Random Seeding for Keypoint Quantization: Application to Large-Scale Image Retrieval

Fengjie Li, Wei Tong, Rong Jin, Member, IEEE, and Anil K Jain, Fellow, IEEE

A preliminary version of this work appeared in the ACM Multimedia 2009 (ACMMM 09) Large-Scale Multimedia Retrieval and Mining Workshop (LS-MMRM).

March 30, 2010
Abstract

Local image features, often referred to as keypoints, have been shown to be extremely effective for large scale near duplicate image retrieval. The most popular approach for keypoint based image retrieval is the clustering-based bag-of-words model. It maps each keypoint in an image to a visual word in a code-book that is usually constructed by a clustering algorithm; each image is then represented by a histogram of visual words. Despite its success, this approach has two main shortcomings: (i) High computational cost of clustering large number of keypoints to construct visual words. Besides, the cost of mapping keypoints to visual words is linear in the number of keypoints, making it expensive to quantize a large image database. (ii) Inconsistent mapping of keypoints. Since the radius of clusters (i.e., the maximum distance of keypoints in a cluster and cluster center) could vary significantly from clusters to clusters, two keypoints can be mapped to the same visual words even if they differ significantly in visual features. We propose a new scheme for keypoint quantization to overcome these shortcomings. Instead of finding visual words by clustering, we randomly select a subset of keypoints as visual words and thereby avoid the high computational cost of clustering. Instead of mapping a keypoint to the closest visual word, we deploy a more consistent criterion in which a keypoint is mapped to a visual word if and only if its distance to the visual word is smaller than a threshold. The proposed random seeding algorithm leads to a more efficient procedure for keypoint quantization, where the computational complexity of keypoint quantization is sublinear in the number of keypoints. Experimental results on near duplicate image retrieval on a database of one million images shows that the proposed quantization scheme is significantly more effective and efficient than the clustering-based approach. This conclusion is further confirmed by our study on large-scale tattoo image retrieval and automatic image annotation.

Index Terms

Image Retrieval, Keypoints, Bag-of-Words Model, Clustering, Quantization

I. INTRODUCTION

Content-based image retrieval (CBIR) has been studied for many years, but the challenge of semantic gap, i.e. the gap between visual similarity and conceptual/perceptual relevance, makes it a much harder problem than most researchers anticipated [1]. However, recent studies have shown that near duplicate image retrieval [2] can be solved effectively by using visual features. Unlike general content-based image retrieval that aims to identify images that are “semantically” relevant to a given query image, the objective of near duplicate image retrieval is to identify images with high visual similarity (see Fig. 1), thereby avoiding the challenge of semantic gap.
Fig. 1. Examples of Near Duplicate Image Retrieval. The first column shows the query images and the subsequent columns are near duplicate images.

A number of studies [2]–[7] have shown that local image features, e.g. SIFT descriptors [8], often referred to as keypoints, are significantly more effective for near duplicate image retrieval than global image features, such as color, texture and shape. The main idea of the keypoint based approach is to extract salient local patches from an image, and represent each local patch by a multi-dimensional feature vector. As a result, each image is represented by a collection of multi-dimensional vectors, which is often referred to as a bag-of-features representation [9].

A number of algorithms have been proposed to measure the distance between two images based on their bag-of-features representations, including similarity based on the optimal partial matching between two sets of key points [10]–[12], pyramid kernel similarity [13], similarity based on the principal angle between two sets of key points [14], and similarity measure based on the match between two distributions [15], [16]. It has been shown [11], [12], [17], [18] that despite its simplicity, the similarity based on the optimal partial matching performs well in comparison to the other similarity measures. It measures the distance between two images by finding the best mapping between the keypoints in the two images that has the overall shortest distance. The main shortcoming of the optimal partial matching is its high computational cost: given a query image, a linear scan is required to compute the similarity between the query and every image in the database, which does not scale well to a large image database.

Several methods have been proposed to improve the computational efficiency of optimal partial
matching [4], [19], [20]. Among them, the bag-of-words model [4] is probably the most popular one due to its empirical success. It takes advantage of the inverted index, which has been successfully used by web search engines to index billions of documents. The key idea is to quantize the continuous high-dimensional space of SIFT features to a vocabulary of visual words, which is typically achieved by a clustering algorithm. By treating each cluster center as a word in a codebook, this approach maps each image feature to its closest visual word, and then represents each image by a histogram of visual words. A number of studies have shown promising performance of this approach for image retrieval [2]–[7], [21] and object recognition [4], [22]–[27].

Despite its success, the bag-of-words model suffers from the following drawbacks:

- **High computational cost in visual vocabulary construction.** One of the key steps in constructing the bag-of-words model is to cluster a large number of keypoints into a relatively smaller number of visual words. For large scale image retrieval, we often need to cluster billions of keypoints into millions of clusters. Although several efficient algorithms [3], [6], [21], [28]–[32] have been developed for large-scale clustering problems, it is still expensive to generate a vocabulary with millions of visual words.

- **High computational cost in keypoint quantization.** Given the constructed visual vocabulary, the next step in bag-of-words model is to map each keypoint in a database image to a visual word, which requires finding the nearest neighbor of every keypoint to the visual words. Since the computational cost for keypoint quantization is linear in the number of keypoints, it is expensive to quantize keypoints for a very large image database to a visual vocabulary. Even with the help of approximate nearest neighbor search algorithms, this step is still costly when good approximation is desired.

- **Inconsistent mapping of keypoints to visual words.** The radius of clusters (i.e., the maximum distance between the keypoints in a cluster and its center) could vary significantly from cluster to cluster. As a result, for clusters with large radius, two keypoints can be mapped to the same visual word even if they differ significantly in visual features, leading to an inconsistent criterion for keypoint quantization and potentially poor performance in image matching.

- **Lack of a theoretic analysis.** Most published studies on the bag-of-words model are motivated by efficiency considerations and are primarily focused on its empirical performance.
Although [33] showed that the similarity between two bag-of-features representations can be interpreted as a matching algorithm between descriptors, it did not establish the relationship between the bag-of-words model and the optimal partial matching. Without a theoretical analysis, the success of the bag-of-words model may only be demonstrated based on empirical performance.

We emphasize that although visual vocabulary construction and keypoint quantization are performed offline, their computational efficiency is still very important because most image databases are constantly updated. For instance, more than 1.5 million photos are uploaded on Flickr every day. Given a large-scale dynamic image database, it is critical to develop methods to efficiently quantize the keypoints and furthermore efficiently create new visual words to capture the visual characteristics of the new images.

We present a new quantization scheme that explicitly addresses the shortcomings of the clustering-based approach for keypoint quantization. Unlike most studies on keypoint quantization that are motivated by computational efficiency, the proposed scheme aims to effectively approximate the optimal partial matching based similarity measure. The main idea is to discretize each keypoint into a binary vector by a set of randomly generated hyper-spheres. Each hyper-sphere is analogous to a cluster, and discretizes a key point into a binary bit: 1 when the keypoint is within the hyper-sphere and 0, otherwise. The bag-of-words representation for an image is now computed as a histogram over the hyper-spheres. Our theoretical analysis shows that despite its simplicity, the image similarity based on the proposed quantization scheme serves as an upper bound on the similarity based on optimal partial matching. To distinguish the proposed scheme from the clustering based approach for keypoint quantization, we refer to it as a random seeding approach for keypoint quantization.

The computational advantage of the proposed approach arises in two aspects: first, the random seeding approach does not require clustering (e.g., K-means algorithm) to identify visual words, leading to a significant saving in computation; second, given the constructed visual vocabulary, the complexity of quantizing \( m \) keypoints into \( n \) visual words is \( O(m \ln n) \) for the clustering-
based approach, and $O(n \ln m)$ for the proposed scheme. Since $m$ is usually at least one order of magnitude larger than $n$, the random seeding approach is significantly more efficient than the clustering-based approaches for keypoint quantization, which is confirmed by our empirical study.

