Extended Laplacian Sparse Coding for Image Categorization

Mouna Dammak, Mahmoud Mejdoub, and Chokri Ben Amar

REGIM: REsearch Groups on Intelligent Machines
University of Sfax, National School of Engineers (ENIS)
Department of Electrical Engineering
Sfax, 3038, Tunisia
{mouna.damak,chokri.benamar}@ieee.org,
mah.mejdoub@gmail.com

Abstract. In image classification task, several recent works show that sparse representation plays a basic role in dictionary learning. However, this approach neglects the spatial relationships in the image space during dictionary learning. However, this approach neglects the neighboring relationship in dictionary learning. To alleviate the impact of this problem, we propose a novel dictionary learning based on Laplacian sparse coding method that profits from the neighboring relationship among the local features. For that purpose, we incorporate the matching between local regions in the Laplacian sparse coding formula. Moreover, we integrate statistical analysis of the distribution of the responses of each local feature to the dictionary basis in the final image representation. Our experimental results prove that our method performs existing background results based on sparse representation.

Keywords: Bag of visual words, Sparse coding, Image categorization, Image spatial information.

1 Introduction

Image classification framework consists in attributing one or more category labels to a given image. It is one of the most fundamental problems in computer vision and pattern recognition. Besides, it has a wide range of applications, such as image and video retrieval, video surveillance, biometrics, etc. In the recent literature, the Bag of Visual Words (BoW)[7] is the most popular approach in image classification task[9,4,3,2]. It has achieved the state-of-the-art performance in several databases. The original BoW [7] is based on K-means method to form the dictionary by quantifying the space of local features into a set of dictionary basis vectors. After that, each local feature is assigned to a single basis vector. We can note that the hard quantization is very strict and leads to error quantization especially if the features are located on the boundary proximity of divers basis vectors.

Sparse coding [15] aims to learn a dictionary and simultaneously find a sparse linear combination of basis vectors from this dictionary to represent the image

C.K. Loo et al. (Eds.): ICONIP 2014, Part III, LNCS 8836, pp. 292–299, 2014.

features. It has consistently enhanced the results on image classification problem by resolving efficiently the problem of hard quantization. Yang et al. [15] proposed Sparse coding SPM (as referred ScSPM). They train the dictionary and compute the sparse codes in the encoding step. In the pooling step, the max pooled responses across different sub-regions are computed.

Sparse coding [15] treats local features independently, ensuing that the sparse codes can vary greatly even for close features. To overcome this drawback, different extensions of sparse coding method [14,6,12] have been suggested recently by adding some regularization or constraints in the sparse coding objective function. The Locality-constrained Linear Coding (LLC) [14] technique considers the locality information in the feature coding process. Contrary to the sparse coding, LLC enforces locality instead of sparsity. It uses the k nearest neighbors of features as the local basis vectors. This leads to smaller coefficient for the basis vectors far away from each local feature. Laplacian Sparse Coding (LSC) [6] learns an unsupervised dictionary, as well as the sparse representation that preserves the conformity of close local descriptors in the data space. This method has used histogram intersection similarity based on k-Nearest Neighbors (KNN) method to construct a Laplacian matrix that tries to preserve the local consistence in the feature space. Only the K-nearest local features are selected to active the Laplacian matrix. This method obtains background results on several object recognition.

After the encoding phase, the pooling step is applied in order to aggregate the encoded features. Two major strategies are used: average pooling and max pooling. The first strategy consists to take the average of the responses over the region in a given visual word. It is applied generally after the BoW encoding step. The second strategy considers the largest responses instead of its average and it is suitable to sparse encoded histograms. These two approaches have two major drawbacks. Firstly, they ignore the spatial information when gathering the local features. As a solution, spatial pyramid representation is used in [7,15,14,6] in order to incorporate the global spatial information into the pooling step. Explicitly, each image is split progressively into finer cells. For every cell, a histogram of basis vector is determined. These histograms are then mixed up using a weighting scheme depending on the level of the spatial pyramid. Secondly, they consider a scalar result for each dictionary basis vector discarding the analysis of the distribution around each visual word. Avila et al. [1] enhance these strategies by proposing Bag of Statistical Sampling Analysis (BoSSA) pooling. It is applied to discretize the distance between K-means clusters and the local features yielding a histogram of distances rather than a scalar. Each bin of this histogram measures the average number of features assigned to a given visual word, which discretized distance falls into this bin.

In this paper, the contributions can be summarized as follows:

1. In the encoding step, we propose a novel sparse coding method in order to enrich the image spatial information during the encoding phase. Compared to LSC that exploits the dependencies between local features only in the

feature space, we propose to exploit the dependencies among them in both feature and image spaces.

