions following (16).

$$S^{JSD}(H^X, H^Y) = \sum_{k=1}^{K} t_k \cdot s_k^{JSD}(H^X, H^Y) \qquad (19)$$

In some other works like [1], instead of similarity between a pair of histograms, distance can also be used for histogram comparison. Several distance measures can be considered. Uri Avni etc. [1] have found the using the L_1 norm distance $D^{L_1}(H^X, H^Y)$ between the visual word histograms of the two images X and Y yields the good results in this database,

$$D_{L_1}(H^X, H^Y) = \sum_{k=1}^{K} |h_k^X - h_k^Y| \qquad (20)$$

similarly, we present $D_{L_1}(H^X, H^Y)$ as linear combination of sub-distance function $d_k^{L_1}(H^X, H^Y) = |h_k^X - h_k^Y|$, which is the distance referring the k-th visual word v_k, as

$$D^{L_1}(H^X, H^Y) = \sum_{k=1}^{K} t_k \cdot d_k^{L_1}(H^X, H^Y) \qquad (21)$$

To this point, the learning of weighting factor is an open general problem for each visual word v_k's sub-similarity functions s_k. In fact, the visual word's discriminating power is implemented by the discriminating power of its sub-similarity function. So we can learn the weighting factor of sub-similarity function as the t_k.

B. Visual Word Weighting Scheme

Without ambiguity, for convenience, we denote the similarity $S(H^X, H^A)$ and distance $D(H^X, H^A)$ between a pair of histograms of images X and A as $S(X, A)$, and the sub-similarity $s_k(H^X, H^A)$ as $s_k(X, A)$, the sub-distance $d_k(H^X, H^A)$ as $d_k(X, A)$. And the general similarity and distance function between X and Y is rewritten as

$$\begin{aligned} S(X, Y) &= \sum_{k=1}^{K} t_k \cdot s_k(X, Y) \\ D(X, Y) &= \sum_{k=1}^{K} t_k \cdot d_k(X, Y) \end{aligned} \qquad (22)$$

Now we describe the proposed approach to learning the weighting factors for visual words t_k in (22). Our learning scheme is similar to the BoostMap approach [13], [14]. However, there are two major differences:

1) The objective in this paper is to learn the weighting factors of visual words , while theirs is to learn an embedding in a normed-space for approximating a computational expensive distance function.
2) As a unsupervised methods not using the objects class information, in BoostMap, an embedding is proximity-preserving when it perfectly preserves proximity relations between triples of objects. The goal of BoostMap is to construct an embedding that is as close to being proximity preserving as possible. However, in our methods, we learn a similarity function that can preserves the inter-class and intra-class similarity as possible, which guarantees correctly the nearest neighbor classification of images in image triples, with the triplet labels based on the class information.

Inspired by BoostMap [13], [14], to learn the weights t_k for (22) generalized from (16) and (19), we consider a triplet of images (X, A, B) from the image database, and l_X is the class label of X. A or B are from the same class as X: $l_X = l_B$, or $l_X = l_A$. At the same time, we add a extra constraint for the selection of triplet (X, A, B): A and B are from different classes, as $l_A \neq l_B$. With this constraint, one of the following two cases must be true:

- $l_X = l_A$ while $l_X \neq l_B$;
- $l_X \neq l_A$ while $l_X = l_B$;

Our training set Ω consists of I triplet of images: $\Omega = ((X_1, A_1, B_1), \cdots, (X_I, A_I, B_I))$. For each such triplet $(X_i, A_i, B_i), i = 1, \cdots, I$, a label is attached: If (X_i, A_i) belongs to the same class, while (X_i, B_i) belongs to different classes, then the corresponding label of the triplet is 1, otherwise, the label is -1. We denote the label of (X_i, A_i, B_i) by y_i:

$$y_i = \begin{cases} +1, \ if \ l_{X_i} = l_{A_i}, \ l_{X_i} \neq l_{B_i}; \\ -1, \ if \ l_{X_i} = l_{B_i}, \ l_{X_i} \neq l_{A_i}. \end{cases} \tag{23}$$

with $S(X, A)$ and $S(X, B)$ are visual histogram similarity functions between (X, A) and (X, B), we can guess whether X is in the same class with A or with B by checking if $S(X, A)$ is larger than $S(X, B)$ or not. With the similarities $S(X_i, A_i)$ and $S(X_i, B_i)$ between visual words' histograms of H^{X_i}, H^{A_i} and H^{B_i}, an ordinal function is defined as

$$\Theta_i = \begin{cases} +1, & S(X_i, A_i) > S(X_i, B_i); \\ -1, & S(X_i, A_i) \leq S(X_i, B_i). \end{cases} \tag{24}$$

Θ_i has two discrete output values: $+1$ and -1, which is a quantized version of a continuous function $\widetilde{\Theta}_i$, defines as:

$$\widetilde{\Theta}_i = S(X_i, A_i) - S(X_i, B_i) \tag{25}$$

when distance is used instead of similarity,

$$\widetilde{\Theta}_i = D(X_i, B_i) - D(X_i, A_i) \tag{26}$$

In spaces where distances can take any value within some range of real numbers, it is typically unusual for an medical image to have the exact same histogram distance to two database images. Consequently, we consider the task of estimating Θ_i to be a binary classification task as y_i.

$$y_i \approx \Theta_i = sign(\widetilde{\Theta}_i) \tag{27}$$

Now, we consider the situation with each individual visual word. With $s_k(X, A)$ and $s_k(X, B)$ are sub-similarity functions between (X, A) and (X, B), we also can guess whether X is in the same class with A or with B by checking if $s_k(X, A)$ is larger than $s_k(X, B)$ or not. More formally, for every sub-similarity function s_k, we define a classifier $\widetilde{\theta}_k(X_i, A_i, B_i)$ as:

$$\widetilde{\theta}_k(X_i, A_i, B_i) = s_k(X_i, A_i) - s_k(X_i, B_i) \tag{28}$$

then $\theta_k(X_i, A_i, B_i) = sign(\widetilde{\theta}_k(X_i, A_i, B_i))$ is an estimate of y_i. $\widetilde{\theta}_k(X_i, A_i, B_i)$ can be regarded as a weak real-valued classifier, and we will employ the real Adaboost learning algorithm [17] to approximate a strong classifier by a number of weak classifiers.

When using distance functions $D(X, Y)$, similar to (28), we can define a weak classifier for $((X_i, A_i, B_i))$ to guess which is closer to X_i between A_i and B_i, as

$$\widetilde{\theta}_k(X_i, A_i, B_i) = d_k(X_i, B_i) - d_k(X_i, A_i) \tag{29}$$

With Adaboost Framework, our goal is to learn a strong classifier $\widetilde{\Theta}_i = \widetilde{\Theta}(X_i, A_i, B_i)$, which is weighted linear combination of binary classifiers $\widetilde{\theta}_k(X_i, A_i, B_i)$ which is learned using only individual visual words,

$$\widetilde{\Theta}(X_i, A_i, B_i) = \sum_{k=1}^{K} t_k \widetilde{\theta}_k(X_i, A_i, B_i) \tag{30}$$

The influence of each base classifier with each visual words's sub-similarity in the ensemble is governed by the weighting vector $\mathbf{t} = [t_1 \ t_2 \ \cdots \ t_K]$ in (30), which will be learned by employing Real-Adaboost algorithm. More specifically, we fit the real-valued classifier $\widetilde{\Theta}(X_i, A_i, B_i)$ by the additive composition of its weak classifiers $\widetilde{\theta}_k(X_i, A_i, B_i) = s_k(X_i, B_i) - s_k(X_i, A_i)$ as

$$\begin{aligned} \widetilde{\Theta}_i &= S(X_i, A_i) - S(X_i, B_i) \\ &= \sum_{k=1}^{K} t_k s_k(X_i, A_i) - \sum_{k=1}^{K} t_k s_k(X_i, B_i) \\ &= \sum_{k=1}^{K} t_k [s_k(X_i, A_i) - s_k(X_i, B_i)] \\ &= \sum_{k=1}^{K} t_k \widetilde{\theta}_k(X_i, A_i, B_i) \end{aligned} \tag{31}$$

where the coefficients $T = \{t_k\}, k = 1, 2, \cdots, K$ are the outputs of the Adaboost learning procedure. They define a histogram similarity function S , which weights the k-th visual word with t_k .

The design aim to classify the triples (X_i, A_i, B_i) as correctly as possible, so that $sign(\Theta(X_i, A_i, B_i)) = y_i$. This is exactly the problem that boosting methods are designed to solve.

