Understanding bag-of-words model: A statistical framework

Article *in* International Journal of Machine Learning and Cybernetics · December 2010

DOI: 10.1007/s13042-010-0001-0

3 authors, including:

Rong Jin
Louisiana State University Health Sciences Center New Orleans
227 PUBLICATIONS **8,926** CITATIONS
SEE PROFILE

Zhi-Hua Zhou
Nanjing University
441 PUBLICATIONS **25,101** CITATIONS
SEE PROFILE

Noname manuscript No.
(will be inserted by the editor)

Understanding Bag-of-Words Model: A Statistical Framework

Yin Zhang · Rong Jin · Zhi-Hua Zhou

Received: date / Accepted: date

Abstract The bag-of-words model is one of the most popular representation methods for object categorization. The key idea is to quantize each extracted key point into one of visual words, and then represent each image by a histogram of the visual words. For this purpose, a clustering algorithm (*e.g.*, K-means), is generally used for generating the visual words. Although a number of studies have shown encouraging results of the bag-of-words representation for object categorization, theoretical studies on properties of the bag-of-words model is almost untouched, possibly due to the difficulty introduced by using a heuristic clustering process. In this paper, we present a statistical framework which generalizes the bag-of-words representation. In this framework, the visual words are generated by a statistical process rather than using a clustering algorithm, while the empirical performance is competitive to clustering-based method. A theoretical analysis based on statistical consistency is presented for the proposed framework. Moreover, based on the framework we developed two algorithms which do not rely on clustering, while achieving competitive performance in object categorization when compared to clustering-based bag-of-words representations.

Keywords Object recognition · Bag of words model · Rademacher complexity

1 Introduction

Inspired by the success of text categorization (Joachims, 1998; McCallum and Nigam, 1998), a bag-of-words representation becomes one of the most popular methods for repre-

Y. Zhang
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
E-mail: zhangyin@lamda.nju.edu.cn

R. Jin
Department of Computer Science & Engineering
Michigan State University, East Lansing, MI 48824
E-mail: rongjin@cse.msu.edu

Z.-H. Zhou
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
E-mail: zhouzh@lamda.nju.edu.cn

senting image content and has been successfully applied to object categorization. In a typical bag-of-words representation, “interesting” local patches are first identified from an image, either by densely sampling (Nowak et al., 2006; Winn et al., 2005) or by a interest point detector (Lowe, 2004). These local patches, represented by vectors in a high dimensional space (*e.g.*, SIFT descriptors (Lowe, 2004)), are often referred to as the key points.

To efficiently handle these key points, the key idea is to quantize each extracted key point into one of *visual words*, and then represent each image by a histogram of the visual words. This vector quantization procedure allows us to represent each image by a histogram of the visual words, which is often referred to as the bag-of-words representation, and consequently converts the object categorization problem into a text categorization problem. A clustering procedure (*e.g.*, K-means) is often applied to group key points from all the training images into a large number of clusters, with the center of each cluster corresponding to a different visual word. Studies (Csurka et al., 2004; Sivic and Zisserman, 2003) have shown promising performance of bag-of-words representation in object categorization. Various methods (Jurie and Triggs, 2005; Lazebnik and Raginsky, 2009; Moosmann et al., 2007; Nister and Stewenius, 2006; Philbin et al., 2008; Tuytelaars and Schmid, 2007; Winn et al., 2005) have been proposed for the visual vocabulary construction to improve both the computational efficiency and the classification accuracy of object categorization. However, to the best of our knowledge, there is no theoretical analysis on the statistical properties of vector quantization for object categorization.

In this paper, we present a statistical framework which generalizes the bag-of-words representation and aim to provide a theoretical understanding for vector quantization and its effect on object categorization from the viewpoint of statistical consistency. In particular, we view

1. each visual word as a *quantization function* $f_k(\boldsymbol{x})$ that is randomly sampled from a class of functions $\mathcal{F}$ by an unknown distribution $\mathcal{P}_{\mathcal{F}}$, and
2. each key point of an image as a random sample from an unknown distribution $q_i(\boldsymbol{x})$.

The above statistical description of key points and visual words allows us to interpret the similarity between two images in bag-of-words representation, the key quantity in object categorization, as an empirical expectation over the distributions $q_i(\boldsymbol{x})$ and $\mathcal{P}_{\mathcal{F}}$. Based on the proposed statistical framework, we present two random algorithms for vector quantization, one based on the empirical distribution and the other based on kernel density estimation. We show that both random algorithms for vector quantization are statistically consistent in estimating the similarity between two images. Our empirical study with object recognition also verifies that the two proposed algorithms (I) yield recognition accuracy that is comparable to the clustering based bag-of-words representation, and (II) are resilient to the number of visual words when the number of training examples is limited. The success of the two simple algorithms validates the proposed statistical framework for vector quantization.

The rest of this paper is organized as follows. Section 2 presents the overview of existing approaches for key point quantization that were used by object recognition. Section 3 presents a statistical framework that generalizes the classical bag-of-words representation, and two random algorithms for vector quantization based on the proposed framework. We show that both algorithms are statistically consistent in estimating the similarity between two images. Empirical study with object recognition reported in Section 4 shows encouraging results of the proposed algorithms for vector quantization, which in return validates the proposed statistical framework for the bag-of-words representation. Section 5 concludes this work.

2 Related Work

In object recognition and texture analysis, a number of algorithms have been proposed for key point quantization. Among them, K-means is probably the most popular one. To reduce the high computational cost of K-means, hierarchical K-means is proposed in (Nister and Stewenius, 2006) for more efficient vector quantization. In (Winn et al., 2005), a supervised learning algorithm is proposed to reduce the visual vocabulary that is initially obtained by K-means, into a more descriptive and compact one. Farquhar *et al.*(2005) model the problem as Gaussian mixture model where each visual words corresponds to a Gaussian component and use the Maximum A Posterior (MAP) approach to learn the parameter. A method based on mean-shift is proposed in (Jurie and Triggs, 2005) for vector quantization to resolve the problem that K-means tends to 'starve' medium density regions in feature space and each key point is allocated to the first visual word similar to it. it. Moosmann *et al.*(2007) use extremely randomized clustering forests to efficiently generate a highly discriminative coding of visual words. To minimize the loss of information in vector quantization, Lazebnik and Raginsky (2009) try to seek a compressed representation of vectors that preserve the sufficient statistics of features. In (Perronnin et al., 2006), images are characterized using a set of category-specific histograms describing whether the content can best be modeled by the universal vocabulary or the specific vocabulary. Tuytelaars and Schmid (2007) propose a quantization method that discretizes a feature space by a regular lattice. van Gemert *et al.*(2008) use kernel density estimation to avoid the problem of 'codeword uncertainty' and 'codeword plausibility'.

