

C1 Units for Scene Classification

Dongjin Song¹ and Dacheng Tao^{2,1}

1. *Biometrics Research Centre, Dept. of Comp., The Hong Kong Polytechnic University*
2. *College of Computer Science, Zhejiang University, P.R. China*
{songdj2008, dacheng.tao}@gmail.com

Abstract

In this paper, we unify C1 units and the locality preserving projections (LPP) into the conventional gist model for scene classification. For the improved gist model, we first utilize the C1 units, intensity channel and color channel of color image to represent the color image with the high dimensional feature, then we project high dimensional samples to a low dimensional subspace via LPP to preserve both the local geometry and the discriminate information, and finally, we apply the nearest neighbour rule with the Euclidean distance for classification. Experimental results based on the USC scene database not only demonstrate that the proposed gist improves the classification accuracy around 7% but also reduce the testing cost around 50 times in comparing with the original gist model proposed by Siagian and Itti in TPAMI 2007.

1. Introduction

In recent years, a large number of approaches for scene classification have been developed and we can classify them into the following three categories:

1. *Low-level visual feature based schemes*, which represent scenes by global visual information [1], including colour, texture, and shape, have been successfully utilized in indoor/outdoor, city/landscape, and forest/mountain applications.
2. *Local feature based schemes* represent scene images with detected interest points [5][9] (or regions) based on some descriptors [2][9]. Local-global features [10] based schemes utilize both the global spatial information and the local descriptors of interest points (or regions) to represent scene images to achieve a robust classification.
3. *Biologically inspired feature based schemes* classify scenes by mimicking the process of visual cortex in recognition tasks. Recent reports

from both neuroscience and computer vision have demonstrated that biologically plausible features [11][13][14] are attractive in visual recognition.

Although Poggio and Bizzi [12] showed that C1 units correspond to complex cells in the visual cortex and they are effective for object recognition, C1 units ignores both the colour and intensity information of an image.

Although the gist feature [13] used for scene classification takes colour, intensity, and orientation information into account, the orientation information extracted by Gabor filters do not fully correspond to complex cells in the visual cortex. In addition, these features are labelled samples drawn from a low dimensional manifold and artificially embedded in a high dimensional ambient space, so both the principal components analysis (PCA) [6] and the independent components analysis (ICA) [4] utilized by Siagian and Itti [13] for dimensionality reduction are not suitable. This is because PCA and ICA do not consider both the non-Euclidean property of biological features and the sample label information.

In this paper, to effectively represent colour scene images, we unify C1 units together with both the colour and intensity information used in the scene classification scheme developed by Siagian and Itti. To efficiently represent the color images of scene in a low dimensional space, we apply the locality preserving projections (LPP) [7] with the supervised setting. Other methods [15] can be utilized here. LPP considers both the non-Euclidean property of the newly unified biological features to preserve the local geometry and the label information to preserve the discriminative information. Finally, we choose the nearest neighbour rule for classification. To test the improved gist model, we compare it with the scene classification algorithm proposed by Siagian and Itti [12] and show improvements in terms of effectiveness and efficiency.

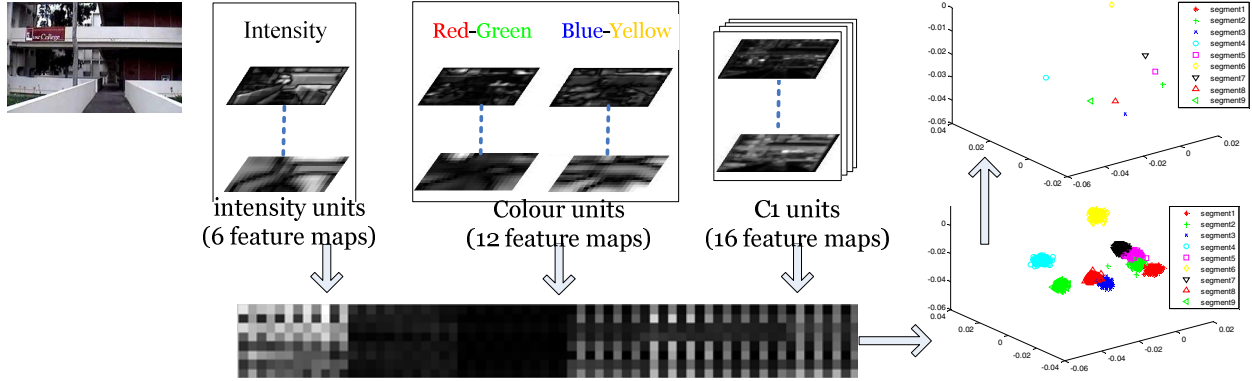


Figure 1. The system diagram of the improved gist model

2. The Improved Gist Model

The improved gist model contains three major components: the newly unified biologically inspired gist feature for colour scene representation, LPP for dimensionality reduction, and the nearest neighbour rule for classification.

