Aggregating Descriptors with Local Gaussian Metrics

Hideki Nakayama
Grad. School of Information Science and Technology
The University of Tokyo
Tokyo, JAPAN
nakayama@ci.i.u-tokyo.ac.jp

Abstract

Recently, large-scale image classification has made a remarkable progress because of the significant advancement in the representation of image features. To realize scalable systems that can handle millions of training samples and tens of thousands of categories, it is crucially important to develop discriminative image signatures that are compatible to linear classifiers. One of the promising approaches to realize this is to encode high-level statistics of local features. Many state-of-the-art large-scale systems are following this approach and have made remarkable progress over the past few years. However, while first-order statistics are frequently used in many methods, the power of higher-order statistics has not received much attention.

In this work, we propose an efficient method to exploit the second-order statistics of local features. For each visual word, the local features of training samples are modeled with a Gaussian, and descriptors from two images are compared using a Fisher vector with respect to the Gaussian. In experiments, we show the promising performance of our method.

1 Introduction

Recently, remarkable progress has been made in the large-scale categorization of images. One of the breakthroughs that has made this possible is the advancement of the representation of image features that are compatible to linear classifiers. Hitherto, most image-categorization systems have used small training datasets and depended on non-linear classifiers such as kernel SVMs. However, these systems cannot be scaled to handle larger data because the computational complexity for training non-linear classifiers is generally $O(N^2) \sim O(N^3)$, where N is the number of training samples [22]. Therefore, linear classification is probably the only choice to accomplish large-scale training within a realistic time frame. However, to successfully apply linear classifiers, we also need to exploit linearly separable image signatures. Moreover, these signatures should have a high discriminative power so as to be able to distinguish tens of thousands of image categories. In this work, we focus on this topic and develop a new feature coding method based on the statistics of local descriptors.

2 Related Work and Our Contribution

Previous work on image-feature coding falls basically into two main approaches. The first is the vector quantization-based approach, also known as bag-of-visual-words (BoVW) [4]. This has been the de-facto standard method for the image-categorization problem for a long time. However, as the original BoVW vector has a strong non-linear property, it must be used with a kernel classifier to obtain reasonable performance. This is prohibitive for the large-scale problems as we have described. Therefore, state-of-the-art BoVW techniques are designed so that the resultant feature vector is directly applicable to linear classifiers. This can be achieved by explicit kernel embedding [14, 17] of a traditional BoVW, or new coding methods based on sparse coding [22, 21].

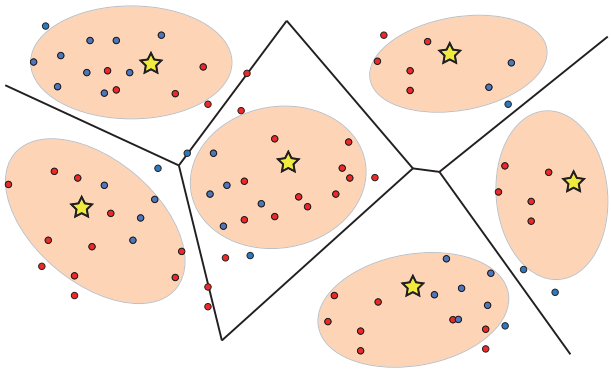

Figure 1: Illustration of our method. Stars represent visual words. Red dots and blue dots represent descriptors from each single image. These descriptors are compared using the Gaussian metrics specific for each visual word.

The second approach goes beyond the simple count statistics of the traditional BoVW technique and utilizes higher-order statistics of local features. For example, the VLAD method [10] uses the sum of difference in the vectors of local features from their nearest visual words. This is interpreted as exploiting the first-order statistics of local features. Super-vector coding [23] concatenates the count statistics and the first-order statistics, whereas the Fisher vector [18] utilizes mean and variance statistics in a sophisticated information geometry framework [9]. Local features from training corpus are modeled with a Gaussian Mixture Model (GMM), and then each image is represented by the deviation from the GMM. In fact, VLAD can be interpreted as a simplified version of the Fisher vector. It has been shown that these high-level statistics can dramatically improve recognition performance with a reduced number of visual words. In addition, as they are carefully derived to be applicable to linear classifiers, they are particularly effective for large-scale problems. For example, the winning systems of the ImageNet large-scale visual recognition challenge 2010 [1] and 2011 [2] used super-vector coding and Fisher vectors, respectively.

