Graph-Guided Fusion Penalty Based Sparse Coding for Image Classification

Xiaoshan Yang[1,2], Tianzhu Zhang[1,2], Changsheng Xu[1,2], and Min Xu[3]

[1] Institute of Automation, Chinese Academy of Science, Beijing, China
[2] China-Singapore Institute of Digital Media, Singapore
[3] iNEXT, Faculty of Engineering and IT, University of Technology, Sydney, Australia
{xiaoshan.yang,tzzhang,csxu}@nlpr.ia.ac.cn, Min.Xu@uts.edu.au

Abstract. In image classification, conventional sparse coding only encodes local features independently. As a result, the similar local features may be encoded into code vectors with large discrepancy. This sensitiveness has became the bottleneck of the traditional sparse coding based image classification methods. In this paper, we propose a novel graph-guided fusion penalty based sparse coding method. To alleviate the sensitiveness of the traditional sparse coding, our approach constrains that the similar local features are encoded into similar code vectors. To achieve this goal, we add the popular graph-guided fusion penalty term into the traditional $l1$-regularized sparse coding formulation. Finally, we adopt the multi-task form of the smoothing proximal gradient method to solve our optimization problem efficiently. Experimental results on 3 benchmark datasets demonstrate the effectiveness of our improved sparse coding method in image classification.

Keywords: image classification, sparse coding, smoothing proximal gradient.

1 Introduction

Image classification is one of the most challenging research tasks in computer vision [13,9]. The improvement on image classification can also benefit other useful applications, such as image search and retrieval [14]. A lot of research efforts devoted to image classification have led to significant progress, and the accuracies on several extremely difficult data sets have been increasing year by year. However, compared with the human recognition abilities, there is still a long way to go for a practical image classification algorithm.

Most of the currently used image classification methods are based on Bag-of-Words [19,18,13,9]. The idea behind these methods is borrowed from text retrieval. In text retrieval, words which represent the minimum semantic units are extracted firstly. Then a quantized code vector which is computed as the frequency of the extracted semantic words is used to represent the text semantic information of a document. In computer vision, the semantic words are replaced by visual words which are comprised of clustered centers in local feature space [21,22]. To be simply, there are five main steps for the conventional Bag-of-Words based image classification methods. **(1) Feature extraction.**

B. Huet et al. (Eds.): PCM 2013, LNCS 8294, pp. 475–484, 2013.

Local features, such as SIFT [15] and SURF [3], are densely extracted for a given image. Actually, each local feature is a statistic histogram of gradients in a local image patch. **(2) Codebook design.** k-means clustering is the most widely used method for constructing codebook. Local features extracted from training images are sampled and clustered into thousands of visual words. Many improved methods have been proposed recently. For example, on-line clustering and mean-shift are combined to create much more effective codebook in [10]. **(3) Feature coding.** Different kinds of coding methods were proposed in the past several years. Yang *et al.* [19] proposed a method called ScSPM which introduces sparse regularization to the soft-assignment coding method. This results in competitive image classification accuracies by only using linear SVM [6]. Wang *et al.* [18] further improved ScSPM with a locality constraint, which leads to an analytical solution to the coding problem and a fast approximated solving method. Liu *et al.* [13] proposed a localized soft-assignment coding(LSC) method which adds a locality constraint to the distance function in traditional soft-assignment coding method, and this idea is similar to the salient coding proposed by Huang *et al.* [9]. **(4) Pooling and concatenating.** This is another key step for extracting high level sematic information from local features. There are three mostly used pooling methods, sum-pooling, average-pooling, and max-pooling. The max-pooling method was proved to be much more effective for sparse feature coding [19,18]. Liu *et al.* [13] proposed a new "mix-order" max-pooling method with max-pooling as its special case. Lazebnik *et al.* [11] proposed a concatenating method called SPM(Spatial Pyramid Matching) which takes advantages of global geometric correspondence among code vectors in a given image. **(5) Classifier construction.** This is a canonical problem in machine learning. Most of the sate-of-the-art image classification methods use SVM classifier due to its efficiency and simplicity. There are also some other methods for classification. Yao *et al.* [20] proposed a combined method of random forest and SVM where a linear SVM is used to replace the ordinary weak classifier at each node of random trees.

