An Efficient Key Point Quantization Algorithm for Large Scale Image Retrieval

Fengjie Li, Wei Tong, Rong Jin, Anil K. Jain and Jung-Eun Lee
Department of Computer Science & Engineering
Michigan State University
East Lansing, Michigan 48824
{lifengji, tongwei, rongjin, jain, leejun11}@cse.msu.edu

ABSTRACT

We focus on the problem of large-scale near duplicate image retrieval. Recent studies have shown that local image features, often referred to as key points, are effective for near duplicate image retrieval. The most popular approach for key point based image matching is the clustering-based bag-of-words model. It maps each key point to a visual word in a code-book that is constructed by a clustering algorithm, and represents each image by a histogram of visual words. Despite its success, there are two main shortcomings of the clustering-based bag-of-words model: (i) it is computationally expensive to cluster millions of key points into thousands of visual words; (ii) there is no theoretical analysis on the performance of the bag-of-words model. We propose a new scheme for key point quantization that addresses these shortcomings. Instead of clustering, the proposed scheme quantizes each key point into a binary vector using a collection of randomly generated hyper-spheres, and a bag-of-words model is constructed based on such randomized quantization. Our theoretical analysis shows that the resulting image similarity provides an upper bound for the similarity based on the optimal partial matching between two sets of key points. Empirical study on a database of $100,000$ images shows that the proposed scheme is not only more efficient but also more effective than the clustering-based approach for near duplicate image retrieval.

1. INTRODUCTION

Although content-based image retrieval (CBIR) has been studied for years, the challenge of semantic gap, i.e. the gap between visual similarity and conceptual/perceptual relevance, has made it a much harder problem than most researchers originally expected [23]. However, recent studies have shown that near duplicate image retrieval [10], whose objective is to identify images with high visual similarity, can be solved effectively. In particular, studies [24, 29, 22, 13, 26, 10] have shown that local image features, e.g. SIFT descriptors [17], often referred to as key points, are effective

for near duplicate image retrieval and visual object recognition than global image features. The key idea is to extract salient local patches from an image, and represent each local patch by a multi-dimensional vector. As a result, each image is represented by a *bag-of-features* [4]. A number of algorithms [8, 18, 2] have been proposed to measure the similarity between two images based on their bag-of-features representations, including similarity based on the optimal partial matching between two sets of key points [30, 18, 2], pyramid kernel similarity [8], similarity based on the principal angle between two sets of key points [32], and similarity measure based on the match between two distributions [11, 20]. It has been shown [8, 18, 2] that despite its simplicity, the similarity based on the optimal partial matching performs well in comparison to the other similarity measurements. However, these optimal partial matching based approaches suffers from high computational complexity: given a query image, a linear scan is required to compute the similarity between the query image and every image in the database, which does not scale well to a large database with millions of images.

Among various approaches that have been proposed to improve the computational efficiency of optimal partial matching, the bag-of-words model [29] is the most popular and probably the most successful one. 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 onto its closest visual word, and represent each image by a histogram of visual words. A number of studies have shown promising performance of this approach for image retrieval [24, 29, 22, 13, 27, 26, 10].

Despite its success, most studies on the bag-of-words model suffer from the following drawbacks: (i) High computational cost of clustering when the number of clusters is very large. In our application, we need to cluster ten million key points into a million clusters. Although a number of algorithms [21, 24, 13, 27, 5, 15, 28] have been proposed for large-scale data clustering, they are still computationally expensive when handling millions or even billions of key points. (ii) Most of the studies on the bag-of-words model are focused on its empirical performance, and lack theoretical analysis.

In this work, we propose a new quantization scheme that explicitly addresses the shortcomings of the clustering-based approaches. The main idea is to quantize each key point by a set of randomly generated hyper-spheres. Each hyper-sphere is analogous to a cluster. Each key point is quantized by a bi-

nary bit, 1 when the key point is within the hyper-sphere and 0 otherwise. The bag-of-words representation for each image $\mathcal{I}$ is computed as a histogram over the hyper-spheres. To distinguish from the clustering based approach for key point quantization, we refer to the proposed scheme as a **random seeding approach** for key point quantization. Compared to the clustering-based method, the main advantage of the proposed random seeding approach is its computational efficiency, which is clearly demonstrated in our empirical study. In addition, we present a theoretical analysis that reveals the relationship between the image similarity based on the optimal partial matching and the bag-of-words model generated by the proposed scheme.

Based on the proposed scheme for key point quantization, we present a two-stage framework for large-scale image retrieval. In such a system, a bag-of-words model is first used to identify a small subset of images that are likely to be similar to a query image. A more precise similarity measure is then applied to re-rank the image subset to further improve the accuracy. We demonstrate the proposed system in the application domain of tattoo image retrieval. We emphasize that although the empirical study is carried out in a specific application domain, the proposed system is generic in that (1) it uses SIFT features for image representation, and (2) the quantization method is designed for the general purpose image indexing.

