SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

January 2016, Vol. 59 012105:1–012105:14
doi: 10.1007/s11432-015-5289-7

Nonnegative correlation coding for image classification

Zhen DONG, Wei LIANG*, Yuwei WU, Mingtao PEI & Yunde JIA

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081*, China*

Received December 9, 2014; accepted January 4, 2015; published online May 18, 2015

Abstract Feature coding is one of the most important procedures in the bag-of-features model for image classification. In this paper, we propose a novel feature coding method called nonnegative correlation coding. In order to obtain a discriminative image representation, our method employs two correlations: the correlation between features and visual words, and the correlation between the obtained codes. The first correlation reflects the locality of codes, i.e., the visual words close to the local feature are activated more easily than the ones distant. The second correlation characterizes the similarity of codes, and it means that similar local features are likely to have similar codes. Both correlations are modeled under the nonnegative constraint. Based on the Nesterov's gradient projection algorithm, we develop an effective numerical solver to optimize the nonnegative correlation coding problem with guaranteed quadratic convergence. Comprehensive experimental results on publicly available datasets demonstrate the effectiveness of our method.

Keywords image classification, correlation coding, nonnegativity, locality, similarity

Citation Dong Z, Liang W, Wu Y W, et al. Nonnegative correlation coding for image classification. Sci China Inf Sci, 2016, 59(1): 012105, doi: 10.1007/s11432-015-5289-7

1 Introduction

Image classification is one of the fundamental problems in computer vision and pattern recognition, and has a multitude of applications, such as image retrieval, human-computer interaction, and web content analysis. The bag-of-features model [1] has been commonly used to generate feature representations for image classification due to its invariance to scale, translation, and rotation. Figure 1 shows the general pipeline of the bag-of-features model consisting of four steps: local features extraction, visual words generation, feature coding, and feature pooling. Among these steps, feature coding is the most important one which greatly influences the image classification system in both accuracy and speed [2]. In this paper, we propose a novel feature coding method called nonnegative correlation coding for image classification. In order to enhance the discriminative power of the codes, we take full advantage of two correlations under the nonnegative constraint. One correlation reflects the locality of codes, and the other characterizes the similarity.

Numerous feature coding methods [1,3–9] have been proposed, and most of them are developed from vector quantization [1]. The vector quantization only selects the visual word closest to the feature to

* Corresponding author (email: liangwei@bit.edu.cn)

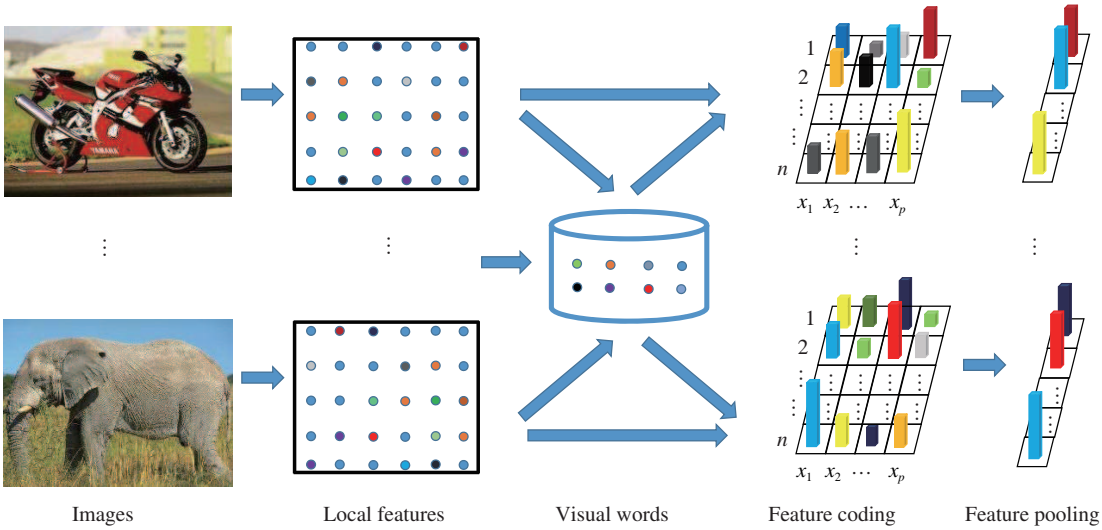

Figure 1 The general pipeline of the bag-of-features image representation framework.

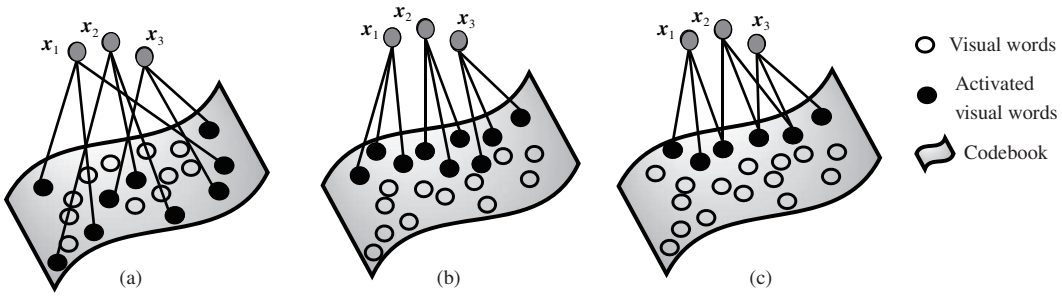

Figure 2 Schematic comparison of visual words selection methods to code local features. (a) The traditional coding scheme without any constraints; (b) only considering locality constraint; (c) our nonnegative correlation coding which considers the properties including nonnegativity, locality, and similarity of codes.

construct the hard assignment. Yang et al. [3] used sparse coding instead of the vector quantization in spatial pyramid matching [10] to obtain nonlinear codes for image classification. Zhang et al. [4] employed the nonnegative constraint to sparse coding to avoid information loss. Yu et al. [5] proposed the local coordinate coding (LCC) method which encourages a feature to be encoded by locally selected visual words, as shown in Figure 2(b). The LCC only uses the geometrical relationship between the feature and visual words, but does not take into account the geometrical relationship between features. Effective utilization of the geometrical relationship between features is beneficial for nearby local features to have similar codes. Gao et al. [6] and Zheng et al. [7] imposed a graph Laplacian constraint on the standard sparse coding to quantize local features more robustly, while both locality and nonnegativity of codes are not considered in their methods.

Our method takes into account the three properties of codes including nonnegativity, locality, and similarity to enhance the discriminative power of the final image representation. We formulate the nonnegative correlation coding as an optimization problem in which the nonnegativity is modeled as a convex constraint. The locality is obtained by minimizing the Euclidean distance between a feature and selected visual words. In order to preserve the similarity of codes, we model the geometrical relationship between features by using the k-nearest neighbor graph based graph Laplacian. We develop an effective numerical solver using the Nesterov's gradient projection algorithm [11] to solve the optimization problem with guaranteed quadratic convergence. As shown in Figure 2(c), our method makes similar features share their neighboring visual words as many as possible and thus generates more discriminative representation for image classification.

