High-Order Topology Modeling of Visual Words for Image Classification

Kaiqi Huang, *Senior Member, IEEE*, Chong Wang, *Student Member, IEEE*, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Modeling relationship between visual words in feature encoding is important in image classification. Recent methods consider this relationship in either image or feature space, and most of them incorporate only pairwise relationship (between visual words). However, in situations involving large variability in images, one cannot capture intrinsic invariance of intra-class images using low-order pairwise relationship. The result is not robust to larger variations in images. In addition, as the number of potential pairings grows exponentially with the number of visual words, the task of learning becomes computationally expensive. To overcome these two limitations, we propose an efficient classification framework that exploits high-order topology of visual words in the feature space, as follows. First, we propose a search algorithm that seeks dependence between the visual words. This dependence is used to construct higher order topology in the feature space. Then, the local features are encoded according to this higher order topology to improve the image classification. Experiments involving four common data sets, namely PASCAL VOC 2007, 15 Scenes, Caltech 101, and UIUC Sport Event, demonstrate that the dependence search significantly improves the efficiency of higher order topological construction, and consequently increases the image classification in all these data sets.**

***Index Terms*— Image classification, feature encoding, visual words, higher-order.**

I. Introduction

IMAGE classification is a fundamental problem in computer vision [1]–[3]. In the last ten years, the Bag-of-Words (BoW) model [4] has been popularly used, while the deep learning – especially the Convolutional Neural Network (CNN) has recently dominated this area and has achieved huge success in large-scale applications such as the ILSVRC dataset [5]. Though deep learning seems a better choice for image classification, there are still some old problems left behind in BoW model. The BoW model has four main steps: feature extraction, codebook learning, feature encoding and pooling. Most studies [6]–[9] on BoW focus on feature encoding, which considers the relationship between each local feature and visual words to reconstruct the feature. These studies particularly found that by using the neighbor visual words of a local feature, the feature can be better reconstructed. Therefore, in this paper, we mainly consider a local feature and its neighbor visual words in feature encoding.

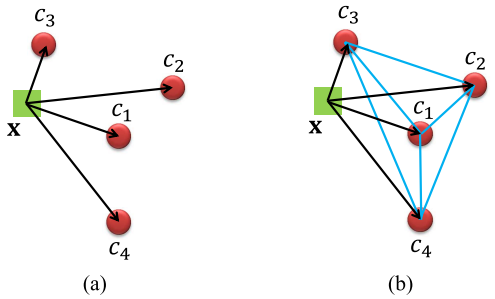

Fig. 1. An illustration of higher-order relationship of visual words for feature encoding. (a) Traditional feature encoding method. (b) feature encoding with higher-order relationship.

One missing element is the inner relationship between visual words, which has been demonstrated to effectively enhance the robustness of feature encoding in studies in physiology [10], [11]. In traditional feature encoding, we cannot ensure that the visual words generated by k-means will always be meaningful. It is likely that some visual words will be very noisy, thus local features will not be reconstructed accurately. To reduce the influence of noise, we consider the higher-order topology (here topology refers to relationship) of visual words. The advantage of this relationship is that it is difficult for noisy visual words to have a strong relationship, while it is easy for meaningful words. Therefore, the influence of noisy visual words can be overcome, and local features can be reconstructed more accurately. Fig.1 gives an example, in which the relationship between the local feature and visual words are considered (black arrows), while the inner relationship of visual words are missing (blue lines). In this paper, we denote this inner relationship as the fully connected network of these neighbor words. A number of previous studies have attempted to consider this relationship in both image space and feature space [12]–[19], but they all have two main limitations: the low-order relationship and high computational cost.

The first limitation is the low-order relationship. Here, "order" means the number of visual words in the relationship, *e.g.,* the order of 2 generates the doublets of words while the

Manuscript received September 11, 2014; revised September 12, 2014, March 20, 2015, and June 11, 2015; accepted June 22, 2015. Date of publication June 23, 2015; date of current version July 13, 2015. This work was supported in part by the National Basic Research Program of China under Grant 2012CB316302, in part by the National Natural Science Foundation of China under Grant 61322209 and Grant 61175007, and in part by the Grant of China Scholarship Council, Australian Research Council under Project DP-120103730 and Project FT-130101457. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Anuj Srivastava.

K. Huang and C. Wang are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China (e-mail: kqhuang@nlpr.ia.ac.cn; cwang@nlpr.ia.ac.cn).

D. Tao is with the Centre for Quantum Computation and Intelligent Systems, University of Technology at Sydney, Sydney, NSW 2007, Australia (e-mail: dacheng.tao@uts.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2449081

order of 3 gives the triplets. Most previous studies consider the visual word relationship in a pairwise way, *e.g.,* $c_1 - c_2$ and $c_2 - c_3$ in Fig.1. However, physiological studies show that the higher-order word relationship are closer in nature to human's recognition, and the pairwise relationship is easily affected by image noise under large image variations. It is therefore difficult to achieve a robust pairwise relationship in feature encoding, and there is a desperate need to focus on higher-order word relationship. In this paper, we focus on higher-order relationship, in which the order can be arbitrarily high and not limited by the number of neighbor words. Even if only the nearest word is considered, we can still use higher-order word relationship to enhance feature encoding.

Another limitation is the high computational cost. A small number of studies have attempted to consider the higher-order relationship, as the number of potential pairings grows exponentially with the number of words, the task of learning becomes computationally expensive. For example in Fig.1, there are 6 and 4 possible word combinations for the order of 2 and 3 respectively. However, we have found that not all visual words should be related to one another, *e.g.,* c_3 and c_4 are far apart and should be removed. On this basis, we consider the removal of unreasonable relationship by proposing an effective word relationship search method.

In this paper, to overcome these two limitations, we propose an efficient classifcation framework that exploits higher-order topology of visual words (HTVWs) in the feature space. The framework has two important steps: (1) *Relationship Searching*: for each visual word, we search its dependence between visual words based on distance and orientation. The dependence is used to construct higher-order topology or relationship in the feature space. (2) *Feature Encoding*: for each local feature, the higher-oder topology is used to better encode the feature. To validate the effectiveness and efficiency, the model is evaluated on four typical datasets [20]–[23], namely PASCAL VOC 2007, 15 Scenes, Caltech 101 and UIUC Sport Event. Results show that the higher-order relationship can improve most recent feature encoding methods consistently and efficiently, and the best performance of the methods of word relationship on these four datasets is reported.

The rest of this paper is organized as follows. We first review related work of visual word relationship in Sec. II. Then, we introduce how to search the related words and construct the higher-order relationship in Sec. III. To validate the effectiveness, detailed experiments and main results are provided in Sec. IV. Finally, we summarize the paper with conclusive remarks in Sec. V.