The rest of the paper is organized as follows: Section 2 reviews the related work on the bag-of-words model and methods for key point quantization; Section 3 describes the proposed scheme for key point quantization, together with a theoretical analysis that validates the proposed approach; Section 4 presents the empirical study of the proposed scheme in the application domain of tattoo image retrieval; Section 5 concludes this work and sets out future research directions.

2. RELATED WORK

One of the major challenges in exploring the bag-of-words model for image retrieval, as pointed out in [24], is how to efficiently construct a vocabulary with hundreds of visual words for a large image database. This requires developing efficient algorithms for large-scale data clustering. Various approaches have been explored for large-scale data clustering, including flat K-means clustering [29], hierarchical clustering [22], and clustering based on approximate nearest neighbor search [21, 24, 13, 27, 10, 16]. However, these approaches are still computationally expensive when handling tens of millions key points, as will be revealed by our study.

In addition to the clustering approach, several random algorithms have been proposed for key point quantization [6, 3]. The key idea of these approaches is to extend Locality Sensitive Hashing [5] to encode the key points by a series of hashing functions. The proposed approach differs from the existing randomized algorithms for key point quantization in its quantization procedure. Furthermore, unlike the existing key point quantization methods that are mainly focused on empirical investigation, we present in detail a theoretical analysis that verifies the proposed randomized approach. Finally, our empirical study shows that the proposed approach is not only more efficient but also more effective for image retrieval than the clustering approach.

3. RANDOM SEEDING FOR KEY POINT QUANTIZATION

In this section, we first present the basic algorithm of random seeding algorithm for key point quantization. We then present an analysis to reveal the relationship between the image similarity based on the optimal partial matching and the bag-of-words model.

Let $\mathcal{D} = (\mathcal{I}_1, \ldots, \mathcal{I}_n)$ be a collection of n images. Each image $\mathcal{I}_i$ is represented by a set of n_i key points, denoted by $X_i = (x_i^1, \ldots, x_i^{n_i})$ where each key point $x_i^j \in \mathbb{R}^d$ is a vector of d dimensions. The first step of the proposed scheme is to randomly sample m key points from all the key points detected in the image collection $\mathcal{D}$, denoted by $c_1, \ldots, c_m$. A hyper-sphere $\mathcal{B}_i$ is created centered at each sampled key point c_i, i.e., $\mathcal{B}_i = \{z \in \mathbb{R}^d : |z - c_i|_2 \leq r\}$, where r is a predefined constant that is derived from the average distance between any two key points in image collection $\mathcal{D}$. Using the hyper-spheres $\{\mathcal{B}_i\}_{i=1}^m$, we quantize each key point x into a binary vector $\vec{b}(x) = (b_1(x), \ldots, b_m(x))$, where each element $b_i(x)$ is 1 if $x \in \mathcal{B}_i$ and 0 otherwise. The bag-of-words representation for image $\mathcal{I}_i$, denoted by $\vec{b}(\mathcal{I}_i) = (b_1(X_i), \ldots, b_m(X_i))$, is computed by adding the binary vectors of all the key points in image $\mathcal{I}_i$, i.e., $b_k(X_i) = \sum_{j=1}^{n_i} b_k(x_i^j)$. Figure 1 shows the basic idea of the proposed scheme for key point quantization and the resulting bag-of-words representation.

Compared to the clustering-based method for key point quantization, the proposed random seeding method is computationally more efficient. This is because unlike clustering-based methods, where the cluster centers are identified by an iterative method, the centers of hyper-spheres used in the proposed scheme are randomly generated. In addition, the proposed approach relies on range search for key point quantization, which is usually more efficient than the nearest neighbor search used in the clustering-based approaches. Besides its computational advantage, the proposed scheme allows for partial matching between key points, because two randomly generated hyper-spheres are allowed to overlap.

We devote the remaining part of this section to a theoretical analysis of the random seeding approach. Our analysis starts with the similarity between two sets of key points that is based on the optimal partial matching. We first show that the partial match based similarity can be upper bounded by a smooth similarity function, and then we derive a scheme of key point quantization that approximates this upper bound similarity function with a small error.