In the remainder of this paper, we mainly focus on feature coding in image classification. As the most popularly used feature coding method, sparse coding has improved the image classification accuracies tremendously in the past few years. However, there are still many problems in conventional sparse coding [2]. As we know, the bottleneck of the conventional sparse coding is that a small change of the local feature will leads to really large variation of the code vector. A principal reason is that the traditional sparse coding only minimize the reconstruction error of each local feature with a sparsity regularization item independently. In other words, conventional sparse coding can not guarantee that the similar local features are encoded into similar code vectors.

In this paper, in order to break through the bottleneck of the conventional sparse coding mentioned in above paragraph, we propose a graph-guided fusion penalty based sparse coding (GFP-SC) algorithm. We introduce the graph-guided fusion penalty into the traditional sparse coding formulation. Thus the code vectors encoded from similar local features are constrained to be similar.

Then we adopt the smoothing proximal gradient method to solve our optimization problem efficiently. In our implementation, we construct a graph with 4-neighborhood structure to describe the contextual structure information among local features in a given local region. Finally, the linear SVM [6] is applied to perform classification. Extensive experimental results on several popular benchmarks show that our GFP-SC outperforms several recently published methods.

This paper is organized as follows. In Section 2, we summarize the works most related to ours. Our improved sparse coding approach is presented in Sections 3. In Section 4, we report and analyze extensive experimental results. At last, a conclusion to our method is given in Section 5.

2 Related Work

In this section, we review 6 mostly used feature coding methods for image classification. Let $\mathbf{b}_i$ ($\mathbf{b}_i \in \Re^d$) denote a visual word or a basis vector, where d is the same as the dimensionality of the local feature. Matrix $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m]$ denotes the visual codebook. The total number of visual words contained in $\mathbf{B}$ is m. Generally, the codebook is trained in advance by clustering. There are also methods which optimize the codebook in coding process [18,8]. Let $\mathbf{x}_i$ ($\mathbf{x}_i \in \Re^d$) be the ith local feature for an image $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$. Let $\mathbf{v}_i$ ($\mathbf{v}_i \in \Re^m$) be the encoded code vector of $\mathbf{x}_i$, each element $\mathbf{v}_{ij}$ is the coefficient with respect to one visual word $\mathbf{b}_j$ contained in codebook $\mathbf{B}$. Define $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n]$. With these denotations, 6 related coding methods can be uniformly written as follows.

Hard-Assignment Coding: In original Bag-of-Words method [5], the simplest hard-assignment or vector quantization was used to construct a bag of keypoints(original name of visual words). The main idea is to count the number of local features assigned to each clustered keypoint. Given a query local feature $\mathbf{x}_i$, the formally expression can be written as

$$\mathbf{v}_{ij} = \begin{cases} 1 & if \quad j = \underset{j=1,\ldots,m}{\arg\min} \|\mathbf{x}_i - \mathbf{b}_j\|_2^2, \\ 0 & otherwise. \end{cases} \tag{1}$$

Soft-Assignment Coding: A local feature $\mathbf{x}_i$ is assigned to just a single visual word by hard-assignment coding while soft-assignment coding assigns it to all visual words in the codebook $\mathbf{B}$ [17] with different weights,

$$\mathbf{v}_{ij} = \frac{\exp(-\alpha \|\mathbf{x}_i - \mathbf{b}_j\|_2^2)}{\sum_{k=1}^{m} \exp(-\alpha \|\mathbf{x}_i - \mathbf{b}_k\|_2^2)}. \tag{2}$$

Sparse Coding: This method [19] is more complicated than Soft-assignment. An optimization problem with sparse regularization constraint needs to be solved for computing the correspondent weight to each visual word in codebook $\mathbf{B}$. If the codebook $\mathbf{B}$ is trained in advance, the optimization problem can be simplified as

$$\mathbf{v}_i = \underset{\mathbf{v}_i \in \Re^m}{\arg\min} \|\mathbf{x}_i - \mathbf{B}\mathbf{v}_i\|_2^2 + \lambda \|\mathbf{v}_i\|_1. \tag{3}$$