Besides the computational efficiency, the random seeding approach deploys a more consistent criterion for keypoint quantization: a keypoint is mapped to a visual word if and only if the distance between the keypoint and the visual word is smaller than the radius $r$ of hypersphere. By choosing a relatively small $r$, the proposed quantization approach is able to ensure a large similarity between a keypoint and its mapped visual word. In contrast, the clustering-based approach maps a keypoint to its closest visual word even when they are separated by a large distance, leading to an inconsistent criterion for keypoint quantization and potentially poor performance in image matching. Figure 2 shows the distributions of distances between keypoints and their nearest cluster centers for the Oxford5K dataset with 10K clusters. It can be seen that these distances have a large variance, indicating a potentially inconsistent criterion for mapping keypoints to visual words. Furthermore, some point-to-center distances are very large, making the clustering-based quantization less accurate.

We integrate the random seeding approach into a recognition/retrieval system for large-scale

---

1The analysis here only considers the time for approximate nearest-neighbor/range search in keypoint quantization. It ignores the time for constructing KD-trees, which in practice is one order of magnitude smaller than approximate nearest-neighbor/range search.
near duplicate image retrieval. We demonstrate the efficacy and efficiency of our system for image retrieval on a database of more than 1 million images. We further validate our system in the domains of tattoo image retrieval and automatic image annotation.

The rest of the paper is organized as follows: Section 2 reviews the related work on image retrieval and efficient methods for keypoint quantization; Section 3 describes the proposed scheme for keypoint quantization, together with a theoretical analysis that validates the proposed approach; Section 4 presents our empirical study of the proposed scheme through a large-scale near duplicate image retrieval experiment. It also validates our conclusions in two application domains, tattoo image retrieval and automatic image annotation. Section 5 makes concluding remarks about our contribution and suggests future research directions.

II. RELATED WORK

Although the bag-of-words model has been shown to be effective for image/object retrieval, one of the major challenges, as pointed out in [3], is how to efficiently construct a vocabulary with millions of visual words for a very large image database. This requires developing efficient algorithms for large-scale data clustering. One of the early studies [4] used flat K-means clustering, but it is unable to handle large vocabularies. In [5], the authors showed that hierarchical clustering is able to greatly increase the size of visual vocabulary and therefore is more suitable for large-scale image retrieval. Recent studies [3], [6], [21], [28] have shown that K-means based on approximate nearest neighbor search can scale to a very large vocabulary. The goal of these studies has been to improve the efficiency of approximate nearest neighbor search by a random algorithm, such as randomized trees [6] or Locality Sensitive Hashing (LSH) [29], [30].

Several algorithms have also been proposed for efficient keypoint quantization that avoid clustering [34], [35]. These methods essentially extend Locality Sensitive Hashing [30], which was originally designed for efficient nearest neighbor search in high dimensional space, to efficiently compute the matching between two sets of keypoints. The main objective of these algorithms is to improve computational efficiency of keypoint quantization, but they have been shown to be less effective than the clustering based approach. In contrast, the proposed random seeding approach is motivated by the approximation of the similarity measure based on optimal partial matching. Our empirical studies show that the proposed approach not only has better retrieval performance but it is computationally more efficient than clustering-based quantization.
Several approaches have been proposed to improve the overall retrieval performance of the bag-of-words model. In [36], relevance feedback is used to improve the retrieval accuracy of a bag-of-words model. In [37], the images returned by a bag-of-words model is reranked by taking account the geometrical constraints in computing the matching scores. The concept of bundle features is proposed in [38] to combine two sets of matching scores, one based on a bag-of-words model and the other based on the matching of image regions. In [39], the authors represented each image by two models, a bag-of-words model and a mixture model; the similarity between two images is calculated as a linear combination of matches based on the two models. In [40], multiple bag-of-words models are constructed and combined to improve the retrieval accuracy. While these approaches improve the retrieval performance, they are all built upon the basic bag-of-words model. Thus, by improving the efficiency and effectiveness of bag-of-words models as proposed here, we will essentially improve the performance of all these approaches for image retrieval.

Finally, since the proposed approach essentially represents each keypoint by a binary vector, it is also related to spectral hashing [41] and Hamming embedding [33] that aim to embed keypoints in a space of binary vectors. The proposed approach differs from the other embedding approaches in that it provides a very sparse set of binary vectors, leading to efficient retrieval algorithms for large image databases.

III. RANDOM SEEDING FOR KEYPPOINT QUANTIZATION

In this section, we first present the random seeding algorithm for keypoint quantization. We then establish the relationship between the proposed scheme for keypoint quantization and the similarity measure based on the optimal partial matching between two sets of keypoints.

Let \( \mathcal{D} = (\mathcal{I}_1, \ldots, \mathcal{I}_T) \) be a collection of \( T \) images. Each image \( \mathcal{I}_i \) is represented by a set of \( n_i \) keypoints, denoted by \( X_i = (x_{i1}^1, \ldots, x_{in_i}^n) \), where each keypoint \( x_{ij} \in \mathbb{R}^d \) is a vector of \( d \) dimensions. We now define a set \( X \) of all the keypoints from all the \( T \) images. The first step of the proposed scheme is to randomly sample \( N \) keypoints from this set \( X \), denoted by \( z_1, \ldots, z_N \). A hyper-sphere \( \mathcal{B}_i \) is created centered at each sampled keypoint \( z_i \), i.e., \( \mathcal{B}_i = \{ x \in \mathbb{R}^d : |x - z_i|_2 \leq r \} \), where \( r \) is a predefined constant that is computed based on the average distance between any two keypoints in image collection \( \mathcal{D} \). Using the hyper-spheres \( \{ \mathcal{B}_i \}_{i=1}^N \), we quantize each keypoint \( x \) into a binary vector \( \tilde{b}(x) = (b_1(x), \ldots, b_N(x)) \), where each element...
$b_i(x)$ is 1 if $x \in B_i$ and 0, otherwise. The bag-of-words representation for image $I_i$, denoted by $\vec{b}(I_i) = (b_1(X_i), \ldots, b_N(X_i))$, is computed by adding the binary vectors of all the keypoints in image $I_i$, i.e., $b_k(X_i) = \sum_{j=1}^{n_i} b_k(x^i_j), k = 1, \ldots, N$.

One of the key properties of the proposed algorithm is its criterion for mapping a keypoint to $N$ visual words: a keypoint $x^i_j$ is mapped to a visual word $z_k$ if and only if the distance between $x^i_j$ and $z_k$ is smaller than the threshold $r$. This is in contrast to the criterion deployed by the clustering approach for keypoint quantization where a keypoint is always mapped to its closest visual word even if the closest cluster center is far from the keypoint. The advantages of this new mapping criterion are threefold:

- **Consistent mapping.** The mapping criterion implies that two keypoints will never be mapped to the same visual word if they are separated by a distance of more than $2r$. By appropriately choosing the threshold $r$, we can ensure that only “similar” keypoints will be mapped to the same visual word, leading to a more consistent mapping criterion than the clustering-based approach.

- **Mapping to multiple visual words.** A keypoint can be mapped to multiple visual words if it is within distance $r$ of more than one visual word. This is similar to the idea of soft membership [42] that allows us to capture the uncertainty in mapping keypoints to visual words. One problem of soft membership is that every keypoint is forced to be associated with multiple visual words even if it is much closer to one visual word than the other visual words. In contrast, in the proposed scheme, soft membership is introduced for a keypoint only when it is close to multiple visual words simultaneously; a keypoint will be mapped to a single visual word if it is far away from others.

