Locally Linear Salient Coding for Image Classification

Mohammadreza Babaee, Gerhard Rigoll
Institute for Human-Machine Communication,
Technische Universität München,
Munich, Germany
{reza.babaee,rigoll}@tum.de

Reza Bahmanyar, Mihai Datcu
Remote Sensing Technology Institute (IMF),
German Aerospace Center (DLR),
Oberpfaffenhofen, Germany
{gholamreza.bahmanyar,mihai.datcu}@dlr.de

***Abstract*—Representing images with their descriptive features is the fundamental problem in CBIR. Feature coding as a key-step in feature description has attracted the attentions in recent years. Among the proposed coding strategies, Bag-of-Words (BoW) is the most widely used model. Recently saliency has been mentioned as the fundamental characteristic of BoW. Base on this idea, Salient Coding (SaC) has been introduced. Empirical studies show that SaC is not able to represent the global structure of data with small number of codewords. In this paper, we remedy this limitation by introducing *Locally Linear Salient Coding (LLSaC)*. This method discovers the global structure of the data by exploiting the local linear reconstructions of the data points. This knowledge in addition to the salient responses, provided by SaC, helps to describe the structure of the data even with a few codewords. Experimental results show that LLSaC obtains state-of-the-art results on various data types such as multimedia and Earth Observation.**

Keywords*—*Content-Based Image Retrieval, Feature Coding, Salient Coding, Locally Linear Embedding

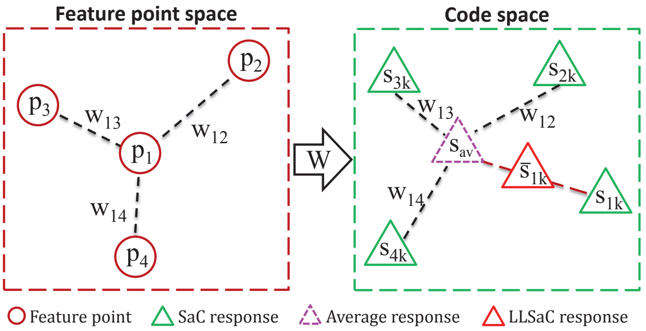

Fig. 1. The main framework of the proposed LLSaC method. The reconstruction weights computed in feature point space are passed to the code space. There, the salient response of each data point is updated by the weighted average of the responses from its neighbors, where the responses are weighted by their corresponding reconstruction weights.

I. Introduction

Exploiting the large volume of the available data (e.g., multimedia, Earth Observation) requires developing efficient CBIR systems. The fundamental problem of any CBIR system is to provide descriptive representations of images. In recent years, Bag-of-Words (BoW) [1] model, a codebook-based image description technique, has been widely used in CBIR and visual indexing problems. BoW model is basically composed of four main steps, e.g., local feature extraction, codebook generation, feature coding, and pooling.

In the first step, various primitive features of an image (e.g., color, texture, shape) are described as vectors of analytical components, so-called *feature vectors*, for every local patches of the image using various methods, e.g., rgb-Hist [2], WLD [3], and SIFT [4]. These vectors form a high-dimensional euclidean space, so-called *feature space*, where each vector is represented as a point there. In the next step, the structure behind the distribution of the feature points is modeled by a *codebook* which is a set of points, so-called *codewords*. The codewords are usually generated by applying a clustering technique (e.g., k-means) on random samples of the feature points. For each image, in the encoding step, a code matrix is generated, where each row shows the responses of a local feature point to different codewords. In order to compute the response values, various coding schemes has been introduced in recent years, for example, voting, reconstruction, and salient -based methods [5]. Finally, the responses of all the local feature points to each codeword are integrated to form a single code value using a pooling technique, e.g., sum, average, and maximum pooling [6]. The output of the pooling step is a histogram with the number of bins equal to the number of the codewords which is then used by learning algorithms.