Locality-Constrained Linear Coding: In the method [18], a locality constraint is introduced to the sparse coding method [19] by multiplying a Euclidean distance weight vector to $\mathbf{v}_i$ in the optimization equation,

$$\begin{aligned} &\mathbf{v}_i = \arg\min_{\mathbf{v}_i \in \Re^m} \|\mathbf{x}_i - \mathbf{B}\mathbf{v}_i\|_2^2 + \lambda \|\mathbf{d}_i \odot \mathbf{v}_i\|_2^2 \\ &s.t. \quad \mathbf{1}^T \mathbf{v}_i = 1. \end{aligned} \tag{4}$$

where $\mathbf{d}_i = exp\left(dist\left(\mathbf{x}_i, \mathbf{B}\right)/\delta\right)$, and $\odot$ represents element-wise multiplication.

Localized Soft-Assignment Coding: As proposed by Liu *et al.*[13], localized soft-assignment coding method mainly focuses on the k nearest visual words for each local feature $\mathbf{x}_i$. Here $\mathbf{N}_k\left(\mathbf{x}_i\right)$ is the k-nearest neighbor of $\mathbf{x}_i$.

$$\mathbf{v}_{ij} = \frac{\exp(-\alpha \|\mathbf{x}_i - \mathbf{b}_j\|_2^2)}{\sum_{\mathbf{b}_m \in \mathbf{N}_k(\mathbf{x}_i), \mathbf{b}_m \neq \mathbf{b}_j} \exp(-\alpha \|\mathbf{x}_i - \mathbf{b}_m\|_2^2)}. \tag{5}$$

Laplacian Sparse Coding: Obviously, all the previously mentioned 5 feature coding methods carry out the coding process independently for each local feature. In contrast, as proposed by Gao *et al.*[8], structure information among the local features is combined in the Laplacian sparse coding method. The structure information is used to construct a Laplacian matrix which is added to the optimization problem of the conventional sparse coding method.

$$\begin{aligned} &\arg\min_{\mathbf{B},\mathbf{V}} \|\mathbf{X} - \mathbf{B}\mathbf{V}\|_F^2 + \lambda \|\mathbf{V}\|_1 + \frac{\beta}{2} \sum_{i,j} \|\mathbf{v}_i - \mathbf{v}_j\|^2 w_{ij} \\ &= \arg\min_{\mathbf{B},\mathbf{V}} \|\mathbf{X} - \mathbf{B}\mathbf{V}\|_F^2 + \lambda \|\mathbf{V}\|_1 + \beta\, Tr(\mathbf{V}\mathbf{L}\mathbf{V}^T) \\ &subject\ to: \|\mathbf{b}_m\|^2 \leq 1, \end{aligned} \tag{6}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix, $\mathbf{W}$ is the similarity matrix comprised of w_{ij}, w_{ij} denotes the relationship among local features $\mathbf{x}_i$ and $\mathbf{x}_j$. $\mathbf{D}_{ii} = \sum_j w_{ij}$, $w_{ii} = 0 (i = 1, 2, ..., n)$, $\|.\|_1$ is the matrix entry-wise $l1$-norm. An approximate method was adopted to solve $\mathbf{V}$ and $\mathbf{B}$.

3 Graph-Guided Fusion Penalty Based Sparse Coding

In this section, we first give the details of our graph-guided fusion penalty based sparse coding (GFP-SC) for image classification. Then we make a discussion about some advantages of our GFP-SC method [8]. Here, we still use the uniform notation in Section 2.