Although many studies have shown encouraging results of the bag-of-words representation for object categorization, none of them provide statistical consistency analysis, which reveals the asymptotic behavior of the bag-of-words model for object recognition. Unlike the existing statistical approaches for key point quantization that are designed to reduce the training error, the proposed framework generalizes the bag-of-words model by the statistical expectation, making it possible to analyze the statistical consistency of the bag-of-words model. Finally, we would like to point out that although several randomized approaches (Moosmann et al., 2007; Nowak et al., 2006; Viitaniemi and Laaksonen, 2008) have been proposed for key point quantization, none of them provides theoretical analysis on statistical consistency. In contrast, we present not only the theoretic results for the two proposed random algorithms for vector quantization, but also the results of the empirical study with object recognition that support the theoretic claim.

3 A Statistical Framework for Bag-of-Words Representation

In this section, we first present a statistical framework for the bag-of-words representation in object categorization, followed by two random algorithms that are derived from the proposed framework. The analysis of statistical consistency is also presented for the two proposed algorithms.

3.1 A Statistical Framework

We consider the bag-of-words representation for images, with each image being represented by a collection of local descriptors. We denote by N the number of training images, and by $X_i = (\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^{n_i})$ the collection of key points used to represent image $\mathcal{I}_i$ where $\boldsymbol{x}_i^l \in$

$\mathcal{X}, l = 1, \ldots, n_i$ is a key point in feature space $\mathcal{X}$. To facilitate statistical analysis, we assume that each key point $\boldsymbol{x}_i^l$ in X_i is randomly drawn from an unknown distribution $q_i(\boldsymbol{x})$ associated with image $\mathcal{I}_i$.

The key idea of the bag-of-words representation is to quantize each key point into one of the visual words that are often derived by clustering. We generalize this idea of quantization by viewing the mapping to a visual word $\boldsymbol{v}_k \in \mathcal{X}$ as a quantization function $f_k(\boldsymbol{x}) : \mathcal{X} \mapsto [0, 1]$. Due to the uncertainty in constructing the vocabulary, we assume that the quantization function $f_k(\boldsymbol{x})$ is randomly drawn from a class of functions, denoted by $\mathcal{F}$, via a unknown distribution $\mathcal{P}_{\mathcal{F}}$. To capture the behavior of quantization, we design the function class $\mathcal{F}$ as follows

$$\mathcal{F} = \{f(\boldsymbol{x}; \boldsymbol{v}) | f(\boldsymbol{x}; \boldsymbol{v}) = I(\|\boldsymbol{x} - \boldsymbol{v}\| \leq \rho), \boldsymbol{v} \in \mathcal{X}\} \tag{1}$$

where indicator function $I(z)$ outputs 1 when z is true, or 0 otherwise. In the above definition, each quantization function $f(\boldsymbol{x}; \boldsymbol{v})$ is essentially a ball of radius ρ centered at $\boldsymbol{v}$. It outputs 1 when a point $\boldsymbol{x}$ is within the ball, and 0 if $\boldsymbol{x}$ is outside the ball. This definition of quantization function is clearly related to the vector quantization by data clustering.

Based on the above statistical interpretation of key points and quantization functions, we can now provide a statistical description for the histogram of visual words, which is the key of bag-of-words representation. Let $\hat{h}_i^k$ denotes the normalized number of key points in image $\mathcal{I}_i$ that are mapped to visual word $\boldsymbol{v}_k$. Given m visual words, or m quantization functions $\{f_k(\boldsymbol{x})\}_{k=1}^m$ that are sampled from $\mathcal{F}$, $\hat{h}_i^k$ is computed as

$$\hat{h}_i^k = \frac{1}{n_i} \sum_{j=1}^{n_i} f_k(\boldsymbol{x}_i^l) = \hat{\mathbb{E}}_i[f_k(\boldsymbol{x})] \tag{2}$$

where $\hat{\mathbb{E}}_i[f_k(\boldsymbol{x})]$ stands for the empirical expectation of function $f_k(\boldsymbol{x})$ based on the samples $\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^{n_i}$. We can generalize the above computation by replacing the empirical expectation $\hat{\mathbb{E}}_i[f_k(\boldsymbol{x})]$ with an expectation over the true distribution $q_i(\boldsymbol{x})$, i.e.,

$$h_i^k = \mathbb{E}_i[f_k(\boldsymbol{x})] = \int d\boldsymbol{x} q_i(\boldsymbol{x}) f_k(\boldsymbol{x}). \tag{3}$$

The bag-of-words representation for image $\mathcal{I}_i$ is expressed by vector $\boldsymbol{h}_i = (h_i^1, \ldots, h_i^m)$.

In the next step, we analyze the pairwise similarity between two images. It is important to note that the pairwise similarity plays a critical role in any pattern classification problems including object categorization. According to the learning theory (Schölkopf and Smola, 2002), it is the pairwise similarity, not the vector representation of images, that decides the classification performance. Using the vector representation $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$, the similarity between two images $\mathcal{I}_i$ and $\mathcal{I}_j$, denoted by $\bar{s}_{ij}$, is computed as

$$\bar{s}_{ij} = \frac{1}{m} \boldsymbol{h}_i^T \boldsymbol{h}_j = \frac{1}{m} \sum_{k=1}^{m} \mathbb{E}_i[f_k(\boldsymbol{x})] \mathbb{E}_j[f_k(\boldsymbol{x})] \tag{4}$$

Similar to the previous analysis, the summation in the above expression can be viewed as an empirical expectation over the sampled quantization functions $f_k(\boldsymbol{x}), k = 1, \ldots, m$. We thus generalize the definition of pairwise similarity in (4) by replacing the empirical expectation with the true expectation, and obtain the true similarity between two images $\mathcal{I}_i$ and $\mathcal{I}_j$ as

$$s_{ij} = \mathbb{E}_{f \sim \mathcal{P}_{\mathcal{F}}} \left[\mathbb{E}_i[f(\boldsymbol{x})] \mathbb{E}_j[f(\boldsymbol{x})] \right] \tag{5}$$

According to the definition in (1), each quantization function is parameterized by a center $\boldsymbol{v}$. Thus, to define $\mathcal{P}_{\mathcal{F}}$, it suffices to define a distribution for the center $\boldsymbol{v}$, denoted by $q(\boldsymbol{v})$. Thus, (5) can be expressed as

$$s_{ij} = \mathbb{E}_{\boldsymbol{v}}\left[\mathbb{E}_i[f(\boldsymbol{x})]\mathbb{E}_j[f(\boldsymbol{x})]\right] \tag{6}$$

3.2 Random Algorithms for Key Point Quantization and their Statistical Consistency

We emphasize that the pairwise similarity in (6) can not be computed directly. This is because both distributions $q_i(\boldsymbol{x})$ and $q(\boldsymbol{v})$ are unknown, which makes it intractable to compute $\mathbb{E}_i[\cdot]$ and $\mathbb{E}_{\boldsymbol{v}}[\cdot]$. In real applications, approximations are needed. In this section, we study how approximations will affect the estimation of pairwise similarity. In particular, given the pairwise similarity estimated by different kinds of approximated distributions, we aim to bound its difference to the underlying true similarity. To simplify our analysis, we assume that each image has at least n key points.