Variety of the possibilities to select a subset of the codewords to describe a feature point as well as various methods to compute the response values make encoding step a hot topic in CBIR. The classic coding method is *Hard Voting (HV)* [1], which counts the number of the nearest neighboring feature points to every codeword as the code value. Using a kernel function, *Soft Voting (SV)* [7] is proposed to not only consider the distances between the feature points and the codewords, but also allow each feature point to be described by more than one codeword. In order to provide more descriptive information about the feature points by the codewords, reconstruction-based methods have been applied to coding scenarios. In these methods, each feature point is reconstructed by a group of codewords constrained by the number of contributed codewords (e.g., *Sparse Coding (SC)* [8]) and the locality of the codewords such as in *LCC* [9] and *LLC* [10]. Considering the locality of the codewords in combination with the maximum pooling in LLC leads to salient representation of the feature points. More precisely, if K nearest codewords are used to

978-1-4799-3990-9/14/$31.00 ©2014 IEEE

code a feature point, closer codewords to the feature point will receive a stronger response than the others. In order to represent the salient characteristics of the feature points and to avoid the computation cost of the LLC method, *Salient Coding (SaC)* [11] and its variants such as *Group Salient Coding (GSC)* [12] have been proposed.

In all the proposed coding methods, each feature point is coded independent of the other points. However, it has been shown in literature [13] that the relations between the neighboring points help discovering the global structure of data. Therefore, in this paper, we propose *Locally Linear Salient Coding (LLSaC)* which is a new variant of SaC. In contrast to the previous coding methods, LLSaC codes the local structures of the feature space instead of only a feature point. In the proposed method, the local structure of the data is discovered by a set of linear coefficients which reconstruct each feature point from its neighbors. The computed coefficients are then used to update the salient response of the feature point, where the salient response is obtained similar to the SaC method. The idea is that if the coefficients can reconstruct a feature point from its neighbors, they should also be able to reconstruct the response of the feature point to a codeword from the responses of its neighboring points. Figure 1 shows the main idea behind our proposed method. In the feature point space, the reconstruction weights w_{ij} are computed based on the linear reconstruction of the point p_1 from its neighbors. Then this knowledge (i.e., reconstruction weights) are passed to the code space. There, the salient response of p_1 to the codeword k, provided by SaC, is updated by the weighted average of the responses of its neighboring points, where the responses are weighted by their corresponding reconstruction weights computed in the feature space.

In order to evaluate the proposed method, it compared to SaC and other coding strategies such as HV, SV, and LLC on 15 natural scenes dataset. Results show that LLSaC improves SaC significantly to outperforms the other coding strategies even for small codebook sizes. Since large number of codewords besides improving learning performance introduces problems such as the storage problem, the curse of dimensionality which increase the computation effort, and the limited degree of freedom [14]. Therefore, developing coding strategies which help learning methods to achieve high accuracies with small codebook sizes make them more scalable. Moreover, it is shown that LLSaC also improves SaC in case of other data types such as Earth Observation which verifies the generality of the proposed method.

Rest of the paper is organized as follows: Section II and III provide brief reviews of Salient Coding and the linear representation of non-linear structures, respectively. Section IV explains the LLSaC method. Results are then discussed in Section V. Finally, the paper is concluded in Section VI.

II. Salient coding

Based on the idea that saliency is a fundamental characteristic of the feature space and the codebook-based coding strategies, *Salient Coding (SaC)* [11] has been introduced. SaC considers the relative distances of the feature points and the codewords in combination with maximum pooling to code the saliency information of the points. In other words, the codeword which is relatively close to a feature point can strongly describe the point independent of the other codewords. Therefore, the salient response s_{ik} of the feature point p_i to the codeword b_k is obtained by:

$$s_{ik} = \begin{cases} \Psi(p_i) & if\ k = \text{argmin}_k \|p_i - b_k\|_2 \\ 0 & else \end{cases} \quad , \tag{1}$$

$$\Psi(p_i) = \Phi(\frac{\sum_t(\|p_i - b_t\|_2 - \|p_i - b_k\|_2)}{\sum_t \|p_i - b_t\|_2}),\ t \in NC(b_k), \tag{2}$$

where Φ is a monotonically decreasing function and $NC(b_k)$ is a set of K nearest codewords to the feature point p_i. According to the Equation 1, each feature point only responds to its nearest codeword which results in a hard assignment strategy [12].

