

A Hierarchical GIST Model Embedding Multiple Biological Feasibilities for Scene Classification

Yina Han, and Guizhong Liu

School of Electronics and Information Engineering, Xian Jiaotong University, China
 ynhan@mailst.xjtu.edu.cn, liugz@xjtu.edu.cn

Abstract

We propose a hierarchical GIST model embedding multiple biological feasibilities for scene classification. In the perceptual layer, spatial layout of Gabor features are extracted in a bio-vision guided way: introducing diagnostic color information, tuning the orientations and scales of Gabor filters, as well as the spacial pooling size to a biological feasible value. In the conceptual layer, for the first time, we attempt to build a computational model for the biological conceptual GIST by kernel PCA based prototype representation, which is specific task orientated as biological GIST, and also in accordance with the unsupervised learning assumption in the primary visual cortex and prototype similarity based categorization in human cognition. Using around 200 dimensions, our model is shown to outperform existing GIST models, and to achieve state-of-the-art performances on four scene datasets.

1. Introduction

One of the most remarkable feats of human visual system is the ability to recognize a complex novel scene very rapidly and accurately [5]. Behavioral studies have shown that observers can rapidly capture the “gist” of the presented scenes without any attention [1, 12]. Hence building a reasonable computational model for GIST is critical for rapid and accurate scene classification. In general, GIST can be represented at both perceptual and conceptual levels [6]. Perceptual GIST refers to the spacial configuration of a scene built during perception [12], and has been modeled as average pooling of various biological relevant low-level features over nonoverlapping subregions arranged on a fixed grid [5, 6, 10]. However, from the biological feasible view, the above representation is inadequate, reflected in empirical filter parameters and pooling size.

Conceptual GIST refers to the inferred semantic information from the perceptual level, which has not been clearly proposed in existing models [5, 6, 10]. Though their adopted dimension reduction technique (such as PCA/ICA) does encode some statistics, it has not sufficiently reveal the semantically related statistics.

In this paper, we propose to improve GIST modeling using a hierarchical representation including both perceptual and conceptual layers. In the perceptual layer, we enrich the most widely used Gabor based model [6], which will be referred as OT-GIST, in several biological feasible aspects: introducing diagnostic color information, tuning the orientations and scales of Gabor filters, as well as the pooling size in a bio-vision guided way. Then, for the first time (to our knowledge), we attempt to build a computational model for the biological conceptual GIST by prototype based representation. Here, Kernel PCA is used to mine the semantic prototypes, which is specific task orientated as biological GIST [6], and also in accordance with the unsupervised learning assumption in the primary visual cortex [4] and prototype similarity based categorization in human cognition [7]. We demonstrate the efficacy of the proposed model through substantial evaluations.

2. The Hierarchical GIST Model

The hierarchical structure including both perceptual and conceptual layers is the inherent property of biological GIST [6, 12]. In our model, perceptual layer is modeled based on OT-GIST [6] as shown in Fig. 1. Specifically, the image is first decomposed by a set of Gabor filters with 4 scales and 8 orientations; then, the output magnitude of each filter is pooled by averaging over 16 nonoverlapping subregions divided by a 4×4 grid. Having the biological relevance of GIST in mind, we further tune the Gabor parameters and the pooling size, as well as introduce color information based on related bio-vision studies instead of empirical values, to

maximize the similarity between our representation and the responses of real cortex to visual information.

In the conceptual layer, which has not been clearly modeled yet, we attempt to model it by the Kernel PCA technology based on the following consideration: the specific task orientated property of biological GIST [6], the unsupervised learning assumption in the primary visual cortex [4], and the prototype similarity based categorization in human cognition [7].