Table 1 Different constraints in different coding schemes

Coding scheme	$\phi(\mathcal{S})$	$\psi(\cdot)$
VQ [1]	–	$\|\boldsymbol{s}_i\|_0 = 1, \mathbf{1}^{\mathrm{T}}\boldsymbol{s}_i = 1$
Sparse Coding [3]	$\sum_{i=1}^{n} \|\boldsymbol{s}_i\|_1$	$\|\boldsymbol{b}_l\|_2 \leqslant 1, \mathbf{1}^{\mathrm{T}}\boldsymbol{s}_i = 1$
LCC [5]	$\sum_{i=1}^{n}\sum_{l=1}^{k} \|\boldsymbol{s}_i^l\|\|\boldsymbol{x}_i - \boldsymbol{s}_l\|_2^2$	$\mathbf{1}^{\mathrm{T}}\boldsymbol{s}_i = 1$
LLC [9]	$\sum_{i=1}^{n}\sum_{l=1}^{k} \left(\boldsymbol{s}_i^l \exp(\|\boldsymbol{x}_i - \boldsymbol{b}_l\|_2/\sigma)\right)^2$	$\mathbf{1}^{\mathrm{T}}\boldsymbol{s}_i = 1$
LScSPM [6], GraphSC [7]	$\sum_{i=1}^{n} \|\boldsymbol{s}_i\|_1 + \mathrm{tr}(\mathcal{S}\boldsymbol{L}\mathcal{S}^{\mathrm{T}})$	$\|\boldsymbol{b}_l\|_2 \leqslant 1$
Our method	$\sum_{i=1}^{n}\sum_{l=1}^{k} \Psi(\boldsymbol{x}_i, \boldsymbol{b}_l)\boldsymbol{s}_i^l + \mathrm{tr}(\mathcal{S}\boldsymbol{L}\mathcal{S}^{\mathrm{T}})$	$\boldsymbol{s}_i \succeq 0, \mathbf{1}^{\mathrm{T}}\boldsymbol{s}_i = 1$

2 Related work

Feature coding has been successfully used for image classification in the past decades. In this section, we only discuss the most relevant literature with our method. Interested readers may refer to [2] for a comprehensive review. Let $\mathcal{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ be a data matrix with n d-dimensional features extracted from an image, $\mathcal{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_k] \in \mathbb{R}^{d\times k}$ be a dictionary where each column represents a visual word, and $\mathcal{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n] \in \mathbb{R}^{k\times n}$ be the coding matrix. The goal of feature coding is to learn a representation such that each input local feature $\boldsymbol{x}_i$ can be well approximated by the dictionary $\mathcal{B}$. The general formulation of feature coding is expressed as

$$\begin{aligned} &\arg\min_{\mathcal{B},\mathcal{S}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \mathcal{B}\boldsymbol{s}_i\|_2^2 + \phi(\mathcal{S}), \\ &\text{s.t.} \quad \psi(\cdot), \end{aligned} \tag{1}$$

where $\|\boldsymbol{x}_i - \mathcal{B}\boldsymbol{s}_i\|_2^2$ measures the approximation error. The $\phi(\mathcal{S})$ and $\psi(\cdot)$ pursue discriminative descriptions, that is, similar/different local features should tend to activate similar/different visual words. The main difference among existing coding models lies in imposing different prior constraints on the generated codes $\mathcal{S}$ via specific $\phi(\mathcal{S})$ and $\psi(\cdot)$. Several constraints are listed in Table 1.

Vector quantization [1] is the simplest coding method. It only selects a single visual word closest to the local feature to construct the hard assignment. For each feature $\boldsymbol{x}_i$, there is only one nonzero coding coefficient. The vector quantization coding is thus given by

$$\begin{aligned} &\arg\min_{\mathcal{S}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \mathcal{B}\boldsymbol{s}_i\|_2^2, \\ &\text{s.t.} \quad \|\boldsymbol{s}_i\|_0 = 1, \mathbf{1}^{\mathrm{T}}\boldsymbol{s}_i = 1, \end{aligned} \tag{2}$$

where the $\|\boldsymbol{s}_i\|_0$ counts the number of nonzero entries in $\boldsymbol{s}_i$. The voting scheme in the vector quantization is highly sensitive to the reconstruction error, which results in the unrecoverable loss of discriminative information. To reduce the quantization loss, soft coding [8, 12] assigns a local feature to all the visual words according to their distances for good classification performances.

The constraint $\|\boldsymbol{s}_i\|_0 = 1$ in Eq. (2) is too restrictive to accurately reconstruct $\boldsymbol{x}_i$. To alleviate this issue, a sparsity regularization term, ℓ_1-norm, is integrated into the objective function of sparse coding:

$$\begin{aligned} &\arg\min_{\mathcal{B},\mathcal{S}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \mathcal{B}\boldsymbol{s}_i\|_2^2 + \|\boldsymbol{s}_i\|_1, \\ &\text{s.t.} \quad \|\boldsymbol{b}_l\|_2 \leqslant 1, \mathbf{1}^{\mathrm{T}}\boldsymbol{s}_i = 1, \end{aligned} \tag{3}$$

where $\|\boldsymbol{s}_i\|_1$ enforces $\boldsymbol{s}_i$ to have a small number of nonzero elements. The goal of the sparse coding is to improve the quality of a sparse representation while maximally preserving the signal fidelity. To attain this goal, many works have been proposed to modify the sparsity constraint. Liu et al. [13] imposed a nonnegative constraint to the sparse coding to represent images for classification. Yu et al. [5] proposed

the LCC method where the sparsity constraint of the sparse coding is replaced by a locality constraint. The LCC explicitly encourages features to be locally encoded. The objective function of the LCC is

$$\begin{aligned} &\arg\min_{\mathcal{B},\mathcal{S}} \sum_{i=1}^{n} \Big(\|\boldsymbol{x}_i - \mathcal{B}\boldsymbol{s}_i\|_2^2 + \sum_{j=1}^{k} |\boldsymbol{s}_i^j| \|\boldsymbol{x}_i - \boldsymbol{b}_j\|_2^2 \Big), \\ &\text{s.t.} \quad \mathbf{1}^{\mathrm{T}} \boldsymbol{s}_i = 1, \end{aligned} \tag{4}$$

where $\boldsymbol{s}_i^j$ is the jth coefficient of $\boldsymbol{s}_i$. The absolute operator of $\boldsymbol{s}_i^j$ in Eq. (4) makes the objective function not differentiable. In order to have an analytical solution, Wang et al. [9] proposed the Locality-constrained Linear Coding (LLC) method which adopts a new constraint function $\sum_{l=1}^{k} \|\boldsymbol{s}_i^l \exp(\|\boldsymbol{x}_i - \boldsymbol{b}_l\|_2/\sigma)\|_2^2$ instead of the $\sum_{l=1}^{k} |\boldsymbol{s}_i^l| \|\boldsymbol{x}_i - \boldsymbol{b}_l\|_2^2$ in Eq. (4). They also provide an approximated implementation of the LLC for fast encoding, in which each local feature is coded on locally selected visual words.