II. Related Work

The studies on modeling the relationship of visual words can be categorized as two main methods. The first one considers the relationship in image space. Morioka and Satoh [19] combined each pair of spatially close visual features together, so that each pair is treated as a data point in a joint pairwise feature space. However, this pairwise relationship only considers the distance but neglects the orientation, which can be important in describing the distribution of visual words more accurately. Therefore, some studies have proposed modeling the spatial arrangement of pairwise visual words [24]. Morioka and Satoh [25] proposed the directional pairwise visual words, a method which divides the direction of the pairwise words according to four orientations. They obtained promising improvements on several benchmarks when combined with the sparse coding based methods [6], [7]. To better describe the orientation distribution, Khan *et al.* [26] considered the spatial orientations of pairwise words by constructing a graph of the words, in which each edge of the graph represents the orientation of two visual words. To describe the distribution more flexible, Yang and Newsam [27] considered different combinations of various distance and orientation restrictions to better adapt to the dataset. Though some improvements are resulted from this work, the distribution of orientation is easily influenced in large image variations because of extensive image noise. To overcome this, researchers have attempted to use some similarity measures which can preserve spatial proximity. For example, to better measure the similarity between local pairwise words, proximity distribution kernels [14] have been used to incorporate geometric information. This study attempts to consider the triplet relationship of visual words, while the triplet version is only straightforward without learning. Inspired by the proximity distribution kernel, Morioka and Satoh [15] proposed the compact correlation coding which has better scale-invariant and robust representation for capturing rich spatial proximity information.

The second method mainly focuses on the relationship of visual words in the feature space. Based on the frequency occurrence of each visual word in the feature space, Winn *et al.* [17] merged pairwise visual words using supervised mapping to encode features discriminatively. Similarly, based on the category distributions in the feature space, Li and Kweon [12] modeled the conceptual relationshios of pairwise words and the discriminative pairs were selected by distributional similarity. To model this semantic relationship more discriminatively, Lazebnik *et al.* [28] proposed the consideration of the category distributions under a maximum entropy framework. These methods have achieved some improvements, although studies show that combining pairwise words in feature space will cause synonymy and polysemy in feature encoding. To overcome this problem, Yuan *et al.* [16] and Zheng *et al.* [18] combined pairwise words to visual phrases and visual synsets respectively. As a result, the semantic of the pairwise words is much stronger for overcoming synonymy and polysemy in feature encoding. However, large image variations will also result in a high amount of noise. To describe the pairwise relationship robustly, Liu *et al.* [13] proposed embedding the pairwise features into a semantic low-dimensional feature space based on diffusion distance, and then adopting kmeans clustering to remove the noise and obtain a semantic pairwise relationship.

Though the above methods have achieved improvements on previous efforts, there are two major challenges remain. First, most previous studies considered the pairwise relationship, which is a low-order relationship. Under large image

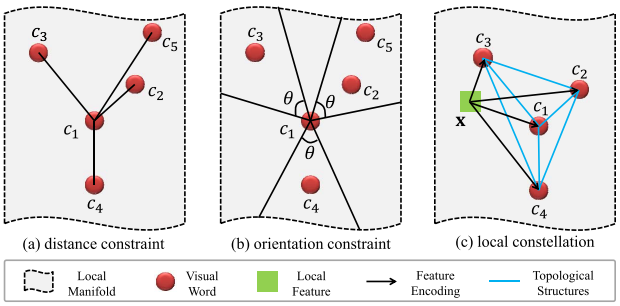

Fig. 2. An illustration of the relationship searching based on distance and orientation constraints. (a) Distance constraint: each related word should not be too far away. (b) Orientation constraint: each related word should occupy a non-overlapping and independent area. (c) The higher-order relationship of visual words.

variations, it is usually difficult for the low-order pairwise relationship to capture the intrinsic invariance of intra-class images. Therefore, it is necessary to consider higher-order relationship to preserve invariance, *e.g.,* this relationship based on the triplets or quadruplets of visual words. Second, previous work has considered that all visual words should be related to one another, but a number of unreasonable relationship exist which make it computationally demanding to consider each combination. Therefore, how to construct the higher-order relationship of visual words has become a challenging problem. In this paper, we propose an efficient classification framework to learn the higher-order relationship of visual words in the feature space.

III. Encoding With Higher-Order Relationship

In this section, we introduce how to learn the higher-order relationship of visual words by an efficient relationshiop searching method. We first give the motivation for the relationship searching in Sec. III-A. Then, in Sec. III-B and Sec. III-C, the relationship searching and feature encoding are elaborated. Lastly, an efficiency analysis is given in Sec.III-D.

A. Motivation

In traditional feature encoding, only the nearest word of a local feature is considered, but it is difficult to reconstruct the feature accurately with one word. Some recent studies have found that locality can better reconstruct the feature, thus the neighbor words of a local feature are considered, *e.g.,* soft coding, sparse coding and fisher vector. However, these studies only consider the relationship between the local feature and its neighbor words, and the inner relationship between visual words are missing. We are therefore motivated to use the higher-order relationship of visual words to enhance feature encoding. The neighbor words for the local feature $\mathbf{x}$ in Fig.2(c) are $\mathbf{c_1}$,...,$\mathbf{c_1}$. The higher-order relationship, as the the fully connected network, between these words is denoted by the blue lines shown in Fig.2(c).

Previous work considers that all the visual words should be related to one another, but a troubesome relationship exists which makes it computationally challenging to consider each combination. To overcome this problem, we pose the question: *instead of considering all the combinations, can we use rules to remove the troublesome relationship*? Starting with this basic idea, we propose the relationship searching method, which is capable of learning higher-order relationship very efficiently.

Algorithm 1 Searching Strategy of $\mathbf{c_i}'$s Related Words

Initialization:

$C = \{\mathbf{c_1}, \mathbf{c_2}, ..., \mathbf{c_M}\},\ B_i = C - \{\mathbf{c_i}\},\ S_i = \{\phi\};$

$\forall\{\mathbf{c_t}\} \in B_i,\ \mathbf{c_{i1}} = \arg\min_{\mathbf{c_t}} \|\mathbf{c_i} - \mathbf{c_t}\|_2;$

$B_i = B_i - \{\mathbf{c_{i1}}\}, S_i = S_i \cup \{\mathbf{c_{i1}}\};$

Iteration:

for $j = 2 : M - 1$ **do**

  $\forall\{\mathbf{c_t}\} \in B_i,\ \mathbf{c_j} = \arg\min_{\mathbf{c_t}} \|\mathbf{c_i} - \mathbf{c_t}\|_2;$

  $B_i = B_i - \{\mathbf{c_j}\};$

  if $\forall\{\mathbf{c_k}\} \in S_i,\ \|\mathbf{c_i} - \mathbf{c_j}\|_2 < \tau\|\mathbf{c_i} - \mathbf{c_{i1}}\|_2 \cap \Delta(\mathbf{c_i}, \mathbf{c_j}, \mathbf{c_k}) > \theta$ **then**

    $S_i = S_i \cup \{\mathbf{c_j}\};$

  end if

end for

B. Relationship Searching

Let $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}] \in \Re^{D\times N}$ be a set of N local features extracted from an image, *e.g.,* SIFT or HOG features [29], [30], $\mathbf{C} = [\mathbf{c_1},\mathbf{c_2},...,\mathbf{c_M}] \in \Re^{D\times M}$ be a visual vocabulary with M visual words. Inspired by similarity modeling in feature encoding, we know that distance and orientation are two important factors, [19], [25]–[27], [31], thus we adopt them as two fundamental principles in searching related words. The basic idea of the algorithm is that for a word $\mathbf{c_i}$, each related word $\mathbf{c_j}$ should be close and occupy an independent area, *i.e.,* the relationship should satisfy the distance and orientation constraints, which are respectively given as follows:

$$\|\mathbf{c_i} - \mathbf{c_j}\|_2 < \tau\|\mathbf{c_i} - \mathbf{c_{i1}}\|_2, \quad s.t.\ j \neq i, \tag{1}$$

$$\Delta(\mathbf{c_i}, \mathbf{c_j}, \mathbf{c_k}) > \theta, \quad s.t.\ j \neq i, \quad \forall \mathbf{c_k} \in S_i. \tag{2}$$

