Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Bilevel Visual Words Coding for Image Classification

Jiemi Zhang, Chenxia Wu, Deng Cai and Jianke Zhu
The State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, China
jmzhang10@gmail.com, chenxiawu0@gmail.com, dengcai@cad.zju.edu.cn, jkzhu@zju.edu.cn

Abstract

Bag-of-Words approach has played an important role in recent works for image classification. In consideration of efficiency, most methods use k-means clustering to generate the codebook. The obtained codebooks often lose the cluster size and shape information with distortion errors and low discriminative power. Though some efforts have been made to optimize codebook in sparse coding, they usually incur higher computational cost. Moreover, they ignore the correlations between codes in the following coding stage, that leads to low discriminative power of the final representation. In this paper, we propose a bilevel visual words coding approach in consideration of representation ability, discriminative power and efficiency. In the bilevel codebook generation stage, k-means and an efficient spectral clustering are respectively run in each level by taking both class information and the shapes of each visual word cluster into account. To obtain discriminative representation in the coding stage, we design a certain localized coding rule with bilevel codebook to select local bases. To further achieve an efficient coding referring to this rule, an online method is proposed to efficiently learn a projection of local descriptor to the visual words in the codebook. After projection, coding can be efficiently completed by a low dimensional localized soft-assignment. Experimental results show that our proposed bilevel visual words coding approach outperforms the state-of-the-art approaches for image classification.

1 Introduction

Bag-of-Words (BoW) approach has received widely attention as the state-of-the-art for image classification [Yu *et al.*, 2009; Wang *et al.*, 2010; Liu *et al.*, 2011; Ponce *et al.*, 2011]. The key steps of BoW model are local feature extraction, codebook generation, feature coding and pooling. Different approaches have worked on each step to improve both its generative property to describe images accurately and its discriminatory power for classification. Despite remarkable progresses, there still remain challenges and disputes especially on effective codebook generation and reasonable coding.

A codebook is a set of visual words that further used as the bases for coding. Most coding methods use k-means clustering to generate the codebook for its efficiency [Sivic and Zisserman, 2003; Yu *et al.*, 2009; Wang *et al.*, 2010; Liu *et al.*, 2011]. The obtained codebooks often lose the cluster size and shape information with distortion errors and low discriminative power [Lazebnik and Raginsky, 2009; van Gemert *et al.*, 2010]. To optimize codebook, sparse coding uses the unsupervised learning to learn an over-complete codebook ensuring sparse representation of local descriptors [Olshausen and Fieldt, 1997; Mairal *et al.*, 2008]. Some supervised learning methods are also proposed to improve the discriminative power of the codebook [Lazebnik and Raginsky, 2009; Boureau *et al.*, 2010]. However, these approaches are too computationally expensive to be scalable to large-scale problems [Shabou and LeBorgne, 2012]. Moreover, in the following coding stage, sparse coding usually tends to select quite different bases for similar descriptors to favor sparsity. As a result, the correlations between codes are overlooked that leads to low discriminative power of the final representation [Yu *et al.*, 2009]. In contrast, it has been shown that localized coding with several nearest local bases would lead to more effective representations with better discriminative power [Wang *et al.*, 2010; Liu *et al.*, 2011].

In this paper, we propose a bilevel visual words coding approach in consideration of representation ability, discriminative power and efficiency. In the bilevel codebook generation stage, the first level visual words are firstly generated using k-means separately in each class to synthesize each class information. Then an efficient spectral clustering is run on the first level visual words to better capture the shapes of visual words and acquire better discriminative power. The output cluster centers with the lower dimensionality consist the second level visual words as the final codebook. To obtain discriminative representation in the coding stage, we design a certain localized coding rule with bilevel codebook to select local bases. To further achieve an efficient coding referring to this rule, an online method is proposed to efficiently learn a projection of local descriptor to the visual words in the codebook. After projection, coding can be efficiently completed by firstly projecting the local descriptor to the same dimensionality with the codebook, secondly coding with a localized

soft-assignment. Experimental results on two typical datasets demonstrate that our proposed coding method with bilevel visual words outperforms the state-of-the-art approaches for image classification.

2 Backgrounds

Considerable research efforts have been devoted to coding techniques in the area of artificial intelligence [Li *et al.*, 2009; Pan *et al.*, 2010; Gong and Zhang, 2011]. To review the state-of-the-art coding methods, we first give some notations. Let us consider a codebook of visual words denoted as $B = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_m] \in \mathbb{R}^{d\times m}$, where d is the dimensionality of a local descriptor and m is the codebook size. Let $\mathbf{x}_i \in \mathbb{R}^d$ be the i-th local descriptor in an image and $\mathbf{z}_i$ be the coding coefficient vector of $\mathbf{x}_i$, with z_{ij} being the coefficient with respect to visual word $\mathbf{b}_j$.