3.1 Distance Measure Based on Optimal Partial Matching

Let the two images $\mathcal{I}_x$ and $\mathcal{I}_y$ be represented by the corresponding sets of key points, 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 measurements that have been proposed, the similarity based on the optimal partial matching [8, 18, 2] is probably the most intuitive one, and has yielded the state-of-the-art performance for both image retrieval and object recognition. It computes the distance between X and Y by the optimal partial matching. 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 distance between X and Y for the given mappings π^1 and π^2, denoted
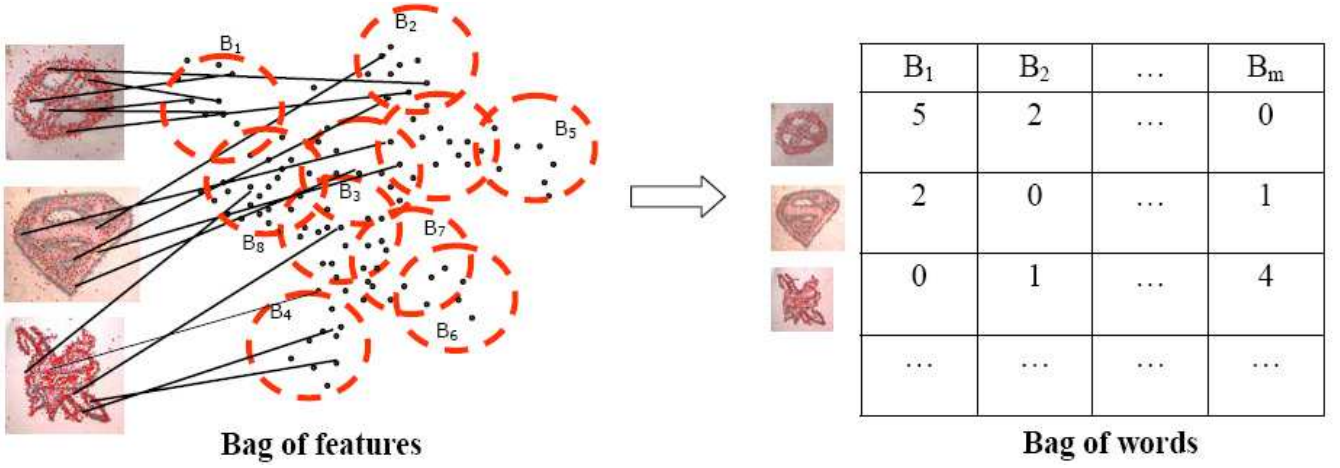
Figure 1: Illustration of the proposed random seeding approach for key point quantization. Each key point is quantized by a collection of randomly generated hyper-spheres $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_m$. The bag-of-words representation is computed as a histogram over the hyper-spheres $\{\mathcal{B}_i\}_{i=1}^m$.

by $d(X, Y; \pi^1, \pi^2)$, is computed as

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

where π_k^1 indicates the index of the point in set X that y_k is mapped to; similarly, π_k^2 indicates the point in set Y that x_k is mapped to. Finally, the distance between X and Y, denoted by $d(X, Y)$, is obtained by minimizing $d(X, Y; \pi^1, \pi^2)$ over the mappings π^1 and π^2. With proper normalization on cardinality, we have:

$$d(X, Y) = \frac{1}{mn} \min_{\pi^1, \pi^2} \sum_{k=1}^m \|x_k - y_{\pi_k^2}\|_2^2 + \sum_{k=1}^n \|y_k - x_{\pi_k^1}\|_2^2. \quad (2)$$

3.2 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 (2) because the distance function $d(X, Y)$ is non-smooth and is defined implicitly through the maximization of the mappings π^1 and π^2. To address this difficulty, we approximate the distance function in (2) by a smooth similarity function, as shown in the following lemma.

LEMMA 1. *For any $\lambda > 0$, we have*

$$-d(X, Y) + \frac{m+n}{mn} \le \frac{2}{\lambda mn} \sum_{i=1}^m \sum_{j=1}^n \exp(-\lambda |x_i - y_j|_2^2). \quad (3)$$

We omit the proof due to the space limitation.

Based on Lemma 1, we define the similarity between X and Y by $s(X, Y)$ as follows

$$s(X, Y) = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \exp\left(-\lambda \|x_i - y_j\|_2^2\right). \quad (4)$$

According to Lemma 1, any two images X and Y are separated by a large distance only if they have a small similarity $s(X, Y)$. Hence, instead of directly computing the distance $d(X, Y)$ between a query image and a database image, we divide the search for visually similar images into two steps.

In the first step, we compute the similarity $s(X, Y)$ between the query image and the images in the database; in the second step, only for the images with large similarity $s(X, Y)$, distance to the given query image is computed explicitly. In the rest of this section, we will show that the similarity $s(X, Y)$ can be computed efficiently by a random seeding algorithm for key point quantization.

3.3 Random Seeding Algorithm

We now show that the similarity measure $s(X, Y)$ defined in (4) can be computed efficiently by a key point 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 computation over the finite number of samples leads to the bag-of-words model with efficient computational algorithms.

3.3.1 Probabilistic Interpretation of Similarity Measure $s(X, Y)$

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 key points. Given a set of key points $X = (x_1, \ldots, x_m)$, we assume that they are sampled from an unknown distribution, denoted by $p(z|X)$. Using the kernel density function estimation [7], the estimate of the unknown distribution $p(z|X)$, denoted by $\widehat{p}(z|X)$, is computed as follows

$$\widehat{p}(z|X) = \frac{1}{m} \sum_{i=1}^m \frac{\mu^{d/2}}{[\pi]^{d/2}} \exp(-\mu|z - x_i|_2^2), \quad (5)$$

where μ specifies the kernel width. the estimate of $\widehat{p}(z|Y)$ is computed similarly.

