Scene Classification using Generalized Local Correlation

Hideki Nakayama Tatsuya Harada Yasuo Kuniyoshi

The University of Tokyo, Grad. School of Information Science and Technology,
Dept. of Mechano-Informatics, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN
{nakayama, harada, kuniyosh}@isi.imi.i.u-tokyo.ac.jp

Abstract

Feature extraction is an important issue for generic image recognition. In recent years, methods based on the bag-of-keypoints technique have been quite successful and are widely used. However, this technique requires the quantization of local patches to build visual words as a preprocessing step, the computational cost of which is enormous. On the other hand, methods based on global image features have been used for a long time. Because global image features can be extracted rapidly, it is relatively easy to use them in practical large-scale systems. However, the performance of global feature methods is usually poor compared to bag-of-keypoints. Therefore, it is essential to develop a more powerful scheme of global feature extraction for achieving practical applications of generic image recognition. In this paper, we show that we can boost the performance of global image features by considering the correlations of local features in addition to the mean. We experimentally verify the effectiveness of our method using standard scene classification benchmark datasets.

1 Introduction

Content-based image recognition and understanding is one of the ultimate goals of computer vision. With the significant advance in computer systems, appearance-based image recognition methods using statistical pattern recognition have been making remarkable progress recently. However, except for some specific applications such as face recognition, few techniques have reached a practical level. The appearance of generic objects and scenes that we see in the real world can vary enormously, even within the same category. In order to allow the recognition of such generic images, image features need to have high expressive power. At the same time, the speed of feature extraction and learning should be as fast as possible, because appearance-based methods will inevitably require an enormous number of training samples to cover the large appearance variations exhibited by real world objects. However, in general, there is a trade-off between performance and speed. The key to achieving practical image recognition is to properly balance this trade-off.

In this paper, we show that we can perform high speed and high performance feature extraction by considering the correlations of local features in addition to the mean. Our method is basically an example of a global feature scheme, and so it does not need heavy preprocessing like bag-of-keypoints. In the experimental section, we verify the effectiveness of our method using standard scene classification benchmark datasets.

2 Bag-of-Keypoints vs. Global Feature

In recent years, local-approach image modeling has been well studied. This approach models an image as a bag of local features, discarding position information. The most well-known and widely-used example of this approach is bag-of-keypoints [1]. The first step of this method is to perform vector quantization of the local features of the training images to obtain centroids, which represent the visual words. The resulting feature is the histogram of visual word occurrences in the image. The benefit of this approach is that we can extract the distribution information of local features effectively. Studies based on this technique have recently obtained very good performance.

However, a major drawback of bag-of-keypoints is that the process of building visual words is computationally quite expensive. Also, the parameter corresponding to the number of visual words affects the generalization ability, and choosing too high a parameter can cause overfitting. Without any prior knowledge of tasks, an experimental tuning process is needed to find the optimal setting for this parameter. Therefore we need to perform trials of vector quantization many times. The computational cost of the learning process thus becomes quite expensive. For this reason, bag-of-keypoints has only been used in relatively small size datasets until now.

On the other hand, global image features, which discard the distribution of local features and describes the whole image feature, have been used for a long time in Content Based Image Retrieval (CBIR) [2]. This can be interpreted as the simple mean of local features densely sampled from an image. Color histogram and edge histogram are representative examples. This scheme does not require a preprocessing step like bag-of-keypoints, and enables extremely fast feature extraction. Therefore, it is relatively easy to apply this scheme to large scale datasets. Also, it is quite stable because it uses only low level statistics (the mean). However, because this approach discards all the information relating to the distribution of local features, it is undeniable that the expression power is inferior to local approach. Although in some datasets global feature methods obtain performance comparable to local approach methods by concatenating multiple descriptors [3], in most cases the performance of global features are worse than that of local approaches.

