LAPLACIAN TENSOR SPARSE CODING FOR IMAGE CATEGORIZATION

Mouna Dammak, Mahmoud Mejdoub, Chokri Ben Amar

REGIM: REsearch Groups on Intelligent Machines,
University of Sfax, National Engineering School of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia

ABSTRACT

To generate the visual codebook, a step of quantization process is obligatory. Several works have proved the efficiency of sparse coding in feature quantization process of BoW based image representation. Furthermore, it is an important method which encodes the original signal in a sparse signal space. Yet, this method neglects the relationships among features. To reduce the impact of this issue, we suggest in this paper, a Laplacian Tensor sparse coding method, which will aim to profit from the relationship among the local features. Precisely, we propose to apply the similarity of tensor descriptors to create a Laplacian Tensor similarity matrix, which can better present in the same time the closeness of local features in the data space and the topological relationship among the spatially near local descriptors. Moreover, we integrate statistical analysis applied to the local features assigned to each visual word in the pooling step. Our experimental results prove that our method prevails or exceeds existing background results.

Index Terms— Sparse Coding, Tensor, Bag of words, Image categorization

1. INTRODUCTION

The concept of the Bag of Words (BoW) [1] approach is to quantize local invariant features into a group of visual words. Then, the image is represented by the occurency vector of the visual words. Several studies [2, 3, 4, 5, 6, 7, 8, 9] have proved the performance of BoW in numerous computer vision applications such as image categorization, image and video retrieval. However, the BoW model represents each local descriptor as a predetermined visual word and symbolizes the local descriptors of an image into a disordered histogram, which may ignore some significant information of local features and misplace spatial information maintained in the local regions of the image. To avoid this issue, Lazebnik et al. [10] have incorporated the spatial information of local regions into the BoW model, using Spatial Pyramid Matching Kernel (SPM). Explicitly, each image is split into progressively finer partitions and Pyramid Match Kernel [11] is applied to match corresponding partitions. Yang et al. [12] have extended SPM by providing Sparse Coding (referred to as ScSPM), and have shown background achievement in image categorization. Furthermore, it tries to detect the optimal weight to be attributed to the visual words for each local feature, replacing k-means with sparse coding. Thus, the preciseness of the quantization process is enhanced. Then, SPM based maximum pooling is applied to fuse all the local features in the image representation. However, local features are treated independently. The reciprocal dependency among local features is disregarded, ensuing that the sparse coding may differ widely even for close features. Different extensions of sparse coding method [13, 12, 14] have been suggested recently by adding some regularization or constraints. Kavukcuoglu et al. [15] proposed a spatial sparse coding for local feature extraction by combining similar filter outputs over spatially neighboring regions. LScSPM learns an unsupervised dictionary, as well as the sparse representation that gardes the conformity of close local descriptors. They have used a histogram intersection similarity method to construct a Laplacian matrix. In this paper, we exploit the tensor descriptors which shows its good approximation of insightful similarity measures between descriptors [16]. The contributions of this study can be summarized as follows: we propose a more robust Laplacian Tensor Sparse Coding for feature quantization. By introducing Laplacian Tensor similarity and tensor descriptors, our new formulation takes into account respectively the consistency of the sparse codes for the close local descriptors and the preservation of the topological relationship among the spatially near local descriptors. Moreover, we employ Bag Of Statistical Sampling Analysis (BOSSA) pooling to take in consideration the statistics of the distribution of the local features assigned to each visual word in the pooling step. The remainder of this paper is organized as follows: In Section 2, we will propose our BoSSA pooling method based on Laplacian Tensor similarity. Experimental results on several datasets are described in Section 3. And, Section 4 concludes.