By assuming that the key points in all the images are sampled from $q(\boldsymbol{v})$, we have an empirical distribution for $q(\boldsymbol{v})$, i.e.,

$$\hat{q}(\boldsymbol{v}) = \frac{1}{\sum_{i=1}^{N} n_i} \sum_{i=1}^{N} \sum_{l=1}^{n_i} \delta(\boldsymbol{v} - \boldsymbol{x}_i^l) \tag{7}$$

where $\delta(\boldsymbol{x})$ is a Dirac delta function that $\int \delta(\boldsymbol{x}) d\boldsymbol{x} = 1$ and $\delta(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \neq \mathbf{0}$. Direct estimation of pairwise similarities using the above empirical distribution is computationally expensive, because the number of key points in all images can be very large. In the bag-of-words model, m visual words are used as prototypes for the key points in all the images. Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ be the m visual words randomly sampled from the key points in all the images. The empirical distribution $\hat{q}(\boldsymbol{v})$ is

$$\hat{q}(\boldsymbol{v}) = \frac{1}{m} \sum_{k=1}^{m} \delta(\boldsymbol{v} - \boldsymbol{v}_k) \tag{8}$$

In the next step, we aim to approximate the unknown distribution $q_i(\boldsymbol{x})$ in two different ways, and show the statistical consistency for each approximation.

3.2.1 Empirically Estimated Density Function for $q_i(\boldsymbol{x})$

First we approximate $q_i(\boldsymbol{x})$ by the empirical distribution $\hat{q}_i(\boldsymbol{x})$ defined as follows

$$\hat{q}_i(\boldsymbol{x}) = \frac{1}{n_i} \sum_{l=1}^{n_i} \delta(\boldsymbol{x} - \boldsymbol{x}_i^l) \tag{9}$$

Given the approximations for distribution $q_i(\boldsymbol{x})$ and $q(\boldsymbol{v})$, we can now compute the approximation of the pairwise similarity s_{ij} defined in (6). For (9), the pairwise similarity, denoted by $\hat{s}_{ij}$, is computed as

$$\begin{aligned} \hat{s}_{ij} &= \hat{\mathbb{E}}_{\boldsymbol{v}}\left[\hat{\mathbb{E}}_i[f(\boldsymbol{x})]\hat{\mathbb{E}}_j[f(\boldsymbol{x})]\right] \\ &= \frac{1}{m} \sum_{k=1}^{m} \left(\frac{1}{n_i} \sum_{l=1}^{n_i} I\Big(\|\boldsymbol{x}_i^l - \boldsymbol{v}_k\| \leq \rho\Big) \right) \left(\frac{1}{n_j} \sum_{l=1}^{n_j} I\Big(\|\boldsymbol{x}_j^l - \boldsymbol{v}_k\| \leq \rho\Big) \right) \end{aligned} \tag{10}$$

To show the statistical consistency of $\hat{s}_{ij}$, we need to bound $|s_{ij} - \hat{s}_{ij}|$. Since there are two approximate distribution used in our estimation, we divide our analysis into two steps. First, we measure $|\bar{s}_{ij} - s_{ij}|$, i.e., the difference in similarity caused by the approximate distribution for $\mathcal{P}_{\mathcal{F}}$. Next, we measure $|\hat{s}_{ij} - \bar{s}_{ij}|$, i.e., the difference caused by using the approximate distribution for $q_i(\boldsymbol{x})$. The overall difference $|s_{ij} - \hat{s}_{ij}|$ is bounded by the sum of the two difference.

We first state the McDiarmid inequality (McDiarmid, 1989), which is used throughout our analysis.

Theorem 1 *(McDiarmid Inequality) Given independent random variables $v_1, v_2, \ldots, v_n, v_i' \in V$, and a function $f : V^n \mapsto \mathbb{R}$ satisfying*

$$\sup_{v_1, v_2, \ldots, v_n, v_i' \in V} |f(\boldsymbol{v}) - f(\boldsymbol{v}')| \leq c_i \tag{11}$$

where $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$ and $\boldsymbol{v}' = (v_1, v_2, \ldots, v_{i-1}, v_i', v_{i+1}, \ldots, v_n)$, then the following statement holds

$$\Pr\Big(|f(\boldsymbol{v}) - \mathbb{E}(f(\boldsymbol{v}))| \geq \epsilon\Big) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right) \tag{12}$$

Using the McDiarmid inequality, we have the following theorem which bounds $|\bar{s}_{ij} - s_{ij}|$.

Theorem 2 *Assuming $f_k(\boldsymbol{x})$, $k = 1, \ldots, m$ are randomly drawn from class $\mathcal{F}$ according to an unknown distribution. And further assuming that any function in $\mathcal{F}$ is universally bounded between 0 and 1. With probability $1 - \delta$, the following inequality holds for any two training images $\mathcal{I}_i$ and $\mathcal{I}_j$*

$$|\bar{s}_{ij} - s_{ij}| \leq \sqrt{\frac{1}{2m} \ln \frac{2}{\delta}} \tag{13}$$

Proof For any $f \in \mathcal{F}$, we have $0 \leq \mathbb{E}_i[f(\boldsymbol{x})]\mathbb{E}_j[f(\boldsymbol{x})] \leq 1$. Thus, for any k, $c_k \leq 1/m$. By setting

$$\delta = 2\exp\left(-2m\epsilon^2\right), \text{ or }, \epsilon = \sqrt{\frac{1}{2m} \ln \frac{2}{\delta}}, \tag{14}$$

we have $\Pr\Big(|\bar{s}_{ij} - s_{ij}| \leq \epsilon\Big) \geq 1 - \delta$.

The above theorem indicates that, if we have the true distribution $q_i(\boldsymbol{x})$ of each image $\mathcal{I}_i$, with a large number of sampled quantization functions $f_k(\boldsymbol{x})$, we have a very good chance to recover the true similarity s_{ij} with a small error. The next theorem bounds $|\hat{s}_{ij} - s_{ij}|$.

Theorem 3 *Assuming each image has at least n randomly sampled key points. Also assuming that $f_k(\boldsymbol{x})$, $k = 1, \ldots, m$ randomly drawn from an unknown distribution over class $\mathcal{F}$. With probability $1 - \delta$, the following inequality is satisfied for any two images $\mathcal{I}_i$ and $\mathcal{I}_j$*

$$|\hat{s}_{ij} - s_{ij}| \leq \sqrt{\frac{1}{2m} \ln \frac{2}{\delta}} + 2\sqrt{\frac{1}{2n} \ln \frac{4m^2}{\delta}} \tag{15}$$

Proof We first need to bound the difference between $\hat{\mathbb{E}}_i[f_k(\boldsymbol{x})]$ and $\mathbb{E}_i[f_k(\boldsymbol{x})]$. Since $0 \leq f(\boldsymbol{x}) \leq 1$ for any $f \in \mathcal{F}$, using McDimard inequality, we have

$$\Pr\left(\left|\hat{\mathbb{E}}_i\left[f_k(\boldsymbol{x})\right] - \mathbb{E}_i\left[f_k(\boldsymbol{x})\right]\right| \geq \epsilon\right) \leq \exp(-2n\epsilon^2) \tag{16}$$

