Modeling Manifold Ways of Scene Perception

Mengyuan Zhu[1], Bolei Zhou[2]

[1]MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems, Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Department of Information Engineering, The Chinese University of Hong Kong
zhumengyuan@126.com, zhoubolei@gmail.com

Abstract. In this paper, under the efficient coding theory we propose a computational model to explore the *intrinsic* dimensionality of scene perception. This model is hierarchically constructed according to the information pathway of visual cortex: By pooling together the activity of local low-level feature detectors across a large regions of the visual fields, we build the population feature representation as the statistical summary of the input image. Then, a large amount of population feature representations of scene images are embedded unsupervisedly into a low-dimensional space called perceptual manifold. Further analysis on the perceptual manifold reveals the topographic properties that 1) scene images which share similar perceptual similarity stay nearby in the manifold space, and 2) dimensions of the space could describe the perceptual continuous changes in the spatial layout of scenes, representing the degree of naturalness, openness, etc. Moreover, scene classification task is implemented to validate the topographic properties of the perceptual manifold space.

Key words: visual perception, hierarchical model, scene classification

1 Introduction

One of the fundamental issues in computational neuroscience concerns how information is encoded and represented by the neural architecture of the brain. As for the neural computation of visual signal, from photoreceptor of retina to the unified scene perception in high-level cortex, the visual processing of input signal is hierarchically constructed. With multilayers of wiring in visual cortex, neural response gradually achieve generalization over the input signal [5].

The neurophysiologic studies [17] indicate that through the hierarchical processing of information in the visual cortex, the extremely high-dimensional input signal is gradually represented by fewer active neurons, which is believed to achieve efficient coding [9]. One of the concepts of efficient coding theory is that with the metabolic constraints the visual cortex relies on the environmental statistics to encode and represent the visual signal [14], that is, a group of neurons should encode information as compactly as possible, in order to most effectively utilize the available computing resources. Mathematically, this is expressed as to maximize the information that neural responses provide about the

visual environment. This theory has been applied to derive efficient codes for natural images and to explain a wide range of response properties of neurons in the visual cortex [14].

On the other hand, studies on the natural image statistics suggest that the natural images are usually embedded in a relatively low dimensional manifold of image space, and there is a large amount of information redundancy within the natural images [15]. According to the efficient coding theory, it could be implied that neural system would be efficiently adapted to reduce information redundancy and extract the underlying low meaningful dimensionality of natural image to form the unified scene perception, in spite of the extremely high-dimensional raw sensory input from the retina [17]. From the functional viewpoint, the hierarchical architecture of visual system could be considered as the multilayered process of nonlinear dimensionality reduction, gradually resulting in sparser and more efficient response in higher-level neurons [13].

In this paper, from the functional view of neural architecture we propose a computational model for visual scene processing. This model termed Perceptual Manifold is *data-drivenly* constructed on the natural image statistics and *hierarchically layered*: By pooling together the activity of local low-level feature detectors across large regions of the visual fields, we build the population feature representation as the statistical summary of the scene image. Then, thousands of population feature representations of scene images are extracted, and to be mapped *unsupervised* along into a low dimensional space called perceptual manifold space. Analysis of this perceptual manifold reveals that scene images which share similar perceptual similarity stay nearby in the manifold space, and the dimensions of the manifold could describe the perceptual continuous changes in the spatial layout of scenes. In addition, scene classification task is performed to validate the topographic property of the perceptual manifold space.

2 Perceptual Manifold Model

A hierarchical model called Perceptual Manifold is proposed. The architecture of the proposed model includes four cortex-like layers of computation: 1) local sparse feature encoding, 2) local higher-order feature encoding, 3)population feature encoding and 4) perceptual manifold embedding (refer to Fig.1).

Different layers of computation abstract their own representations of input signals, which accounts for the different hierarchical stages of neuronal response to visual stimulus [17]. It is assumed that through the hierarchial layers of computation, the dimensionality of image representation is gradually reduced, that is, $M < N < K$, leading to more general organization of image representation.