Using the estimates $\widehat{p}(z|X)$ and $\widehat{p}(z|Y)$, we compute the similarity between X and Y as follows

$$\widehat{s}(X, Y) = \int dz \widehat{p}(z|X) \widehat{p}(z|Y). \quad (6)$$

The following lemma shows the relationship between $s(X, Y)$

in (4) and $\widehat{s}(X,Y)$ in (6).

LEMMA 2. *The following relationship holds for $s(X,Y)$ in (4) and $\widehat{s}(X,Y)$ in (6) if $\mu = 2\lambda$*

$$\widehat{s}(X,Y;2\lambda) = \frac{\mu^{d/2}}{[2\pi]^{d/2}} s(X,Y;\lambda). \qquad (7)$$

The lemma follows directly from the definition.

As revealed in Lemma 2, except for the constant factor $[\mu/(2\pi)]^{d/2}$, the two similarity measures $s(X,Y)$ and $\widehat{s}(X,Y)$ are equivalent. In the following analysis, we will use $\widehat{s}(X,Y)$ for similarity measure. The key advantage of using $\widehat{s}(X,Y)$, instead of $s(X,Y)$, is that we can view $\widehat{s}(X,Y)$ as an expectation of $p(z|X)p(z|Y)$ over hidden variable z, which is crucial for deriving a scheme for key point quantization.

3.3.2 Similarity Measure as Expectation

To view $\widehat{s}(X,Y)$ as an expectation, i.e., $E_z[p(z|X)p(z|Y)]$, we need to introduce an appropriate distribution for z. Note that in (6), $\widehat{s}(X,Y)$ is defined as an integration over the unbounded space of z. Hence, it is inappropriate to define a uniform distribution for z, which is sometimes referred to as inappropriate prior in Bayesian analysis. To address this difficulty, we assume that $|x|_2 \leq R$ for any key point x that is detected for images in a given database. We introduce a domain Q for z that is defined as follows

$$Q = \{z : |z|_2 \leq \gamma R\},$$

where $\gamma \geq 1$ is a constant that will be determined empirically. We then approximate $\widehat{s}(X,Y)$ by $\widehat{s}_1(X,Y)$, which is defined as an integration over the domain Q, i.e.,

$$\widehat{s}_1(X,Y) = \int_{z \in Q} dz \widehat{p}(z|X)\widehat{p}(z|Y). \qquad (8)$$

The following lemma gives a bound on the difference between $\widehat{s}(X,Y)$ and $\widehat{s}_1(X,Y)$.

LEMMA 3. *The following inequality holds for any $\gamma > 0$*

$$\frac{|\widehat{s}(X,Y) - \widehat{s}_1(X,Y)|}{\widehat{s}(X,Y)} \leq$$

$$\exp(-2\mu\gamma(\gamma - 2)R^2)\left(\left[\frac{\pi}{2\mu}\right]^{1/2} + \gamma R\right)^{d-1} \qquad (9)$$

The proof is omitted due to the space limitation.

As indicated by Lemma 3, a large value of γ will lead to a small value for $|\widehat{s}(X,Y) - \widehat{s}_1(X,Y)|/\widehat{s}(X,Y)$, implying that $\widehat{s}_1(X,Y)$ is close to $\widehat{s}(X,Y)$. This is shown by the following corollary.

COROLLARY 1. *If*

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

we have

$$\frac{|\widehat{s}(X,Y) - \widehat{s}_1(X,Y)|}{\widehat{s}(X,Y)} \leq \delta$$

The proof is omitted due to the space limitation.

With the above analysis, we can now focus on the similarity measure $\widehat{s}_1(X,Y)$ since it is close to $\widehat{s}(X,Y)$ with

sufficiently large γ. Since the domain of z in $\widehat{s}_1(X,Y)$ is bounded, we can introduce a uniform distribution for z, i.e.,

$$q(z) = I(z \in Q)/vol(Q),$$

where $vol(Q)$ 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 $\widehat{s}_1(X,Y)$ as the expectation of $\widehat{p}(z|X)\widehat{p}(z|Y)$ over random variable z, i.e.,

$$\widehat{s}_1(X,Y) = vol(Q)E_z[\widehat{p}(z|X)\widehat{p}(z|Y)] \propto \int_z q(z)\widehat{p}(z|X)\widehat{p}(z|Y) \quad (10)$$

In the remaining analysis, we will drop the factor $vol(Q)$ for the sake of simplicity.