- **Eliminating outlier keypoints.** A keypoint is ignored by the proposed algorithm if its distances to all the $N$ visual words are larger than the threshold $r$. The underlying rationale is that if a keypoint is far away from all the visual words, it is very likely to be an outlier and therefore should be skipped in the construction of the bag-of-words model. By eliminating the outlying keypoints, we not only improve the accuracy of image retrieval but also its efficiency as a smaller number of keypoints need to be considered for image matching.

\(^2\)Although this problem can be solved using kernel function [43], it is computationally prohibitive when both the number of keypoints and the number of visual words are large.

March 30, 2010
Algorithm 1 Random Seeding Approach for Keypoint Quantization

Input:
- Image collection $\mathcal{D} = \{I_1, \ldots, I_T\}$, with each image $I_i$ represented by a set of keypoints
- Number of visual words $N$
- Distance threshold $r$

Visual Vocabulary Construction
- Generate visual words $z_1, \ldots, z_N$ by randomly sampling $N$ keypoints from $X$, which is the set of all keypoints of all $T$ images in $\mathcal{D}$

Keypoint Quantization
1) Construct a randomized KD tree $T$ for all the keypoints detected in $\mathcal{D}$.
2) For $i = 1, \ldots, N$
   - Use KD tree $T$ and approximate nearest neighbor search to efficiently find the keypoints within the distance $r$ of visual word $z_i$
   - Compute the histogram of visual word $z_i$ for all the images in $\mathcal{D}$

To efficiently implement the proposed random seeding approach, we use the Fast Library for Approximate Nearest Neighbors (FLANN)$^3$, which implements the randomized KD-trees [32], to efficiently decide which subset of keypoints are within distance $r$ of a given visual word. Algorithm 1 shows the major steps of the proposed algorithm. Compared to the clustering-based approach, there are two main features of Algorithm 1:

- It constructs the visual vocabulary by randomly sampling keypoints from the image collection $\mathcal{D}$ and thus avoids the high computational cost of clustering.
- It constructs randomized KD-trees for all the keypoints in $\mathcal{D}$, and iterates over the visual words to find the subset of keypoints that are within a distance $r$ of a visual word. As a result, the computational cost for quantizing $m$ keypoints into $N$ visual words is $O(N \log m)$. In contrast, the clustering-based approach constructs randomized KD-trees for visual words, and iterates over keypoints to find the nearest visual word for each keypoint, leading to $O(m \log N)$ complexity for keypoint quantization. Since $m$ is usually at least an order of

magnitude larger than $N$, the proposed algorithm is more efficient than the clustering-based approach for keypoint quantization. The above analysis ignores the time for constructing randomized KD-trees, which is usually an order faster than mapping keypoints to visual words.

The theoretical foundation of our quantization scheme is based on the relationship between the image similarity resulting from this quantization scheme and the image similarity resulting from optimal partial matching. We first show that the partial matching based similarity can be upper bounded by a smooth similarity function. We then derive the proposed scheme of keypoint quantization that approximates this upper bound similarity function with a small error.

A. Distance Measure Based on Optimal Partial Matching

Let two images $I_x$ and $I_y$ be represented by the corresponding sets of keypoints, denoted by $X = (x_1, \ldots, x_m)$ and $Y = (y_1, \ldots, y_n)$, where each $x_i$ and $y_j$ is a vector in $d$-dimensional space. Among various similarity measures that have been proposed, the similarity based on the optimal partial matching [11]–[13] is probably the most intuitive one, and has yielded the state-of-the-art performance for both image retrieval and object recognition. Let $\pi^1 : \{1, \ldots, m\} \mapsto \{1, \ldots, n\}$ map each point in set $X$ to a point in $Y$, and $\pi^2 : \{1, \ldots, n\} \mapsto \{1, \ldots, m\}$ map each point in set $Y$ to a point in $X$. Then the optimal partial matching distance between $X$ and $Y$ for the given mappings $\pi^1$ and $\pi^2$, denoted by $d(X, Y; \pi^1, \pi^2)$, is computed as

$$d(X, Y; \pi^1, \pi^2) = \sum_{k=1}^{m} \|x_k - y_{\pi^1_k}\|_2^2 + \sum_{k=1}^{n} \|y_k - x_{\pi^2_k}\|_2^2,$$

(1)

where $\pi^1_k$ indicates the index of the point in set $Y$ to which $x_k$ is mapped; similarly, $\pi^2_k$ indicates the point in set $X$ to which $y_k$ is mapped. Finally, the optimal partial matching distance between $X$ and $Y$, denoted by $d(X, Y)$, is obtained by minimizing $d(X, Y; \pi^1, \pi^2)$ over the mappings $\pi^1$ and $\pi^2$.

B. Approximating the Optimal Partial Matching by A Smooth Similarity Function

It is in general difficult to derive an appropriate scheme of key point quantization directly from (1) because the distance function $d(X, Y)$ is non-smooth and is defined implicitly through the

**Note that we generalize the optimal partial matching in [13] by making it to be a symmetric distance measure.**
minimization of the mappings $\pi^1$ and $\pi^2$. To address this difficulty, we approximate the distance function in (1) by a smooth similarity function, as shown in the following lemma.

**Lemma 1:** For any $\lambda > 0$, we have

$$-d(X, Y) + (m + n)d_0 \leq \frac{2}{\lambda} \sum_{i=1}^{m} \sum_{j=1}^{n} \exp(-\lambda\|x_i - y_j\|_2^2 - d_0),$$

(2)

where $d_0 = \min_{i,j} \|x_i - y_j\|_2^2$.

Proof of this lemma can be found in the Appendix I. Note that the equality in Lemma 1 is satisfied when $d_0 = |x_k - y_{\pi^1 k}|_2^2 = |y_k - x_{\pi^2 k}|_2^2$ for all $k$ and as $\lambda$ goes to infinity. The lemma significantly simplifies the similarity measure, leading to a potentially efficient implementation of image retrieval systems. In figure 3 we show the difference between the similarity based on optimal partial matching and the upper bound presented in Lemma 1. In this example, we randomly generate a set $X$ that contains 1,000 data points in $R^{128}$, and each feature is an integer uniformly drawn from interval $[1, 255]$. The second bag $Y$ is generated by perturbing data points in $X$ by adding a scaled Gaussian noise $\gamma N(0, 1)$ to every feature, where $\gamma > 0$ controls the noise level and is varied from 0 to 50. We set $\lambda = 25$ in this study. We observed from Figure 3 that when the perturbation is small (i.e., $X$ and $Y$ are similar), the upper bound provides a good approximation to the similarity based on optimal partial matching (measured in $-d(X, Y)$); the gap between the two similarity measures becomes large when the similarity between $X$ and $Y$ is small. Since we are mostly interested in near duplicate image retrieval, it is therefore much more important to make a good approximation when the optimal partial matching based distance is small, making the result of Lemma 1 valuable in our application domain.

Using Lemma 1, we introduce the following similarity measure between point sets $X$ and $Y$

$$s(X, Y) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \exp\left(-\lambda\|x_i - y_j\|_2^2\right).$$

(3)

Note that compared to Lemma 1, we set $d_0 = 0$ in the above similarity measure and introduce the factor $1/mn$, which is useful for simplifying our analysis as shown later. Instead of directly computing the distance $d(X, Y)$ between a query image and a database image, we will compute the similarity $s(X, Y)$, and rank images in a database in the descending order of their similarity.