2.1 Local Sparse Feature Encoding

Experimental studies have shown that the receptive fields of simple cells in the primary visual cortex produce a sparse representation of input signal [14]. Efficient coding method [1] assumes that the image patch is transformed by a set of

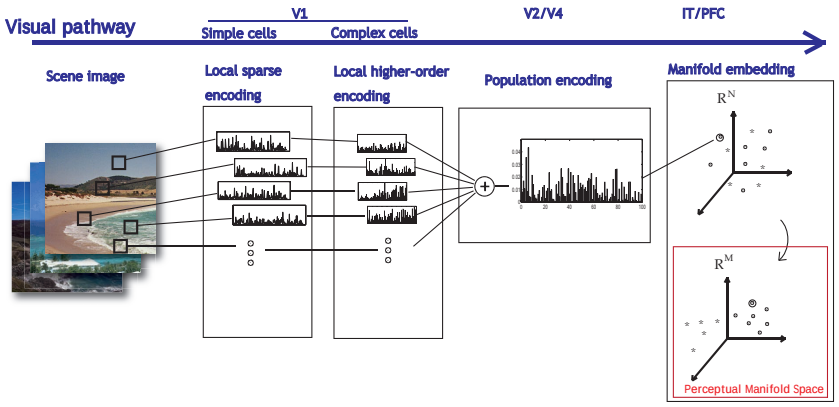

Fig. 1. Schematic diagram of the Perceptual Manifold model. There are four hierarchical layers of sensory computation to form the final perceptual space, which resemble the information pathway in visual cortex.

linear filters w_i to output response u_i. In matrix form,

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

Or equivalently in terms of generative process, $\mathbf{x}$=$\mathbf{A}\mathbf{u}$=$\mathbf{W}^{-1}\mathbf{u}$. Then, the filter response u_i are assumed to be statistically independent,

$$p(\mathbf{u}) = \prod_i p(u_i) \tag{2}$$

where $p(u_i) \propto \exp(-|u_i|)$. Let $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ...\mathbf{w}_K]$ be the learned filter functions, and K is the number of basis functions, so that the dim($\mathbf{u}$)=K. Fig. 2 shows a subset of the filter functions $\mathbf{w}$. These filter functions resemble the receptive field properties of simple cells, i.e., they are spatially localized, oriented and band-pass in different spatial frequency bands. A vectorized image patch $\mathbf{x}$ can be decomposed into those statistically independent bases, in which only a small portion of bases are activated at one time. They are used as the first layer of local feature extraction in our framework, so that representation of local image patches in first layer is $\mathbf{u}$. This layer of computation resembles the simple cells in V1 [16].

2.2 Local Higher-order Feature Encoding

Higher-level visual neurons encode statistical variations that characterize local image regions, these results provide a functional explanation for nonlinear effects in complex cells [3]. Thus the coefficients of local basis $\mathbf{A}$ are further assumed to follow a generalized Gaussian distribution,

$$p(u) = \mathcal{N}(0, \lambda, q) = z\exp(-|\frac{u}{\lambda}|^q), \tag{3}$$

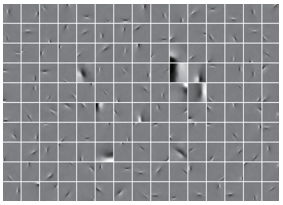

Fig. 2. A subset of filter basis **W** trained from natural image patches.

