Scene Gist: A Holistic Generative Model of Natural Image

Bolei Zhou[1] and Liqing Zhang[2]

[1] MOE-Microsoft Laboratory for Intelligent Computing and Intelligent Systems, and Department of Biomedical Engineering, Shanghai Jiao Tong University, No.800, Dongchuan Road, Shanghai, China
zhoubolei@gmail.com

[2] MOE-Microsoft Laboratory for Intelligent Computing and Intelligent Systems, and Department of Computer Science and Engineering, Shanghai Jiao Tong University, No.800, Dongchuan Road, Shanghai, China
zhang-lq@cs.sjtu.edu.cn

Abstract. This paper proposes a novel generative model for natural image representation and scene classification. Given a natural image, it is decomposed with learned holistic basis called *scene gist* components. This gist representation is a global and adaptive image descriptor, generatively including most essential information related to visual perception. Meanwhile prior knowledge for scene category is integrated in the generative model to interpret the newly input image. To validate the efficiency of the scene gist representation, a simple nonparametric scene classification algorithm is developed based on minimizing the scene reconstruction error. Finally comparison with other scene classification algorithm is given to show the higher performance of the proposed model.

Keywords: image representation, natural image statistics, scene classification.

1 Introduction

One of the extraordinary capabilities of the human visual system is its ability to rapidly group elements from a complex natural scene into the holistic and semantic percept. The studies of cognitive psychology have shown that human can recognize the category of natural scene in less than 150ms when a novel scene image is presented [1,2].

The gist of a novel scene is recognized at a single glance, independent of its spatial complexity. How is this remarkable feat accomplished? One prominent view of scene recognition is based on the idea that a scene is organized from a collection of objects. This notion depicts visual processing as a hierarchical organization of local modules of increasing complexity(gradually from edge to shape, object, then to global scene percept)[3]. On the other hand, psychological results suggest that a scene may be initially represented as a global entity and segmentation of region or object appears at a later stage after the formation of scene gist [2,4].

H. Zha, R.-i. Taniguchi, and S. Maybank (Eds.): ACCV 2009, Part II, LNCS 5995, pp. 395–404, 2010.

Motivated by the psychological evidence of scene gist components existed in the visual percept of natural image, we propose a novel holistic representation for natural image. The scene gist representation is an *global* image descriptor, adapted to natural image statistics to realize a most compressive encoding. Moreover, the marked performance in the scene classification task proves the superiority of this scene gist representation.

1.1 Related Works

Image representation or descriptor is of fundamental importance to the research of computer vision. It directly deals with organization of pixels, and plays a key role for extracting feature for later processing like feature classification and object recognition. Standard image descriptors, such as SIFT [5], bank of Gabor wavelet and image pyramid [6], have been widely used in feature extraction. In recent years, there have appeared another kind of *adapted* image descriptors drawing our attention. Early work on natural image statistics reveals that the natural image signal is highly non-gaussian and contains much information redundancy[7]. This leads to the headway in the Independent Component Analysis [8] or sparse coding [9] for natural image representation. The adapted basis share the similar response properties to the simple cell in the primary visual cortex. Summarily, the central concept of efficient coding is straightforward: if we want to efficiently capture the feature and reduce the redundancy, the image representation should reflect intrinsic structural properties of natural image [8].

Recent works on scene image modeling are mainly based on local approach, such as bag of words model like pLSA [10] and LDA [11], those methods are mainly through the hierarchical organization of local information to formulate the percepts of scenes. On the other hand, psychological results indicate that human visual system is more likely to rely on global approach to recognize the category of scenes [12]. Oliva et al[4] propose a global representation called Spatial Envelope. However, their holistic modeling of scene is only based on the amplitude of Fourier component coefficients for gray image, which ignores the influence of color information on the visual scene perception [13], and their model is not generative.

Fig. 1. Example images from 8 scene categories in the dataset [4] we work on

Fig.1 shows the example scene images from the scene dataset[4] our model is implemented on. The rest of the paper is organized as follows. Section 2 extends our scene gist model in detail. In Section 3, the scene gist representation is applied to the scene classification experiment, and comparison is given. Section 4 concludes this holistic generative model, and discusses its subspace property related to image manifold research.