The codebook B usually consists of the cluster centers generated by the efficient k-means clustering. Then the common objective of coding methods is to quantize the local descriptor to visual words by the least square fitting:

$$\arg\min_{\mathbf{z}_i} \|\mathbf{x}_i - B\mathbf{z}_i\|_2^2. \tag{1}$$

Different constraints or regularizers for this objective lead to different coding methods. The traditional Vector Quantization (VQ) [Cosman *et al.*, 1996] used the non-negative constraints $\|\mathbf{z}_i\|_{\ell_1} = 1$ and $\mathbf{z}_i \succeq 0$ to force the coding weight to be 1, and the cardinality constraint $\|\mathbf{z}_i\|_{\ell_0} = 1$ to make the coefficient only have one non-zero element. Then the quantization could be simply performed by a hard assignment: each local descriptor is assigned to the nearest visual word. This coding method would lead to unreasonable mapping results because of the variability, such as image noise, varying scene illumination and non-affine changes in the measurement regions. Any of these will make the mapping results drift a lot from the original one [Philbin *et al.*, 2008].

Sparse Coding (SC) [Lee *et al.*, 2006; Mairal *et al.*, 2008] is one alternative method which yields better results in many applications. It utilizes the idea of soft-assignment and uses a sparsity regularization term (*i.e.* the ℓ_1 norm of $\mathbf{z}_i$) to relax the constraint $\|\mathbf{z}_i\|_{\ell_0} = 1$:

$$\arg\min_{B,\mathbf{z}_i} \|\mathbf{x}_i - B\mathbf{z}_i\|_2^2 + \lambda\|\mathbf{z}_i\|_{\ell_1}. \tag{2}$$

As the codebook B is usually over-complete, the sparsity regularization is necessary to ensure that the under-determined system has an unique solution. Compared with vector quantization approach, sparse coding achieves much less quantization error. However, it is very likely to select quite different visual words for similar descriptors to favor sparsity, leading to a weak correlations between codes [Wang *et al.*, 2010].

Unlike the sparse coding, Locality-constrained Linear Coding (LLC) [Wang *et al.*, 2010] enforces locality instead of sparsity, which leads to smaller coefficient for the basis vectors farther from a local descriptor $\mathbf{x}_i$:

$$\begin{aligned} &\arg\min_{\mathbf{z}_i} \|\mathbf{x}_i - B\mathbf{z}_i\|_2^2 + \lambda\|\mathbf{d}_i \odot \mathbf{z}_i\|_2^2 \\ &s.t. \quad \mathbf{1}^T\mathbf{z}_i = 1, \end{aligned} \tag{3}$$

where $\odot$ denotes the element-wise multiplication, and $\mathbf{d}_i \in \mathbb{R}^m$ is the locality adaptor that gives different freedom for each visual word proportional to its similarity to the input descriptor $\mathbf{x}_i$. Specifically, $\mathbf{d}_i = [\exp(\frac{dist(\mathbf{x}_i,\mathbf{b}_1)}{\delta}), \ldots, \exp(\frac{dist(\mathbf{x}_i,\mathbf{b}_m)}{\delta})]^T$, where $dist(,)$ denotes the Euclidean distance and δ is a positive parameter adjusting the weight decay speed for the locality adaptor. In practice, a approximation is proposed to improve the computational efficiency by ignoring the second term in Eq. (3). It directly selects the k nearest visual words of $\mathbf{x}_i$ to minimize the first term by solving a much smaller linear system. This gives the coding coefficient for the selected k visual words and other coefficients are set to zero.

3 Coding with Bilevel Visual Words

In this section, we will introduce a novel coding method for image classification using bilevel visual words, that can better reflect the shapes of visual words with better discriminative power. We firstly introduce the bilevel structure to generate codebook, where k-means and an efficient spectral clustering are run in each level by taking both class information and the shapes of each visual word cluster into account. A corresponding localized coding rule is then designed with bilevel codebook to select local bases. To achieve a faster coding, an online projection learning method is further proposed to learn the localized coding rule between local descriptor and the visual words from the bilevel structure.