By setting

$$2\exp(-2n\epsilon^2) = \frac{\delta}{2m^2}, \text{ or, } \epsilon = \sqrt{\frac{1}{2n}\ln\left(\frac{4m^2}{\delta}\right)}$$

with probability $1-\delta/2$, we have $|\hat{\mathbb{E}}_i[f_k(\boldsymbol{x})]-\mathbb{E}_i[f_k(\boldsymbol{x})]| \leq \epsilon$ and $|\hat{\mathbb{E}}_j[f_k(\boldsymbol{x})]-\mathbb{E}_j[f_k(\boldsymbol{x})]| \leq \epsilon$ for all $f_k(\boldsymbol{x})_{k=1}^m$ simultaneously. As a result, with probability $1-\delta/2$, for any two image $\mathcal{I}_i$ and $\mathcal{I}_j$, we have

$$\begin{aligned}
|\hat{s}_{ij} - \bar{s}_{ij}| &\leq \frac{1}{m}\sum_{k=1}^{m} |\hat{\mathbb{E}}_i^k\hat{\mathbb{E}}_j^k - \mathbb{E}_i^k\mathbb{E}_j^k| \\
&\leq \frac{1}{m}\sum_{k=1}^{m} |(\hat{\mathbb{E}}_i^k - \mathbb{E}_i^k)\hat{\mathbb{E}}_j^k| + |\mathbb{E}_i^k(\hat{\mathbb{E}}_j^k - \mathbb{E}_j^k)| \\
&\leq \frac{1}{m}\sum_{k=1}^{m} |\hat{\mathbb{E}}_i^k - \mathbb{E}_i^k| + |\hat{\mathbb{E}}_j^k - \mathbb{E}_j^k| \\
&\leq 2\epsilon = 2\sqrt{\frac{1}{2n}\ln\left(\frac{4m^2}{\delta}\right)}
\end{aligned} \tag{17}$$

where $\hat{\mathbb{E}}_i^k$ stands for $\hat{\mathbb{E}}_i[f_k(\boldsymbol{x})]$ for simplicity. According to Theorem 2, with probability $1-\delta/2$, we have

$$|\bar{s}_{ij} - s_{ij}| \leq \sqrt{\frac{1}{2m}\ln\frac{2}{\delta}} \tag{18}$$

Combining (17) and (18), we have the result in the theorem. With probability $1-\delta$, the following inequality is satisfied

$$|\hat{s}_{ij} - s_{ij}| \leq \sqrt{\frac{1}{2m}\ln\frac{2}{\delta}} + 2\sqrt{\frac{1}{2n}\ln\frac{4m^2}{\delta}} \tag{19}$$

Remark Theorem 3 reveals an interesting relationship between the estimation error $|s_{ij} - \hat{s}_{ij}|$ and the number of quantization functions (or the number of visual words). The upper bound in Theorem 3 consists of two terms: the first term decreases at a rate of $O(1/\sqrt{m})$ while the second term increases at a rate of $O(\ln m)$. When the number of visual words m is small, the first term dominates the upper bound, and therefore increasing m will reduce the difference $|\hat{s}_{ij} - s_{ij}|$. As m becomes significantly larger than n, the second term will dominate the upper bound, and therefore increasing m will lead to a larger $|\hat{s}_{ij} - s_{ij}|$. This result appears to be consistent with the observations on the size of the visual vocabulary: a large vocabulary tends to performance well in object categorization; but, too many visual words could deteriorate the classification accuracy.

Finally, we emphasize that although the idea of vector quantization by randomly sampled centers was already discussed in (Jurie and Triggs, 2005; Viitaniemi and Laaksonen, 2008), to the best of our knowledge, this is the first work that presents its statistical consistency analysis.

3.2.2 Kernel Density Function Estimation for $q_i(\boldsymbol{x})$

In this section, we approximate $q_i(\boldsymbol{x})$ by a kernel density estimation. To this end, we assume that the density function $q_i(\boldsymbol{x})$ belongs to a family of smooth functions $\mathcal{F}_D$ that is defined as follows

$$\mathcal{F}_D = \left\{ q(\boldsymbol{x}) : \mathcal{X} \mapsto \mathbb{R}_+ \middle| \big\langle q(\boldsymbol{x}), q(\boldsymbol{x}) \big\rangle_{\mathcal{H}_\kappa} \leq B^2, \int q(\boldsymbol{x}) d\boldsymbol{x} = 1 \right\} \tag{20}$$

where $\kappa(\boldsymbol{x}, \boldsymbol{x}') : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}_+$ is a local kernel function with $\int \kappa(\boldsymbol{x}, \boldsymbol{x}') d\boldsymbol{x}' = 1$. B controls the functional norm of $q(\boldsymbol{x})$ in the reproducing kernel Hilbert space $\mathcal{H}_\kappa$. An example of $\kappa(\boldsymbol{x}, \boldsymbol{x}')$ is RBF function, i.e. $\kappa(\boldsymbol{x}, \boldsymbol{x}') \propto \exp(-\lambda d(\boldsymbol{x}, \boldsymbol{x}')^2)$, where $d(\boldsymbol{x}, \boldsymbol{x}') = \|\boldsymbol{x} - \boldsymbol{x}'\|_2$. Then, the distribution $q_i(\boldsymbol{x})$ is approximated by a kernel density estimation $\tilde{q}_i(\boldsymbol{x})$ defined as follows

$$\tilde{q}_i(\boldsymbol{x}) = \sum_{l=1}^{n_i} \alpha_i^l \kappa(\boldsymbol{x}, \boldsymbol{x}_i^l), \tag{21}$$

where $\alpha_i^l (1 \leq l \leq n_i)$ are the combination weight that satisfy (i) $\alpha_i^l \geq 0$, (ii) $\sum_{l=1}^{n_i} \alpha_i^l = 1$, and (iii) $\alpha_i K_i \alpha_i \leq B^2$, where $K_i = [\kappa(\boldsymbol{x}_i^l, \boldsymbol{x}_i^{l'})]_{n_i \times n_i}$.