3. Perceptual Layer

3.1. Color opponent features

Color information is a key diagnostic cue of scene categories [6]. Given a color image X , with r , g , and b being the red, green, and blue values, four broadly-tuned color channels are created: $R = r - (g + b)/2$, $G = g - (r + b)/2$, $B = b - (r + g)/2$ and $Y = r + g - 2(|r - g| + b)$. According to the color opponent theories [10], that neurons in cortex are excited by the differences between colors rather than each individual one, and such chromatic opponency exists for red/green (RG) and blue/yellow (BY) pairs, our model represents each color image as one intensity channel and two ‘‘color double-opponent’’ channels:

$$I(x, y) = \frac{1}{3}(r(x, y) + g(x, y) + b(x, y)) \quad (1)$$

$$RG(x, y) = \|R(x, y) - G(x, y)\| \quad (2)$$

$$BY(x, y) = \|B(x, y) - Y(x, y)\| \quad (3)$$

3.2. Biological feasible Gabor filters

The kernels of Gabor filters have been proven to be similar to the profile of cortical simple cell receptive fields [9], and can be defined as follows:

$$g_{\mu, \nu}(z) = \frac{\|k_{\mu, \nu}\|^2}{\sigma^2} e^{-\frac{\|k_{\mu, \nu}\|^2 \|z\|^2}{2\sigma^2}} [e^{ik_{\mu, \nu} \cdot z} - e^{-\frac{\sigma^2}{2}}] \quad (4)$$

where μ and ν define the orientation and scale, $z = (x, y)$, $\|\cdot\|$ denotes the norm operator, and the wave vector $k_{\mu, \nu}$ is defined as: $k_{\mu, \nu} = k_{\nu} e^{i\phi_{\mu}}$. k_{ν} is the modulation frequency. It has been shown that the frequency bandwidth of simple cells in the visual cortex is about one octave [9]. In order to ensure the decomposition of an image into aliasing negligible octaves, k_{ν} should be: $k_{\nu} = \pi/2^{\nu}$; $\nu \in \{0, \dots, N - 1\}$. The number of scales N is determined by the image size. In our model, the images are rescaled to 256×256 squares irrespective of their aspect ratio, thereby $N = \log_2(256) = 8$. $\phi_{\mu} = \mu\Delta\phi$; $\mu \in \{0, \dots, M - 1\}$ is the orientation

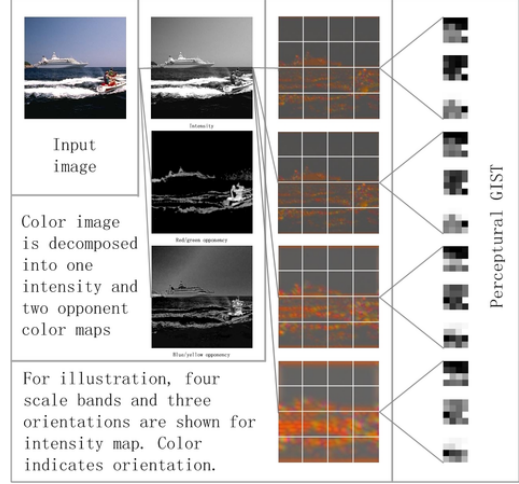


Figure 1. Perceptual layer overview

that filters tuned to. It has been shown that the resolution of the orientation tuning ability of the human visual system is as high as 5° . Here we set $\Delta\phi = 9^\circ$, and the number of orientations M is $180^\circ/9^\circ = 20$. In summary, we would use Gabor filters of eight different scales, $\nu \in \{0, \dots, 7\}$, and twenty orientations, $\mu \in \{0, \dots, 19\}$.

For the case of color image, I , together with RG and BY is decomposed by the proposed Gabor filters respectively:

$$O_{l, \mu, \nu}(x, y) = l(x, y) * g_{\mu, \nu}(x, y) \quad (5)$$

where $l \in \{I, RG, BY\}$, $\mu \in \{0, \dots, 19\}$, $\nu \in \{0, \dots, 7\}$, and $*$ denotes convolution operator, resulting in $20 \times 8 = 160$ gray maps, and $2 \times 20 \times 8 = 320$ color maps.