**C. Random Seeding Algorithm**

We now show that the similarity measure $s(X, Y)$ defined in (3) can be computed efficiently by the proposed keypoint quantization scheme. The overall idea is to first interpret the similarity
measure $s(X,Y)$ as an expectation over a hidden variable $z$. We then approximate $s(X,Y)$ with a high accuracy by replacing the computation of expectation with an average over a finite number of samples. We finally show that the approximation over the finite number of samples leads to the bag-of-words model that can be efficiently computed.

1) Similarity Measure $s(X,Y)$ as Expectation: In order to interpret the similarity measure $s(X,Y)$ as an expectation over a hidden variable $z$, we consider a probabilistic model for the similarity measure between two sets of keypoints. We assume that the given set of keypoints $X = (x_1, \ldots, x_m)$ are i.i.d. samples from an unknown distribution, denoted by $p(z|X)$. Using the kernel density function estimation [44], the estimate of $p(z|X)$, denoted by $\hat{p}(z|X)$, is computed as follows

$$\hat{p}(z|X) = \frac{1}{m} \sum_{i=1}^{m} \frac{\mu^{d/2}}{[\pi d/2]} \exp\left(-\mu|z - x_i|^2\right), \quad (4)$$

where $\mu$ specifies the kernel width and $d$ is the dimensionality. Similarly, the estimate of the unknown distribution for a set $n$ of keypoints $Y$, denoted by $\hat{p}(z|Y)$, is computed as

$$\hat{p}(z|Y) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mu^{d/2}}{[\pi d/2]} \exp\left(-\mu|z - y_i|^2\right). \quad (5)$$

Fig. 3. An example showing the relationship between the optimal partial matching similarity and the upper bound in Lemma 1. It shows that when two sets of keypoints (corresponding to two images) are similar, the upper bound in Lemma 1 provides a good approximation to optimal partial matching.
The following lemma shows that the similarity measure \( s(X, Y) \) can be expressed in \( \hat{p}(z|X) \) and \( \hat{p}(z|Y) \).

**Lemma 2:** The following relationship holds if \( \mu = 2\lambda \)

\[
s(X, Y) = \frac{[2\pi]^{d/2}}{\mu^{d/2}} \int \hat{p}(z|X)\hat{p}(z|Y)dz
\]  
(6) 

Proof can be found in Appendix II.

To view \( \int \hat{p}(z|X)\hat{p}(z|Y)dz \) as an expectation, we need to introduce an appropriate distribution for \( z \). Note that \( \int \hat{p}(z|X)\hat{p}(z|Y)dz \) is defined as an integration over the unbounded space of \( z \). Hence, it is inappropriate to define a uniform distribution \( f \) or \( \hat{\gamma} \) as follows

\[
\hat{\gamma} = \sqrt{\frac{2}{\pi}} \mu^{d/2} \int \hat{p}(z|X)\hat{p}(z|Y)dz
\]

As indicated by Lemma 3, a large value of \( \gamma \) will lead to a small value for \( |\int \hat{p}(z|X)\hat{p}(z|Y)dz-\hat{s}_1(X, Y)| \), implying that \( \hat{s}_1(X, Y) \) is close to \( \int \hat{p}(z|X)\hat{p}(z|Y)dz \). This is shown by the following corollary.

**Corollary 1:** If

\[
\gamma \geq \max \left( 4, \frac{\pi^{1/2}}{(2\mu)^{1/2}R^d} \left( (d-1) + \sqrt{(d-1)^2 + 4\mu \ln(1/\delta)} \right) \right)
\]  

\( \delta \)This assumption is reasonable since each dimension of the SIFT descriptor is upper-bounded.

March 30, 2010 DRAFT

https://mc.manuscriptcentral.com/tpami-cs
we have
\[
\frac{|s(X, Y) - \zeta \hat{s}(X, Y)|}{s(X, Y)} \leq \delta
\]
The result follows immediately from Lemma 3.

With the above analysis, we can now focus on the similarity measure \( \hat{s}(X, Y) \). Since the domain of \( z \) in \( \hat{s}(X, Y) \) is bounded, we can now introduce a uniform distribution for \( z \), i.e.,
\[
q(z) = \frac{1}{\text{vol}(Q)} I(z \in Q),
\]
where \( \text{vol}(Q) \) stands for the volume of domain \( Q \) and \( I(x) \) is an indicator function that outputs 1 when \( x \) is true and 0 otherwise. Using the distribution \( q(z) \), we can interpret \( \hat{s}(X, Y) \) as the expectation of \( \hat{p}(z|X)\hat{p}(z|Y) \) over random variable \( z \), i.e.,
\[
\hat{s}(X, Y) = \text{vol}(Q) \mathbb{E}_z[\hat{p}(z|X)\hat{p}(z|Y)]
\]
(9)

In the remaining analysis, we will drop the factor \( \text{vol}(Q) \) for the sake of simplicity.

2) Keypoint Quantization by Random Samples: We can further approximate the expectation interpretation in (9) by replacing the distribution \( q(z) \) with a finite number of samples. Let \( z_1, z_2, \ldots, z_N \) be \( N \) samples randomly drawn from the distribution \( q(z) \). Using the samples \( \{z_k\}_{k=1}^N \), we approximate \( \hat{s}(X, Y) \) by \( \hat{s}_2(X, Y) \) that is defined as follows
\[
\hat{s}_2(X, Y) = \frac{\mu^d}{mn\pi^d} \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^N \frac{1}{N} \exp(-\mu|x_i|^2 - \mu|z_k - y_j|^2) \]
(10)
\[
= \frac{\mu^d}{mnN\pi^d} \sum_{k=1}^N \left( \sum_{i=1}^m \exp(-\mu|x_i|^2) \left( \sum_{j=1}^n \exp(-\mu|z_k - y_j|^2) \right) \right). \]

It is important to note that, in the expression of \( \hat{s}_2(X, Y) \), by replacing the expectation over the distribution \( q(z) \) with the average over \( N \) samples, we essentially represent each point set \( X = (x_1, \ldots, x_m) \) by a vector
\[
\vec{f}(X) = (f_1(X), \ldots, f_N(X))
\]
(11)
where \( f_k(X), k = 1, \ldots, N \) is defined as
\[
f_k(X) = \frac{1}{m} \sum_{i=1}^m \exp(-\mu|x_i|^2).
\]
Using the notation \( \vec{f}(X) \) and \( \vec{f}(Y) \), we can write \( \hat{s}_2(X, Y) \) as
\[
\hat{s}_2(X, Y) = \vec{f}(X) \cdot \vec{f}(Y),
\]
i.e., the similarity \( \hat{s}_2(X, Y) \) can be computed as the dot product between two histograms. The theorem below shows that with a sufficiently large number \( N \), \( \hat{s}_1(X, Y) \) and \( \hat{s}_2(X, Y) \) will differ only by a small value.

**Theorem 2:** With probability \( 1 - \delta \), we have

\[
|\hat{s}_2(X, Y) - \hat{s}_1(X, Y)| \leq \sqrt{\frac{1}{N} \ln \frac{2}{\delta}}
\]

Proof can be found in Appendix V. The result in Theorem 2 only applies to two images. We furthermore generalize it to a collection of images. Let \( D = (I_1, \ldots, I_T) \) be a collection of \( T \) images and \( X_i = (x_i^1, \ldots, x_i^{n_i}) \) be the set of keypoints associated with image \( I_i \). We have the following corollary showing that for any two images \( I_i \) and \( I_j \) in \( D \), their similarities \( \hat{s}_1(X_i, X_j) \) and \( \hat{s}_2(X_i, X_j) \) are close to each other.