Now we introduce the details of the learning procedure. It is known that the Adaboost algorithm [17] takes a number of rounds. At each round, a weak learner is trained with the weighted version of the original training samples. After that, each training sample is re-weighted according to the

confidence of being correctly classified by the weak learner. Briefly, the samples that are falsely classified by the weak classifier will be assigned to larger weights after this iteration and vice versa. The procedure continues until the testing error no longer decreases or some pre-determined number of rounds is reached.

To use AdaBoost algorithm, we should first specify the input as follows:

- Training triple data set with class labels $\Pi = \{(o_i, y_i)\}, i = 1, \cdots, I, o_i \in \Omega, y_i \in \{+1, -1\}$. We denote $o_i = (X_i, A_i, B_i)$ as a triple in the training set. Ω is the set of triple:

$$\begin{aligned}\Omega =&\{(X_i, A_i, B_i)|(l_{X_i} = l_{A_i}\ and\ l_{X_i} \neq l_{B_i})\\ &or\ (l_{X_i} = l_{B_i}\ and\ l_{X_i} \neq l_{A_i})\}\end{aligned} \tag{32}$$

- weak classifiers set $\Phi = \{\widetilde{\theta}_k(o) = \widetilde{\theta}_k(X, A, B)\}, k = 1, \cdots, K$ as follows

$$\widetilde{\theta}_k(X, A, B) = s_k(X, A) - s_k(X, B) \tag{33}$$

where K is the dimension of the visual word histogram, also the number of sub-similarity functions, and $s_k(X, A)$ is the sub-similarity in the k-th bin of the visual word histogram between objects X and A, defined in (14) or (18).

Then, at each Adaboost learning round r, a weight $D_r(i)$ will be assigned to each triplet o_i, satisfying $\sum_{i=1}^{I} D_r(i) = 1$. In the beginning, all weights are initialized equally: $D_r(i) = \frac{1}{M}$.

In the beginning, the composition coefficients of $T = t_k, k = 1, \cdots, K$, of the weak classifiers $\widetilde{\theta}_k$ are set to 0. At the r-th round, we try to select a weak classifier $\widetilde{\theta}_{k^*}$ from the pool $\Phi = \{\widetilde{\theta}_k\}, k = 1, \cdots, K$ that best minimizes the overall empirical training error. To quantify this notion, a error measure ε_k was proposed [17]:

$$\varepsilon_k = \sum_{i=1}^{I} D_t(i) \left[y_i \neq \widetilde{\theta}_k(o_i) \right] \tag{34}$$

In (34), $k = 1, 2, ..., K$ and i ranges from $1, 2, ..., I$, where K is the dimension of the visual histogram, i.e., the number of weak classifier and I is the number of triplets (training data). Generally speaking, ε_k represent the benefit of adding the k-th weak classifier $\widetilde{\theta}_k(o)$ to the current classifier composition in minimizing the empirical training error. The smaller the ε_k , the larger the benefit. When $\varepsilon_k < 0.5$, adding $\widetilde{\theta}_k(o)$ actually deteriorates the classification performance. Therefore, at the r-th iteration, we choose the weak classifier $k^* : k^*$ that minimizes ε_k:

$$\widetilde{\theta}_{k^*} = \underset{\widetilde{\theta}_k \in \Phi}{argmin}\ \varepsilon_k \tag{35}$$

The overall Adaboost learning procedure is summarized in Algorithm 1. This step-sequence is interleaved until $\varepsilon_k > 0.5$.

Algorithm 1 The Adaboost algorithm for learning the visual word weighting factors.

Require: $\Pi = \{(o_i, y_i)\}, i = 1, \cdots, I$, the training data set;
Require: $\Phi = \{\widetilde{\theta}_k(o)\}, k = 1, \cdots, K$, the weak classifiers set.
Initialize training weights: $D_1(i) = \frac{1}{I}$;
Initialize classifier weights: $t_k = 0$;
for $r = 1, \cdots, R$ **do**
Compute the error ε_k for each weak classifier with respect to the distribution $D_r(i)$ using (34);
Find the classifier $\widetilde{\theta}_{k^*}$ that minimizes the error for round r using (35);
if $\varepsilon_k \geq 0.5$ **then**
Stop;
end if
Update weight t_{k^*} for $\widetilde{\theta}_{k^*}$:

$$t_{k^*} = t_{k^*} + \frac{1}{2} ln \frac{1 - \varepsilon_{k^*}}{\varepsilon_{k^*}} \tag{36}$$

where ε_{k^*} is the weighted error rate of classifier $\widetilde{\theta}_{k^*}$.
Re-set the weights associated with the training samples o_i:

$$D_{t+1}(i) = D_t(i) \frac{exp(-t_{k^*} \widetilde{\theta}_{k^*}(o_i) y_i)}{Z_r(\widetilde{\theta}_{k^*}, t_{k^*})} \tag{37}$$

end for
Output the visual word weighting factors: $\{t_k\}, k = 1, \cdots, K$.

V. Experiments

In this section, we test the proposed techniques through extensive experiments on medical image retrieval. The experimental results suggest that the proposed algorithms have superior performance in general and are especially suited for bag-of-feature based medical image retrieval tasks.

A. Experiment I: X-Ray Image Retrieval on ImageCLEFmed dataset

In this section we evaluate the proposed methods using the ImageCLEFmed 2007 and 2008 datasets. For comparison, the first experimental setup is exactly the same to the one conducted in [1], with similar description. We first investigate the sensitivity to various parameters that define the system, using ImageCLEFmed 2007 dataset. We then show classification and retrieval experiments on large radiograph archives using ImageCLEFmed 2008 dataset.

1) Experiment Setup on ImageCLEFmed 2007 Dataset: A database of 12,000 categorized radiographs, which is the basis for the ImageClef 2007 medical image classification competition [18], is used in the first experiment. A set of 11,000 images are used for training, and 1000 serve for testing. There are 116 different categories within the archive, differing in either the examined region, the image orientation with

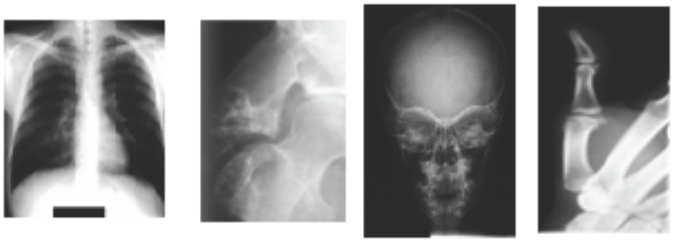
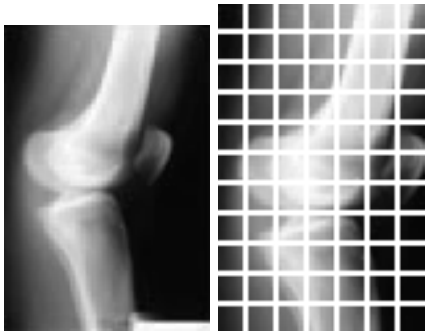


Fig. 5. Sample images from ImageClef medical annotation challenge



(a) (b)

Fig. 6. Representing a ImageCLEFmed image (a) as a bag of collection image patches (b).

respect to the body or the biological system under evaluation. Several of these images are presented in Fig. 5.

We start by representing an input image as a collection of small patches, as is shown in Fig. 6. In this step images are sampled with a dense grid. We extract a patch around every pixel, using a patch size of 9×9 pixels. Patches along the border of the image are considered as noise and are ignored. The intensity values within a patch are normalized to have zero mean and unit variance. Patches that have a single intensity value of black are ignored. To reduce both the algorithms computational complexity and the level of noise, we apply a principal component analysis procedure (PCA) and reduce the data dimensionality from 81 to 7.

The next step of our system is to learn a dictionary of visual words based on a representative set on images. To accelerate the learning process we randomly take a subset of all the patches. The main step in the dictionary building procedure is clustering the patches, using the k-means algorithm, to form a small-size dictionary of visual words. Using the generated dictionary, each image X is represented as a histogram of visual words using NN assignment in [1]. Here, we compare our QP patch assignment against NN assignment in [1]. The MA and SA assignments are also valuated in the experiments.