3.3. Resolution of spatial configuration

For spatial layout information which is diagnostic in scene categorization [6], each map $O_{l, \mu, \nu}(x, y)$; $l \in \{I, RG, BY\}$, $\mu \in \{0, \dots, 20\}$, $\nu \in \{0, \dots, 8\}$ is pooled by averaging on nonoverlapping subregions arranged on a $R \times R$ grid, which can be formalized as:

$$G_{l, \mu, \nu}(k, l) = \frac{256^2}{R^2} \sum_{x=\frac{256k}{R}}^{\frac{256(k+1)}{R}-1} \sum_{y=\frac{256l}{R}}^{\frac{256(l+1)}{R}-1} O_{l, \mu, \nu}(x, y) \quad (6)$$

where $k, l \in \{1, \dots, R\}$ are indices of the subregions. R controls the spatial resolution. Seeing that scene categories can be identified based on a resolution as low as 2 cycles/images in color, while 4 cycles/images in gray-scale [6], in our model we set $R = 4$ for color images and $R = 8$ for gray-scale images. Concatenating

all $G_{l,\mu,\nu}(k, l)$ forms the final ‘‘long’’ perceptual GIST vector, \mathcal{G}^p , namely 7680 dimensions for color images and 10240 dimensions for gray-scale images. The rich perceptual information is beneficial to further semantic inference in conceptual layer.

4. Conceptual Layer

4.1. Task-Wise Prototype

Since the content of GIST is constrained by categorization task [6], and categories are essentially defined by similarity to prototypes [7]. This requires to determine the most informative prototypes for each specific task as well as appropriate distance metric on the perceptual space, which is inherently non-linear and low dimensional [5]. Here we use KPCA [8] based method. Fig. 2 shows a toy example of the 2-D feature visualization (reduced by MDS) for two categorization tasks (indexed by rows) projected to four prototypes (indexed by columns), where colors represent similarities between features and prototypes.

4.2. KPCA based prototype representation

Given a specific categorization task $\mathcal{T} = \{\mathcal{G}_1^p, \mathcal{G}_2^p, \dots, \mathcal{G}_M^p\}$, where \mathcal{G}_i^p , $i \in \{1, \dots, M\}$ represents the perceptual GIST feature from the training set, M is the number of training data. Generally, the topology within \mathcal{T} is nonlinear. By nonlinearly mapping Φ , $\Phi(\mathcal{T})$ can be decomposed linearly by PCA. Taking the first l most informative eigenvectors as the prototypes, which best explain the variance in $\Phi(\mathcal{T})$, the first l principle components of any $\Phi(\mathcal{G}_x^p)$ naturally measure the similarity to the prototypes by Cosine Distance.

Within the KPCA methodology [8], the centered kernel matrix $\tilde{\mathbf{K}}$ corresponding to \mathcal{T} is first defined as:

$$\begin{aligned} \tilde{\mathbf{K}}(i, j) &= ((\Phi(\mathcal{G}_i^p) - \bar{\Phi}) \cdot (\Phi(\mathcal{G}_j^p) - \bar{\Phi})) \\ &= (\tilde{\Phi}(\mathcal{G}_i^p) \cdot \tilde{\Phi}(\mathcal{G}_j^p)) \\ &= \tilde{k}(\mathcal{G}_i^p, \mathcal{G}_j^p) \end{aligned} \quad (7)$$

where $\bar{\Phi} = \frac{1}{M} \sum_{i=1}^M \Phi(\mathcal{G}_i^p)$. It can easily be shown that $\tilde{\mathbf{K}} = \mathbf{H}\mathbf{K}\mathbf{H}$, where $\mathbf{H} = \mathbf{I} - \frac{1}{M} \mathbf{1}\mathbf{1}^t$ is the centering matrix, \mathbf{I} is the $M \times M$ identity matrix, $\mathbf{1} = (1, \dots, 1)^T$ is an $M \times 1$ vector. The symmetric matrix $\tilde{\mathbf{K}}$ is further decomposed as:

$$\tilde{\mathbf{K}} = \mathcal{A}\mathbf{\Lambda}\mathcal{A}^t \quad (8)$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_M)$ is a eigenvalue matrix with $(\lambda_1 \geq \lambda_2 \geq \dots \lambda_M)$. The column vector of \mathcal{A} , namely $\mathbf{a}^i = (a_1^i, \dots, a_N^i)^T$ is the eigenvector corresponding to λ_i .