where $z = q/(2\lambda\Gamma[1/q])$ is the normalizing constant, and mostly λ=1. In [2], the variance λ value is assumed to be generated by variance basis as follows:

$$\log\boldsymbol{\lambda} = \mathbf{B}\mathbf{v}, \tag{4}$$

where $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, ...\mathbf{b}_N]$ are variance basis functions trained from thousands of natural image patches [1], N is the number of variance basis functions and $\mathbf{v}$ is the higher-order representation of local image patches, so that the $\dim(\mathbf{v})$=N, where N<K. The transformation from sparse representation $\mathbf{u}$ to the higher-order representation $\mathbf{v}$ is determined by maximizing the posterior distribution for a given $\mathbf{u}$,

$$\hat{\mathbf{v}} = \text{argmax}p(\mathbf{u}|\mathbf{B}, \mathbf{v})p(\mathbf{v}). \tag{5}$$

where $p(\mathbf{v})$=$\prod_j p(v_j)$ and $p(v_j) \propto \exp(-|v_j|)$. In the simulation, $\hat{\mathbf{v}}$ is derived by gradient ascent [2]. For simplicity, here the nonlinear transformation of $\mathbf{u}$ to $\mathbf{v}$ is denoted as a general function f, where $\mathbf{v}$=$f(\mathbf{u})$. This is second layer of computation in our framework, and through it $\mathbf{v} = [v_1, v_2, ..., v_N]^\top$ becomes the representation of local image patches. This layer of computation resembles the complex cells in V1 [3].

2.3 Population Feature Encoding

The neurophysiologic study [17] suggests that on the population level in extrastriate visual areas II(V2) and IV(V4), a normalized pooling mechanism might be used to extract the global response of the stimulus. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, ...]$ denote the sample matrix, where $\mathbf{x}_n$ is the vectorized image patch sampled from one scene image. After the first two layers of local encoding: $\mathbf{u}_n = \mathbf{W}\mathbf{x}_n$ and $\mathbf{v}_n = f(\mathbf{u}_n)$, the population feature component for the i^{th} feature of $\mathbf{v}_n$ is,

$$p_i = \frac{\sum_n([\mathbf{v}_n]_i)^2}{\sum_{i=1}^N \sum_n([\mathbf{v}_n]_i)^2} \tag{6}$$

where $[\mathbf{v}_n]_i$ indicates the i^{th} element of the vector $\mathbf{v}_n$. Thus, $\mathbf{p} = [p_1, p_2, ..., p_N]^\top$ indicates the normalized population feature response of scene image, which accounts for the holistic representation of scene image in the third layer of computation. This layer of computation resembles the population coding in V2/V4.

[1] Refer to [2] for visualization of the variance basis

2.4 Perceptual Space Embedding

To explore the intrinsic dimensionality of scene perception, the Local Linear Embedding [10] is applied as the method of nonlinear dimensionality reduction to a large amount of the population feature responses of different scenes:

First step: compute the weight ω_{ij} that best linearly reconstructs $\mathbf{p}_i$ from its neighbor $\mathbf{p}_j$, minimizing:

$$\varepsilon(\omega) = \sum_i |\mathbf{p}_i - \sum_j \omega_{ij}\mathbf{p}_j|^2 \tag{7}$$

Second step: compute the low-dimensional embedding vectors $\mathbf{q}_i$ best reconstructed by ω_{ij}, minimizing:

$$\phi(\mathbf{q}) = \sum_i |\mathbf{q}_i - \sum_j \omega_{ij}\mathbf{q}_j|^2 \tag{8}$$

The resulting embedding space $\mathbb{R}^M$ is called perceptual manifold space, as the final layer of computation in our architecture. And $\mathbf{q} = [q_1, q_2, ..., q_M]^\top$ is the representation of scene perception for a specific image, in which $\dim(\mathbf{q})=M$ and $M < N < K$. This layer of computation is believed to exist in the inferotemporal cortex(IT) or the prefrontal cortex(PFC), which involve in forming the perception of objects and scenes [12].

The implementation and analysis of the perceptual manifold space are presented in the following experiment section.