2.1. Biologically inspired gist feature

In this framework, we unify C1 units, colour and intensity units as the biologically inspired feature for scene image representation. The C1 units are utilized here to replace the original orientation channel in the scene classification algorithm developed by Siagian and Itti [13]. This is because the orientation information extracted by Gabor filters does not fully correspond to complex cells in the visual cortex.

C1 units (16 feature maps): The C1 units correspond to complex cells in the visual cortex [14]. By using a maximum operation, C1 units pool over S1 units and only keep the max response of a local area of S1 units from the same orientation and scale. The S1 units correspond to simple cells in S1 layer of the visual cortex. Gabor functions are similar to the receptive field profiles in the mammalian cortical simple cells so they are utilized for feature extraction. The Gabor mother function is:

$$F(x, y) = \exp\left(-\left(x_0 + \gamma^2 y_0^2\right) / \left(2\delta^2\right)\right) \cos\left(2\pi x_0 / \lambda\right)$$

where $x_0 = x \cos \theta + y \sin \theta$, $y_0 = -x \sin \theta + y \cos \theta$, the range of x and y decides the scales of Gabor filters, and θ controls orientations. Here we arrange the S1 filters to form a pyramid of 8 scales and to span a range of sizes from 7×7 to 21×21 pixels in the steps of two pixels. We consider four orientations: 0° , 45° , 90° , and 135° , thus leading to 32 different S1 receptive field types total. Since the step between the

two scales is only two pixels, we can get precise information of the orientations. The size of local area is decided by the scale band index of S1 units. We use 4 bands from 8×8 to 14×14 grid with the step of two pixels. Because the maximum operation shows some tolerance for shift and size and it also highlights the orientation information, it provides us more precise and robust features.

Colour units and Intensity units (12 feature maps):

The colour units are inspired by the ‘‘colour double-opponent’’ system in cortex [8]. Neurons are excited by a colour (e.g., blue) and inhibited by another colour (e.g., yellow) in the centre of receptive field, so are neurons in the surround. Meanwhile, the intensity units correspond to the neurons of mammals which are sensitive to dark centres on bright surrounds or bright centres on dark surrounds [8]. Herein five channels are use: $R = r - (g + b) / 2$, $G = g - (r + b) / 2$,

$$B = b - (r + g) / 2, Y = r + g - 2(|r - g| + b), \text{ and}$$

$$I = (r + g + b) / 3.$$

For each channel (R, G, B, Y and I), dyadic Gaussian pyramids are also used for generating nine spatial scales with a ratio from 1:1 (level 0) to 1:256 (level 8). To get feature maps, the centre-surround operation is performed between centre levels ($c = 2, 3, 4$) and surround levels ($s = c + d$, with $d = 3, 4$), i.e., six feature maps are computed at levels of 2-5, 2-6, 3-6, 3-7, 4-7, and 4-8. Intensity units form an independent channel, and colour pairs form two new channels as red-green (R-G) and blue-yellow (B-Y) channels. The feature map, i.e., the across scales difference between two corresponding centre and surround maps is also obtained by firstly interpolated the surround map to the same size of relevant centre map and then subtracted by the relevant centre map point-by-point, i.e.,

$$RG(c, s) = \left| (R(c) - G(c)) - \text{Interp}_{s-c}(R(s) - G(s)) \right|,$$

$$BY(c, s) = \left| (B(c) - Y(c)) - \text{Interp}_{s-c}(B(s) - Y(s)) \right|,$$

and $I(c, s) = |I(c) - \text{Interp}_{s-c}I(s)|$.

Scene image representation: In this paper, a colour image is represented by 34 feature maps, and we decompose each feature map into 4 by 4 grid sub-regions. All sub-regions have identical length and width. Then, the mean value of each sub region is calculated for final representation, i.e., 16 mean values are utilized to represent each feature map and 544 values are obtained for image representation.