Using the kernel density function, we approximate the pairwise similarity for (21) as follows

$$\tilde{s}_{ij} = \hat{\mathbb{E}}_{\boldsymbol{v}} \left[\tilde{\mathbb{E}}_i[f(\boldsymbol{x})] \tilde{\mathbb{E}}_j[f(\boldsymbol{x})] \right] = \frac{1}{m} \sum_{k=1}^{m} \left(\sum_{l=1}^{n_i} \alpha_i^l \theta\big(\boldsymbol{x}_i^l, \boldsymbol{v}_k\big) \right) \left(\sum_{l=1}^{n_j} \alpha_j^l \theta\big(\boldsymbol{x}_j^l, \boldsymbol{v}_k\big) \right) \tag{22}$$

where function $\theta(\boldsymbol{x}, \boldsymbol{v})$ is defined as

$$\theta(\boldsymbol{x}, \boldsymbol{v}) = \int d\boldsymbol{z} I\big(d(\boldsymbol{z}, \boldsymbol{v}) \leq \rho\big) \kappa(\boldsymbol{x}, \boldsymbol{z}) \tag{23}$$

To bound the difference between $\tilde{s}_{ij}$ and s_{ij}, we follow the analysis (Shawe-Taylor and Dolia, 2007) by viewing $\mathbb{E}_i[f(\boldsymbol{x})]\mathbb{E}_j[f(\boldsymbol{x})]$ as a mapping, denoted by $g : \mathcal{F} \mapsto \mathbb{R}_+$, i.e.,

$$g(f; q_i, q_j) = \mathbb{E}_i[f(\boldsymbol{x})]\mathbb{E}_j[f(\boldsymbol{x})] \tag{24}$$

The domain for function g, denoted by $\mathcal{G}$, is defined as

$$\mathcal{G} = \big\{ g : \mathcal{F} \mapsto \mathbb{R}_+ \big| \, \exists q_i, q_j \in \mathcal{F}_D \quad \text{s.t. } g(f) = \mathbb{E}_i[f(\boldsymbol{x})]\mathbb{E}_j[f(\boldsymbol{x})] \big\} \tag{25}$$

To bound the complexity of a class of functions, we introduce the concept of *Randemacher complexity* (Bartlett and Wang, 2002):

Defination 1 *(Randemacher Complexity) Suppose $x_1, \ldots, x_n$ are sampled from a set $\mathcal{X}$ with i.i.d. Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. The Randemacher complexity of $\mathcal{F}$ is defined as*

$$R_n(\mathcal{F}) = \mathbb{E}_{x_1, \ldots, x_n, \sigma} \left(\sup_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right) \tag{26}$$

where σ_i is independent uniform ± 1-valued random variables.

Assuming at least n key points are randomly sampled from each image, we have the following lemmas that bounds the complexity of domain $\mathcal{G}$:

Lemma 1 *The Rademacher complexity of function class $\mathcal{G}$, denoted by $R_m(\mathcal{G})$, is bounded as*

$$R_m(\mathcal{G}) \leq 2BC_\kappa \frac{\mathbb{E}_f\left[\int d\boldsymbol{x}|f(\boldsymbol{x})|\right]}{\sqrt{m}} \tag{27}$$

where $C_\kappa = \max_{\boldsymbol{x},\boldsymbol{z}} \sqrt{\kappa(\boldsymbol{x},\boldsymbol{z})}$

Proof Denote $F = \{f_1, \dots, f_m\}$, according to the definition, we have

$$\begin{aligned}
R_m(\mathcal{G}) &= \mathbb{E}_{\sigma,F}\left[\sup_{g\in\mathcal{G}} \frac{2}{m}\sum_{k=1}^{m} \sigma_k g(f_k)\right] \\
&= \mathbb{E}_F\left[\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}} \frac{2}{m}\sum_{k=1}^{m} \sigma_k g(f_k)\middle| F\right]\right] \\
&= \mathbb{E}_F\left[\mathbb{E}_\sigma\left[\sup_{q_i,q_j\in\mathcal{F}_D} \frac{2}{m}\sum_{k=1}^{m} \sigma_k \mathbb{E}_i[f_k]\mathbb{E}_j[f_k]\middle| F\right]\right] \\
&\leq \mathbb{E}_F\left[\mathbb{E}_\sigma\left[\sup_{\|\boldsymbol{\omega}_i\|\leq B} \frac{2}{m}\sum_{k=1}^{m} \sigma_k \mathbb{E}_i[f_k]\middle| F\right]\right] \\
&= \frac{2}{m}\mathbb{E}_F\left[\mathbb{E}_\sigma\left[\sup_{\|\boldsymbol{\omega}_i\|\leq B} \left\langle \boldsymbol{\omega}_i, \sum_{k=1}^{m} \sigma_k \varPhi_k\right\rangle\middle| F\right]\right]
\end{aligned}$$

where $\varPhi_k = \big(\langle\phi_1(\cdot), f_k(\cdot)\rangle, \langle\phi_2(\cdot), f_k(\cdot)\rangle, \dots\big)$ and $\phi_k(x)$ is an eigen function of $\kappa(x,x')$

$$\begin{aligned}
&\leq \frac{2B}{m}\mathbb{E}_F\left[\mathbb{E}_\sigma\left[\left\|\sum_{k=1}^{m} \sigma_k \varPhi_k\right\|\middle| F\right]\right] \\
&= \frac{2B}{m}\mathbb{E}_F\left[\mathbb{E}_\sigma\left[\left(\sum_{k,t} \sigma_k\sigma_t\big\langle\varPhi_k, \varPhi_t\big\rangle\right)^{\frac{1}{2}}\middle| F\right]\right] \\
&\leq \frac{2B}{m}\mathbb{E}_F\left[\left(\sum_{k,t} \mathbb{E}_\sigma\left[\sigma_k\sigma_t\big\langle\varPhi_k, \varPhi_t(\boldsymbol{x})\big\rangle\middle| F\right]\right)^{\frac{1}{2}}\right] \\
&= \frac{2B}{m}\mathbb{E}_F\left[\left(\sum_{k} \mathbb{E}_\sigma\left[\sigma_k^2\big\langle\varPhi_k, \varPhi_k\big\rangle\middle| F\right]\right)^{\frac{1}{2}}\right] \\
&= \frac{2B}{m}\mathbb{E}_F\left[\left(\sum_{k} \big\langle\varPhi_k, \varPhi_k\big\rangle\right)^{\frac{1}{2}}\right] \\
&= \frac{2B}{m}\mathbb{E}_F\left[\left(\sum_{k} \int d\boldsymbol{z}d\boldsymbol{x}f_k(\boldsymbol{x})f_k(\boldsymbol{z})\kappa(\boldsymbol{x},\boldsymbol{z})\right)^{\frac{1}{2}}\right] \\
&\leq 2BC_\kappa \frac{\mathbb{E}_f\left[\int d\boldsymbol{x}|f(\boldsymbol{x})|\right]}{\sqrt{m}}
\end{aligned} \tag{28}$$

where the first inequality is because $\mathbb{E}_j[f_k] \leq 1$, the second inequality is from Cauchy's inequality, the third and fourth inequalities are from Jensen's inequality. The last equality follows

$$\langle \Phi_k, \Phi_k \rangle = \sum_i \left\langle \phi_i(\cdot), f_k(\cdot) \right\rangle^2 = \int dz dx f_k(\boldsymbol{x}) f_k(\boldsymbol{z}) \kappa(\boldsymbol{x}, \boldsymbol{z}) \tag{29}$$

From (Bartlett and Wang, 2002), we have the following lemmas:

Lemma 2 *(Theorem 12 in (Bartlett and Wang, 2002)) For $1 \leq q < \infty$, let $\mathcal{L} = \{|f - h|^q : f \in \mathcal{F}\}$, where h and $\|f - h\|_\infty$ is uniformly bounded. We have*

$$R_n(\mathcal{L}) \leq 2q\|f - h\|_\infty \left(R_n(\mathcal{F}) + \frac{\|h\|_\infty}{\sqrt{n}} \right) \tag{30}$$

Lemma 3 *(Theorem 8 in (Bartlett and Wang, 2002)) With probability $1 - \delta$ the following inequality holds*

$$\mathbb{E}\phi(Y, f(X)) \leq \hat{\mathbb{E}}_n \phi(Y, f(X)) + R_n(\phi \circ \mathcal{F}) + \sqrt{\frac{8 \ln(2/\delta)}{n}} \tag{31}$$

where $\phi(x, y)$ is the loss function, n is the number of samples and $\phi \circ \mathcal{F} = \{(x, y) \mapsto \phi(y, f(x)) - \phi(y, 0) : f \in \mathcal{F}\}$.