2. BOSSA POOLING BASED ON LAPLACIAN TENSOR SIMILARITY

2.1. Sparse Coding for Codebook Generation

In BoW model, k-means clustering is broadly utilized for the codebook building in which, the local feature space $X = [x_1, \ldots, x_N]$ $\left(x_i \in R^D\right)$ is split into K clusters $s =$

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

$[s_1, \ldots, s_K]$, and the conforming centers $U = [u_1, \ldots, u_K] \in R^{D \times K}$ generate the codebook. Each local feature is assigned to one cluster only. Notice that k-means clustering aims at determining these clusters and reducing the inter-class error. This can be expressed by an optimization problem formula:

$$\min_{U,S} \sum_{i=1}^{K} \sum_{x_j \in s_i} \|x_j - u_i\|^2 = \min_{U,V} \sum_{i=1}^{N} \|x_i - U v_i\|^2 \quad (1)$$
$$subject\ to\ : Card\,(v_i) = 1,\ |v_i| = 1,\ v_i \geq 0, \forall i$$

Where $V = [v_1, v_2, \ldots, v_N]$ is a matrix of weight vectors (where $v_i \in R^{k*1}$and $v_{k,i}$ is the weight of the vector x_i in the cluster u_k). Yet, the restriction that every local feature is only attributed to a single visual word is extremely strict, particularly for the points situated at the frontier of various clusters. The hard restriction $Card\,(v_m) = 1$ on v_m is often released for the soft assignment method. Moreover, to avert each feature to be contributed to overly many clusters, the sparse restriction on the weight vector v_m is always incorporated to the objective function. Then, we reach the optimization issue of sparse coding:

$$min \|X - UV\|^2 + \lambda \sum_i \|v_i\|_1 \quad (2)$$
$$subject\ to\ :\ |u_j| \leq 1 \forall j = 1, \ldots, K$$

2.2. Tensor Sparse Coding

2.2.1. Problem Definition

Sparse coding has proved its efficiency in feature quantization process. Yet, the major drawbacks of this method is that it neglects the consistency of the sparse codes for the close local descriptors. Indeed, several close local features may be attributed to distinct visual words of the codebook because of the susceptibility of quantization. Besides it ignores the spatial relationship between the local feature vectors. But, current studies have proved that the dependency among the features is significant for image classification [17, 18].

2.2.2. Objective Function

To enhance the characterization of the relationship among the local features and reduce the vulnerability of sparse coding, we introduce the Laplacian Tensor similarity matrix L into the optimization equation 2. The Laplacian application ensures that we obtain similar sparse codes for close local descriptors. Furthermore, with the application of the tensor similarity matrix We ensure the encoding of spatial relationship between feature vectors that have nearest locations in the image (see section 2.3). We can reformulate the resulting optimization problem as eq 3:

$$min \|X - UV\|_F^2 + \lambda \sum_i \|v_i\| + \frac{\beta}{2} tr\left(VLV^T\right) \quad (3)$$

Where the Laplacian matrix L is defined as $L = DW$. $W_{i,j}$ evaluates the similarity between the tensor descriptors of the samples x_i and x_j having the sparse code v_i and v_j respectively. Matrix D defined by $D_{ii} = \sum_j W_{i,j}$, provides a natural measure on the data samples.

By simple algebra formulation, the objective function 3 can be reduced to:

$$min \|X - UV\|_F^2 + \lambda \sum_i \|v_i\| + \frac{\beta}{2} \sum_{i,j} \|v_i - v_j\|^2 W_{i,j} \quad (4)$$

The optimization problem is split into two phases : (i) Learning codebook and sparse codes of sample features: For that, we randomly choose samples of some local features selected from the training set to construct the Laplacian Tensor matrix and learn the codebook U. (ii) Learning sparse codes for a new feature : For that, we calculate its k nearest neighbours in the samples and construct a similarity vector W_i optimizing the next objective:

$$\min_U \|x - Uv\|_F^2 + \lambda \|v\|_1 + \beta \sum_i \|v - v_i\|^2 W_i \quad (5)$$

Note that x and v are the new local feature and the sparse code respectively. Subscript i indexes the sample feature and Wi is computed using the similarity between the tensor descriptor of x and the tensor descriptors of x_i.