A nearest neighbor classifier is a reasonable choice and is used in [1], which is also a choice in our experiments for comparison. Given a query image the retrieval is based on finding the nearest image in the labeled training set. Following [1], we use the L_1 norm distance to for comparing the word histograms of the two images. We tested our boosted visual word weighting factor t_k with the L_1 norm distance

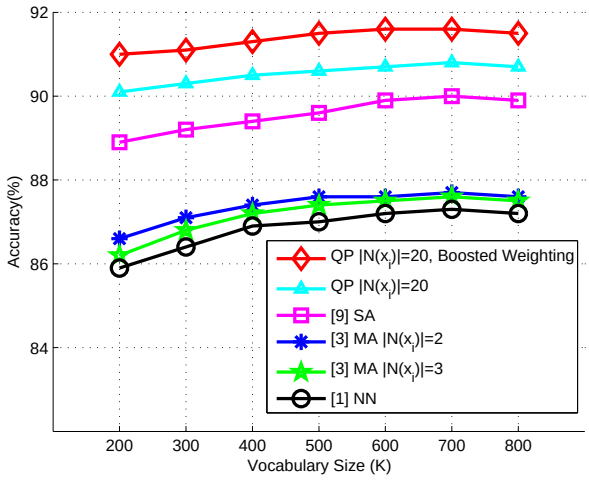
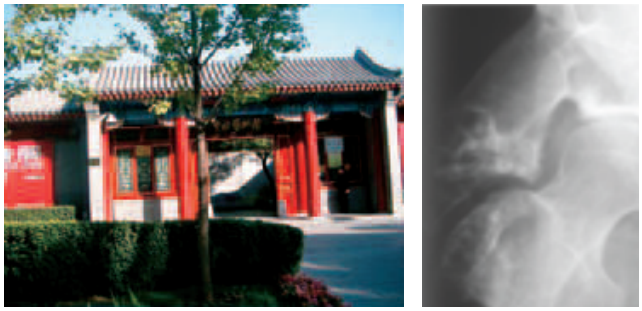


Fig. 7. Effect of vocabulary size, for K-NN classifiers with different assignment methods.

as $D^{L1}_{Boosted} = \sum_k t_k \cdot d^{L1}_k$, and compare it to the original L_1 norm distance $D^{L1} = \sum_k d^{L1}_k$. The classification is based on a set of manually categorized images. We run 20 cross-validation experiments trained on 10000 images and verified on 1000 randomly drawn test images.

2) Parameters Validation on ImageCLEFmed 2007 Dataset: There are two main free parameters for the proposed approach, the number of visual words K in the visual vocabulary, the neighbor size $N(x)$ for QP assignment. In order to show the proposed approach is applicable in a reasonable range for parameters, we test the performance of the proposed approach on a range of parameter values.

- **Validating the Visual Vocabulary Size** K. In evaluating vocabulary size, we tune vocabulary sizes K in our experiments. The vocabulary sizes we consider are $\{200, 300, 400, 500, 600, 700, 800\}$. The results for all types of local assignment evaluated for various vocabulary sizes are given in Fig. 7.
  As illustrated in Fig. 7, increasing the vocabulary size increases the classification performance and the performance of the four ambiguity types seems to converge up to 700 words. Based on the these experiments, a dictionary size of 700 visual words was selected.
  In Fig. 7, the vocabulary sizes are relatively small. The largest vocabulary in Fig. 7 has 800 elements and only 0.08 percent of all features used for clustering are comprised. The behavior of relatively small vocabularies may not be identical to relatively large vocabularies. With vocabulary sizes that are relatively large compared to the total number of training image features, different assignment type performance may diverge different again. Different from bag-of-features based natural image retrieval system, whose typical vocabularies range in the thousands of visual words [8], the vocabulary here seems much too small. In fact, for typical natural image retrieval, the image's content is much more complex and



(a) (b)

Fig. 8. Illumination of natural image (a) and medical image (b).

we need a large size visual vocabulary, while in medical image retrieval tasks, the contents of medical images are somehow simple and thus we don't need so large vocabulary. We illuminated this difference between natural images and medical images in Fig. 8 for the readers' easy understanding. In (a), there are house, tree, road and grass etc, so we need more visual words to represent this image, while in (b), there a single bone in it, thus only a small vocabulary is needed. More evidence can be found in reference [19]. Avni et al. obtained the best performance in ImageClef 2008 medical image retrieval challenge using a small visual vocabulary of 700 visual words.

We continue the analysis comparing our method with the different representative methods of each of the assignment strategies according to the concepts related in local feature assignment. In order to perform this comparison, we use the MA with $|N(x)| = 2$ and 3, uncertainty SA with $\sigma = 200$ which shows a much greater power as a soft assignment when classifiers are compared to kernel and codebook [9]. The kernel size σ for SA is chosen via 10-fold cross validation, following [9]. For each test, we recompute the k-NN classification ($k = 3$) accuracy between QP with $|N(x)| = 10$ and each assignment methods. From Fig. 7, it must be noted that QP assignment with proper $|N(x)|$ performs generally better than most of the commonly used assignment strategy. Taking into account the number of times a method achieves the best mean performance, QP assignment is the first choice for all the vocabulary size, followed SA, MA with $|N(x)| = 2$, and MA with $|N(x)| = 3$. The MA of local image patch descriptors to visual words slightly improves the accuracy of the search at the cost of an increased search time, due to the impact of the method on the visual word vector sparsity. For instance, for $V = 30,000$ visual words, the number of multiplications performed is seven times higher for MA 3 than for the simple NN assignment. It should be used for applications requiring high accuracy. Note that the number of assignments $|N(x)|$ must be small, e.g., 2 or 3, as we have observed that the accuracy decreases for larger values. Regarding the worst performances, traditional NN assignment is the last option. Observe that QP and SA assignment are never the last choice.

To validate if the boosted visual word weighting factor improves the L_1 norm distance, we also depicts the averaged classification accuracy of the weighted version of QP assignment methods using different K in Fig. 7. The trends of the curves are similar to those in original QP assignment. That is, the averaged classification points of boosted weighting methods show a tendency to increase as the number of the vocabulary size K. Among them, the best three averaged classification points 91.60% are reached at the $K = 700$ vocabulary using boosted weighting strategy.

- **Validating the Neighbor Size** $|N(x)|$. For evaluation of the Neighbor Size $|N(x)|$, a performance curve for QP assignment is plotted showing the k-NN classification versus $|N(x)|$, which is the number of the most nearest visual words $v_k \in V$ of a visual local feature x. The curves on the ImageCLEFmed 2007 Dataset using image patch as local feature are shown in Fig. 9. From the results, it is clear that larger Neighbor Size $|N(x)|$, especially when $|N(x)| = 1 \sim 8$, outperform smaller ones. The recognition rate achieved by all neighboring sizes when $|N(x)| \geq 20$ is relatively stable for a wide range. The best performance is obtained by setting $|N(x)|$ to 20-29. Because a shorter neighbors list $N(x)$ results in an overall speedup across all algorithms that we consider, we fix the neighboring size $|N(x)|$ to 20 visual candidates in all experiments.

  It is very interesting to note that, for large $|N(x)|$, the QP assignment algorithm performs similar to a smaller one. This phenomenon is because in the QP assignment, the contribution function of a visual feature x to $v_k \in N(x)$ is the weights used in the linear reconstruction of the x from v_ks, among which only a few ones are valid with $C^{QP}(v_k, x) \neq 0$, although it is learned from a large set of neighbors. To illuminate this, we define a minimum subset of $N(x)$ for a local feature (image patch here) x, whose numbers' contribution function $C^{QP}(v_k, x) \neq 0$, call as critical neighbors (CN). The CNs of a local feature is a minimum set to reconstitute the x, while the visual word $v_k \in N(x)$ and $v_k \notin CN(x)$ is not necessary. Here, we give two example image patches x_1 and x_2 in the first row of Fig. 10(b) with a vocabulary V of size 10 as Fig. 10(a). their QP contributions $C^{QP}(v_i, x)$ and the CN in $N(x)$ with different neighboring size $N(x)$ is show blow the pathes. As we can see from Fig. 10, with the increasing $N(x)$, the $C^{QP}(v_i, x)$ does not change continuously. One interesting fact is, with in a interval of $N(x)$, the $C^{QP}(v_i, x)$s are identical. For example, $N(x) = 5 \sim 7$ and $N(x) = 8 \sim 10$ for x_1, while $N(x) = 4 \sim 7$ and $N(x) = 8 \sim 10$ for x_1. This means that the $C^{QP}(v_i, x)$s do not benefit from enlarging the $N(x)$ in the interval, while spend more computation time, since the time complex is $O(|N(x)|^3)$. This can also be observed from the CNs of x_1 and x_2, which is also identical for the $N(x)$ in some interval. For example, when we varying the $N(x)$ in $4 \sim 7$ for x_2, the CN do