2.2. Locality preserving projections (LPP)

LPP is a linearization of Laplacian eigenmap, which preserves proximity relationships by manipulations on an undirected weighted graph and indicates neighbour relations of pairwise points. The supervised setting of LPP considers the class label information to improve the performance of LPP for classification tasks.

Because the dimensionality of the aforementioned biologically inspired gist feature is artificially high, it is essential to discover its intrinsic dimension by a suitable subspace selection algorithm. LPP is a good choice here because: 1) it preserves the local geometry of samples, 2) it considers the class label information under the supervised setting, and 3) it works efficiently. In this paper, we apply LPP to reduce the new gist feature dimensionality from 544 to 4. This step can improve both the effectiveness and the efficiency in comparing with the original gist model [13].

2.3. Nearest neighbour rule for classification

The nearest neighbour rule combined with the Euclidean measure is used for classification. It has four advantages: 1) it is a stable classifier so it usually achieves a good performance; 2) it is naturally nonlinear so nonlinearization (e.g., kernelization) is not required; 3) we only need to compare the L_2 (i.e., Euclidean) distance between with a test sample with the center of a class during classification for the LPP pre-processed samples; and 4) we can conduct the testing without training.

The detailed procedure is as following: 1) calculate centers of samples from all 9 classes independently; 2) calculate the distance between a test sample and each of the 9 centers; and 3) assign a label to the test sample based on the nearest neighbor rule, i.e., the test sample has the same label as the nearest center.

3. Performance Evaluation

Experiments have been conducted on the USC scene dataset [13], which contains 375 video clips from three

sites (ACB, AnF and FDF). Each site consists of 9 segments to describe different parts of a site. The baseline results are obtained from [13].

In this paper, the procedure for constructing both the training and testing video clips is identical to [13]. The only difference is that in the test stage, they constructed 4 trails by dividing the video clips into 4 groups to test the performance of their system on different lighting conditions. We combine all test video clips of 4 lighting conditions together for testing.

3.1. Step by step evaluation on the ACB site

Table 1. Step by Step Evaluation on Site ACB

Methods:	ACB
Baseline (Gist) (80 dimension) [13]	87.96%
NN Gist (4 dimension)	45.51%
C1 NN Gist (4 dimension)	73.84%
LPP NN Gist (4 dimension)	94.55%
The improved gist model (4 dimension)	95.37%

This experiment justifies the effectiveness of each component in the improved gist model independently, i.e., the newly unified biologically inspired gist feature, the supervised LPP for subspace selection, and the nearest neighbour rule for classification. The baseline for comparison is [13]. In the following tests, we uniformly down sample the ACB site in both the training and testing sets to reduce the time cost.

In baseline [13], feature dimension is reduced by PCA/ICA to 80. In “NN Gist”, we reduce the feature dimension to 4, because it is the minimal effective dimension both for the improvement of performance and the test speed. For other tests, we also reduce the feature dimension to 4 for fair comparisons. As shown in Table 1, both C1 units and LPP can significantly improve the classification accuracy. The combination can further increase the accuracy.

To justify the effectiveness of the nearest neighbour rule, we replace the three-layer neural network in the baseline with the nearest neighbour rule and term this procedure as “NN Gist”. We justify the effectiveness of C1 units by replacing the Gabor orientation channel in “NN Gist” with the C1 units and term this procedure as “C1 NN Gist”. We justify the proposed LPP in scene classification by replacing PCA/ICA stage in “NN Gist” with LPP and term this procedure as “LPP NN Gist”. Finally, we run the improved gist model.

3.2. Large scale dataset evaluation

We compare the improved gist model with [13] based on all the three sites of the USC dataset. Table 2 shows the average classification accuracies for all three sites in the USC dataset. In comparing with the baseline, the classification accuracy of each site has

been improved dramatically around 7 percent. Both in the baseline and our new framework, the classification results on AnF is lower than the other two sites due to the obstacles to classifier the vegetation dominated segments. For each site, the performance comparison is given in Figure 2. From left to right, the figure shows the evaluation in Sites ACB, AnF, and FDF, respectively. In each subfigure, we can see that the improved gist model significantly outperform the original gist model [13], i.e., the baseline. (In addition, the performance on the 6th segment of AnF is bad not only because the center of the 6th segment is very close to the center of the 7th segment in the training stage but also may because that the projected results of the 6th segment by LPP is more dispersive than other segments) In addition, the improved gist model is much more efficient than the baseline and improves the test speed around 50 times.