2.3. Tensor Similarity Matrix

The originality in our study is that we employ tensor descriptors to calculate the similarity between two local features considering the spatial relationship information. For that, we describe the spatial relations between features by the construction of a local graph around every feature. To form the graph, we consider the 8 spatial neighbours as mentioned in figure 1. For every local feature vector x_i, we compute the tensor descriptor by considering all the feature vectors forming its graph. The rows of the tensor descriptor corresponds to the feature vectors forming its representative graph (the 8 spatial local feature vectors neighbours of x_i) (see Figure 1).

Given two tensor descriptors X_A and X_B of x_i and x_j respectively. S_{AB} is the similarity between them. It is defined as $\exp\left(-\frac{dist(X_A, X_B)}{t}\right)$ where

- $dist$ is the distance between the two tensors X_A and X_B, which means the summation of (1) the euclidean distance between x_i and x_j and (2) the mean pairwise euclidean distances between the matched descriptors in

X_A and X_B. For each descriptor in X_A, the matching is carried out by finding the closet descriptor in X_B (in the sense of the euclidean distance)

- t is the mean of the pairwise distances between the tensor descriptors.

In the following step, we use KNN method to form the tensor similarity matrix W. Especially, if X_B (the tensor related to x_j) is in the K nearest neighbors of X_A (the tensor related to x_i) taking into consideration the tensor similarity, then we fix $W_{i,j} = W_{j,i} = 1$, otherwise, we fix $W_{i,j}$ = 0.

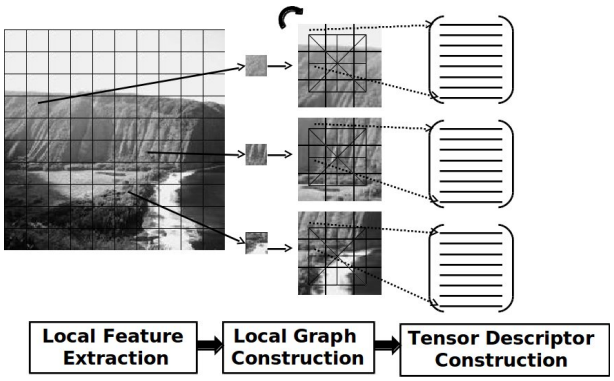

Fig. 1. The construction of Tensor descriptors

2.4. BOSSA Pooling Based Image Representation

We exploit the BOSSA technique [19] because it takes into consideration the statistics of the distribution of the local features assigned to each visual word. The distribution of local descriptors around each visual word is estimated by discretizing for each cluster u_m the weights $v_{m,i}$ over B bins and counting the number of local descriptors falling into each bin. Thus, for each visual word u_m we obtain a local histogram h_m. $h_{m,b}$ corresponds to the number of local descriptors x_j whose $v_{m,j}$ falls into the b^{th} bin. To this local histogram representation is added a scalar t_m, encoding the information regarding the number of visual descriptors d_i corresponding to each visual word u_m. We notice that t_m value corresponds to a classical BoW term. Afterwards, the local histogram $z_m$is L_1 normalized. After calculating a local histogram z_m for each visual word u_m, we concatenate them to construct the image representation z. So, the image representation z can be expressed as follows:

$$z = [[z_{m,b}], t_m] \ ; (m, b) \in \{1, \ldots, K\} * \{1, \ldots, B\} \quad (6)$$

where z is a vector of size $K * (B + 1)$.