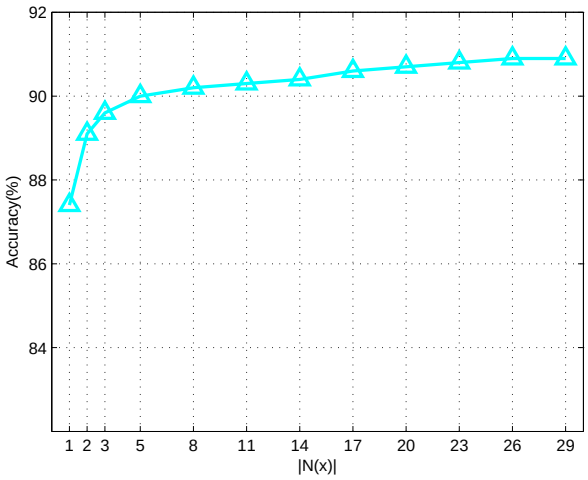


Fig. 9. We consider the effect of varying the neighboring size $N(x)$ on our evaluation methodology. We plot the mean k-NN classification accuracy (%) obtained by QP assignment (using 20 cross-validated reference metrics, as described in Section V-A1) for several different neighboring sizes $N(x)$ on the ImageCLEFmed 2007.

not change at all which means the additional 3 visual word have no contribution for x_2 in QP reconstruction, and sequently, do not effect the assignment of x. So it's not necessary to use a large $N(x)$.

Another interesting phenomena is that, for some patch (like x_1), patch neighbor that is further away contributes more to the QP then a neighbor that is closer, e.g. the difference between step 4 and step 5 for image patch x_1. As we explained in section III, by saying 'a close neighboring visual word', we implicitly measure the similarity between a local feature and a visual word using a pairwise distance, which neglects the context information in the visual words neighborhood. When we utilize the contextual similarity to model the contribution function using QP assignment, we may have a quite different results. One the other hand, when we enlarge the radius of the neighborhood $|N(x)|$, adding new visual words to $N(x)$, the structure of the local neighborhood might change. For example, when $|N(x)$ is enlarged from 4 to 5, the neighborhood of x_1 is changed, and x_1 find a new neighbor which is not included by the previous $N(x_1)$–the first visual word, and more importantly, with it, x_1 can reconstruct itself better than using the last one collaborating with other visual words. In this situation, QP assignment move the weight from last one to the first visual word, although the last one seems quiet similar to x_1. However, we must note that, the neighborhood structure keep stable in most cases (e.g. for x_1, when $|N(x)|$ is enlarged from 5 to 6, and 6 to 7, etc.). The main task, is to find a the minimum neighborhood size $|N(x)|$ with the same neighborhood structure (CN), to neglect the redundant visual words.

- **Validating Different Multiple Assignment Strategies for Various Vocabularies**. We can use different data set for clustering to generate different various vocabularies.

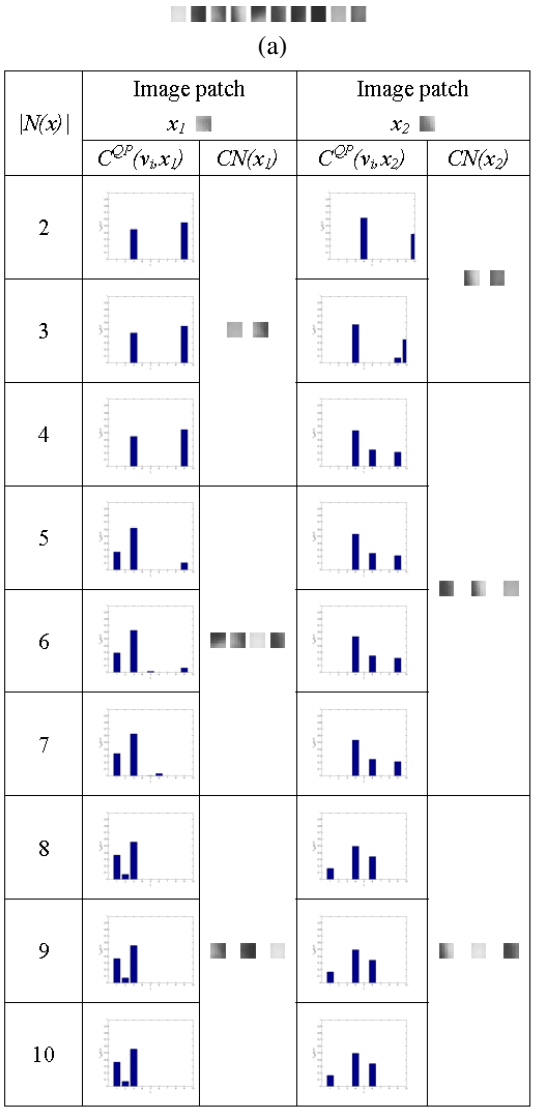


Fig. 10. Illumination of image patches x_1 and x_2's QP Contributions $C^{QP}(v_i, x)$ to the visual words with varying neighboring size $|N(x)|$

The data set used for the clustering may have an impact on the accuracy. In this section, we also test multiple assignment QP vs. NN, MA and SA for various vocabularies on set ImageCELF 2007. The visual vocabularies are generated by clustering two medical image sets: ImageCELF 2007 itself, ImageCELF 2008 [1], and a texture data set CUReT [23]. The visual vocabulary sizes are all set to be $K = 700$. The classification results are shown in Tables I. For ImageCELF 2007 data set, we compare in column "clustering data set" k-means clustering on two uncorrelated data set (ImageCELF 2008 and CUReT) with k-means clustering on the evaluation data set itself (ImageCELF 2007). In all the cases, the results are improved by generating the visual vocabulary with a subset of the data set on which the experiments are performed. Moreover, compared to using a texture dataset to generate the visual vocabulary, using another medical

TABLE I
CLASSIFICATION RATE (%) OF MULTIPLE ASSIGNMENT QP VS NN, MA AND SA FOR VARIOUS VOCABULARIES ON IMAGECELF 2007.

Assignment	Clustering Data Sets		
	ImageCELF 2007	ImageCELF 2008	CUReT
[1] NN	87.3000	85.3000	81.7000
[8] MA	87.7000	85.6000	82.1000
[9] SA	90.0000	87.6000	84.0000
QP	90.8000	88.4000	84.8000

TABLE II
COMPARISON OF DIFFERENT VISUAL WORDS WEIGHTING METHODS ON IMAGECLEFMED 2007: BOOSTED WEIGHTING VS. TF, TF.IDF AND [27] WEIGHTS.

Assignment	Weighting			
	[15] tf	[15] tf.idf	[27] Weight	Boosted
[1] NN	87.3000	87.3000	88.1000	**88.3000**
[8] MA	87.7000	87.6000	88.5000	**88.7000**
[9] SA	90.0000	89.8000	**91.0000**	90.7000
QP	90.8000	90.6000	91.4000	**91.6000**

image set (ImageCELF 2008) seems a better choice for a medical image task. Base on these results, we use the visual vocabulary generated by clustering local feature dataset of ImageCELF 2007.

3) Comparison of Weighting Methods to Different Assignment Strategies on ImageCLEFmed 2007 Dataset: Using the selected parameters, we carry out expriments on ImageCLEFmed 2007 Dataset to compare our proposed QP multiple assignment and boosted visual words weighting strategies against other state-of-the-art. According to the reviewer's advices, we consider the following visual word weighting strategies as comparision of boosted weighting:

- **tf**. The most popular term frequency representation only adopts the raw term frequency (tf) in the document [15].
- **tf.idf**. A conventional inverse collection frequency factor (idf) [15], as introduced in 3. Here we use idf to weight tf, resulting tf.idf as a bag level representation.
- **[27] Weights**. A vocabulary weights learning method proposed by [27], which is very relevant to this paper.

The results are summarized in Table II. These results show that for all assignment strategies, except for the SA, for which [27] method has the best result, the developed method, boosted weighting algorithm, have larger discriminative power compared against tf and tf.idf. Generally, tf and tf.idf are comparable, tf is relatively better. In most cases, for different weighting strategies, the classification performances of unweighed tf and the idf weighted tf—tf.idf with different assignment methods are generally inferior to those in weighted with using boosted or [27] methods. Thus, though the visual vocabulary of data in the data set are not large, classification can still benefit from discriminative weighting factors learning algorithms. In 3 out of 4 cases, our boosted algorithm outperform [27] algorithm. However, the classification result by using boosted weighting factor learning is already quite competitive as only the [27] algorithm have better performances when SA is used as assignment strategy. We attribute this result to the nonlinear property of the SA assignment in the local features.