3.1 Codebook Generation

In most previous works, in consideration of efficiency, the initial codebook usually consists of the cluster centers obtained by k-means clustering method on randomly selected local descriptors from the training images. The obtained codebooks often lose the cluster size and shape information with distortion errors and low discriminative power [Lazebnik and Raginsky, 2009; van Gemert *et al.*, 2010].

To deal with this weakness while preserving the efficiency, we propose to generate the codebook in two levels as illustrated in Fig. 1. Firstly k-means is run on local descriptors from each class separately. After that, we have h cluster centers from each class. The total $n = h \times c$ cluster centers consist the first level visual words $B_1 \in \mathbb{R}^{d\times n} = [\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, ..., \tilde{\mathbf{b}}_m]$ if we have c classes image dataset. Secondly, we run an efficient spectral clustering [Ng *et al.*, 2001] on these first level visual words to generate the second level visual words as the final codebook $B_2 \in \mathbb{R}^{d_e\times m} = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_m]$, each column of which is the cluster center output by the spectral clustering with the lower dimensionality d_e.

As we know, spectral clustering can adapt to a wider range of geometries and detect non-convex patterns and linearly non-separable clusters [Ng *et al.*, 2001; Filippone *et al.*, 2008]. Also it can detect the cluster (arbitrary shapes) structure of data [Cai *et al.*, 2010]. So it shows better performance than k-means. The spectral clustering can be thought as a two-step approach [Belkin and Niyogi, 2001]. The first step is unfolding the data manifold using the manifold learning algorithms and the second step is performing typically k-means on the flat embedding for the data points. Therefore, the ob-

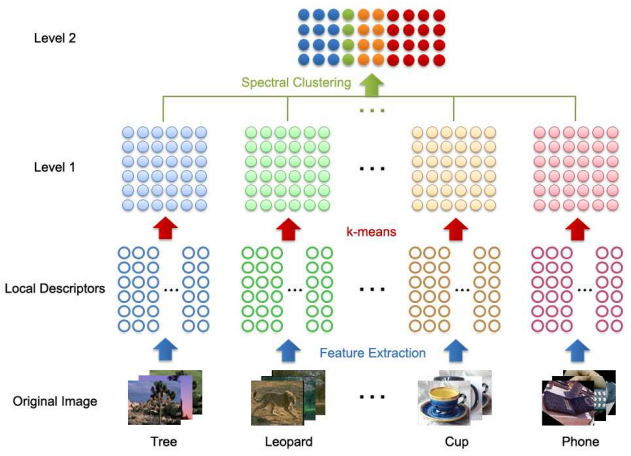


Figure 1: Bilevel visual words generation process. Firstly k-means is run on local descriptors from each class separately, where the output cluster centers from each class consist the first level visual words. Secondly an efficient spectral clustering is run on the first level visual words, where the output cluster centers with lower dimensionality consist the second level visual words.

tained lower dimensional cluster centers can better reflect the cluster information. In the first step, we usually embed the data by retaining the top eigenvectors of a graph Laplacian, which is defined on the affinity matrix of data points. Since it is computational infeasible to directly build the large affinity matrix for all local descriptors from the training dataset, we instead run the spectral clustering on the first level visual words, which synthesize each class information.

Efficient Spectral Clustering

Consider a graph with n vertices where each vertex corresponds to a first level visual word. We define the affinity matrix W on the graph as the dot-product weighting, *i.e.*, $W_{ij} = \tilde{\mathbf{b}}_i^\top \tilde{\mathbf{b}}_j$ and $W = B_1^\top B_1$. The Laplacian matrix $L = D - W$, where D is the degree matrix whose diagonal elements are the column/row sums of W, $D = \sum_j W_{ij}$ and other elements are all zero. It is easy to check that the eigenvectors of the normalized Laplacian $D^{-1/2}LD^{-1/2}$ corresponding to the smallest eigenvalues are the same as the eigenvectors of $D^{-1/2}WD^{-1/2}$ corresponding to the largest eigenvalues [Ng *et al.*, 2001]. Then the embeddings can be obtained by solving the eigen-decomposition of $D^{-1/2}WD^{-1/2} = \hat{B}_1^\top \hat{B}_1 \in \mathbb{R}^{n\times n}$, where $\hat{B}_1 = B_1 D^{-1/2}$. As we know, it is commonly an $\mathbb{O}(n^3)$ problem that is not scalable to the number of first level visual words.