Aforementioned coding schemes are applied on local features independently. Gao et al. [6, 14] and Zheng et al. [7] imposed a graph Laplacian constraint on the sparse coding:

$$\begin{aligned} &\arg\min_{\mathcal{B},\mathcal{S}} \|\mathcal{X} - \mathcal{B}\mathcal{S}\|_2^2 + \sum_{i=1}^{n} \|\boldsymbol{s}_i\|_1 + \mathrm{tr}(\mathcal{S}\boldsymbol{L}\mathcal{S}^{\mathrm{T}}), \\ &\text{s.t.} \quad \|\boldsymbol{b}_l\|_2 = 1, \end{aligned} \tag{5}$$

where $\boldsymbol{L}$ is the Laplacian matrix [15]. In this way, local features are quantized more robustly from the viewpoint of global similarity between codes, i.e., similar features tend to have similar codes. Recently, Shabou and LeBorgne [16] presented a locality-constrained and spatially regularized coding method which preserves locality constraints both in the feature space and the spatial domain of the image. Wang et al. [17] proposed a linear distance coding method uniting the distance vector and the original feature vector to capture discriminative information for image classification. Zhang et al. [18] encoded local features by the proposed low-rank sparse learning for image classification. Zhang and Ma [19] reported a new image classification method by leveraging the low-rank matrix decomposition and Laplacian group sparse coding.

3 Nonnegative correlation coding

In order to enhance the discriminative power of the final representation, we take full advantages of three properties of codes as prior constraints including nonnegativity, locality, and similarity.

Nonnegative constraint. The nonnegative constraint is motivated by the fact that the responses of the complex cells in visual receptive fields are nonnegative values [20]. Besides, in the standard sparse coding scheme, sparse coefficients often have some negative elements. The succeeding max pooling strategy often prevents these negative elements appearing in the final representation, because most of the coefficients in sparse coding are zero. This issue means that some useful information is lost, hindering the final classification performance. Therefore, we impose the nonnegative normalization constraints $\mathbf{1}^{\mathrm{T}}\boldsymbol{s}_i = 1$ and $\boldsymbol{s}_i \succeq 0$ to maintain the unified range of values for all $\boldsymbol{s}_i$.

In the fast approximation of the LLC, Wang et al. [9] only utilizes a few visual words close to each feature to reconstruct the feature. These visual words may happen to lie at the same side of the feature, as shown in Figure 3(a). In this case, the reconstruction is impossible just using nonnegative coefficients. Therefore, they reported that the nonnegative constraint could decrease the LLC performance. Different from their work, each visual word has the probability to be activated in our method, implying that the nonnegative constraint can be available, as shown in Figure 3(b).

Locality constraint. Yu et al. [5] proved that each feature on the low-dimensional manifold can be approximated by a linear combination of its nearby visual words. Therefore, the visual words close to the feature should be activated easily to preserve locality. If a visual word $\boldsymbol{b}_j$ is much closer to $\boldsymbol{x}_i$ than other words, $\boldsymbol{s}_i^j$ (the response of $\boldsymbol{s}_i$ on the visual word $\boldsymbol{b}_j$) will be much stronger than other entries in $\boldsymbol{s}_i$.

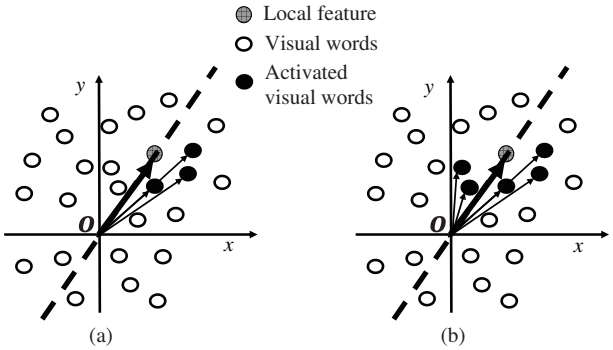

Figure 3 Comparison between the fast approximation of the LLC and the nonnegative correlation coding. (a) In the fast approximation of the LLC, only a few visual words close to the local feature are selected to reconstruct the feature, the reconstruction is impossible via nonnegative coefficients in this case; (b) the nonnegative correlation coding method encodes a local feature over the whole dictionary.

The locality is modeled by

$$\begin{aligned} &\arg\min_{\boldsymbol{s}_i} \sum_{l=1}^{k} \Psi(\boldsymbol{x}_i, \boldsymbol{b}_l) \boldsymbol{s}_i^l, \\ &\text{s.t. } \mathbf{1}^{\mathrm{T}} \boldsymbol{s}_i = 1,\ \boldsymbol{s}_i \succeq 0, \end{aligned} \tag{6}$$

where $\Psi(\boldsymbol{x}_i, \boldsymbol{b}_l) = \exp(\|\boldsymbol{x}_i - \boldsymbol{b}_l\|_2/\sigma)$ gives the distance measure between the feature and each visual word. The parameter σ determines the weight decay speed for the locality adaptor.

Similarity constraint. The manifold assumption implies that close-by features tend to have similar codes and distant ones are less likely to take similar codes. The geometrical structure of the manifold is significant for discrimination [15]. This structure can be approximated by a graph with n vertices where each vertex corresponds to a feature $\boldsymbol{x}_i$. The edge weight matrix $\boldsymbol{W}$ of the graph is defined in two ways, and the first one is by using the cosine of the angle between two features:

$$\boldsymbol{W}_{ij} = \begin{cases} \dfrac{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j}{\|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\|}, & \text{if } \boldsymbol{x}_i \in N_\varepsilon(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in N_\varepsilon(\boldsymbol{x}_i), \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The Euclidean distance between the features $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can also be used to construct the edge weight matrix, and $\boldsymbol{W}$ is thus defined as

$$\boldsymbol{W}_{ij} = \begin{cases} \dfrac{1}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2}, & \text{if } \boldsymbol{x}_i \in N_\varepsilon(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in N_\varepsilon(\boldsymbol{x}_i), \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

In both Eqs. (8) and (9), $N_\varepsilon(\boldsymbol{x}_i)$ represents the set of ε nearest neighbors of $\boldsymbol{x}_i$. The performances of the two definitions of the edge weight matrix are compared in Section 5. As defined above, if two features $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are close to each other, the weight $\boldsymbol{W}_{ij}$ will be large. The similarity constraint is implemented by

$$\begin{aligned} &\arg\min_{\mathcal{S}} \frac{1}{2} \sum_{i,j=1}^{n} \|\boldsymbol{s}_i - \boldsymbol{s}_j\|^2 \boldsymbol{W}_{ij}, \\ &\text{s.t. } \mathbf{1}^{\mathrm{T}} \boldsymbol{s}_i = 1,\ \boldsymbol{s}_i \succeq 0. \end{aligned} \tag{9}$$

By introducing the Laplacian matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, where $\boldsymbol{D}$ is a diagonal matrix whose elements are column (or row) sums of $\boldsymbol{W}$, Eq. (9) can be rewritten in a matrix form:

$$\begin{aligned} &\arg\min_{\mathcal{S}} \operatorname{tr}(\mathcal{S} \boldsymbol{L} \mathcal{S}^{\mathrm{T}}), \\ &\text{s.t. } \mathbf{1}^{\mathrm{T}} \boldsymbol{s}_i = 1,\ \boldsymbol{s}_i \succeq 0. \end{aligned} \tag{10}$$

Eq. (9) implies that if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are close (i.e., $\boldsymbol{W}_{ij}$ is large), $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are also close to each other.