Based on the above lemmas, we have the following theorem

Theorem 4 *Assume that the density function $q_i(\boldsymbol{x}), q_j(\boldsymbol{x}) \in \mathcal{F}_D$. Let $\tilde{q}_i(\boldsymbol{x}), \tilde{q}_j(\boldsymbol{x}) \in \mathcal{F}_D$ be an estimated density function from n sampled key points. We have, with probability $1 - \delta$, the following inequality holds*

$$\begin{aligned} \mathbb{E}_f[|g(f; \tilde{q}_i, \tilde{q}_j) - g(f; q_i, q_j)|] &\leq \hat{\mathbb{E}}_f[|g(f; \tilde{q}_i, \tilde{q}_j) - g(f; \hat{q}_i, \hat{q}_j)|] \\ &+2 \left(2BC_\kappa \frac{\mathbb{E}_f\left[\int d\boldsymbol{x}|f(\boldsymbol{x})|\right]}{\sqrt{m}} + \frac{1}{\sqrt{m}} \right) + \sqrt{\frac{\ln(8/\delta)}{2m}} + 2\sqrt{\frac{\ln(8m^2/\delta)}{2n}} \end{aligned} \tag{32}$$

Proof From Lemma 3, with probability $1 - \delta/2$, we have

$$\begin{aligned} &\mathbb{E}_f[|g(f; \tilde{q}_i, \tilde{q}_j) - g(f; q_i, q_j)|] \\ &\quad \leq \hat{\mathbb{E}}_f[|g(f; \tilde{q}_i, \tilde{q}_j) - g(f; q_i, q_j)|] + R_m(|\mathcal{G} - g(f; q_i, q_j)|) + \sqrt{\frac{8 \ln(4/\delta)}{m}} \end{aligned} \tag{33}$$

Since $0 \leq g(f; q_i, q_j) \leq 1$, using the results in Lemma 1 and 2, we have

$$\begin{aligned} R_m\big(|\mathcal{G} - g(f; q_i, q_j)|\big) &\leq 2 \left(R_m(\mathcal{G}) + \frac{1}{\sqrt{m}} \right) \\ &\leq 2 \left(2BC_\kappa \frac{\mathbb{E}_f\left[\int d\boldsymbol{x}|f(\boldsymbol{x})|\right]}{\sqrt{m}} + \frac{1}{\sqrt{m}} \right) \end{aligned} \tag{34}$$

Hence, we have, with probability $1 - \delta/2$ the following inequality holds

$$\begin{aligned} \mathbb{E}_f[|g(f; \tilde{q}_i, \tilde{q}_j) - g(f; q_i, q_j)|] &\leq \hat{\mathbb{E}}_f[|g(f; \tilde{q}_i, \tilde{q}_j) - g(f; q_i, q_j)|] \\ &+2 \left(2BC_\kappa \frac{\mathbb{E}_f\left[\int d\boldsymbol{x}|f(\boldsymbol{x})|\right]}{\sqrt{m}} + \frac{1}{\sqrt{m}} \right) + \sqrt{\frac{8 \ln(4/\delta)}{m}} \end{aligned} \tag{35}$$

Next, we aim to bound $\hat{\mathbb{E}}_f[|g(f;\tilde{q}_i,\tilde{q}_j) - g(f;q_i,q_j)|]$. Note that

$$\hat{\mathbb{E}}_f[|g(f;\tilde{q}_i,\tilde{q}_j) - g(f;q_i,q_j)|] = \frac{1}{m}\sum_{k=1}^{m} |g(f_k;\tilde{q}_i,\tilde{q}_j) - g(f_k;q_i,q_j)|$$

$$\leq \frac{1}{m}\sum_{k=1}^{m}\Big(|g(f_k;\tilde{q}_i,\tilde{q}_j) - g(f_k;\hat{q}_i,\hat{q}_j)| + |g(f_k;\hat{q}_i,\hat{q}_j) - g(f_k;q_i,q_j)|\Big) \quad (36)$$

Using the same logistics in the proof of Theorem 3, we have, with probability $1-\delta/2$

$$\frac{1}{m}\sum_{k=1}^{m} |g(f_k;\hat{q}_i,\hat{q}_j) - g(f_k;q_i,q_j)| \leq \sqrt{\frac{\ln(8/\delta)}{2m}} + 2\sqrt{\frac{\ln(8m^2/\delta)}{2n}} \quad (37)$$

From the above results, we have, with probability $1-\delta/2$, the following inequality holds

$$\frac{1}{m}\sum_{k=1}^{m} |g(f_k;\tilde{q}_i,\tilde{q}_j) - g(f_k;q_i,q_j)|$$
$$\leq \frac{1}{m}\sum_{k=1}^{m} |g(f_k;\tilde{q}_i,\tilde{q}_j) - g(f_k;\hat{q}_i,\hat{q}_j)| + \sqrt{\frac{\ln(8/\delta)}{2m}} + 2\sqrt{\frac{\ln(8m^2/\delta)}{2n}} \quad (38)$$

Combining the above results together, we have, with probability $1-\delta$, the following inequality holds

$$\mathbb{E}_f[|g(f;\tilde{q}_i,\tilde{q}_j) - g(f;q_i,q_j)|]$$
$$\leq \hat{\mathbb{E}}_f[|g(f;\tilde{q}_i,\tilde{q}_j) - g(f;\hat{q}_i,\hat{q}_j)|] + 2\left(2BC_\kappa\frac{\mathbb{E}_f\left[\int d\boldsymbol{x}|f(\boldsymbol{x})|\right]}{\sqrt{m}} + \frac{1}{\sqrt{m}}\right)$$
$$+\sqrt{\frac{\ln(8/\delta)}{2m}} + 2\sqrt{\frac{\ln(8m^2/\delta)}{2n}} \quad (39)$$

In our empirical study, we will use RBF kernel function for $\kappa(\boldsymbol{x},\boldsymbol{x}')$ with $\alpha_i^l = 1/n_i$. The corollary below shows the bound for this choice of kernel density estimation.