As $\hat{B}_1^\top \hat{B}_1$ shares the non-zero eignvalues with $\hat{B}_1 \hat{B}_1^\top \in \mathbb{R}^{d\times d}$, we can solve the eign-problem of $\hat{B}_1 \hat{B}_1^\top$ instead, which takes only $\mathbb{O}(d^3 + d_e dn)$, that is only linear to the number of first level visual words. Since the dimensionality of the used local descriptor d (128 for SIFT and 64 for SURF) is much smaller than n, the computational time could be $\mathbb{O}(d_e dn)$, much less than $\mathbb{O}(n^3)$. In detail, we denote a diagonal matrix $\Sigma \in \mathbb{R}^{d_e\times d_e}$ whose diagonal elements are the selected top d_e eigenvalues as $\sigma_1, \sigma_2, \cdots, \sigma_{d_e}$ where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{d_e} \geq 0$. The corresponding top eignvectors of $\hat{B}_1^\top \hat{B}_1$ and $\hat{B}_1 \hat{B}_1^\top$ are $V \in \mathbb{R}^{d_e\times n}$ and $U \in \mathbb{R}^{d_e\times d}$. Considering the singular value decomposition of $\hat{B}_1$, the eigenvectors V and U have the following relationship:

$$V = \Sigma^{-\frac{1}{2}} U \hat{B}_1. \tag{4}$$

Thus, we could firstly get U by $\mathbb{O}(d^3)$ computations for eign-decomposition of $\hat{B}_1 \hat{B}_1^\top$ and get the final embeddings of the first level visual words $V = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n]$ by $\mathbb{O}(d_e dn)$ computations for above equation. After that, we run k-means on V to m clusters with d_e-dimensional cluster centers as the second level visual words B_2 consisting our final codebook.

3.2 Coding Rule

In this section, we will discuss how to code the local descriptor on bilevel visual words. Since the dimensionality are different between the second level visual words and the local descriptor, we cannot directly use the existing localized coding methods for single level. A simple solution is to learn a linear projection $P \in \mathbb{R}^{d\times d_e}$ by a regression between V and $P^\top B_1$, then project the local descriptor into the same subspace of second level visual words for coding:

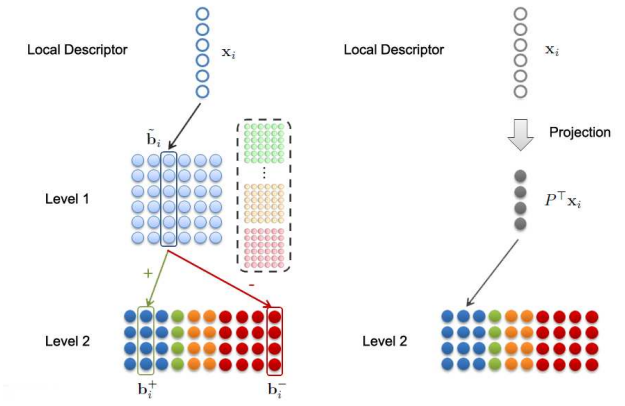
$$\arg\min_{\mathbf{z}_i} \|P^\top \mathbf{x}_i - B_2 \mathbf{z}_i\|_2^2. \tag{5}$$

However, directly using this projection does not consider the localized coding rule, *i.e.*, selecting the several nearest visual words as bases. Most previous works have demonstrated that carefully choosing the visual words (local bases) using this localized rule to quantize the descriptor is the key to the success of coding [Liu *et al.*, 2011; Wang *et al.*, 2010]. Similarly, we can give a localized rule for the bilevel structure as in Fig. 2(a). A local descriptor $\mathbf{x}_i$ can be firstly mapped to its nearest first level visual word $\tilde{\mathbf{b}}_i$ in its own class:

$$\begin{aligned} \tilde{\mathbf{b}}_i &= \arg\min_{\tilde{\mathbf{b}}_l} \|\tilde{\mathbf{b}}_l - \mathbf{x}_i\|_2^2 \\ s.t. \quad & y(\tilde{\mathbf{b}}_l) = y(\mathbf{x}_i), \end{aligned} \tag{6}$$

where $y(\mathbf{x}_i), y(\tilde{\mathbf{b}}_l)$ denotes the class label they come from. Then $\mathbf{x}_i$ is further mapped to those near second level visual words according to the distances between the second level visual words and $\mathbf{v}_i$, the embedding of $\tilde{\mathbf{b}}_i$.

However, for a novel input descriptor, we cannot know its class label that $\tilde{\mathbf{b}}_i$ cannot be obtained. Though it could be solved by a nearest neighbor search among all first level visual words, it incurs much more distance computations as well. To both consider the computation efficiency and the localized coding rule. We propose an online method to learn the linear projection P of the local descriptor to the second level visual words based on the localized coding rule for the bilevel structure. In detail, for a certain local descriptor $\mathbf{x}_i$, we denote its ideal second level visual word for coding as $\mathbf{b}_i^+$ and in contrast $\mathbf{b}_i^-$ indicates the second level visual word.