Taking above constraints into account, the nonnegative correlation coding is formulated as

$$\begin{aligned}&\arg\min_{\mathcal{S}} \sum_{i=1}^{n} \Big(\|\boldsymbol{x}_i - \mathcal{B}\boldsymbol{s}_i\|_2^2 + \lambda_1 \sum_{l=1}^{k} \Psi(\boldsymbol{x}_i, \boldsymbol{b}_l)\boldsymbol{s}_i^l \Big) + \lambda_2 \mathrm{tr}(\mathcal{S}\boldsymbol{L}\mathcal{S}^{\mathrm{T}}),\\ &\text{s.t. } \mathbf{1}^{\mathrm{T}}\boldsymbol{s}_i = 1,\ \boldsymbol{s}_i \succeq 0,\end{aligned} \tag{11}$$

where λ_1 and λ_2 are positive regularization parameters to control the weights of the locality and similarity of $\mathcal{S}$, respectively. By taking full advantage of these three constraints, similar features share their neighboring visual words as many as possible. The nonnegative correlation coding is thus able to generate more discriminative representation.

4 Optimization

Given a dictionary $\mathcal{B}$, we update each vector $\boldsymbol{s}_i$ individually while holding all the other vectors $\{\boldsymbol{s}_j\}_{j\neq i}$ fixed. The model for optimizing $\boldsymbol{s}_i$ is

$$\begin{aligned}&\arg\min_{\boldsymbol{s}_i} g(\boldsymbol{s}_i) = \|\boldsymbol{x}_i - \mathcal{B}\boldsymbol{s}_i\|_2^2 + \lambda_1 \Psi(\boldsymbol{x}_i, \mathcal{B})^{\mathrm{T}}\boldsymbol{s}_i + \lambda_2 \boldsymbol{L}_{ii}\boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{s}_i + \boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{h}_i,\\ &\text{s.t.} \qquad \mathbf{1}^{\mathrm{T}}\boldsymbol{s}_i = 1,\ \boldsymbol{s}_i \succeq 0,\end{aligned} \tag{12}$$

where $\Psi(\boldsymbol{x}_i, \mathcal{B}) = [\Psi(\boldsymbol{x}_i, \boldsymbol{b}_1), \ldots, \Psi(\boldsymbol{x}_i, \boldsymbol{b}_k)]^{\mathrm{T}}$ and $\boldsymbol{h}_i = 2\lambda_2(\sum_{j\neq i} \boldsymbol{L}_{ij}\boldsymbol{s}_j)$. The convex constraint set of Eq. (12) constitutes a multinomial simplex $C = \{\boldsymbol{s}_i \in \mathbb{R}^k : \mathbf{1}^{\mathrm{T}}\boldsymbol{s}_i = 1,\ \boldsymbol{s}_i \geqslant 0\}$. We employ Nesterov's gradient projection method [11], a first-order optimization procedure, to solve Eq. (12). A key step of the Nesterov's gradient projection is to efficiently project a vector $\boldsymbol{s}_i$ onto the simplex C.

4.1 Euclidean projection onto the simplex

The Euclidean projection of a vector $\boldsymbol{v} \in \mathbb{R}^k$ onto C is given by

$$\begin{aligned}&\Pi_C(\boldsymbol{v}) = \arg\min_{\boldsymbol{v}'} \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{v}'\|_2^2,\\ &\text{s.t.} \qquad \mathbf{1}^{\mathrm{T}}\boldsymbol{v}' = 1,\ \boldsymbol{v}' \geqslant 0.\end{aligned} \tag{13}$$

The Lagrangian of the problem in Eq. (13) is

$$\mathcal{L}(\boldsymbol{v}', \mu, \boldsymbol{\omega}) = \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{v}'\|_2^2 + \mu\Big(\sum_{i=1}^{k} \boldsymbol{v}_i' - 1 \Big) - \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{v}', \tag{14}$$

where μ is a Lagrange multiplier and $\boldsymbol{\omega}$ is a vector of nonnegative Lagrange multipliers. By setting the derivative of (14) respect to $\boldsymbol{v}_i'$ to zero, we have $\partial\mathcal{L}/\partial v_i' = v_i' - v_i + \mu - \omega_i = 0$ where v_i', v_i, and ω_i are the i-th element of $\boldsymbol{v}'$, $\boldsymbol{v}$, and $\boldsymbol{\omega}$, respectively. The complementary slackness KKT condition implies that whenever $v_i' > 0$ we have $\omega_i = 0$. Thus, we get $v_i' = \max\{v_i - \mu, 0\}$ where $\mu = \big(\sum_{i=1}^{\rho} \boldsymbol{z}_i - 1\big)/\rho$ and $\rho = \max\Big\{ i \in \mathbb{N}_+ |\ \boldsymbol{z}_i - \frac{1}{i}\big(\sum_{r=1}^{i} \boldsymbol{z}_r - 1\big) > 0, i \leqslant k \Big\}$. $\boldsymbol{z}$ denotes the vector obtained by sorting $\boldsymbol{v}$ in a descending order. The Euclidean projection algorithm is summarized in Algorithm 1. The projection operator $\Pi_C(\cdot)$ is implemented efficiently in $O(k \log k)$ [21].

Algorithm 1 Euclidean projection onto the simplex

Require: A vector $\boldsymbol{v} \in \mathbb{R}^k$;
Ensure: A vector $\boldsymbol{v}' = [\boldsymbol{v}_1', \boldsymbol{v}_2', \ldots, \boldsymbol{v}_k']^{\mathrm{T}}$ such that $\boldsymbol{v}_i' = \max\{\boldsymbol{v}_i - \mu, 0\}$;
1. Sort $\boldsymbol{v}$ into $\boldsymbol{z}$ such that $\boldsymbol{z}_1 \geqslant \boldsymbol{z}_2 \geqslant \cdots \geqslant \boldsymbol{z}_k$;
2. Compute $\rho = \max\Big\{ i \in [1:k] :\ \boldsymbol{z}_i - \frac{1}{i}\big(\sum_{r=1}^{i} \boldsymbol{z}_r - 1\big) > 0 \Big\}$;
3. Compute $\mu = \frac{1}{\rho}\big(\sum_{i=1}^{\rho} \boldsymbol{z}_i - 1\big)$;

4.2 Nesterov's gradient projection

We use the Nesterov's gradient projection method to solve (12). The first-order Taylor expansion of $g(\boldsymbol{s}_i)$ at $\boldsymbol{v}$ is

$$\mathcal{Q}_{\beta,\boldsymbol{v}}(\boldsymbol{s}_i) = g(\boldsymbol{v}) + \nabla g(\boldsymbol{v})^{\mathrm{T}}(\boldsymbol{s}_i - \boldsymbol{v}) + \frac{\beta}{2}\|\boldsymbol{s}_i - \boldsymbol{v}\|_2^2, \tag{15}$$

where $\nabla g(\boldsymbol{v}) = 2\mathcal{B}^{\mathrm{T}}\mathcal{B}\boldsymbol{v} - 2\mathcal{B}^{\mathrm{T}}\boldsymbol{x}_i + \lambda_1\Psi(\boldsymbol{x}_i, \mathcal{B}) + 2\lambda_2\boldsymbol{L}_{ii}\boldsymbol{v} + \boldsymbol{h}_i$ is the gradient of $g(\boldsymbol{s}_i)$ at $\boldsymbol{v}$. By setting the deviation of (15) to 0, we obtain

$$\arg\min_{\boldsymbol{s}_i \in C} \mathcal{Q}_{\beta,\boldsymbol{v}}(\boldsymbol{s}_i) = \Pi_C\Big(\boldsymbol{v} - \frac{1}{\beta}\nabla g(\boldsymbol{v})\Big), \tag{16}$$

where $\Pi_C(\boldsymbol{v})$ is the Euclidean projection of $\boldsymbol{v}$ onto C.