Thus, bag-of-keypoints and global features provide different points in the trade-off between performance and speed. In this paper, we propose a new scheme of boosting the performance of global image features, without paying a much greater computational cost. To do this we add basic distribution information by calculating the correlation of local features. These are also low level statistics, like the mean.

3 Generalized Local Correlation

Here, we have N training images. Suppose there are $p^{(j)}$ d-dimensional local features $\boldsymbol{v}_k^{(j)}(k \leq p^{(j)})$ in an image $I^{(j)}(j \leq N)$. Normal global image features are interpreted as using the mean of the local features $\boldsymbol{\mu}^{(j)} = \frac{1}{p^{(j)}} \sum_k^{p^{(j)}} \boldsymbol{v}_k^{(j)}$. Our method supplements this mean vector with the correlations of the local features. For example, we can obtain $\frac{1}{2}d(d+1)$ 1st order correlations. We concatenate the mean and correlations as the feature vector of the image $I^{(j)}$. Let $R^{(j)} = \frac{1}{p^{(j)}} \sum_k^{p^{(j)}} \boldsymbol{v}_k^{(j)} \boldsymbol{v}_k^{(j)T}$ denote the auto-correlation matrix of local features. Then we get the feature vector,

$$\boldsymbol{x}^{(j)} = \begin{pmatrix} \boldsymbol{\mu}^{(j)} \\ upper(R^{(j)}) \end{pmatrix}. \tag{1}$$

Here, *upper()* is a function that enumerates the components in the upper triangular part of a matrix. We call this feature as the Generalized Local Correlation (GLC), to emphasize that this scheme can be applied to any generic local descriptor. Of course we can extract even higher order correlations. However, in this paper we use at most 1st order correlations, because the dimension of the feature vector becomes exponentially large when we use higher-order correlations, making it difficult to prevent overfitting.

Moreover, when the dimension of local features d is large, even the number of the 1st order correlations becomes quite large. To address this problem we perform dimensionality reduction using PCA. Let R denote the auto-correlation matrix of local features extracted from all training images, then

$$R = \frac{1}{\sum_j^N p^{(j)}} \sum_j^N p^{(j)} R^{(j)}. \tag{2}$$

We can obtain the projection matrix U by solving an eigen value problem as follows,

$$RU = U\Omega \quad (U^T U = I). \tag{3}$$

Here, Ω is a diagonal matrix having eigenvalues as the elements. We cut off the principal component space at a proper dimension m, and use the first m eigen vectors as the projection matrix U_m. The resultant feature vector can be obtained as follows,

$$\boldsymbol{x}^{(j)} = \begin{pmatrix} \boldsymbol{\mu}^{(j)} \\ upper(U_m^T R^{(j)} U_m) \end{pmatrix}. \tag{4}$$

GLC can also be interpreted as the mean of polynomial combinations of local features. Therefore, it follows the scheme of conventional global image features, and its advantageous properties such as position invariance and additive property are preserved. This is one of the major merits of GLC.

4 Implementation

4.1 Local Feature Extraction

Keypoint Detection: Generally, local feature methods involve two steps. The first one is keypoint detection, and the second one is the feature description at the keypoints. The best known method of local feature description is SIFT [4], which uses Difference of Gaussian filters to perform keypoint detection. However, for image classification, keypoint detection based on filters does not always work effectively. Nowak *et al.* [5], in a study comparing image classification performance achieved by various keypoint detection methods on several datasets, showed that random keypoint detection achieved the best performance. Also, Fei-Fei *et al.* [6] performed classification on 13 image scene datasets, and showed that grid-based keypoint detection gave the best performance. Considering these insights, we perform keypoints detection based on a grid. This strategy is called Dense Sampling.

Feature Description: We use the SIFT descriptor [4] as the local feature descriptor. We space the keypoint M pixels apart, and extract 128-dimensional local feature (Gray-SIFT) from each region of $L \times L$ pixels having the keypoint at the center. For a color image, we extract SIFT descriptions independently from each RGB component and concatenate them to get 384-dimensional local feature (RGB-SIFT). Also, it was shown by Bosch *et al.* [7] that multiscale SIFT feature description can improve the robustness against a scale change. We follow this strategy in our experiments.