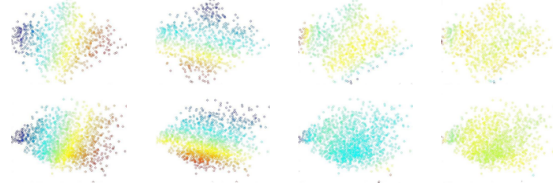


Figure 2. Toy example for two tasks

Since our concern is the similarity measure and the eigenvector matrix $\mathcal{V} = (\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_M)$ satisfies $\mathcal{V} = \mathcal{T}\mathbf{H}\mathcal{A}$, the l prototypes matrix for task $\Phi(\mathcal{T})$, namely $\mathcal{V}_l = (\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_l)$, does not need to be computed explicitly. The l -dimensional concept \mathcal{G}_x^c can be directly inferred from perceptual \mathcal{G}_x^p as:

$$\begin{aligned} \mathcal{G}_x^c &= \mathcal{V}_l^T \tilde{\Phi}(\mathcal{G}_x^p) = \mathcal{A}^T \mathbf{H} \mathcal{T}^T (\Phi(\mathcal{G}_x^p) - \bar{\Phi}) \\ &= \mathcal{A}^T \mathbf{H} (\mathbf{k}_{\mathcal{G}_x^p} - \frac{1}{l} \mathbf{K} \mathbf{1}) \end{aligned} \quad (9)$$

Specifically, each component $\mathcal{G}_{x,i}^c$, $i \in \{1, \dots, l\}$ of vector \mathcal{G}_x^c is computed as:

$$\mathcal{G}_{x,i}^c = (\mathcal{V}_i \cdot \Phi(\mathcal{G}_x^p)) = \sum_{k=1}^l a_k^i k(\mathcal{G}_x^p, \mathcal{G}_k^p) \quad (10)$$

where $\mathbf{k}_{\mathcal{G}_x^p} = (k(\mathcal{G}_x^p, \mathcal{G}_1^p) k(\mathcal{G}_x^p, \mathcal{G}_2^p) \dots k(\mathcal{G}_x^p, \mathcal{G}_l^p))^T$, and the polynomial kernel with fractional power, $k(x, y) = (x \cdot y)^d$, $0 < d < 1$, is adopted in our model.

5. Experiments

The proposed GIST model is evaluated for both scene and object classification tasks. All the experiments are repeated ten times with randomly selected training and testing images. The final result is reported as the mean and standard deviation of each individual one. Categorization is done with a rbf SVM.

First we carry out three groups of experiments on four scene datasets: OT [5], VS [11], FP [1] and LSP [3] following the same protocol as [5, 11, 1, 3] that is 100 images per class for training (except for sky/clouds in VS with 30 for training) and the rest for testing. The first group is to compare our model with other GIST models: OT1 [5], OT [6], RM [12], and SI [10], implemented as in [10]. The second group is to compare our model with two popular BoF based models: pLSA-BoF and SPM-BoF, implemented as in [3]. The third group is to compare our model directly with the reported results in [5, 1, 3, 11] on the same dataset.