Corollary 5 *When the kernel function $\kappa(\boldsymbol{x},\boldsymbol{x}') = \left(1/(2\pi\sigma^2)\right)^{d/2}\exp\left(-\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2/(2\sigma^2)\right)$ and $\alpha_i^l = 1/n_i$, the bound in Theorem 4 becomes*

$$\mathbb{E}_f[|g(f;\tilde{q}_i,\tilde{q}_j) - g(f;q_i,q_j)|] \leq \Big(1/(2\pi\sigma^2)\Big)^{d/2}\Big(1-\exp\Big(-\rho^2/(2\sigma^2)\Big)\Big)$$
$$+2\frac{2\mathbb{E}_f\left[\int d\boldsymbol{x}|f(\boldsymbol{x})|\right]/\sqrt{n_i}+1}{\sqrt{m}} + \sqrt{\frac{\ln(8/\delta)}{2m}} + 2\sqrt{\frac{\ln(8m^2/\delta)}{2n}} \quad (40)$$

Remark Theorem 4 bounds the true expectation of the difference between the similarity estimated by kernel density function and the true similarity. Similar to Theorem 3, this bound also consists of a term decreasing at a rate of $O(1/\sqrt{m})$ and a term increasing at a rate of $O(\ln m)$. What's more, we can see in order to minimize the true expectation of the difference between the similarity estimated by kernel density function and the true similarity, we need to minimize the empirical expectation of the difference between the similarity estimated by kernel density function and the similarity estimated by empirical density function. If $\kappa(\boldsymbol{x},\boldsymbol{v})$ decreases exponentially as $d(\boldsymbol{x},\boldsymbol{v})$ decreases, such as Gaussian kernel, we have $\theta(\boldsymbol{x},\boldsymbol{v})$ close to 1 when $d(\boldsymbol{x}, \boldsymbol{v}) \leq \rho$ while $\theta(\boldsymbol{x}, \boldsymbol{v})$ close to 0 when $d(\boldsymbol{x}, \boldsymbol{v}) > \rho$. In such circumstance, setting $\alpha_i^l = 1/n_i$ for all $1 \leq l \leq n_i$ is a good choice for the approximation and is also very efficient since we do not need to learn α.

Note that although the idea of kernel density estimation was already proposed in some studies (*e.g.*,(van Gemert et al., 2008)), to the best of our knowledge, this is the first work that reveals the statistical consistency of kernel density estimation for the bag-of-words representation.

4 Empirical Study

In this empirical study, we aim to verify the proposed framework and the related analysis. To this end, based on the discussion in Section 3.2, we present two random algorithms for vector quantization that are shown in Algorithm 1. We refer to the algorithm based on empirical distribution as “Quantization via Empirical Estimation”, or **QEE** for short, and to the algorithm based on kernel density estimation as “Quantization via Kernel Estimation”, or **QKE** for short. Note that since both vector quantization algorithms do not rely on the clustering algorithms to identify visual words, they are in general computationally more efficient. In addition, both algorithms have error bounds decreases at the rate of $O(1/\sqrt{m})$ when the number of key points n is large, indicating that they are robust to the number of visual words m. We emphasize that although similar random algorithms for vector quantization have been discussed in (Farquhar et al., 2005; Philbin et al., 2008; van Gemert et al., 2008; Nowak et al., 2006; Viitaniemi and Laaksonen, 2008), the purpose of this empirical study is to verify that

- simple random algorithms deliver similar performance of object recognition as the clustering based algorithm, and
- the random algorithms are robust to the number of visual words, as predicted by the statistical consistency analysis.

Finally in the implementation of QKE, to efficiently calculate θ-function, we approximate it as (Abramowitz and Stegun, 1972)

$$\theta \approx \frac{2(\tilde{d} - \tilde{\rho})^2 - 1}{4\sqrt{\pi}(\tilde{d} - \tilde{\rho})^3 \exp(\tilde{d} - \tilde{\rho})^2} - \frac{2(\tilde{d} + \tilde{\rho})^2 - 1}{4\sqrt{\pi}(\tilde{d} + \tilde{\rho})^3 \exp(\tilde{d} + \tilde{\rho})^2} \tag{41}$$

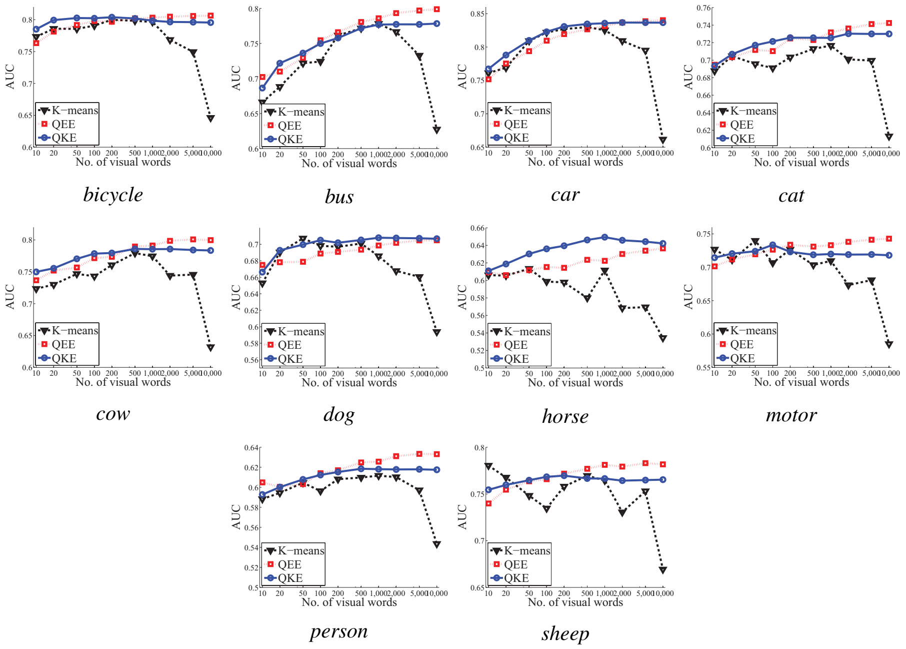
where $\tilde{d} = d/\sigma$, $\tilde{\rho} = \rho/\sigma$ and σ is the width of the Gaussian kernel.

Two data sets are used in our study: PASCAL VOC Challenge 2006 data set (Everingham et al., 2006) and Graz02 data set (Opelt et al., 2006). *PASCAL06* contains $5,304$ images from ten classes. We randomly select 100 images for training and 500 for testing. The *Graz02* data set contains 365 bike images, 420 car images, 311 people images and 380 background images. We randomly select 100 images from each class for training, and use the remaining for testing. By using a relatively small number of examples for training, we are able to examine the sensitivity of a vector quantization algorithm to the number of visual words. On average $1,000$ key points are extracted from each image, and each key point is represented by the SIFT local descriptor (Vedaldi and Fulkerson, 2008). For *PASCAL06* data set, the binary classification performance for each object class is measured by the area under the ROC curve (AUC). For *Graz02* data set, the binary classification performance for each object class is measured by the accuracy. Results averaged over ten random trials are reported.