2 Scene Gist Generative Model

Psychological study [1] has indicated that human visual system integrates enough information for the category of a scene in about 150ms. What is the underlying computational mechanism in the visual cortex? We consider this problem from the view of signal analysis and reconstruction. A discrete-time input signal $\mathbf{x}$ can be holistically viewed as an N×1 column vector in $\mathbb{R}^N$(we treat a scene image data by vectorizing it into a long one-dimensional vector). To sense and extract the gist of scenes for visual system is like a dynamic filtering and reconstruction process between input signal $\mathbf{x}\in \mathbb{R}^N$ and intrinsic signal representation $\mathbf{s}\in \mathbb{R}^M$, that is:

$$\mathbf{s} = \mathbf{W}\mathbf{x} \tag{1}$$

$$\hat{\mathbf{x}} = \mathbf{A}\mathbf{s} \tag{2}$$

where filter basis $\mathbf{W}$ is a projection from the image pixel space $\mathbb{R}^N$ to a representational space $\mathbb{R}^M$, the reconstruction basis $\mathbf{A}$(if $\mathbf{W}$ is full-rank square matrix, $\mathbf{A}$=$\mathbf{W}^{-1}$) recovers the image pixels from a given representational space, and $\hat{\mathbf{x}}$ is the reconstructed image signal.

Our generative model learns nearly-optimized holistic components $\mathbf{A}$ and $\mathbf{W}$ to represent the natural image. When the image signal is projected on these basis to enable M≪N, and to minimize $\|\mathbf{x} - \hat{\mathbf{x}}\|_2$, then, we term $\mathbf{s}$ as the scene gist representation, $\mathbf{A}$ and $\mathbf{W}$ as scene gist components. Before further expending our scene gist representation, two important issues on information redundancy should be discussed, which constitute a theoretical foundation for our model.

2.1 Information Redundancy Revisited

Efficient coding is a general framework under which many mechanisms of our visual processing can be interpreted. Barlow [14] first proposed the efficient coding hypothesis for the purpose of visual processing as to removes information redundancies in the sensory input. The research of natural image statistics [7] also indicates that there is a large amount of redundancy in the visual signal. To sum up, two kinds of information redundancy exist in visual signal processing:

Perceptual redundancy. Perceptual redundancy relates to the human prior knowledge about the visual world. In a sense, prior knowledge is the redundant

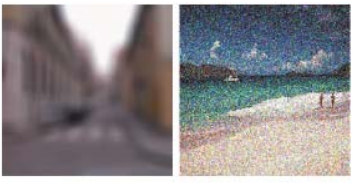

Fig. 2. Natural scene images are heavily blurred and noised respectively, we can recognize them because of prior knowledge for street and coast(better view in color edition)

information that should be suppressed by a coding system [14]. However, human visual system relies heavily on the prior knowledge to interpret the input signal, especially when the input signal is incomplete or noisy. Fig.2 shows two blurred and heavily noised natural scene images, we still can easily recognize them as street and coast because of our prior knowledge on the spatial layout of two scene categories.

Our scene gist model would integrate the prior knowledge in the learned gist components **A** and **W**, to help encode and interpret newly input signal.

Computational redundancy. The other kind is computational redundancy. One of the fundamental problems in computer vision is the curse of dimensionality. The high dimensionality of image weakens the performance of algorithms like object recognition [15]. Hopefully despite of the high dimensionality of image, there are two regularities of natural image that could be utilized. First, natural images are usually embedded in a relatively low dimensional subspace of images, and there are common spatial patterns along the ensemble of the same scene category [16]. Second, for the specific purpose of visual task such as the scene classification, the necessary scale of images might be low.