3.3.3 Key Point Quantization by Random Samples

In the next step, we further approximate the expectation interpretation in (10) by replacing the distribution $q(z)$ with a finite number of samples. We denote by $z_1, z_2, \ldots, z_N$ N empirical samples randomly drawn from the distribution $q(z)$. Using the empirical samples $\{z_k\}_{k=1}^N$, we approximate $\widehat{s}_1(X,Y)$ by $\widehat{s}_2(X,Y)$ that is defined as follows

$$\widehat{s}_2(X,Y) = \qquad (11)$$

$$= \frac{\mu^d}{mn\pi^d} \sum_{i=1}^m \sum_{j=1}^n \frac{1}{N} \sum_{k=1}^N \exp(-\mu|z_k - x_i|_2^2 - \mu|z_k - y_j|_2^2)$$

$$\leq \frac{\mu^d}{mnN\pi^d} \sum_{k=1}^N \left(\sum_{i=1}^m \exp(-\mu|z_k - x_i|_2^2)\right)\left(\sum_{j=1}^n \exp(-\mu|z_k - y_j|_2^2)\right)$$

It is important to note that, in the expression of $\widehat{s}_2(X,Y)$, by replacing the expectation over the distribution $q(z)$ with the average over the empirical samples, we essentially represent each image $X = (x_1, \ldots, x_m)$ by a vector

$$\vec{f}(X) = (f_1(X), \ldots, f_N(X)) \qquad (12)$$

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

$$f_k(X) = \frac{1}{m} \sum_{i=1}^m \exp(-\mu|z_k - x_i|_2^2).$$

The theorem below shows that with a sufficiently large number of samplings (i.e., N is sufficiently large), $\widehat{s}_1(X,Y)$ and $\widehat{s}_2(X,Y)$ will differ only by a small value.

THEOREM 2. *With probability $1 - \delta$, we have*

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

Theorem 2 directly follows the Hoeffding inequality [9].

The above analysis only applies to two images. We furthermore generalize it to a collection of images. We denote by $\mathcal{D} = (X_1, \ldots, X_T)$ the collection of T images where $X_i = (x_i^1, \ldots, x_i^{n_i})$ is a collection of key points. We have the following corollary showing that for any two images X_i and X_j in $\mathcal{D}$, their similarities $\widehat{s}_1(X_i, X_j)$ and $\widehat{s}_2(X_i, X_j)$ are close to each other.

COROLLARY 3. *Given a collection $\mathcal{D}$ of T images, with probability $1 - \delta$, the following inequality holds for any two images X_i and X_j in $\mathcal{D}$*

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

The above corollary is simply proved by a union bound.

Although (12) provides a vector representation for each image X (i.e., a set of key points), 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 now represent each image X by a vector

$$\vec{h}(X) = (h_1(X), \ldots, h_N(X)), \qquad (13)$$

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

$$h_k(X) = \sum_{i=1}^{m} \sum_{j=1}^{m} \max(\exp(-\mu|z_k - y_j|_2^2) - \eta, 0) \qquad (14)$$

Accordingly, we introduce the similarity $\widehat{s}_3(X, Y)$ to approximate $\widehat{s}_2(X, Y)$ as follows:

$$\widehat{s}_3(X, Y) = \left[\frac{\mu}{\pi}\right]^d \frac{1}{mnN} \sum_{k=1}^{N} h_k(X) h_k(Y) \qquad (15)$$

The following proposition shows that $\widehat{s}_3(X, Y)$ is close to $\widehat{s}_2(X, Y)$ if the threshold η is small.

PROPOSITION 1. *For any $\eta > 0$, we have*

$$|\widehat{s}_3(X, Y) - \widehat{s}_2(X, Y)| \leq \eta(\mu/\pi)^d$$

The advantage of using $\widehat{s}_3(X, Y)$ is that to quantize key point x_i, we will only consider the subset of $z_1, \ldots, z_N$ that is in the range of $\frac{1}{\mu} \ln[1/\eta]$ of x_i. Any sample z_k whose distance to x_i is larger than $\frac{1}{\mu} \ln[1/\eta]$ will not make any contribution to the quantization of x_i, and therefore will be ignored. The range search can be carried out efficiently by using any of the existing techniques such as approximate kd-trees or Locality Sensitive Hashing (LSH).

3.4 Implementation

To generate the bag-of-words model $\vec{h}(X)$ in (13), the first step is to sample z_i within a hyper-sphere. In practice, sampling z_i uniformly within a hyper sphere is a nontrivial problem. In our implementation, we acquire the samples $z_1, \ldots, z_N$ by first sampling N points from all the key points in the image database $\mathcal{D}$, and then scale the sampled key points by a factor γ. This practice essentially assumes that all the key points 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 γ to be large, $\gamma = 1$ does yield desirable performance in our empirical study.

The second step of generating the bag-of-words model is to compute element $h_k(X)$ in the bag-of-words model $\vec{h}(X)$. Note that $h_k(X)$ defined in (14) 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 be $I(\exp(-\mu|z_k - y_j|_2^2) \geq \eta)$, 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 $I(|z_k - y_j|_2 \leq r)$, where $r = \sqrt{[\ln(1/\eta)]/\mu}$. In practice, we set $r = 0.75\bar{r}$, where $\bar{r}$ is the average distance between any two key points in image collection $\mathcal{D}$. We estimate $\bar{r}$ by randomly sampling $10,000$ key point pairs from the entire collection.