**Corollary 3:** Given a collection \( D \) of \( T \) images, with probability \( 1 - \delta \), the following inequality holds for any two images \( I_i \) and \( I_j \) in \( D \)

\[
|\hat{s}_2(X_i, X_j) - \hat{s}_1(X_i, X_j)| \leq \sqrt{\frac{1}{N} \ln \frac{T(T-1)}{\delta}}
\]

The above corollary is easily proved by a union bound.

Although (11) provides a vector representation for each keypoint set, it is overall not a sparse representation, which makes it difficult to implement an efficient retrieval algorithm that scales well to a large image database. To address this challenge, we introduce a threshold \( \eta > 0 \) and set the coefficient \( \exp(-\mu |z_k - x_i|_2^2) \) to be zero whenever \( \exp(-\mu |z_k - x_i|_2^2) \) is smaller than the threshold. This is equivalent to replacing the coefficient \( \exp(-\mu |z_k - x_i|_2^2) \) with \( \max(\exp(-\mu |z_k - x_i|_2^2) - \eta, 0) \). As a result, we represent each keypoint set \( X = (x_1, \ldots, x_m) \) by a vector

\[
\vec{h}(X) = (h_1(X), \ldots, h_N(X))
\]

where \( h_k(X), k = 1, \ldots, N \) is defined as

\[
h_k(X) = \sum_{i=1}^{m} \max(\exp(-\mu |z_k - x_i|_2^2) - \eta, 0)
\]

Note that by introducing the threshold \( \eta \) into the bag-of-words model, the result in Corollary 3 is relaxed by a factor \( \eta m \). The threshold \( \eta \) essentially balances the tradeoff between efficiency and efficacy: a large \( \eta \) will result in a sparse bag-of-words representation of images, leading to efficient retrieval of gallery images; on the other hand, a small \( \eta \) will result in a more accurate approximation of the optimal partial matching similarity, leading to accurate retrieval of visually similar images.
3) **Implementation:** To generate the bag-of-words model $\vec{h}(X)$ in (12), the first step is to sample centers $z_i$’s from the domain $Q$. In practice, sampling $z_i$ uniformly from a high dimensional domain is a nontrivial problem. In our implementation, we acquire the samples $z_1, \ldots, z_N$ by first sampling $N$ keypoints from all the keypoints in the image database $D$, and then scaling the sampled keypoints by a factor $\gamma$. This practice essentially assumes that all the keypoints are uniformly distributed within a hyper-sphere. Although this assumption may not be true, it significantly reduces the computational cost of the proposed algorithm, and also yields good performance. We set $\gamma = 1$ in our experiment. Although our theoretical result requires $\gamma$ to be large, $\gamma = 1$ does yield desirable performance in our empirical study.

The second step in the bag-of-words model is to compute element $h_k(X)$ in $\vec{h}(X)$. Note that $h_k(X)$ defined in (13) is a real number, making it difficult to implement it by a typical text search engine that requires integer elements in a bag-of-words model. We thus simplify the function $\max(0, \exp(-\mu|z_k - y_j|^2)^2 - \eta)$ to $I(\exp(-\mu|z_k - y_j|^2)^2 \geq \eta)$, where $I(z)$ is an indicator function that outputs 1 if $z$ is true and 0, otherwise. Note that function $I(\exp(-\mu|z_k - y_j|^2)^2 \geq \eta)$ is equivalent to $I(|z_k - y_j|^2 \leq r)$, with $r = \sqrt{\ln(1/\eta)/\mu}$. In practice, we set $r$ to be proportional to the average distance between any two keypoints in image collection $D$. In order to estimate the average pairwise distance, we randomly sample 10,000 keypoint pairs from the entire collection.

To efficiently compute $I(|z_k - y_j|^2 \leq r)$, which requires identifying the subset of keypoints in the image collection $D$ that are within the range $r$ of each center $z_k$. We use the Fast Library for Approximate Nearest Neighbors (FLANN) in Algorithm 1.

### IV. Experiments

To verify both the efficiency and efficacy of the random seeding based quantization scheme, we conduct experiments on near duplicate image retrieval on the Oxford5K dataset. Following the experimental settings in [37], we introduce a distractor (background) database consisting of 1 million images downloaded from Flickr. We further validate the proposed quantization scheme by conducting two additional experiments in the domain of tattoo image retrieval and automatic image annotation. The tattoo image retrieval experiment is performed on a dataset of 100,000 tattoo images, where the objective is to find images with the similar tattoo pattern as the one

---


---
given in the query. For the experiment on automatic image annotation, we used the ESP dataset that consists of over 100,000 manually annotated images, and the objective is to automatically annotate images with appropriate keywords that reflect the visual content of the images. Table I summarizes the three experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Query Set</th>
<th>Gallery Set</th>
<th>Total # of Keypoints</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Near duplicate image retrieval</td>
<td>55 Oxford Images</td>
<td>Oxford5K + Flickr1M</td>
<td>14.97M + 798M</td>
<td>mAP</td>
</tr>
<tr>
<td>Tattoo image retrieval</td>
<td>995 Tattoos</td>
<td>Tattoo60K + Esp40K</td>
<td>10.8M</td>
<td>CMC</td>
</tr>
<tr>
<td>Automatic image annotation</td>
<td>1,000 ESP Images</td>
<td>Esp110K</td>
<td>12.8M</td>
<td>F1</td>
</tr>
</tbody>
</table>

**Table I**

**Summary of experiments**

In all the three experiments, we obtain the features based on Harris detector and SIFT descriptor, using the feature extractor in [45] with the default parameters. In the baseline method, hierarchical K-means is applied to cluster keypoints into visual words, and then each keypoint is mapped to its nearest visual word. We used the implementation of large-scale data clustering in [28], which has been optimized for efficient SIFT keypoint matching. Since both the baseline method and the proposed random seeding quantization depend on efficient large scale nearest neighbor search, we used the FLANN [28] in both the methods. The parameter setting for hierarchical K-means is tuned using the auto-tune feature provided in the library, by using 10% of the dataset and setting the accuracy level to be 95%. The search parameters for the proposed method are set accordingly to make the results comparable. For the implementation of the random seeding algorithm, unless otherwise specified, radius $r$ is set to 50% of the average pairwise distance between any two keypoints. We estimate the average pairwise distance by a sampling approach described in Section III-C.3. After obtaining the bag-of-words representations (either by the clustering-based method or by random seeding), Okapi BM25, a state-of-the-art text retrieval model [46], was used to compute the similarity between a query and gallery images.

7http://www.gwap.com/gwap/gamesPreview/espgame/
For automatic image annotation, our retrieval system is first used to identify the gallery images that are visually similar to a test image; a $k$-NN classifier [47] is then applied to determine the annotation words for the test image. All the experiments were performed on a Xeon 2.8G server with 144GB memory. In the following sections, we will present results of individual experiments.

A. Near Duplicate Image Retrieval

In this experiment, we first perform retrieval experiments on the Oxford5K dataset, a benchmark dataset for image retrieval. We then expand the database with 1 million Flickr images (i.e., Flickr1M), and report retrieval results on this large image database. We also evaluate the scalability of the random seeding approach for keypoint quantization by varying the number of gallery images.

1) Dataset: The Oxford5K dataset [48] is comprised of 5,062 images, among which 55 images serve as queries. We extracted a total of 14.97 million keypoints from this dataset. The distractor Flickr dataset (i.e., Flickr1M) contains 999,771 images that are represented by 798 million keypoints. In the large scale image retrieval experiment, these two datasets are combined together, resulting in a dataset with a total of 1,004,833 images.

2) Evaluation: Following [33], [37], we use the mean average precision (mAP) as our evaluation metric. It computes average precision for each query based on its precision-recall curve, and averages the average precision over all the queries.