Eqn. (1) indicates that each related word $\mathbf{c_j}$ must be close to $\mathbf{c_i}$, in which $\mathbf{c_{i1}}$ is the nearest word of $\mathbf{c_i}$, and τ is a constant controlling the tolerable distance. The example in Fig. 2(a) shows a situation in which $\mathbf{c_5}$ is too far from $\mathbf{c_1}$, thus it cannot satisfy the distance constraint. Eqn. (2) indicates that each related word should have a non-overlapping and independent area, in which S_i is the set of $\mathbf{c_i}'s$ related words, and θ is a constant determining the orientation of each area, as shown in Fig. 2(b). Furthermore, $\Delta(\mathbf{c_i}, \mathbf{c_j}, \mathbf{c_k})$ denotes the angle between $\mathbf{c_j}-\mathbf{c_i}$ and $\mathbf{c_k}-\mathbf{c_i}$, as defined in Eqn. (3). Fig. 2(b) shows that $\mathbf{c_5}$ cannot satisfy the orientation constraint because it locates in the area of $\mathbf{c_2}$.

$$\Delta(\mathbf{c_i}, \mathbf{c_j}, \mathbf{c_k}) = \arccos \frac{\langle \mathbf{c_j}-\mathbf{c_i}, \mathbf{c_k}-\mathbf{c_i} \rangle}{\|\mathbf{c_j}-\mathbf{c_i}\|_2 \bullet \|\mathbf{c_k}-\mathbf{c_i}\|_2}. \tag{3}$$

Based on the above two constraints, we give the relationship searching algorithm in Alg. 1. For each word $\mathbf{c_i}$, all other

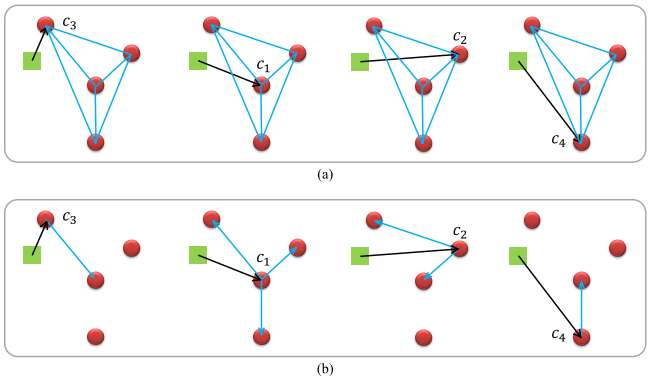

Fig. 3. (a) The higher-order relationship in previous studies; (b) the simplified higher-order relationship in our method.

words are first arranged by ascending Euclidean distance, and the nearest $\mathbf{c_{i1}}$ is added to S_i. Then, other words are checked in order according to the distance and orientation constraints, and the word that satisfies both constraints is added to S_i. Finally, the set of $\mathbf{c_i}'$s related words S_i is obtained. By applying the algorithm to $\mathbf{c_1}$,,...,$\mathbf{c_M}$, the set of related words $\mathbf{S}$ for the vocabulary $\mathbf{C}$ is defined as

$$\begin{aligned} \mathbf{S} &= [\mathbf{S_1}, \mathbf{S_2}, ..., \mathbf{S_M}] \in \Re^{D \times \sum_{i=1}^{M} Q_i} \\ \forall\, \mathbf{S_i} &= \left[\mathbf{c_{i1}}, \mathbf{c_{i2}}, ..., \mathbf{c_{iQ_i}}\right] \in \Re^{D \times Q_i}, \end{aligned} \tag{4}$$

wherein Q_i denotes the number of the related words for $\mathbf{c_i}$, and $\mathbf{c_{ij}}$ is the j^{th} dependent word of $\mathbf{c_i}$. With the related words obtained, the fully connected network can be transformed into Fig.3(b). The troublesome relationship in previous studies, shown in Fig.3(a), is removed to improve efficiency.

C. Feature Encoding

In this section, we define the higher-order relationship and use it in various feature encoding methods. We take sparse coding as an example. General sparse coding can be formulated as follows:

$$\mathbf{R_i} = \arg\min_{\mathbf{R_i}} \left\| \mathbf{x_i} - \mathbf{C}\mathbf{R_i}^T \right\|_2^2 + \lambda T\left(\mathbf{R_i}\right), \tag{5}$$

where $\mathbf{R_i} = \left[(r_1)_i, (r_2)_i, ..., (r_M)_i\right] \in \Re^{1 \times M}$ represents the feature quantization of $\mathbf{x_i}$ on the vocabulary $\mathbf{C}$, $T\left(\mathbf{R_i}\right)$ denotes an arbitrary constraint of sparsity, and λ is a sparsity coefficient of the constraint. To use the higher-order relationship in sparse coding, we first define the relationship. In this paper, the higher-order relationship of words are defined as $\Delta\left(\mathbf{C}, \mathbf{x_i}, \mathbf{S}\right)$, which can be factorized as follows:

$$\begin{aligned} \Delta\left(\mathbf{C}, \mathbf{x_i}, \mathbf{S}\right) &\in \Re^{1 \times \sum_{k=1}^{M} Q_k} \\ &= \left[\Delta\left(\mathbf{c_1}, \mathbf{x_i}, \mathbf{S_1}\right), ..., \Delta\left(\mathbf{c_M}, \mathbf{x_i}, \mathbf{S_M}\right)\right]. \end{aligned} \tag{6}$$

In Eqn.7, the higher-order relationship of visual words $\Delta\left(\mathbf{C}, \mathbf{x_i}, \mathbf{S}\right)$ is constructed by $\Delta\left(\mathbf{c_k}, \mathbf{x_i}, \mathbf{S_k}\right), \forall k = 1, ..., M$, which is defined as:

$$\begin{aligned} \Delta\left(\mathbf{c_k}, \mathbf{x_i}, \mathbf{S_k}\right) &\in \Re^{1 \times Q_k} \\ &= \left[\Delta\left(\mathbf{c_k}, \mathbf{x_i}, \mathbf{c_{k1}}\right), ..., \Delta\left(\mathbf{c_k}, \mathbf{x_i}, \mathbf{c_{kQ_k}}\right)\right], \end{aligned} \tag{7}$$

in which $\Delta\left(\mathbf{c_k}, \mathbf{x_i}, \mathbf{c_{k1}}\right)$ has been defined before in Eqn.3. Based on Eqn.7, we can see that the higher-order relationship of visual words in this paper is actually the orientation relationship among the related neighbor words. Distance and orientation are two important factors for describing the distribution of visual words. Distance is already considered in the neighbor words, thus we only need to use the orientations to describe the distribution more accurately.