3.1 Our Formulation

The sparse coding method shown in Equation (3) can be equivalently rewritten as

$$\arg\min_{\mathbf{V} \in \Re^{m \times n}} \|\mathbf{X} - \mathbf{B}\mathbf{V}\|_F^2 + \lambda \|\mathbf{V}\|_1, \tag{7}$$

where $\|.\|_F$ is the matrix Frobenius norm(entry-wise $l2$-norm), and $\|.\|_1$ is the matrix entry-wise $l1$-norm. As discussed in Section 2, the contextual structure information among neighborhood local features is not considered in traditional sparse coding. In order to make the full use of the contextual structure information among local features, we add a graph-guided fusion penalty to Equation (7). Then, we get Equation (8)

$$\arg\min_{\mathbf{V}\in\Re^{m\times n}} \|\mathbf{X}-\mathbf{B}\mathbf{V}\|_F^2 + \gamma\Omega_G(\mathbf{V}) + \lambda\|\mathbf{V}\|_1, \tag{8}$$

where γ is the regularization parameter for structured sparsity, the graph penalty is defined as

$$\Omega_G(\mathbf{V}) = \sum_{e=(i,j)\in E} w_{ij}\|\mathbf{v}_i - \mathbf{v}_j\|_1, \tag{9}$$

here, the weight w_{ij} is computed according to the visual similarity among two query local features

$$w_{ij} = \exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|_2^2}{2\delta^2}\right), \tag{10}$$

e is one edge which connected node i and j in the graph, and the edges set of the graph is defined as

$$E = \left\{(i,j) \mid \big(i<j\big) \wedge \big(\mathbf{p}_i \in N_k(\mathbf{p}_j) \vee \mathbf{p}_j \in N_k(\mathbf{p}_i)\big), i \; or \; j \in \{1,...,n\}\right\}, \tag{11}$$

where $\mathbf{p}_i$ and $\mathbf{p}_j$ denote the spatial coordinates of local features $\mathbf{x}_i$ and $\mathbf{x}_j$ in an image respectively, $N_k(\mathbf{p})$ denotes the spatial k nearest neighbors of the $\mathbf{p}$. This graph-guided fusion penalty gives a structure regularization using contextual information among the local features.

3.2 Solutions

Equation (9) can be rewritten as the matrix form

$$\Omega_G(\mathbf{V}) = \left\|\mathbf{C}\mathbf{V}^T\right\|_1, \tag{12}$$

where $\mathbf{C} \in \Re^{|E|\times n}$ is the incident matrix defined as

$$\mathbf{C}_{(i,j),k} = \begin{cases} w_{ij}, & if \;\; k=i \\ -w_{ij}, & if \;\; k=j \\ 0 & otherwise. \end{cases} \tag{13}$$

As mentioned in [4], the Equation (12) can be further rewritten as

$$\Omega_G(\mathbf{V}) = \max_{\mathbf{A}\in\mathcal{Q}} \left\langle \mathbf{C}\mathbf{V}^T, \mathbf{A}\right\rangle = \max_{\mathbf{A}\in\mathcal{Q}} Tr(\mathbf{V}\mathbf{C}^T\mathbf{A}), \tag{14}$$

where $\mathbf{A} \in \mathcal{Q} = \{\mathbf{A} \mid \|\mathbf{A}\|_\infty \le 1, \mathbf{A} \in \Re^{|E|\times m}\}$ is a matrix of auxiliary variables. The difficulty to solve the optimization problem in Equation (8) is the

non-smoothness of the graph penalty $\Omega_G(\mathbf{V})$. Here we introduce smooth approximation to $\Omega_G(\mathbf{V})$ as follows

$$f_u(\mathbf{V}) = \max_{\mathbf{A}\in\mathcal{Q}} \left\langle \mathbf{C}\mathbf{V}^T, \mathbf{A} \right\rangle - ud\left(\mathbf{A}\right), \tag{15}$$

where u is the smoothness parameter and $d(\mathbf{A})$ is defined as $\frac{1}{2}\|\mathbf{A}\|_F^2$. It is easy to prove that $f_u(\mathbf{V})$ is convex and smooth. Replace $\Omega_G(\mathbf{V})$ in Equation (8) with $f_u(\mathbf{V})$, we get the final optimization equation

$$\arg\min_{\mathbf{V}\in\Re^{m\times n}} \|\mathbf{X} - \mathbf{B}\mathbf{V}\|_F^2 + \gamma f_u(\mathbf{V}) + \lambda\|\mathbf{V}\|_1. \tag{16}$$

Then, above optimization Equation (16) can be solved by smoothing proximal gradient method.