4) Large Scale Medical Image Retrieval on ImageCLEFmed 2008 Dataset: In ImageClef 2008 a large-scale medical image retrieval competition was conducted [19]. A database of over 66,000 images was used with 30 query topics. Each topic is composed of one or more example images and a short textual description in several languages. The objective is to return a ranked set of 1000 images from the complete database, sorted by their relevance to the presented queries. The retrieved results were manually judged for relevance by medical experts.

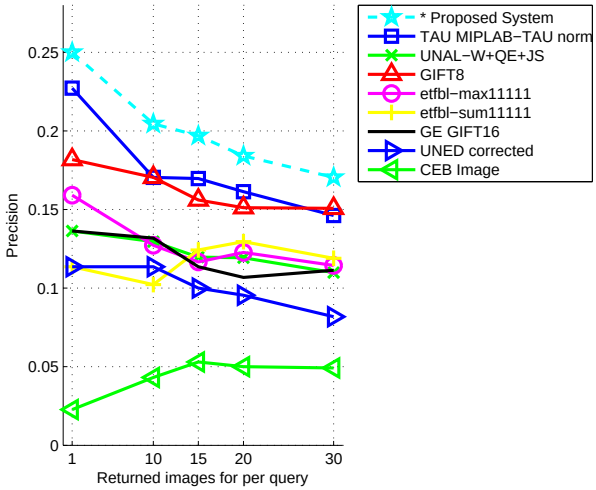
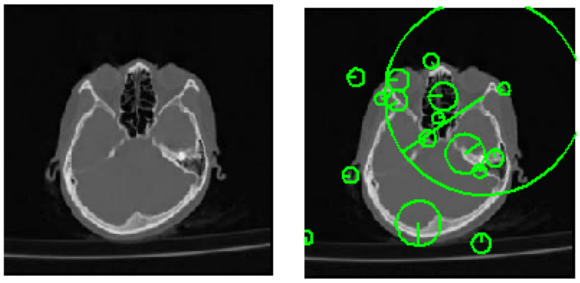


Fig. 11. Precision vs returned images graph of visual retrieval systems on ImageClef 2008 medical database. Average precision shown for first 5, 10, 15, 20 and 30 returned images for each query image. Our proposed system's result marked with dashed lines.

Based on the above experiments, a dictionary size of 700 visual words was selected, where each word contains 7 PCA coefficients. Fig. 11 shows the performances of our proposed retrieval system using QP assignment with $|N(x)| = 20$ and boosted weighting strategy, marked with (*), along with visual retrieval algorithms submitted by additional groups [19]. From Fig. 11, it must be noted that our system performs generally better than all of the other automatic purely visual retrieval systems making it the first choice in ImageCLEFmed 2008 data set. It is also observed that the performance of the proposed system decrees with the addition of more and more returned images, whenever such an decrees is possible.

B. Experiment II: Medical Image Retrieval on 304 CT Set

The second test data set in this paper contains 304 CT images, so we call it 304 CT set. These images are first used in [2], and compose 6 body parts of different people: foot, abdomen, kidney, lung, head, and heart. These images extracted randomly from 6 CT series, and each subset images is sampled from continuous slices in one series. Thus each subset images can be treated as similar in content. Totally, there are 16 such subsets. One example image in 304 CT Images Set is shown in Fig. 12 (a). For simulate the real condition while can compute result conveniently, we randomly add rotation (in $[-60°, 60°]$) and scale (in $[0.5, 2]$) on the images in the experiments.



(a) (b)

Fig. 12. One Image in 304 CT (a) and its SIFT local keypoints (b).

Following [2], we extract SIFT (Scale Invariant Feature Transform) feature at local keypoints [5], [6] as local features of the medical images, as is shown in Fig. 12 (b). This strategy uses a keypoint detector based on the identification of interesting points in the location-scale space. This is implemented efficiently by processing a series of difference-of-Gaussian images. The final stage of this algorithm calculates a rotation invariant descriptor using predefined orientations over a set of blocks. We use SIFT points with the most common parameter configuration: 8 orientations and 4×4 blocks, resulting in a descriptor of 128 dimensions.

The next step is to make visual word vocabulary V. This step usually uses k-means clustering method, and use cluster centers as visual vocabulary term. We take the local SIFT feature descriptors of all the images in training dataset as the input to k-means to get all cluster centers v_k, and derive a vocabulary $V = \{v_k\}, k = 1, \cdots, K$ of $|V| = K$ visual words. We do not spend additional effort to tune the value of K, but simply set it to 2400.

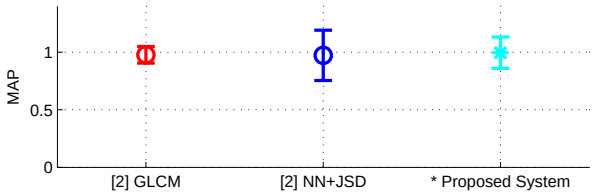
After making visual words, images in database file are transformed into TDM (Term-document matrix), in which an image is represented as a file document vector [2], or as refereed as a histogram in our paper. The file document vector is build using the NN assignment strategy in [2]. Here, we use the QP assignment to build the histogram in our proposed system.

In this group of experiments, we use the Jensen-Shannon divergence $S^{JSD} = \sum_{k=1}^{K} s_k^{JSD}$ as the baseline similarity measure as in [2]. To improve the retrieval performance, we apply our boosted weighting strategy to enhance the original JSD similarity measure, resulting $S_{Boosted}^{JSD} = \sum_{k=1}^{K} t_k s_k^{JSD}$ by learning the weighting factor t_k.

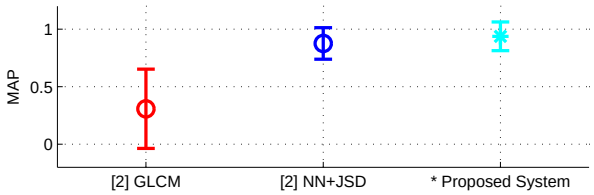
Because the system proposed by [2] used NN assignment and JSD similarity for comparison of histograms, we will refer it as 'NN+JSD' in our experiment results in Fig. 13. Here, we also use widely used co-occurrence feature as baseline to compare different methods' performance, which will be refereed as 'GLCM' in Fig. 13.

In this experiment, MAP (Mean Average Precision) [20] value is used for evaluate performance between different methods. Average Precision (AP) is computed for every query image q as

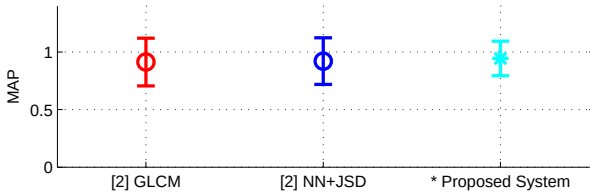
$$AP(q) = \frac{1}{N_R} \sum_{n=1}^{N_R} P_q(R_n) \tag{38}$$



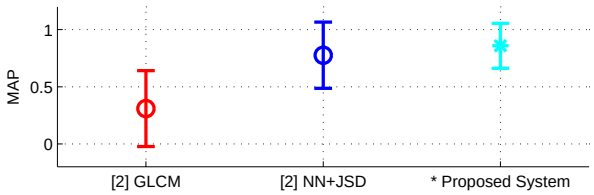
(a) Images without transformation



(b) Add rotation transformation on images



(c) Add scale transformation on images



(d) Add both rotation and scale transformation on images

Fig. 13. Comparison between different transformations on 304 CT set.

where $P_q(R_n)$ is precision recall function value. MAP is the mean of AP:

$$MAP = \frac{1}{|Q|} \sum_{q \in Q} AP(q) \tag{39}$$

For 16 classes in this experiment, randomly choose 3 images in per class as query images, and use the average of this 3 query images' precision as the final precision. We use MAP value as the final value as widely adopted in retrieval experiments. Fig. 13 shows the overall results.

From Fig. 13, we can draw that totally speaking SIFT-based methods (NN+JSD and proposed system) are better than co-occurrence feature method (GLCM), especially when faced on rotation and scale transformation which is usually met in medical image.