To provide a baseline, we also investigate the performance of our method using edge histograms and color histograms as local descriptors.

4.2 Classification Method

A common tool for classification in recent work on generic image recognition has been the Support Vector Machine (SVM). In spite of its widespread use, the SVM has a major drawback: the computational cost for learning increases in proportion to the square of the number of training samples. One of the goals of our research is system scalability, so the computational cost of the SVM is not acceptable. Instead, we use Probabilistic Linear Discriminant Analysis (PLDA) [8], a probabilistic interpretation of LDA. The learning process of LDA involves only the solution of a generalized eigenvalue problem. The computational complexity of learning is linear in the number of the training samples. This makes it possible to train the system relatively fast, even in a large scale problem.

Let K denote the number of target classes, Σ_w denote the within-class covariance matrix, and Σ_b denote the between-class covariance matrix. LDA is formulated as the following generalized eigenvalue problem.

$$\Sigma_b W = \acute{\Sigma}_w W \Lambda \quad (W^T \acute{\Sigma}_w W = I). \tag{5}$$

Here, $\acute{\Sigma}_w = \Sigma_w + \alpha I$. α is a parameter to decide the amplitude of the regularization matrix, which is used to prevent overfitting. Currently, we tune α experimentally. The regularization matrix is unnecessary if we have enough samples, but the larger the feature dimension used, the larger the number of required samples.

Let $n = N/K$ denote the number of samples in each class, and $\boldsymbol{\mu}_x$ denote the mean of an image feature over the entire dataset. The following projection maps an image feature $\boldsymbol{x}$ to a point in the latent space:

$$\boldsymbol{u} = \left(\frac{n-1}{n}\right)^{1/2} W^T(\boldsymbol{x} - \boldsymbol{\mu}_x). \tag{6}$$

The covariance of the latent values is given by the following expression:

$$\Psi = max\left(0, \frac{n-1}{n}\Lambda - \frac{1}{n}\right). \tag{7}$$

Using this structure, we classify a newly input sample $\boldsymbol{x}_s$ by maximum likelihood estimation. We assume that $\boldsymbol{u}_s$, the projected point of $\boldsymbol{x}_s$, is generated from a certain class C with probability:

$$p(\boldsymbol{u}_s|\boldsymbol{u}^C_{1...n}) = \mathcal{N}\left(\boldsymbol{u}_s|\frac{n\Psi}{n\Psi+I}\bar{\boldsymbol{u}}^C, I + \frac{\Psi}{n\Psi+I}\right). \tag{8}$$

Here, $\boldsymbol{u}^C_{1...n}$ are latent values of n independent training samples that belong to class C, and $\bar{\boldsymbol{u}}^C$ is the mean of them. We classify $\boldsymbol{x}_s$ to the class which has the largest value of eq.(8).

Although LDA is a classical multivariate analysis method, we can perform optimal classification with probabilistic background using the scheme of PLDA. Moreover, in the classic LDA setting, the dimension of the discriminant space becomes a problem. However, we do not need to do this in PLDA because it automatically weights each dimension according to the discriminant criterion.

5 Data Sets

We experiment with two commonly used scene classification benchmark datasets. One is by Oliva *et al.* [9] (OT8), and one is by Lazebnik *et al.* [10] (LSP15). OT8 consists of 2,688 color images of eight classes shown in Fig. 1. Each class has 260∼410 sample images. We also use OT4N which contains four natural scene classes of OT8 (coast, forest, mountain, open country), and OT4MM which contains four man-made scene classes (highway, inside city, tall building, street).

LSP15 consists of gray images of OT8 plus seven additional classes shown in Fig. 2. In all, it has 4,492 gray images. LSP15 has the largest number of target classes among scene datasets currently in use.