In this work, we focus on the second approach. In short, we attempt to include more statistics of local features to further improve the performance of this approach. Because the standard Fisher vector assumes a GMM with diagonal matrices, statistical information related to correlation is missed, which we believe is the key for discrimination. We take the standard SIFT descriptor [13] as an example. As SIFT consists of local edge histograms, correlations of elements correspond to specific middle-level shape patterns over an image. This type of information is thought to provide rich discriminative cues for classification. Of course, in theory, the Fisher vector could utilize such information by modeling a GMM with full covariance matrices; however, the cost would be prohibitive for real problems.

To the best of our knowledge, the VLAT method [19] has been the only method that successfully uses full second-order statistics (variance and co-variance). This method is an extension of VLAD and concatenates the elements of higher-order tensor products. It has been shown that the inner products of VLAT vectors approximate the sum kernel on bags [7] for a Gaussian kernel when sufficiently high-order tensors are exploited. This fact theoretically supports the adequacy of using VLAT with linear classifiers, although, at most the second-order tensors are considered in practice. In this work, we propose an alternative approach for encoding second-order statistics using local Gaussian metrics and explain it in the next section.

3 Our Approach

Our approach is essentially a hybrid of the Fisher approach and VLAT (Fig. 1). First, we compute K visual words $\{\boldsymbol{c}_n\}_{n=1}^{K}$ via k-means clustering as in the usual BoVW. Local features of training samples in each Voronoi cell are modeled with a Gaussian, and descriptors of two images in this area are compared using the Fisher vector with respect to the Gaussian. Although we fit a single Gaussian independently for each visual word (Voronoi cell), we estimate the full covariance matrix of the Gaussian. This is a major difference from the standard implementation of the Fisher vector,

where only the diagonal elements of Gaussians are estimated when fitting a GMM. In this way, we can efficiently exploit the second order statistics of local features with a theoretically supported metric. We name our method the Vectors of Locally Aggregated Gaussian-gradients (VLAG).

3.1 Embedding Local Gaussian Metrics

Nakayama *et al.* [15] proposed a method of modeling whole local features from images with a single Gaussian and derived its Fisher vector-like representation for linear classification. Here we extend this method to handle the local structure of a feature space.

Let $\boldsymbol{x} \in R^d$ denote a local feature. For each area (cell) spanned by a specific visual word $\boldsymbol{c}_n$, we first encode the local features with their sum and correlations centered to $\boldsymbol{c}_n$.

Specifically, for an image I_i,

$$\boldsymbol{\eta}_{i,n} = \left(\begin{array}{c} \sum\limits_{\boldsymbol{x}_i \text{ such that } \mathrm{NN}(\boldsymbol{x}_i)=\boldsymbol{c}_n} (\boldsymbol{x}_i - \boldsymbol{c}_n) \\ upper\left(\sum\limits_{\boldsymbol{x}_i \text{ such that } \mathrm{NN}(\boldsymbol{x}_i)=\boldsymbol{c}_n} (\boldsymbol{x}_i - \boldsymbol{c}_n)(\boldsymbol{x}_i - \boldsymbol{c}_n)^T \right) \end{array} \right), \tag{1}$$

where $upper()$ is the flattened vector of the components in the upper triangular part of a symmetric matrix. Therefore, the dimension of this signature is $d + d(d+1)/2$. The local representation of this is the same as that of VLAT. In reality, this signature corresponds to the gradient vector of local features taking a Gaussian as the generative model [15]. We further normalize this vector with the inverse of the Fisher information matrix as in the Fisher vector framework, which we denote by G_n in the following.