III. Linear representation of non-linear structures

Using linear coefficients to represent the non-linear structure of data has been introduced first by Saul and Roweis [13] in their proposed neighborhood preserving dimensionality reduction method, the so-called *Locally Linear Embedding (LLE)*. The idea is that every original N-dimensional data point $p_i \in \mathbb{R}^N$ can be reconstructed by a linear combination of its neighboring points $p_j \in \mathbb{R}^N$, given a set of weighs $w_{ij} \in W$. To compute the weights that best reconstruct the data points, the following cost function should be minimized,

$$E(W) = \sum_i |p_i - \sum_j w_{ij} p_j|^2. \tag{3}$$

In this function, w_{ij} determines the contribution of p_j to the reconstruction of p_i. Thus, each row of the matrix W should sum to one, $\sum_j w_{ij} = 1$. Moreover, in order to allow only the contributions of the neighbors, for every non-neighboring points $w_{ij} = 0$. The optimal weights are obtained in closed form by solving a least square problem. For more details about computing optimal weights, we refer readers to [15].

IV. Locally linear salient coding

In this section we explain our proposed *Locally Linear Salient Coding (LLSaC)* method which is a new variant of SaC method. SaC has been introduced to use the local saliency of points. However, as it is mentioned in the original article [11], for small codebooks SaC is worse than other coding techniques such as HV and SV. Since SaC codes each feature point independent of the other points, the small number of codewords cannot represent the structure of the entire feature space. Therefore, SaC is highly sensitive to the codebook size and only outperforms the other schemes for large codebooks.

In order to overcome the limitation of SaC in representing the structure of the feature space with small number of codewords, we introduce LLSaC in this paper. LLSaC codes the local patches neighboring to every feature points in order to provide codewords with more informative responses.

LLSaC discovers the structure of the feature space based on a set of linear coefficients which construct each feature point from its neighbors. The idea of linear representation of non-linear structures has been introduced in [13] for locally linear embedding of high-dimensional points into a lower-dimensional space. Based on this idea, we claim that if the point p_i is reconstructed from its neighbors p_j using linear coefficients w_{ij}, the response s_{ik} of p_i to the codeword b_k should also be reconstructed from the responses (s_{jk}) of its neighboring points,

$$s_{ik} \simeq \sum_{j} w_{ij} s_{jk}, \;\; j \in NN(p_i), \tag{4}$$

where $NN(p_i)$ is the set of $\bar{K}$ nearest feature points to the point p_i. Thus, first, the responses of the feature points to the codewords are computed using SaC method. Then, the salient responses of every feature points p_i (i.e., s_{ik}) is updated by:

$$\bar{s}_{ik} = \frac{1}{2}(\sum_{j} w_{ij} s_{jk} + s_{ik}), \;\; j \in NN(p_i). \tag{5}$$

Finally, the updated salient codes $(\bar{s}_{ik})$ are integrated using maximum pooling to form the final image descriptor.

V. Experiments and Results

In this section, LLSaC is compared to SaC and other coding schemes such as HV, SV, and LLC. In order to be consistent with the previous feature coding articles (e.g., [5], [11], [16]), we use the coding toolkit developed by [16]. Moreover, in order to compare LLSaC with SaC more precisely, the experiments are run on 15 natural scenes dataset[1] and the results are compared to the reported results in the original paper [11]. In addition to this dataset, LLSaC is evaluated on an Earth Observation dataset, the so-called UCMerced-LandUse dataset[2].