In comparison to various bag-of-feature algorithms in [2],

our proposed system employing the QP assignment and the boosted visual word weighting yields the satisfactory performance in general. In particular, the proposed system yields the best performance among 4 dataset with different transformations on the 304 CT set benchmark, as shown in Fig. 13. Since 304 CT set is collected from a real word task [2], the intrinsic structures underlying the local features and visual words is generally unknown although it seems plausible that the features described by SIFT have rather few degrees of freedom. Results achieved by QP assignment suggest that this local feature set may be of linear neighboring structures. Given the fact that our algorithm dominates the performance on all the benchmarks, we conclude that our system with QP assignment and boosted visual weighting would be highly competitive with the existing co-occurrence feature method and traditional bag-of-feature algorithms for some specific tasks.

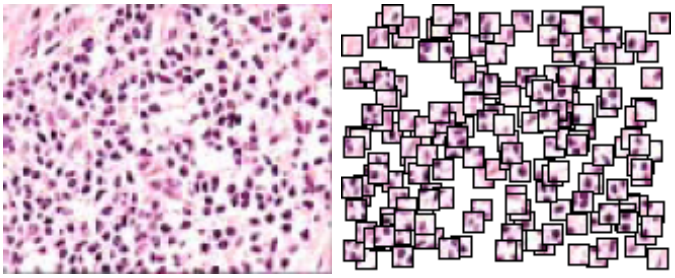
C. Experiment III: Basal-Cell Carcinoma Image Classification

The third medical image dataset has been previously used in an unrelated clinical study to diagnose a special skin cancer known as basal-cell carcinoma [21]. Basal-cell carcinoma is the most common skin disease in white populations and its incidence is growing world wide [22]. Pathologists confirm whether or not this disease is present after a biopsied tissue is evaluated under microscope. The database is composed of 1,502 images annotated by experts into 18 categories. One example image is shown in Fig. 14 (a). Each label corresponds to a histopathology concept which may be found in a basal-cell carcinoma image. An image may have one or several labels, that is to say, different concepts may be recognized within the same image and the other way around.

Following, the work in [21], in this experiment, the two local feature detection strategies with their corresponding feature descriptor have both been evaluated:

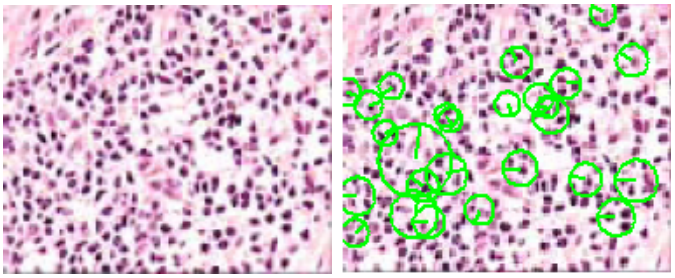
- **Image patches**. The first strategy is dense random sampled image patches. The goal of this strategy is to select points in the image plane randomly and then, define a block of pixels around that coordinate. The size of the block is set to 9×9 pixels, and the descriptor for these blocks is the vector with explicit pixel values in gray scales, as is shown in Fig. 14. This descriptor will be called raw block, but it is also known as, image patch, texton or raw pixel descriptor [23]. As in [1], we call it image patch in this paper.
- **SIFT**. The second strategy is based on Scale-Invariant Feature Transform (SIFT) points [5], [6]. This strategy uses a keypoint detector based on the identification of interesting points in the location-scale space, as is shown in Fig. 15.

The visual vocabulary or codebook V is built using a clustering or vector quantization algorithm. The k-means algorithm is used in this work to find a set of centroids in the local features dataset. An important decision in the construction of the codebook is the selection of its size K, that is, how many codeblocks are needed to represent



(a) (b)

Fig. 14. Representing a Basal-Cell Carcinoma image (a) as a collection of randomly sampled image patches (b).



(a) (b)

Fig. 15. Representing a Basal-Cell Carcinoma image (a) as a collection of key points with SIFT descriptors(b).

image contents. The final image representation is calculated by counting the histogram of each codeblock in the set of local features of an image, using the transitional NN assignment as in [21] or QP assignment in our system. This representation is known as term frequencies (TF) in text applications and is also adopted in this work. Those are the standard image representations, commonly used for image categorization. In addition, the inverse document frequency (IDF) has also been used as weighting scheme to produce a new image representation in [21], which is given in (3). Classifiers used in this experiment are Support Vector Machines (SVM), that receives as input a data representation implicitly defined by ahistogram intersection kernel (4) function as $S^{HIK}(X,Y) = \sum_{k=1}^{K} s_k^{HIK}(X,Y)$, where $s_k^{HIK}(X,Y) = min(f^X k, f_k^Y)$. We improve it with our boosted visual word weighting factor t_k as $S_{Boosted}^{HIK}(X,Y) = \sum_{k=1}^{K} t_k \cdot s_k^{HIK}(X,Y)$, and compare it with the original tf, tf.idf weighting methods in [21], and the weighting methods proposed in [27].

The collection is split into 2 datasets, the first one for training and validation, and the second one for testing. The dataset partition is done using stratified sampling in order to preserve the original distribution of examples in both datasets. This is particularly important due to the high imbalance of classes. In the same way, the performance measures reported in this work are precision, recall and F-measure to evaluate the detection rate of positive examples, since the class imbalance may produce trivial classifiers with high accuracy that do not recognize any positive sample. In addition, since one image can be classified in many classes simultaneously, the classification strategy is based on binary classifiers following the one-against-all rule. Experiments to evaluate different

configurations of the bag of features approach have been done. For each experiment, the regularization parameter of the SVM is controlled using 10-fold cross validation in the training dataset, to guarantee good generalization on the test dataset. Reported results are calculated on the test dataset and averaged over all 18 classes.

- **Evaluation of Codebook Size** K. The first evaluation focuses on the codebook size K. We have tested six different codebook sizes, starting with 50 codeblocks and following with 150, 250, 500, 750 and 1000. Fig. 16 shows a plot for codebook size K vs. F-measure using two different local feature descriptors.
  As can be seen from the Fig. 16, for the methods using image patches as local feature, the performance increase and then decrease while the vocabulary size K increase. The best performance is archived at $K = 500$ for our system. This behavior is explained by the intrinsic structure of histopathology images: they are usually composed of some kinds of textures, that is, the number of distinctive patterns in the collection is limited. Perhaps surprisingly, for SIFT based methods, the classification performance decreases while the vocabulary size K increases. The readers might doubt that the performance goes down when the vocabulary grows is very strange, in Fig. 16. This has never been reported before at such small vocabulary sizes in natural image retrieval. I suspect something is wrong here. As we argued before, the medical image only needs a small vocabulary size, not like the natural images. This is especially for the Basal-Cell Carcinoma dataset, in which most images have much simpler patterns. As we can see from Fig. 14 (a), the image is composed of some sample texture patterns. In this case, if we use a larger visual vocabularies, it means the k-means runs a small number of iterations and the visual words are not compact enough.
  In our experience, these parameter settings performed well; nevertheless, small changes in K can affect the performance of each technique as is shown in Fig. 16. Approaches for estimating K in bag-of-features method are currently under investigation.
  For comparison, the results obtained when using the methods proposed by [21], which uses the NN assignment and absolute term frequencies (tf) without any visual weighting are provided in Fig. 16. These results have been computed with the same local descriptors as in the our proposed system. The proposed classification approach based on QP assignment and boosted weighting for HIK similarity yields better performance than a transitional bag-of-feature based on NN assignment and absolute (tf) [21], although the latter is already quite sophisticated.
- **Comparison of Local Descriptor**. The second factor to evaluate is the feature descriptor. As is shown in Fig. 16, the image patch descriptor has obtained a better performance in terms of F-measure among all vocabulary sizes except $K = 50$.
  Table III shows the performance summary of the different configurations evaluated in this work. In bold are the best values of precision, recall and F-measure, showing that patch-based strategies are more effective in general.
  An important question here is why SIFT points, that are provided with robust invariant properties, are not a good choice with this type of histopathology images. First, there is some evidence that they were not designed to find the most informative patches for image classification [21], and second, it is possible that all the attempts to increase the invariance of features in histopathology images, lead to a loss of discriminative information. Furthermore, we also suspect that the reason that SIFT does not perform as well is because of the interest point detector. Thus We also plot the results using dense-sampling instead of interest point detection for SIFT in Fig 16. It is interesting to notice that when the dense-sampling is combined with SIFT descriptor, the performance do get better, but is still not as well as row blocks. However, we can notice that it can benefit from a larger vocabulary size, i.e. its best classification results are got when $V = 750$. This is because when dense-sampling is used to SIFT, more local features are generated and thus makes the vocabulary larger.
  The nature of the descriptor is also a determinant factor in this behavior since the performance of the interest point detection based SIFT points decreases faster than the performance of raw blocks. This suggests that a SIFT-based vocabulary requires less visual words to express all different patterns in the image collection, which is consistent with the rotation and scale invariance properties of that descriptor. On the other hand, a dense-sampled SIFT and image patches-based visual vocabulary requires a larger size because it is representing the same visual patterns using different visual words.
- **Comparison of Visual Word Weighting Strategies to Multiple Assignments**. The next aspect in this evaluation is the image representation, i.e. the use of absolute term frequencies (tf) [7], the use of the weighted scheme provided by inverse document frequencies (tf.idf) [7], the used of weighted scheme provided by [27] to the NN, MA, SA and our QP assignment methods. The results are reported in Table III. Note that in this group of experiments, we utilize image patches as local features, since according to the prevenient experiments, image patches performs better than SIFT as local features.
  Table III lists the classification accuracy for different assignment and weighting configurations. The results Table III show that for different assailment strategies, boosted weighting outperforms tf, tf.idf, and [27] method. There is no significant difference between tf and tf.idf based on statistical hypothesis test. According to the results presented in Table III, it is not clear when idf improves the original tf's classification performance. Moreover, our Boosted weighting achieves accuracy performance similar to that of classifiers trained on whole training sets using [27] method. Generally speaking, Boosted weighting performs equally as well as [27] method, but much better than tf and tf.idf. This may be due to the different strategies used in different weighting