We compare three vector quantization methods: K-means, QEE and QKE. Note that we do not include more advanced algorithms for vector quantization in our study because the

Algorithm 1 The QEE/QKE algorithm for generating bag-of-words representation

```
Input:
    $X = \{X_1, \ldots, X_N\}$: a collection of $N$ training images
    $m$: the number of sampled cluster centers
    $\rho$: threshold used by quantization functions
Output:
    $H = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N\}$: bag-of-words representation for training images
Process:
Sample $m$ centers $\{\boldsymbol{v}_k\}_{k=1}^m$ from the key points in $X$
for $i = 1$ to $N$ do
    for $k = 1$ to $m$ do
        $h_i^k = \sum_{l=1}^{n_i} I\left(d(\boldsymbol{x}_i^l, \boldsymbol{v}_k) \leq \rho\right)/n_i$ (for QEE)
        $h_i^k = \sum_{l=1}^{n_i} \theta(\boldsymbol{x}_i^l, \boldsymbol{v}_k)/n_i$ with $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \left(1/(2\pi\sigma^2)\right)^{d/2} \exp\left(-d(\boldsymbol{x}, \boldsymbol{x}')^2/(2\sigma^2)\right)$ (for QKE)
    end for
    Set $\boldsymbol{h}_i = (h_i^1, \ldots, h_i^m)$
end for
```



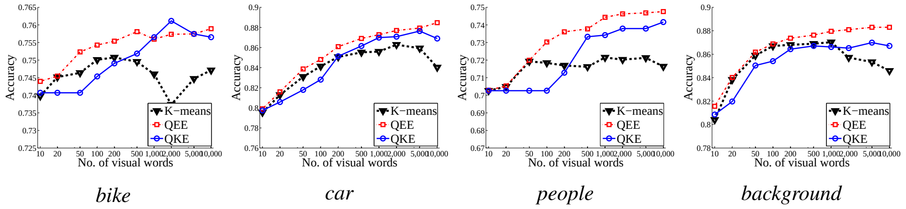
bicycle *bus* *car* *cat*

cow *dog* *horse* *motor*

person *sheep*

Fig. 1 Comparison of different quantization methods with varied number of visual words on *PASCAL06*.

objective of this study is to validate the proposed statistical framework for bag-of-words representation and the analysis on statistical consistency. Threshold ρ used by quantization functions $f(\boldsymbol{x})$ is set as $\rho = 0.5 \times \bar{d}$, where $\bar{d}$ is the average distance between all the key points and the randomly selected centers. A RBF kernel is used in QKE with the kernel width σ is set as $0.75\bar{d}$ according to our experience. Binary linear SVM is used for each classification problem. To examine the sensitivity to the number of visual words, for both data sets, we varied the number of visual words from 10 to $10,000$, as shown in Figure 1 and 2.



bike *car* *people* *background*

Fig. 2 Comparison of different quantization methods with varying number of visual words on *Graz02*.

First, we observe that the proposed algorithms for vector quantization yield comparable if not better performance than the K-means clustering algorithm. This confirms the proposed statistical framework for key point quantization is effective. Second, we observe that the clustering based approach for vector quantization tends to perform worse, sometimes very significantly, when the number of visual words is large. We attribute this instability to the fact that K-means requires each interest point belongs to exactly one visual word. If the number of clusters is not appropriate, for example, too large compared to the number of instances, two relevant key points may be separated into different clusters although they are both very near to the boundary. It will lead to a poor estimation of pairwise similarity. The problem of "hard assignment" was also observed in (Philbin et al., 2008; van Gemert et al., 2008). In contrast, for the proposed algorithms, we observe a rather stable improvement as the number of visual words increases, consistent with our analysis in statistical consistency.

5 Conclusion

The bag-of-words model is one of the most popular representation methods for object categorization. The key idea is to quantize each extracted key point into one of visual words, and then represent each image by a histogram of the visual words. For this purpose, a clustering algorithm (*e.g.*, K-means), is generally used for generating the visual words. Although a number of studies have shown encouraging results of the bag-of-words representation for object categorization, theoretical studies on properties of the bag-of-words model is almost untouched, possibly due to the difficulty introduced by using a heuristic clustering process. In this paper, we present a statistical framework which generalizes the bag-of-words representation. In this framework, the visual words are generated by a statistical process rather than using a clustering algorithm, while the empirical performance is competitive to clustering-based method. A theoretical analysis based on statistical consistency is presented for the proposed framework. Moreover, based on the framework we developed two algorithms which do not rely on clustering, while achieving competitive performance in object categorization when compared to clustering-based bag-of-words representations.

Bag-of-words representation is a popular approach to object categorization. Despite its success, few studies are devoted to the theoretic analysis of the bag-of-words representation. In this work, we present a statistical framework for key point quantization that generalizes the bag-of-words model by statistical expectation. We present two random algorithms for vector quantization where the visual words are generated by a statistical process rather than using a clustering algorithm. A theoretical analysis of their statistical consistency is presented. We also verify the efficacy and the robustness of the proposed framework by applying it to object recognition. In the future, we plan to examine the dependence of the proposed algorithms on the threshold ρ, and extend QKE to weighted kernel density estimation.

Acknowledgements We want to thank the reviewers for helpful comments and suggestions. This research is partially supported by the National Fundamental Research Program of China (2010CB327903), the Jiangsu 333 High-Level Talent Cultivation Program and the National Science Foundation (IIS-0643494). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, NY, 1972.

P. L. Bartlett and M. Wang. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

G. Csurka, C. Dance, L. Fan, J. Williamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, Prague, Czech Republic, 2004.

M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf, 2006.

J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving "bag-of-keypoints" image categorisation. Technical report, University of Southampton, 2005.

T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, pages 137–142, Chemnitz, Germany, 1998.

F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, pages 604–610, Beijing, China, 2005.

S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31 (7):1294–1309, 2009.

D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI Workshop on Learning for Text Categorization*, Madison, WI, 1998.

C. McDiarmid. On the method of bounded differences. *In Surveys in Combinatorics 1989*, pages 148–188, 1989.

F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 985–992. MIT Press, Cambridge, MA, 2007.

D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2161–2168, New York, NY, 2006.

E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *Proceedings of the 9th European Conference on Computer Vision*, pages 490–503, Graz, Austria, 2006.

A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. Generic object recognition with boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):416–431, 2006.

F. Perronnin, C. Dance, G. Csurka, and M. Bressian. Adapted vocabularies for generic visual categorization. In *Proceedings of the 9th European Conference on Computer Vision*, pages 464–475, Graz, Austria, 2006.

J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008.

B. Schölkopf and A. J. Smola. *Learning with kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.

J. Shawe-Taylor and A. Dolia. A framework for probability density estimation. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, pages 468–475, San Juan, Puerto Rico, 2007.

J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, pages 1470–1477, Nice, France, 2003.

T. Tuytelaars and C. Schmid. Vector quantizing feature space with a regular lattice. In *Proceedings of the 11th IEEE International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brazil, 2007.

J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. In *Proceedings of the 10th European Conference on Computer Vision*, pages 696–709, Marseille, France, 2008.

A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

V. Viitaniemi and J. Laaksonen. Experiments on selection of codebooks for local image feature histograms. In *Proceedings of the 10th International Conference series on Visual Information Systems*, pages 126–137, Salerno, Italy, 2008.

J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, pages 1800–1807, Beijing, China, 2005.