Fig.3 demonstrates those two regularities of natural image. Fig.3a is the chart of procedure we take: first downsample original 256×256 natural image to 32×32, then resize it to 256×256, so that information of the resized image is equal to the 32×32 image. And Fig.3b shows the result for example images from 8 different categories, orderly, coast, mountain, forest, open country, street, inside city, tall building and highway (refer to Fig. 1 for the example images with original scale). We can see that the low 32×32 scale preserves enough information for recognition. In Fig. 3d, we average 200 natural scene images from 8 different category ensembles separately, arranged in the same order as Fig.3b, we could still recognize the category of those image even the images are heavily blurred and averaged. This example illustrates that from image regularity we still could distinguish the certain semantic information such as the scene category. Moreover, when we average images from two or more different ensembles of scene category together, like in Fig.3c, there is no statistical regularities among the random averaged images, that is to say, the statistical regularities only exist within the

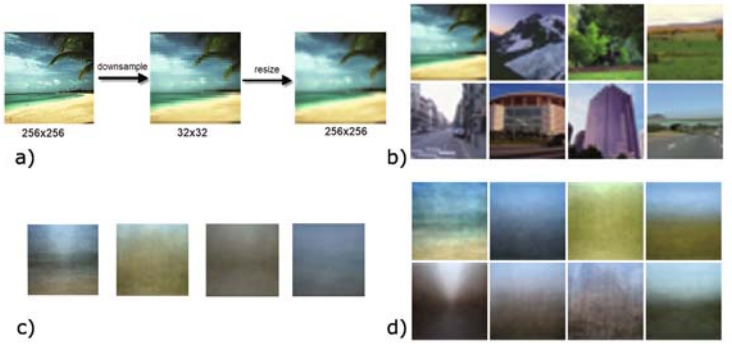

Fig. 3. (a)downsample procedure. (b)32×32 scale images still preserve enough information for scene recognition. (c)average images from different scene category, statistical regularities is not recognizable. (d)average images from same scene category, there exists clearly statistical regularities.(better view in color edition)

same scene category. Those statistical regularities are learned in our model as scene gist components.

2.2 Learning Gist Component

Since we have demonstrated that there are statistical regularities within scene images sampled from same category ensemble, our assumption is that if those regularities were learned as prior knowledge, we could construct highly efficient representation for natural scenes.

Let $\mathbf{X}=[\mathbf{x}^1,\mathbf{x}^2,\dots]$ be the matrix of images from the ensemble of one scene category, $\mathbf{W}=[\mathbf{w}_1,\mathbf{w}_2,\dots]$ be the filter basis, $\mathbf{W}_T$ be the first T rows of $\mathbf{W}$, and $\mathbf{A}_T$ be the first T columns of $\mathbf{A}$. Since given the number of T, the optimal $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{W}}$ should minimizes the reconstruction error,

$$\tilde{\mathbf{W}} = \arg\min_{\mathbf{W}} \|\mathbf{X} - \mathbf{A}_T(\mathbf{W}_T\mathbf{X})\|_F, \tag{3}$$

where $\|\cdot\|_F$ is the Frobenius norm, defined as: $\|\mathrm{Y}\|_F = \sqrt{\mathrm{Tr}(\mathrm{Y}^\top\mathrm{Y})}$. According to Eckart-Young theorem [17], the optimal solution to Eq .3 is the PCA basis of the sample matrix, that is, $\mathbf{W}$ is ensemble of the eigenvectors for sample covariance matrix, and $\mathbf{A}=\mathbf{W}^{-1}$. Given the threshold T, and a scene image $\mathbf{x}$, then

$$\hat{\mathbf{s}} = \mathbf{W}_T\mathbf{x} \tag{4}$$

we can reconstruct the scene image to minimize the reconstruction error by:

$$\hat{\mathbf{x}} = \mathbf{A}_T\hat{\mathbf{s}} \tag{5}$$

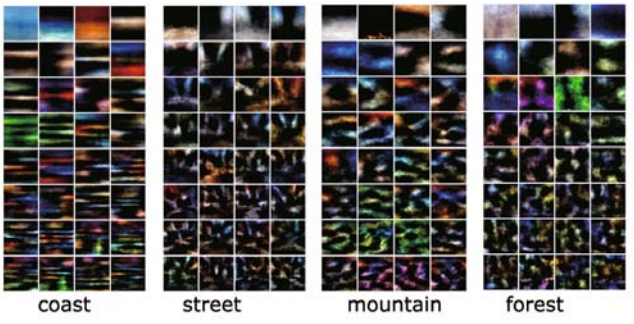

Fig. 4. First 32 Gist components $\mathbf{A}_T^i$ from 4 scene categories i

We learn gist components $\mathbf{A}_T$ and $\mathbf{W}_T$ respectively from 8 category ensembles based on PCA components.[1] In Fig.4, we show first 32 gist components $\mathbf{A}^i$ from 4 scene categories *i*. Obviously it reveals that the gist components have holistic spatial property for corresponding scene categories(refer the example scene images in Fig.1), we can see that the gist components are the holistic components for every scene category.

Because of the adaptive property and orthogonality of PCA basis, the energy of the image signal focuses on the first few principal basis. Fig.5 illustrates that two coast images are projected to the coast PCA basis, the amplitude of coefficients is focused on first few gist components, and the series of small images below the horizontal-axle are the reconstructed images by increasing the threshold T, from 5, 10, 20, 50, 100, 200, 500. Empirically with T around 200, the perceptual loss can be hardly perceived.

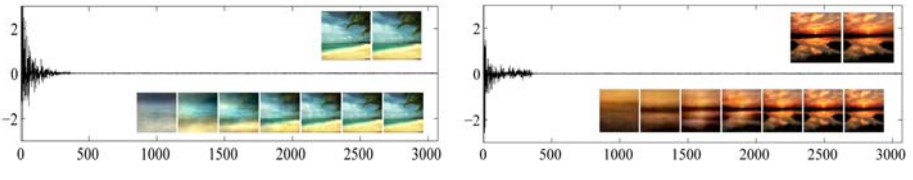

Fig. 5. Two coast images are projected to the 3072 PCA basis. The vertical axle is the coefficient of every PCA basis. The two images above the horizontal axle are the original image and the downsampled image. The series of scene image below is reconstructed by tuning threshold T, from 5,10,20,50,100,200 to 500.

[1] We learn PCA components respectively on down-sampled 32×32 RGB images from different category ensembles, so that learned PCA components is 32×32×3=3072 dimensional.

2.3 Discriminative Property of Gist Components

We have learned the gist component pair ($\mathbf{W}_T^i$,$\mathbf{A}_T^i$) for different scene categories i, then, what is the difference between ($\mathbf{W}_T^i$,$\mathbf{A}_T^i$) and ($\mathbf{W}_T^j$,$\mathbf{A}_T^j$), while i$\neq$ j? It is found that the gist components have discriminative sparse property between different categories: When one scene image is projected to the PCA components from the same scene category, the coefficients of gist components appear to be sparse(focused on first few components), otherwise are not. Fig. 6 gives two examples: a coast scene image is projected to gist components of mountain, and the other one is a mountain scene image projected to gist components of coast, in both conditions the coefficients of basis are not sparse. Then, threshing the coefficients would bring on the reconstructed image signals both great energy and perceptual loss, as shown in Fig. 6.

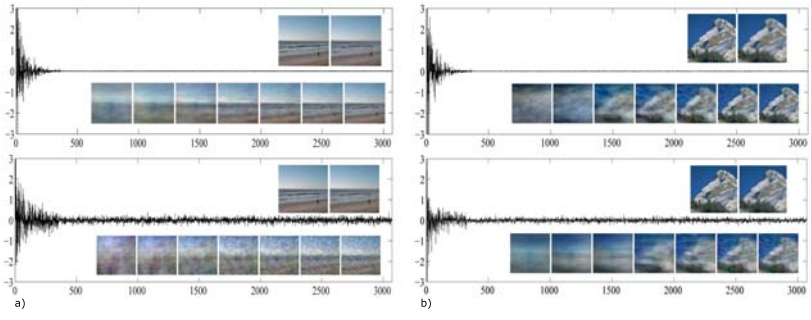

Fig. 6. a) and b) Threshing the gist component coefficients would lead to great signal energy and perceptual loss if scene image is projected to gist components of other scene category

In the following Experiment section, we apply this discriminative property of gist components to develop a simple nonparametric classification algorithm, based on minimizing the scene reconstruction loss.