To solve Eq. (12), a sequence $\{\boldsymbol{s}_i^{(t)}\}$ is generated by performing the Euclidean projection in (16): $\boldsymbol{s}_i^{(t+1)} = \Pi_C(\boldsymbol{v}^{(t)} - \nabla g(\boldsymbol{v}^{(t)})/\beta_t)$ where $\boldsymbol{v}^{(t)} = \boldsymbol{s}_i^{(t)} + \alpha_t\big(\boldsymbol{s}_i^{(t)} - \boldsymbol{s}_i^{(t-1)}\big)$. In the Nesterov's gradient projection, choosing proper parameters α_t and β_t is significant for the convergence property. Similar to [22], we set $\alpha_t = (\delta_{t-1} - 1)/\delta_t$ with $\delta_t = \big(1 + \sqrt{1 + 4\delta_{t-1}^2}\big)/2$, $\delta_0 = 0$ and $\delta_1 = 1$. β_t is selected by finding the smallest nonnegative integer j such that $g(\boldsymbol{s}_i^{(t+1)}) \leqslant \mathcal{Q}_{\beta_t,\boldsymbol{v}^{(t)}}(\boldsymbol{s}_i^{(t+1)})$ with $\beta_t = 2^j\beta_{t-1}$. The Nesterov's gradient projection algorithm for optimizing Eq. (12) is detailed in Algorithm 2.

Algorithm 2 Nesterov's gradient projection algorithm for optimizing (12)

Require: Samples set $\mathcal{X} \in \mathbb{R}^{d\times n}$, dictionary $\mathcal{B} \in \mathbb{R}^{d\times k}$, Laplacian Matrix $\mathcal{L} \in \mathbb{R}^{n\times n}$, parameters $\lambda_1, \lambda_2 \in \mathbb{R}$;
Ensure: $\{\boldsymbol{s}_i^*\}_{i=1}^n$;
  for $i = 1, 2, \ldots, n$ **do**
    initialize $\boldsymbol{s}_i^{(0)} = \boldsymbol{s}_i^{(1)} = 1/k$, $\delta_0 = 0$, $\delta_1 = 1$, $\beta_0 = 1$, $\boldsymbol{d} = \boldsymbol{h}_i = \boldsymbol{0}_{k\times 1}$;
    for $j = 1, 2, \ldots, k$ **do**
      $\boldsymbol{d}^j = \exp(\|\boldsymbol{x}_i - \boldsymbol{b}_j\|/\sigma)$;
    end for
    $\boldsymbol{d} = \boldsymbol{d}/\|\boldsymbol{d}\|_1$;
    for $j = 1, 2, \ldots, n$ **do**
      $\boldsymbol{h}_i = \boldsymbol{h}_i + \boldsymbol{L}_{ij}\boldsymbol{s}_j$;
    end for
    $\boldsymbol{h}_i = 2\lambda_2(\boldsymbol{h}_i - \boldsymbol{L}_{ii}\boldsymbol{s}_i)$;
    for $t = 1, 2, \ldots$ **do**
      $\alpha_t = (\delta_{t-1} - 1)/\delta_t$, $\boldsymbol{v}^{(t)} = \boldsymbol{s}_i^{(t)} + \alpha_t\big(\boldsymbol{s}_i^{(t)} - \boldsymbol{s}_i^{(t-1)}\big)$;
      for $j = 1, 2, \ldots$ **do**
        $\beta = 2^j\beta_{t-1}$, $\boldsymbol{v} = \boldsymbol{v}^{(t)} - \frac{1}{\beta}\Big(2\mathcal{B}^{\mathrm{T}}\mathcal{B}\boldsymbol{v}^{(t)} - 2\mathcal{B}^{\mathrm{T}}\boldsymbol{x}_i + \lambda_1\boldsymbol{d} + 2\lambda_2\boldsymbol{L}_{ii}\boldsymbol{v}^{(t)} + \boldsymbol{h}_i\Big)$;
        $\boldsymbol{s}_i^{(t)} = \Pi_C(\boldsymbol{v})$ using Algorithm 1;
        if $g(\boldsymbol{s}_i) \leqslant \mathcal{Q}_{\beta_t,\boldsymbol{v}^{(t)}}(\boldsymbol{s}_i)$ **then**
          update $\beta_t = \beta$, $\boldsymbol{s}_i^{(t+1)} = \boldsymbol{s}_i^{(t)}$;
          break;
        end if
      end for
      $\delta_{t+1} = \frac{1+\sqrt{1+4\delta_t^2}}{2}$;
    end for
    Output $\boldsymbol{s}_i^* = \boldsymbol{s}_i^{(t)}$ for each $\boldsymbol{s}_i$.
  end for

5 Experiments

We test the proposed method on four public datasets: UIUC-sport [23], Caltech-101 [24], Caltech-256 [25], and Pascal VOC 2007 [26]. The benefits of some key components of our method are also shown in this section.

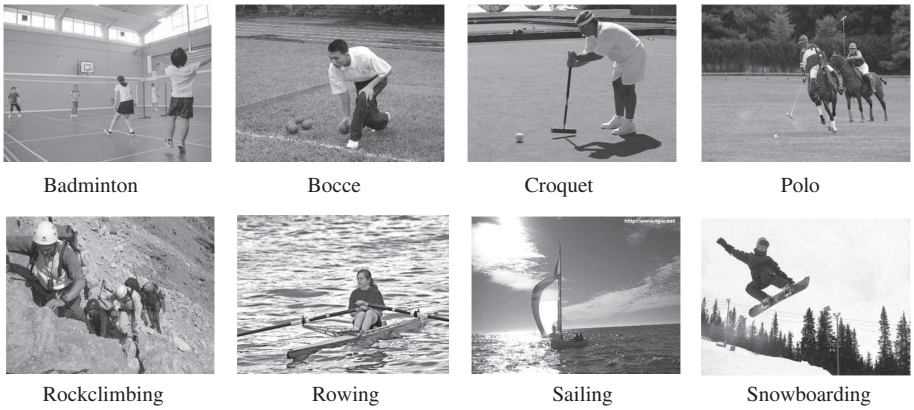

Figure 4 The example images of the UIUC-sport dataset.

5.1 Experimental set-up

We adopt the commonly used SIFT descriptor [27] as the low-level feature. To be consistent with previous work, the 128-dimensional SIFT feature is densely extracted from images on a grid with a step of 8 pixels under the scale of 16×16. We randomly select 500000 features to learn the dictionary whose size is 1024. After all features are encoded, the spatial pyramid matching [3,10] with levels of $[1 \times 1, 2 \times 2, 4 \times 4]$ is performed. In training and testing procedure, the one-vs-all linear SVM is used for its advantages in speed and excellent performance in maximum feature pooling based image classification. For each dataset, the training images are randomly selected and the results reported are the averages of 10 independent experiments. In our coding method, the most important parameters are λ_1 and λ_2. For different datasets, the value of the two parameters are different. As our observation, the proposed method achieves the best performance when λ_1 is 0.2–0.4 and λ_2 is 0.3–0.5, respectively. Specifically, for the UIUC sport and Caltech-256 datasets, we set $\lambda_1 = 0.3$ and $\lambda_2 = 0.4$, in the Caltech-101 dataset, λ_1 and λ_2 are set 0.2 and 0.5 respectively, and we set $\lambda_1 = 0.2$ and $\lambda_2 = 0.3$ in the Pascal VOC 2007 dataset.