With the consideration of this higher-order relationship of words, we enhance the general sparse coding as follows:

$$\begin{aligned} \mathbf{V_i} = \arg\min_{\mathbf{V_i}} &\left\| \mathbf{x_i} - \mathbf{S}\mathbf{V_i}^T \right\|_2^2 \\ &+ \left[\lambda T\left(\mathbf{V_i}\right) + \alpha \left\| \mathbf{V_i} \odot \Delta\left(\mathbf{C}, \mathbf{x_i}, \mathbf{S}\right) \right\|_2^2\right], \end{aligned} \tag{8}$$

wherein $\mathbf{S} = [\mathbf{S_1}, \mathbf{S_2}, ..., \mathbf{S_M}]$ is the set of dependent words for vocabulary $\mathbf{C}$, as shown in Eqn. (4), and $\mathbf{V_i}$ is the feature quantization of $\mathbf{x_i}$ on $\mathbf{S}$:

$$\begin{aligned} \mathbf{V_i} &= \left[(\mathbf{V_1})_\mathbf{i}, ..., (\mathbf{V_M})_\mathbf{i}\right] \in \Re^{1 \times \sum_{k=1}^{M} Q_k} \\ \forall\ (\mathbf{V_k})_\mathbf{i} &= \left[(v_{k1})_i, ..., \left(v_{kQ_k}\right)_i\right] \in \Re^{1 \times Q_k}. \end{aligned} \tag{9}$$

In Eqn. (8), α is the coefficient of the higher-order relationship constraint, and $\odot$ denotes the element-wise multiplication.

Following the feature encoding step, the image representation $\mathbf{F} \in \Re^{1 \times \sum_{k=1}^{M} Q_k}$ can be constructed by operating maximum pooling [7], [32] over all the N local features. This is shown in Eqn. (10), wherein $\max_{column}$ preserves the maximum quantization of each column.

$$\mathbf{F} = \max_{column}\left[\mathbf{V_1}^\mathbf{T}, \mathbf{V_2}^\mathbf{T}, ..., \mathbf{V_N}^\mathbf{T}\right]^T. \tag{10}$$

In particularly, general sparse coding with the higher-order relationship has feasible solutions. According to lasso [33] and elastic net [34], they can solve the optimization with $L1$ regularization, $L2$ regularization or the combination of both regularizations. These two regularizations have been successfully used in sparse coding based methods, *e.g.,* sparse coding (SC) [7], local-constrained linear coding (LLC) [6] and over-complete sparse coding (OSC) [35]. In this paper, the penalization of higher-order relationship is considered as L2 regularization, thus the relationship model can be combined easily with many sparse coding based methods. It is known that the relationship model can not only encode visual feature precisely, but also retain the efficiency of the sparse coding based methods.

Other encoding methods are also very popular. We consider four methods: two typical methods are hard quantization (HQ) [4] and soft quantization (SQ) [36], and the other two recent methods are super vector coding (SVC) [8] and improved fisher kernel (IFK) [9].

The higher-order topology (or higher-order relationship) model is used in HQ and SQ as follows:

$$\begin{aligned} (r_k)_i &= \frac{\exp\left(-\sigma_d \times \left\|\mathbf{x_i} - \mathbf{c_k}\right\|_2^2\right)}{\sum_{k=1}^{K} \exp\left(-\sigma_d \times \left\|\mathbf{x_i} - \mathbf{c_k}\right\|_2^2\right)} \\ \left(v_{kq}\right)_i &= (r_k)_i \times \exp\left(-\sigma_\Delta \times \Delta\left(\mathbf{c_k}, \mathbf{x_i}, \mathbf{c_{kq}}\right)^2\right), \end{aligned} \tag{11}$$

in which σ_d and σ_Δ represent the Gaussian smoothing factors of distance and orientation respectively, and r_k denotes the feature quantization on vocabulary $\mathbf{C}$. Specifically, if $K = 1$, Eqn. (11) is for HQ, otherwise it is for SQ, *e.g.,* $K = 5$ is validated to perform the best [36].

Due to the similar formulation of SVC and IFK, we only list the higher-order relationship for IFK:

$$(\mathbf{r_k})_\mathbf{i} = \sum_{k=1}^{K} \phi_i(k) \left[\frac{\mathbf{x_i} - \mathbf{c_k}}{\sigma_\mathbf{k}}, \frac{(\mathbf{x_i} - \mathbf{c_k})^2}{{\sigma_\mathbf{k}}^2} - 1 \right]$$
$$\left(\mathbf{v_{kq}}\right)_\mathbf{i} = (\mathbf{r_k})_\mathbf{i} \times \exp\left(-\sigma_\Delta \times \Delta\left(\mathbf{c_k}, \mathbf{x_i}, \mathbf{c_{kq}}\right)^2\right) \quad (12)$$

wherein $\phi_i(k)$ and $\sigma_\mathbf{k}$ are $\mathbf{c_k}$'s cluster weight and covariance matrix respectively [9]. One difference in Eqn. (12) is that $\mathbf{r_k}$ and $\mathbf{v_{kq}}$ are all vectors which are determined by the difference encoding in SVC and IFK.

D. Efficiency Analysis

In this section, we give the efficiency analysis of the relationship searching and feature encoding. Previous work models the pairwise relationship based on the occurrence frequency of similar local features. If we consider the higher-order relationship of these previous methods, the frequency of occurrence of each diffenrent feature will lead to a different relationship. Therefore, the computational complexity of the previous methods is about $O(NM)$, in which N and M are the number of local features and visual words respectively. However, usually N is much larger than M because of the dense feature extraction, which will make the relationship construction very slow. Compared to the previous studies, the proposed relationship searching only considers about M, thus the complexity is about $O(M^2 logM)$, which significantly improves efficiency. Our method is comparable with the original coding methods in respect of the efficiency of the coding stage. In the sparse coding-based methods, we use the LARS algorithm in the SPAM toolbox to solve the optimization with lasso or elastic net. Our method may be a little slow compared to the original sparse coding, because the local feature is encoded by the set of dependent words S, which is usually K times larger than the vocabulary size C. However, the efficiency of the other coding methods is comparable to the original coding methods as a result of the simple multiplication operation. the efficiency of the higher-order relationship is higher than in previous studies in respect of the overall consideration of the trelationship construction and feature encoding,

IV. Experiments and Results

To demonstrate the effectiveness and efficiency of the higher-order topology of visual words (HTVWs) model, here we call it the higher-order relationship (HOR) simply. We evaluate it in this section on four typical datasets of different categories and levels of difficulty. We first give the overall results, and then evaluate the higher-order relationship, feature coding and efficiency respectively.

TABLE I
The Detailed Settings of Vocabulary Construction, Coding and Pooling Methods. K Is the Number of the Nearest Visual Words for a Local Feature in Feature Encoding

Coding	K	Vocabulary	Vocabulary Size	Pooling
HQ	1	K-means	$16 \sim 8192$	Maximum
SQ	5	K-means	$16 \sim 8192$	Maximum
LLC	5	K-means	$16 \sim 8192$	Maximum
OSC	-	K-means + GMM	$16 \sim 8192$	Maximum
SVC	1	K-means	$16 \sim 256$	Average
IFK	5	K-means + GMM	$16 \sim 256$	Average

A. Detailed Settings

1) Datasets and Evaluation: We use four typical datasets: 15 Scenes, Caltech 101, UIUC Sport Event and PASCAL VOC 2007. We use common settings on the first three datasets: 100 training images and the rest for testing on 15 Scenes; 30 training images and the remainder (at most 50) for testing on Caltech 101; 70 training images and 60 for testing on UIUC Sport Event. Experiments are implemented over 10 random splits of the data, and the mean accuracy and standard deviation are reported. The PASCAL VOC 2007 dataset incorporates 20 object categories and includes 5011 training/validation and 4952 testing images. Average precision is reported.

2) Descriptor: We use SIFT local features generated by the VLFeat toolbox [37]. The features are extracted from local patches densely located at every 4 pixels on an image under three feature scales, defined by setting the width of the SIFT spatial bins to 4, 6 and 8 pixels. The low contrast SIFT features for other options are detected and dropped when the magnitude is below a certain threshold, based on the average magnitude of gradient. Tthe *'fast'* option is selected to achieve a faster extraction algorithm.