3 Experiment

Scene classification task is to assign each test image to one category of scene, Fig.1 shows the example scene images from the dataset[4], which includes 8 categories of scenes. The performance is illustrated by a confusion table, and overall performance is measured by the average value of the diagonal entries of the confusion table, see Fig. 7 and Table 2.

3.1 Gist Subspace Classification Method

For each scene category i, we have learned gist components $\mathbf{W}_T^i$ and $\mathbf{A}_T^i$ on corresponding scene category ensembles. Then given one scene image $\mathbf{x}\in \mathbb{R}^N$,

$\mathbf{s}_i$=$\mathbf{W}_T^i\mathbf{x}$, $\hat{\mathbf{x}}_i$=$\mathbf{A}_T^i\mathbf{s}_i$. Relying on the discriminative property of gist components shown in Section 2.3, we can classify $\mathbf{x}$ by assigning it to the scene category i that minimizes the reconstruction error between $\mathbf{x}$ and $\hat{\mathbf{x}}_i$:

$$\min_i \varepsilon_i(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}_i\|_2 = \|\mathbf{x} - \mathbf{A}_T^i(\mathbf{W}_T^i\mathbf{x})\|_2, \tag{6}$$

Algorithm below summarizes the complete scene classification procedure.

Table 1. Gist Subspace classification algorithm

Algorithm : Scene classification

1:**Input**: k pair of gist components ($\mathbf{W}_T^1$,$\mathbf{A}_T^1$),($\mathbf{W}_T^2$,$\mathbf{A}_T^2$),. . . ,($\mathbf{W}_T^k$,$\mathbf{A}_k^1$) for k scene categories, where $\mathbf{W}_T^i \in \mathbb{R}^{T\times N}$,$\mathbf{A}_T^i \in \mathbb{R}^{N\times T}$. And a test scene image $\mathbf{x}\in \mathbb{R}^N$.
2:Compute the gist representation $\mathbf{s}_i$=$\mathbf{W}_T^i\mathbf{x}$, for every pair of components i.
3:Compute the reconstruction errors $\varepsilon_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{A}_T^i(\mathbf{W}_T^i\mathbf{x})\|_2$, for every pair of components i.
4:**Output**: identity($\mathbf{x}$)=$\arg\min_i \varepsilon_i(\mathbf{x})$

For the sake of comparison with other methods, we learn our gist components of each categories from downsampled 48×48 gray images, which ignore the influence of color information, then the threshold T is set empirically as 150, so that the learned gist components $\mathbf{W}_{150}^i \in \mathbb{R}^{150\times 2304}$,$\mathbf{A}_{150}^i \in \mathbb{R}^{2304\times 150}$. Experiment has been repeated ten times with different 200 (75% of each scene category ensemble) randomly selected images for learning scene gist components and the entire ensemble of images for test images for scene classification.

3.2 Result

Fig. 7 shows the confusion table for scene classification. The average performance is 88.75%. In Table 2, we compare our algorithm with other two methods[4,10],

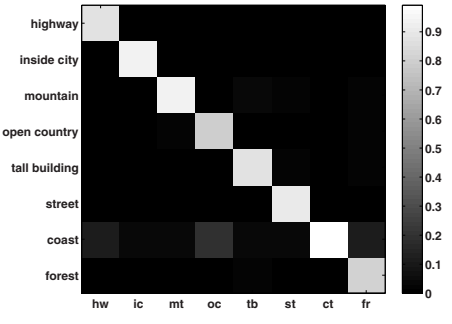

Fig. 7. Confusion table of scene classification. The average performance is 88.75%.