To efficiently compute the bag-of-words model, we need to efficiently compute $I(|z_k - y_j|_2 \leq r)$, which requires identifying the subset of key points in the image collection $\mathcal{D}$ that are within the range r of each center z_k. To this end, we use the Fast Library for Approximate Nearest Neighbors (FLANN)[1], which implements the randomized kd-trees [28], for efficient range search.

4. EXPERIMENTS

We aim to verify both the efficiency and the efficacy of the proposed scheme for key point quantization in the domain of large-scale tattoo image retrieval. We choose the tattoo image retrieval for evaluation because the relevance judgments of retrieved images can be determined objectively. In particular, a retrieved tattoo image is judged as relevant to a query tattoo image only when both images contain the same tattoo symbol. In addition, the large size of the tattoo image database used in our study, which contains more than 100K images, allows us to evaluate both the efficiency and efficacy of proposed approach for large-scale image retrieval.

Below we first describe the task of tattoo image retrieval, which is an application of near duplicate image retrieval in law enforcement. We then present the experimental setup, including the image data set and the evaluation protocol used in our study. We finally present the comparative study using both the clustering-based method and the proposed scheme for key point quantization.

4.1 Introduction to Tattoo Image Retrieval

People have used tattoos 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 [2] showed the rising popularity of tattoos amongst the younger section of the population, i.e., about 36% of Americans in the age group 18 to 29 have at least one tattoo.

Tattoos engraved on the human body have been successfully used to assist in human identification in forensics and law enforcement applications. For instance, tattoos were used to identify victims of the 9/11 terrorist attacks and the Asian tsunami in 2004 [14]. Criminal identification using tattoos is another important application. Law enforcement agencies routinely photograph and catalog tattoo patterns for the purpose of identifying victims and convicts (who often use aliases). As an example, Los Angeles county police department maintains a database of about 2 million tattoo images. The ANSI/NIST-ITL 1-2000 standard [1] defines eight major classes (e.g., human face, animal, and symbols) and 80 subclasses for categorizing tattoos. Manual searches are performed by matching the class label of a query tattoo with labels of tattoos in a database. This matching pro-

---

[1] http://www.cs.ubc.ca/~mariusm/index.php/FLANN/FLANN

[2] Tattoo Facts and Statistics, Oct. 2006, http://www.vanishingtattoo.com/tattoo_facts.htm.

cess based on human-assigned class labels is subjective, has limited performance, and is very time-consuming. This motivated us to develop a image retrieval system for automated tattoo image matching. Given a query tattoo image, our system identifies the tattoo images from a large database that are the photos of the same tattoo as the query.

Although tattoo image retrieval can be classified as a near duplicate image retrieval problem since the objective is to find the same tattoo image in a given database, we emphasize that it is a significantly more challenging problem than near-duplicate image detection [10, 31, 16]. Note that there are two main reasons why a law enforcement database may have multiple images of a tattoo: (i) a person with a tattoo has been arrested more than once; every time an arrest is made, the tattoo image is captured. (ii) individuals belonging to the same gang tend to share the same tattoo symbol. Because of these reasons, we usually observe large variance in the visual appearance of the images of the same tattoo, making it a challenging image retrieval problem.

4.2 Experimental Setup

Data set.
The database used here consists of $101,745$ images, among which $61,745$ are tattoo images that were provided by the authors of [12] and the remaining $40,000$ images were randomly selected from the ESP game data set [3]. The purpose of adding images from the ESP game data set is to verify the capability of the developed system in distinguishing tattoo images from non-tattoo images. On average, about 100 key points are detected from each image. So we have more than 10 million key points for the entire image collection.

Evaluation methodology.
To evaluate the retrieval performance of the system for tattoo image retrieval, one duplicate image of the same tattoo is used as a query image to retrieve its visually similar image(s) in the database. In the database of tattoo images, we have selected 995 tattoo images as queries, and have manually identified the tattoo images in the database that are visually similar to each query image. Retrieved images are ranked in the descending order of their similarity, and cumulative matching characteristics (CMC) curve [19] is adopted as the evaluation metric in our study. This metric accumulates the correct number of retrieved images as the rank increases. For a given rank position k, its CMC score is computed as the percentage of queries whose matched images are found in the first k retrieved images. The CMC score is similar to recall, a common metric used in information retrieval. We use CMC curve, instead of precision & recall curve, because CMC curve is the most widely used evaluation metric in biometric and forensic analysis.