The Gaussian parameters of the local features from the entire training set in the n-th cell are as follows.

$$\boldsymbol{\mu}_n = \frac{1}{|T_n|} \sum_{\boldsymbol{x} \text{ such that } \mathrm{NN}(\boldsymbol{x})=\boldsymbol{c}_n} (\boldsymbol{x} - \boldsymbol{c}_n), \tag{2}$$

$$C_n = \frac{1}{|T_n|} \sum_{\boldsymbol{x} \text{ such that } \mathrm{NN}(\boldsymbol{x})=\boldsymbol{c}_n} (\boldsymbol{x} - \boldsymbol{c}_n - \boldsymbol{\mu}_n)(\boldsymbol{x} - \boldsymbol{c}_n - \boldsymbol{\mu}_n)^T, \tag{3}$$

where $|T_n|$ is the number of sample features. Using them, we can explicitly obtain G_n in a closed form. For details, refer to [15]. Thus, descriptors of an image I_i in the n-the cell are encoded as the Fisher vector of a local Gaussian.

$$\boldsymbol{\zeta}_{i,n} = G_n^{1/2} \boldsymbol{\eta}_{i,n}. \tag{4}$$

Finally, all $\{\boldsymbol{\zeta}_{i,n}\}_{n=1}^K$ are concatenated, and subsequently, are power-law normalized and L2 normalized [18]. This is the resultant image signature of our proposed VLAG method.

4 Experiments

4.1 Setup

We validate our method using standard benchmarks for image-categorization problems. Namely, we use 15 scene [12], Caltech-101 [6], and Caltech-256 [8] datasets. If images are large, they are resized to 10K pixels. For the 15 scene dataset, we use 100 samples per class for training and the remaining samples for testing. For the Caltech-101 and Caltech-256 datasets, we use 30 samples for training and 50 for testing, for each class. Performance is evaluated by the mean of the classification rate for each class. We report the average score of five trials, randomly replacing the samples.

To extract local features, we use the SIFT descriptor [13] for all experiments in this work. We extract local features from 24×24 patches on regular grids with a spacing of 5 pixels. Further, SIFT descriptors are compressed into 32 dimensions using PCA, except for the usual BoVW baseline. For BoVW, we apply histogram intersection kernel and use a non-linear SVM to provide a fair baseline in terms of the recognition accuracy. For other representations, we directly use a linear SVM. We use libsvm [3] and liblinear [5] packages for the implementations of non-linear and linear SVMs, respectively. Note that we do not include any spatial information because we focus on the performance of feature coding methods.

Table 1: Comparison of classification rate on the 15 scene dataset (%).

Dictionary size	200	500	1000	2000	4000	8000			
BoVW	69.1	74.6	76.4	77.4	77.5	77.6			
Dictionary size	1	2	4	8	16	32	64	128	256
VLAD	47.9	56.0	60.4	64.8	68.5	70.5	75.1	74.5	75.9
Fisher vector	63.3	68.0	72.1	75.1	76.7	77.8	79.3	79.6	**80.0**
VLAT	71.9	73.0	74.5	75.7	76.8	77.8	77.1		
VLAG (Ours)	74.5	75.7	76.9	77.6	78.8	79.7	**80.1**		

Table 2: Comparison of classification rate on the Caltech-101 dataset (%).

Dictionary size	200	500	1000	2000	4000	8000			
BoVW	42.0	45.4	47.3	47.8	48.1	47.4			
Dictionary size	1	2	4	8	16	32	64	128	256
VLAD	17.3	24.7	31.6	35.9	41.8	44.6	48.5	49.9	51.8
Fisher vector	31.5	38.5	42.5	47.5	51.2	52.6	54.9	55.5	56.2
VLAT	43.2	45.7	48.2	50.8	53.8	55.7	57.4		
VLAG (Ours)	46.2	49.4	51.0	54.2	56.4	57.9	**58.3**		

Table 3: Comparison of classification rate on the Caltech-256 dataset (%).