3. EXPERIMENTS

3.1. Implementation details

For both datasets Scene-15 and Caltech-256, we extract dense SIFT features [10, 9, 20]. To be compatible with previous researches [12, 10], we exploit the same properties to withdraw SIFT feature. We set the step size and patch size to 8 and 16 respectively. We fix the codebook size to 1024, and choose $(1.0 \sim 1.2) * 10^5$ features randomly to construct codebook for both datasets. To preserve global spatial information, we apply SPM $((1 * 1), (2 * 2), (4 * 4))$ and $((1 * 1), (2 * 2), (3 * 1))$ for Scene- 15 and Caltech-256 respectively without weighting each level. When computing the encoding for each spatial region obtained by the pyramidal representation, the contribution of local features is considered either through max-pooling (TScSPM), in which case each bin in the encoding is assigned a value equal to the maximum across feature encoding in that region or through BOSSA pooling (TScBOSSA). For Bossa pooling, the number of bins B is fixed to 4. There are two important parameters in our objective formula λ: the sparsity of the sparse codes and β: the weight on the closeness restriction. These values are fixed by cross validation: in Caltech 256, we fix $\beta = 0.1$, $\lambda = 0.3$ and in Scene, we fix $\beta = 0.2$, $\lambda = 0.4$. To compute the Tensor similarity matrix, we fix number of nearest neighbors to 5 for the KNN. For classification, we apply the non-linear $chi2-$kernel for BOSSA pooling (TScBOSSA) and the linear kernel SVM for max pooling (TScSPM).

3.2. Scene-15 Dataset

The Scene-15 is a dataset developed by Lazebnic et al. [10]. The images have the same sizes $256 * 256$. It involves 4492 images split into 15 categories, each category from 260 to 440. We take 100 hazardous images per category to train the system and the remainder are used to the test. We repeat the same treatement 10 times and report the mean category accuracy. We record the performance based on diverse approaches in table 1. The latter shows that our TScSPM can attain an extremely high performance on this dataset and exceed ScSPM by approximately 11% by joining the Laplacian Tensor constraint. The reason may be that, this dataset includes considerable textures in every region, which brings about the instability in sparse coding method. By the Laplacian Tensor term, close regions will be encoded into similar sparse codes. So, we can represent the image more precisely. We can also notice that our TScSPM exceeds LScSPM. This is the outcome of the incorporation of the spatial relationship between local feature vectors thanks to the Tensor similarity matrix application.

Number of training images	15	30	45	60
Method	Average Categorization Rate(%)			
KSPM [22]	-	34.10	-	-
KC [23]	-	27.17 ± 0.46	-	-
ScSPM[12]	27.73 ± 0.51	34.02 ± 0.35	37.46 ± 0.55	40.14 ± 0.91
LScSPM[14]	29.99 ± 0.15	35.74 ± 0.1	38.47 ± 0.51	40.32 ± 0.32
TScSPM (our)	32.12 ± 0.17	38.02 ± 0.15	40.53 ± 0.2	42.46 ± 0.25
TScBOSSA (our)	34.14 ± 0.25	40.09 ± 0.2	42.56 ± 0.12	44.39 ± 0.34

Table 2. Performance Comparison on Caltech-256 Dataset

Method	Average Classification rate(%)
KSPM [10]	81.4 ± 0.5
ScSPM[12]	80.28 ± 0.93
HIK+OCSVM [21]	84 ± 0.46
LScSPM[14]	89.75 ± 0.5
TScSPM (our)	90.78 ± 0.5
TScBOSSA (our)	91.79 ± 0.5

Table 1. Performance Comparison on Scene-15 Dataset

3.3. Caltech-256 Dataset

The Caltech-256 dataset contains 30607 images in 257 diverse object categories. There are many improvements involving higher intra-class changeability and higher changeability in object poses and locations. We delete the clutter category. We assess our approach under the four various settings: 15, 30, 45 and 60 training images. The results of this dataset are mentioned in table 2. From this table, we observe that our method outperforms background performance on this dataset. This demonstrates that by the application ofLaplacian Tensor similarity matrix, the relationships among local feature vectors can be better presented.

3.4. Impact of BoSSA pooling on our new encoding method

In this experiment, we compare the two different pooling strategies : Max-pooling and BoSSA in our new encoding Tensor Sparse Coding. TScBOSSA exceeds both TScSPM and LScSPM representations. If we compare TScBOSSA with TScSPM, we will observe an increase about 1% and 2% for Scene-15 and Caltech-256 respectively. These results confirm the advantages introduced by TScBOSSA representation. We conclude that the combination of BoSSA pooling with Tensor Laplacian Sparse Coding enhances the image classification results.