5.2 UIUC-sport dataset

The UIUC-sport dataset [23] contains 1579 images of 8 categories including badminton, bocce, croquet, polo, rock climbing, rowing, sailing, and snow boarding. The number of images ranges from 137 to 250 per category. Figure 4 shows some example images of this dataset. For fair consideration, we randomly select 70 images from each class as training data and 60 images for test as the same with previous methods.

The comparison results are shown in Table 2, where the results of other methods are available from published papers conveniently. The results indicate that our method outperforms other coding methods. The underlying reason is that more constrains are considered in our method. It should be noted that LLC and GraphSC optimize the codebook and codes simultaneously, while our method only utilizes the dictionary generated by K-means. The confusion matrix is shown in Figure 5 where "RC", "BM", "BO", "CQ", "PO", "RO", "SA", and "SB" represent "RockClimbing", "Badminton", "Bocce", "Croquet", "Polo", "Rowing", "Sailing", and "Snowboarding", respectively. From the confusion matrix, we find that most of the misclassifications are between "bocce" and "croquet", the probable reason is that they have much visually similarity, e.g. they all have images with a man squatting and touching a ball.

We also evaluate the effects of the three constraints in our method on this dataset. First, the nonnegative constraint is removed from Eq. (11), and a simplified coding method called Correlation Coding (C-Coding) is obtained. The C-Coding has an analytical solution. The classification result of this method on UIUC-sport dataset is reported in Table 3. The effectiveness of the nonnegative constraint can be clearly seen from the performance difference between the nonnegative correlation coding (NC-coding) and the C-Coding. As analyzed above, the nonnegative constraint is quite important in avoiding information loss while encoding a local feature on the whole dictionary for image classification. The locality and similarity constraints are then removed from the NC-Coding model respectively, and two simplified cod-

Table 2 Comparison results on UIUC-sport dataset

Method	Classification accuracy (%)
Liu [8]	82.29
ScSPM [3]	82.74
HIK+one class SVM [28]	83.54
LLC [9]	83.09
LScSPM [6]	85.18
LR-Sc$^+$SPM [4]	86.69
LR-LGSC [19]	87.14
Our Method	87.36

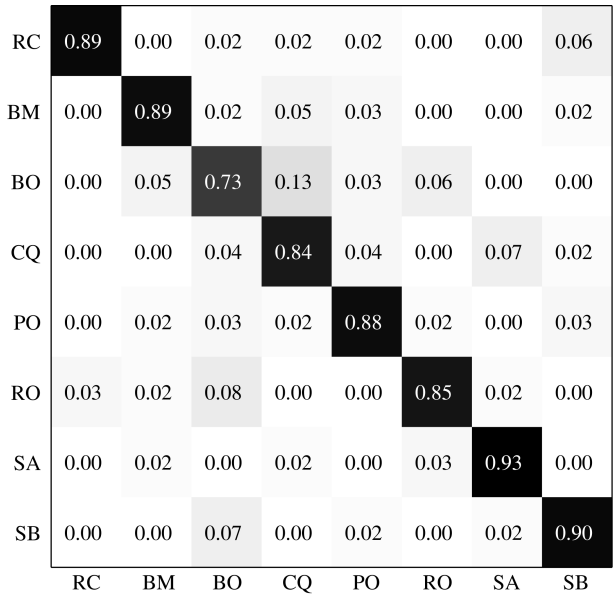

Figure 5 The confusion matrix of nonnegative correlation coding on UIUC-sport dataset.

Table 3 Classification accuracies of NC-Coding and three simplified methods on UIUC-sport dataset

Method	Classification accuracy (%)
C-Coding	81.38 ± 0.46
NSC-Coding	83.22 ± 0.42
NLC-Coding	85.89 ± 0.67
NC-Coding	87.36 ± 0.39

ing methods are obtained: Nonnegative Similarity Correlation Coding (NSC-Coding) and Nonnegative Locality Correlation Coding (NLC-Coding). Here, we just need to set λ_1 and λ_2 to 0 in Eq. (11), respectively. The performances of these two methods are also shown in Table 3. As expected, the NC-Coding method outperforms the NSC-Coding and NLC-Coding methods, which demonstrates the effectiveness of locality and similarity constraints in obtaining discriminative representation for image classification.

The computational cost of the proposed method is also evaluated in this part. We compare the accelerated proximal gradient (APG) approach [29] and the Nesterov's gradient projection (NPG) method for solving (12) to convergence. 50 images are randomly selected and the average coding times are reported in Table 4. Both the experiments are conducted on the computer with a 3.40 GHz CPU and 16.0 GB memory. From Table 4, we can clearly observe that NPG is quite fast than APG.

Table 4 Running time comparison of different optimization methods

Optimization method	Average time (s)
APG	718.02
NPG	477.71

Table 5 Comparison results on Caltech-101 dataset

Method	Classification accuracy by using different size of training sample (%)					
	5	10	15	20	25	30
SRC [30]	48.80	60.10	64.90	67.70	69.20	70.70
D-KSVD [31]	49.60	59.50	65.10	68.60	71.10	73.00
LR-Sc$^+$SPM [4]	–	–	69.58	–	–	75.68
CCLR-Sc$^+$SPM [19]	–	–	70.86	–	–	76.62
ScSPM [3]	–	–	67.00	–	–	73.20
LCSRC [16]	–	–	–	–	–	73.23
LLC [9]	51.15	59.77	65.43	67.74	70.16	73.44
Liu [8]	–	–	–	–	–	74.21
LR-LGSC [32]	55.73	64.09	68.15	72.11	74.2	76.52
Our Method	52.47	63.33	67.91	72.32	74.12	76.61

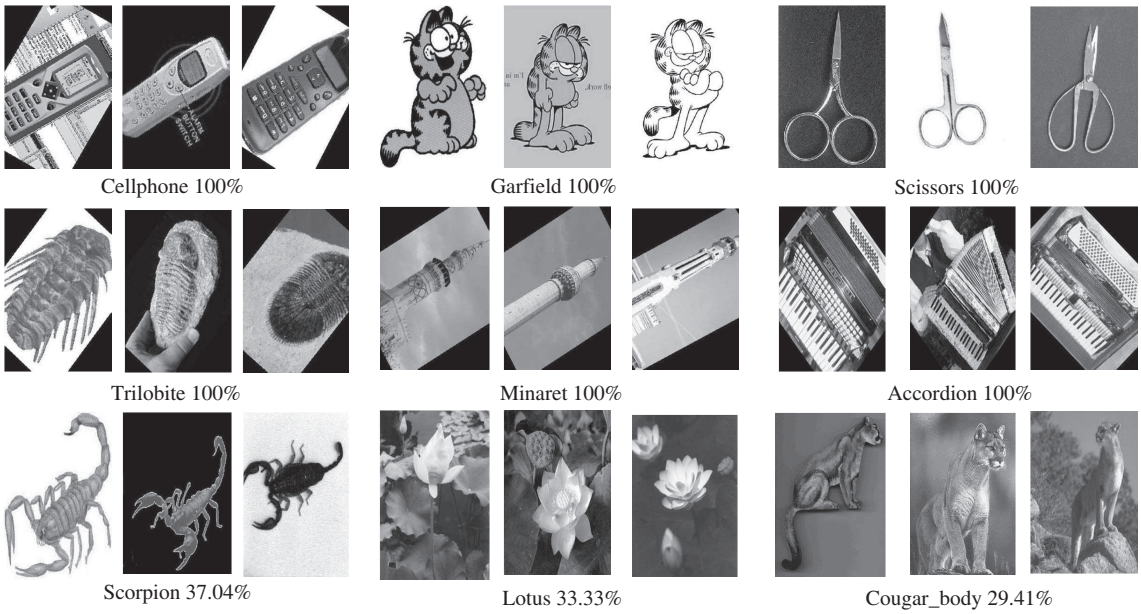

Figure 6 Example images from classes with highest and lowest classification accuracies from the Caltech-101 dataset.