3) Experimental Results: We first report the experimental results on the Oxford5K dataset, followed by the experiments using the Oxford5K+Flickr1M dataset. Both experiments use the same set of 55 query images from the Oxford5K dataset.

a) Results for the Oxford5K dataset: We generate visual words either by clustering or by random seeding approach. We set the number of visual words to 10K, 100K, and 1 million. Table II reports the computation time for generating bag-of-words models by the clustering-based method and by the random seeding method. To facilitate the comparison of the two methods, we divide the running time into two parts, i.e., the time for constructing visual vocabulary (vocabulary construction time) and the time for mapping keypoints to visual words (quantization
time). Note that since the random seeding algorithm finds visual words by randomly sampling a set of keypoints, it is basically independent of vocabulary construction. We observe that not only it is computationally expensive to find visual words by clustering, it also takes considerably longer time for the clustering based method to map keypoints to visual words compared to random seeding approach. We also observe that the running time for the clustering-based method remains almost constant as the number of visual words increases, while it increases noticeably for the random seeding approach. This is because of the specific procedure used in Algorithm 1 for keypoint quantization, which takes $O(N \log m)$ to quantize $m$ keypoints into $N$ visual words. In contrast, it takes the clustering-based method $O(m \log N)$ to accomplish the same quantization task. The weaker dependency of the clustering method on $N$ makes its running time less sensitive to the number of visual words.

<table>
<thead>
<tr>
<th>Method</th>
<th>Vocabulary Construction</th>
<th>Quantization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vocabulary 10K</td>
<td>9,178s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>42,984s⁹</td>
</tr>
<tr>
<td></td>
<td>Vocabulary 100K</td>
<td>9,941s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>43,726s</td>
</tr>
<tr>
<td></td>
<td>Vocabulary 1M</td>
<td>10,405s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>44,128s</td>
</tr>
<tr>
<td>Random Seeding</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vocabulary Construction</td>
<td>5s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,425s</td>
</tr>
<tr>
<td></td>
<td>Quantization</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2,232s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8s</td>
</tr>
</tbody>
</table>

**TABLE II**

TIME FOR GENERATING BAG-OF-WORDS MODELS FOR THE OXFORD5K DATASET.

Table III summarizes the retrieval accuracy (mAP) and query time (i.e., time for running text retrieval to identify the gallery images that are visually similar to the query image) with different number of visual words. We observe that for both the methods, the retrieval accuracy is improved from $\sim 30\%$ to over $50\%$ as the number of visual words is increased from 10K to 1 million, which is consistent with the previous studies [33], [37]. In all the cases, we observe that the random seeding algorithm is more accurate than the clustering based method, although the gap is reduced as the number of visual words increases. We attribute this performance difference due to the inconsistent mapping of the clustering-based method. In particular, when the number of visual words increases, the performance gap between the two methods reduces.

⁹It could be significantly larger than the time reported in other studies due to differences in the NN search parameter settings and/or the library used. We also note that the quantization time is significantly larger than the vocabulary construction time since hierarchical clustering is used for vocabulary construction.
visual words is small, a keypoint is more likely to be mapped to a visual word even if they are separated by a large distance. The occurrence of inconsistent mapping in the clustering-based method is reduced as the number of visual words increases, leading to a smaller gap in retrieval accuracy. The query time result shows that the random seeding algorithm is more efficient in retrieving visually similar images than the clustering-based approach, although the advantage is diminished as the number of visual words increases. We attribute this phenomena to eliminating outlier keypoints. In particular, when the number of visual words is small, it is very likely for that a keypoint will be far away from all the visual words; those outlying keypoints are ignored by the proposed method in the quantization step, leading to efficient and effective retrieval. As the number of visual words increases, the number of outlier keypoints decreases, leading to a longer retrieval time. In fact, we found that about 84% keypoints were ignored when the number of visual words is 10K, and only 46% keypoints were ignored when the number of visual words is increased to 1 million\textsuperscript{10}.

To further examine the mapping criterion employed by the random seeding algorithm, we perform the retrieval experiments by using the random seeding algorithm but with the visual words identified by the clustering algorithm, referred to as clustering seeding. The retrieval performance of this approach is reported in the last row of Table III. We observe that the clustering seeding method performs worse than the random seeding approach under all settings. This result can be explained by the fact that some clusters have large variance (cover a widely spread set of keypoints), which contradicts the mapping criterion (range search that maps keypoints to a visual word only when their distance is smaller than a threshold) used by the random seeding method.

\textbf{b) Robustness:} We examine the sensitivity of the random seeding method w.r.t. radius $r$, the only parameter in the proposed keypoint quantization. Figure 4 shows how the retrieval accuracy of the random seeding algorithm is affected by $r$. The retrieval accuracy appears to be stable when $r$ is sufficiently large (i.e., $r \geq 0.5R$, where $R$ is the estimated average pairwise keypoint distance). We attribute the poor performance of small $r$ to the effect of eliminating too many outlier keypoints in the quantization step, leading to potentially poor bag-of-words model

\textsuperscript{10}We note that a random down-sampling of keypoints usually degrade the retrieval performance. For instance, we ran the same experiment with the clustering-based approach but with only 54% of keypoints randomly sampled from the original dataset. We observed that the mAP of the clustering based approach dropped from 0.52 to 0.47.
TABLE III

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP (query time)</th>
<th>Vocabulary 10K</th>
<th>Vocabulary 100K</th>
<th>Vocabulary 1M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering</td>
<td>0.28 (0.39s)</td>
<td>0.41 (0.19s)</td>
<td>0.52 (0.08s)</td>
<td></td>
</tr>
<tr>
<td>Random Seeding</td>
<td>0.35 (0.05s)</td>
<td>0.45 (0.06s)</td>
<td>0.54 (0.08s)</td>
<td></td>
</tr>
<tr>
<td>Cluster Seeding</td>
<td>0.32 (0.06s)</td>
<td>0.38 (0.06s)</td>
<td>0.50 (0.09s)</td>
<td></td>
</tr>
</tbody>
</table>

RETRIEVAL ACCURACY (mAP) AND QUERY TIME (SECOND) FOR THE NEAR DUPLICATE IMAGE RETRIEVAL ON THE OXFORD5K DATASET. “CLUSTER SEEDING” IN THE LAST ROW REFERS TO RANGE SEARCH QUANTIZATION USING CLUSTER CENTERS

for image representation.

![Graph showing mAP vs. Radius](Fig. 4. The effect of radius r on the retrieval accuracy of the random seeding method for keypoint quantization. R is the average pairwise keypoint distance.)

c) Oxford5K+Flickr1M Retrieval: We now test the random seeding method on the large data base Oxford5K+Flickr1M that combines the two data sets. Following the setup in [37], we use a vocabulary of 1 million visual words generated from the Oxford5K dataset. Table IV shows the retrieval accuracy, query time, and time for keypoint quantization for the Oxford5K+Flickr1M dataset. Note that we exclude the clustering time from the keypoint quantization time since the vocabulary has already been constructed from the Oxford5K dataset. We observe that the random seeding algorithm significantly outperforms the clustering method.
We also examine the scalability of the random seeding algorithm in keypoint quantization. In this experiment, we fix the vocabulary to be 1 million visual words generated from the Oxford5K dataset, and gradually add Flickr images into the Oxford5K dataset. Figure 5 reports the quantization time for both the methods with the numbers of gallery images. As the size of image database increases, the quantization time of the random seeding method increases at a slower rate than that of the clustering-based method, making it more suitable for handling large-scale dynamic image databases.

![Quantization Time Graph](image)

Fig. 5. The quantization time of the proposed random seeding algorithm and the clustering based approach. The vocabulary is fixed to be 1M visual words generated from the Oxford5K dataset. Background images from Flickr dataset are gradually added to the gallery. The vocabulary construction time is not shown here.