(a) Localized bilevel coding rule. (b) Coding with projection.

Figure 2: A projection is learned referring to the localized bilevel coding rule: a certain local descriptor $\mathbf{x}_i$ is firstly mapped to its nearest first level visual word $\tilde{\mathbf{b}}_i$ and further mapped to those near second level visual words $\mathbf{b}_i^+$ according to the distances between the second level visual words and the embedding of $\tilde{\mathbf{b}}_i$. Then coding can be completed by firstly projecting the local descriptor into the same dimensionality with the second level visual words.

$\mathbf{b}_i^+, \mathbf{b}_i^-$ can be chosen according to the localized coding rule for the bilevel visual words as illustrated in Fig. 2(a):

$$\mathbf{b}_i^+ \in \{\mathbf{b}_l | \mathbf{b}_l \in \mathbb{N}_k(\mathbf{v}_i)\}, \quad \mathbf{b}_i^- \in \{\mathbf{b}_l | \mathbf{b}_l \notin \mathbb{N}_k(\mathbf{v}_i)\}; \quad (7)$$

where $\mathbf{v}_i$ is the embedding of the nearest first level visual word $\tilde{\mathbf{b}}_i$ of $\mathbf{x}_i$ in its class and $\mathbb{N}_k(\mathbf{v}_i)$ denotes the k-nearest second level visual word of $\mathbf{v}_i$.

After projection, we want the distance of $P^\top \mathbf{x}_i$ to $\mathbf{b}_i^+$ is smaller than that to $\mathbf{b}_i^-$:

$$\|P^\top \mathbf{x}_i - \mathbf{b}_i^-\|_2^2 - \|P^\top \mathbf{x}_i - \mathbf{b}_i^+\|_2^2 > \eta, \quad (8)$$

where η is a positive parameter which controls the safe distance. To give a more effective loss function, we define the hinge loss:

$$l_P(\mathbf{x}_i, \mathbf{b}_i^+, \mathbf{b}_i^-) = \max(0, \eta + \|P^\top \mathbf{x}_i - \mathbf{b}_i^+\|_2^2 - \|P^\top \mathbf{x}_i - \mathbf{b}_i^-\|_2^2). \quad (9)$$

After further expanding items in $\|.\|_2^2$, we could have the final compact form:

$$\begin{aligned} &l_P(\mathbf{x}_i, \mathbf{b}_i^+, \mathbf{b}_i^-) = \max(0, r_i - 2\mathbf{x}_i^\top P(\mathbf{b}_i^+ - \mathbf{b}_i^-)), \\ &r_i = \eta + \|\mathbf{b}_i^+\|_2^2 - \|\mathbf{b}_i^-\|_2^2. \end{aligned} \quad (10)$$

3.3 Projection Learning

To minimize the global loss L_P, we propose an algorithm inspired by the Passive-Aggressive family of algorithms [Crammer *et al.*, 2006]. The algorithm iteratively learns the projection matrix P for each triplet $(\mathbf{x}_i, \mathbf{b}_i^+, \mathbf{b}_i^-)$. In each iteration, we solve the following convex problem with a soft margin:

$$\begin{aligned} P^t = \arg\min_P \frac{1}{2}\|P - P^{t-1}\|_{\mathbb{F}}^2 + \alpha\xi \\ s.t. \quad l_P(\mathbf{x}_i, \mathbf{b}_i^+, \mathbf{b}_i^-) \le \xi, \ \ \xi \ge 0, \end{aligned} \quad (11)$$

where $\|.\|_{\mathbb{F}}$ is the Frobenius norm, ξ is the slack variable and the parameter α controls the trade-off between changes of P and loss minimization. In t-th iteration, we want to minimize the loss on current triplet $(\mathbf{x}_i, \mathbf{b}_i^+, \mathbf{b}_i^-)$ while P is close to that of the last iteration.