3 Experiments

For the training of image basis $\mathbf{A}$(or $\mathbf{W}$) and variance basis $\mathbf{B}$, 150000 20×20 gray image patches from a standard set of ten 512×512 natural images [8] are extracted. The number of 20×20 filter basis $\mathbf{W}$ is 400, equivalently K=400, and the number of variance basis function $\mathbf{B}$ is limited to 100, equivalently N=100. For manifold embedding layer, the dataset of scene images used here comes from [4], which contains 3890 images from 13 semantic categories of natural scenes, like coast and forest, etc. All 3890 images are normalized to 128×128 pixels before layers of computation. The dimensionality of manifold space M is tuned empirically as 15, so that the manifold embedding is $\mathbb{R}^{100} \rightarrow \mathbb{R}^{15}$. Thus, through the whole process of multilayered computations, the representative space changes as $\mathbb{R}^{400} \rightarrow \mathbb{R}^{100} \rightarrow \mathbb{R}^{15}$.

In the following part, the topographic properties of the perceptual space are analyzed at first. Then to validate those properties of perceptual space, scene classification task is performed in this perceptual space.

3.1 Perceptual Space Analysis

For visualization of the perceptual space, data points of scene images from four scene categories are described by the first three principal component coordinates

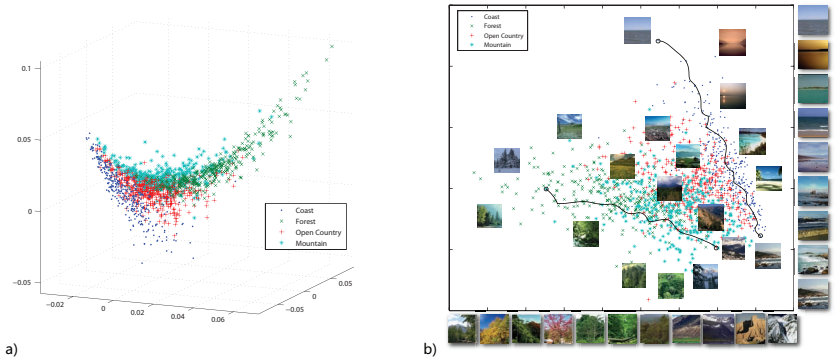

Fig. 3. Data points of scene images from four categories are visualized by the first three coordinates and first two coordinates of perceptual space. **a)** Clustering and nonlinear geometric property of the data points in the 2D perceptual space. **b)** Representative scenes are shown next to the corresponding data points in different parts of the 3D perceptual space. The bottom and right sets of images correspond to points along the two pathes(linked by solid line), illustrating particular perceptual changes in scene images.

(Fig. 3a) and first two principal component coordinates of perceptual space (Fig. 3b). In Fig. 3a, we can see that there are clustering and nonlinear geometric properties among the pool of data points. In Fig. 3b, representative scenes are shown next to the corresponding data points in different regions of the perceptual space. The bottom and right sets of images correspond to points along the two pathes(linked by solid line), illustrating particular degree of perceptual changes within the scene images.

As we can see, the topographic properties of perceptual space are related to the perceptual dimensions(degree of naturalness, openness, expansion, etc) supervised trained in [7], which represent the dominant spatial dimensions of a scene. Our Perceptual Manifold model is layered in a bottom-up way to find the *intrinsic* dimensionality of scene perception. Our finding supports the viewpoint that the shape of a scene could be described by a few perceptual dimensions [6]. Moreover, the topographic properties give further implication that human visual system might be adapted to both *extract* and *integrate* the lower perceptual dimensions to form the holistic scene perception. After that, scene classification task is performed to validate the topographic properties of perceptual manifold space.

3.2 Scene Classification

Scene classification task is to classify each image from testing set into one category of scenes. The dataset contains 13 categories of scenes, 100 images from each scene class are as training set, and the rest are as testing set. Both testing

set and training set of images have been embedded in perceptual manifold space before the classification, so that all images are represented as 15 dimensional feature vectors. A 13-way linear SVM classifier is trained on the training set, then it is applied to classify images from testing set.