<table>
<thead>
<tr>
<th></th>
<th>mAP</th>
<th>Query Time (seconds)</th>
<th>Quantization Time (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering</td>
<td>0.33</td>
<td>0.84</td>
<td>591.2</td>
</tr>
<tr>
<td>Random Seeding</td>
<td>0.45</td>
<td>0.33</td>
<td>78.7</td>
</tr>
</tbody>
</table>

**TABLE IV**

RETRIEVAL ACCURACY (mAP), QUERY TIME, AND QUANTIZATION TIME (i.e., TIME FOR MAPPING KEYPOINTS TO VISUAL WORDS) FOR THE CLUSTERING-BASED METHOD AND THE RANDOM SEEDING METHOD.
B. Tattoo Image Retrieval

Body tattoos have been used by individuals for over 5,000 years to differentiate themselves from others. Historically, the practice of tattooing was limited to particular groups such as motorcycle bikers, soldiers, sailors, and members of criminal gangs. A recent study published in the Journal of the American Academy of Dermatology in 2006\(^{11}\) showed the rising popularity of tattoos amongst the younger section of the population; about 36% of Americans in the age group 18 to 29 have at least one tattoo.

Law enforcement agencies routinely photograph and catalog tattoo patterns for the purpose of identifying victims and convicts. The ANSI/NIST-ITL 1-2000 standard \(^{49}\) defines eight major classes (e.g., human face, animal, symbols, ...) and 80 subclasses for categorizing tattoos. Given a query image of a tattoo, a search is performed by matching the class label of the query tattoo with labels of tattoos in a database. This matching process based on human-assigned class labels is subjective and has poor retrieval performance. This motivated Lee et al. \(^{50}\) to develop an image retrieval system for automatic tattoo image matching. Given a query tattoo image, the system attempts to retrieve all the tattoo images from a database that are the photos of the same tattoo as the query.

1) Dataset: The database used here consists of 101,745 images, among which 61,745 are tattoo images that were provided by the Michigan Forensics Lab \(^{50}\); the remaining 40,000 images were randomly selected from the ESP game data set\(^{12}\). The purpose of adding images from the ESP game dataset is to verify the capability of the tattoo image retrieval system in distinguishing tattoo images from non-tattoo images. On average, about 108 keypoints are detected from each image, leading to a total of more than 10 million keypoints for the entire image collection.

2) Evaluation: To evaluate the retrieval performance of the tattoo image retrieval system, one duplicate image of the same tattoo is used as a query image to retrieve all of its visually similar image(s) in the database. For the tattoo image database used here, we used 995 tattoo images as queries, and manually identified the tattoo images in the database that are visually similar to each query. Cumulative matching characteristics (CMC) curve \(^{51}\), which accumulates the correct


\(^{12}\)http://www.gwap.com/gwap/gamesPreview/espgame/
number of retrieved images as the rank increases, was used as the evaluation metric. For a given rank position \( k \), the CMC score is computed as the percentage of queries whose matched images are found in the first \( k \) retrieved images. We use CMC curve, instead of precision & recall curve, because it is one of the most widely used evaluation metric in biometric and forensic analysis.

3) Experimental Results: Figure 6 shows the CMC curve of the tattoo retrieval system using the proposed random seeding algorithm and the clustering algorithm for keypoint quantization. Compared to the clustering approach, we observe an overall 5% improvement in the CMC score of the proposed random seeding algorithm. Figure 7 shows retrieved images for four queries.

![CMC Curve](image)

Fig. 6. The CMC curves of the tattoo retrieval system using the proposed random seeding algorithm and the clustering algorithm for keypoint quantization. The number of visual words is set to be 1 million.

Next, we evaluate the efficiency of the proposed scheme for keypoint quantization. It takes the proposed algorithm 3,476 seconds to quantize all the keypoints (i.e., 10.8 million keypoints in total) in the tattoo image collection into one million visual words (i.e., hyper-spheres). On the other hand, the running time of the clustering-based method is 31,305 seconds, which is about 9 times longer than the proposed random seeding algorithm. This result shows that the proposed scheme is not only more effective but also more efficient than the clustering-based algorithm for keypoint quantization. We also observe that both the methods have similar retrieval time, which is 0.05 second per query.
C. Automatic Image Annotation

We also validate the proposed method on an automatic image annotation task. We implement a $k$NN classification based approach for automatic image annotation, which was shown to be particularly effective for large-scale image annotation [47]. More specifically, to identify the annotation words for a given test image, we first search for the images in the training set that are visually similar to the test image. We then annotate the test image with the key words that are most frequently used by the first $k$ retrieved images.

1) Dataset: The ESP game dataset was used in our study. It consists of 109,720 images. On average, 117 keypoints were extracted for each image, leading to a total of 12.8 million keypoints for the entire database. We remove the rare annotation words that are used by less than ten images in the database. This gives us a vocabulary of 5,395 unique annotation keywords. On average, each image is annotated by 7.5 words. We randomly selected 1,000 images from the dataset to form the test set, and used the remaining 108,720 images for training.

2) Evaluation: We evaluate the performance of our system by comparing the predicted annotation words to the ground truth. The metric F1, which is widely used in multi-label learning [52], is used for evaluation. It is computed as follows: for a given test image, we first rank all the key words in the descending order of their occurrences in the first $k$ images that are retrieved by the system. The first $n_l$ words (with a random tie-breaker) in the ranking
list, denoted by \( P = \{ w_1, \ldots, w_{n_l} \} \), are the predicted annotation words for the test image. Let \( Q = \{ w'_1, \ldots, w'_{n_t} \} \) be the set of true annotations for the given test image. The \( F_1 \) measure for this test image is computed as the harmonic mean of precision \( p \) and recall \( r \), i.e., \( F_1 = \frac{2pr}{p+r} \). Given multiple test images, we first compute \( F_1 \) for each test image and then average the \( F_1 \) scores over all the test images as the final evaluation measure. Since the \( F_1 \) measure depends on the number of predicted annotation words \( n_l \), which is usually difficult to determine, we report the \( F_1 \) measures for different numbers of predicted annotation words.

3) Experimental Results: First, we evaluate the annotation accuracy of our system based on different methods for keypoint quantization. One of the key parameter used by the \( k \)NN based image annotation system is \( k \), i.e., the number of nearest neighbor images used for predicting the annotation words for a test image. We estimate this parameter by a cross validation procedure which uses 20\% of training images to validate the parameter \( k \). The optimal \( k \) found by cross validation for both quantization methods is 50. Figure 8 shows the \( F_1 \) values of the \( k \)NN based image annotation system that uses two different quantization methods, with \( n_l \) varied from 1 to 10. We observe that the proposed random seeding algorithm significantly outperforms the clustering based approach in the annotation accuracy.

Fig. 8. The results of a \( k \)NN based image annotation system using the proposed random seeding algorithm for keypoint quantization as well as the clustering based method. \( k \) is set to be 50 based on cross validation, and the number of visual words is set to be 1 million.

Finally, we evaluate the efficiency of the proposed random seeding algorithm for keypoint
quantization. It takes 3,376 seconds for the random seeding method to quantize all the keypoints, while the quantization time for the clustering based method is 45,008 seconds. This again confirmed that the proposed approach is more efficient than the clustering based method for large-scale keypoint quantization.