2. In the pooling step, inspired by the BoSSA [1] method that applies a statistical analysis on the distances between the local features and the k-means clusters, we develop a novel pooling method based on performing statistical analysis for the sparse codes.

2 Laplacian Sparse Coding Formula

Sparse coding method aims to reduce the problem of hard quantization. It finds a sparse linear combination of basis vectors for each image feature. Given the local feature space $X = [x_1, \ldots, x_N]$, $x_i \in \Re^{D\times 1}$, K basis vectors $U = [u_1, \ldots, u_K] \in \Re^{D\times K}$ generate the dictionary and the matrix of the sparse codes $V = [v_1, v_2, \ldots, v_N]$ where $v_i \in \Re^{\mathrm{k}\times 1}$ and v_{ik} is the weight of the vector x_i in the basis vector u_k, the optimization problem of sparse coding can be rewritten as follows:

$$\min_{U,V} \|X - UV\|_F^2 + \lambda \sum_i \|v_i\|_1 \qquad (1)$$
$$\mathit{subject\ to}\ \|u_j\| \le 1; \forall j = 1, \ldots, K$$

The first term in Eq.1 is the reconstruction error, and the second term is used to control the sparsity of the sparse codes v_i. λ is the tradeoff parameter used to balance the sparsity and the reconstruction error. Sparse coding has proved its efficiency in feature quantization process. Yet, the major drawbacks of this coding method is that it neglects the consistency of the sparse codes for the close local descriptors.

Gao et al [6] proposed Laplacian sparse coding to incorporate the similarity among the local features in the feature space. They added a regularization term in the objective function of sparse coding. Given two local features x_i and x_j as well as their sparse codes v_i and v_j respectively, $W_{i,j}$ measures the similarity between these features, the function objective of LSC is described as follows:

$$\min_{U,V} \|X - UV\|_F^2 + \lambda \sum_i \|v_i\|_1 + \frac{\beta}{2} \sum_{i,j} \|v_i - v_j\|^2 W_{i,j} \qquad (2)$$

Then, the formula 2 can be reformulated as:

$$\min_{U,V} \|X - UV\|_F^2 + \lambda \sum_i \|v_i\|_1 + \beta tr\left(VLV^{\mathrm{T}}\right) \qquad (3)$$
$$\mathit{subject\ to:}\ \|u_m\|^2 = 1$$

where β is the weight on the closeness restriction and L defines the Laplacian matrix.

3 Proposed Approach

In this section, we describe the details of the extended approach based on Laplacian sparse coding. First, the local features are extracted using dense SIFT [8] features. Then, the local regions are built around local features in order to incorporate the local spatial information during the sparse coding process. The local features are encoded, via our proposed approach, to sparse codes taking into account the consistency between the sparse codes and the local regions centred around their corresponding local features. Furthermore, we apply our proposed Sparse BoSSA Pooling (SBP) to give the final image representation. Finally, a multi-class non-linear SVM classifiers is trained for image category prediction. These steps are detailed in the next sections.

3.1 Feature Extraction

Several works [11] prove that sampling density is better than interest points. SIFT descriptor demonstrates its excellent results in image classification [13,5,6,15,14]. For that, we implement in our experiment dense SIFT features. Given a local region, SIFT descriptor is computed as 16 histograms of 8 gradient orientations. It gives a 128-dimensional vectors.

3.2 Proposed Extension of Laplacian Sparse Coding

Given the local feature space $X = [x_1, \ldots, x_N]$, $x_i \in \Re^{D\times 1}$ extracted as described in section 3.1. In order to take into account the local spatial information during the encoding phase, we form the local regions $R(x_i)$ centred around each local feature x_i. We consider the eight spatial neighbours $E(x_i)$ for each local feature x_i to form the local region $R(x_i) = \{x_i, E(x_i)\}$ as showed in Figure 1.