In the following, we try to solve the optimization problem in Eq. (11). It can be easily found that $P^t = P^{t-1}$ satisfying Eq. (11) when $r_i - 2\mathbf{x}_i^\top P(\mathbf{b}_i^+ - \mathbf{b}_i^-) < 0$. Otherwise, $l_P(\mathbf{x}_i, \mathbf{b}_i^+, \mathbf{b}_i^-) = r_i - 2\mathbf{x}_i^\top P(\mathbf{b}_i^+ - \mathbf{b}_i^-)$. Then the Lagrangian $\mathcal{L}_{P,\xi,\beta,\lambda}$ is introduced:

$$\mathcal{L}_{P,\xi,\beta,\lambda} = \frac{1}{2}\|P - P^{t-1}\|^2 + \alpha\xi + \beta(r_i - 2\mathbf{x}_i^\top P(\mathbf{b}_i^+ - \mathbf{b}_i^-) - \xi) - \lambda\xi, \quad (12)$$

where $\beta \ge 0, \lambda \ge 0$ are the Lagrangian multipliers. The optimal solution can be obtained by setting the gradient of $\mathcal{L}_{P,\xi,\beta,\lambda}$ to zero:

$$\begin{aligned} &\frac{\partial \mathcal{L}_{P,\xi,\beta,\lambda}}{\partial P} = P - P^{t-1} - \beta Q_i = 0 \\ &\Rightarrow P = P^{t-1} + \beta Q_i, \end{aligned} \quad (13)$$

where $Q_i = 2\mathbf{x}_i(\mathbf{b}_i^+ - \mathbf{b}_i^-)^\top$. This is the update function of P. Next we need to know the value of β. Firstly we set the derivative of the Lagrangian with respect to ξ to 0:

$$\frac{\partial \mathcal{L}_{W,\xi,\beta,\lambda}}{\partial \xi} = \alpha - \beta - \lambda = 0. \quad (14)$$

According to Eq. (13)(14), we can rewrite the Lagrangian in Eq. (12) as follows:

$$\begin{aligned} \mathcal{L}_\beta &= \frac{1}{2}\beta^2\|Q_i\|^2 + \beta r_i - 2\beta\mathbf{x}_i^\top(P^{t-1} + \beta Q_i)(\mathbf{b}_i^+ - \mathbf{b}_i^-) \\ &= -\frac{1}{2}\beta^2\|Q_i\|^2 + \beta r_i - 2\beta\mathbf{x}_i^\top P^{t-1}(\mathbf{b}_i^+ - \mathbf{b}_i^-). \end{aligned} \quad (15)$$

Then we set the derivative of the Lagrangian with respect to β to 0:

$$\begin{aligned} &\frac{\partial \mathcal{L}_\beta}{\partial \beta} = -\beta\|Q_i\|^2 + r_i - 2\mathbf{x}_i^\top P^{t-1}(\mathbf{b}_i^+ - \mathbf{b}_i^-) = 0 \\ &\Rightarrow \beta = \frac{r_i - 2\mathbf{x}_i^\top P^{t-1}(\mathbf{b}_i^+ - \mathbf{b}_i^-)}{\|Q_i\|^2}. \end{aligned} \quad (16)$$

From Eq. (14) and $\lambda \ge 0$, we can also have $\beta \le \alpha$. As when $r_i - 2\mathbf{x}_i^\top P(\mathbf{b}_i^+ - \mathbf{b}_i^-) < 0$, $P^t = P^{t-1}$, we have $\beta = 0$. Considering all the conditions for β, the final update function should be:

$$\begin{aligned} P &= P^{t-1} + \beta Q_i \\ \beta &= \min\left\{\alpha, \frac{l_{P^{t-1}}(\mathbf{x}_i, \mathbf{b}_i^+, \mathbf{b}_i^-)}{\|Q_i\|^2}\right\}. \end{aligned} \quad (17)$$

It is shown that the accumulative loss of this type of iteratively learning is bounded and likely to be small [Crammer *et al.*, 2006].

Once obtaining the projection matrix P, we can simply solve the coding problem in Eq. (5) as illustrated in Fig. 2(b). Firstly we project the local descriptor using the learned P. Then the several nearest visual words are selected as local

bases according to the Euclidean distances between $P^\top \mathbf{x}_i$ and the visual words in B_2. Finally the coefficient can be obtained by LLC [Wang *et al.*, 2010] or localized soft-assignemnt [Liu *et al.*, 2011]. In practice, we suggest using a localized soft-assignment coding and the coefficient would be:

$$z_{ij} = \frac{exp(-\theta \hat{dist}(P^\top \mathbf{x}_i, \mathbf{b}_j))}{\sum_{l=1}^{m} exp(-\theta \hat{dist}(P^\top \mathbf{x}_i, \mathbf{b}_l))},$$
$$\hat{dist}(P^\top \mathbf{x}_i, \mathbf{b}_l) = \begin{cases} dist(P^\top \mathbf{x}_i, \mathbf{b}_l) & \text{if } \mathbf{b}_l = \mathbb{N}_k(P^\top \mathbf{x}_i), \\ \infty & \text{otherwise.} \end{cases} \tag{18}$$

4 Experiment

In this section, the experiments will verified that the bilevel structure codebook can improve the classification performance and our proposed coding method can achieve the leading classification accuracy and coding efficiency compared with the state-of-the-art approaches.