V. Conclusion

We have proposed an efficient algorithm for keypoint quantization needed in the bag-of-words model. It is designed to approximate the similarity measure based on the optimal partial matching between two sets of keypoints. The main idea is to quantize each keypoint into a sparse vector of real numbers by a set of randomly generated hyper-spheres. Our theoretical analysis shows that the similarity based on the proposed quantization algorithm provides an upper bound on the optimal partial matching based similarity. The proposed random seeding algorithm is more effective than the clustering-based approach because it deploys a more consistent criterion for keypoint quantization, which ensures that only similar keypoints are mapped to the same visual words. It is also more efficient than the clustering-based approach because (1) it avoids the high computational cost of large-scale clustering, and (2) its quantization time for mapping keypoints to visual words is sublinear in the number of keypoints, making it more suitable for large image databases than the clustering-based approach. Experimental results on large-scale near duplicate image retrieval, tattoo image retrieval, and automatic image annotation confirm that the proposed random seeding algorithm is more effective than the state-of-the-art clustering based approach for keypoint quantization. Besides efficacy, our empirical studies also indicate that the proposed scheme is more efficient than the clustering based method for keypoint quantization and in retrieving visually similar images. In the future, we plan to investigate different sampling methods to further improve the quantization.

APPENDIX I

PROOF OF LEMMA 1

Proof: We have

$$-d(X, Y) \leq \sum_{k=1}^{m} \max_{1 \leq i \leq n} (-|x_k - y_i|^2) + \sum_{k=1}^{n} \max_{1 \leq i \leq m} (-|y_k - x_i|^2)$$
\[
\begin{align*}
\leq & \sum_{k=1}^{m} \frac{1}{\lambda} \ln \left( \sum_{i=1}^{n} \exp \left( -\lambda |x_k - y_i|^2 \right) \right) + \sum_{k=1}^{m} \frac{1}{\lambda} \ln \left( \sum_{i=1}^{n} \exp \left( -\lambda |y_k - x_i|^2 \right) \right) \\
\leq & \frac{2}{\lambda} \sum_{k=1}^{m} \sum_{i=1}^{n} \exp \left( -\lambda \|x_i - y_j\|^2 - d_0 \right) - d_0 \\
\text{where } d_0 &= \min_{k,i} |x_k - y_i|^2,
\end{align*}
\]

\section*{Appendix II}

\section*{Proof of Lemma 2}

\textbf{Proof:} We have

\[
\int dz \hat{p}(z|X) \hat{p}(z|Y) = \frac{\mu^d}{mn \pi^d} \sum_{i=1}^{m} \sum_{j=1}^{n} \int dz \exp \left( -\mu |z - x_i|^2 - \mu |z - y_j|^2 \right)
\]

\[
= \frac{\mu^d}{mn \pi^d} \sum_{i=1}^{m} \sum_{j=1}^{n} \int dz \exp \left( -\mu |x_i - y_j|^2 - 2\mu |z - x_i + y_j|^2 \right)
\]

\[
= \frac{\mu^d}{mn [2\pi]^{d/2}} \sum_{i=1}^{m} \sum_{j=1}^{n} \exp \left( -\mu |x_i - y_j|^2 \right) = \frac{\mu^d}{mn [2\pi]^{d/2}} s(X, Y)
\]

\section*{Appendix III}

\section*{Proof of Lemma 3}

\textbf{Proof:} We compute the difference between \( \int dz \exp \left( -\mu \|z - x_i\|^2 + |z - y_j|^2 \right) \) and \( \int_{z \in Q} dz \exp \left( -\mu \|z - x_i\|^2 + |z - y_j|^2 \right) \), i.e.,

\[
\Delta_{i,j} = \int_{\gamma R}^{+\infty} r^{d-1} dr d\Omega \exp \left( -\mu \left( |z - x_i|^2 + |z - y_j|^2 \right) \right)
\]

\[
= \int_{\gamma R}^{+\infty} r^{d-1} dr d\Omega \exp \left( -\mu \left( |\gamma R \hat{z} - x_i| + (r - \gamma R) \right) \right)
\]

\[
\leq \exp \left( -2\mu (\gamma - 1)^2 R^2 \right) \int_{\gamma R}^{+\infty} r^{d-1} dr d\Omega \exp \left( -2\mu (r - \gamma R)^2 \right)
\]

\[
= \exp \left( -2\mu (\gamma - 1)^2 R^2 \right) \int_{0}^{+\infty} (r + \gamma R)^{d-1} dr d\Omega \exp \left( -2\mu r^2 \right)
\]

\[
= \exp \left( -2\mu (\gamma - 1)^2 R^2 \right) \left( \frac{\pi R^2}{2\mu} \right)^{1/2} \gamma R^{d-1}
\]

March 30, 2010

DRAFT

https://mc.manuscriptcentral.com/tpami-cs
Using $\Delta_{i,j}$ computed above, we have the result in the lemma by noticing that $|x_i - y_j|^2 \leq 4R^2$.

**APPENDIX IV**

**PROOF OF COROLLARY 1**

**Proof:** Since $\gamma \geq \max(4, \pi^{1/2}/[(2\mu)^{1/2}R])$, we have

$$\frac{|\hat{s}(X,Y) - \hat{s}_1(X,Y)|}{\hat{s}(X,Y)} \leq \exp(-\mu \gamma^2 R^2) (2\gamma R)^{d-1}$$

To ensure the difference is bounded by $\delta$, it is sufficient to ensure

$$-\mu \gamma^2 R^2 + (d - 1) \ln(2\gamma R) \leq \ln \delta$$

Using $\ln(\gamma R) \leq \gamma R - 1$, it suffices to show

$$-\mu \gamma^2 R^2 + (d - 1) \gamma R \leq \ln \delta$$

which implies

$$\gamma \geq \frac{1}{2\mu R} \left( (d - 1) + \sqrt{(d - 1)^2 + 4\mu \ln[1/\delta]} \right)$$

**APPENDIX V**

**PROOF OF THEOREM 2**

**Proof:** Since $0 \leq p(z|X)p(z|Y) \leq 1$, according to Hoeffding inequality [53], we have

$$\Pr \left( \left| \hat{E}_N[p(z|X)p(z|Y)] - E_z[p(z|X)p(z|Y)] \right| \geq \varepsilon \right) \leq 2 \exp(-2N\varepsilon^2)$$

where $\hat{E}_N[.]$ stands for the expectation over $N$ random samples $\{z_k\}_{k=1}^N$. By setting $\delta = 2 \exp(-2N\varepsilon^2)$, we have, with the probability $1 - \delta$, the following inequality:

$$|\hat{s}_2(X,Y) - \hat{s}_1(X,Y)| \leq \sqrt{\frac{1}{N} \ln \frac{2}{\delta}}$$
ACKNOWLEDGMENT

This work is supported in part by ARO grant No. W911NF-08-1-0403. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ARO. Anil Jain’s research is partially supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-2008-000-10008-0).

REFERENCES


Fengjie Li received his B.S. degree from Xi’dian University in 2002 and M.S. from Mississippi State University in 2006. He is currently a Ph.D student at the Department of Computer Science & Engineering at Michigan State University. His research interests includes Information Retrieval techniques and Machine Learning algorithms for large scale high dimensional data.
Wei Tong received the BS and MS degrees in Computer Science and Engineering from Northeastern University, China in 2000 and 2004 respectively. He is currently a Ph.D candidate in the Department of Computer Science and Engineering at Michigan State University. His research interests include machine learning, information retrieval, image annotation and computer vision.

Rong Jin is an Associate Professor in the Computer and Science Engineering Department at Michigan State University. He is working in the areas of statistical machine learning and its application to information retrieval. He has published over eighty conference and journal articles on the related topics. Dr. Jin holds a B.A. in Engineering from Tianjin University, an M.S. in Physics from Beijing University, and an M.S. and Ph.D. in Computer Science from Carnegie Mellon University. He received the NSF Career Award in 2006.