A. Datasets

The 15 natural scenes dataset is a collection of 4485 gray value images of outdoor and indoor scenes. The images are grouped into 15 non-equal size categories, where each contains between 200 and 400 images. In our experiments, 100 images from each category are randomly selected as training samples and the rest are used to test the learned model.
UCMerced-LadUse is a collection of 2100 multi-spectral images of land-use scenes. The images are categorized into 21 classes, where each class contains 100 images. For our experiments, 40 images from each class are randomly selected for training and the rest are used for testing. Figure 2 shows some representative samples of the datasets.

B. Experimental setups

In order to be consistent with the experimental setup in the previous works in feature coding area (e.g., [5], [11], [16]), the 128 dimensional SIFT descriptors are extracted densely for every 4 pixels. In order to provide richer description of the image features, SIFT is extracted for three scales: 16×16, 24×24, and 32×32. Then, k-means clustering is applied to samples of the local feature points to generate codebooks of various sizes. In coding step, the parameter K (i.e., the number of nearest codewords) is fixed to K = 5 for LLSaC, SaC, and LLC according to [11].

(a) (b)

Fig. 2. Representative samples from (a) 15 natural scenes and (b) UCMerced-LandUse datasets.

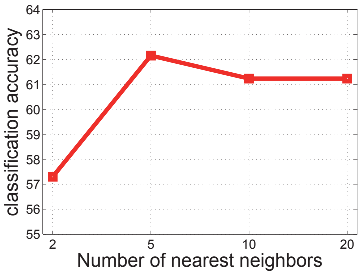

Fig. 3. Performance of LLSaC on 15 natural scenes dataset for different $\bar{K}$.

In LLSaC, the number of neighbors which reconstruct the feature points is fixed to $\bar{K} = 5$ based on an empirical study. We study the influence of $\bar{K}$ to the performance of our proposed method in classification of 15 natural scene image collection. Figure 3 shows the performances for $\bar{K} = 2, 5, 10, 20$ under the codebook size of 16. The figure indicates that the small number of neighbors cannot provide enough information about the structure of the data. However, using too many neighbors affects the locality of the reconstruction weights.

In order to compare the performances of the coding strategies, they are used to classify images using a classification method, so-called SVM[3]. The setup parameters of SVM such as cost and gamma are set to 1 according to [11]. Then the classification accuracies are reported for various codebook sizes. For each codebook size the experiments are run 10 times and the average result is presented.

C. Results and discussions

In order to evaluate the proposed method, its performance on classification tasks are compared to the other coding methods for two datasets. Figure 4 shows the classification accuracies for LLSaC, SaC, HV, SV, and LLC. As the graph shows, LLSaC outperforms all the other methods under various

[1] http://www-cvr.ai.uiuc.edu/ponce_grp/data
[2] http://vision.ucmerced.edu/datasets/landuse.htmls
[3] http://www.csie.ntu.edu.tw/cjlin/liblinear/

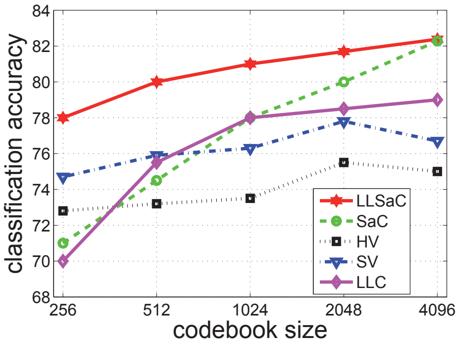

Fig. 4. Performance comparison of LLSaC and other coding schemes under different codebook sizes on 15 natural scenes dataset. Results for SaC, HV, SV, and LLC are reported from [11].