4. CONCLUSION

In this research, we suggest a more perfect sparse coding method called Laplacian Tensor Sparse Coding which can be exploited to learn the codebook and quantize local features more precisely. The proposed method ensures the consistency of the sparse codes for the close local descriptors and the preservation of the topological relationship among the spatially near local descriptors. We also apply BoSSA pooling in order to improve the pooling step on the BOW construction. Experimental results proved the efficiency of our approach.

5. REFERENCES

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proceedings of the International Conference on Computer Vision*, 2003, vol. 2, pp. 1470–1477.

[2] Mouna Dammak, Mahmoud Mejdoub, Mourad Zaied, and Chokri Ben Amar, "Feature vector approximation based on wavelet network," in *ICAART (1)*, 2012, pp. 394–399.

[3] Gaurav Sharma, Frederic Jurie, and Cordelia Schmid, "Discriminative spatial saliency for image classification," in *Computer Vision and Pattern Recognition, CVPR 2012, June, 2012*, Providence, RI, Etats-Unis, 2012.

[4] Manel Sekma, Mahmoud Mejdoub, and Chokri Ben Amar, "Human action recognition using temporal segmentation and accordion representation," in *Computer Analysis of Images and Patterns*, 2013, pp. 563–570.

[5] Rahat Khan, Cécile Barat, Damien Muselet, and Christophe Ducottet, "Spatial orientations of visual word pairs to improve bag-of-visual-words model," in *BMVC*, 2012.

[6] Najib Ben Aoun, Mahmoud Mejdoub, and Chokri Ben Amar, "Graph-based approach for human action recognition using spatio-temporal features," *J. Visual Communication and Image Representation*, vol. 25, no. 2, pp. 329–338, 2014.

[7] Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce, "Learning mid-level features for recognition," in *International Conference on Computer Vision and Pattern Recognition*. 2010, IEEE.

[8] Mouna Dammak, Mahmoud Mejdoub, and Chokri Ben Amar, "A survey of extended methods to the bag of visualwords for image categorization and retrieval," in *VISAPP*, 2014, pp. 676–683.

[9] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *BMVC*, 2011.

[10] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[11] Kristen Grauman and Trevor Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *In ICCV*, 2005, pp. 1458–1465.

[12] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2009.

[13] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1–8.

[14] Shenghua Gao, Ivor Wai Hung Tsang, Liang Tien Chia, and Peilin Zhao, "Local features are not lonely: Laplacian sparse coding for image classification," in *CVPR*, 2010.

[15] K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun, "Learning invariant features through topographic filter maps," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 1605–1612.

[16] Xian-Hua Han, Yen wei Chen, and Xiang Ruan, "Multilinear supervised neighborhood embedding of a local descriptor tensor for scene/object recognition," *Image Processing, IEEE Transactions on*, vol. 21, no. 3, pp. 1314–1326, 2012.

[17] Ni Bingbing, Yan Shuicheng, Wang Meng, Ashraf A. Kassim, and Tian Qi, "High order local spatial context modeling by spatialized random forest," *IEEE Transactions on Image Processing*, vol. 22, no. 2, 2013.

[18] Liang Zheng and Shengjin Wang, "Visual phraselet: Refining spatial constraints for large scale image search," *Signal Processing Letters, IEEE*, vol. 20, no. 4, 2013.

[19] Sandra Eliza Fontes de Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo de Albuquerque Araújo, "Bossa: Extended bow formalism for image classification," in *ICIP*, 2011, pp. 2909–2912.

[20] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, 2004.

[21] Jianxin Wu and J.M. Rehg, "Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 630–637.

[22] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.

[23] Jan C. Gemert, Jan-Mark Geusebroek, Cor J. Veenman, and Arnold W. Smeulders, "Kernel codebooks for scene categorization," in *ECCV*, 2008, pp. 696–709.