Table 2. The comprehensive classification results on USC scenes dataset.

	ACB	AnF	FDF
Baseline	87.96%	84.21%	88.62%
New framework	95.37%	88.32%	97.40%

4. Conclusion

In this paper, we proposed an improved gist model for natural scene classification. The new model is constructed by a newly unified biologically inspired gist feature, the locality preserving projections (LPP), and the nearest neighbour classification rule. In the new unified biologically inspired gist feature, we applied the C1 units, which correspond to complex cells in the visual cortex and can describe the visual orientation responses precisely. Although the dimensionality of the new unified biologically inspired gist feature is artificially high, its intrinsic dimension is very low. Therefore, LPP is applied for dimensionality reduction. In the LPP subspace, we show the Euclidean

property of samples. Based on thorough experiments, we show that the improved gist model can improve the classification accuracy around 7% and the testing speed around 50 times. Therefore, it is ready for real applications.

Acknowledgements

The work was supported by Hong Kong Research Grants Council General Research Fund (under project number 528708), the Competitive Research Grants for Newly Recruited Junior Academic Staff 2007-2008 with the Hong Kong Polytechnic University (under project number A-PC0A), and National Science Foundation of China (under project number 60703037).

References

- [1] M. Boutell, C. Brown, and J. Luo, "Review of the State of the Art in Semantic Scene Classification," Univ. Rochester, 2002.
- [2] J. G. Daugman, "Two-Dimensional Spectral Analysis of Cortical Receptive Field Profile," *Vision Res.*, 1980.
- [3] R. Epstein, A. Harris, D. Stanley, and N. Kanwisher, "The Parahippocampal Place Area: Perception, Encoding, or Memory Retrieval?" *Neuron*, 2000.
- [4] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley, 2001.
- [5] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," *Alvey Vision Conference*, 1988.
- [6] H. Hotelling, "Analysis of A Complex of Statistical Variables into Principal Components," *J. of Educational Psy.*, 1933.
- [7] X. He and P. Niyogi, "Locality Preserving Projections," *NIPS*.
- [8] L. Itti, C. Koch and E. Niebur, "A Model of Saliency-based Visual Attention for Rapid Scene Analysis," *IEEE T-PAMI*, vol. 20, pp. 1254-1259, 1998.
- [9] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *IJCV*, vol. 60, pp. 91-110, 2004.
- [10] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *IEEE CVPR*, 2006.
- [11] Aude Oliva and Antonio Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope" *IJCV*, vol. 42, no. 3, pp. 145-175, 2001.
- [12] T. Poggio and E. Bizzi, "Generalization in Vision and Motor Control," *Nature*, pp. 768-774, 2004.
- [13] C. Siagian and L. Itti, "Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention," *IEEE T-PAMI*, pp. 300-312, 2007.
- [14] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Object Recognition with Cortex-like mechanisms," *IEEE T-PAMI*, 2007.
- [15] X. Li, et al., "Discriminant Locally Linear Embedding with High Order Tensor Data," *IEEE T-SMC-B*, 2008.

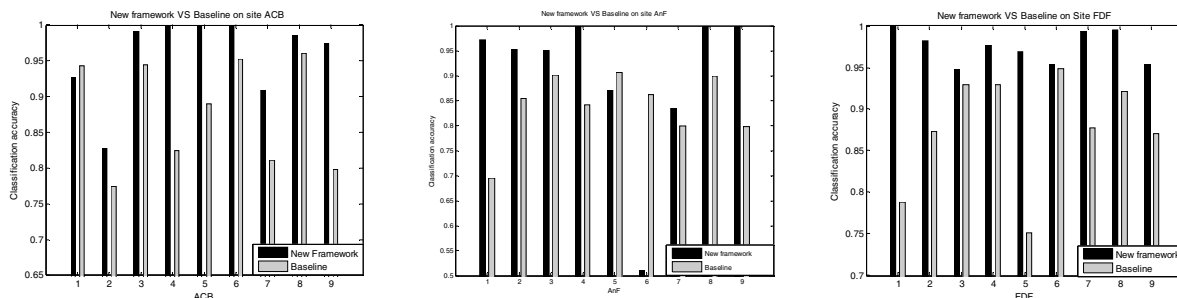


Figure 2. Large scale performance evaluation