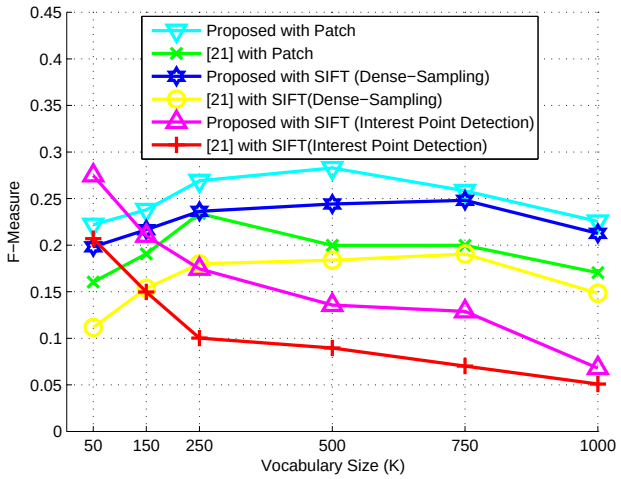


Fig. 16. Vocabulary size K vs. F-Measure for two bag of features representation using SIFT points and image patches.

methods. Boosted and [27] method tend to select learn weights that lead to maximizing margin between classes. The possible explanation for this difference lies in their different theoretical rationales. The idf is a unsupervised method, which don't utilize the class information, while our boosted weighting is learned under the supervision of the class relationships of triplets.

The summary of the bag-of-features based medical image classification experiments using the four different local feature assignment strategies is also presented in Table III. The methods are the convictional NN assignment used by [21], MA [8], SA [9] and out QP assignment Method. These experiments are shown for different weighting methods, and they clearly and consistently illustrate the outperformance of QP assignment with respect to NN, MA and SA for almost all the weighting methods. The outperformance of our QP assignment essentially comes from the inclusion of the contextual visual words; in almost all cases, a small context size $|N(x)| = 20$ was sufficient in order to improve the performance of the boosted weighting, and a few larger for the other weighting methods. On the one hand, this corroborates the fact that the boosted weighting provides state-of-the-art performances, and on the other hand, their performances can be consistently improved by including neighboring contextual visual words of a local feature.

In summary, the use of QP assignment as a base visual descriptor assignment generally leads to the better performance than transitional NN strategy for bag-of-feature algorithms regardless of visual weighting methods. In contrast to the baseline tf and tf.idf visual word weighting's performance, boosted weighting with constantly makes the improvement for both local descriptors regardless of NN or QP assignment.

TABLE III
PERFORMANCE MEASURES FOR THE DIFFERENT VISUAL WORDS WEIGHTING METHODS ([21] TF, [21] TF.IDF, [27] WEIGHTING AND BOOSTED WEIGHTING) TO DIFFERENT LOCAL ASSIGNMENT STRATEGIES ([21] NN, [8] MA, [9] SA AND QP ASSIGNMENTS).

Assignment	Weighting	Precision	Recall	F-Measure
[21] NN	[21] tf	0.610	0.162	0.234
[21] NN	[21] tf.idf	0.634	0.152	0.231
[21] NN	[27] Weighting	0.6487	0.1604	0.2572
[21] NN	Boosted	0.6539	0.1697	0.2695
[8] MA	[21] tf	0.6129	0.1634	0.2580
[8] MA	[21] tf.idf	0.6233	0.1545	0.2476
[8] MA	[27] Weighting	0.6499	0.1689	0.2681
[8] MA	Boosted	0.6547	0.1705	0.2705
[9] SA	[21] tf	0.6371	0.1653	0.2625
[9] SA	[21] tf.idf	0.6380	0.1668	0.2645
[9] SA	[27] Weighting	0.6669	0.1745	0.2766
[9] SA	Boosted	0.6728	0.1774	0.2808
QP	[21] tf	0.6382	0.1666	0.2642
QP	[21] tf.idf	0.6474	0.1691	0.2682
QP	[27] Weighting	0.6675	0.1763	0.2789
QP	Boosted	**0.6790**	**0.1787**	**0.2829**

VI. CONCLUSION

This paper has considered the problem of assigning the local descriptors to visual words, and the weighting of visual words to improve the discriminating power in bag-of-feature based medical image retrieval. We have introduced a method to enable efficient multiple assignment of local feature descriptors to visual words for bag-of-feature methods, and experiments show good results for a variety of data sets, representations, and base descriptor. Our other main contribution is a new visual word weighting factor learning algorithm to construct theoretically sound discriminate weighting strategy functions—for both similarity and distance comparison of histograms representation of images. Experiments demonstrate our technique's accuracy and flexibility for a number of large-scale medical image search tasks.

In future work, we intend to explore online extensions to our algorithm that will allow the local features' assignment and the visual word weighting to be processed in an incremental fashion, while still allowing intermittent queries. We are also interested in considering generalizations of our medical image retrieval method to accommodate alternative assignment and weights learning algorithms within our framework.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for providing valuable comments to the improvement in technical contents and paper presentation. The work was supported by the Major State Basic Research Development Program of China (973 Program) under grant no. 2010CB834303 and a grant from King Abdullah University of Science and Technology.

REFERENCES

[1] Avni, U; Greenspan, H; Sharon, M, et al. X-Ray image categorization and retrieval using patch-based visualwords representation. 2009 IEEE INTERNATIONAL SYMPOSIUM ON BIOMEDICAL IMAGING: FROM NANO TO MACRO, VOLS 1 AND 2 Pages: 350-353 Published: 2009.

[2] Zhi, Li-Jia ; Zhang, Shao-Min; Zhao, Da-Zhe; Zhao, Hong; Lin, Shu-Kuan. Medical image retrieval using SIFT feature. Proceedings of the 2009 2nd International Congress on Image and Signal Processing, CISP'09, 2009.

[3] Jegou, Herve; Schmid, Cordelia; Harzallah, Hedi, et al. Accurate image search using the contextual dissimilarity measure. IEEE Trans Pattern Anal Mach Intell Volume: 32 Issue: 1 Pages: 2-11 Published: 2010 Jan

[4] Varma, Manik; Zisserman, Andrew. A statistical approach to material classification using image patch exemplars. IEEE Trans Pattern Anal Mach Intell Volume: 31 Issue: 11 Pages: 2032-47 Published: 2009 Nov

[5] Lowe, DG. Distinctive image features from scale-invariant keypoints INTERNATIONAL JOURNAL OF COMPUTER VISION Volume: 60 Issue: 2 Pages: 91-110 Published: NOV 2004

[6] Pun, CM; Lee, MC. Log-polar wavelet energy signatures for rotation and scale invariant texture classification. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE Volume: 25 Issue: 5 Pages: 590-603 Published: MAY 2003

[7] Jianxin Wu; Rehg, J.M. Beyond the Euclidean distance: Creating effective visual codebooks using the Histogram Intersection Kernel. 2009 IEEE 12th International Conference on Computer Vision (ICCV) Pages: 630-7 Published: 2009.