Implementation of the Tattoo Image Retrieval System.
Similar to many implementations of the bag-of-words model for image retrieval [24], we have a two-stage retrieval system. In the first stage of the retrieval system, a bag-of-words model is obtained for each image, and a text retrieval system is used to rank all the images in the database for a given query. In our implementation, the Lemur text re-
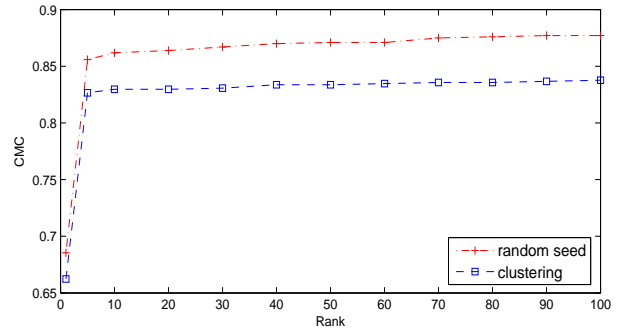
Figure 2: The CMC curves of the two-stage retrieval system using the proposed random seed algorithm and the clustering algorithm for key point quantization. The number of visual words is set to be 1 million. The number of candidate images that will be re-ranked in the second stage is set to be 1000.

trieval system [4] with Okapi retrieval model [25] is used to compute the similarity between a query image and images in the database. The objective of the first stage is to identify a small number of images that are likely to share a large visual similarity with the query image. In our experiment, the first 1000 most similar images are identified as the candidate matches for a query image.

The second stage of the system re-ranks the 1000 similar images obtained from the previous stage. We use the image matching algorithm presented in [12] to recompute the similarity between the query image and the $1,000$ image candidates. Unlike the bag-of-words model, this matching algorithm [12] takes into account the geometric relationship among key points when computing the matching score between two images, which was shown to be effective for tattoo image retrieval. More detailed description of the tattoo image matching algorithm can be found in [12].

To evaluate the effectiveness of the proposed quantization scheme, we compare it to the clustering-based approach, which is the most popular approach for key point quantization. To quantize key points, the clustering-based approach first performs hierarchical k-means clustering over all the key points, and then quantizes each key point to the closest cluster center. Random kd-trees [21] are used to efficiently identify the nearest cluster center for each key point. Based on our experience, we set the number of clusters to be 1 million for both the proposed scheme and the clustering-based approach for key point quantization.

4.3 Experimental Results

All the experiments are performed on a dual core Dell machine with 16G memory. We first report the retrieval accuracy, followed by the computational efficiency of key point quantization.

Retrieval accuracy.
Figure 2 shows the CMC curve of the two-stage retrieval system using the proposed random seed algorithm and the clustering algorithm for key point quantization, respectively.

Figure 3: Example query images (one per row) and the top ten images retrieved by the proposed two-stage system using the random seed algorithm for key point quantization. Note that for the last two queries, more than one image with similar tattoo were retrieved.

Compared to the clustering approach, we observe about 3% improvement in the CMC score of the proposed random seed algorithm.

To further confirm our result, we vary the number of images that are retrieved in the first stage by a text search engine that are then re-ranked by a more accurate image similarity measure in the second stage. Figure 4 shows the CMC curves when the number of re-ranked images is set to 100 and 500. Again, the proposed algorithm outperforms the clustering method in both cases.

Efficiency of key point quantization.

To evaluate the efficiency of the proposed random seed algorithm for key point quantization, we measure its running time. To quantize all the key points in the image collection (i.e., 10 million key points in total) into a million visual words, the overall running time of the proposed algorithm is $3,333$ seconds. The clustering algorithm takes $25,669$ seconds to quantize the same number of key points into one million clusters, which is about 8 times slower than the proposed algorithm. This result clearly indicates that the proposed key point quantization scheme is significantly more efficient than the clustering algorithm, making it scalable to large image databases with millions of images. We also observe that both quantization methods result in similar response times for retrieving the matched images for each query, about one second per query. Finally, we also conducted experiments with 100K visual words and observed similar results.

5. CONCLUSIONS

We have proposed an efficient algorithm for key point quantization. The main idea is to quantize each key point into a sparse vector of real numbers by a set of randomly generated hyper-spheres. The proposed algorithm is com-

putationally more efficient compared to the clustering-based key point quantization in that (i) the center of each hyper-sphere is randomly determined, avoiding an iterative search for cluster centers, and (ii) the radius of each hyper-sphere is fixed to be constant, avoiding nearest neighbor search. In addition, our theoretical analysis shows that the similarity based on the proposed quantization algorithm provides an upper bound for the optimal partial matching based similarity. Empirical study on a data set of $100,000$ images with more than ten million key points shows that (i) our approach yields better retrieval performance than the clustering-based approach, and (ii) it reduces the running time of key point quantization by a factor of 8. In the future, we plan to investigate other approaches for key point quantization, such as random projection.

6. ACKNOWLEDGEMENTS

7. REFERENCES

[1] *ANSI/NIST-ITL 1-2007 standard: Data Format for the Interchange of Fingerprint, Facial, & Other Biometric Information*. 2007.

[2] S. Boughhorbel, J.-P. Tarel, and F. Fleuret. Non-mercer kernels for svm object recognition. In *British Machine Vision Conference*, 2004.

[3] O. Chum, J. Philbin, and A. Zisserman. Near duplicate image detection: min-hash and tf-idf weighting. In *Proceedings of the British Machine Vision Conference*, 2008.