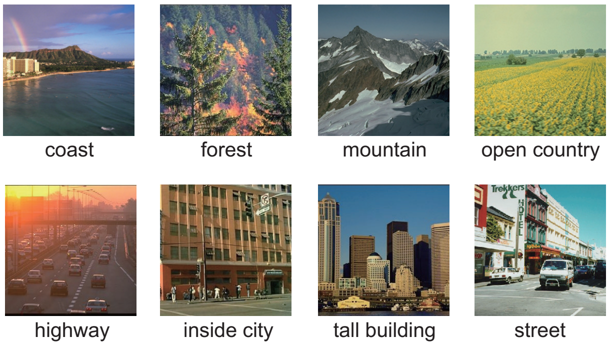

Figure 1: Sample images from the OT8 datasets.

6 Experiment

We randomly choose 100 training images for each class in OT8 and LSP15, 250 in OT4N and OT4MM. We use the remaining samples as test data, and calculate the mean of the classification rate of each class.

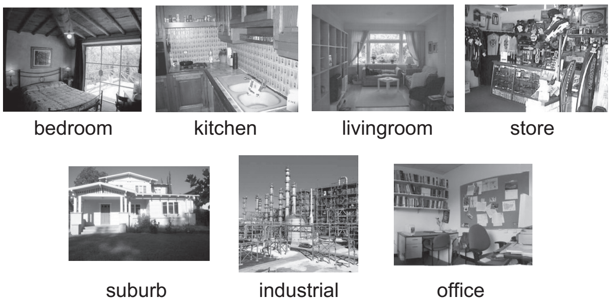

Figure 2: Additional seven classes in LSP15.

This score is far averaged over many trials replacing the training and test samples randomly. In this paper, we use the average over 100 trials.

6.1 Baseline Performance in OT8

Here, we investigate the effectiveness of GLC using three different local descriptors in OT8. We use the local edge histogram, the color histogram, and the SIFT descriptor. For the edge histogram, we extract 72-dimensional gradient direction histogram from gray scale images. For the color histogram, we use the standard 84-dimensional HSV color histogram from color images. We use 36 dimensions for H, 32 dimensions for S, and 16 dimensions for V. For these two descriptors, we fix the parameters of the sliding window as $L = 10$ and $M = 5$. We extract GLC as eq.(1).

As for SIFT descriptor, we fix the parameters as $L = 16$ and $M = 5$. However, because the dimension of SIFT descriptor is large, we perform dimensionality reduction using PCA beforehand, and then extract the 1st order GLC as shown in eq.(4). We use $m = 30$ PCA vectors.

Table 1 shows the performance of each local feature description. "Mean" is the case in which only the mean of local feature is used. This can be interpreted as 0th order GLC, and is very similar to normal global feature. "GLC (1st)" is the case in which we also use the local correlations proposed in this paper. As these results show, the performance considerably improves in each type of local descriptor when GLC is used. Also, it is shown that the performance of SIFT descriptor is significantly better than those of two baseline descriptors.

Table 1: Baseline performance in OT8 (%).

Descriptor	Mean	GLC (1st)
Edge Hist	66.5	73.6
Color Hist	44.9	51.8
Gray-SIFT	73.1	84.8
RGB-SIFT	77.7	86.7

6.2 Comparison with previous works

Next, we compare the performance of our method with pervious works in OT8, LSP15, OT4N, and OT4MM. We use RGB-SIFT as the descriptor for OT8, OT4N and OT4MM, and Gray-SIFT for LSP15.

We extract GLC from four different scales, $L = 8, 16, 24, 32$. These scale parameters are the same ones used by Bosch *et al.* [7].

Recently it has been shown that recognition performance can be improved by adding spatial information using hierarchical partitioning of images [10]. Therefore, much recent work uses Spatial Information (SI). However, using this type of information can be thought of as task fitting and does not always guarantee performance improvement. Our objective is to compare the generic performance of the systems. Therefore, we summarize both cases of previous works, using SI and not using SI. Note that our method does not use SI.