[8] Jegou, H; Douze, M; Schmid, C Improving Bag-of-Features for Large Scale Image Search. INTERNATIONAL JOURNAL OF COMPUTER VISION Volume: 87 Issue: 3 Pages: 316-336 Published: 2010.

[9] van Gemert, Jan C; Veenman, Cor J; Smeulders, Arnold W M, et al. Visual word ambiguity. IEEE Trans Pattern Anal Mach Intell Volume: 32 Issue: 7 Pages: 1271-83 Published: 2010 Jul.

[10] Jegou, H.; Harzallah, H.; Schmid, C. A contextual dissimilarity measure for accurate and efficient image search. CVPR '07. IEEE Conference on Computer Vision and Pattern Recognition Pages: 9-16 Published: 2007 2007.

[11] Roweis, ST; Saul, LK. Nonlinear dimensionality reduction by locally linear embedding. SCIENCE Volume: 290 Issue: 5500 Pages: 2323-+ Published: 2000.

[12] Jingdong Wang; Fei Wang; Changshui Zhang, et al. Linear neighborhood propagation and its applications. IEEE Transactions on Pattern Analysis and Machine Intelligence Pages: 1600-15 Published: 09 2009 Sept. 2009.

[13] Athitsos, V; Alon, J; Sclaroff, S, et al. BoostMap: An embedding method for efficient nearest neighbor retrieval. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE Volume: 30 Issue: 1 Pages: 89-104 Published: 2008.

[14] Athitsos, V.; Alon, J.; Sclaroff, S., et al. BoostMap: A method for efficient approximate similarity rankings. Conference Information: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Date: Washington, DC USA. Source: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Pages: II-268-75 Vol.2—2001 Published: 2004.

[15] Man Lan; Chew Lim Tan; Jian Su, et al. Supervised and traditional term weighting methods for automatic text categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence Pages: 721-35 Published: 04 2009 April 2009.

[16] Puzicha, J.; Buhmann, J.M.; Rubner, Y., et al. Empirical evaluation of dissimilarity measures for color and texture. Proceedings of the Seventh IEEE International Conference on Computer Vision—Proceedings of the Seventh IEEE International Conference on Computer Vision Pages: 1165-72 vol.2—2 vol.xxvii+1258 Published: 1999.

[17] Schapire, RE; Singer, Y Improved boosting algorithms using confidence-rated predictions. MACHINE LEARNING Volume: 37 Issue: 3 Pages: 297-336 Published: 1999.

[18] Muller, H.; Deselaers, T.; Deserno, T.M., et al. Overview of the ImageCLEFmed 2007 medical retrieval and medical annotation tasks. Advances in Multilingual and Multimodel Information Retrieval. 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007. Revised Selected Papers Pages: 472-91 Published: 2008.

[19] Muller, H.; Kalpathy-Cramer, J.; Kahn, C.E. Jr., et al. Overview of the ImageCLEFmed 2008 medical image retrieval task. Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008. Revised Selected Papers Pages: 512-22—xxiv+1002 Published: 2009.

[20] Meij, E; Trieschnigg, D; de Rijke, M, et al. Conceptual language models for domain-specific retrieval. INFORMATION PROCESSING & MANAGEMENT Volume: 46 Issue: 4 Pages: 448-469 Published: 2010.

[21] Caicedo, J.C.; Cruz, A.; Gonzalez, F.A. Histopathology image classification using bag of features and kernel functions. Artificial Intelligence in Medicine. Proceedings 12th Conference on Artificial Intelligence in Medicine, AIME 2009 Pages: 126-35—xix+439 Published: 2009.

[22] Hahn, H; Wicking, C; Zaphiropoulos, PG, et al. Mutations of the human homolog of Drosophila patched in the nevoid basal cell carcinoma syndrome. CELL Volume: 85 Issue: 6 Pages: 841-851 Published: 1996.

[23] Varma, Manik; Zisserman, Andrew A statistical approach to material classification using image patch exemplars. IEEE Trans Pattern Anal Mach Intell Volume: 31 Issue: 11 Pages: 2032-47 Published: 2009 Nov.

[24] Yang, XW; Bai, X; Latecki, LJ, et al. Improving Shape Retrieval by Learning Graph Transduction. COMPUTER VISION - ECCV 2008, PT IV, PROCEEDINGS Volume: 5305 Pages: 788-801 Published: 2008.

[25] Bai, X; Yang, XW; Latecki, LJ, et al. Learning Context-Sensitive Shape Similarity by Graph Transduction. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE Volume: 32 Issue: 5 Pages: 861-874 Published: 2010.

[26] Yang, JC; Yu, K; Gong, YH, et al. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. CVPR: 2009 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, VOLS 1-4 Pages: 1794-1801 Published: 2009.

[27] Cai, HP; Yan, F; Mikolajczyk, K Learning Weights for Codebook in Image Classification and Retrieval. 2010 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR) Pages: 2320-2327 Published: 2010.

Jingyan Wang received his B.A. degree in 2007 from Central South University, Changsha, China. Currently, he is a PhD. candidate at the Shanghai Institute of Applied Physics, Chinese Academy of Science, Shanghai, China.

His research interests include Machine Learning and its applications on Bioinformatics, Medical Imaging and Biometrics. Jingyan Wang is a Student Member of IEEE and IEEE Engineering in Medicine and Biology Society.

Yongping Li received MS degree in the Electrical Engineering in 1989 from the Graduate School of Chinese Academy of Sciences. He worked as research fellow at Shanghai Institute of Applied Physics (SINAP) for various research projects on computerized instrumentations from 1989-1995. He joined the Center for Vision Speech and Signal Processing, Univ. of Surrey, UK in 1996 with a Royal Society fellowship, and received PhD in Pattern Recognition, with emphasis on Biometrics, in 2000 under the supervision of Professor Josef Kittler. He worked as a Principal Researcher on speech signal processing at a startup company in Silicon Valley, California from 2001 to 2002. From 2003 to present, he is a research professor and PhD student supervisor at Shanghai Institute of Applied Physics, and a part-time PhD student supervisor at the Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences.

His current research interests include digital signal processing for MCU and DSP based instrumentations, biometric algorithms for face, voiceprint and fingerprint recognition, multiple biometrics, biometric solution integration, biometrics system performance evaluation, testing, and standardization. Dr. Li is a member of the IEEE, IEEE computer society and IEEE instrumentation and measurement.

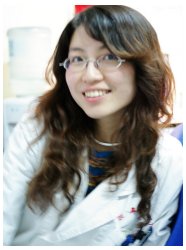
Ying Zhang received the B.E. degree in computer science and technology from Henan University, China in 2004. She is currently pursuing the M.E. degree in signal and information processing at Shanghai Institute of Applied Physics, Chinese Academy of Science, Shanghai, China.

Her research interests include Medical Image Retrieval, Machine Learning and Face Recognition.

Chao Wang received Ph.D. Degree in CS from Graduate of Chinese Academy of Sciences and the Master degree in CS from Shandong University, China. Currently, he is a Postdoc at Oregon Graduate Institute (OGI), Oregon Health & Science University (OHSU), US.

His research interests include Medical Image Processing, Machine Learning, Computer Vision, Pattern Recognition. Especially, his main studies are on face recognition.

Honglan Xie received her PhD Degree from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences in 2004. She is currently a researcher in Shanghai Institute of Applied Physics, Chinese Academy of Science. Her research interest includes synchrotron radiation X-ray imaging and its applications on biomedical imaging.



Guoling Chen Guoling Chen received her Master Degree of Surgery (Otonasolaryngology) in 2010 and Bachelor Degree of Clinical Medicine in 2008 both from Shanghai Medical College of Fudan University, Shanghai, China. She is currently a physician in Zhongshan Hospital, Shanghai, China. Her research interests include Medical Imaging and Computer Aided Diagnosis (CAD).

Xin Gao earned his Ph.D. in Computer Science from the University of Waterloo, Canada, in 2009. He received his bachelor's degree in Computer Science from Tsinghua University, China, in 2004. Currently, Xin Gao is an Assistant Professor of Computer Science in the Mathematical and Computer Sciences and Engineering Division at King Abdullah University of Science and Technology, Saudi Arabia. He is also an adjunct faculty member in David R. Cheriton School of Computer Science at the University of Waterloo.

Xin Gao's research interests are in bioinformatics, computational biology, algorithms and machine learning. He is interested in designing algorithms and developing machine learning techniques to solve problems in structural biology, systems biology and biological sequence analysis.