3) Vocabulary: Vocabulary size is set to be $[2^4, ..., 2^{13}]$, that is,ten scales in total. The sizes for SVC and IFK are set to be $[16, 32, 64, 128, 256]$ because they construct image representation in a much larger dimensional space. We construct these vocabularies by standard k-means, in which 10 million random local features are used to guarantee the effectiveness of the clustering. Additionally,k-means is used to initialize the encoding methods which use Gaussian Mixture Model (GMM) and vocabulary learning [6], [9], [35], as shown in Table.I.

4) Coding and Pooling: We use six popular encoding methods and two important pooling methods, and the detailed settings are shown in Table. I. The coding methods are HQ, SQ, LLC, OSC, SVC and IFK, in which the number of the nearest words K is set as 5 in SQ, LLC and IFK for the best performance [6], [36], [38], and two cluster mixtures are used in OSC [35]. We use two polling methods: average pooling for SVC and IFK [8], [9], and maximum pooling is used for others to achieve better performance [6], [32], [36].

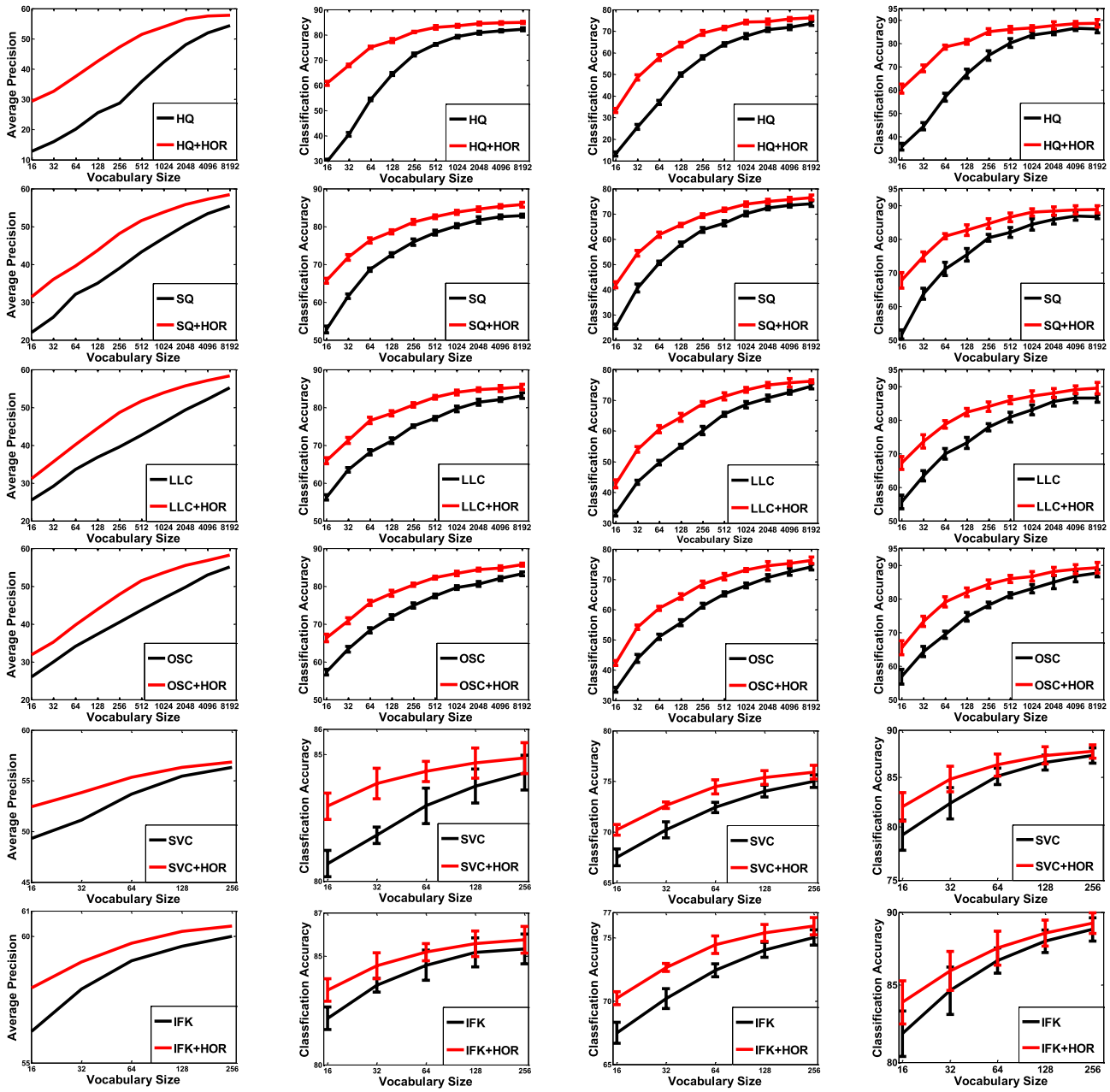

Fig. 4. The performance of the four typical feature encoding methods by using higher-order relationship (HOR) on four popular datasets. The feature encoding methods are Hard Quantization (HQ), Soft Quantization (SQ) [36], Local- constrained Linear Coding (LLC) [6], Over-Complete Sparse Coding (OSC) [35], Super Vector Coding (SVC) [8] and Improved Fisher Kernel (IFK) [9]. Datasets (from left to right): Col.1: PASCAL VOC 2007; Col.2: 15 Scenes; Col.3: Caltech 101; Col.4: UIUC Sport Event.

5) Higher-Order Relationship: The parameters in the relationship searching algorithm of the distance constraint τ and the orientation constraint θ in Eqn. (1) and Eqn. (2) are first set as 2.0 and 120° respectively by parameter selection in Sec.IV-E1. Then, for the feature encoding in Eqn. (8), the penalty coefficients of the sparsity λ and the higher-order relationship α are determined by a five-fold cross-validation, which gives $\lambda = 0.1$ and $\alpha = 0.05$. Lastly, we set the Gaussian smoothing factor σ_d and the number of a visual word's related words Q_i in Eqn. (11) as 10 and 3 respectively, accoding to the parameter selection in Sec.IV-E2.

6) Partition and Training: For the 15 Scenes, Caltech 101 and UIUC Sport Event datasets, we use the common settings of the spatial pyramid matching (SPM) [21] under the image partitions of 1×1, 2×2 and 4×4; while for PASCAL VOC 2007, the partition of 1×1, 2×2 and 3×1 is used instead [6], [38], [39]. In the training phase, all classifiers are trained by the rule of one-vs-all, and the lLIBLINEAR with a five-fold validation is adopted for high efficiency, low memory cost and good generalization.

B. Overall Results

In this section, we give the overall results of the HOR in many recent encoding methods. Fig. 4 shows the classification precision of six encoding methods. Based on these results, there are three main observations.