Table 2 shows the result of performance comparison. [7, 10] extract bag-of-keypoints using SIFT descriptor, and perform classification via SVM etc. [11] estimates a part-based generative model of images using Conditional Random Field (CRF), and performs classification and segmentation of an image simultaneously. However, its computational cost is even higher than that of bag-of-keypoints.

It is notable that even in the "Mean" case our approach obtains a comparable performance to those of previous works in "no SI". This is probably because global feature methods are well-suited to the task of scene classification. Also, it appears that multiscale SIFT description increases performance. Furthermore, in case of "GLC (1st)", our method considerably outperforms previous works in "no SI", and even achieves performance close to those in "with SI". Thus, in spite of the simplicity of our feature description and classification methods, it is shown that our system has a performance comparable to state-of-the-art methods.

Table 2: Comparison of the performance in four scene datasets (%).

Dataset	Mean	GLC	Previous	
		(1st)	no SI	with SI
OT8	81.9	88.4	82.3 [11]	90.2 [11]
			82.5 [7]	87.8 [7]
OT4N	86.7	91.8	90.7 [7]	93.9 [7]
				89.0 [9]
OT4MM	89.6	93.9	91.7 [7]	94.8 [7]
				89.0 [9]
LSP15	70.7	79.6	72.7 [7]	83.7 [7]
			74.8 [10]	81.4 [10]

6.3 Discussion of Scalability

Here we estimate the computational cost of final feature extraction and preprocessing respectively. Let p denote the number of local features in an image, d denote the dimension of local features, and V denote the number of visual words in bag-of-keypoints scheme. The computational cost of the final image feature extraction per image is $O(pd^2)$ for our method and $O(pVd)$ for bag-of-keypoints. In most work, V is substantially larger than d. Also, V must be made much larger as the task becomes larger and more complicated.

Furthermore, the bag-of-keypoints method requires a preprocessing step in which the local features are clustered using the K-means algorithm. The computational cost of this process becomes greatly enlarged with the scale of the task, because the number of training samples and V both increase. Moreover, it uses a massive amount of memory to store the local features of all the training samples. Our method does not require a substantial preprocessing step. The only preparation necessary is to find the PCA matrix, and this operation is linear in the number of training samples. Also, it requires a small amount of memory because it needs to preserve only the covariance matrix in memory. Thus, our method is not only accurate but also quite fast and highly scalable.

7 Conclusion

In this paper, we proposed a method to boost the performance of global image feature methods by using local feature correlations. In experiments, we showed that GLC substantially improves performance for a variety of local descriptors. In the quantitative comparison using OT8 and LSP15, we obtained a comparable performance to state-of-the-art methods based on local approach like bag-of-keypoints. In addition, our method is much faster and more scalable than bag-of-keypoints. Therefore, our method is a highly practical one which achieves a good balance in the trade-off between accuracy and speed.

Our future work is to investigate the effectiveness of our method in other benchmarks of various tasks such as object recognition and texture recognition. Also, we will test our method in large scale datasets and try to consider remaining issues such as using higher-order correlations.

References

[1] 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 Compuer Vision*, pp. 1–22, 2004.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 1–32, 2000.

[3] N. Hervé and N. Boujemaa. Image annotation: which approach for realistic databases? In *Proc. ACM CIVR*, 2007.

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

[5] E. Nowak, F. Jurie, and B. Trigges. Sampling strategies for bag-of-features image classification. In *Proc. ECCV*, pp. 490–503, 2006.

[6] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. IEEE CVPR*, pp. 524–531, 2005.

[7] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 712–727, 2008.

[8] S. Ioffe. Probabilistic linear discriminant analysis. In *Proc. ECCV*, pp. 531–542, 2006.

[9] A. Oliva and A. Torallba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, Vol. 42, No. 3, pp. 145–175, 2001.

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

[11] Y. Wang and S. Gong. Conditional random field for natural scene categorization. In *Proc. BMVC*, 2007.