Table 2. Comparison of our algorithm with other scene classification methods

Method	Performance
Gist Subspace	88.75%
Spatial Envelope[4]	83.75%
pLSA[10]	86.65%

from which we can see our Gist Subspace classification method achieves the best performance. This experiment demonstrates the computational efficiency of the scene gist representation.

4 Discussion and Conclusion

The scene gist components $\mathbf{A}_T$ and $\mathbf{W}_T$ include the prior knowledge through the learning process. We conjecture those adapted components may act like the specialized neurons responsible for encoding spatial layout of the scene image in the visual cortex. Our work supports the assumption that there is close relationship between the natural image statistics and the neural representation[7].

On the other hand, there is in-depth implication beyond the discriminative property of gist components. The learning procedure for gist components is based on the Principal Component Analysis. Mathematically, PCA, as the classical linear technique for dimensionality reduction, is to discover the intrinsic structure of data lying on or near a *linear* low-dimensional subspace in the high-dimensional input space. So that different ensembles of scene category may be different subspaces $\mathbb{R}^M$ embedded in high dimensional space $\mathbb{R}^N$, where $M \ll N$, and the different scene gist components approximate the unit basis spanning each subspace. When a natural image is projected to the subspace of the same scene category, energy of signal concentrates on first few basis, otherwise it does not(refer to Fig. 6). That is why the Gist Subspace classification algorithm works. Our findings correspond with previous work on manifold learning [18] that the natural images are embedded in a low-dimensional manifold. Our further work would focus on generalizing a perceptual-meaningful manifold structure for natural images.

To conclude, we propose a holistic generative model of natural image. The scene classification experiment demonstrates its efficiency on representing natural scene image. Moreover, its inner connection to the more general modeling of natural image is discussed.

Acknowledgements

The work was supported by the Science and Technology Commission of Shanghai Municipality (Grant No. 08511501701), the National Basic Research Program of China (Grant No. 2005CB724301) and the Okawa Foundation Research Grant Program, Japan. The first author would like to dedicate this work to his passed-away dear Yan Zhang, and also thank Xiaodi Hou, Qiaochu Tang and Shiteng Suo for their valuable discussion and comments.

References

1. Fabre-Thorpe, M., Delorme, A.: A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. J. Cogn. Neurosci. 13(2), 171–180 (2001)
2. Oliva, A.: Gist of the scene. In: Itti, L., Rees, G., Tsotsos, J.K. (eds.) The Encyclopedia of Neurobiology of Attention, pp. 251–256. Elsevier, San Diego (2005)
3. Marr, D.: Vision: A computational investigation into the human representation and processing of visual information. W.H. Freeman, New York (1982)
4. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42, 145–175 (2001)
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
6. Forsyth, D.A., Ponce, J.: Computer Vision: A Modern Approach. Prentice Hall, Englewood Cliffs (2002)
7. Simoncelli, E.P., Olshausen, B.: Natural image statistics and neural representation. Annual Review of Neuroscience 24, 1193–1216 (2001)
8. Bell, A.J., Sejnowski, T.J.: The "independent components" of natural scenes are edge filters. Vision Res. 37(23), 3327–3338 (1997)
9. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381(6583), 607–609 (1996)
10. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)
11. Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proc. CVPR, pp. 524–531 (2005)
12. Schyns, P.G., Oliva, A.: From blobs to boundary edges: Evidence for time and spatial scale dependent scene recognition. Psychological Science 5, 195–200 (1994)
13. Castelhano, M.S., Henderson, J.M.: The influence of color on the perception of scene gist. Journal of experimental psychology, Human perception and performance 34(3), 660–675 (2008)
14. Barlow, H.: Redundancy reduction revisited. Network 12(3), 241–253 (2001)
15. Riesenhuber, M., Poggio, T.: Models of object recognition. Nature neuroscience 3(suppl.), 1199–1204 (2000)
16. Torralba, A., Oliva, A.: Statistics of natural image categories. Network (Bristol, England) 14(3), 391–412 (2003)
17. Deprettere, E.F. (ed.): SVD and signal processing: algorithms, applications and architectures. North-Holland Publishing Co., Amsterdam (1988)
18. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)