3.3 Discussion

In this subsection, we give a detailed comparison between our coding method and the traditional sparse coding. Besides, we give the difference between our coding method and the Laplacian sparse coding which is, to our best knowledge, the most similar published method to ours.

Our GFP-SC vs. Traditional Sparse Coding: The basic idea of our method is shown in (a) of Figure 1. Naturally, the densely sampled SIFT features in an image are connected with each other by their spatial and visual information. Traditional sparse coding discards the contextual structure information while our coding method effectively retains all this kind of information. Besides, (b) of Figure 1 shows a toy example about the detailed difference between traditional sparse coding and our coding method. What is noteworthy is that local features x, y and z are three adjacent local features sampled in (a), but z is much more similar with y than x in visual space. In traditional sparse coding, coding processes for x, y and z are carried out independently. Without loss of generality, we assume that visual words $\{b_1, b_2, b_3\}$ are assigned to x, $\{b_1, b_2, b_4\}$ are assigned to y and $\{b_4, b_5, b_6\}$ are assigned to z. In this case, we can see that the visual words $\{b_4, b_5, b_6\}$ assigned to z are much more different from $\{b_1, b_2, b_4\}$ assigned to y than $\{b_1, b_2, b_3\}$ assigned to x, though z is much more similar with y than x in visual space. However, if we consider the constraint of graph-guided fusion penalty, the visual words $\{b_4, b_5, b_6\}$ assigned to local feature y are the same as visual words of z. Thus, our GFP-SC method will definitely alleviate the sensitiveness of the traditional sparse coding.

Our GFP-SC vs. Laplacian Sparse Coding: From Equation (8) and Equation (6), it is easy to find that both Laplacian sparse coding and our coding methods introduce new regularization items to take the contextual structure information into account. Here, we show the differences between our graph-guided fusion penalty based sparse coding method and the Laplacian sparse coding method. **(1)** From equations (6) and (9), we can see that the $l1$ norm is used in our method to construct the graph-guided fusion penalty while $l2$ norm is

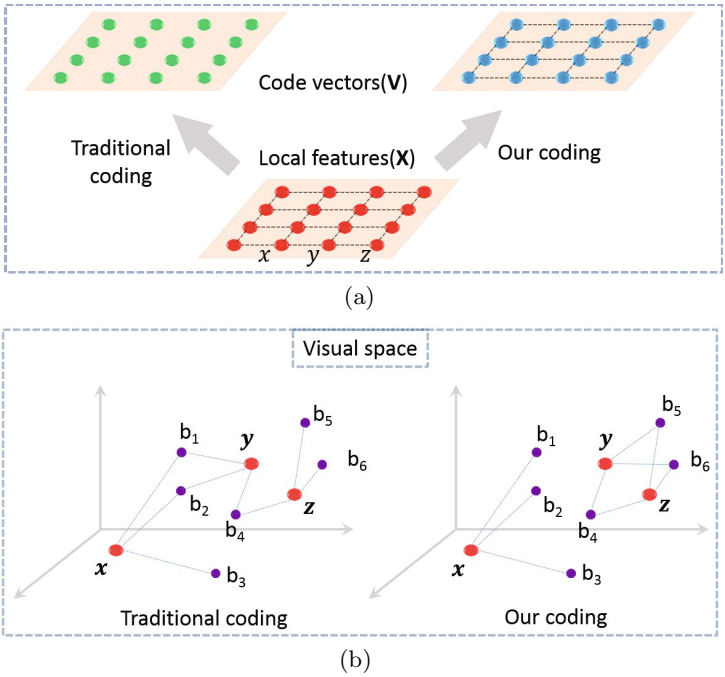

Fig. 1. Illustration of the difference between traditional sparse coding and our improved sparse coding method (GFP-SC). **(a)** Traditional sparse coding discards the contextual structure information while our coding method retains all this kind of information. **(b)** Note that x, y, z are adjacent local features sampled in (a), but z is much more similar with y than x in visual space. Our method is more likely to assign similar visual words (purple points) to similar local features in visual space.