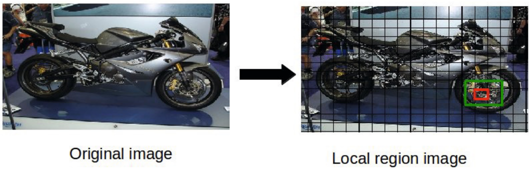

Fig. 1. Illustration of the local spatial information extraction process

After that, we aim to learn the unsupervised dictionary and to compute the sparse code for each feature. In the classical Laplacian sparse coding, $W_{i,j}$ computes the similarity between local features x_i and x_j in order to realize the consistency between local features and sparse codes. In this paper, we propose to compute the similarity between x_i and x_j taking into account the similarity between their spatial neighborhood in the image. Explicitly, we fix $W_{i,j} = 1$

if the local region $R(x_i)$ is among the k-nearest neighbour of the local region $R(x_j)$, otherwise, we fix $W_{i,j} = 0$. To compute the similarity between $R(x_i)$ and $R(x_j)$, we define the similarity measure $S(R(x_i), R(x_j))$ as the summation of (1) the histogram intersection similarity between x_i and x_j and (2) the mean pairwise similarities between the matched local features in $E(x_i)$ and $E(x_j)$. For each local feature in $E(x_i)$, the matching is carried out by finding the closet local feature in $E(x_j)$ (in the sense of the histogram intersection similarity).

3.3 Proposed Sparse BoSSA Pooling Method

In the previous section, we have trained the unsupervised dictionary and we have coded each local feature by a sparse code. In this section, we will represent the final vector of a given image $I = \{x_i\}_1^M$ via these sparse codes. To measure the distribution of the responses of each local descriptor to the dictionary's vector basis, we adapt BoSSA [1] pooling strategy to our new sparse encoding scheme as referred sparse BoSSA pooling. For that purpose, we built a histogram $h_{k,b}$ of size B for each k^{th} basis vector. Each bin of this histogram represents the occurrences of the absolute value of the sparse code weights that fall into this bin. The formula describes the computation of a given histogram h_k for an image I.

$$h_{k,b} = card\left(v_i \ \mid \ x_i \in I \ and \ v_k^{min} + s \times b \leq |v_{ik}| \leq v_k^{max} + s \times (b+1)\right)$$

where

$$s = \frac{v_k^{max} - v_k^{min}}{B} \ and \ b \in [0, \ldots, B-1]$$

B denotes the number of bins, v_k^{min} and $v_k^{\max}$ limit the range of activated sparse code weights $|v_{i,k}|$ over all descriptors x_i extracted from the images of the learning set and the step s corresponds to the length of the bin.

4 Experiments

4.1 Experimental Protocol

In our experiments, we extract densely SIFT features from 8×8 patches using a spatial stride equal to 4. After that, we form a local region for each local feature. Then, we learn the dictionary and we compute the sparse code for each local feature implementing our encoding method. Furthermore, we apply SPR in order to preserve the global spatial information and we apply SBP in each subregion. For fair comparison to [15,6], the splits of the SPR is $[(1 \times 1), (2 \times 2), (4 \times 4)]$. Also, the number of basis vectors is fixed to 1024 and the number of bins is fixed to $B = 3$. Two settings are included in our objective function λ: the sparsity of the sparse codes and β: the weight on the closeness restriction. λ and β are fixed by cross validation: we fix $\beta = 0.1$, $\lambda = 0.3$ for UIUC Sport and Caltech-256, and we fix $\beta = 0.2$, $\lambda = 0.4$ in Corel dataset. In the classification step, we train the histograms with the $chi - square$ non-linear SVM.

4.2 Datasets

We evaluate our approach for four datasets: UIUC-Sport, Corel-10 Dataset and caltech-256. For fair comparison, we keep the identical experimental properties as [15,6]. Table 1 summarizes the characteristics for all the datasets: the number of classes, the number of the images in the dataset, the number of training images per class and the number of test images.

Table 1. The general description of the datasets

	UIUC-sport	Corel	Caltech-256
# of classes	8	10	257
# of images	1792	1000	30607
# of training	70	50	15/30/45/60
# of test	remainder	50	remainder

4.3 Results

Impact of the Spatial Context (SC). Table 2 depicts the influence of the spatial context added in the regularization term of the objective function. We observe that the integration of the dependencies between local features both in feature space and image space is more important than the integration of only the dependencies between local features in the feature space.

Table 2. Impact of the spatial context on classification accuracy

Methods	UIUC-Sport	Corel	caltech-256
Without SC	$85.18 \pm .46$	$88.76 \pm .94$	$35.74 \pm .1$
With SC	$86.6 \pm .42$	$90.15 \pm .76$	$38.35 \pm .46$

Impact of SBP Pooling on Our New Encoding Method. In this experiment, we study the impact of our sparse BoSSA pooling method on classification accuracy. Table 3 shows that the proposed pooling method enhances the classification accuracy in all datasets. These results confirm the advantages introduced by SBP representation.