Dictionary size	200	500	1000	2000	4000	8000			
BoVW	18.6	20.8	22.0	22.6	22.7	22.4			
Dictionary size	1	2	4	8	16	32	64	128	256
VLAD	5.9	8.2	10.4	13.4	16.4	18.7	20.4	21.3	22.4
Fisher vector	12.6	16.4	18.6	22.1	23.7	25.0	25.7	26.6	27.3
VLAT	18.7	20.2	21.4	23.2	25.2	26.5	27.9		
VLAG (Ours)	20.2	22.4	23.4	25.3	27.7	28.7	**29.5**		

4.2 Experimental Results

Tables 1, 2, and 3 summarize the results for the 15 scene, Caltech-101, and Caltech-256 datasets, respectively. We empirically found that power-law normalization is also effective for VLAD and VLAT. Therefore, we apply it to them when it improves their performance. We did not test VLAG and VLAT with more than 64 visual words as the dimension of their feature vectors become too large. In addition, we observed that 64 visual words already gave a satisfactory performance.

Overall, VLAG and VLAT achieve high accuracy with a small number of visual words. It is surprising that by just using a few or several visual words, their performances are well comparable to those of BoVW with a non-linear kernel using thousands of words. This result shows the power of higher-level statistics for feature coding. Moreover, we observed that VLAG consistently outperforms VLAT, indicating the importance of embedding a better metric to fully exploit the statistical properties of local features.

5 Conclusion

In this work, we proposed a novel method of feature coding using the second-order statistics of local descriptors. We compared our method with closely related methods and showed its effectiveness. Because our image signature is rather high-dimensional but applicable to linear classifiers, it is expected to be more powerful when used for large-scale problems. In future, we would like to apply our method to large-scale problems and compare it with other state-of-the-art feature coding methods such as locality-constrained linear coding [21]. In addition, it would be of interest to include spatial information [12, 11] and introduce efficient compression techniques [20, 16] to realize practical systems.

References

[1] A. Berg, J. Deng, and L. Fei-Fei. ImageNet large scale visual recognition challenge 2010. http://www.image-net.org/challenges/LSVRC/2010/.

[2] A. Berg, J. Deng, S. Satheesh, H. Su, and L. Fei-Fei. ImageNet large scale visual recognition challenge 2011. http://www.image-net.org/challenges/LSVRC/2011/.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 2011.

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

[5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Journal of Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[7] P. Gosselin, M. Cord, and S. Philipp-Foliguet. Kernels on bags for multi-object database retrieval. In *Proc. ACM CIVR*, pages 226–231, 2007.

[8] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical report, California Institute of Technology, 2007.

[9] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Proc. NIPS*, 1999.

[10] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Proc. IEEE CVPR*, pages 3304–3311, 2010.

[11] J. Krapac, J. Verbeek, and F. Jurie. Modeling spatial layout with Fisher vectors for image categorization. In *Proc. IEEE ICCV*, 2011.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE CVPR*, pages 2169–2178, 2006.

[13] D. Lowe. Object recognition from local scale-invariant features. In *Proc. IEEE ICCV*, pages 1150–1157, 1999.

[14] S. Maji and A. Berg. Max-margin additive classifiers for detection. In *Proc. IEEE ICCV*, 2009.

[15] H. Nakayama, T. Harada, and Y. Kuniyoshi. Global Gaussian approach for scene categorization using information geometry. In *Proc. IEEE CVPR*, 2010.

[16] R. Negrel, D. Picard, and P. Gosselin. Compact tensor based image representation for similarity search. In *Proc. IEEE ICIP*, 2012.

[17] F. Perronnin, J. Sánchez, and Y. Liu. Large-scale image categorization with explicit data embedding. In *Proc. IEEE CVPR*, pages 2297–2304, 2010.

[18] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proc. ECCV*, 2010.

[19] D. Picard and P.-H. Gosselin. Improving image similarity with vectors of locally aggregated tensors. In *Proc. IEEE ICIP*, 2011.

[20] J. Sánchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In *Proc. IEEE CVPR*, 2011.

[21] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Proc. IEEE CVPR*, pages 3360–3367, 2010.

[22] J. Yang, K. Yu, Y. Gong, and T. S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Proc. IEEE CVPR*, 2009.

[23] X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. In *Proc. ECCV*, 2010.