Two widely used data sets are evaluated: Caltech-101 [Fei-Fei *et al.*, 2004] with 9144 images of 101 categories objects plus one background category and 15-Scenes [Lazebnik *et al.*, 2006] with 4485 images of 15 categories scenes. For each dataset, we randomly sample some images to form the training set for each class and repeat the classification process for 10 times and then the average classification accuracy over all classes under different training set is used for evaluation. The Histogram of Oriented Gradient (HOG) [Dalal and Triggs, 2005] is used as the local descriptor and is extracted using the same method in [Wang *et al.*, 2010]: the features are extracted from patches densely located by every 8 pixels on the image, under three scales, 16×16, 25×25 and 31×31 respectively. The dimension of each HOG descriptor is 128. We also employ the spatial pyramid mathching (SPM) algorithm [Lazebnik *et al.*, 2006] in the pooling stage. The SPM method captures the spatial information: for each spatial subregion, the codes of the descriptors are pooled together to get the corresponding pooled feature. These pooled features from each sub-region are concatenated and normalized as the final image feature representation. In this experiment, we used the max pooling and the ℓ^2 normalization as in [Yang *et al.*, 2009]. Same as [Wang *et al.*, 2010], 4×4, 2×2 and 1×1 sub-regions are used for SPM. The codebook size is set to 2048 for both datasets. For classification task, we use a linear SVM and implement it with the LibSVM toolbox.

We mainly compare the following three coding methods: Localized Soft-Assignment Coding (LSAC) with the codebook obtained by the single level k-means clustering (1km-LSAC) [Liu *et al.*, 2011], LSAC with the codebook obtained by the bilevel k-means clustering (2km-LSAC) and our proposed Bilevel Visual Words Coding (BVWC) approach. For all these three coding methods, we fix the neighborhood size as 5 and θ as 10 suggested in [Liu *et al.*, 2011]. For bilevel structure, we empirically set the number of cluster for each class $h = 500$. For our method, the dimensionality of the second level visual words is set to $d_e = 32$. Moreover, the classification results using several other state-of-the-art coding methods are also compared in our experiments.

Table 1: Image classification results on Caltech-101 dataset

training images for each class	15	30
[Zhang *et al.*, 2006]	59.1	66.20
[Lazebnik *et al.*, 2006]	56.40	64.40
[Greg Griffin and Perona, 2007]	59.0	67.60
[Boiman *et al.*, 2008]	65.00	70.40
[Jain *et al.*, 2008]	61.00	69.10
[Yang *et al.*, 2009]	67.00	73.20
LLC	64.11	71.13
1km-LSAC	64.32	70.13
2km-LSAC	65.11	70.81
BVWC	**67.51**	**74.43**

Table 2: Image classification results on 15-Scenes dataset

training images for each class	50	100
[Lazebnik *et al.*, 2006]	-	81.4
[Yang *et al.*, 2009]	-	80.28
VQ	74.48	78.36
SC	76.34	79.42
LLC	77.35	79.69
1km-LSAC	77.53	81.07
2km-LSAC	78.17	81.76
BVWC	**79.03**	**82.53**

4.1 Results

Table 1 and Table 2 give the classification results with different sizes of training images on two tested datasets. The table is divided into two sections. The bottom section lists the approaches implemented by ourselves. Firstly, we find that coding with bilevel visual words show better performance than that with the single level visual words. It is because class information is firstly synthesized by the first level clustering that leads to less noises in the second level clustering to generate the final codebook. Secondly, the leading performance of our proposed method has demonstrated that our used spectral clustering method outperforms k-means in the second level clustering as the data manifold is learned through the embedding. It also shows our learned projection referring to the localized coding rule for bilevel visual words can map the local descriptor to the desired visual words. The top section of the table lists different versions of the existing coding schemes reported in the literature. Our proposed coding method also outperforms these state-of-the-art approaches.