TABLE II

AVERAGE PRECISION OF THE HIGHER-ORDER RELATIONSHIP AND THE PAIRWISE METHODS ON PASCAL VOC 2007. DLPC [25] AND DLPB [25] ARE THE DIRECTIONAL LOCAL PAIRWISE RELATIONSHIP, AND HYBRID [15] IS THE COMBINATION OF CCC [15] AND LPC [15]

Methods	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
LPC	68.8	55.3	42.8	65.8	19.7	60.5	73.8	54.0	51.5	39.2
DLPC	66.9	56.3	40.3	63.0	21.7	56.9	72.9	52.1	50.0	35.9
DLPB	71.2	60.5	44.8	63.9	20.0	56.5	74.0	54.2	50.4	41.9
CCC	72.3	56.9	43.3	62.1	23.9	58.9	74.1	58.2	51.8	40.5
Hybrid	72.5	59.4	47.8	66.8	27.1	63.6	76.9	58.2	54.6	41.7
Ours	**72.5**	**64.2**	**47.9**	**70.0**	**27.3**	**65.8**	**78.6**	**60.3**	**56.1**	**48.4**

Methods	table	dog	horse	motorbike	person	plant	sheep	sofa	train	tvmonitor	Mean AP
LPC	46.1	43.5	73.2	59.8	82.3	21.0	38.0	48.0	72.2	48.7	53.2
DLPC	46.6	40.8	75.3	60.0	81.4	25.4	39.4	48.4	70.7	47.3	52.6
DLPB	47.1	42.5	75.6	60.6	81.2	24.2	42.8	51.5	69.0	48.8	54.0
CCC	49.9	41.7	73.6	59.9	80.8	24.2	39.8	43.0	74.1	48.9	53.9
Hybrid	54.6	45.9	76.6	62.6	83.3	28.8	42.4	50.6	77.1	51.8	57.1
Ours	52.6	44.3	**78.0**	**66.8**	**84.6**	**29.2**	**48.6**	**55.1**	76.4	**55.0**	**59.1**

The HOR model is easily applied in recent feature encoding methods, and consistently enhances them under all vocabulary sizes. On all the datesets, HOR achieves improvement of at least 2% over the six encoding methods, which demonstrates its strong generalization. HOR can also obtain large and stable improvements under both small and large vocabulary sizes, demonstrating the robustness of the higher-order relationship.

The classification precision of HQ, SQ, LLC and MSC for small vocabulary sizes, *e.g.,* smaller than 256, can largely be improved by the higher-order relationship model. On all four datasets, HOR obtains an improvement between 5% and 10%, *e.g.,* HOR improves HQ by more than 10% at a size of 256 on PASCAL VOC 2007; under the same size, OSC is improved by almost 5% on Caltech 101. The reason for this large improvement may be that the original encoding for small sizes will lose most relationship information because only a few visual words are sparsely distributed, but the relationship model can take full advantage of the local and global information among these few words for better image representation.

Under large vocabulary sizes, *e.g.,* larger than 4096, HOR shows a stable improvement over the encoding methods of HQ, SQ, LLC and MSC. Specifically, HOR consistently improves feature encoding between 2% and 4%, *e.g.,* HOR improves LLC by almost 4% at the size of 8192 on PASCAL VOC 2007; under the same size, SQ is improved by 3% on 15 Scenes. The reason for a stable improvement rather than large improvement may be that there are many more visual words in large vocabularies, and their distribution will become much denser. As a result, the higher-order relationship model will concentrate on the local information and lose some meaningful global cues, thus the improvement will not be as large as it is in the case of small vocabularies.

Table. II and Table. III show the classification precision of the original feature encoding methods, pairwise relationship methods and HOR on the four datasets. It can be clearly observed that in all the testing datasets, HOR obtains the highest classification precision. On the challenging PASCAL VOC 2007 database, HOR obtains the highest performance on 17 object categories, and it achieves average precision of 59.1%, which is 2% higher than CCC [15] and more than 5% higher than the other pairwise relationship methods. The improvements are particularly impressive in certain categories, *e.g.,* 7% on cow, 6% on sheep and 5% on sofa.

TABLE III

CLASSIFICATION ACCURACY OF THE FEATURE ENCODING METHODS, THE LOCAL PAIRWISE RELATIONSHIP METHODS AND THE HIGHER-ORDER TOPOLOGY MODEL ON 15 SCENES, CALTECH 101 AND UIUC SPORT EVENT DATASETS

Methods	15 Scenes	Caltech 101	UIUC Sport
HQ [32], [38]	80.8 ± 0.4	74.41 ± 1.04	86.62 ± 1.14
SQ [36], [38]	83.0 ± 0.7	75.93 ± 0.57	86.92 ± 1.31
SC [32]	84.1 ± 0.5	73.2 ± 0.54	-
LLC [6], [38]	83.45 ± 0.5	73.44	86.52 ± 1.3
OSC [35]	83.25 ± 0.5	74.09 ± 0.95	87.62 ± 1.13
SVC [8], [38]	84.78 ± 0.61	75.52 ± 1.03	87.26 ± 0.81
IFK [9], [38]	85.34 ± 0.68	77.78	88.80 ± 0.81
LPC [19]	83.40 ± 0.58	72.94 ± 0.54	-
DLPC [25]	83.80 ± 0.32	71.60 ± 0.65	-
DLPB [25]	85.22 ± 0.55	77.63 ± 0.63	-
CR [31]	82.9 ± 0.86	74.25	-
CCC [15]	-	-	-
Ours	$\mathbf{85.76 \pm 0.43}$	$\mathbf{77.78 \pm 0.86}$	$\mathbf{89.23 \pm 0.57}$

For the other three datasets, HOR also demostrates the best results in the evaluation, as listed in Table. III. On the 15 Scenes database specially, HOR obtains classification accuracy of 85.76%, which beats DPLB and is more than 2% higher than other encoding and topological methods. On the UIUC Sport Event, the accuracy of 89.23% obtained by HOR is an improvement over IFK and is about 2% higher than other methods. Lastly, on the Caltech 101 dataset, HOR achieves the accuracy of 77.78%, which is much higher than most of feature encoding and pairwise topological methods.

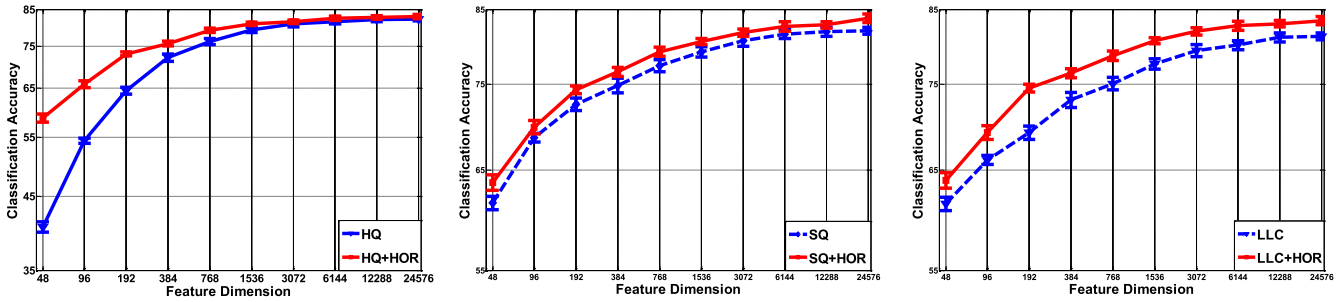

Fig. 5. The comparison of serval feature encoding methods under the codebook with higher-order relationship and the codebook with equal feature dimension on the 15 Scenes dataset.

In summary, the higher-order relationship model is effective in datasets with different categories and levels of difficulty, and it better represents images than the original feature encoding methods and the pairwise relationship methods.