Table 3. Impact of sparse BoSSA pooling on image classification accuracy

Methods	UIUC-Sport	Corel	Caltech-256
Without SBP	86.6 ± .42	90.15 ± .76	38.35 ± .46
With SBP	87.85 ± .46	91.33 ± .94	39.64 ± .53

Table 4. Performance Comparison on Caltech-256 Dataset

Number of training images	15	30	45	60
Method	Average Classification rate(%)			
BoW [10]	23.5 ± 0.42	29.1 ± 0.38	32.17 ± 0.53	34.21 ± 0.24
ScSPM[15]	27.73 ± 0.51	34.02 ± 0.35	37.46 ± 0.55	40.14 ± 0.91
LLC [5]	27.74 ± 0.32	32.07 ± 0.24	35.09 ± 0.44	37.79 ± 0.42
LSC[6]	29.99 ± 0.15	35.74 ± 0.1	38.47 ± 0.51	40.32 ± 0.32
Our	33.72 ± 0.7	39.64 ± 0.53	42.16 ± 0.51	44.03 ± 0.63

Comparison with State-of-the-Art. We compare our approach to different image classification methods in the literature. The SPM baseline method and baseline methods based on sparse coding: ScSPM, LLC, LSC. Table 5 and 4 show that our method exceeds background performance on divers datasets. This demonstrates that the proposed method can improve the classical Laplacian sparse coding by taking into account the locality constraint among the local features in the encoding phase and the statistical distribution of the sparse code weights in the pooling step.

Table 5. Performance Comparison on UIUC-Sport and Corel datasets

Methods	UIUC-Sport	Corel
SPM [7]	79.98 ± 1.67	-
ScSPM [15]	82.74 ± 1.46	86.6 ± 1.01
LLC [14]	83.09 ± 1.3	87.93 ± 1.04
LSC [6]	85.18 ± 0.46	88.76 ± 1.04
Our	87.85 ± 0.46	91.33 ± 0.49

5 Conclusion

In this study, we aim to enhance the image classification task. For that, we propose a new sparse encoding method in order to improve the dictionary learning and the sparse coding process. Indeed, the incorporation of spatial locality among the features in the image space ensures the consistency between the sparse codes and the local regions centred around their corresponding local features. Furthermore, we propose a new pooling scheme that adapt BoSSA pooling on the novel sparse codes. This enables us to take into account the distribution of the sparse codes weights around each vector basis. Experimental results proves the efficiency of the proposed approach.

References

1. Avila, S., Thome, N., Cord, M., Valle, E., de Albuquerque Araújo, A.: Bossa: Extended bow formalism for image classification. In: 18th IEEE International Conference on Image Processing, pp. 2909–2912 (2011)
2. Ben Aoun, N., Mejdoub, M., Ben Amar, C.: Graph-based approach for human action recognition using spatio-temporal features. J. Visual Communication and Image Representation 25(2), 329–338 (2014)
3. Dammak, M., Mejdoub, M., Ben Amar, C.: A survey of extended methods to the bag of visual words for image categorization and retrieval. In: 9th International Conference on Computer Vision Theory and Application, pp. 676–683 (2014)
4. Dammak, M., Mejdoub, M., Zaied, M., Amar, C.B.: Feature vector approximation based on wavelet network. In: ICAART, vol. (1), pp. 394–399 (2012)
5. Gao, S., Tsang, I.W.H., Chia, L.T.: Sparse representation with kernels. IEEE Transactions on Image Processing 22(2), 423–434 (2013)
6. Gao, S., Tsang, I.W.H., Chia, L.T., Zhao, P.: Local features are not lonely: Laplacian sparse coding for image classification. In: 23th IEEE Conference on Computer Vision and Pattern Recognition, pp. 3555–3561 (2010)
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2169–2178. IEEE Computer Society (2006)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2) (2004)
9. Mejdoub, M., Ben Amar, C.: Classification improvement of local feature vectors over the knn algorithm. Multimedia Tools and Applications 64(1), 197–218 (2013)
10. Morioka, N., Satoh, S.: Learning directional local pairwise bases with sparse coding. In: Proceedings of the British Machine Vision Conference, pp. 1–11 (2010)
11. Nowak, E., Jurie, F., Triggs, B.: Sampling stra9th european conference on computer vision,tegies for bag of features features image classification. In: 9th European Conference on Computer Vision, pp. 490–503 (2006)
12. Ren, W., Huang, Y., Zhao, X., Huang, K., Tan, T.: Local hypersphere coding based on edges between visual words. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 190–203. Springer, Heidelberg (2013)
13. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: A comparison of color features for visual concept classification. In: ACM International Conference on Image and Video Retrieval (2008)
14. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: 23th IEEE Conference on Computer Vision and Pattern Recognition, pp. 3360–3367 (2010)
15. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: 22th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1794–1801 (2009)