5.3 Caltech-101 dataset

The Caltech-101 dataset [24] contains 101 object categories, such as cameras, chairs, flowers, vehicles, etc. All the categories are with significant variances in shape and cluttered backgrounds. This dataset has 9144 images in all, and the image number varies from 31 to 800 per category. Moreover, it is individually added to an extra "background" category, i.e., BACKGROUND_Google.

For each category, we randomly select $5, 10, 15, ..., 30$ images for training and test on the remaining as the same with previous work. The results compared with previous methods are listed in Table 5, which indicates that our method outperforms all the other methods. Figure 6 shows some example images from classes with the highest and lowest classification accuracy of our method on the Caltech101 dataset. The figure indicates that our method performs less successful on classes with large intra-class variations.

Table 6 Comparison results on Caltech-256 dataset

Method	Classification accuracy by using different size of training sample (%)			
	15	30	45	60
SPM [10, 25]	–	34.10	–	–
LR-Sc$^+$SPM [4]	35.31	–	–	–
ScSPM [3]	27.73	34.02	37.46	40.14
LLC [9]	34.36	41.19	45.31	47.68
LScSPM [6, 14]	29.99	35.74	38.47	40.32
Our Method	36.84	44.17	48.92	50.28

Table 7 Comparison results of two edge weight matrixes

$\boldsymbol{W}$ generation method	Classification accuracy by using different size of training sample (%)			
	15	30	45	60
Using Euclidean distance	33.31	42.76	47.35	48.47
Using cosine	36.84	44.17	48.92	50.28

5.4 Caltech-256 dataset

We also test our method on the Caltech-256 dataset [25] which is a challenging dataset for object recognition. Different from the Caltech-101 dataset where the objects are often in the center of the image, the Caltech-256 dataset presents much higher intra-class variability including the variability in object size, pose, and location. The Caltech-256 dataset consists of 256 categories and a background class in which none of the image belongs to the 256 categories. Each class contains at least 80 images, to provide a total of 29780 images.

Following the common experimental setting, we tested our coding method on 15, 30, 45, and 60 training images per class respectively. The comparison results are displayed in Table 6. As can be seen from the table, the classification accuracies of our method are better than the LLC and the LScSPM under all cases. It demonstrates that the proposed coding method is able to preserve the locality and similarity of codes simultaneously and is more effective for image classification.

We also evaluate the performance difference between two ways of constructing edge weight matrix described in Section 3 on the Caltech-256 dataset. The edge weight matrix $\boldsymbol{W}$ generated by using the cosine and the Euclidean distance between two local features are used in the objective function respectively, and the classification accuracies under different training samples are listed in Table 7. As shown in the table, the cosine outperforms the Euclidean distance, which is probably due to the sensitiveness of the Euclidean distance to the values of the local features. Therefore, the cosine is more appropriate for measuring similarity.

5.5 Pascal VOC 2007 dataset

The Pascal VOC 2007 dataset [26] is a typical dataset for image classification with 9963 images from 20 categories of objects (i.e., aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, people, potted plant, sheep, soft, train, and tv/monitor). Figure 7 shows some example images of this dataset. As shown in the figure, the images of the dataset which are all daily photos obtained from Flickr vary significantly in size, viewpoint, illumination, scale, pose, and deformation. Therefore, the Pascal VOC 2007 is a challenging dataset for image classification. The training and testing samples have been well divided for convenient evaluation and fair comparison.

Classification performance is measured by the mean average precision (mAP) which is the standard metric adopted by the Pascal challenge. The proposed coding method is compared with LLC [9], the best result of VOC2007 competition (Best'07) [26], the re-implementation of Fisher coding (FC) [33] and Super vector (SV) [34] by [35], and the CCLR-Sc$^+$SPM [32]. The comparison results are shown in Table 8. The proposed method outperforms LLC, which demonstrates the effectiveness of our coding method. Our method is also comparable to the Fisher coding and CCLR-Sc$^+$SPM. The Fisher coding

Figure 7 Example images from the Pascal VOC 2007 dataset.

Table 8 Comparison results on Pascal VOC 2007 dataset

Object class	Classification accuracy (%)					
	LLC [9]	Best'07 [26]	FK [33, 35]	SV [34, 35]	CCLR-Sc$^+$SPM [32]	Our method
Aeroplane	74.8	77.5	80.0	74.3	**80.2**	78.7
Bicycle	65.2	63.6	67.4	63.8	67.1	**67.6**
Bird	50.7	**56.1**	51.9	47.0	52.7	54.3
Boat	70.9	**71.9**	70.9	69.4	71.3	70.5
Bottle	28.7	**33.1**	30.8	29.1	31.5	31.0
Bus	68.8	60.6	72.2	66.5	71.9	**72.4**
Car	78.5	78.0	79.9	77.3	**80.4**	80.1
Cat	61.7	58.8	61.4	60.2	61.8	**62.1**
Chair	54.3	53.5	**56.0**	50.2	55.7	54.8
Cow	48.6	42.6	**49.6**	46.5	49.6	47.9
Dining table	51.8	54.9	**58.4**	51.9	56.2	57.4
Dog	44.1	45.8	44.8	44.1	44.7	**45.9**
Horse	76.6	77.5	78.8	77.9	79.1	**79.4**
Motorbike	66.9	64.0	**70.8**	67.1	69.3	70.1
People	83.5	85.9	85.0	83.1	84.8	**86.2**
Potted plant	30.8	**36.3**	31.7	27.6	31.9	31.2
Sheep	44.6	44.7	51.0	48.5	**48.6**	44.4
Soft	53.4	50.9	56.4	51.1	**56.6**	56.5
Train	78.2	79.2	**80.2**	75.5	79.9	78.6
Tv/monitor	53.5	53.2	**57.5**	52.3	56.3	55.3
mAP	59.3	59.4	**61.7**	58.2	61.5	61.2

inspired by Fisher kernel integrates the power of both generative and discriminative models, and it can thus preserve more information than reconstruction-based coding methods. The CCLR-Sc$^+$SPM is a new image classification framework involving dictionary learning and data de-noising by correlation constrained low-rank and sparse matrix decomposition, while our method only focuses on the coding procedure. The comparable results show the effectiveness of our coding method for image classification.