used in the Laplacian sparse coding. In this case, code vectors computed from similar local features by our method are almost identical except several elements where related visual words are different. **(2)** We construct a sparse matrix $\mathbf{C}$ to regularize the code vectors other than the Laplacian matrix $\mathbf{L}$ in the Laplacian sparse coding method. **(3)** Only the structure information in local feature space is considered in Laplacian sparse coding method while our method combines the structure information in both feature space and spatial space as shown in Equations (9), (10) and (11) respectively. **(4)** Due to the time complexity, only an approximate solver was given in [8]. The contextual structure information is introduced from a subset of sampled local features. In contrast, our formulation can be solved using the smoothing proximal gradient method efficiently, where all the local features are encoded simultaneously. Thus the more complete contextual structure information can be retained to the code vectors after coding.

4 Experiments

To validate the effectiveness of our GFP-SC method for image classification, we compare it with 4 state-of-the-art local feature coding methods which are denoted

as: ScSPM(linear spatial pyramid matching using sparse coding)[19], LLC(locality-constrained linear coding)[18], LSC(localized soft-assignment coding)[13] and LScSPM(laplacian sparse coding)[8].

4.1 Implementation Details

Since there are always tens of thousands of local features contained in each image, it is time consuming to solve Equation (16) for all SIFT features in each image. In this paper, we adopt a simple trick to speed up our structured coding algorithm. We use SLIC (Simple Linear Iterative Clustering) algorithm [1] to segment each image into smaller regions. Then, our graph-guided fusion penalty based sparse coding (GFP-SC) algorithm is applied in each image region. We choose the parameter of the initial image region size for the SLIC method to be 60. Bigger initial region size for the SLIC method leads to higher accuracy but also more computation time [1].

To fairly compare with ScSPM, LLC and LSC, we use the publicly available source codes provided by the authors. For these three feature coding methods and our GFP-SC, we use the same experimental setup in all 5 steps (mentioned in Section 1) of the image classification except the feature coding step. We use same image size and same training/testing splits on each dataset. In feature extraction, single scale SIFT features (16×16) with step size 4 are extracted. We also use the same codebook trained by k-means clustering in advance. Max-pooling is used for pooling and a 3 levels SPM with 1×1, 2×2, 4×4 cells is used for concatenating. For the linear SVM used in ScSPM, LLC, LSC and our GFP-SC methods, we adopt the same code from Liblinear [6] instead of the different implementations provided by authors of the three baseline methods.

To compare with LScSPM, we borrow the results reported in [8], because the authors do not share their codes for comparison.

4.2 Results and Analysis

The experiments are conducted on three popularly used datasets in the literatures: Caltech101 [7], UIUC 8-Sports [12], 10 class Corel dataset [16]. On each dataset, all the results shown in Table 1 are calculated by averaging the accuracies of 10 random training/testing splits.

Following are details of the three datasets. **Caltech101** [7] contains 9144 images of objects with 101 classes. Each image is resized to be no bigger than 300 in height and width. We use 30 images for training and remaining images for testing in each split. The accuracies are shown in the first column of Table 1. **UIUC 8-Sports** dataset contains 1579 images of 8 category sport events [12]. Each image is resized to be no bigger than 400 in height and width due to the high resolution of images in this dataset. We use 70 images for training and 60 images for testing in each split. The accuracy results are shown in the second column of Table 1. **10-Corel** dataset contains 1000 images with 10 categories [16]. Each image is resized to be no bigger than 300 in height and width. We use 50 images for training and 50 images for testing in each split. The accuracy results are shown in the third column of Table 1.