The detailed comparison results are shown in Table 1. Observe that the performances of OT1, OT, RM and SI are similar, which is consistent with [10], while

Table 1. Comparison of scene categorization performances

Model	OT	OT-4N	OT-4M	VS	FP	LSP
OT1-Gist	82.8 ± 0.7	85.1 ± 1.0	90.2 ± 2.1	56.4 ± 5.3	74.8 ± 1.1	71.5 ± 0.7
OT-Gist	83.3 ± 0.8	86.2 ± 1.1	91.6 ± 1.9	59.8 ± 4.8	76.6 ± 0.9	72.0 ± 0.6
RM-Gist	81.5 ± 0.8	83.3 ± 0.8	88.3 ± 1.2	57.2 ± 4.5	71.3 ± 0.7	70.2 ± 0.7
SI-Gist	82.3 ± 0.7	84.6 ± 1.0	88.7 ± 1.1	58.1 ± 3.1	73.4 ± 0.7	71.2 ± 0.6
pLSA-BoF	70.4 ± 0.8	71.0 ± 0.7	81.8 ± 1.1	61.1 ± 3.4	65.9 ± 1.4	63.3 ± 1.2
SPM-BoF	87.5 ± 0.7	86.9 ± 0.8	91.9 ± 0.6	66.6 ± 7.2	81.7 ± 0.4	76.8 ± 0.5
Authors	83.7 [5]	89.0 [5]	89.0 [5]	74.1 [11]	65.2 [1]	81.4 [3]
Ours	88.7 ± 0.7	91.3 ± 1.2	95.7 ± 1.3	74.4 ± 4.3	81.5 ± 0.7	76.6 ± 0.4

our model achieves better performances, especially for VS dataset, which is more ambiguous in semantics [11]. This demonstrates the “gist” of a scene can be better captured by our model. Compared to BoF models, our implementation of SPM-BoF is not able to reproduce the results reported in [3] probably due to some engineering problem. Following our own baseline, our model achieves similar performances to the state-of-the-art SPM-BoF. But as a global representation, our model is more compact (with around 200 dimensions) than local regions based SPM-BoF model (with 4200 dimensions). Moreover, the rbf SVM used for our model is more efficient than histogram intersection SVM for SPM-BoF model.

Table 2. Comparison on Caltech-101

Model	15 training images/cat.	30 training images/cat.
Serre et al. [9]	35	42
Mutch & Lowe [4]	51	56
Wolf et al. [13]	51.2	–
Huang et al. [2]	49.8 ± 1.2	–
OT-GIST	48.7 ± 1.0	56.1 ± 1.1
Ours	53.7 ± 0.8	60.6 ± 1.1

To further compare with other biologically-motivated models with their directly reported results, we evaluate our model on the Caltech-101 [1] dataset with 101 object categories and 1 background. Following the standard setup, training on 15 and 30 images per category and testing on the rest. As shown in Table. 2, our model achieves slightly better performances.

The above results have proved the validity of introducing statistical conceptual layer for modeling GIST. However, performances on dataset with indoor scenes, and especially on Caltech-101, are not satisfactory. Therefore, our next step is to build conceptual GIST with more discriminative properties.

References

- [1] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. *CVPR*, pages 524–531, 2005.
- [2] Y. Huang, K. Huang, L. Wang, D. Tao, T. Tan, and X. Li. Enhanced biologically inspired model. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2008.
- [3] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *CVPR*, pages 2169–2178, 2006.
- [4] J. Mutch and D. G. Lowe. Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, 80(1):45–57, 2008.
- [5] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, May 2001.
- [6] A. Oliva and A. Torralba. Building the gist of a scene: the role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.
- [7] E. Rosch. Natural categories. *Cognitive Psychology*, 4:328–350, 1973.
- [8] B. Schölkopf, A. J. Smola, and K. R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.
- [9] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:411–426, 2007.
- [10] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):300–312, Feb 2007.
- [11] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, 72(2):133–157, 2007.
- [12] L. L. Walker and J. Malik. When is scene recognition just texture recognition. *Vision Research*, 44:2301–2311, 2003.
- [13] L. Wolf, S. M. Bileschi, and E. Meyers. Perception strategies in hierarchical vision systems. *CVPR*, pages 2153–2160, 2006.