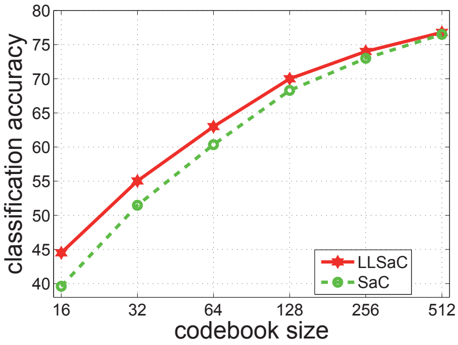

Fig. 5. Performance comparison of LLSaC and SaC under different codebook sizes on UCMerced-LandUse dataset.

codebook sizes. However, as the dictionary size increases the performance of LLSaC converges to the results of SaC. Since LLSaC provides the codewords with the responses from local structures of the feature space, even small number of codewords can discover the global structure of the data. Therefore, LLSaC is more robust under the change in the codebook size. However, as the number of codewords increases enough (about 80% of the number of local feature points in each image), they can represent the structure of the data with no need for extra information from the local structures. Consequently, LLSaC and SaC performs similarly for large number of codewords.

Figure 5 shows the improvement achieved by updating the SaC codes using the responses of the neighboring codes in LLSaC on UCMerced-LandUse dataset. Since various kinds of data (e.g., multimedia, Earth Observation) result in different topologies in feature space, evaluating the coding strategies on Earth Observation data verifies the generality of the methods. Results indicate that LLSaC surpasses SaC also on this dataset.

VI. Conclusion

In this paper we propose *Local Linear Salient Coding (LLSaC)*, a new variant of Salient Coding (SaC). This method remedies the limitations of SaC in representing the structure of the feature space by small number of codewords. LLSaC discovers the global structure of the data by exploiting the local linear reconstructions of the data points. This knowledge is then used to update the salient responses resulted by SaC. Experimental results indicate that LLSaC is able to describe the structure of the feature space even with a few codewords.

In this paper, the locally linear reconstruction technique is applied to SaC; however, this technique can be seen as an independent wrapper which can be applied to other codebook-based coding strategies. This allows the coding methods to use the local information of the feature space in addition to the responses of each individual feature point to the codewords. Thus, for future works, we suggest to apply this technique to other feature coding strategies such as LLC, SV, and the variants of SaC.

Acknowledgment

The authors would like to thank Munich Aerospace for the financial support of this research.

References

[1] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[2] K. E. A. Van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1582–1596, Sept 2010.

[3] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao, "Wld: A robust local image descriptor," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1705–1720, Sept 2010.

[4] D. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2, 1999, pp. 1150–1157 vol.2.

[5] Y. Huang, Z. Wu, L. Wang, and T. Tan, "Feature coding in image classification: A comprehensive study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. PrePrints, p. 1, 2013.

[6] Y.-L. Boureau, J. Ponce, and Y. Lecun, "A theoretical analysis of feature pooling in visual recognition," in *27TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, HAIFA, ISRAEL*, 2010.

[7] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in *ECCV 2008, PART III. LNCS*. Springer, 2008, pp. 696–709.

[8] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009, pp. 1794–1801.

[9] K. Yu, T. Zhang, and Y. Gong, in *NIPS*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., pp. 2223–2231.

[10] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IN: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN CLASSIFICATOIN*, 2010.

[11] Y. Huang, K. Huang, Y. Yu, and T. Tan, "Salient coding for image classification," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 1753–1760, 2011.

[12] Z. Wu, Y. Huang, L. W. 0001, and T. Tan, "Group encoding of local features in image classification." in *ICPR*. IEEE, 2012, pp. 1505–1508.

[13] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *SCIENCE*, vol. 290, pp. 2323–2326, 2000.

[14] H. Liu, Z. Yang, Z. Wu, and X. Li, "A-optimal non-negative projection for image representation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 1592–1599.

[15] L. K. Saul and S. T. Roweis, "An Introduction to Locally Linear Embedding," Tech. Rep., 2000.

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