[4] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints.
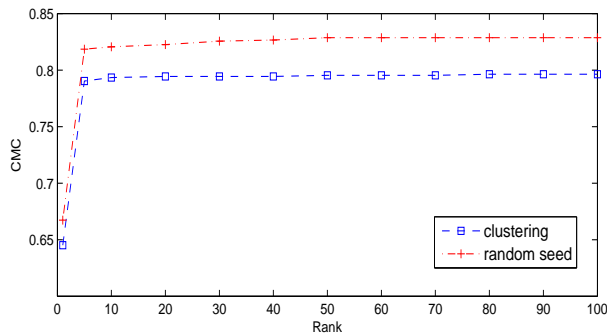
(a) The number of re-ranked images = 100
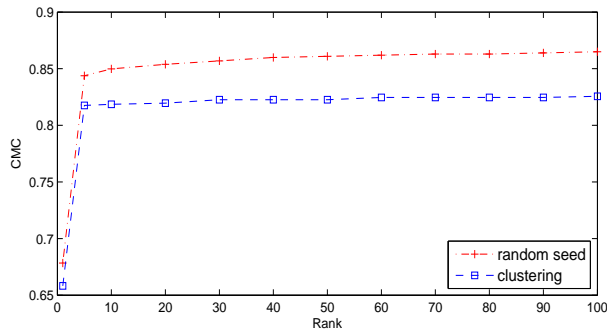
(b) The number of re-ranked images = 500

Figure 4: The CMC curves of the two-stage retrieval system using the proposed random seed algorithm and the clustering algorithm for key point quantization. The number of visual words is set to be 1 million. The number of candidate images re-ranked in the second stage is set to be 100 in (a) and 500 in (b).

In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[5] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, 2004.

[6] W. Dong, Z. Wang, M. Charikar, and K. Li. Efficiently matching sets of features with random histograms. In *Proceeding of ACM Multimedia*, pages 179–188, 2008.

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.

[8] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *J. Mach. Learn. Res.*, 8:725–760, 2007.

[9] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[10] Y. Ke, R. Sukthankar, and L. Huston. Efficient near-duplicate detection and sub-image retrieval. In *Proceeding of ACM Multimedia*, pages 869–876, 2004.

[11] R. Kondor and T. Jebara. A kernel between sets of vectors. In *Proceedings of the International Conference on Machine Learning*, pages 361–368, 2003.

[12] J.-E. Lee, A. K. Jain, and R. Jin. Scars, marks and tattoos (SMT): Soft biometric for suspect and victim identification. In *Biometric Consortium Conference and Technology Expo*, 2008.

[13] V. Lepetit, P. Lagger, and P. Fua. Randomized trees for real-time keypoint recognition. In *CVPR*, pages 775–781, 2005.

[14] E. Lipton and J. Glanz. Limits of DNA research pushed to identify the dead of sept. 11. *New York Times*, April 2002.

[15] T. Liu, A. Moore, A. Gray, and K. Yang. An investigation of practical approximate nearest neighbor algorithms. In *Neural Information Processing Systems*, pages 825–832, 2004.

[16] T. Liu, C. Rosenberg, and H. Rowley. Clustering billions of images with large scale nearest neighbor search. In *IEEE Workshop on Applications of Computer Vision (WACV)*, page 28, 2007.

[17] D. Lowe. Distinctive image features form scale-invariant keypoints. In *International Journal of Computer Vision*, volume 60, pages 91–110, 2004.

[18] S. Lyu. Mercer kernels for object recognition with local features. In *CVPR*, pages 223–229, 2005.

[19] H. Moon and P. J. Phillips. Computational and performance aspects of pca-based face recognition algorithms. *Perception*, 30:303–321, 2001.

[20] P. Moreno, P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Neural Information Processing Systems*, 2003.

[21] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISSAPP (1)*, pages 331–340, 2009.

[22] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, pages 2161–2168, 2006.

[23] T. Pavlidis. Limitations of cbir. In *ICPR*, 2008.

[24] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, pages 1–8, 2007.

[25] S. E. Robertson, S. Walker, and M. Hancock-Beaulieu. Okapi at trec-7. In *Proceedings of the Seventh Text REtrieval Conference*, 1998.

[26] G. Shakhnarovich, T. Darrell, and P. Indyk. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*. MIT Press, 2006.

[27] C. Silpa-Anan and R. Hartley. Localization using an imagemap. In *Proceedings of the 2004 Australasian Conference on Robotics & Automation*, 2004.

[28] C. Silpa-Anan and R. Hartley. Optimised kd-trees for fast image descriptor matching. In *CVPR*, pages 1–8, 2008.

[29] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[30] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *CVPR*, pages 257–264, 2003.

[31] B. Wang, Z. Li, M. Li, and W.-Y. Ma. Large-scale duplicate detection for web image search. In *ICME*, pages 353–356, 2006.

[32] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, 4:913–931, 2003.