Besides, our coding method can be combined with CCLR-Sc$^+$SPM by replacing the nonnegative sparse coding and LLC with our coding method.

6 Conclusion

In this paper, the nonnegative correlation coding method has been proposed to transform low-level features into high-level representations for image classification. The nonnegative correlation coding employs the nonnegativity, locality, and similarity of codes as constraints to reduce information loss in coding. By imposing these constrains, our method makes similar local features share their neighboring visual words as many as possible and thus enhances the discriminative power of the final representation. The Nesterov's gradient projection algorithm with guaranteed quadratic convergence is quite fast for solving the nonnegative correlation coding. Experimental results show that the proposed method is effective in image classification.

Acknowledgements This work was supported in part by National Basic Research Program of China (973) (Grant No. 2012CB720000), National Natural Science Foundation of China (NSFC) (Grant No. 61203291), and Specialized Research Fund for the Doctoral Program of Chinese Higher Education (Grant No. 20121101110035). The authors are grateful to Min YANG and Yang HE for helpful discussions.

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Sivic J, Zisserman A. Video google: A text retrieval approach to object matching in videos. In: Proceedings of the 9th Computer Vision Conference. Nice: IEEE, 2003. 1470–1477
2 Huang Y Z, Wu Z F, Wang L, et al. Feature coding in image classification: a comprehensive study. IEEE Trans Pattern Anal Mach Intel, 2013, 36: 493–506
3 Yang J, Yu K, Gong Y, et al. Linear spatial pyramid matching using sparse coding for image classification. In: Proceedings of the 22th Computer Vision and Pattern Recognition Conference. Miami: IEEE, 2009. 1794–1801
4 Zhang C, Liu J, Tian Q, et al. Image classification by non-negative sparse coding, low-rank and sparse decomposition. In: Proceedings of the 24th Computer Vision and Pattern Recognition Conference. Colorado Springs: IEEE, 2011. 1673–1680
5 Yu K, Zhang T, Gong Y, et al. Nonlinear learning using local coordinate coding. In: Proceedings of the 24th Advances in Neural Information Processing Systems. Vancouver: NIPS, 2009. 2223–2231
6 Gao S, Tsang I, Chia L, et al. Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. IEEE Trans Pattern Anal Mach Intel, 2013, 35: 92–104
7 Zheng M, Bu J, Chen C, et al. Graph regularized sparse coding for image representation. IEEE Trans Image Process, 2011, 20: 1327–1336
8 Liu L Q, Wang L, Liu X W, et al. In defense of soft-assignment coding. In: Proceedings of the 13th Computer Vision Conference. Barcelona: IEEE, 2011. 2486–2493
9 Wang J, Yang J, Yu K, et al. Locality-constrained linear coding for image classification. In: Proceedings of the 23th Computer Vision and Pattern Recognition Conference. San Francisco: IEEE, 2010. 3360–3367
10 Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 19th Computer Vision and Pattern Recognition Conference. San Francisco: IEEE, 2006. 2169–2178
11 Nesterov Y, Nesterov I U E. Introductory Lectures on Convex Optimization: A Basic Course. Berlin: Springer, 2004
12 van Gemert J C, Veenman C J, Smeulders A W, et al. Visual word ambiguity. IEEE Trans Pattern Anal Mach Intel, 2010, 32: 1271–1283
13 Liu Y, Wu F, Zhang Z, et al. Sparse representation using nonnegative curds and whey. In: Proceedings of the 23th Computer Vision and Pattern Recognition Conference. San Francisco: IEEE, 2010. 3578–3585
14 Gao S, Tsang I W, Chia L- T, et al. Local features are not lonely–Laplacian sparse coding for image classification. In: Proceedings of the 23th Computer Vision and Pattern Recognition Conference. San Francisco: IEEE, 2010. 3555–3561
15 Chung F R. Spectral Graph Theory. Washington DC: American Mathematical Society, 1997
16 Shabou A, LeBorgne H. Locality-constrained and spatially regularized coding for scene categorization. In: Proceedings of the 25th Computer Vision and Pattern Recognition Conference. Providence: IEEE, 2012. 3618–3625
17 Wang Z L, Feng J S, Yan S C, et al. Linear distance coding for image classification. IEEE Trans Image Process, 2013, 22: 537–548
18 Zhang T, Ghanem B, Liu S, et al. Low-rank sparse coding for image classification. In: Proceedings of the 14th Computer Vision Conference. Sydney: IEEE, 2013. 281–288

19 Zhang L, Ma C. Low-rank decomposition and Laplacian group sparse coding for image classification. Neurocomputing, 2014, 135: 339–347
20 Hoyer P O. Modeling receptive fields with non-negative sparse coding. Neurocomputing, 2003, 52: 547–552
21 Duchi J, Shalev-Shwartz S, Singer Y, et al. Efficient projections onto the l1-ball for learning in high dimensions. In: Proceedings of the 25th Machine Learning Conference. Helsinki: ACM, 2008. 272–279
22 Nesterov Y. A method of solving a convex programming problem with convergence rate O(1/k2). Soviet Math Dok, 1983, 27: 372–376
23 Li L J, Li F F. What, where and who? Classifying events by scene and object recognition. In: Proceedings of the 11th Computer Vision Conference. Rio de Janeiro: IEEE, 2007. 1–8
24 Li F F, Fergus R, Perona P, et al. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. Comput Vis Image Underst, 2007, 106: 59–70
25 Griffin G, Holub A, Perona P, et al. Caltech-256 Object Category Dataset. Technical Report. Pasadena: California Institute of Technology, 2007
26 Everingham M, van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge. Int J Comput Vis, 2010, 88: 303–338
27 Lowe D G. Distinctive image features from scale-invariant keypoints. Int J Comput Vis, 2004, 60: 91–110
28 Wu J, Rehg J M. Beyond the euclidean distance: creating effective visual codebooks using the histogram intersection kernel. In: Proceedings of the 12th Computer Vision Conference. Kyoto: IEEE, 2009. 630–637
29 Bao C, Wu Y, Ling H, et al. Real time robust l1 tracker using accelerated proximal gradient approach. In: Proceedings of the 25th Computer Vision and Pattern Recognition Conference. Providence: IEEE, 2012. 1830–1837
30 Wright J, Yang A Y, Ganesh A, et al. Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intel, 2009, 31: 210–227
31 Zhang Q, Li B. Discriminative k-svd for dictionary learning in face recognition. In: Proceedings of the 23th Computer Vision and Pattern Recognition Conference. San Francisco: IEEE, 2010. 2691–2698
32 Zhang C J, Liu J, Liang C, et al. Image classification by non-negative sparse coding, correlation constrained low-rank and sparse decomposition. Comput Vis Image Underst, 2014, 123: 14–22
33 Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization. In: Proceedings of the 20th Computer Vision and Pattern Recognition Conference. Minneapolis: IEEE, 2007. 1–8
34 Zhou X, Yu K, Zhang T, et al. Image classification using super-vector coding of local image descriptors. In: Proceedings of the 11th European Conference on Computer Vision. Berlin: Springer , 2010. 6315: 141–154
35 Chatfield K, Lempitsky V, Vedaldi A, et al. The devil is in the details: an evaluation of recent feature encoding methods. In: Proceedings of the 22nd British Machine Vision Conference. Dundee: BMVA Press, 2011. 1–12