5 Conclusion

In this paper, we proposed a novel coding method using bilevel visual words. In the codebook generation stage, we generated the first level visual words using k-means separately in each class to synthesize each class information. An efficient spectral clustering was then run on the first level visual words to generate the second level visual words with lower dimensionality, which can better capture the shapes of visual words. We further designed a corresponding localized coding rule with bilevel codebook to select local bases. Based on this rule, an online method was proposed to efficiently learn a projection of local descriptor to the visual words in

the codebook. As a result, coding can be efficiently completed by firstly projecting the local descriptor to the same dimensionality with the codebook, secondly coding with a localized soft-assignment. We have shown the effectiveness and the efficiency of our proposed bilevel visual words coding approach for the image classification in the experiments.

Acknowledgments

This work was supported by the National Basic Research Program of China(973 Program) under Grant 2011CB302206 and National Nature Science Foundation of China (Grant Nos: 61222207, 91120302, 61103105).

References

[Belkin and Niyogi, 2001] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 2001.

[Boiman *et al.*, 2008] Oren Boiman, Eli Shechtman, and Michal Irani. In defense of nearest-neighbor based image classification. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[Boureau *et al.*, 2010] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2010.

[Cai *et al.*, 2010] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *SIGKDD*, 2010.

[Cosman *et al.*, 1996] Pamela C. Cosman, Robert M. Gray, and Martin Vetterli. Vector quantization of image subbands: a survey. *IEEE Transactions on Image Processing*, 5(2):202–225, 1996.

[Crammer *et al.*, 2006] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 2006.

[Dalal and Triggs, 2005] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2005.

[Fei-Fei *et al.*, 2004] 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 *Workshop on Generative-Model Based Vision*, 2004.

[Filippone *et al.*, 2008] Maurizio Filippone, Francesco Camastra, Francesco Masulli, and Stefano Rovetta. A survey of kernel and spectral methods for clustering. *Pattern Recogn.*, 41:176–190, 2008.

[Gong and Zhang, 2011] Pinghua Gong and Changshui Zhang. A fast dual projected newton method for l1-regularized least squares. In *IJCAI*, 2011.

[Greg Griffin and Perona, 2007] Alex Holub Greg Griffin and Pietro Perona. Caltech-256 object category dataset. In *7694*, 2007.

[Jain *et al.*, 2008] Prateek Jain, Brian Kulis, and Kristen Grauman. Fast image search for learned metrics. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[Lazebnik and Raginsky, 2009] Svetlana Lazebnik and Maxim Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 31(7):1294–1309, 2009.

[Lazebnik *et al.*, 2006] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition (CVPR)*, 2006.

[Lee *et al.*, 2006] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient sparse coding algorithms. In *NIPS*, pages 801–808, 2006.

[Li *et al.*, 2009] Bin Li, Qiang Yang, and Xiangyang Xue. Can movies and books collaborate? cross-domain collaborative filtering for sparsity reduction. In *IJCAI*, 2009.

[Liu *et al.*, 2011] Lingqiao Liu, Lei Wang, and Xinwang Liu. In defense of soft-assignment coding. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[Mairal *et al.*, 2008] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[Ng *et al.*, 2001] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2001.

[Olshausen and Fieldt, 1997] Bruno A. Olshausen and David J. Fieldt. Sparse coding with an overcomplete basis set: a strategy employed by v1. *Vision Research*, 37:3311–3325, 1997.

[Pan *et al.*, 2010] Weike Pan, Evan Wei Xiang, Nathan Nan Liu, and Qiang Yang. Transfer learning in collaborative filtering for sparsity reduction. In *AAAI*, 2010.

[Philbin *et al.*, 2008] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[Ponce *et al.*, 2011] Víctor Ponce, Mario Gorga, Xavier Baró, and Sergio Escalera. Human behavior analysis from video data using bag-of-gestures. In *IJCAI*, 2011.

[Shabou and LeBorgne, 2012] A. Shabou and H. LeBorgne. Locality-constrained and spatially regularized coding for scene categorization. In *CVPR*, 2012.

[Sivic and Zisserman, 2003] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2003.

[van Gemert *et al.*, 2010] J.C. van Gemert, C.J. Veenman, A.W.M. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(7):1271 –1283, 2010.

[Wang *et al.*, 2010] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas S. Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367, 2010.

[Yang *et al.*, 2009] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition (CVPR)*, 2009.

[Yu *et al.*, 2009] Kai Yu, Tong Zhang, and Yihong Gong. Nonlinear learning using local coordinate coding. In *NIPS*, pages 2223–2231, 2009.

[Zhang *et al.*, 2006] Hao Zhang, Alexander C. Berg, Michael Maire, and Jitendra Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2006.