Table 1. Accuracies on 3 data sets

Method	Caltech101	Sports8	Corel10
ScSPM [19]	73.0±1.1	86.0±0.8	87.8±1.0
LLC [18]	72.3±0.8	86.4±1.6	88.4±1.3
LSC [13]	73.2±1.2	86.2±1.5	88.3±0.8
GFP-SC	**74.1±1.0**	**87.9±1.4**	**89.1±1.4**
LScSPM [8]	–	85.3±0.5	88.4±0.8

From Table 1, we can see that the classification accuracies for the first three baseline coding methods are almost the same. Compared with these three baseline methods, our coding method makes 1% increases of the accuracies on all three datasets. Due to the difficulty of these datasets, a slightly even smaller than 1% improvement is acceptable as reported by recent top-level literatures in computer vision. Thus we believe our graph-guided fusion penalty based sparse coding method is effective by incorporating the structured contextual information ignored in other coding methods. Authors of [8] only give an approximate algorithm where the codebook $\mathbf{B}$ and $\mathbf{V}$ are optimized iteratively. Nevertheless, our coding method still shows competitive results on the last two datasets by only using a fixed codebook $\mathbf{B}$.

5 Conclusions

In this paper, we propose a novel graph-guided fusion penalty based sparse coding for image classification. Combined with the contextual structure information among local features, our algorithm tends to encode the similar local features into similar code vectors. Our method is implemented based on the the smoothing proximal gradient method. Experimental results on three popularly used datasets demonstrate the effectiveness of our coding algorithm compared with four state-of-the-art feature coding methods.

Acknowledgments. This work is supported by National Natural Science Foundation of China (61225009, 61003161), and Beijing Natural Science Foundation (4131004), also by the Singapore National Research Foundation under International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34(11), 2274–2282 (2012)

2. Bao, B.-K., Zhu, G., Shen, J., Yan, S.: Robust image analysis with sparse representation on quantized visual features. IEEE Transactions on Image Processing 22(3), 860–871 (2013)
3. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
4. Chen, X., Lin, Q., Kim, S., Carbonell, J.G., Xing, E.P.: Smoothing proximal gradient method for general structured sparse learning. In: UAI, pp. 105–114 (2011)
5. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22 (2004)
6. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: Liblinear: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
7. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: Workshop on Generative-Model Based Vision, CVPR (2004)
8. Gao, S., Tsang, I.W.-H., Chia, L.-T., Zhao, P.: Local features are not lonely - laplacian sparse coding for image classification. In: CVPR, pp. 3555–3561 (2010)
9. Huang, Y., Huang, K., Yu, Y., Tan, T.: Salient coding for image classification. In: CVPR, pp. 1753–1760 (2011)
10. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: ICCV, pp. 604–610 (2005)
11. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2006)
12. Li, L.-J., Li, F.-F.: What, where and who? classifying events by scene and object recognition. In: ICCV, pp. 1–8 (2007)
13. Liu, L., Wang, L., Liu, X.: In defense of soft-assignment coding. In: ICCV, pp. 2486–2493 (2011)
14. Liu, S., Feng, J., Song, Z., Zhang, T., Lu, H., Xu, C., Yan, S.: Hi, magic closet, tell me what to wear? In: ACM Multimedia (2012)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
16. Lu, Z., Ip, H.H.-S.: Image categorization with spatial mismatch kernels. In: CVPR, pp. 397–404 (2009)
17. van Gemert, J.C., Geusebroek, J.-M., Veenman, C.J., Smeulders, A.W.M.: Kernel codebooks for scene categorization. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 696–709. Springer, Heidelberg (2008)
18. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T.S., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR, pp. 3360–3367 (2010)
19. Yang, J., Yu, K., Gong, Y., Huang, T.S.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR, pp. 1794–1801 (2009)
20. Yao, B., Khosla, A., Li, F.-F.: Combining randomization and discrimination for fine-grained image categorization. In: CVPR, pp. 1577–1584 (2011)
21. Zhang, T., Liu, J., Liu, S., Ouyang, Y., Lu, H.: Boosted exemplar learning for human action recognition. In: ICCV Workshop on Video-oriented Object and Event Classification (2009)
22. Zhang, T., Liu, J., Liu, S., Xu, C., Lu, H.: Boosted exemplar learning for action recognition and annotation. IEEE Transactions on CSVT 21(7), 853–866 (2011)