C. Higher-Order Evaluation

In this section, we evaluate the influence of higher-order relationship model in codebook generation. Here, we use HQ, SQ and LLC for testing on the 15 Scenes dataset. We use the codebook with the same dimensionality for fair comparison, due to the higher feature dimensionality of the proposed feature representation, Fig.5 shows the classification accuracy of these three encoding methods under a series of feature dimension. It can be seen that the higher-order relationship model achieves consistent improvement, which demonstrates that the relationship searching is more effective in encoding methods than enlarging the size of the codebook.

D. Efficiency Evaluation

In this section, we show that the higher-order relationship model can be very efficient in enhancing image classification. For the detailed experimental settings, we compare the HOR model with the most efficient pairwise relationship method LPC [19], and the computational complexity, which contains the relationship construction and feature encoding, will be evaluated individually. The setup of feature extraction and vocabulary clustering is the same as the settings in Sec. IV-A, and the number of the nearest neighbors in the feature encoding is set as 5 in both methods. Efficiency is evaluated on the UIUC Sport Event database, and we use the vocabulary sizes of 16 $\sim$ 8192 based on soft quantization (SQ). All the experiments are implemented on a computer server with an Intel X5650 CPU (2.67 GHz and 12 cores), and a combination of $Matlab/C++$ is in programming.

Fig. 6 and Fig. 7 show the efficiency (s) of the relationship construction and the feature encoding respectively. The relationship construction contains the pairwise feature construction and vocabulary clustering on all the images of the dataset, while distance computation and feature assignment are considered in feature encoding only for one image. It can be clearly observed in Fig. 6 that the relationship construction in the pairwise method costs much more time than the

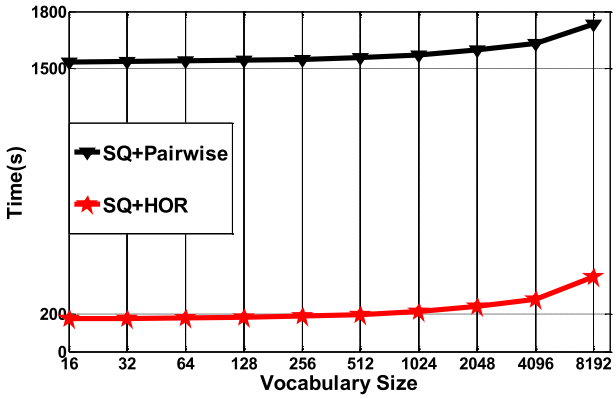

Fig. 6. Efficiency comparison of the relationship construction on all the images of the UIUC Sport Event database.

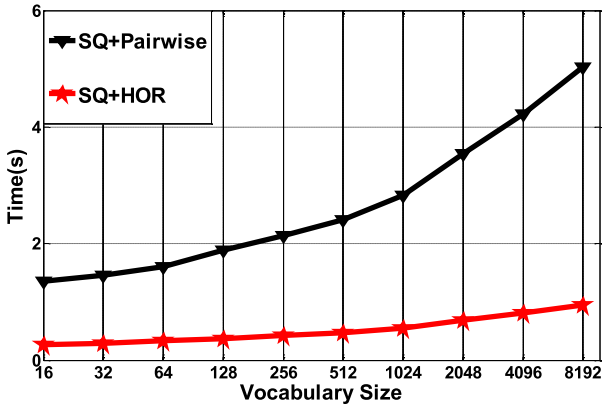

Fig. 7. Efficiency comparison of the feature encoding on the same image of the UIUC Sport Event database.

proposed model, *e.g.,* the LPC costs 1500s while the proposed model costs only 200s, which is approximately 7 times less. The reason for this is that the pairwise methods have to construct the pairwise local features, which is time-consuming and cost about 1350s in our experimental setup. In contrast, the proposed model operates directly on the original single local features, and the self-incremental relationship searching algorithm is also very efficient, costs only 50s at the size of 8k. Fig. 7 shows that the HOR model is much more efficient in the feature encoding than the pairwise method. We can see that across all vocabulary sizes, the feature encoding by HOR is no more than 1s, and has the potential

TABLE IV

CLASSIFICATION ACCURACY UNDER DIFFERENT COMBINATIONS OF THE DISTANCE CONSTRAINT(τ) AND THE ORIENTATION CONSTRAINTS (θ) ON THE UIUC SPORT EVENT DATABASE. THE VOCABULARY SIZE IS SET AS 1024 AND HQ IS SELECTED FOR EVALUATION

Accuracy	$\tau = 1.2$	$\tau = 1.4$	$\tau = 1.6$	$\tau = 1.8$	$\tau = 2.0$	$\tau = 2.2$	$\tau = 2.4$	$\tau = 2.6$	$\tau = 2.8$	$\tau = 3.0$
$\theta = 30°$	85.81	85.71	86.15	**85.33**	85.25	84.60	86.21	85.06	84.23	85.27
$\theta = 60°$	85.46	85.85	84.85	85.75	85.60	**86.67**	86.33	85.57	86.52	85.96
$\theta = 90°$	85.42	84.96	85.08	85.43	**85.56**	85.42	85.17	85.23	85.15	85.13
$\theta = 120°$	85.23	85.84	85.52	85.62	**87.43**	84.13	85.56	85.38	85.96	85.31
$\theta = 150°$	85.00	85.77	85.35	**86.18**	85.70	84.97	85.08	85.42	84.79	85.29

to be applied in practical conditions. However, the pariwise method costs almost 5 times more than HOR because each feature has to generate 5 pairwise features, and encoding these more pairwise features will be time-consuming. As a result, all these experiments demonstrate that the higher-order relationship model can be very efficient in enhancing image classification.

E. Parameter Selection

In this subsection, we attempt to give some practical guidelines of the higher-order relationship model by parameter selection. We study the impact of parameters from two aspects: relationship searching and feature encoding.

1) Relationship Searching: In this section, we attempt to select two parameters in the phase of relationship searching: the distance constraint τ in Eqn. (1) and the orientation constraint θ in Eqn. (2). The distance constraint τ for these two parameters is set as $[1.2, 1, 4, ..., 2.8, 3]$, ten values in total; while the orientation constraint θ is set as $[30°, 60°, 90°, 120°, 150°]$. We test all the above combinations at the size of 1024 on the UIUC Sport Event dataset, and Table. IV shows the classification accuracy of all these combinations based on hard quantization (HQ). It is evident that the best accuracy 87.43 occurs at $\theta = 120°$ and $\tau = 2.0$, thus we use these two values on all of our experiments. Another important observation is that for each orientation, most of the high accuracy is concentrated on the middle part of the distance constraint. The probable reason for this is that the small τ just generates a few related words, which cannot well capture the local relation; In contrast, if τ is too large, there will be too many dependent words which may lead to over-fitting of the proposed model.

2) Feature Encoding: We attempt to select two important parameters in the feature encoding: the number of dependent words for each visual word Q_i in Eqn. (8) and the Gaussian smoothing factor σ_Δ in Eqn. (11). They are evaluated based on soft quantization (SQ) at the vocabulary size of 512 on the UIUC Sport Event dataset. The classification accuracy for Q_k is shown in Fig. 8 using 1 to 20 respectively. It can be seen that $Q_k = 3$ obtains the best performance, and the advantage is that the smaller Q_k leads to faster computation and less memory usage. If Q_k is too large, the division of dependent area will be too fine, which may lead to over-fitting. For σ_Δ, as shown in Fig. 9, $\sigma_d = 10$ obtains the best performance. If σ_Δ is too large or too small, the quantization $\mathbf{V_i}$ will be either too smooth or too sharp.