The average accuracy of classification for our method is 68.9%. The average accuracy for baseline methods LDA[4] is 64.0%. Even though the Perceptual Manifold model is not designed specifically for the scene classification task, our model achieves good performance. Scene classification task well validates the topographic properties of the perceptual manifold space. And it further reveals that there is neural correlation between the perceptual space and semantic space in human cognitive process [6].

4 Discussion

4.1 On the Dimensionality of Perceptual Manifold Space

The choice of reduced dimensionality M for manifold space is theoretically and experimentally important. First, the local linear embedding [10] itself relies on the amount of observation samples and the setting of reduced dimensionality to search for intrinsically low-dimensional structures embedded nonlinearly in high-dimensional observations. And the theoretic analysis of this point could be found in the studies on manifold learning [11], which goes beyond the scope of this paper. Second, the topographical properties of perceptual manifold are influenced by the dimensionality M value, here we illustrate that by the correlation between dimensionality M value and performance of scene classification, as shown in Fig. 4. From that we can see, the dimensionality to engender the topographic properties of perceptual manifold is rather low. For the limitation of paper length, more theoretic analysis would be included in our further work.

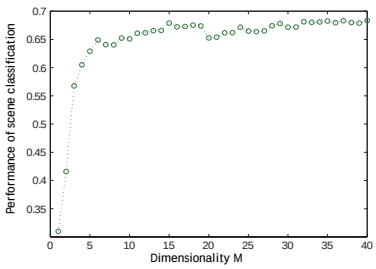

Fig. 4. The correlation between dimensionality M and scene classification performance.

5 Conclusion

A novel hierarchical model of scene perception termed Perceptual Manifold is introduced in this paper. Through the cortex-like layers of computation, dimensionality of input visual signals is gradually reduced, and it finally leads to the formation of perceptual manifold space. In this perceptual manifold space, there exist topographic properties that **1)** data points of perceptual similarly scene

images stay nearby in this perceptual manifold space and **2)** dimensions of the perceptual space could describe the meaningful continuous changes in the spatial layout of scene images. Scene classification task is performed to validate the topographic properties of the perceptual manifold space.

6 Acknowledgement

The work was supported by the National Natural Science Foundation of China (Grant No. 90920014) and the NSFC-JSPS International Cooperation Program (Grant No. 61111140019).

References

1. A. J. Bell and T. J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Res*, 1997.
2. Y. Karklin and M. S. Lewicki. A hierarchical bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Comp.*, 17(2):397–423, February 2005.
3. Y. Karklin and M. S. Lewicki. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457:83–86, January 2009.
4. F.-F. Li and P. Perona.
5. J. J. Nassi and E. M. Callaway. Parallel processing strategies of the primate visual system. *Nat Rev Neurosci*, 10(5):360–372, May 2009.
6. A. Oliva. Gist of the scene. In L. Itti, G. Rees, and J. K. Tsotsos, editors, *The Encyclopedia of Neurobiology of Attention*, pages 251–256. Elsevier, San Diego, CA, 2005.
7. A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 2001.
8. B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996.
9. B. A. Olshausen and D. J. Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487, August 2004.
10. S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000.
11. L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha, and D. D. Lee. Spectral methods for dimensionality reduction. *Semisupervised Learning. MIT Press: Cambridge, MA*, 2006.
12. T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans Pattern Anal Mach Intell*, 29(3):411–426, 2007.
13. S. H. Seung and D. D. Lee. Cognition: The manifold ways of perception. *Science*, 2000.
14. E. P. Simoncelli and B. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 2001.
15. A. Srivastava, A. B. Lee, E. P. Simoncelli, and S. c. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 2003.
16. J. v. van Hateren and A. v. d. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Royal Soc. Lond. B*, 265:359–366, 1998.
17. M. Weliky, J. Fiser, R. H. Hunt, and D. N. Wagner. Coding of natural scenes in primary visual cortex. *Neuron*, 2003.