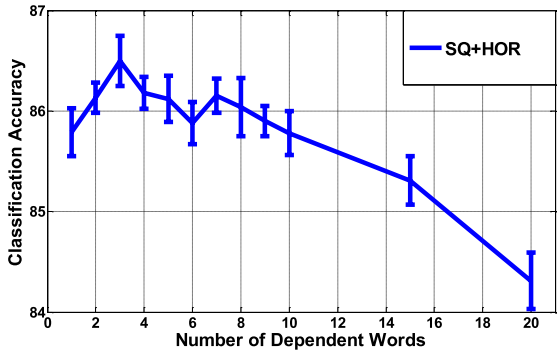

Fig. 8. Classification accuracy based on the different values of the number of dependent words Q_k.

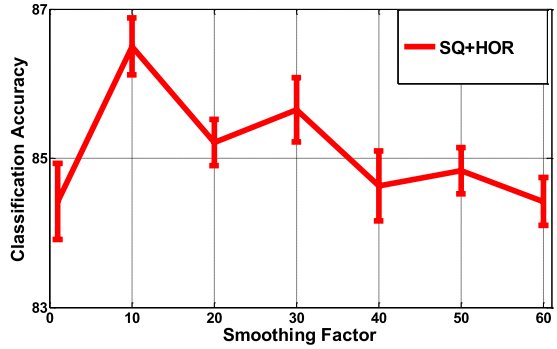

Fig. 9. Classification accuracy based on the different values of the Gaussian smoothing factor σ_Δ.

V. CONCLUSIONS

In this paper, the higher-order relationship of visual words has been studied in feature encoding for image classification, and we have proposed an efficient classification framework. The framework has two important steps, namely relationship searching and feature encoding. The framework has two main steps: (1) For each visual word, we search its related words

based on distance and orientation. With the related words obtained, the higher-order relationship can be learned much more efficiently; (2) For each local feature, we use its neighbor visual words and their higher-order relationship jointly to encode the feature. Experiments on four typical datasets, namely PASCAL VOC 2007, 15 Scenes, Caltech 101 and UIUC Sport Event, demonstrate that the proposed model consistently and efficiently enhances most recent feature encoding methods, and achieves the best performance of all word relationship methods on these four datasets.

REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *Assoc. Comput. Mach. Comput. Surv.*, vol. 40, no. 5, pp. 1–60, 2008.

[2] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, May 2001.

[3] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 34, no. 3, pp. 334–352, Aug. 2004.

[4] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2004, pp. 1–16.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Jun. 2007, pp. 248–255.

[6] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.

[7] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1794–1801.

[8] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 141–154.

[9] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.

[10] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas, "Nonlinear dimensionality reduction techniques for classification and visualization," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 645–651.

[11] H. S. Seung and D. D. Lee, "The manifold ways of perception," *Science*, vol. 290, no. 5500, pp. 2268–2269, 2000.

[12] T. Li and I.-S. Kweon, "Measuring conceptual relation of visual words for visual categorization," in *Proc. 16th IEEE Int. Conf. Image Process.*, Nov. 2009, pp. 2057–2060.

[13] J. Liu, Y. Yang, and M. Shah, "Learning semantic visual vocabularies using diffusion distance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 461–468.

[14] H. Ling and S. Soatto, "Proximity distribution kernels for geometric context in category recognition," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[15] N. Morioka and S. Satoh, "Compact correlation coding for visual object categorization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1639–1646.

[16] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: From visual words to visual phrases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[17] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 1800–1807.

[18] Y.-T. Zheng, M. Zhao, S.-Y. Neo, T.-S. Chua, and Q. Tian, "Visual synset: Towards a higher-level visual representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[19] N. Morioka and S. Satoh, "Building compact local pairwise codebook with joint feature space clustering," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 692–705.

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[21] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2169–2178.

[22] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Los Alamitos, CA, USA, Jun. 2004, pp. 178–186.

[23] L.-J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[24] O. A. B. Penatti, E. Valle, and R. da S. Torres, "Encoding spatial arrangement of visual words," in *Proc. Iberoamer. Congr. Pattern Recognit.*, 2011, pp. 240–247.

[25] N. Morioka and S. Satoh, "Learning directional local pairwise bases with sparse coding," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 32.1–32.11.

[26] R. Khan, C. Barat, D. Muselet, and C. Ducottet, "Spatial orientations of visual word pairs to improve bag-of-visual-words model," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 89.1–89.11.

[27] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1465–1472.

[28] S. Lazebnik, C. Schmid, and J. Ponce, "A maximum entropy framework for part-based texture and object recognition," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 832–838.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[31] Y. Huang, K. Huang, C. Wang, and T. Tan, "Exploring relations of visual codes for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1649–1656.

[32] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. CVPR*, Jun. 2010, pp. 2559–2566.

[33] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[34] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc. Ser. B, Statist. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005.

[35] J. Yang, K. Yu, and T. Huang, "Efficient highly over-complete sparse coding using a mixture model," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 113–126.

[36] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2486–2493.

[37] A. Vedaldi and B. Fulkerson. (2008). *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. [Online]. Available: http://www.vlfeat.org/

[38] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: An evaluation of recent feature encoding methods," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 76.1–76.12.

[39] Y. Huang, K. Huang, Y. Yu, and T. Tan, "Salient coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1753–1760.

Kaiqi Huang (SM'09) is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, China. His current research interests include computer vision, pattern recognition, and biological-based vision. He has authored over 100 papers in important international journals and conference, such as the IEEE TIPAMI, T-IP, T-SMC-B, TCSVT, *Pattern Recognition*, Computer Vision and Image Understanding, ECCV, CVPR, ICIP, and ICPR. He received the Best Student Paper Awards from ACPR10, the winner prizes of the detection task in both PASCAL VOC10 and PASCAL VOC11, the Honorable Mention Prize of the classification task in PASCAL VOC11, and the Winner Prize of classification task with additional data in ILSVRC 2014. He was the Deputy General Secretary of the IEEE Beijing Section (2006–2008).

Chong Wang received the B.Sc. from the Beijing University of Posts and Telecommunications, China. He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, China. His current research interests include pattern recognition, computer vision, and machine learning. He has published several top and international conference and journal papers in ICIP, ICPR, ECCV, and TIP. In 2010 and 2011, he has participated in the famous PASCAL VOC challenge and won prizes in both years. Besides, he also won the championship of the classification task with additional data in ILSVRC 2014.

Dacheng Tao (F'15) is currently a Professor of Computer Science with the Centre for Quantum Computation & Intelligent Systems, and the Faculty of Engineering and Information Technology, University of Technology, Sydney. He mainly applies statistics and mathematics to data analytics problems. His research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and over 100 publications in prestigious journals and prominent conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the *Journal of Machine Learning Research*, the *International Journal of Computer Vision*, the Conference on Neural Information Processing Systems, the International Conference on Machine Learning, the Computer Vision and Pattern Recognition Conference, the International Conference on Computer Vision, the European Conference on Computer Vision, the International Conference on Artificial Intelligence and Statistics, the International Conference on Data Mining, and ACM Conference on Knowledge Discovery and Data Mining. He received several best paper awards, such as the Best Theory/Algorithm Paper Runner Up Award in the IEEE ICDM'07, the Best Student Paper Award in the IEEE ICDM'13, and the 2014 ICDM 10 Year Highest-Impact Paper Award.